African Journal of Biotechnology Vol. 9 (4), pp. 438-442, 25 January, 2010
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB09.1024
ISSN 1684–5315 © 2010 Academic Journals

Full Length Research Paper

The impact of information quantity and strength of relationship between training set and validation set on accuracy of genomic estimated breeding values

M. Saatchi*, S. R. Miraei-Ashtiani, A. Nejati Javaremi, M. Moradi-Shahrebabak and H. Mehrabani-Yeghaneh

Department of Animal Science, University College of Agriculture and Natural Resource, University of Tehran, Karaj, Iran.

Accepted 20 August, 2009

Recent advances in genomic selection are a revolution in animal breeding. A genome consisting 10 chromosomes each with 100 cM in length with 100 equally spaced markers (1 cM) were simulated. After 50 generations of random mating in a finite population (N_e = 100) in order to create sufficient linkage disequilibrium, population was expanded to two different population sizes of 500 and 1000. This structure was conserved until generation 59. Only females of generations 51 to 58 had phenotypic records and were included in the training set. The generation 59 was assumed as juveniles without any phenotypic records (validation set). Two measures of heritability (h^2 = 0.1 and h^2 = 0.5) were considered. Each simulation was replicated 10 times and results were averaged across replications. The results showed that using individuals of more recent generations in training set led to higher accuracy of genomic estimated breeding values (GEBVs) than individuals from more distant generations. However, increase in the amount of phenotypic records in training set even from individuals of older generations will increase accuracy of GEBVs. Number of phenotypic records in training set was shown to have important role in accuracy of GEBVs especially for low heritability traits.

Key words: Genomic selection, GEBVs, training set, validation set, generation distance.

INTRODUCTION

Traditional methods of genetic evaluation depend on phenotypic and pedigree information. This elongates the time needed for availability of the phenotypic records in most farm species such as dairy cattle. This leads to reduced genetic improvement rate due to longer generation interval (Schaeffer, 2006). One of the most important goals of modern breeding programs is to utilize genotypic information at DNA level to increase genetic progress by reducing the generation interval and improving the accuracy of estimated breeding values. Currently, it is possible to genotype individuals for tens of thousands of single nucleotide polymorphisms (SNP) loci with gene chip array (Goddard and Hayes, 2007). These markers can be used to genomic estimated breeding values (GEBVs) as proposed by Meuwissen et al. (2001).

*Corresponding author. E-mail: saatchi.mahdi@gmail.com. Tel: +989121915461. Fax: +982612246752.

In genomic selection method, marker effects are estimated first with a reference data set containing individuals with marker genotypes and trait phenotypes (training set); then GEBV of juveniles with marker genotype and no trait phenotype (validation set) would be the sum of the corresponding marker effects over all loci. GEBVs can be estimated as soon as DNA can be obtained, which reduces generation interval and increases genetic progress (König et al., 2009). Schaeffer (2006) compared a strategy that utilizes genomic estimated breeding values with a traditional progeny testing strategy under a typical Canadian-like dairy cattle situation. He concluded that costs of proving bulls were reduced by 92% and genetic change was increased by a factor of 2, due to reduced generation interval. Another advantage of genomic selection for animal breeding is its ability to control inbreeding (Daetwyler et al., 2007). One of the most important factors that affect accuracy of GEBVs is number of phenotypic records in the training set. Meuwissen et al. (2001) compared effect of three different

Table 1. Population structure and parameters used in the simulation.

Parameter	Value
Number of chromosome	10
Number of SNP markers per chromosome	100
Genome length	1000 cM
Marker distance (cM)	1
Number of QTL	50
QTL effects	Normal distribution
Recombination	Haldane map function
Number of generation	59
Generation 1 to 50, create LD	50 male,50 female
Generation 51 to 59	500 and 1000 individuals
Training set	Females of generation 51 to 58
Validation set	Females of generation 59
Heritability	0.1 and 0.5

numbers of phenotypic records in training set on accuracy of estimated GEBV in validation set. They used three different statistical methods and showed that in all statistical methods accuracy of estimated GEBVs will increase by increasing the number of phenotypic records in the training set. Similar results were reported by other studies (e. g. Calus and Veerkamp, 2007; Muir, 2007). Goddard (2009) developed a formula that calculates expectation of the accuracy of GEBV by deterministic model. In this formula accuracy of GEBVs estimation has direct relationship to number of individuals in the training set.

