

# Bayesian model combining linkage and linkage disequilibrium analysis for low density-based genomic selection in animal breeding

Fabyano Fonseca Silva<sup>a</sup>, Elcer Albenis Zamora Jerez<sup>a</sup>, Marcos Deon Vilela de Resende<sup>b</sup>, José Marcelo Soriano Viana<sup>c</sup>, Camila Ferreira Azevedo<sup>d</sup>, Paulo Sávio Lopes<sup>a</sup>, Moysés Nascimento<sup>d</sup>, Rodrigo Oliveira de Lima<sup>e</sup> and Simone Eliza Facioni Guimarães<sup>a</sup>

<sup>a</sup>Department of Animal Science, Universidade Federal de Viçosa, Viçosa, Brazil; <sup>b</sup>EMBRAPA Forestry Research, Colombo, Brazil; <sup>c</sup>Department of General Biology, Universidade Federal de Viçosa, Viçosa, Brazil; <sup>d</sup>Department of Statistics, Universidade Federal de Viçosa, Viçosa, Brazil; <sup>e</sup>Department of Plant Science, Universidade Federal de Viçosa, Viçosa, Brazil

## ABSTRACT

We combined linkage (LA) and linkage disequilibrium (LDA) analyses (emerging the term 'LALDA') for genomic selection (GS) purposes. The models were fitted to a simulated dataset and to a real data of feed conversion ratio in pigs. Firstly, the significant QTLs (quantitative trait locus) were identified through LA-based mixed models considering the QTL-genotypes as random effects by means of genotypic identity by descent matrix. This matrix was calculated at the positions of significant QTLs (based on LA) allowing to include the QTL-genotype effects additionally to SNP (single nucleotide polymorphism) markers (based on LDA) and additive polygenic effects in several GS models (Bayesian Ridge Regression – BRR; Bayes A – BA; Bayes B – BB; Bayes C – BC and Bayesian LASSO – BL). These models combining all mentioned effects were denominated LALDA. Goodness-of-fit and predictive ability analyses were performed to evaluate the efficiency of these models. For the real data, although slightly, the superiority of the LALDA models was verified in comparison to traditional LDA models for GS. For the simulated dataset, the models presented similar results. For both LDA and LALDA frameworks, BA showed the best fitting through Deviance Information Criterion and higher predictive ability in the simulated and real datasets.

## ARTICLE HISTORY

Received 4 May 2017  
Accepted 5 December 2017

## KEYWORDS

Genetic architecture; genome prediction; genotypic IBD matrix; SNP marker

## 1. Introduction

Linkage disequilibrium analysis (LDA) is the basic concept behind genomic selection (GS), since the associations between markers and QTLs (quantitative trait locus) is an attribute of the population as a whole. For this reason, it is expected that these associations be shared by all individuals of the population and be preserved for several generations. On the other hand, linkage analysis (LA) takes into account only the linkage disequilibrium within families (or specific crosses), which is undone through recombination after few generations.

Wientjes et al. (2013) reported that prediction accuracy in GS depends on linkage disequilibrium from recent familiar structures, signalling for the contribution of LA for GS. Furthermore, LA can be used to identify chromosomal regions that are inherited from a common ancestor by using identity by descent (IBD) matrices. According to Bercovici et al. (2010), these regions are better candidates for detection of causal mutations affecting the phenotypes, and therefore, can be exploited under a genome-enabled prediction viewpoint.

Recently, both sources of information (LA and LDA) have been combined for GS purposes (Boichard et al. 2012; Luan et al. 2012; Wientjes et al. 2013) and showed satisfactory results, even outperforming the LDA models (traditional GS models) in relation to predictive ability in simulated and real

data. These cited studies used only Genomic best linear unbiased prediction (GBLUP)-based models. However, to the best of our knowledge, there are no reports in literature approaching the efficiency of the Bayesian GS models (Bayes A, B, Cpi, LASSO and Ridge) in the presence of QTL-genotype effects obtained from LA.

