

## Short communication

## A note on accuracy of Bayesian LASSO regression in GWS

Fabyano Fonseca Silva <sup>a,\*</sup>, Luis Varona <sup>b</sup>, Marcos Deon V. de Resende <sup>c</sup>, Júlio Sílvia S. Bueno Filho <sup>d</sup>,  
Guilherme J.M. Rosa <sup>e</sup>, José Marcelo Soriano Viana <sup>f</sup>

<sup>a</sup> Departamento de Estatística, Universidade Federal de Viçosa, Av. P.H. Hoff, 36570–000, Viçosa, Brazil

<sup>b</sup> Unidad de Genética Cuantitativa y Mejora Animal. Facultad de Veterinaria, Universidad de Zaragoza, 50013 Zaragoza, Spain

<sup>c</sup> Embrapa Florestas/Universidade Federal de Viçosa, Estrada da Ribeira, km 111, 83411–000, Colombo, Brazil

<sup>d</sup> Departamento de Ciências Exatas, Universidade Federal de Lavras, Campus da UFLA, 37200–000, Lavras, Brazil

<sup>e</sup> Department of Animal Science, University of Wisconsin, 460 Animal Science Building, Madison, USA

<sup>f</sup> Departamento de Biologia Geral, Universidade Federal de Viçosa, Av. P.H. Hoff, 36570–000, Viçosa, Brazil

## ARTICLE INFO

## Article history:

Received 27 April 2011

Received in revised form 5 September 2011

Accepted 9 September 2011

## Keywords:

Genome wide selection

Penalized regression

SNP markers

## ABSTRACT

Several genome wide selection (GWS) statistical methods have been proposed in the last years, and among these stands out the Bayesian LASSO (BL), which is a penalized regression method based on the regularization parameter ( $\lambda$ ) estimates. In general, the posterior mean values for  $\lambda$  are those that minimize the residual sum of squares (RSS) while controlling the L1 norm (absolute values) of the regression coefficients. However, another option is to use fixed values of  $\lambda$ , which is independent of this minimization process. Nevertheless, the most important aim of GWS is to make predictions about genomic breeding values (GBV =  $u$ ) for individuals that have not been measured directly for the trait, and for this reason the parameter to maximize should be the accuracy ( $r_{u,\hat{u}}$ ). Thus, a question can arise as to whether such estimated  $\lambda$  values that minimize RSS are the same as that which maximize  $r_{u,\hat{u}}$ . In order to answer this question, this paper aims to provide methodological and computational resources in order to evaluate the influence of BL regularization parameter estimates on the correlation between true and estimated GBV (accuracy) depending on genetic structure of the target trait (few or many QTLs and low or medium heritability). In general, it is possible to report, on average, that GBV prediction is robust in relation to the  $\lambda$  estimation, since the different values for  $\lambda$  lead to similar accuracy values. Moreover, the fixed  $\lambda$  values grid request high computational costs, implying that the random  $\lambda$  method is more attractive, since it is much faster to use just one Gibbs sampler run, while the grid must to use one run for each fixed  $\lambda$  value.

© 2011 Elsevier B.V. All rights reserved.

## 1. Introduction

Recently, several genomic wide selection (GWS) methods have been developed based upon the predictive ability of marker effects, and consequently enabling the genomic breeding value (GBV) estimates of livestock animals. According to de los Campos et al. (2009), with whole-genome scans, 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, and then these authors proposed the use of Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) regression, which provides good features of subset-selection with the shrinkage theory.

In relation to the LASSO history, the original method was proposed by Tibshirani (1996), and the Bayesian approach was presented by Park and Casella (2008) and modified and applied to GWS by de los Campos et al. (2009), and ever

\* Corresponding author. Tel.: +55 31 38991790; fax: +55 31 38993880.

E-mail addresses: [fabyanofonseca@ufv.br](mailto:fabyanofonseca@ufv.br) (F.F. Silva), [lvarona@unizar.es](mailto:lvarona@unizar.es) (L. Varona), [deon@cnpf.embrapa.br](mailto:deon@cnpf.embrapa.br) (M.D.V. de Resende), [jssbueno@ufla.br](mailto:jssbueno@ufla.br) (J.S.S.B. Filho), [grosa@wisc.edu](mailto:grosa@wisc.edu) (G.J.M. Rosa), [jmsviana@ufv.br](mailto:jmsviana@ufv.br) (J.M.S. Viana).

since, the success of this methodology has been reported by several authors (Cleveland et al., 2010; de los Campos et al., 2010; Pérez et al., 2010) in different fields of science.

