Indian Journal of Animal Sciences **87** (3): 324–328, March 2017/Article

Genomic accuracy in different genetic architecture and genomic structure

F ALA NOSHAHR[1], S A RAFAT[2], R IMANY-NABIYYI[3], S ALIJANI[4] and C ROBERT GRANIE[5]

Faculty of Agriculture, Tabriz University, Tabriz, 29 Bahman Boulevar 516 66 Iran
and
Institut national de la recherche agronomique, BP 27, 31326 Castanet-Tolosan France

Received: 29 June 2016; Accepted: 7 September 2016

ABSTRACT

Genomic selection has been widely implemented in national and international genetic evaluation in the animal industry, because of its potential advantages over traditional selection methods and the availability of commercial high density single nucleotide polymorphism (SNP) panels.Considerable uncertainty currently exists in determining which genome-wide evaluation method is the most appropriate. We hypothesize that genome-wide methods deal differently with the genetic architecture of quantitative traits and genomes. A genomic linear unbiased prediction method (GBLUP) and a genomic nonlinear Bayesian variable selection methods (BayesA and BayesB) were compared using stochastic simulation across three effective population sizes (N_e). Thereby, a genome with three chromosomes, 100 cM each was simulated. For each animal, a trait was simulated with heritability of 0.50, three different marker densities (1000, 2000 and 3000 markers) and number of QTL was assumed to be either 100, 200 or 300. Data were simulated with two different distributions for the QTL effect which were uniform and gamma (a= 1.66, b=0.4). Marker density, number of QTL and QTL effect distributions significantly affected the genomic accuracy with different N_e. BayesB produced estimates with higher accuracies in traits influenced by a low number of QTL, high marker density, gamma QTL effect distribution and with high N_e.

Key words: BayesA, BayesB, GBLUP, Genomic accuracy, Marker density, N_e

Genome-wide evaluation combines traditional approaches to the prediction of genetic values with the use of high throughput genotype data such as SNP (Meuwissen *et al.* 2001). In breeding programs, estimating breeding values with high accuracy is one of the main objectives. The accuracy of predicted genetic values from genome-wide evaluation can be substantially higher than that of traditional methods provided that enough phenotypic records are available for estimating marker effects (Daetwyler *et al.* 2010, Goddard 2008, Hayes *et al.* 2009). 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 (Daetwyler *et al.* 2007, Dekkers 2007). Accuracy of MEBV depends on genetic architecture of the trait, number of QTL, marker density panels, the heritability of the trait, the size of the training population, the distribution of QTL effect, the method used to estimate marker effects and LD between markers and QTL (Meuwissen *et al.* 2001, Goddard 2008).

There are two main approaches in genomic selection for estimating breeding values. The first approach assumes that all SNPs have effects on the trait variance and the second approach assumes that only 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 (Nejati Javaremi *et al.* 1997). 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 other have small contributions. In this approach, Bayesian methods (e.g., BayesB, BayesC and Lasso) have usually been used (Tibshirani 1996).

Factors affecting the accuracy of the genomic selection in different N_e are largely unknown. We hypothesise in this paper that the relative utility of genome-wide evaluation methods depends significantly on both the genomic structure of the population and the genetic trait architecture.Thus, the main objective of this study was to compare GBLUP, and non-linear variable selection methods, BayesA and BayesB, using simulated data across a range of population and trait genetic architectures to

Present address: [1]Ph.D student (f.alanoshahr@gmail.com), [2,4]Associate Professor (abbasrafat@hotmail.com, saeidsbry @yahoo.com), Department of Animal Sciences. [3]Associate Professor (imany@tabrizu.ac.ir), Department of Mathematical Sciences. [5](christele.robert-granie @toulouse.inra.fr).

investigate the effects of marker density, number of QTL and N_e on the accuracy of MEBVs.

MATERIALS AND METHODS

Simulation: The populations were simulated using the QMSim software (Sargolzaei and Schenkel 2009) based on forward-in-time process. A genome consisted of three chromosome with a length of 100 cM was simulated and 1,000, 2,000 and 3,000 SNPs were equally spaced over the chromosome. Three different numbers of QTL (100, 200 and 300) 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 loci were assumed to be biallelic for both SNPs and QTL with allele frequencies equal to 0.50 (Table 1).