Towards this orientation, we aimed to propose and test a new class of Bayesian LALDA models (combining LA and LDA) for genomic selection implemented through free software (QXPAK e R). These models were fitted to a simulated dataset and to a real data of feed conversion ratio (FCR) in an experimental outbred pig population.

## 2. Material and methods

### 2.1. Simulated data

We used a public simulated dataset (Usai et al. 2014). The base generation (GEN0) was composed by 1020 (20 males and 1000 females) unrelated individuals. Each one of the next four non-overlapping generations (GEN1, GEN2, GEN3 and GEN4) consisted of 20 males and 1000 females from GEN0 by randomly mating each male with 50 females. The pedigree was composed by 4100 individuals (males only from GEN0 and all the individuals from GEN1 to GEN4). The simulated genome comprised 5

chromosomes, each one with a size of 99.95 Mb carrying 2000 equally distributed SNPs (single nucleotide polymorphism) (1 SNP per 0.05 Mb). Currently, this number of 10,000 SNPs can be considered as low-density for the majority of livestock species. A total of 50 SNPs were randomly sampled and treated as true QTLs. The sum of the QTLs' additive effects over each individual was defined as true breeding values (TBV). The simulated trait (Y1) was generated assuming heritability equal to 0.35. The phenotype of each individual was given by TBV plus the random residual effect sampled from a normal distribution with mean zero and variance that ensures the heritability equal to 0.35. Data of 3000 females from GEN1 to GEN3 were used in the training population, so that the rest of the population (1020 individuals from GEN4) were considered for validation. All used data are available in the following electronic address: <http://qtl-mas-2012.kassiopeagroup.com/en/dataset.php>.

## 2.2. Real data: experimental population and phenotypic data

The use of animals was approved by the Animal Care and Use Committee at Department of Animal Science (RG Number 181 444) of the Universidade Federal de Viçosa, Brazil.

The phenotypic data were obtained from an experiment carried out in Viçosa, Minas Gerais State. A three-generation resource population was created and managed as described by Hidalgo et al. (2013) and Verardo et al. (2014). Briefly, 2 naturalized Piau breed grandsires were crossed with 18 granddams from a commercial line composed of Large White, Landrace and Pietrain breeds in order to produce the F1 generation, from which 11 F1 sires and 54 F1 dams were selected. These F1 individuals were crossed to produce the F2 population, of which 341 animals were phenotyped for FCR. This trait was chosen due to its economic importance and to the difficulty to be measured, which justify the use of GS. The animals were weighted from birth and weaned at 21 days old. After weaning, they received commercial feed ad libitum. They were kept in individual pens from 77 to 105 days old, where the daily feed intake (kg/day), the daily weight gain (kg/day) and FCR (kg/kg) traits were measured. The phenotypic mean, median, standard deviation, maximum and minimum values for FCR were, respectively, 2.78, 2.71, 0.59, 1.53 and 5.25.

## 2.3. DNA extraction, genotyping and SNP quality control

Genomic DNA was extracted from the white cells of parental, F1 and F2 animals; more details can be found in Band et al. (2005). The low-density customized SNPChip with 384 markers was based on the Illumina Porcine SNP60 BeadChip (San Diego, CA, USA). These SNPs were selected according to QTL positions that were previously identified in this population by using meta-analyses (Silva et al. 2011) and fine mapping (Hidalgo et al. 2013; Verardo et al. 2014). Thus, although a small number of markers have been used, the customized SNPchip based on previously identified QTL positions ensures appropriate coverage of the relevant genome regions in this population.

From these, 66 SNPs were discarded because of a low-genotyping call rate (<0.95), and from the remaining 318 SNPs, 81 were discarded due to a minor allele frequency (MAF) < 0.05. Thus, 237 SNP markers were distributed on the *Sus scrofa* chromosomes (SSC) as follows: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (30), SSC17 (25) and SSCX (13). The average distance between markers within each chromosome was equal to 5.17, 2.37, 2.25, 3.93, 2.68 and 11.0 Mb; respectively for SSC1, SSC4, SSC7, SSC8, SSC17 and SSCX.

