Effect of marker density and trait heritability on the accuracy of genomic prediction over three generations

N. Gorgani Firozjah, H. Atashi[*], M. Dadpasand and M. J. Zamiri

Department of Animal Science, Shiraz University, Shiraz, Iran.

* Corresponding author, E-mail address: hadiatashi@gmail.com

Abstract   The aim of this study was to determine the effect of marker density, level of heritability, number of QTLs, and size of training set on the genomic accuracy over three generations. Thereby, a trait was simulated with heritability of 0.10, 0.25 or 0.40. For each animal, a genome with 20 chromosomes, 1 Morgan each, was simulated. Different marker densities (2000, 4000 and 6000 markers) and 400 and 600 randomly distributed QTLs were simulated. Marker density, size of training set, and heritability level significantly affected the genomic accuracy ($P < 0.05$). Increasing the marker density from 4000 to 6000 did not affect the genomic accuracy, likewise there was no difference between genomic accuracy of the first, and second validation sets (generations 8 and 9). The results showed that 4000 markers may be appropriate for genomic evaluation, and that the estimated marker effects can be used for at least two subsequent generations although the marker effects should be re-estimated for the third generation.

Keywords: marker density, heritability, genomic breeding value

Introduction

Genomic selection was first introduced in 1998 (Visscher and Haley, 1998), then its methods and principles were presented by Meuwissen et al. (2001). Genomic evaluation is carried out in two steps: (a) estimating the effect of marker on a particular trait in the training set (population having phenotypic records); (b) predicting the genomic breeding values for animals in the validation set (population without phenotypic records). This method relies on computing genomic estimated breeding values (GEBV) using a large number of single nucleotide polymorphism (SNP) markers. The accuracy of genomic predictions depends on the level of linkage disequilibrium (LD) between the markers and QTLs, number of animals in the reference population, heritability of the trait, and distribution of QTLs (Hayes et al., 2009). Marker density panel is one of the most important factors affecting accuracy of genomic prediction (Solberg et al., 2008; Habier et al., 2009; Meuwissen, 2009; Weigel et al., 2009). In genomic selection, all genetic variance is explained by the markers which are scattered in the whole genome. Genomic selection can reduce the costs of genetic evaluation and increase genetic improvement by reducing the generation interval and increasing the accuracy of selection (Scheffer, 2006; Hayes et al., 2009). Even though the advancement in molecular technology makes it possible to use these high-density SNP markers in genomic selection, the high cost of genotyping a large population prevents the broader implementation of genomic selection in many livestock species including the native cattle breeds. The solution is to reduce genomic evaluation costs through reducing the number of markers, reducing the number of animals that should be genotype (reducing the size of the reference population), using the estimated markers effects for two or more generations, and finally a combination of the items expressed. The present simulation study was carried out to investigate the effect of number of markers (2000, 4000, and 6000), level of heritability (0.10, 0.25, and 0.40), number of QTLs (400 and 600), and size of the training set (1000 and 2000) on the genomic accuracy over three generations.

Materials and Methods

Simulation

The populations were simulated using the QMSim software (Sargolzaei and Schenkel 2009) based on forward-in-time process. The base population consisted of 1000 unrelated animals (500 males and 500 females). The individuals were randomly selected as parents and randomly mated for 1000 discrete generations. Then, the population size was gradually decreased from 1000 individuals in generation 1001 to 300 individuals in gene-

ration 2020 to create the linkage disequilibrium (LD) in the historical population. The number of individuals of both sexes were remained the same and the mating system was based on random union of gametes randomly sampled from male and female gametes pools.

In the next step, a population (the expanded population) was generated using randomly selected 300 founders (150 males and 150 females) from the last generation of the historical population. To enlarge the population, 10 generations were simulated with five offspring per dam and an exponential growth of the number of dams. The mating system was based on the random union of gametes with no selection. Subsequently, 40 males and 2000 females from the last generation of the expanded population were randomly mated to generate another 10 generations. Individuals of the training set (1000 and 2000 individuals) were randomly selected from the 6th and 7th generations. The validation set were all individuals from the generations eight, nine and ten.

