IJAS Iranian Journal of Applied Animal Science

A Comparison of the Sensitivity of the BayesC and Genomic Best Linear Unbiased Prediction (GBLUP) Methods of Estimating Genomic Breeding Values under Different Quantitative Trait Locus (QTL) Model Assumptions

Research Article

M. Shirali[1,2*], S.R. Miraei-Ashtiani[1], A. Pakdel[1], C. Haley[2,3], P. Navarro[3] and R. Pong-Wong[2]

[1] Department of Animal Science, Faculty of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
[2] Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, United Kingdom
[3] Medical Research Council Human Genetics (MRC) Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, United Kingdom

Received on: 28 Mar 2014
Revised on: 11 May 2014
Accepted on: 31 May 2014
Online Published on: Mar 2015

*Correspondence E-mail: masoud.shirali@roslin.ed.ac.uk

Online version is available on: www.ijas.ir

ABSTRACT

The objective of this study was to compare the accuracy of estimating and predicting breeding values using two diverse approaches, GBLUP and BayesC, using simulated data under different quantitative trait locus (QTL) effect distributions. Data were simulated with three different distributions for the QTL effect which were uniform, normal and gamma (1.66, 0.4). The number of QTL was assumed to be either 5, 10 or 20. In total, 9 different scenarios were generated to compare the markers estimated breeding values obtained from these scenarios using t-tests. In comparisons between GBLUP and BayesC within different scenarios for a trait of interest, the genomic estimated breeding values produced and the true breeding values in a training set were highly correlated (r>0.80), despite diverse assumptions and distributions. BayesC produced more accurate estimations than GBLUP in most simulated traits. In all scenarios, GBLUP had a consistently high accuracy independent of different distributions of QTL effects and at all numbers of QTL. BayesC produced estimates with higher accuracies in traits influenced by a low number of QTL and with gamma QTL effects distribution. In conclusion, GBLUP and BayesC had persistent high accuracies in all scenarios, although BayesC performed better in traits with low numbers of QTL and a Gamma effect distribution.

KEY WORDS BayesC, breeding value, GBLUP, number of QTL, QTL effect distribution.

INTRODUCTION

In breeding programs, estimating breeding values with high accuracy is one of the main objectives. In recent years, the improvement in genotyping technologies and genomic selection methods has resulted in greater accuracy in estimating breeding values. The term, genomic selection was introduced by Visscher and Haley (1998) and its methodology was outlined by Meuwissen *et al.* (2001). Selection based on genome wide distributed markers estimated breeding values (MEBVs) resulted in increased genetic progress, due to improvement in the accuracy of estimations of MEBVs, reduction in the generation interval (Meuwissen *et al.* 2001) and reduction in inbreeding rates, due to emphasis on MEBVs rather than family information (Woolliams *et al.* 2002; Daetwyler *et al.* 2007; Dekkers, 2007). The accuracy in obtaining MEBVs determines the success rate in breeding programs. The accuracy with which MEBVs are

estimated depends on several factors including the genetic architecture of the trait, the applied method in estimating marker effects and the distribution of quantitative trait locus (QTL) variance (Meuwissen *et al.* 2001; Goddard, 2008; Solberg *et al*. 2008; Shirali *et al*. 2012).

There are two main approaches in genomic selection for estimating breeding values. The first approach assumes that all single nucleotide polymorphisms (SNPs) have effects on the trait variance and the second approach assumes that just some SNPs contribute to the trait variance. In the first approach, genomic best linear unbiased prediction (GBLUP) methods including a form of ridge regression (Meuwissen *et al.* 2001) are applied, which instead of a pedigree relationship matrix, the marker relationship matrix is used (NejatiJavaremi *et al.* 1997; Villanueva *et al.* 2005; Hayes *et al.* 2009). The second approach assumes that a limited number of SNPs contribute to the trait variances and that among these affecting SNPs, only few of them make large contributions to trait variance and the rest have small contributions. In this approach, Bayesian methods (e.g., BayesB, BayesC and Lasso) have usually been used (Tibshirani, 1996; Meuwissen *et al.* 2001). The Bayesian methods use one prior for the QTL effect distribution and another prior for the number of QTL (Meuwissen *et al.* 2001). However, the true distribution of the QTL effects is unknown for many quantitative traits. Goddard (2008) found higher accuracies by using a gamma (1.66, 0.4) prior distribution for the QTL effects compared to a normal prior distribution. Briefly, the GBLUP assumption is that the genetic model is an infinitesimal model and all SNPs have effects on the trait of interest, while the BayesC approach assumes that some QTL have a high impact on trait variance and the rest have no effect on the trait. Both of these approaches may be suitable for different traits due to their assumptions regarding the genetic background of the trait of interest. The objective of this study was to investigate the effects of QTL effect distribution and number of QTL on the accuracy of MEBVs using the GBLUP and BayesC approaches with simulated data in quantitative traits with continuous phenotypes under different genetic architectures.