## 2.4. Mixed model for QTL detection through LA

In order to implement the proposed LALDA model, first we have to fit the Fernando and Grossman (1989) model (1) aiming to identify significant QTLs under a LA approach. This model was fitted separately for each chromosome.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{w} + \mathbf{e}, \quad (1)$$

where  $\mathbf{y}$  is the vector (with dimension equal to  $n$ , the number of phenotyped animals) of FCR records;  $\mathbf{X}$  is the incidence matrix of fixed effects vector ( $\boldsymbol{\beta}$ ) (sex, lot and halotane gene genotypes);  $\mathbf{Z}$  is the incidence matrix of polygenic ( $\mathbf{u}$ ) and QTL-genotype ( $\mathbf{w}$ ) effects, being  $\mathbf{w} = [w_1, w_2, \dots, w_n]'$  and  $\mathbf{e}$  represents the vector of residual terms. Under independence between  $\mathbf{u}$ ,  $\mathbf{w}$  and  $\mathbf{e}$ , the following distributions were assumed:  $\mathbf{u} \sim N(0, \sigma_u^2 \mathbf{A})$ , where  $\mathbf{A}$  is the pedigree-based relationship matrix;  $\mathbf{w} \sim N(0, \sigma_w^2 \mathbf{Q})$ , where  $\mathbf{Q}$  is the genotypic IBD (Nagamine 2005) matrix (a covariance matrix which elements are the probabilities that individuals are identical by descent conditional on pedigree and SNP marker information); and  $\mathbf{e} \sim N(0, \sigma_e^2 \mathbf{I}_n)$ .

Model (1) was fitted through QXPAK.5.05 software (Pérez-Enciso and Misztal 2011). For the simulated dataset (with 10,000 SNPs) we need to use the options MEMORY\_RAM = yes and TRANSPOSE, which are indicated for large scale genotyping data. The  $\mathbf{Q}$  matrices are stored in files called 'zran.x', where  $x$  represents the SNP positions scanned every cM. Significant QTLs are identified at each tested position by using a likelihood ratio test (LRT). In this context, model (1) is the full model, whereas this model without  $\mathbf{w}$  effect is the null model. The  $p$ -values for this test is calculated at each position  $x$ . Once the significant QTL have been identified at position  $x$ , the  $\mathbf{Q}$  matrix stored at this position will be used to compose the Bayesian LALDA models introduced in the next section.

## 2.5. Bayesian LALDA models for genomic selection

It is important to emphasize that if a significant QTL is located exactly in the physical position of a given SNP marker, such marker will be excluded from the LDA component of the LALDA model, being considered only in the LA component. It avoids that the same marker be used simultaneously in both components (LA and LDA) of LALDA model, which can cause multicollinearity problems.

In order to implement LALDA models, the traditional genomic selection Bayesian regression models (so called Bayesian alphabet) have been adapted to bear the QTL-genotype ( $\mathbf{w}$ ) effect from LA. In addition to this effect, the proposed

LALDA model (2) also contains the SNP ( $g_i$ ) and additive polygenic ( $\mathbf{u}$ ) effects. The inclusion of  $\mathbf{u}$  is due to the small number of SNP markers used here. Assuming only one significant QTL from LA, we have:

$$\mathbf{y}^* = \mathbf{1}\mu + \sum_{i=1}^m \mathbf{x}_i g_i + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{w} + \mathbf{e}, \quad (2)$$

where  $\mathbf{y}^*$  is the vector of FCR records pre-corrected for fixed effects (sex, lot and halotane gene genotypes);  $\mathbf{1}$  is a vector of ones and  $\mu$  is the overall mean;  $g_i$  is the additive effect of SNP marker  $i$ ;  $\mathbf{x}_i$  is the incidence vector of each marker (assuming the values 2, 1 or 0 that represents the SNP genotypes AA, Aa and aa, respectively) and  $m$  is the number of markers;  $\mathbf{Z}$  is the incidence matrix of polygenic ( $\mathbf{u}$ ) and QTL-genotype ( $\mathbf{w}$ ) effects, assuming, respectively,  $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{A})$  and  $\mathbf{w} \sim N(\mathbf{0}, \sigma_w^2 \mathbf{Q})$ ; and  $\mathbf{e}$  is the residual vector,  $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ . In summary, the terms  $\mathbf{Z}\mathbf{w}$  and  $\sum_{i=1}^m \mathbf{x}_i g_i$  represent, respectively, the LA and the LDA components of the proposed LALDA model.