In practice, one of the limiting factors to increase number of genotyped individuals in training set is the cost of genotyping. Moreover, individuals with phenotypic records may belong to different generations. For example, in dairy cattle, bulls' DNA exists from several generations. The main question is whether and how generation distance between individuals in training set and validation set affect the accuracy of GEBVs. On the other hand, heritability of the trait of interest has been shown to affect the accuracy of genomic selection (Calus and Veerkamp, 2007; Goddard, 2009).

Therefore, the objective of this paper was to investigate the effect of (i) number of phenotype records in training set, and (ii) strength of relationship between training set and validation set, on accuracy of genomic estimated breeding values. Different scenarios were simulated by including different number of individuals of different generations in training set. Also, traits with low and high heritability were compared.

MATERIALS AND METHODS

Simulation

A genome consisting 10 chromosomes each with 100 cM in length with 100 equally spaced SNP's (every 1 cM) and a total number of 50 QTL's (that scattered on chromosomes randomly) was generated for each individual. This small genome size was chosen to decrease calculation time. Both SNP and QTL were assumed to be biallelic with equal initial allelic frequencies. For these simulations, gene substitution effects for each QTL were assigned randomly from a standard normal distribution, a ~ N (0, 1). Fifty QTLs covered total genetic variance and individual true breeding values. Only additive genetic effect was considered.

An effective population size of 100 individuals was simulated, of which 50 were male and 50 were female. This structure was followed by 50 generations of random mating, implying that each individual had on average two offspring in the next generation (variance of family size was two). The paternal and maternal haplotypes for each individual were generated based on Haldane mapping function to generate recombinant haplotypes. Sires and dams in the base generation were assumed to be unrelated. Fifty generations of random mating were practiced to generate sufficient linkage disequilibrium (LD) between loci. Two LD measurements, r^2 and D', were used to calculate LD in generation 50, as average of all syntenic marker loci. Markers with a minor allele frequency of < 0.05 were discarded.

After the first 50 generations, 9 additional generations (51 to 59) were simulated. Population was expanded to obtain intended population size in generation 51. Population size was constant until generation 59. Two different population sizes (500 and 1000 individuals with equal number of males and females) in each of the last 9 generations were simulated. Only females of generations 51 through 58 (250 or 500 females in each generation) had trait phenotype and, thus, were included in the training set according to different scenarios. To investigate the effect of generation distance between training set and validation set on accuracy of GEBVs, females from different generations (distant and recent generations) were included in training set. The validation data contained individuals from generation 59. For simplification, no selection was considered. Population structure and parameters used in the simulation are presented in Table 1.

Models

For calculation of GEBV, the simple mixed model estimation method suggested by Meuwissen et al. (2001) was used assuming that all loci explained an equal amount of variance (That is, the variance per locus $\sigma^2 m$, is $\sigma^2 m = \sigma^2 a / n$ where $\sigma^2 a$ is the total

Table 2. Mean (±SE) of homozygosity and linkage disequilibrium (D' and r^2) between markers in generation 50.

Parameter	Mean ± SE
D'	0.59 ± 0.002
r^2	0.17 ± 0.001
Homozygosity	0.61 ± 0.001

genetic variance and *n* is the number of marker loci). Meuwissen et al. (2001) termed this method as best linear unbiased prediction (BLUP) method. This assumption (equal variance over all loci) is clearly unrealistic. Genetic variance may not be equal across markers, for example, major genes may exist on some chromosomes. However, BLUP is quick, easy to program and as Meuwissen et al. (2001) demonstrated, BLUP performs almost as well as the much more advanced and time consuming Bayesian methods.

The model to estimate the marker effects was

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zm} + \mathbf{e} \quad (1)$$

Where, y is the vector of observations, b is the vector of means, m is the vector of random marker effects, e is the vector of random residual effects, X and Z are coefficient matrices. Row elements of Z consist of 0, 1 and 2 for marker genotype. Then, the expected value of y is 1μ and the variance of y is $\mathbf{V(y)} = \mathbf{ZIZ'}\sigma^2_m + \mathbf{I}\sigma^2_e$, (assuming equal variance for each marker).

The mixed model equation (MME) for BLUP is

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{I}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \quad (2)$$

We considered $\alpha = \sigma^2 e / \sigma^2 m$ as Meuwissen et al. (2001). After obtaining solution for vector m, GEBV was estimated as

$$\mathrm{GEBV}_i = \mathbf{Z}_i \hat{\mathbf{m}}_i \quad (3)$$

The genetic variance was determined as variance of true breeding values among individuals in generation 51 through 58.