For each animal, a genome with 20 chromosomes, 1 Morgan each, was simulated. Three marker density panels (2000, 4000, and 6000 markers) were simulated with equal allelic frequency in the first generation of the historical population. The markers were randomly scattered across the genome, and none of the markers directly affected the trait. A total of 400 and 600 QTLs were randomly distributed among the markers. Effects at the QTLs were sampled from a gamma distribution with shape 0.4.The mutation rate of the markers and QTLs was assumed 2.5×10^{-5} per locus per generation (Solberg et al., 2008).

A trait with heritability of 0.10, 0.25 or 0.40 and phenotypic variance of 1.0 was simulated. The true breeding value (TBV) of each individual was equal to the sum of the QTL allele substitution effects, assuming only additive QTL effects. Phenotypes were generated by adding residuals, randomly drawn from a normal distribution with mean equal to zero, to the TBVs. For all scenarios, 10 replicates were simulated.

Statistical model

Prediction of marker effects

The following ridge regression BLUP model which is a popular approach for estimating the marker effects in genomic evaluations (Meuwissen et al., 2001; Habier et al., 2007) was used to estimate SNP effects using a program written in the R software (R Development Core Team, 2011).

$$y = 1\mu + X`g + e$$

where, y is the vector of phenotypes values, μ is the overall mean, X is the matrix of marker genotypes for

each animal (the SNP genotypes were coded as the number of copies of one SNP allele, i.e. 0, 1 or 2), g is the vector of marker effects, and e is a vector of random errors with distribution of $N(0, \sigma_e^2)$. No additional information, such as marker location, polygenic effects, or pedigree was used in the model.The mixed model equation used to estimate the markers effects was:

$$\begin{bmatrix} 1_n`1_n & 1_n`X \\ X`1_n & X`X + I\lambda \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1_n`y \\ X`y \end{bmatrix}$$

where, X is the matrix of marker genotypes for each animal, I is an identity matrix (total number of marker $\times$ total number of markers), λ is the ridge regression factor $= \frac{\sigma_e^2}{\sigma_g^2}$, σ_e^2 is the residual variance, σ_g^2 is variance common to each marker effect $= \frac{\sigma_a^2}{2\sum_{i=1}^{n} p_i q_i}$ and n is the total number of markers (Meuwissen et al., 2001; Habier et al., 2007).

Once the marker effects were estimated, the GEBV was computed for the animals in the validation set as: **GEBV = $X\hat{g}$**, in which, X is the matrix of marker genotypes for each animal, and $\hat{g}$ is the vector of estimated marker effects. The accuracy of GEBV was calculated as the correlation between GEBVs and TBVs.

The effects of heritability levels, marker density panels, and the number of QTLs on the accuracy of genomic predictions over three subsequent generations were evaluated using PROC GLM, and the average accuracies of GEBV were compared using the least squares means (LSM) procedure at $P<0.05$ (SAS, 2003).

Linkage Disequilibrium

The linkage disequilibrium (LD) measure r^2, square of the correlation of alleles at two loci, was used for measuring LD (Hill and Robertson 1968).

$$r^2 = \frac{D^2}{f(A) \, f(a) \, f(B) \, f(b)}$$

where, $D = f(AB) - f(A) \, f(B)$, and f(AB), f(A), f(a), f(B), f(b) are observed frequencies of haplotypes AB and of alleles A, a, B, b, respectively.

Results and Discussion

On average, 756 (37.8%) of 2000, 1506 (37.7%) of 4000, and 2260 (37.7%) of 6000 markers were polymorphic (MAF > 0.05) in the recent generations, while the corresponding values for QTLs were 160.50 (40.1%) out of 400 and 241.60 (40.3%) out of 600.