All models treated here assume the same prior distribution for the residual variance, a scaled inverted chi-squared distribution given by  $\sigma_e^2 | v_e, S_e \sim v_e S_e \chi_{v_e}^{-2}$ . The differences between these models will be given by the prior distributions assumed for the marker effects.

The ridge regression (BRR) model is the Bayesian version of RR-BLUP proposed by Meuwissen et al. (2001), which assumes the same genetic variance for all markers; i.e.  $g_i \sim N(0, s_g^2)$ ,  $\sigma_g^2 = \sigma_{g_1}^2 = \sigma_{g_2}^2 = \dots = \sigma_{g_m}^2$ . The prior distribution for the marker genetic variance is the following scaled inverted chi-squared distribution,  $\sigma_g^2 | v_g, S_g \sim v_g S_g \chi_{v_g}^{-2}$ . The Bayes A (BA) model, differently from BRR, assumes one specific variance ( $\sigma_{g_i}^2$ ) for each marker  $i$ ,  $g_i \sim N(0, \sigma_{g_i}^2)$ , whose prior distribution is given by  $\sigma_{g_i}^2 | v_{g_i}, S_{g_i} \sim v_{g_i} S_{g_i} \chi_{v_{g_i}}^{-2}$ . Thus, the joint prior distribution for the genetic variance of all markers is given by  $\prod_{i=1}^m \sigma_{g_i}^2 | v_{g_i}, S_{g_i}$ .

Alternative approaches for BRR and BA models are Bayes B (BB), Bayes  $C\pi$  (BC $\pi$ ) and Bayesian Lasso (BL), which assume variable selection given by the shortening of the regression coefficients (marker effects). BB and BC $\pi$  extend BA and BRR, respectively, by introducing an additional parameter  $\pi$  (probability of the marker effect be equal to zero), and this parameter is treated as unknown and it is assigned a Beta prior  $\pi \sim \text{Beta}(p_0, \pi_0)$ , with  $p_0 > 0$  and  $\pi_0 \in [0, 1]$  (Pérez and de los Campos 2014). The BB model assumes as prior for the marker effects a normal mixture distribution given by  $g_i | \pi \sim (1 - \pi)N(0, \sigma_{g_i}^2) + \pi N(0, \sigma_{g_i}^2 = 0)$ . Similarly to BA,  $\sigma_{g_i}^2$  denotes that each SNP has its own variance with prior distribution given by  $\sigma_{g_i}^2 | v_{g_i}, S_{g_i} \sim v_{g_i} S_{g_i} \chi_{v_{g_i}}^{-2}$ . In the BC $\pi$  model, the prior distribution for marker effects is also given by a normal mixture distribution,  $g_i | \pi \sim (1 - \pi)N(0, \sigma_{g_i}^2) + \pi N(0, \sigma_{g_i}^2 = 0)$ , however, similarly to BRR, this model assumes the same genetic variance for all markers with prior distribution given by  $\sigma_g^2 | v_g, S_g \sim v_g S_g \chi_{v_g}^{-2}$ . Differently from BB and BC, BL provides regression coefficients estimators that solve the following optimization problem:  $\min \left\{ \sum_i (\mathbf{y}_i - \mathbf{x}_i g_i)^2 + \lambda \sum_i |g_i| \right\}$ , where

the term  $\sum_i |g_i|$  is the sum of the absolute values of the regression coefficients and  $\lambda$  is the parameter that controls the shortening these coefficients.

The BGLR (Pérez and de los Campos 2014) package of R software was used to implement all considered models (see the codes at supplementary material). Analyses were carried out using Monte Carlo Markov Chain (MCMC) algorithms saving every 5th cycle from a total of 130,000 iterations, after 30,000 of burn-in. Global convergence was checked by Geweke's Z criterion and visual inspection of trace plots.