Table 1. Population structure and simulation parameters

Parameter	Value
Number of chromosome	3
Genome length	300 cM
Effective population size (N_e)	50, 100 and 200
Number of QTL	100, 200 and 300
Number of SNP markers	1,000, 2,000 and 3,000
QTL effects distributions	Uniform and gamma (1.66, 0.4)
Heritability	0.5
Training set	All individuals of generation 51 and 52
Validation set	All individuals of generations 53 to 60

The simulation started with an initial population of 100 N_e individuals and followed by $0.5N_e$ and $2N_e$ discrete generations, denoted as historical generations. In the initial population and each historical generation, males and females were randomly selected to form N_e matings and produced N_e offspring with $0.5N_e$ males and $0.5N_e$ females. The parent's gametes were simulated assuming LD based on the Haldane mapping function 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. Subsequent to the LD populations, 10 more generations (51 to 60) were constructed. The base population consisted of 1,000 unrelated animals (500 males and 500 females). In this study, generation 51 and 52 was assumed as a training population and the other generations (53 to 60) as validation populations. In simulating training and validation populations, three QTL number (100, 200 and 300), three marker densities (1,000, 2,000 and 3,000) and heritability level of 0.50 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 two different assumed distributions for the QTL effect were uniform and gamma (a=1.66, b=0.4). Overall these assumptions for simulated traits for this study had different genetic architectures. The mutation rate of the markers and QTLs was assumed 2.5×10^{-5} per locus per generation (Solberg *et al.* 2008).

Estimating the breeding values: Three methods, GBLUP, BayesA and BayesB, were used to estimate QTL, SNPs effects and genomic breeding values. The main difference between these three applied approaches is in their assumptions regarding genetic models of the trait. The genomic estimated breeding values (GEBV) for individuals in validation generations for three GBLUP, BayesA and BayesB methods were predicted using the model:

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

Where, n, number of SNPs across the genome; $X_{i,}$ design matrix which refers to individual genotypes for SNPs; g_i, is the 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. In GBLUP, this assumption has been shown to be unrealistic. The additive genetic effects of SNPs (g) were assumed to have a normal distribution $N(0, \sigma_g)$ where g was the realized relationship vector 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.

BayesA method: In this model like GBLUP, all SNPs were assumed to have some effect, however, assumed that some of the SNPs were in linkage disequilibrium with QTL of moderate to large effect. The SNP effects sampled from a normal distribution with the variance for each SNP sampled from an inverse scaled Chi-square distribution. The shape of the distribution that the SNP effects were sampled from is dependent on the degrees of freedom used for the inverse scaled Chi-square distribution (Habier *et al.* 2007).

BayesB method: BayesB assumes that many of the SNPs are in genomic regions where there are no QTL and thus have zero effects, whilst a small proportion of SNPs are in LD with QTL and consequently do have an effect. This structure means that those effects that are non-zero can be thought of as those in stronger LD with the QTL. In fact, if the number of times a SNP is included in the model is recorded, the posterior probability of that SNP being linked to a QTL can be calculated. A Gibbs sampling algorithm was implemented to obtain samples from the joint posterior distribution. For each analysis, a Markov chain Monte Carlo (MCMC) with 210,000 cycles with WinBUGS software ran and the first 10,000 cycles were discarded as burn-in period. Estimates at every 5th iteration were sorted as a sample, resulting in a total 40,000 samples.

RESULTS AND DISCUSSION

In current simulation analysis, calculated average LD values between all SNPs (r^2) in the last generation of the LD population (generation 50) was 0.191±0.011. This indicates that 87% of the expected LD had been achieved in this simulation. The genomic accuracy, the correlations

between TBVs and GEBVs, for different marker densities (1,000, 2,000 and 3,000), different number of QTLs (100, 200 and 300), different N_e (50, 100 and 200) with two QTL effect distributions, uniform (Table 2) and gamma (Table 3) were showed.

