Impact of Accounting for Polygenic Effects on the Accuracy of Genomic Evaluations in Livestock Breeding

Abbas ATEFI[1] (✉)
Abdol Ahad SHADPARVAR[2]
Navid GHAVI-HOSSEINZADEH[2]

Summary

To investigate the accuracy of genomic breeding values, different scenarios were defined by accounting for polygenic effects, a different number of quantitative trait loci (30, 90, 150), and three levels of heritability (0.15, 0.25, and 0.4). The Bayes B method was used to estimate marker effects. A historical population was simulated stochastically, which consisted of 100 animals at first 100 generations, then the population size gradually increased to 1000 animals during the next 100 generations. The animals in generation 201 with known genotypic and phenotypic records were assigned as the reference population, and animals of generation 202 were considered as the validation population. The genome was comprised of one chromosome with 100 cM length and 500 markers that were distributed through the genome randomly. Picking up the information that was not captured by linkage disequilibrium (LD), including polygenic effects in the predictions increased the accuracy of genomic evaluations. As the trait heritability went from 0.15 to 0.40, the average genomic accuracy increased from 0.48 to 0.64. An increment in the number of quantitative trait loci (N_{QTL}) declined the accuracy of the Bayes B method. This study suggests that the highest accuracy (0.74) was achieved when additive genotypic effects were coded by a few quantitative trait loci and a lot of small effects included in the prediction of genomic breeding values.

Key words

accuracy, genomic prediction, Bayes B method, number of quantitative trait loci, heritability

[1] Ministry of Agriculture- Jahad, Islamic Republic of Iran

[2] Department of Animal Science, Faculty of Agricultural Science, University of Guilan, Rasht, Islamic Republic of Iran

✉ Corresponding author: abbas.atefi@gmail.com

Received: March 5, 2020 | Accepted: July 12, 2021

Introduction

Most of the economically important traits in livestock have a complex architecture defined by a large number of genes and affected by environmental factors. Statistical methods such as BLUP combined phenotypes and pedigree information to estimate the genetic merit (breeding values) of the selected candidates under Fischer's infinitesimal model. According to this model, phenotypes are expressed by an infinite number of loci, each with an infinitesimal additive effect. In the last decades, due to advances in molecular technologies and statistical methods, several chromosomal regions that influence quantitative traits have been detected. Moreover, the finite amount of DNA in the mammalian genome suggests that there must be a finite number of loci that control the expression of quantitative traits (between 20,000 and 35,000 genes) (Ewing and Green, 2000), in contrast with Fischer's infinitesimal model. The genomic selection method first suggested by Meuwissen et al. (2001) used dense marker information to enhance the potential for improving the accuracy of genetic values estimation. In genomic selection, genotypic information is used to select elite individuals to produce the next generation. The marker effects are estimated in a reference population, in which the individuals have both known phenotypes and genotypes. The estimated marker effects are used to estimate the genomic breeding values of selection candidates. However, many different methods were suggested to estimate the SNPs effects (Meuwissen et al. 2001, de los Campos et al. 2009, 2010, VanRaden, 2008, Gianola et al. 2006). Many studies have shown that factors such as the size of the reference data set (Meuwissen et al., 2001, VanRaden and Sullivan, 2010), trait heritability, the number of loci affecting the trait (Daetwyler et al., 2008), the degree of genetic relationships between training and validation samples (Habier et al., 2007) and the distributions of allele frequencies (Clark et al., 2011) affect the accuracy of genomic evaluations (Hayes et al. 2010, De los Campos et al. 2013). The standard genomic evaluation methods utilize genetic markers information that is in LD with at least a QTL. Accounting for polygenic effects in addition to marker effects reduces the number of false-positive QTL by decreasing fake associations. This approach may reduce prediction errors and therefore enhance the accuracy of the genomic evaluations. These advantages may be achieved due to better use of LD information and also better trapping relationship information that is not captured by LD. This study aimed to evaluate the accuracy of genomic prediction with approaches that account for polygenic effects, different numbers of QTLs, and three levels of heritability.