### 2.6. Comparing Bayesian LALDA models for genomic selection

All mentioned models (BRR, BA, BB, BC e BL) from (2) were compared in terms of goodness-of-fit considering the whole dataset through Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002). As previously mentioned in the subsection 2.1, for the simulated dataset the training and validation populations were composed by 3000 and 1020 individuals, respectively. For the real dataset, the predictive ability evaluation was implemented by means of five-fold cross-validation analysis. For this, of the 5 subsamples (4 subsamples with 68 animals and 1 subsample with 69 animals), a single subsample was retained as the validation data for testing the predictive ability, and the remaining 4 subsamples were used as training data. The cross-validation process was repeated five times, with each of the subsamples used exactly once as the validation data.

The predictive ability was reported as the correlation between the pre-corrected phenotype ( $\mathbf{y}^*$ ) from validation dataset and the estimated total breeding value (EBV) denoted as  $\hat{\mathbf{y}}^*$ . From LALDA models,  $\hat{\mathbf{y}}^* = \sum_{i=1}^l \mathbf{x}_i \hat{g}_i + \mathbf{Z}\hat{\mathbf{u}} + \mathbf{Z}\hat{\mathbf{w}}$ , and from the reduced model (LDA),  $\hat{\mathbf{y}}^* = \sum_{i=1}^l \mathbf{x}_i \hat{g}_i + \mathbf{Z}\hat{\mathbf{u}}$ .

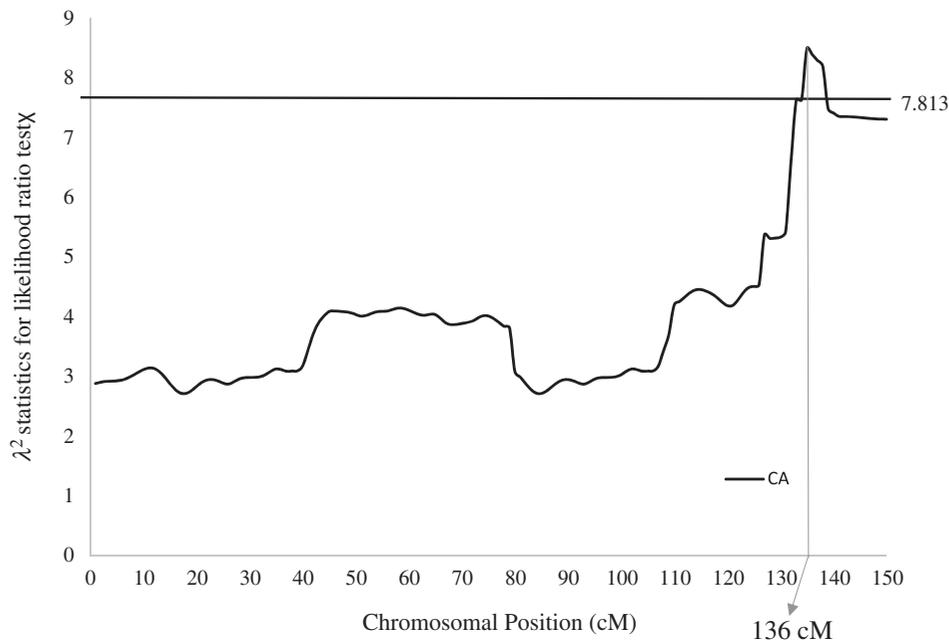
### 2.7. Heritability estimates

After identifying the best model ('Bayesian alphabet' under LALDA or LDA approaches), the variance components and heritability were estimated using the whole dataset. For LALDA model, the heritability estimate was obtained as follows:

$$\hat{h}^2 = \frac{\hat{\sigma}_a^2 + \hat{\sigma}_u^2 + \hat{\sigma}_w^2}{\hat{\sigma}_a^2 + \hat{\sigma}_u^2 + \hat{\sigma}_w^2 + \hat{\sigma}_e^2}, \quad (3)$$

where  $\hat{\sigma}_a^2$  is the estimate of the genetic variance explained by all SNP markers based on LDA,  $\hat{\sigma}_w^2$  is the estimate of the genetic variance explained by the QTL based on LA,  $\hat{\sigma}_u^2$  is the estimate of the additive polygenic variance, and  $\hat{\sigma}_e^2$  the estimate of the residual variance.