Genomic accuracy under different marker density and N_e: The results showed that, the relative genomic accuracy increased with the decrease of QTL numbers, increase of marker densities and with the increase of N_e. Increasing the marker density from 1,000 to 3,000, increased the average genomic accuracy with two QTL effect distributions and three N_e levels (Fig. 1a) and in all scenarios SE<0.03 (Fig. 1b). Increasing the accuracy of genomic breeding values by increasing marker density to 3,000 increased the linkage disequilibrium between markers and QTL. The results of this study were agreement with the results of Solberg *et al.* (2008) and Habier *et al.* (2007). Solberg *et al*. (2008) reported that increasing marker density from 100 to 800 markers at each Morgan, genomic accuracy increased from 69 to 86%. Increasing the number of markers increased the LD between genes and markers, and thus increased the accuracy of genomic evaluations.

Also by increasing the effective population size, genomic accuracy of breeding values also increased. The reason can be attributed to increase in the number of known data (number of phenotypic records in the base population) versus the number of unknowns variables (SNP effects). When the number of observations in the base population are greater, the SNP effects will be more precisely estimated and eventually genomic breeding values will be more accurate. In general, in the same number of QTL and density of markers in both uniform and gamma QTL effect distribution, estimated accuracies of GEBV for high N_e (200), were higher than the moderate (100) and low N_e (50), respectively.

Genomic accuracy under different numbers of QTL and QTL effect distributions: In current study, increasing the number of QTLs from 100 to 300, decreased the average

Table 2. The estimated genomic accuracy for different effective population size (N_e), three marker densities and numbers of QTL (N_{OTL}) with uniform QTL effect. SE<0.03 in all scenarios.

		Statistical method								
		GBLUP			BayesA			BayesB		
		Marker density								
N_e	N_{OTL}	1,000	2,000	3,000	1,000	2,000	3,000	1,000	2,000	3,000
	100	0.437	0.458	0.469	0.441	0.462	0.474	0.563	0.579	0.592
50	200	0.428	0.443	0.455	0.432	0.447	0.461	0.557	0.568	0.584
	300	0.420	0.436	0.442	0.417	0.430	0.435	0.542	0.555	0.573
	100	0.729	0.756	0.783	0.732	0.758	0.789	0.821	0.832	0.868
100	200	0.718	0.729	0.735	0.725	0.733	0.745	0.813	0.820	0.831
	300	0.709	0.715	0.722	0.706	0.711	0.718	0.733	0.745	0.751
	100	0.751	0.762	0.785	0.757	0.774	0.791	0.831	0.847	0.879
200	200	0.746	0.752	0.778	0.750	0.761	0.786	0.825	0.838	0.851
	300	0.733	0.741	0.769	0.725	0.732	0.758	0.817	0.826	0.833

Table 3. The estimated genomic accuracy for different effective population size (N_e), three marker densities and numbers of QTL (N_{OTL}) with gamma QTL effect. SE<0.03 in all scenarios.

		Statistical method								
		GBLUP			BayesA			BayesB		
		Marker density								
N_e	N_{OTL}	1,000	2,000	3,000	1,000	2,000	3,000	1,000	2,000	3,000
	100	0.450	0.464	0.475	0.458	0.473	0.481	0.574	0.585	0.599
50	200	0.444	0.451	0.466	0.453	0.468	0.475	0.567	0.572	0.587
	300	0.436	0.442	0.457	0.427	0.436	0.443	0.555	0.564	0.569
	100	0.733	0.757	0.791	0.744	0.763	0.797	0.844	0.867	0.886
100	200	0.720	0.733	0.747	0.731	0.746	0.759	0.830	0.843	0.855
	300	0.714	0.722	0.730	0.712	0.719	0.725	0.818	0.827	0.839
	100	0.789	0.807	0.815	0.793	0.811	0.825	0.849	0.875	0.893
200	200	0.780	0.794	0.805	0.784	0.802	0.817	0.833	0.862	0.873
	300	0.771	0.782	0.795	0.764	0.771	0.780	0.821	0.845	0.854