MATERIALS AND METHODS

Simulation

A genome consisted of one chromosome with a length of 100 cM was simulated and one thousand SNPs were equally spaced over the chromosome. Three different numbers of QTL (5, 10 and 20) were considered and QTLs were uniformly distributed over the chromosome. One hundred individuals, including 50 males and 50 females, were simulated for the base population (zero generation). These individuals were assumed to be biallelic for both SNPs and QTL with allele frequencies equal to 0.50 (Table 1). For the first generation, one male and one female were randomly chosen from the base population as parents. The parent's gametes were simulated assuming linkage equilibrium (LD) based on the Haldane mapping function (Haldane, 1919) to generate recombinant gametes and were randomly combined to create the individual. The first generation structure was followed through to the 50th generation of random mating to make linkage disequilibrium populations. The occurred recombination in the chromosome had a poisson distribution. For each generation, LD was measured using r^2 which was the average LD of all SNPs. Subsequent to the LD populations four more generations (51 to 54) were constructed. The population sizes for each of these four generations were considered to be 500 individuals, consisted of 250 males and 250 females for each population. In this study, generation 51 was assumed as a training population and the other generations (52 to 54) as validation populations. In simulating training and validation populations, three QTL (5, 10 and 20) were assumed to be influencing the trait of interest. This indicates the genetic background of the trait by the proportion of the SNPs that influenced the trait. Furthermore, the three different distributions were assumed for the QTL effect were uniform, normal and gamma (1.66, 0.4) (Table 2). Overall these assumptions for simulations generated traits for this study that had different genetic architectures.

Estimating the breeding values

Two methods, GBLUP and BayesC, were used to estimate SNPs effects and genomic breeding values. The main difference between these two applied approaches is in their assumptions regarding genetic models of the trait. GBLUP assumes an infinitesimal model is the genetic model of the trait of interest, and the BayesC assumes a QTL model. The genomic MEBVs (GEBVs) for individuals in validation generations (52-54) for both GBLUP and BayesC methods were predicted using the model:

$$GEBV = \sum_{i}^{n} X_i \hat{g}_i$$

Where:

n: number of chromosome across the genome.

X_i: design matrix which refers to individual genotypes for chromosome i.

g_i: vector of SNPs effects in chromosome i.

GBLUP method

The GBLUP approach was based on simple mixed model and assumed that all SNPs had equal effects on genetic variance of the considered trait.

Table 1 Population structure and simulated parameters

Parameter	Value
Number of chromosome	1
Number of SNP markers per chromosome	1000
Genome length	100cM
Marker distance (cM)	0.1
Number of QTL	5, 10 and 20
QTL effects distributions	Normal, gamma (1.66; 0.4) and uniform
Recombination	Haldane map function
Number of generation	54
LD populations	50
LD population size per generation	100 individuals (50 males and 50 females)
Number of generation for population	4 (generation 51 to 54)
Population size	500 individuals (250 males and 250 females)
Training set	All individuals of generation 51 (500 individuals include 250 males and 250 females)
Validation set	All individuals of generations 52 to 54 (1500 individuals include 750 males and 750 females)
Heritability	0.3

QTL: quantitative trait locus and LD: linkage equilibrium.

Table 2 Scenarios with different numbers of QTL and different distributions of QTL variances

Scenario	Number of QTL	Distribution of QTL variance
NE5	5	Normal
GE5	5	Gamma
UE5	5	Uniform
NE10	10	Normal
GE10	10	Gamma
UE10	10	Uniform
NE20	20	Normal
GE20	20	Gamma
UE20	20	Uniform

QTL: quantitative trait locus.
GE: gamma; NE: normal and UE: uniform.