In (3), for the methods assuming the same variance for all markers (BRR and BC $\pi$ ), the variance component  $\hat{\sigma}_a^2$  was accessed as  $\hat{\sigma}_a^2 = 2\hat{\sigma}_g^2 \sum_{n=1}^m \hat{p}_i(1 - \hat{p}_i)$ ; and for those assuming different variances (BA, BB and BL) as  $\hat{\sigma}_a^2 = 2 \sum_{n=1}^m \hat{p}_i(1 - \hat{p}_i)\hat{\sigma}_{g_i}^2$ . The term  $p_i$  is the MAF of marker  $i$ .



**Figure 1.** QTL profile on chromosome 8 for feed conversion rate in an outbreed pig population.

### 3. Results

#### 3.1. QTL detection through LA and model comparisons

For the simulated data, the 50 true QTLs postulated in the simulation were detected through LA. As the QTLs are assumed as specific SNPs (see subsection 2.1), these markers were excluded from the LDA component of the LALDA model, being considered only in the LA component.

With respect to real data, [Figure 1](#) shows the profile of the LRT test across SSC8, since there were no significant QTLs in other studied chromosomes. From this figure, we found a significant QTL at position 136 cM (maximum LRT statistics).

[Table 1](#) shows the DIC values for all fitted models (BRR, BA, BB, BC and BL under LDA and LALDA viewpoints) considering both simulated and real datasets.

For the simulated dataset, the LALDA and LDA frameworks showed exactly the same goodness-of-fit measures. On the other hand, for the real data, the LALDA model outperformed the LDA (lower DIC values) for all 'Bayesian alphabet' models (BRR, BA, BB, BC and BL). Additionally, the BA, followed by the

BB and BL models presented lower DIC in both LALDA and LDA models in the two evaluated datasets. Thus, models assuming one specific variance per marker fitted better to the both datasets.

[Table 2](#) brings the results of cross-validation analysis for predictive ability. For the real dataset, we calculated the correlation between the observed and predicted values; whereas for the simulation dataset, the correlation between observed and TBV was used for this aim. Although the performance of LALDA and LDA models has been similar when considering the real dataset, the results followed the trend observed for goodness-of-fit analysis ([Table 1](#)). In summary, the LALDA approach was slightly superior to the LDA for all 'Bayesian alphabet' models (BRR, BA, BB, BC and BL). The BA model showed the best predictive ability, reinforcing the relevance to assume one specific variance per marker for the genomic analysis of FCR trait in the studied population.

With respect to the simulated data, since the true QTLs were assumed as specific SNP markers, the fact of exploiting the marker effect under a LA or LDA approach did not influence

**Table 1.** Deviance Information Criterion (DIC) values for LALDA (combining LA and linkage disequilibrium analysis) and LDA Bayesian models considering real data of feed conversion rate in an outbreed pig population and simulated data.

Dataset	Model	Linkage pattern	
		LALDA <sup>a</sup>	LDA <sup>b</sup>
Real	BRR	3288.0	3297.2
	Bayes A	3284.6	3293.8
	Bayes B	3286.8	3296.1
	Bayes $C\pi$	3291.7	3299.0
	Bayesian LASSO	3286.1	3295.7
	BRR	4828.6	4827.1
Simulated	Bayes A	4820.7	4820.9
	Bayes B	4822.3	4822.1
	Bayes $C\pi$	4829.5	4829.1
	Bayesian LASSO	4823.8	4823.1

<sup>a</sup>LALDA: Bayesian regression + additive polygenic + QTL-genotype effects.

<sup>b</sup>LDA: Bayesian regression + additive polygenic.

**Table 2.** Predictive ability (correlation between observed and predicted values) for LALDA (combining LA and LDA) and LDA Bayesian models considering feed conversion rate in an outbreed pig population.