As haplotyping would increase computation time with little or no gain in accuracy at high marker density (Calus et al., 2008), we used genotypes rather than haplotypes. Different scenarios were compared by the accuracy of the estimated genomic breeding values for individuals without a phenotypic record (generation 59). Accuracies were calculated as the correlation between simulated and estimated breeding values. Each simulated data set was replicated 10 times and results were averaged across replicates.

RESULTS

Simulated data

After fifty generations of random mating in finite population (N_e = 100), considered linkage disequilibrium between markers was created. Two linkage disequilibrium measurement, D' and r^2, were used to measure the amount of LD between pairs of markers in the individuals of generation 50. Results are presented in Table 2.

Trait with high heritability

Results from different scenarios are presented in Table 3. As it is shown in Table 3, in all scenarios with equal number of phenotypic records, when individuals in training set belong to generations close to validation set, accuracy of GEBV is higher than when they belong to generations far from validation set. For example, when data from 1000 individuals of generations 57 and 58 (generations close to validation set) were used in training set, the accuracy of GEBV were higher than when an equal number (1000 individuals) of individuals from generations 51 and 52 were used in training set (0.634 versus 0.555). Similar results are obtained for similar comparisons between generations.

Even with less number of phenotypic records from individuals of more recent generations had higher accuracy of GEBV than more phenotypic records of individuals from more distant generations. For example 2000 records of individuals from generation 55 to 58 in the training set led to higher accuracy of GEBV (0.706) in comparison with 3000 records of individuals from older generations of 51 through 56 (0.695).

These results indicate that accuracy of GEBVs in validation set increase by adding number of individuals in the training set even if these individuals belong to older generations. For example including information of individuals from generations 51 and 52 to training set with information of individuals from generations 53 to 58 had a small increasing effect on accuracy of GEBV from 0.737 to 0.749.

Traits with low heritability

Results from different scenarios are presented in Table 4. As it is seen in Table 4, results for traits with low heritability were similar to the results of traits with high heritability. It means that using individuals from recent generations in the training set has positive effect on accuracy in comparison to the information from individuals of older generations. However, for the low heritability traits, number of records had higher impact than the distance between generations of training set and validation set. For example, using 2000 individuals from generations 55 to 58 in training set was not better than using 3000 individuals from generations 53 to 58 (in comparison with high heritability traits). Results for low heritability trait confirmed that accuracy of GEBV will increase by increasing the number of phenotypic records in the training set even if new records are from more distant generations.

Table 3. Accuracy of GEBVs in validation set according to different number of individuals (250 or 500 in each generation) from different generations in training set for trait with $h^2 = 0.5$.

Training set	51 and 52	57 and 58	51 - 54	55 - 58	51 - 56	53 - 58	51 - 58
Number of individuals	1000	1000	2000	2000	3000	3000	4000
Accuracy*	0.555	0.634	0.639	0.706	0.695	0.737	0.749
SE**	0.028	0.013	0.019	0.012	0.015	0.010	0.010
Number of individuals	500	500	1000	1000	1500	1500	2000
Accuracy*	0.464	0.573	0.552	0.630	0.613	0.666	0.689
SE**	0.034	0.029	0.037	0.026	0.030	0.025	0.025

*Correlation coefficient of True Breeding Value (TBV) and GEBV.
**Standard error.

Table 4. Accuracy of GEBVs in validation set according to different number of individuals (250 or 500 in each generation) from different generations in training set for trait with $h^2 = 0.1$.

Training set	51 and 52	57 and 58	51 - 54	55 - 58	51 - 56	53 - 58	51 - 58
Number of individuals	1000	1000	2000	2000	3000	3000	4000
Accuracy*	0.267	0.335	0.347	0.398	0.405	0.431	0.460
SE**	0.028	0.030	0.034	0.030	0.031	0.032	0.033
Number of individuals	500	500	1000	1000	1500	1500	2000
Accuracy*	0.246	0.310	0.323	0.374	0.381	0.419	0.448
SE**	0.036	0.029	0.038	0.022	0.033	0.023	0.026

*Correlation coefficient of True Breeding Value (TBV) and GEBV.
**Standard error.