Materials and Methods

Simulation

Various scenarios were defined according to all combinations of accounting for polygenic effects, three different levels of heritability, and QTL numbers. Prediction accuracy, the correlation between the predicted genomic breeding values and the true values were estimated for each scenario. Parameter estimation was performed via the Gibbs Sampler algorithm implemented in the BGLR package of R software (Perez and De los Campos, 2014).

A historical population of 100 effective numbers with an equal sex ratio was simulated using the QMSim software (Sargolzaei and Schenkel, 2009), assuming three heritabilities of 0.15, 0.25, or 0.4. During the first 100 historical generations, random mating was performed; then, to arrive at a mutation-drift balance, 100 more generations were simulated in which the population size increased to 1000 individuals gradually (the average number of progenies per dam was equal to two). After the last historical generation, the recent population was constructed by the random selection of 1000 individuals and two successive generations were generated by random mating. The animals in generation 201 with known genotypes and records for the trait constructed the training population. The animals of generation 202 formed the validation population which it was assumed there were no phenotypic records.

The genome was comprised of one chromosome of 100 cM, and 500 marker loci and QTL were randomly distributed on the chromosome. All marker loci and QTL were bi-allelic with equal initial allelic frequencies. The number of segregating QTL affecting the trait was set at 30, 90, or 150. The Marker and QTL allele frequencies were assumed to be equal in the 200th generation. The mutation rate of the markers and QTLs were assumed to be 2.5×10^{-5} per locus per generation. To calculate the true breeding values, the additive effect of the QTLs was sampled from the gamma distribution with a shape parameter of 0.4 and a scale parameter of 1.66. The true breeding values were calculated as below:

$$TBV_i = \sum_{j=1}^{n=N_{QTL}} Q_{ij} q_j \qquad [1]$$

where $\mathbf{Q}_{ij}$ is an incidence vector indicating QTL alleles at locus $\boldsymbol{j}$ for animal $\boldsymbol{i}$ and $\boldsymbol{q}_j$ is a vector of QTL alleles effects at locus j. The phenotypic values were calculated as the sum of true breeding values and errors that were sampled from a normal distribution $N(0, \sigma^2_e)$.

The Bayes B model was used to estimate marker effects and to account for polygenic effects, a modified Bayes B model, described below, was used.

Model

Bayes B

The Bayes B method was first described in the study of Meuwissen et al (2001). Bayes B is likely the most accepted model, besides the lack of its formulation. Bayes B assumes a normal prior distribution on the marker effects with zero mean and variance. Then, a mixture of distributions is assumed on this variance that is equal to zero with probability π and distributed as a chi-square distribution with probability 1-π.

$\sigma^2_{\beta i} = 0$ with probability π,

$\sigma^2_{\beta i} \sim \chi^{-2}$ (df,s²) with probability (1-π).

In his formulation assuming a zero variance implies the absence of uncertainty about the marker effect, and therefore the inference lacks Bayesian sense. Furthermore, the selection of π is arbitrary with no justification as well as the choice of the hyperparameters in the inversed chi-square distribution that causes the flaws of this method. However, Bayes B is one of the most used methods and provides high accurate predictions, especially for those traits coded by large effect genes as fat percentage.

Modified Bayes B

The Bayes B model was extended to include a polygenic effect (Solberg et al., 2009):

$$Y = \mu + a + \sum_{j=1}^{p} X_j g_j + e \quad [2]$$

where **y** is the vector of phenotypes, **μ** is the overall mean, **a** is the vector of polygenic effects, **Σ** is the summation over all marker loci from 1 to *p*, X_j is a design matrix for the j^{th} marker, g_j is the vector of the j^{th} marker effect and ***e*** is the residual term. The variance of **a** was Var(**a**) = $A\sigma^2_a$, where **A** (1000 × 1000) is the additive relationship matrix, calculated based on five generations of pedigree from generation 196 to 200 using the algorithm of (Meuwissen and Luo,1992). Polygenic effects were sampled in the MCMC chain using Gibbs sampling and assuming a prior N(0, σ^2_a) following Sorenson and Gianola (2002), and σ^2_a was estimated using a scaled inverted chi-squared prior distribution with -2 degrees of freedom, which implies a non-informative flat prior distribution (Sorenson and Gianola, 2002).