Dataset	Model	Linkage pattern	
		LALDA <sup>a</sup>	LDA <sup>b</sup>
Real	BRR	0.28	0.26
	Bayes A	0.32	0.29
	Bayes B	0.30	0.28
	Bayes $C\pi$	0.27	0.25
	Bayesian LASSO	0.29	0.27
	BRR	0.68	0.67
Simulated	Bayes A	0.73	0.73
	Bayes B	0.70	0.71
	Bayes $C\pi$	0.67	0.66
	Bayesian LASSO	0.71	0.71

<sup>a</sup>LALDA: Bayesian regression + additive polygenic + QTL-genotype effects.

<sup>b</sup>LDA: Bayesian regression + additive polygenic.

the predictive ability. However, in terms of Bayesian models, the same pattern observed for real data was provided for simulated data (i.e. BA, followed by the BB and BL models presented higher predictive abilities).

### 3.1. Heritability estimates

For the real dataset, since the LALDA model via Bayes A outperformed the other models in terms of goodness-of-fit and predictive ability through cross-validation, this model was chosen to provide heritability estimates according to Equation (3). LALDA and LDA presented the same performance for the simulated dataset, and the Bayes A model outperformed the others, thus the Equation (3) was used to estimate the heritability in both real and simulated datasets.

Considering the real dataset, the heritability estimated for FCR was equal to 0.23, with posterior standard deviation equal to 0.038. This estimate is consistent with those reported in literature. When using the simulated dataset (assuming  $h^2=0.35$ ), it is possible to infer that all models underestimated the heritability. The RR-BLUP, BC, BLASSO, BB and BA provided estimates equal to 0.18, 0.20, 0.25, 0.27 and 0.29, respectively.

## 4. Discussion

### 4.1. QTL detection through LA and model comparisons

For the simulated dataset, although a relative 'high' number of QTLs (a total of 50) have been postulated in the simulation, the LA model was able to detect these ones. As mentioned earlier, since the true QTLs were assumed as being specific SNPs, the LA was applied for a kind of 'optimal situation' (the marker is the QTL), thus improving the chances to detect all QTLs.

In terms of the real dataset, in comparison with other usually traits in pig breeding, we note that there are relatively few studies approaching QTL detection for FCR in crossed pig populations. The reason is the difficulty of measurement and the high cost involved in feed intake phenotyping. Similar to the results presented here, Beeckmann et al. (2003) also reported significant QTL for FCR on SSC8 (between the 96.3 and 106 cM) using animals from crosses of pure Meishan, Pietrain and European wild pig breeds. Zhang et al. (2009) using Duroc White  $\times$  Erhualian Chinese crossbred populations identified significant QTLs for FCR in SSC8 at position 96 cM that explained 3.27% of the phenotypic variance. The QTL position for FCR do not exactly match the positions of the SNPs markers on the physical map. The two most closest SNP markers, ALGA0049550 and ALGA0050287, are located, respectively, at positions 132.3 and 138.2 cM. (ALGA0049550) e 138.2 (ALGA0050287) on SSC8. Thus, the QTL-genotype effect at position 136 cM was included in the Bayesian models for genomic prediction as the LA component simultaneously with all SNP markers (which composed the LDA component). The IBD matrix stored by QXPAK5.05 at the mentioned position was considered as the covariance matrix associated to QTL-genotype effect.

The DIC is relevant only to investigate the models' properties and to infer about parameter estimates. However,

the genomic selection depends on predictive ability concept, which measure the prediction efficiency of non-observed phenotypes based on known parameter estimates (in this case the SNP marker effect estimates). Thus, models chosen by goodness-of-fit measures may not be the best models to make predictions of new observations. According to Bishop (2006), the main reason for this problem is the overfitting, which can be evaluated through cross-validation analysis.