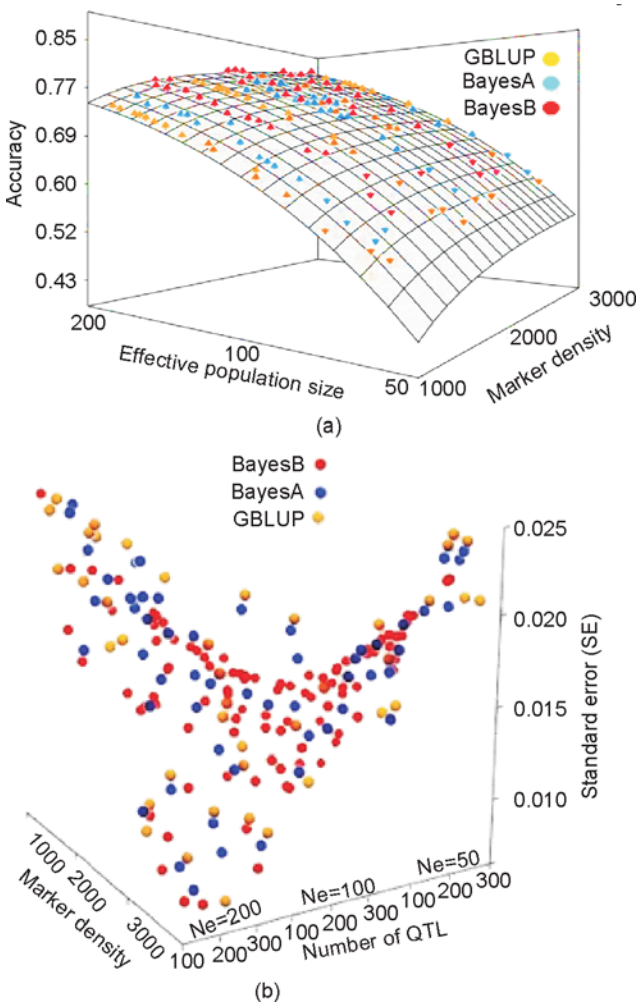

Fig. 1. Genomic accuracy of three methods viz. GBLUP, BayesA and BayesB with three marker density levels in three effective population size (a) and distributions of SE (b).

genomic accuracy in all three N_e and two QTL effect ditributions (Fig. 2). The results of the current study were in agreement with Daetwyler *et al.* (2010) who found a decrease in the accuracy with an increase in the number of QTL. 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 increase. 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.

In addition, the gamma distributions of QTL effects resulted in better accuracy in three methods. Shirali *et al.* (2015) also reported better accuracy using BayesC estimation for gamma distribution in QTL effect. When the distribution of the gene effects is gamma, some genes have major effects and a high percentage of genes are close to

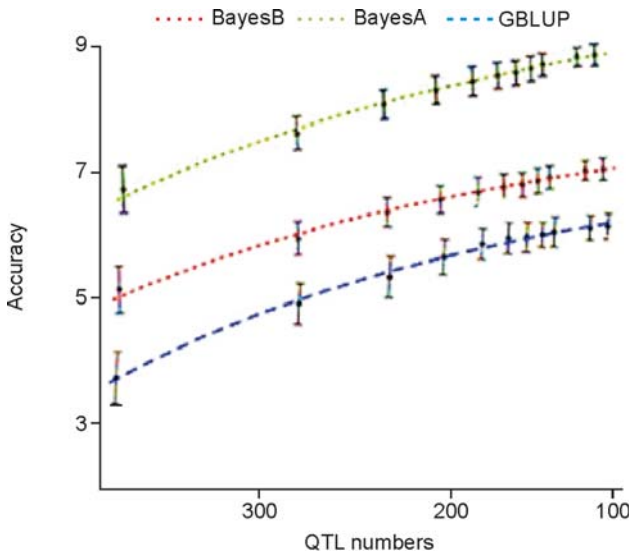

Fig. 2. Genomic accuracy of three methods viz. GBLUP, BayesA and BayesB with three numbers of QTL.

zero impact. So the Bayesian methods compared with non-Bayesian methods are better. These effects can be due to two possible reasons; first, the prior, the QTL effect and the QTL effect are all gamma distributions. Goddard *et al.* (2011) 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 and uniform distribution, resulting in more accurate estimation of GEBVs for traits which are influenced by a number of QTL with high effects (Daetwyler *et al.* 2010, Nadaf and Pong-Wong 2011).