In summary, the LASSO popularity is due in part to a key feature of the procedure: shrinkage of the vector of regression coefficients toward zero with the possibility of setting some coefficients identically equal to zero, resulting in a simultaneous estimation and variable selection procedure (Hans, 2009). Actually, LASSO is a penalized regression method, and the intensity of the penalization is given by the regularization parameter values, widely known by  $\lambda$  ( $\lambda \geq 0$ ).

In respect to  $\lambda$ , Park and Casella (2008) and Yi and Xu (2008) recognized that the performance of the Bayesian LASSO (BL) depends critically on the tuning of  $\lambda$ . Thus, different prior distributions have been proposed, like conjugate Gamma (Park and Casella, 2008) and Beta (de los Campos et al., 2009), and the posterior mean values are those that minimize the residual sum of squares (RSS) while controlling the L1 norm (absolute values) of the regression coefficients. However, other option to these prior, is to use a mass-point at some value (i.e., fixed values) of  $\lambda$ , which is independent of this minimization process.

Since one of the most important aims of GWS is to make predictions about GBV for individuals that have not been measured directly for the trait, the accuracy, given by the correlation ( $r_{u,\hat{u}}$ ) between the true and estimated GBV ( $u$ ) must be maximized. Thus, a question can arise as to whether such estimated  $\lambda$  values that minimize RSS are the same as that which maximize  $r_{u,\hat{u}}$ . However, in some cases, larger  $r_{u,\hat{u}}$  may occur simultaneously with larger bias, so in order to point to this problem the mean squared error (MSE) may be used:  $MSE = (1/N) \sum_{i=1}^N (u_i - \hat{u}_i)^2$ .

This paper aims to provide methodological and computational resources in order to evaluate the influence of BL regularization parameter estimates on the correlation between true and estimated GBV (accuracy) depending on genetic structure of the target trait (few or many QTLs and low or medium heritability).

## 2. Material and methods

### 2.1. Simulated data set

A population with an effective size of 100 was simulated for 1000 generations. After 1000 generations, the actual size of the population increased up to 1000 (500 per sex) and remained at 1000 for three discrete and consecutive generations. During the whole process, all individuals were generated with one gamete from a random father and one from a random mother. Therefore the data set for the estimation of the marker effects consisted of the 3000 individuals from the last three generations (generations 1001, 1002 and 1003).

The genome was assumed to consist of 10 chromosomes each 100 cM long and 1000 loci/chromosome were located at random map positions. Mutations were generated at a rate of  $2.5 \times 10^{-3}$  per locus per generation at the marker loci and at a rate of  $2.5 \times 10^{-5}$  at the QTL loci following Meuwissen et al. (2001).

Under the assumption of biallelic SNP and QTL, two distinct situations were considered: 9995 SNP markers plus 5 QTLs, and

9950 SNP markers plus 50 QTLs. In both situations, the effect of each QTL ( $a_j$ ) was given by:  $a_j = \sqrt{\sigma_a^2/2pq}$ , being  $p$  the allele frequency,  $q = 1 - p$  and  $\sigma_{ai}^2$  values were generated from  $\chi^{-2}(v, S)$  distribution (Toro and Varona, 2010) using the *rvinvchisq* R (R Development Core Team, 2010) function. The total additive genetic variance ( $\sigma_a^2$ ) was  $\sigma_a^2 = \sigma_{a1}^2 + \sigma_{a2}^2 + \dots + \sigma_{aQ}^2 = 17.5$ , being  $Q$  the number of QTLs (5 or 50). For  $Q = 5$  was generated 5 random numbers from  $\chi^{-2}(5, 5)$ , whose sum was 17.5, and for  $Q = 50$  was generated 50 random numbers from  $\chi^{-2}(5, 0.5)$ , whose sum was 17.5. Under this approach, were considered other two distinct situations related to low ( $h^2 = 0.1$ ) and medium ( $h^2 = 0.3$ ) heritabilities, whose residual variance ( $\sigma_e^2$ ) were, respectively, 175 and 40. Thus, each one of the 3000 individuals was genotyped ( $x_{ij} = 1, 0, -1$  for the SNP genotypes AA, Aa and aa at each locus, respectively) and phenotyped, being the phenotype value of each individual  $i$  given by:

$$y_i = u_i + e_i = \sum_{j=1}^{1000} x_{ij} a_j + e_i, e_i \sim N(0, \sigma_e^2). \quad (1)$$

