Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

European Journal of Experimental Biology, 2013, 3(2):42-47

ISSN: 2248 –9215
CODEN (USA): EJEBAU

Accuracy of genomic prediction using RR-BLUP and Bayesian LASSO

Honarvar M.[1] and Rostami M.[2]

[1]*Department of Animal Science, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran*
[2]*Department of Animal Agriculture, Chaloos Branch, Islamic Azad University, Mazandaran, Iran*

ABSTRACT

We compared the accuracies of two genomic-selection prediction methods as affected by marker density and quantitative trait locus (QTL) number. Methods used to derive genomic estimated breeding values (GEBV) were random regression best linear unbiased prediction (RR–BLUP) and a Bayesian LASSO (Least Absolute Shrinkage and Selection Operator). In this study the genome comprised four chromosomes of 250 cM each. Also considering the number of markers 1000, 2000 and 5000 and the number of QTLs 4, 10, 20 and 40 and heritability of 5, 10 and 25 percent were compared.. In all scenarios Bayesian LASSO was more accurate than RR-BLUP, also increasing the number of QTLs, the evaluation accuracy decreases slightly which this reduction is greater in the lower heritability. The correlation between true breeding value and the genomic estimated breeding value in target generations applying RR-BLUP and Bayesian LASSO decreased from 0.918 to 0.807 and 0.933 to 0.847 respectively.

Key words: Accuracy, Genomic, RR-BLUP, Bayesian LASSO

INTRODUCTION

Genomic selection is a form of marker-assisted selection in which genetic markers covering the whole genome are used so that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker. The major limitation to the implementation of genomic selection has been the large number of markers required and the cost of genotyping these markers. Recently both these limitations have been overcome in most livestock species following the sequencing of the livestock genomes, the subsequent availability of hundreds of thousands of single nucleotide polymorphisms (SNP), and dramatic developments in SNP genotyping technology [1].

the availability of many thousands of SNPs spread across the genome for different livestock species opens up possibilities to include genome-wide marker information in prediction of total breeding values, to perform genomic selection. Compared to traditional breeding practice, including genomic information yields a considerable increase in selection responses for juvenile animals that do not have phenotypic records [2] and potentially can reduce the costs of a breeding program up to 90% [3]. As a result of these developments there are many livestock breeding companies planning to implement genomic selection in the near future. The purpose of this paper is to review the requirements for maximum benefits to be derived from genomic selection.

Under a SNPs markers whole-genome scans approach, many markers are likely to be located in regions that are not involved in the determination of traits of interest. On the other hand, some markers may be in linkage disequilibrium with some QTL, or in regions harboring genes involved in the infinitesimal component of the trait. This suggests that differential shrinkage of marker effects should be a feature of the model, then an alternative is the use of LASSO (Least Absolute Shrinkage and Selection Operator) regression, which provide good features of subset-selection (i.e., variable selection) with the shrinkage theory. de los Campos et al. [4] proposed a Bayesian approach of LASSO regression in genome-wide selection(GWS), and ever since, the success of this methodology has been reported by several authors [5].

MATERIALS AND METHODS

Statistical models

Two methods were used to estimate SNP effects: random regression BLUP (RR-BLUP) and Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) regression Other than the requirement that markers are located across the genome, no additional information, such as marker location or pedigree, is required by the methods. The basic model can be denoted as:

$$y_i = g(\boldsymbol{X_i}) + e_i$$

Where y_i is the Estimated breeding value (EBV), X_i is a 1 × p vector of SNP genotypes, $g(X_i)$ is a function relating genotypes to EBVs and e_i is a residual term. The SNP genotypes are coded as variates according to the number of copies of one SNP Allele, i.e. 0, 1 or 2. We denote with **X** the matrix containing the column vectors $\boldsymbol{X_k}$ of SNP genotypes at locus *k* (*k* =1, 2, ..., *p*).

RR-BLUP

In this method, SNP effects are assumed random [2], with $g(x_i)$ having the form:

$$g(x_i) = \sum_{k=1}^{p} x_{ik}\,\beta_k$$

Where β_k the effect associated with SNP k is, x_k is set up as described above for additive effects. The regression coefficients are found by solving the normal equations,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X} + \boldsymbol{I\lambda})^{-1}\boldsymbol{X'y}$$

Where λ is constant for all SNPs. Differences in shrinkage between SNP still arise as a result of variation in allele frequency. Meuwissen et al. [2] and Habier et al. [6]. have calculated λ for their simulated data from known genetic and residual variances. With no knowledge of these variance components and analyzing EBV data, an appropriate value for the shrinkage parameter can be obtained by cross-validation. When EBV have a variety of reliabilities then the regression can be weighted accordingly so that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'R^{-1}X} + \boldsymbol{I\lambda})^{-1}\boldsymbol{X'R^{-1}y}$$

Where **R** is a diagonal matrix of Weights. they were treated as homogeneous, i.e. **R** = **I**.