Genomic accuracy under different methods: Among the three studied methods, the greatest genomic accuracy was obtained in low number QTL (100), high marker density, gamma QTL effect ditribution and large number of N_e (200) with BayesB method. Many studies had shown that Bayes method is more accurate in comparison with BLUP which is consistent with the present results (Solberg *et al.* 2008). One weakness of the GBLUP method is that it consider same proportion of the variance for all markers in the genomic prediction of breeding values. While in the Bayesian method based on prior distribution, different weights are assigned to the marker density. However, GBLUP has an advantage over BayesA at high number of QTL (300), but this advantage decreased as number of QTL decreased. However in this study, GBLUP had an advantage over BayesA at high number of QTL (300), but this advantage decreased by reducing the number of QTL. When the number of QTL affecting the trait is high, GBLUP is similar or even better than Bayesian methods (Daetwyler *et al.* 2010, Wimmer *et al.* 2013).

Habier *et al.* (2007) compared different methods of genomic breeding values evaluation and showed that Bayes method had high accuracy for any number of markers. In GBLUP method, equal variance in all markers is considered and it is no longer necessary to have preliminary information

on the variance of the markers effects (what is needed in the Bayes approach). This method is simpler than the Bayesian method and requires less computation. This method is also more affected by family relationships among people. Solberg *et al*. (2008) in a study used 1,000 phenotypic records in the reference group for a trait with a heritability of 0.5. They used Bayes method for estimating the effects of the markers and reported that the genomic evaluation accuracy of validation set is 0.66. This advantage could be due to the statistical method used. It has been reported in some studies that Bayes methods are better than the BLUP method.

In summary, the extent of linkage disequilibrium have major impact on the accuracy of MEBV. Based on the findings of this simulation study, low QTL number, as well as high dense marker panels, aiming to increase the level of linkage disequilibrium between markers and QTL, will likely be needed for successful implementation of genomic selection. To implement genomic selection with LD panels, a training population of sufficient size is necessary. By using a dense marker map covering all chromosomes, it is possible to accurately estimate the breeding value of animals that have no phenotypic record of their own and no progeny. The GBLUP method of analysis was as good as the BayesA method for traits influenced by a high number of QTL. BayesB produced estimates with higher accuracies in traits influenced by low number of QTL and with a gamma QTL effect distribution. Also in the low N_e, the BayesB method was more efficient than other two study methods. Higher accuracy can be obtained with Bayesian methods because this methods better takes into account the variable contribution of individual QTL. Based on this, Bayesian methods should be preferred over GBLUP.

REFERENCES

Daetwyler H D, Pong-Wong R, Villanueva B and Woolliams J A. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Journal of Animal Breeding and Genetics* **185**: 1021–31.

De los Campos G, Vazquez A I, Fernando R L and Daniel S. 2013b. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genetics* **7**(7): e1003608.

Goddard M E, Hayes B J and Meuwissen T H E. 2011. Using the genomic relationship matrix to predict the acuracy of genomic selection. *Journal of Animal Breeding and Genetics* **128**: 409–21.

Habier D, Fernando R L, Kizilkaya K and Garrick D J. 2007. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**: 186–93.

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

Sargolzaei M and Schenkel F S. 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**: 680–81.

Shirali M, Miraei-Ashtiani S R, Pakdel A, Haley C and Pong-Wong R. 2015. 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. *Iranian Journal of Animal Science* **5**(1):41–46.

Solberg T R, Sonesson A K, Woolliams J A and Meuwissen T H E. 2008. Genomic selection using different marker types and densities. *Journal of Animal Science* **86**: 2447–54.

Wimmer V, Lehermeier C, Albrecht T, Auinger H J, Wang Y and Schön C C. 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Journal of Animal Breeding and Genetics* **195**: 573–87.