The elements of the *X* for each individual depend on the number of alleles present in its genotype. For example, for i^{th} individual having genotypes AA, Aa, or aa at j^{th} marker locus, the X_{ij} element in *X* was assigned equal to 2, 1, or 0, respectively. In this study, a Bayesian approach (Bayes B) was used to estimate marker effects.

Results and Discussion

The accuracy of all scenarios was presented in Table 1. Because of picking up the information that was not captured by LD, including polygenic effects in the predictions increased the accuracy of genomic evaluations. The highest improvement of accuracy due to involving polygenic effects (0.07) occurred for the trait with the highest heritability (h^2=0.4) and the lowest number of QTLs (N_{QTL}=30).

Table 1. Accuracy of genomic prediction with approaches that account for polygenic effects, different number of QTLs, and three levels of heritability

Heritability	Number of QTLs	Without polygenic effects	With polygenic effects
30	0.51	0.54	
90	0.47	0.50	0.15
150	0.44	0.47	
30	0.62	0.66	
90	0.57	0.62	0.25
150	0.52	0.58	
30	0.67	0.74	
90	0.63	0.67	0.4
150	0.59	0.62	

In a simulation study, Piyasatian and Dekkers (2013) showed that when LD was low, the increment of the accuracy due to the inclusion of polygenic effects was noticeable. They declared that polygenic effects increased the LD signal and captured the remaining relationship information that was not captured by SNPs, depending on the extent of LD across chromosomes and training population size. Accounting for polygenic effects in a genomic model influenced the estimated variances by picking up the part of the genetic variance that was not captured by the common genomic model. A simulation study by Calus and Veerkamp (2007) showed a slight increase in accuracy by including polygenic effects on the genomic approach, but this depended on the extent of LD between adjacent SNPs. Kapell et al. (2012) considered various growth, behavioral and physiological traits in mice and showed that involving polygenic effects had little effect on the prediction ability of the genomic approach. Legarra et al. (2008) and De los Campos et al. (2009) reported an increased prediction ability using the genomic model relative to the polygenic model, but little difference between a solely genomic model and a combined genomic-polygenic model was found.

The accuracy of genomic evaluation improved from 0.48 to 0.64 as heritability increased from 0.15 to 0.40 (Fig. 1a). The positive correlation between genomic evaluation accuracy and heritability was reported in previous studies (Atefi et al, 2018, Wang et al. 2019, Calus and Veerkamp 2007, Kolbehdari et al. 2007, Martinez et al. 2018). Since marker effects are estimated using the relationship between phenotype and genotypic markers, estimation of marker effects and therefore genomic breeding values (GEBVs) will be more accurate for traits with high heritability.

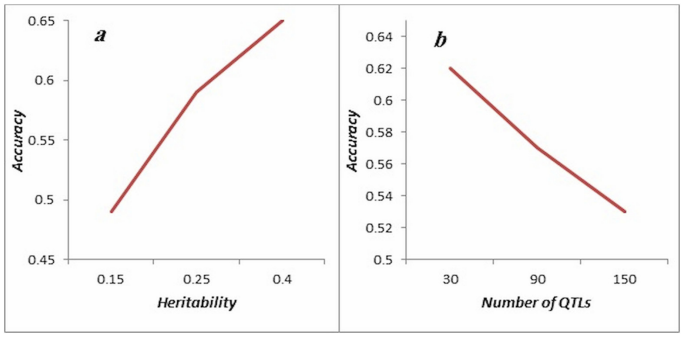

Figure 1. Accuracy of genomic evaluations for a) Different levels of heritability, b) Different number of QTLs

In a simulation study, where three levels for heritability (0.5, 0.3, and 0.1) in combination with other population structures and genetic architecture of the trait were investigated, the results showed an increasing trend in the accuracy of GEBVs when heritability of the trait increased (Atefi et al, 2018).