In GBLUP, this assumption has been shown to be unrealistic. However, Meuwissen *et al.* (2001) showed that GBLUP is easy, fast, simple and intelligible and it is still one useful approach in genome-wide association studies (GWAS). In the GBLUP approach, the following model was applied using ASReml (Gilmour *et al.* 1995):

$y = \mu 1_n + X_i g_i + e$

Where:
y: vector of phenotypic values.
μ: overall mean.
n: number of records.
1_n: vector of n ones.
g_i: represents the additive genetic effects of the i^{th} SNP.
X_i: design matrix for the i^{th} SNP.
e: residual error with normal distribution.

The additive genetic effects of SNPs (g) were assumed to have a normal distribution $N(0, \sigma_g)$ where g was the realized relationship matrix for all loci. The g was calculated based on the identical-by-state probabilities between a pair of individuals for all individuals in the training and validation populations.

The total allelic relationship between each pair of individuals was calculated based on the method of NejatiJavaremi *et al.* (1997). The mixed model equation to obtain breeding values is (Henderson, 1975):

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

Where:
$\alpha = (\sigma^2_e/\sigma^2_g)$.
I: identity matrix.

BayesC method

The BayesC method was performed using the model:

$$y = \mu 1_n + \sum_{i=1}^{m} X_i g_i + e$$

Where:
y: phenotypic value vector.
μ: overall mean.
n: number of records.
1_n: vector of n ones.
X_i: represents the vector of genotypes (1, 2 and 3 for genotypes 00, 10/01 and 11, respectively) of the i^{th} SNP.
g_i: allelic substitution effect for SNP *i*.
e: vector of residual distributed with $N(0, \sigma_e)$.

In the BayesC approach, the model assumption was for an allelic substitution effect for each SNP with a mixture distribution in which a portion of SNPs (π) have effects on the trait with $N(0, \sigma_{snps})$ distribution and the rest have no effect on the trait. Gibbs sampling was used for implementation of the model. A flat prior was applied to calculate the parameters π, σ^2_{snps} and σ^2_e.

For each analysis, a Markov chain Monte Carlo (MCMC) with 210000 cycles ran and the first 10000 cycles were discarded as burn-in period. Estimates at every 5^{th} iteration were sorted as a sample, resulting in a total 40000 samples.

Comparison of the methods to estimate breeding values

Estimations from each method were compared based on the accuracy and the mean square error of prediction (MSEP) of GEBV. The correlation between GEBV and true ge-

nomic breeding value (TGBV) was used as measure of accuracy. MSEP is the average of the squared prediction errors of MEBVs. To calculate the accuracy and MSEP, the GEBV estimates of each individual in the training population were used following Coster *et al.* (2010). In addition, the accuracy and MSEP were also estimated for the validation generations. GEBVs from BayesC and GBLUP in all scenarios were compared with each other using t-tests.

RESULTS AND DISCUSSION

In the simulation analysis, calculated average LD between all SNPs (r^2) in last generation of the LD population (generation 50) was 0.185 ± 0.012. This indicates that 87% of the expected LD had been achieved in this simulation. The expected LD based on Sved (1971) is 0.211.

Accuracy under different numbers of QTL

Using GBLUP, no differences in the accuracy of estimations were detected by increasing in the number of QTL. However, the BayesC method was more accurate under the scenario with 5 QTL and with the gamma QTL effect distribution (GE5) than under scenarios with 10 and 20 QTL with the gamma effect distribution (GE10 and GE20, respectively). This indicates that the BayesC approach performed differently under different numbers of QTL when the QTL effect distribution is gamma.

Accuracy under different distributions of the QTL effect

Table 3 indicates that the average accuracy of MEBVs calculated by BayesC and GBLUP while considering different distributions of the QTL effect. The result indicates that in all scenarios for different distributions of the QTL effect and different numbers of QTL, the BayesC estimations were significantly ($P<0.05$) better than those by GBLUP except under uniform distribution with 20 QTL (UE20) scenario. The accuracies of MEBVs estimations were significantly ($P<0.05$) higher for BayesC compared to GBLUP when comparing the following scenarios: normal and uniform distribution of QTL effects with 5 QTL scenarios (NE5 and UE5, respectively), gamma and normal distributions of QTL effects with 10 QTL (GE10 and NE10, respectively) and gamma distribution of the QTL effect with 20 QTL (GE20) scenarios.