In this study, the genomic accuracy, the correlation between TBVs and GEBVs, for different marker densities

Effect of marker density and trait heritability on genomic prediction

Table 1. The estimated genomic accuracy (SE) for a training set of 1,000, three marker density panels, and two levels of QTLs over three subsequent generations

Generation	8		9		10	
QTL	400	600	400	600	400	600
Marker density panels						
2000	34.3(0.78)^{ab,B}	36.7(0.78)^{a,A}	31.3(0.78)^{bc,B}	33.3(0.78)^{abc,A}	30.1(0.78)^{c,B}	31.1(0.78)^{bc,A}
4000	38.0(0.78)^{ab,AB}	39.3(0.78)^{a,A}	34.8(0.78)^{bc,AB}	35.9(0.78)^{abc,A}	33.0(0.78)^{c,AB}	33.8(0.78)^{c,A}
6000	39.6(0.78)^{a,A}	40.5(0.78)^{a,A}	37.2(0.78)^{ab,A}	37.0(0.78)^{ab,A}	34.4(0.78)^{b,A}	34.7(0.78)^{b,A}

Different capital letters indicate significant differences (P < 0.05) within columns. Different small letters indicate significant differences (P < 0.05) within rows.

Table 2. The estimated genomic accuracy (SE) for a training set of 2,000, three marker density panels, and two levels of QTLs over three subsequent generations

Generation	8		9		10	
QTL	400	600	400	600	400	600
Marker density panels						
2000	42.1(0.73)^{ab,B}	43.8(0.73)^{a,B}	39.0(0.73)^{bc,B}	40.4(0.73)^{abc,B}	37.1(0.73)^{c,B}	37.7(0.73)^{c,B}
4000	46.1(0.73)^{ab,A}	46.8(0.73)^{a,AB}	43.2(0.73)^{bc,A}	43.9(0.73)^{abc,AB}	41.4(0.73)^{c,A}	41.6(0.73)^{c,A}
6000	48.4(0.73)^{ab,A}	48.7(0.73)^{a,A}	45.0(0.73)^{bc,A}	45.1(0.73)^{abc,A}	42.2(0.73)^{c,A}	42.8(0.73)^{c,A}

Different capital letters indicate significant differences (P < 0.05) within columns. Different small letters indicate significant differences (P < 0.05) within rows.

(2000, 4000 and 6000), different number of QTLs (400 and 600), different levels of heritability (0.10, 0.25, and 0.40), and different sizes of training sets (1000 and 2000), over three subsequent generations were investigated. With increasing size of the training set, from 1000 to 2,000 individuals (P<0.05), the average genomic accuracy increased from 35.3 (0.19) to 43.1% (0.19) which is in a close agreement with previous reports (Meuwissen et al., 2001; Callus and Veerkamp 2007). Increasing the number of markers from 2,000 to 6000 increased the average genomic accuracy from 33.8 (0.32) to 37.3% (0.32) with a training set of 1000 individuals, and from 40 (0.30) to 45.4% (0.30) with a tra-

ining set of 2000 individuals. Solberg et al. (2008) reported that for population with an effective size of 100, training set of 1000 and a heritability of 0.50, the accuracy of genomic predictions increased from 0.69 to 0.86 as marker density increased from 100 to 800 markers per Morgan. Although with increasing marker density from 2000 to 4000 the accuracy of genomic evaluation increased (*P*<0.05), but increasing marker density from 4000 to 6000, did not affect the genomic accuracy (*P*≥0.05).