The amount of accuracy decreased as the number of QTLs increased. The lowest accuracy (0.53) was achieved for N_{QTL}=150 and the highest value (0.62) was for N_{QTL}=30 (Fig. 1b).

In a simulation study, increasing the number of QTLs from 0.03 Me to 1 Me (Me is the number of independent chromosome segments), decreased the accuracy of the Bayes B method from 0.739 to 0.344 (Daetwyler et al. 2008). However, different trends were reported by researchers. For instance, Gorgani Firozjah et al (2014) reported that the accuracy of all investigated scenarios decreased by increasing the number of QTLs from 400 to 600.

Different trends of genomic accuracy due to increasing the number of QTLs may be related to the interaction between the number of QTLs with other components of genetic architecture, i.e., the interaction among QTLs alleles and interaction between QTLs and non-genetic factors.

Bayesian methods define a prior density for SNPs in which a high proportion of SNPs has a null effect (π), while other SNPs have large or moderate effects. Therefore, Bayesian methods have greater accuracy for the traits controlled by a few QTLs (Wang et al. 2019).

This study suggests that the highest accuracy (0.74) was achieved when trait variation was extremely specified by additive genotypic effects, and additive genotypic effects were coded by a few QTLs and a lot of small effects provided that all these small effects were included in the estimation of GEBVs.

Conclusion

The accuracy of GEBVs was affected by three investigated factors in this study i.e., trait heritability, number of QTLs, and the inclusion of polygenic effects in the genomic evaluations. The accuracy of the Bayes B method was increased as the heritability increased. Conversely, the increment of the number of QTLs decreased the accuracy of GEBV. Including polygenic effects in the genomic evaluation improved the accuracy of GEBV due to capturing information that could not be captured only by LD. The highest accuracy (0.74) was obtained for the trait with the highest heritability which its variance defined by the lowest number of QTLs (30) and plenty of minor genes. The results of this study emphasize that considering polygenic effects in addition to genetic markers improved the accuracy of genomic breeding values by exploiting relationship information that could not be captured by LD.

References

Atefi A., Shadparvar A. A., Ghavi Hossein-Zadeh N. (2018). Accuracy of Genomic Prediction under Different Genetic Architectures and Estimation Methods. Iran J Appl Anim Sci 8 (1): 43-52

Calus M., Veerkamp R. C. (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 (6): 362-368. doi:10.1111/j.1439-0388.2007.00691.x

Clark S. A., Hickey J. M., Van der Werf J. H. (2011). Different Models of Genetic Variation and Their Effect on Genomic Evaluation. Genet Sel Evol 43 (18). doi: 10.1186/1297-9686-43-18

Daetwyler, H. D., Villanueva, B., Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3 (10): e3395. doi:10.1371/journal.pone.0003395

De los Campos G., Gianola D., Rosa G. J., Weigel K. A., Crossa J. (2010). Semi-Parametric Genomic-Enabled Prediction of Genetic Values Using Reproducing Kernel Hilbert Spaces Methods. Genet. Res 92 (4): 295-308. doi:10.1017/S0016672310000285

De los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Calus M. P. (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193 (2): 327-345. doi: 10.1534/genetics.112.143313

De Los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K., Cotes J. M. (2009). Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. Genetics 182 (1): 375-385. doi: 10.1534/genetics.112.143313

Ewing B., Green P. (2000). Analysis of Expressed Sequence Tags Indicates 35,000 Human Genes. Nat Genet 25 (2): 232-234. doi: 10.1038/76115