Bayesian Lasso Regression

Given phonotypical measurements and genotype information, we could obtain the preconditioned response $y^{\sim}$ based on the generic form of linear regression. However in genome-wide association studies, a number of covariates, which are either discrete or continuous, may be measured for each subject. In order to estimate genetic effects precisely by adjusting for these covariates, a GWAS model that takes into account the effects of important covariates would be more appropriate. Therefore, we describe the preconditioned value $y^{\sim}_i$ of a quantitative trait for subject i as

$$\tilde{y}_i = \mu + \boldsymbol{X_i^T}\alpha + \boldsymbol{Z_i^T}\beta + \boldsymbol{\xi_i^T}\,\boldsymbol{a} + \boldsymbol{\zeta_i^T}\boldsymbol{d} + \epsilon_i, \qquad i = 1, \dots, n$$

Where μ is the overall mean, Xi is the *d1-dimensional* vector of discrete covariates for subject i, $\alpha = (\alpha 1, \cdots, \alpha d1)^T$ is the vector of regression coefficients for discrete covariates, Zi is the d2 -dimensional vector of continuous covariates for subject *i*, $\beta = (\beta 1, \cdots, \beta d2)^T$ is the vector of regression coefficients for continuous covariates,

$a = (a1, \cdots, ap)^T$ and $d = (d1, \cdots, dp)^T$ are the p-dimensional vectors of the additive and dominant effects of SNPs, respectively, ξ_i and ζ_i are the indicator vectors of the additive and dominant effects of SNPs for subject i, and i is the residual error assumed to follow a $N(0, \sigma^2)$ distribution. The *j-th* elements of ξ_i and ζ_i are defined as

$$\xi_{ij} = \begin{cases} 1, if\ the\ genotype\ of\ SNP\ j\ is\ AA \\ 0, if\ the\ genotype\ of\ SNP\ j\ is\ Aa \\ -1, if\ the\ genotype\ of\ SNP\ j\ is\ aa \end{cases}$$

$$\zeta_{ij} = \begin{cases} 1, if\ the\ genotype\ of\ SNP\ j\ is\ Aa \\ 0, if\ the\ genotype\ of\ SNP\ jis\ AA\ or\ aa \end{cases}$$

The Bayesian lasso is implemented with a hierarchical model, in which scale mixtures of normal are used as prior distributions for the genetic effects and exponential priors are considered for their variances, and then solved by using the Markov chain Monte Carlo (MCMC) algorithm. Our approach obviates the choice of the lasso parameter by imposing a diffuse hyper-prior on it and estimating it along with other parameters and is particularly powerful for selecting the most relevant SNPs for GWASs where the number of predictors exceeds the number of observations[7].

Table1. The parameters of the simulated genetic model

Number of chromosomes	4
Mutation rate	2.5×10^{-5}
Distribution of additive mutational effects	Gamma(1.66, 0.4)
Dominance of QTL effects	0
Population structure	
Generations 1–1000	Ideal*, N = 100
Generation 1001	Ideal*, N= 200
Generation 1002	20 half-sib families, N = 2000
Generation 1003 and later	Ideal*, N =2000
Marker genotyping	Generations 1001 and later
Phenotypic recording	Generations 1001 and 1002

* *Ideal denotes a population structure where the effective size equals the actual population size. This structure is simulated by giving every male (female) in generation t–1 an equal probability of becoming the sire (dam) of animal i in generation t, which implies no selection and random mating of males and females [2].*