### 2.2. Regression model and cross validation

Under a matrix notation, the presented model (1) can be rewritten as:  $y = X\alpha + e$ , where  $y = [y_i]$ ,  $X = [x_{ij}]$ ,  $\alpha = [a_j]$  and  $e = [e_i]$ . In general terms, the Bayesian LASSO is a penalized procedure that minimizes RSS subject to the non-differentiable constraint expressed in terms of the norm of the coefficients, being the estimator (posterior means) given by:  $\hat{\alpha}_L = \arg \min_{\beta} (\hat{y} - X\alpha)'(\hat{y} - X\alpha) + \lambda \sum_{j=1}^{1000} |\alpha_j|$ . We can see that the regularization parameter  $\lambda$  plays a central role, and under a Bayesian perspective, the  $\lambda$  estimation depends on prior distributions. In the present study, the Park and Casella (2008) prior,  $\lambda^2 \sim \text{Gamma}(\alpha_1, \alpha_2)$ , was used and compared with the mass-point at some value, i.e. to elicit a fixed value for  $\lambda$ .

Since the idea behind using SNP technology in animal breeding is reasoned on the assumption that the GBV can be accurately predicted based on their genotypes through estimated SNP effects, the objective function is the correlation ( $r_{u,\hat{u}}$ ) between the true and estimated GBV, which must be maximized. Under a BL approach, the estimated  $\lambda$  values are those that minimize the residual sum of squares (RSS) but not necessarily those that maximize  $r_{u,\hat{u}}$ . Thus, fixed  $\lambda$  values, selected from a set of values by either statistical method, for example a cross validation, should result in higher  $r_{u,\hat{u}}$  values.

About  $r_{u,\hat{u}}$ , it must be clear that in practical terms the true value  $u$  is unknown, so the observed phenotypes (adjusted for fixed and/or polygenic effects) of the validation data sets can be used in replacement of  $u$ . In this way, as in the present study the main objective was to evaluate the Bayesian LASSO robustness to  $\lambda$  estimation, the use of  $u$  makes sense.

In order to analyze this hypothesis, a simplistic simulation study was performed. This one was composed according to genetic structure of the target trait, considering four different scenarios: 5 QTLs and low heritability,  $h^2 = 0.1$  (I); 50 QTLs and low heritability,  $h^2 = 0.1$  (II); 5 QTLs and medium heritability,  $h^2 = 0.3$  (III); and 50 QTLs and medium heritability,  $h^2 = 0.3$  (IV). For each one of these scenarios, the  $r_{u,\hat{u}}$  values were calculated obeying a 3-fold cross validation methodology, being  $r_{u,\hat{u}}$  calculated using  $u$  and  $\hat{u}$  from validation data set presented above.

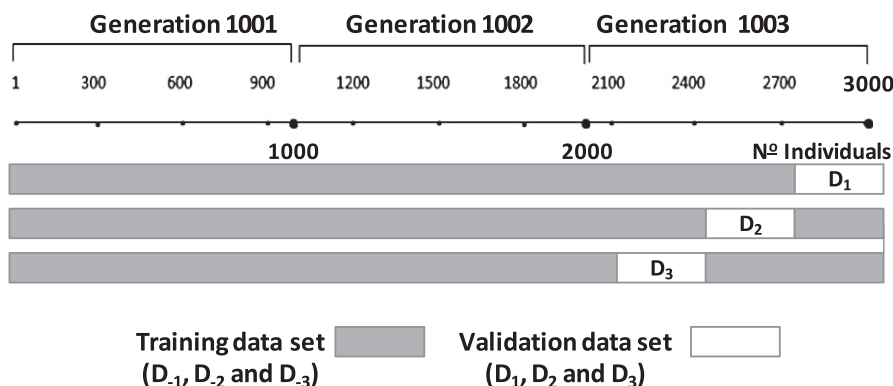


Fig. 1. Illustrative scheme of the data sets used in the cross validation analysis.