Increasing the number of QTLs from 400 to 600 increased the average genomic accuracy from 34.7 (0.26) to 35.8 % (0.26) with a training set of 1000 individuals,

Table 3. The estimated genomic accuracy (SE) for a training set of 1,000, three marker density panels, and three levels of heritability over three subsequent generations

Generation	8			9			10		
h^2	0.10	0.25	0.40	0.10	0.25	0.40	0.10	0.25	0.40
Marker density panels									
2000	29.0(0.96)^{cde,A}	36.7(0.96)^{ab,B}	40.7(0.96)^{a,B}	27.4(0.96)^{de,A}	33.8(0.96)^{bc,B}	35.7(0.96)^{ab,B}	26.3(0.96)^{e,A}	32.0(0.96)^{bcd,B}	33.6(0.96)^{bc,B}
4000	31.6(0.96)^{de,A}	40.3(0.96)^{ab,AB}	44.0(0.96)^{a,AB}	28.9(0.96)^{ef,A}	37.3(0.96)^{bc,AB}	39.9(0.96)^{ab,AB}	26.3(0.96)^{f,A}	34.2(0.96)^{cd,AB}	39.7(0.96)^{ab,A}
6000	31.6(0.96)^{d,A}	42.6(0.96)^{ab,A}	46.1(0.96)^{a,A}	28.7(0.96)^{de,A}	39.6(0.96)^{bc,A}	43.0(0.96)^{ab,A}	26.3(0.96)^{e,A}	37.5(0.96)^{c,A}	39.9(0.96)^{bc,A}

Different capital letters indicate significant differences (P < 0.05) within columns. Different small letters indicate significant differences (P < 0.05) within rows.

Table 4. The estimated genomic accuracy (SE) for a training set of 2,000, three marker density panels, and three levels of heritability over three subsequent generations

Generation	8			9			10		
h^2	0.10	0.25	0.40	0.10	0.25	0.40	0.10	0.25	0.40
Marker density panels									
2000	36.6(0.9)^{cde,A}	43.5(0.9)^{b,B}	48.7(0.9)^{a,B}	34.4(0.9)^{de,A}	41.0(0.9)^{bc,B}	43.6(0.9)^{b,B}	32.9(0.9)^{e,A}	38.7(0.9)^{cd,B}	40.6(0.9)^{bc,B}
4000	39.6(0.9)^{de,A}	48.1(0.9)^{ab,AB}	51.7(0.9)^{a,AB}	37.0(0.9)^{ef,A}	46.0(0.9)^{bc,A}	47.6(0.9)^{abc,AB}	34.5(0.9)^{f,A}	43.2(0.9)^{cd,AB}	46.8(0.9)^{bc,A}
6000	40.2(0.9)^{d,A}	51.3(0.9)^{ab,A}	54.1(0.9)^{a,A}	36.9(0.9)^{de,A}	47.4(0.9)^{bc,A}	50.9(0.9)^{ab,A}	34.8(0.9)^{e,A}	45.4(0.9)^{c,A}	47.5(0.9)^{bc,A}

Different capital letters indicate significant differences (P < 0.05) within columns. Different small letters indicate significant differences (P < 0.05) within rows.

Table 5. The estimated genomic accuracy (SE) for atraining set of 1,000, two levels of QTLs and three levels of heritability over three subsequent generations

Generation	8		9		10	
QTL	400	600	400	600	400	600
h^2						
0.10	29.2(0.78)[ab,C]	32.3(0.78)[a,B]	26.9(0.78)[bc,B]	29.7(0.78)[ab,B]	25.3(0.78)[c,B]	27.3(0.78)[bc,B]
0.25	39.2(0.78)[a,B]	40.5(0.78)[a,A]	36.7(0.78)[ab,A]	37.1(0.78)[ab,A]	34.5(0.78)[b,A]	34.6(0.78)[b,A]
0.40	43.5(0.78)[ab,A]	43.7(0.78)[a,A]	39.7(0.78)[bc,A]	39.5(0.78)[c,A]	37.7(0.78)[c,A]	37.7(0.78)[c,A]

Different capital letters indicate significant differences (P < 0.05) within columns. Different small letters indicate significant differences (P < 0.05) within rows.