The data Simulation: Data sets with heritability of 5, 10 and 25 percent at different marker densities were simulated to allow comparison of the different models, in terms of accuracy of predicted breeding values. An effective population size of 100 animals was simulated, of which half of the animals were female and the other half male. This structure was kept constant for 1000 generations. Mating was performed by drawing the parents of an animal randomly from the animals of the previous generation. The considered genome comprised four chromosomes of 250 cM each. The number of segregating QTL affecting the trait was set at 4, 10, 20 and 40 and the number of markers was 1000, 2000 and 5000 for the traits with heritability of 0.05 and 0.25. QTL loci were randomly determined, with all possible positions on the genome having equal chance Simulating a whole All marker loci with a minor allele frequency in generations 1001-1003 of 0.02 were discarded. Different marker densities were created for each simulated data set. To arrive at a mutation-drift balance, populations were simulated for 1000 generations at an effective size of 100. After these 1000 generations, the actual size of the populations was increased, to 200 (100 males and100 females) in generation 1001, and to 2000 (20 half-sib families of size 100 each) in generations 1002 and 1003. The animals in generations 1001 and 1002 were marker genotyped and recorded for the trait. Phenotypic records were obtained by adding a normally distributed error term with variance 1 to the genetic value of the individuals. The 2000 animals of generation 1003 are assumed to be juveniles that did not (yet) have

phenotypic records and their breeding values were estimated using marker information only. The breeding value of each animal was the sum of the effects of the QTL alleles that it carried. To obtain the phenotype, we added a normal error deviate with variance calculated to achieve the desired heritability. The statistical methods will be compared for their accuracy of predicting the true genetic values of the animals in generation 1003 [2].

RESULTS AND DISCUSSION

The correlation between the actual and the estimated breeding values in the generation of 1003, considering the number of QTLs 4, 10, 20 and 40 and the number of markers 1000, 2000 and 5000 for the trait with heritability of 0.25 is presented in Table 2. Regarding to Table 2, by increasing the number of markers the evaluation accuracy is increased and by decreasing it the accuracy decreases. By increasing the number of QTLs the accuracy of evaluation decreased slightly. For example, when the number of markers was considered to be 1000, by increasing the number of QTLs from 4 to 40, the accuracy of evaluation using RR-BLUP and Bayes methods decreased from 0.875 to 0.859 and 0.889 to 0.870 respectively. On the other hand, by increasing the number of markers the effect of the number of QTLs on evaluation accuracy decreased.

Table2. The correlation between the actual and the estimated breeding values in the generation of 1003 with different number of marker and QTL

QTL	SNP	RR-BLUP	Bayes
4	1000	0.8751	0.8892
	2000	0.9024	0.9184
	5000	0.9111	0.9299
10	1000	0.8779	0.9059
	2000	0.9091	0.9247
	5000	0.9179	0.9330
20	1000	0.8721	0.8956
	2000	0.9114	0.9261
	5000	0.9171	0.9376
40	1000	0.8590	0.8707
	2000	0.8994	0.9136
	5000	0.9205	0.9335

The values of evaluation accuracy using RR-BLUP and Bayes methods by considering 1000, 2000 and 5000 markers and for heritability of 5, 10 and 25 percent are shown in Table 3. From Table 3 one can see that the evaluation accuracy of both methods increases by increasing the number of markers and heritability and vice versa. The rate of variation in the evaluation accuracy related to the number of markers in lower heritability is less than high heritability.

Table 3. The correlation between the actual and the estimated breeding values in the generation of 1003 with different number of marker and heritability

SNP	h^2	RR-BLUP	Bayes
1000	5	0.8043	0.8130
	10	0.8398	0.8576
	25	0.8779	0.9059
2000	5	0.8400	0.8424
	10	0.8678	0.8820
	25	0.9091	0.9247
5000	5	0.8461	0.8534
	10	0.8758	0.9031
	25	0.9179	0.9330

Meuwissen et al. [2] presented that the greater the number of markers, the markers effects will be satisfied more accurately and as a result the evaluation accuracy of genomic breeding values will be higher. These results are consistent with the results in this work. On the other hand, Nilson et al. [8] presented that by increasing the trait

heritability from 0.2 to 0.4, the genomic evaluation accuracy increases by about 4 percent which is consistent with the results from Kolbehdari et al.[9] but is contrary to expectations of Meuwissen et al [2].

The evaluation accuracy values by RR-BLUP and Bayes methods for 4, 10, 20 and 40 QTLs and 5, 10 and 25 percent heritability are shown in Table 4. As presented, increasing the number of QTLs, the evaluation accuracy decreases slightly which this reduction is greater in the lower heritability.

Table 4. The correlation between the actual and the estimated breeding values in the generation of 1003 with different number of QTL and heritability

QTL	h^2	RR-BLUP	Bayes
4	5	0.8389	0.8502
	10	0.8691	0.8788
	25	0.9111	0.9299
10	5	0.8461	0.8534
	10	0.8758	0.9031
	25	0.9179	0.9330
20	5	0.8480	0.8562
	10	0.8762	0.8919
	25	0.9171	0.9376
40	5	0.8415	0.8589
	10	0.8637	0.8967
	25	0.9205	0.9335

The accuracy rate in the target generation (1003 to 1009) using BLUP and Bayes methods are shown in Table 5. The number of markers and QTLs are considered to be 5000 and 10 respectively. As expected, in both cases, the evaluation accuracy rate decreases across the generations.