Without loss of generality for each considered scenario, the 900 last individuals from generation 1003 were divided into 3 data sets ( $D_k$ ) of 300 individuals,  $D_1$ ,  $D_2$  and  $D_3$ , which constituted the validation data sets. So, each one of the three training data sets, with 2700 individual, was given by  $D_{-k}$ , denoting that the  $D_k$  was removed (Fig. 1). The justification is that it does not make sense to use information of the generation 1003 to predict phenotype in the generations 1001 and 1002, since there is no interest on predicting the past.

The BLR (de los Campos et al., 2009) package of R software (R Development Core Team, 2010) was used in order to implement the BL regression for each one of the  $D_{-k}$  data set in each considered scenario (I, II, III and IV). The regression model was fitted by two distinct ways, estimating and fixing the  $\lambda$  values. In the first, the Park and Casella (2008) Gamma prior was used, and in the second way, fixed values of  $\lambda$ , varying between 1 and 290 (1, 10, 20, ..., 290), were used. Altogether, for each data set of each scenario, 31 analyses were realized, 30 with fixed and 1 with estimated  $\lambda$  values.

From the estimates of additive SNP effects, breeding values ( $u_i$ ) were calculated for the 300 individuals of the validation data set according to Falconer and Mackay (1996):  $\hat{u}_i = \sum_{j=1}^{1000} [I(x_{ij} = 1)(2q_j\hat{a}_j) + I(x_{ij} = 0)(q_j\hat{a}_j - p_j\hat{a}_j) + I(x_{ij} = -1)(-2p_j\hat{a}_j)]$ , where  $x_{ij}$  is an indicator function of the genotype of the  $j^{\text{th}}$  marker of the  $i^{\text{th}}$  individual that takes the values 1, 0, -1 when the genotypes are AA, Aa or aa, respectively. Moreover,  $p_j$  and  $q_j$  are the allelic frequencies (A or a) for the  $j^{\text{th}}$  marker in the training population and  $\hat{\beta}$  is the estimated additive SNP effects from this same population. Since the true GBV ( $u_i$ ) is

known in each validation data set, the accuracy (simple correlation between  $u_i$  and  $\hat{u}_i$ ) can be obtained for each one of these data sets and each one of the scenario and  $\lambda$  values.

### 3. Results and discussion

We can note in Table 1 that the differences between  $\hat{r}_{u,\hat{u}}$  values for all scenarios were slight, thus it is possible to report, on average, that the GBV prediction is robust in relation to the regularization parameter estimation, since the different values for  $\lambda$  lead to almost the same values of  $\hat{r}_{u,\hat{u}}$ . This fact can be illustrated by the accuracy curves in Fig. 2, which can be considered flat, mainly in the presence of a small number of genes (5 in this context). Furthermore, the MSE were proportional to  $\hat{r}_{u,\hat{u}}$ , showing that the problem related with larger  $\hat{r}_{u,\hat{u}}$  values occurring simultaneously with larger bias has not been verified.

In relation to heritability, although the results showed close accuracy values, the highest differences were observed for low heritability scenarios (I and II), being in these cases the accuracy of fixed  $\lambda$  method higher than those from random  $\lambda$  method. With respect to this fact, Jiménez-Montero et al. (2010) also had related better performance of Bayesian LASSO with random  $\lambda$  estimation for higher heritabilities.