and from 42.7 (0.24) to 43.4% (0.24) with a training set of 2000 individuals. According to Zhang et al. (2011), genomic accuracy increased slightly as the number of QTLs increased. There was a significant difference between genomic accuracies of the eighth and tenth generations (*P*< 0.05), but no differences were recorded between the eighth and ninth generation, which is in close agreement with previous studies (Meuwissen et al., 2001; Habier et al., 2007). Increasing the level of heritability from 0.10 to 0.40 increased the average genomic accuracy from 36.3 (0.30) to 47.9% (0.30) with a training set of 2000 individuals, and from 28.5 (0.32) to 40.3% (0.32) with a training set of 1000 individuals. For heritability values of 0.10 and 0.50, genomic accuracies of 48% and 67% (Callus et al., 2008) and 40% and 70% (Saatchi et al., 2010) were reported. Hayes et al. (2009) reported that with traits of low heritability, more phenotypic records in the training setwere needed to obtain a certain level of genomic accuracy. The least squares means of genomic accuracy in different marker panels and different numbers of QTLs over three generations are presented in Table 1 (training set of 1000 individuals) and Table 2 (training set of 2000 individuals). With the training set of 2000, and for both QTL densities, increasing the marker number from 2000 to 6000, increased the genomic accuracy, but not with the training set of 1000 and 600 QTLs (*P* ≥ 0.05). Goddard (2009) reported that with a large number of QTLs, each one having a small effect, a large number of individuals in the training set are required to estimate the marker effects with high accuracy.

The least squares means of the genomic accuracy under different marker density panels and heritability levels, and over three generations are presented in Table 3 (with a training set of 1000 individuals) and Table 4 (with a training set of 2000 individuals). Under all situations, the average genomic accuracy increased when heritability values increased from 0.10 to 0.40, (*P*< 0.05). With heritability values of 0.25 and 0.40, but not 0.10, increasing the number of markers from 2000 to 6000, significantly increased the average genomic accuracy. The least squares means of the genomic accuracy for different QTL number, different heritability level, and over three generations are presented in Table 5 (training set of 1000 individuals) and Table 6 (training set of 2000 individuals). Under different QTL densities, increasing the heritability increased the average genomic accuracy (P<0.05) which is in a close agreement with Zhang et al. (2011).

The results of this study showed that increasing the marker density to 4000, improved the genomic accuracy for traits with high and moderate heritability, but not for low heritable traits. On the other hand, there was no difference between the genomic accuracy estimates of the first and second validation sets (generations eight and nine). Therefore, it seems that 4000 markers maybe sufficient for genomic evaluation where the estimated marker effect scan be used for at least two subsequent generations; however, the marker effects needs to be re-estimated for the third generation.

References

Calus, M.P.L., Veerkamp, R.F. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in gen-

Table 6. The estimated genomic accuracy (SE) for atraining set of 1,000, two levels of QTLs and three levels of heritability over three subsequent generations

Generation	8		9		10	
QTL	400	600	400	600	400	600
h^2						
0.10	37.7(0.73)[ab,C]	40.0(0.73)[a,B]	34.9(0.73)[bc,B]	37.3(0.73)[ab,B]	33.3(0.73)[c,B]	34.8(0.73)[bc,B]
0.25	47.3(0.73)[a,B]	48.0(0.73)[a,A]	44.8(0.73)[ab,A]	44.8(0.73)[ab,A]	42.6(0.73)[b,A]	42.2(0.73)[b,A]
0.40	51.6(0.73)[a,A]	51.4(0.73)[a,A]	47.6(0.73)[b,A]	47.1(0.73)[b,A]	44.8(0.73)[b,A]	45.1(0.73)[b,A]

Different capital letters indicate significant differences (P < 0.05) within columns. Different small letters indicate significant differences (P < 0.05) within rows.

omic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* 124, 362-368.

Calus, M.P.L., Meuwissen, T.H.E., De Roos, A. P. W., Veerkamp, R. F. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178, 553-561.

Goddard, M.E. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136, 245-257.

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.

Habier, D., Fernando, R.L., Dekkers, J. C. M. 2009. Genomic selection using low-density marker panels. *Genetics* 182, 343–353.

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

Hill, W.G., Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38, 226-231.