The BayesC analysis indicated that the greatest accuracy ($P<0.01$) is achieved compared to GBLUP under gamma distributions of the effect and with 5 QTL. Furthermore, in the BayesC approach the scenario with gamma distributions of the QTL s and 5 QTL (GE5) had significantly greater accuracy of estimation compared to the uniform and normal distributions of the QTL effects with 5 QTL (UE5 and NE5, respectively). In addition, BayesC under normal distribution of the QTL effect with 10 QTL (NE10) performed significantly more accurately than with the uniform distribution of QTL effects with 10 QTL (UE10). Furthermore, the type of QTL effect distribution did not show any significant effect on the accuracy of MEBVs estimations using the GBLUP approach. In the current study, 5, 10 and 20 QTL were considered to simulate traits variance and 1000 marker simulated to analyse the trait. In fact 0.5, 1 and 2 present of markers are in complete LD with simulated QTL.

In this study, GBLUP and BayesC methods of analysis were compared for diverse populations with different trait genetic architectures. The results indicated that GBLUP had a consistently high accuracy in all QTL distributions and in scenarios with different numbers of QTL. The results are in agreement with the study of Shirali *et al.* (2012) who reported that different QTL variance distributions and different numbers of QTL had no effects on accuracies of GBLUP estimations. The maximum accuracy of BayesC estimates was achieved for the lowest number of QTL. This indicates that BayesC has an advantage over GBLUP for analysis of traits that are influenced by a low number of QTL. The results of the current study are in agreement with Daetwyler *et al.* (2010) who found a decrease in the accuracy with an increase in the number of QTL. The current study found that the accuracy of BayesC dropped as the number of QTL increased, which resulted in no advantage over GBLUP when the number of QTL was less than 20, with a uniform distribution. By increasing the number of QTL for a trait, the average variance of each QTL for the trait of interest will decrease and the estimation of the QTL effect will be less accurate. With uniform QTL effect distribution, by increasing the number of QTL the proportional contribution of each QTL on the trait will be very low and therefore some of their effects will be missed and missing heritability will be increased. Shirali *et al.* (2012) reported that in traits influenced by a high number of QTL, the QTL variance distribution does not influence the accuracies when using the BayesC method. This can be due to the fact that by increasing the number of QTLs, the effect of each QTL on the trait will decrease and thus estimated QTL effects will be small and the QTL effect distribution will be more similar to a uniform distribution. Therefore, the effect of each QTL for the trait of interest will decrease resulting in difficulties for the BayesC model to detect the QTL effects. The BayesC method provided more accurate estimates under different QTL distributions compared to GBLUP. However, Shirali *et al.* (2012) suggested that GBLUP can provide estimations as accurate as BayesC under different QTL variance distributions. BayesC estimated QTL effects are more accurate in simulated QTL effect distributions than in QTL variance distributions.

Table 3 Average (standard error) accuracies of markers estimated breeding values (MEBVs) when analysed by either the genomic best linear unbiased prediction (GBLUP) method or the BayesC method for each generation (GNR) for different scenarios (SCN) in QTL effect distribution

SCN	GNR	BayesC	GBLUP	SCN	GNR	BayesC	GBLUP	SCN	GNR	BayesC	GBLUP
GE5	51	0.912 (0.031)	0.828 (0.027)	GE10	51	0.847 (0.040)	0.813 (0.028)	GE20	51	0.863 (0.044)	0.831 (0.026)
	54	0.883 (0.047)	0.741 (0.050)		54	0.773 (0.065)	0.706 (0.043)		54	0.804 (0.076)	0.753 (0.048)
NE5	51	0.877 (0.054)	0.817 (0.047)	NE10	51	0.859 (0.030)	0.822 (0.025)	NE20	51	0.853 (0.025)	0.836 (0.024)
	54	0.826 (0.084)	0.719 (0.065)		54	0.795 (0.056)	0.727 (0.045)		54	0.787 (0.050)	0.762 (0.041)
UE5	51	0.857 (0.040)	0.825 (0.029)	UE10	51	0.832 (0.021)	0.819 (0.019)	UE20	51	0.842 (0.030)	0.834 (0.032)
	54	0.799 (0.063)	0.731 (0.056)		54	0.740 (0.043)	0.715 (0.036)		54	0.765 (0.057)	0.757 (0.058)

GE: gamma; NE: normal and UE: uniform.