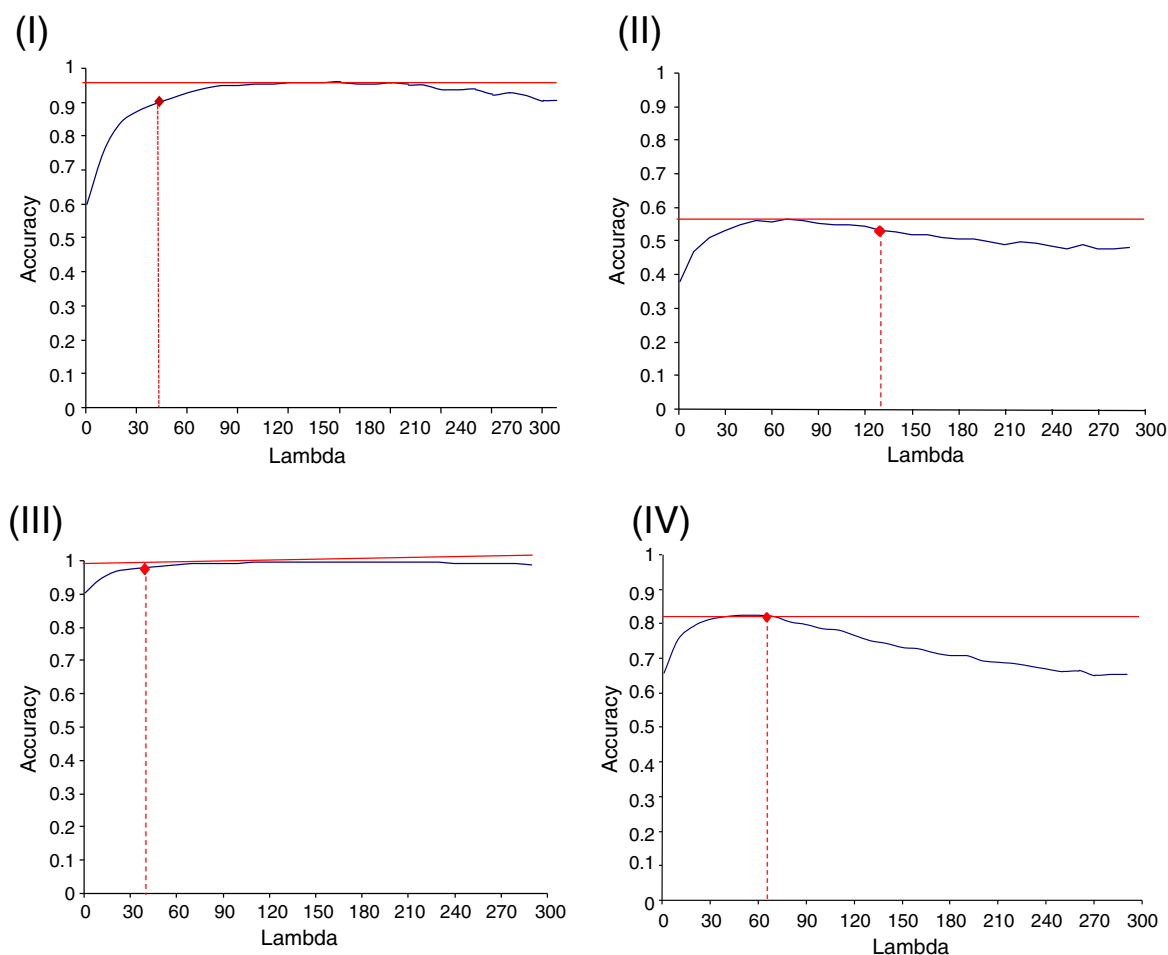
Another relevant feature of Table 1 is that the estimated  $\lambda$  values from random method were smaller than the best fixed  $\lambda$  values for the 5 gene scenarios (I and III), and on the other hand, for the 50 gene scenarios (II and IV) the opposite behavior was observed. In general terms, it can be explained by the fact that for small number of significant genes the

Table 1

Means and standard deviations for regularization parameter ( $\lambda$ ), accuracy ( $\hat{r}_{u,\hat{u}}$ ) and MSE values calculated from 3 data sets of the cross validation study.

| Scenarios | Random parameter      |                       |                      | Fixed parameter      |                       |                       |
|-----------|-----------------------|-----------------------|----------------------|----------------------|-----------------------|-----------------------|
|           | $\hat{\lambda}$       | $\hat{r}_{u,\hat{u}}$ | MSE                  | $\lambda_{\max}$     | $\hat{r}_{u,\hat{u}}$ | MSE                   |
| I*        | 43.3567<br>(5.9876)   | 0.9088<br>(0.0430)    | 40.95<br>(10.5601)   | 131.1600<br>(1.1435) | 0.9614<br>(0.0456)    | 88.7421<br>(10.3452)  |
| II        | 131.0234<br>(19.3456) | 0.5319<br>(0.0633)    | 76.0987<br>(16.9876) | 69.8760<br>(1.9876)  | 0.5720<br>(0.0976)    | 124.0980<br>(27.9877) |
| III       | 41.1914<br>(3.6999)   | 0.9846<br>(0.0113)    | 41.1645<br>(8.2309)  | 149.8740<br>(1.9765) | 0.9867<br>(0.0076)    | 80.1254<br>(12.9000)  |
| IV        | 63.0802<br>(7.2345)   | 0.8149<br>(0.0654)    | 59.1264<br>(7.0987)  | 52.6785<br>(2.0098)  | 0.8106<br>(0.0108)    | 73.9807<br>(11.0198)  |

\* I: 5 genes and  $h^2 = 0.1$ ; II: 50 genes and  $h^2 = 0.1$ ; III: 5 genes and  $h^2 = 0.3$ ; IV: 50 genes and  $h^2 = 0.3$ .