Meuwissen, T.H.E. 2009. Accuracy of breeding values of unrelatedindividuals predicted by dense SNP genotyping. *Genetics Selection Evolution* 41, 35.

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.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Saatchi, M., Miraei-Ashtiani, S.R., Nejati-Javaremi, A., Moradi-Shahrebabak, M., Mehrabani-Yeghaneh, H. 2010. The impact of information quantity and strength of relationship between training set and validation set on accuracy of genomic estimated breeding values. *African Journal of Biotechnology* 9, 438-442.

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

SAS. 2003. User's Guide: Statistics, Version 9.1 Edition. Inst., Inc., Cary, NC.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* 123, 218-223.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H.E. 2008. Genomic selection using different marker types and densities. *Journal of Animal Science* 86, 2447-2454.

Visscher, P.M., Haley, C.S., 1998. Strategies for marker-assisted selection in pig breeding programs. In: *Proceedings of the 6ᵗʰWorld Congress on Genetics Applied to Livestock Production*. Australia 23, 503-510.

Weigel, K.A., de los Campos, G., Gonzalez-Recio, O., Naya, H., Wu, X.L., Long, N., Rosa, G.J.M., Gianola, D. 2009. Predictive ability of direct genomic values for lifetime net merit of Holsteinsires using selected subsets of single nucleotide polymorphism markers. *Journal of Dairy Science* 92, 5248-5257.

Zhang, Z., Zhang, Q., Ding, X. 2011. Advances in genomic selection in domestic animals. *Chinese Science Bulletin* 56, 2655-2663.

Communicating editor: Ali K. Esmailizadeh

اثر تراکم نشانگر و وراثت‌پذیری برصحت انتخاب ژنومیک در سه نسل متوالی

ن. گرگانی فیروزجاه، هـ. آتشی*، م. دادپسند و م.ج. ضمیری

نویسنده مسئول، پست الکترونیک: hadiatashi@gmail.com

چکیده   هدف این پژوهش، بررسـی تاثیر تراکم‌های متفاوت نشـانگر، سـطوح متفاوت وراثت‌پذیری، افزایش اندازه‌ی جمعیـت کاندیدا، تراکم‌های متفاوت QTL، بر صـــحت ارزیابی ژنومیک در ســـه نســـل پی‌درپی بود. یک صـــفت با وراثت‌پذیری ۱۰، ۲۵ و ۴۰ درصـد شبیه‌سازی شد. برای هر حیوان یک ژنوم ۲۰ کروموزمی و با اندازه‌ی ۲۰ مورگان (هر کروموزم یک مورگان) شبیه‌سازی شد. شمار نشانگرها در ژنوم، ۲۰۰۰، ۴۰۰۰ و ۶۰۰۰ و فاصله‌ی بین آن‌ها به‌ترتیب یک، ۰/۵ و ۰/۳۳ سـانتی‌مورگان بود. شـمار QTLها، ۴۰۰ و ۶۰۰ در نظر گرفته شد که به‌گونه‌ی تصادفی در طول ژنوم توزیع شـــدند. تراکم نشـــانگرها، وراثت‌پذیری و اندازه‌ی جمعیت مرجع بر میانگین کمترین مربعات صـــحت برآورد ارزش‌های اصـلاحی ژنومیک تاثیر داشـتند ($P < 0.05$). یافته‌های این پژوهش نشـان داد که صـحت انتخاب در حالت اسـتفاده از ۴۰۰۰ یا ۶۰۰۰ نشـانگر و همچنین در نسـل‌های اول و دوم جمعیت کاندیدا (نسـل‌های ۸ و ۹) تفاوت معنی‌دار با یکدیگر نداشـتند. بنابراین می‌توان از ۴۰۰۰ نشـانگر، برای برآورد ارزش‌های اصـلاحی ژنومیک برای دسـت‌کم دو نسل (جمعیت کاندیدا) استفاده کرد، اما برای نسل سوم جمعیت کاندیدا، اثر نشانگرها باید دوباره برآورد شود.