Therefore the MEBVs in different QTL effect distributions have a higher accuracy rather than in QTL effect distributions. Gamma distributions of QTL effects resulted in better accuracy in BayesC. Shirali *et al.* (2012) also reported better accuracy using BayesC estimation for gamma distribution in QTL variance.

These effects can be due to two possible reasons; first, the prior, the QTL effect and the QTL variance are all gamma distributions. As a result, BayesC would have a better estimation of any SNP effect. Goddard (2008) reported that BayesC under gamma prior provide better accuracies for MEBVs and this is in agreement with the current study.

Second, gamma distribution captures QTL with very high effects compared to a normal distribution, resulting in more accurate estimation of GEBVs for traits which are influenced by a number of QTL with high effects. The correlations between GEBVs from BayesC and TEBV were above 80% in all scenarios, which is in agreement with Nadaf and Pong-Wong (2011), Daetwyler *et al.* (2010) and Solberg *et al.* (2008).

CONCLUSION

The GBLUP method of analysis was as good as the BayesC method fortraits influenced by a high number of QTL and with a uniform QTL effect distribution, such as traits with a polygenic genetic model. GBLUP had a consistently high accuracy in scenarios with all QTL effect distributions and all numbers of QTL. BayesC produced estimates with higher accuracies in traits influenced by low number of QTL and with a gamma QTL effect distribution.

ACKNOWLEDGEMENT

The first author acknowledges financial support from Iranian Ministry of Science, Research and Technology to support study at the Roslin institute.

REFERENCES

Coster A., Bastiaansen J.W.M., Calus M.P.L., Van Arendonk J.A.M. and Bovenhuis H. (2010). Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genet. Select. Evol.* **42,** 9-15.

Daetwyler H.D., Pong-Wong R., Villanueva B. and Woolliams J.A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*. **185,** 1021-1031.

Daetwyler H.D., Villanueva B., Bijma P. and Woolliams J.A. (2007). Inbreeding in genome-wide selection. *J. Anim. Breed. Genet.* **124,** 369-376.

Dekkers J.C.M. (2007). Prediction of response from marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* **124,** 331-341.

Gilmour A.R., Thompson R. and Cullis B.R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*. **51,** 1440-1450.

Goddard M. (2008). Genomic selection prediction of accuracy and maximisation of long term response. *Genetica*. **136,** 245-257.

Haldane J.B.S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet*. **8,** 299-309.

Hayes B.J., Visscher P.M. and Goddard M.E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91,** 47-60.

Henderson C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*. **31,** 423-447.

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. **157,** 321-322.

Nadaf J. and Pong-Wong R. (2011). Applying different genomic evaluation approaches on QTLMAS2010 dataset. *BMC Proc*. **5(3),** 9-16.

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

Shirali M., Miraei-Ashtiani S.R., Pakdel A., Haley C. and Pong-Wong R. (2012). Comparison between Bayesc and GBLUP in

estimating genomic breeding values under different QTL variance distributions. *Iranian J. Anim. Sci.* **43,** 261-268.

Solberg T.R., Sonesson A.K., Woolliams J.A. and Meuwissen T.H.E. (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* **86,** 2447-2454.

Sved J.A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol*. **2,** 125-141.

Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B Met.* **58,** 267-288.

Villanueva B., Pong-Wong R., Fernandez J. and Toro M.A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* **83,** 1747-1752.

Visscher P.M. and Haley C.S. (1998). Strategies for marker-assisted selection in pig breeding programmes. Pp. 503-510 in Proc. 6th World Cong. Genet. Appl. Livest. Prod. WCGALP, Armidale, Australia.

Woolliams J.A., Pong-Wong R. and Villanueva B. (2002). Strategic optimisation of short- and long-term gain and inbreeding in MAS and non-MAS schemes. Pp. 23-25 in Proc. 7th World Cong. Genet. Appl. Livest. Prod. Montpellier, France.