**Fig. 2.** Behavior of accuracy ( $r_{u,\hat{u}}$ ) as a function of lambda values (regularization parameter) according to the four scenarios (I, II, III and IV). The vertical line indicates the estimates for random regularization parameter ( $\hat{\lambda}$ ).

shrinkage effect was suppressed under a random  $\lambda$  approach, since the operation field of this parameter is reduced, i.e. there are few variables to be selected. In this context, when 50 genes were considered, the estimation process was able to impose its shrinkage power.

Still in relation to  $\lambda$  magnitude, in theory, higher values imply stronger penalization, which results in more marker regression coefficients being shrunk toward zero (Hans, 2009), so being the values of  $\hat{r}_{u,\hat{u}}$  very similar between the two methods (random and fixed  $\lambda$ ), this penalization does not affect the BL prediction, reinforcing the robustness of Bayesian LASSO.

In summary, the presented results are in agreement with those obtained by Pérez et al. (2010), who indicate that the fully Bayesian analysis (posterior means of  $\lambda$ ) was as good as the best obtained when cross-validation was used for choosing  $\lambda$ , being at the neighborhood of the optimal values found from a grid of values. Thus, in general, in the present study also it is possible to report that the fixed  $\lambda$  values grid request high computational costs, implying that the random  $\lambda$  method is more attractive, since it is much faster to use just one Gibbs sampler run, while the grid must use one run for each fixed  $\lambda$  value.

In relation to BL predictive ability, Cleveland et al. (2010) using a simulated dataset showed that the BL with random  $\lambda$  produced accuracy values of 0.9166 when all individuals were used in the analysis, and, on average terms, these results are in agreement with those presented in the present study (scenarios I and III). On the other hand, Legarra et al. (2010) used random  $\lambda$  BL in the analysis of real data of Holstein bulls genotyped with the Illumina Bovine SNP50 Bead-Chip, and obtained accuracy values ranging from 0.30 for protein yield to 0.73 for fat percentage. These results are approximately in line with ours concerning scenarios II and IV, which are associated with a large number of genes controlling the trait.

In order to finish the Table 1 discussion, as commented earlier (Section 2.2) in real situations the true value  $u$  is unknown, so the results about the accuracy of random or fixed  $\lambda$  values must be discussed in terms of adjusted phenotypes using cross validation approaches.

#### Appendix A. Supplementary data

Supplementary data to this article can be found online at [doi:10.1016/j.livsci.2011.09.010](https://doi.org/10.1016/j.livsci.2011.09.010).

## References

- Cleveland, M.A., Forni, S., Deeb, N., Maltecca, C., 2010. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. *BMC Proc.* 4 (Suppl 1), S6.
- 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, 375–385.
- de los Campos, G., Gianola, D., Allison, D.B., 2010. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* 11, 880–886.
- Falconer, D.S., Mackay, T.F.C., 1996. *Introduction to Quantitative Genetics*. Longman, London.
- Hans, C., 2009. Bayesian LASSO regression. *Biometrika* 96, 835–845.
- Jiménez-Montero, J.A., Gonzalez-Recio, O., Alenda, R., 2010. Genotyping strategies for genomic selection in dairy cattle. *Proceedings of the 10th WCGALP*, Leipzig, Germany, pp. 2–146.
- Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., Ducrocq, V., 2010. Aptitude of Bayesian Lasso for genomic selection. *Proceedings of 9th WCGALP*, pp. 1–8.
- Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.
- Park, T., Casella, G., 2008. The Bayesian LASSO. *J. Am. Stat. Assoc.* 103, 681–686.
- Pérez, P., de los Campos, G., Crossa, J., Gianola, D., 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome* 3, 106–116.
- R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B* 58, 267–288.
- Toro, M.A., Varona, L., 2010. A note on mate allocation for dominance handling in genomic selection. *Genet Sel Evol.* 42, 33.
- Yi, N., Xu, S., 2008. Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179, 1045–1055.