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

Received: 17 Aug. 2014, accepted: 3 Oct. 2014, published online: 12 Nov. 2014

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 (Schefter, 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-
To create the linkage disequilibrium (LD) in the historical population, the number of individuals of both sexes were retained 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 typic variance of 1.0 was simulated. The true breeding values (TBVs) were randomly distributed among the markers. Effects directly affected the trait. A total of 400 and 600 QTLs were randomly sampled 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, \( \mu \) 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 \( N(0, \sigma^2_e) \). No additional information, such as marker location, polygenic effects, or pedigree was used in the model. The mixed model used to estimate the markers effects was:

\[
\begin{bmatrix}
1_n & 1_n X \\
X & 1_n \\
\end{bmatrix}
\begin{bmatrix}
\hat{u} \\
\hat{g} \\
\end{bmatrix} =
\begin{bmatrix}
1_n y \\
X y \\
\end{bmatrix}
\]

where, \( X \) is the matrix of marker genotypes for each animal, \( I \) is an identity matrix (total number of marker × total number of markers), \( \lambda \) is the ridge regression factor \( \frac{\sigma^2_g}{\sigma^2_e} \), \( \sigma^2_e \) is the residual variance, \( \sigma^2_g \) is variance common to each marker effect \( \frac{\sigma^2_g}{2 \sum_{n=1}^{p} p_i q_i} \) 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

<table>
<thead>
<tr>
<th>Generation</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL density panels</td>
<td>400</td>
<td>600</td>
<td>400</td>
</tr>
<tr>
<td>2000</td>
<td>34.3(0.78)ab,B</td>
<td>36.7(0.78)cA</td>
<td>31.3(0.78)cB</td>
</tr>
<tr>
<td>4000</td>
<td>38.0(0.78)ab,AB</td>
<td>39.3(0.78)cA</td>
<td>34.8(0.78)cB</td>
</tr>
<tr>
<td>6000</td>
<td>39.6(0.78)cA</td>
<td>40.5(0.78)cA</td>
<td>37.2(0.78)ab,A</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Generation</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL density panels</td>
<td>400</td>
<td>600</td>
<td>400</td>
</tr>
<tr>
<td>2000</td>
<td>42.1(0.73)c,A</td>
<td>43.8(0.73)cB</td>
<td>39.0(0.73)cB</td>
</tr>
<tr>
<td>4000</td>
<td>46.1(0.73)cA</td>
<td>46.8(0.73)cB</td>
<td>43.2(0.73)cB</td>
</tr>
<tr>
<td>6000</td>
<td>48.4(0.73)cA</td>
<td>48.7(0.73)cA</td>
<td>45.0(0.73)cA</td>
</tr>
</tbody>
</table>

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 2000 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 2000 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 training 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

<table>
<thead>
<tr>
<th>Generation</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker density panels</td>
<td>0.10</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>2000</td>
<td>29.0(0.96)cd,A</td>
<td>36.7(0.96)cB</td>
<td>40.7(0.96)cB</td>
</tr>
<tr>
<td>4000</td>
<td>31.6(0.96)c,A</td>
<td>40.3(0.96)cB</td>
<td>44.0(0.96)c,A</td>
</tr>
<tr>
<td>6000</td>
<td>31.6(0.96)c,A</td>
<td>42.6(0.96)cA</td>
<td>46.1(0.96)cA</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Generation</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker density panels</td>
<td>0.10</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>2000</td>
<td>36.6(0.9)de,A</td>
<td>43.5(0.9)de,B</td>
<td>48.7(0.9)cA</td>
</tr>
<tr>
<td>4000</td>
<td>39.6(0.9)de,A</td>
<td>48.1(0.9)de,B</td>
<td>51.7(0.9)c,A</td>
</tr>
<tr>
<td>6000</td>
<td>40.8(0.9)c,A</td>
<td>51.3(0.9)c,B</td>
<td>54.1(0.9)c,A</td>
</tr>
</tbody>
</table>

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 set were 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 6000 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 geno-

---

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

<table>
<thead>
<tr>
<th>Generation</th>
<th>QTL</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>600</td>
<td>400</td>
</tr>
<tr>
<td>h²</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td></td>
<td>29.2(0.78)c</td>
<td>32.3(0.78)c</td>
<td>26.9(0.78)c</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>39.2(0.78)c</td>
<td>40.5(0.78)cA</td>
<td>36.7(0.78)cA</td>
</tr>
<tr>
<td>0.40</td>
<td></td>
<td>43.5(0.78)cA</td>
<td>43.7(0.78)cA</td>
<td>39.7(0.78)cA</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Generation</th>
<th>QTL</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>600</td>
<td>400</td>
</tr>
<tr>
<td>h²</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td></td>
<td>37.7(0.73)c</td>
<td>40.0(0.73)c</td>
<td>34.9(0.73)c</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>47.3(0.73)c</td>
<td>48.0(0.73)cA</td>
<td>44.8(0.73)cA</td>
</tr>
<tr>
<td>0.40</td>
<td></td>
<td>51.6(0.73)cA</td>
<td>51.4(0.73)cA</td>
<td>47.6(0.73)cA</td>
</tr>
</tbody>
</table>

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


**Communicating editor:** Ali K. Esmailizadeh
اثر تراکم نشانگر و وراثتپذیری بر صحت انتخاب زنومیک در سه نسل متوالی
ن. غرگانی فیروزجاه، ه - آتش، م. دادپسند و ج. ضمیری
نویسنده، مسئول، پست الکترونیک: hadiatashi@gmail.com

چکیده
هدف این پژوهش بررسی تاثیر تراکم نشانگر، سطوح متغیر و وراثتپذیری، افزایش اندازه‌ی جمع‌آوری و مقدار QTL جمعیت کاندیدا، تراکم‌های متغیر و وراثت‌پذیری 200000 نسل‌پزشکی در سه نسل پدروپیسی بود. یک صفت و ریزی وراثت‌پذیری 0.10 و 0.40 درصد شیب‌سازی شد. برای هر حیوان یک زنوم 20 کروموزوم و با اندازه‌ی 2005 مورگان (هر کروموزوم یک مورگان) شیب‌سازی شد. شمار نشانگران در زنوم 200000 و 400000 و فاصله‌ی بین آنها به ترتیب یک، 0.5 و 0.33 سانتیمترگان بود. شمار QTL‌ها 4000 و 6000 در نظر گرفته شد که به‌گونه‌ای تصادفی در طول زنوم توزیع شدند. تراکم نشانگران، وراثت‌پذیری و اندازه‌ی جمعیت مرحله‌ی میانگین کمترین مرحله‌ی وجود آزاد از ورود اثرهای اصلاحی زنومیک تاثیر داشتند (0.05 < P < 0.10) با غرفه‌ای این پژوهش نشان داد که صحت انتخاب در حالت استفاده از 10000 یا 10000 نشانگر و همچنین در نسل‌های اول و دوم جمعیت کاندیدا (نسل‌های 8 و 9) تفاوت معنی‌دار با یکدیگر نداشتند. بنابراین می‌توان از 10000 نشانگر، برای برآورد اثرهای اصلاحی زنومیک برای دستکم دو نسل (جمعیت کاندیدا) استفاده کرد، اما برای نسل سوم جمعیت کاندیدا، اثر نشانگرها باید دوباره برآورده شود.