DISCUSSION

Accuracy of estimated breeding values based on marker distance and number of phenotypic records in the training set is similar to other studies. Meuwissen et al. (2001) used equally spaced (1 cM between adjacent markers) markers in a simulation study with different sizes of phenotypic records for a trait with heritability of 0.5 in the training set to estimate genomic EBV's for the validation set. Based on BLUP evaluation they obtained GEBVs with accuracies of 0.579, 0.659 and 0.732 for the 500, 1000 and 2200 records in the training set, respectively. These measures of accuracy are close to our results of 0.573 for 500 records; 0.639 and 0.630 for 1000 records; and 0.706 and 0.689 for 2000 records. Also, as shown in our study, increasing the number of phenotypic records in the training set leads to increased measures of accuracy in the validation set. It is expected that higher amount of information leads to better estimates of marker effects. Solberg et al. (2006) in a simulation study used 1000 phenotypic records in the training set for a trait with heritability of 0.5 and genomic structure similar to our study. Instead they used Bayes-B method to estimate marker effects. Accuracy of GEBVs of the progeny in the training set considered as validation set was 0.663. Small advantage (in comparison with 0.634 when generations 57 and 58 were included in the training set) may be due to the statistical method used in their evaluation. Advantage of Bayesian method to BLUP evaluation has been shown in some studies (Meuwissen et al., 2001; Hayes et al., 2009).

Our results show that using information of generations closer to the validation set leads to more accurate GEBVs compared to using information of more distant generations in the training set. This may be due to: (i) weaker relationship between individuals of training set and validation set, (ii) higher amounts of recombination and changes in haplotypes structure, and (iii) reduction in LD between markers and QTLs, through higher number of generations between individuals of training set and validation set. Habier et al. (2007) indicated that genomic selection uses genomic relationship among individuals and LD between markers and QTL to improve accuracy of GEBVs. They showed increase in accuracy of evaluation is partly due to using genomic relationship information among individuals. In an earlier study Nejati-Javaremi et al. (1997) replaced pedigree based relationship by marker-based total allelic relationship and documented its impact on reducing prediction error variance, hence, increasing accuracy of evaluation.

In genomic selection, effects of QTL are distributed among adjacent marker loci. In other words, some degrees of co-linearity exist among neighboring markers. With increasing distance between generations in training set and generation of validation set, because of higher amounts of recombination occurrence and because of

change in haplotypes, the accuracy of evaluation decreases. Meuwissen et al. (2001) used Bayesian method and obtained accuracy of 0.848 for the GEBVs of individuals in the training set. They showed that the accuracy of GEBVs decreased to 0.804, 0.768, 0.758, 0.734 and 0.718 in 5 subsequent generations, respectively. Muir (2007) showed that after several generations following estimation of marker effect the accuracy of GEBVs reduces and these effects should be re-estimated.

Our study also shows that GEBVs for the traits with higher heritability is more accurate than GEBVs for the traits of lower heritability. Similar results have been reported elsewhere (Willumsen et al., 2009; Goddard, 2009).

As expected, the importance of the number of phenotypic records is shown to be more important for the traits with lower heritability. Hayes et al. (2009) used the formula of Goddard (2009) to estimate the accuracy of GEBVs and indicated that although higher number of records is required to reach at a certain level of accuracy for a trait with low heritability, but this relationship is not linear. In general it is safe to comment that increasing the number of phenotypic records in the training set leads to higher accuracy of evaluation even if these records belong to more distant generations. However, if cost of genotyping is an issue it may be recommended to use genotypes (and phenotypic information) of individuals from more recent generations.

REFERENCES

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.

Calus MPL, Meuwissen THE, Deroos APW, Veerkamp RF (2008). Accuracy of genomic selection using different methods to define haplotype. Genetics, 178: 553-561.

Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007). Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124: 369-376.

Goddard ME (2009). Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136:245-257.

Goddard ME, Hayes BJ (2007). Genomic selection. J. Anim. Breed. Genet. 124: 323-330.

Habier D, Fernando RL, Dekkers JCM (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389-2397.

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

König S, Simianer H, Willam A (2009). Economic evaluation of genomic breeding programs. J. Dairy Sci. 92: 382-391.

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: 342-55.

Nejati-Javaremi A, Smith C, Gibson JP (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75: 1738-1745.

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

Solberg TR, Sonesson A, Wooliams J, Meuwissen THE (2006). Genomic selection using different marker types and density. 8th World Congress on Genet. Appl. to Livestock Prod. August 13-18, Belo Horizonte, Brazil.

Willumsen TM, Janss L, Lund MS (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. 126: 3-13.