Gianola D., Fernando R. L., Stella A. (2006). Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics 173 (3): 1761-1776. doi: 10.1534/genetics.105.049510

Gorgani Firozjah N., Atashi H., Dadpasand M., Zamiri M. (2014). Effect of Marker Density and Trait Heritability on the Accuracy of Genomic Prediction over Three Generations. J Livest Sci Technol 2 (2): 53-58. doi: 10.22103/jlst.2014.880

Habier D., Fernando R., Dekkers J. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177 (4): 23972389. doi: 10.1534/genetics.107.081190

Hayes B. J., Pryce J., Chamberlain A. J., Bowman P. J., Goddard M. E. (2010). Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet 6 (9): e1001139. doi:10.1371/journal.pgen.1001139

Kapell D. N., Sorensen D., Su G., Janss L. L., Ashworth C. J., Roehe R. (2012). The Efficiency of Genomic Selection Using Bayesian Multi-Marker Models for Traits Selected to Reflect a Wide Range of Heritabilities and Frequencies of Detected Quantitative Traits Loci in Mice. BMC Genet 13: 42. doi:10.1186/1471-2156-13-42

Kolbehdari D., Schaeffer L., Robinson J. (2007). Estimation of Genome-Wide Haplotype Effects in Half-Sib Designs. J Anim Breed Genet 124 (6): 356-361. doi:10.1111/j.1439-0388.2007.00698.x

Legarra A., Robert-Granié C., Manfredi E., Elsen J. M. (2008). Performance of Genomic Selection in Mice. Genetics 180: 611–618. doi: 10.1534/genetics.108.088575

Martínez R., Burgos-Paz W., Bejarano D., Reyes P., Rocha J. F. (2018). Genomic Predictions and Accuracy of Weight Traits in a Breeding Program for Colombian Brahman. Proceedings of the World Congress on Genetics Applied to Livestock Production, Volume Genetic Gain - Strategies for Local Breeds 2, Auckland, New Zealand, 644 pp.

Meuwissen T., Luo Z. (1992). Computing Inbreeding Coefficients in Large Populations. Genet Sel Evol 24:305-313.

Meuwissen T., Hayes B., Goddard G. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157 (4): 1819-1829

Pérez P., De Los Campos G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198 (2):483-495. doi: 10.1534/genetics.114.164442

Piyasatian N., Dekkers J. (2013). Accuracy of Genomic Prediction when Accounting for Population Structure and Polygenic Effects. Animal Industry Report 10 (1). doi: 10.31274/ans_air-180814-1252

Sargolzaei M., Schenkel F. S. (2009). QMSim: A Large-Scale Genome Simulator for Livestock. Bioinformatics 25 (4): 680-681. doi: 10.1093/bioinformatics/btp045

Solberg T. R., Sonesson A. K., Woolliams J. A., Odegard J., Meuwissen T. (2009). Persistence of Accuracy of Genome-Wide Breeding Values over Generations when Including a Polygenic Effect. Genet Sel Evol 41: 53. doi: 10.1186/1297-9686-41-53

Sørensen D, Gianola D. (2002). Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. Springer-Verlag, New York, USA. 740 pp. doi: 10.1007/b98952

VanRaden P. M. (2008). Efficient Methods to Compute Genomic Predictions. J. Dairy Sci 91: 4414–4423. doi: 10.3168/jds.2007-0980

VanRaden P. M., Sullivan P. G. (2010). International Genomic Evaluation Methods for Dairy Cattle. Genet Sel Evol 42 (1):7. doi: 10.1186/1297-9686-42-7

Wang X., Miao J., Chang T., Xia J., An B., Li Y., Xu L., Zhang L., Gao X., Li J., Gao H. (2019). Evaluation of GBLUP, BayesB and Elastic Net for Genomic Prediction in Chinese Simmental Beef Cattle. PLoS ONE 14 (2): e0210442. doi: 10.1371/journal.pone.0210442