Table 5. The correlation between the actual and the estimated breeding values in the generation of 1003 -1009

Generations	RR-BLUP	Bayes
1003	0.9179	0.9330
1004	0.8795	0.9121
1005	0.8522	0.8928
1006	0.8392	0.8776
1007	0.8242	0.8693
1008	0.8101	0.8540
1009	0.8074	0.8476

The correlation between true breeding value and the genomic estimated breeding value in target generations (1003 to 1009)applying RR-BLUP and Bayesian LASSO decreased from 0.918 to 0.807 and 0.933 to 0.847 respectively. The use of phenotypic and family tree information is a common method in animal breeding. The use of genetic information of individuals in molecular level to increase genetic progress by reducing the generation gap and improving the evaluation accuracy of breeding values is one of the main goals of modern biological technologies in animal breeding [10]. In general, the genomic selection is a form of selection assisted by marker which uses the genetic markers covering the whole genome [11] and it makes a notable increase , in comparison with conventional breeding, in response to a selection of young animals that do not have sufficient phenotypic records from relatives [2].

Increasing of the genomic data increases available information for genetic evaluation. This information contains molecular genotypes in loci. The observed genotypes in markers are available and thus the model can include the unknown effect of each of the markers. In this study, BLUP mixed model and Bayes method were used to evaluate the markers effects. In BLUP model, presented by Meuwissen et al. [2] a constant and equal variance is attributed to all of the loci. It is also assumed that each marker has a small effect and no marker has a very large effect. Many studies have shown that Bayes method is more accurate in comparison with BLUP which is consistent with the present results [12]

Habier et al [6] compared different methods of genomic breeding values evaluation and showed that Bayes method has high accuracy for any number of markers. In BLUP 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.

Meuwissen et al [2] in a simulation study, used 500 phenotypic records in the reference group in order to estimating the genomic breeding values of people in validation set for a trait with heritability 0.5 and they reported that the accuracy of single trait genomic evaluation by using BLUP method is 0.57. Solberg et al [12] in a study used 1000 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 [13].

CONCLUSION

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. For evaluation of traits which collecting the phenotypic records from them is difficult or impossible, such as traits that only appear in females, disease resistance and traits with low heritability, using the Genomic selection is suggested. Methods that assumed a prior distribution for the variance associated with each chromosome segment gave more accurate predictions of breeding values even when the prior was not correct.

REFERENCES

[1] Park T., and Casella, G. *J. Am. Stat. Assoc*, **2008**,103,681-686.
[2] Meuwissen T.H.E., Hayes B.J., Goddard M.E, *Journal of Genetics*, **2001**, 177, 1819-1829.
[3] Schaeffer, L. R. *J, Anim. Breed. Genet*, **2006**, 123,218–223.
[4] Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes, *Genetics*,***2009***, 182(1),375-385.
[5] Silva, F.F. et al, *Livestock Science*,***2011*** (10.1016/j.livsci.2011.09.010)
[6] Habier D., Fernando R. L., Dekkers J.C.M, *Journal of Genetics*, **2007**, 177, 2389-2397.
[7]Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li and Rongling Wu, *Bioinformatics*, **2011,** . 27 (4), 516-523**.**
[8] Nielsen H.M., Sonesson A.K., Yazdi H., Meuwissen T.H.E, *Aquaculture*,**2009**, 289,259–264.
[9] Kolbehdari D., Schaeffer L. R., Robinson J. A. B, *Journal of Animal Breeding and Genetics*, **2007**, 124,356 – 361.
[10] VanRaden P., Van Tassell C., Wiggans G., Sonstegard T., Schnabel R., Taylor J., Schenkel F, *Journal of Dairy Science*, **2009**, 92,16-24.
[11] Goddard M.E., Hayes B.J, *Journal of Animal Breeding and Genetics*, **2007**, 124, 323 – 330.
[12] Solberg T.R., Sonesson A.K., Woolliams J.A., Meuwissen T.H.E, *Journal of Animal Science*, **2008**, 86(10),41-53.
[13] Hayes B.J., Bowman P.J., Chamberlain A., Goddard M.E, *Journal of Dairy Science*, **2009**, 92,433-443.