The best performance of LALDA framework for real dataset (Tables 1 and 2) can be due mainly to the small number of markers used here, since it enabled to exploit a relevant region (136 cM) not directly considered in the LDA because there is no SNP marker physically located at this position. This kind of extra information added to the model may have contributed to the improvement of the model performance. However, for the simulated dataset, this advantage of LALDA models was not observed because the true QTLs were assumed as the markers themselves. For this situation which the significant QTL positions from LA are coincident with the physical SNP map position, we adopted the LALDA considering these markers only in the LA component (QTL-genotype effect). In this context, the fitted model assumed that linkage disequilibrium occurs only within families (or specific crosses), and can be undone through recombination after few generations. Since the comparisons between the mentioned LALDA with the traditional LDA models resulted in similar results, we can infer that LA modelling was not relevant for the simulated dataset.

The best performance (goodness-of-fit and predictive ability analysis) of BA model for both datasets, independently of the LD approach (LDA or LALDA), suggests that the assumption of one variance per marker is suitable (since BA outperformed BRR and BC $\pi$ ) and that variable selection was not effective (since BA outperformed BB and BL).

Although there are few references comparing the LALDA with the LDA models, Boichard et al. (2012) studying the entire Holstein cattle population from France, reported that predictive ability for total milk production, total protein production, total fat production, milk protein percentage, milk fat percentage and fertility were 0.60 and 0.56, 0.57 and 0.55, 0.66 and 0.59, 0.73 and 0.73, 0.81 and 0.72, and 0.39 and 0.35 respectively for LALDA (denominated as QTL-BLUP) and LDA (GBLUP) models. In general, although the authors used a much larger number of animals and markers, the LALDA model proved to be satisfactory. Luan et al. (2012) studying populations of Italian dairy cattle reported the importance of identity-by-state information for the accuracy of genomic selection. Comparing the predictive ability of models by combining information from LA and LDA, the authors found values of 0.60 and 0.59, 0.60 and 0.58 and 0.63 and 0.61, respectively, for the LALDA and LDA models, considering the traits total milk production, total protein production and total fat production. Although the models presented similar performance, the authors concluded that LA component explains why the prediction equations derived for one breed may not predict accurate genome-wide breeding values when applied to other breeds, since family structures differ among breeds.

## 4.2. Heritability estimates

According to a general review by Rothschild and Ruvinsky (2010), the heritability for FCR in pigs is generally moderate (ranging from 0.12 to 0.58, with an average of 0.30) and suggest that selection would be successful. Under a genomic selection approach, Jiao et al. (2014) working with Duroc pigs reported the heritability estimated obtained through Bayes A equal to 0.32 (0.09). Lopez et al. (2016) reported heritability estimates for FCR closest to 0.30, and concluded that genomic selection for FCR is clearly superior to the conventional scheme in terms of monetary genetic gain and profit. According to Akanno et al. (2013), when using genomic selection for feed conversion in pigs assuming heritability equal to 0.32, the accuracy of selection index ranged from 0.38 to 0.66, whereas the conventional selection provides accuracies between 0.10 and 0.64.

## 5. Conclusions

Bayesian LALDA models proposed here enable to infer about the nature of LD at specific regions of interest, thus implying in new insights on whole genome prediction such as genetic architecture information and genomic breeding in crossed populations. The available theoretical support and the related computational features characterize these models as a new methodology for genomic selection.

## Acknowledgements

The authors would like to thank FAPEMIG (PPM-00563-16) and CNPq (445539/2014-3 and 309188/2014-8).

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

Akanno EC, Schenkel FS, Friendship RM, Robinson JAB. 2013. Relative economic returns from selection schemes for a nucleus swine breeding program. *Livest Res Rural Dev.* 53:13.

Band GDO, Guimarães SEF, Lopes PS, Schierholt AS, Silva KM, Pires AV, Benevenuto Júnior AA, Gomide LAM. 2005. Relationship between the Porcine Stress Syndrome gene and pork quality traits of F2 pigs resulting from divergent crosses. *Genet Mol Biol.* 28:88–91.

Beeckmann P, Moser G, Bartenschlager H, Reiner G, Geldermann H. 2003. Linkage and QTL mapping for *Sus scrofa* chromosome 8. *J Anim Breed Genet.* 120:66–73.

Bercovici S, Meek C, Wexler Y, Geiger D. 2010. Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics* 26:i175–i182.

Bishop CM. 2006. *Pattern recognition and machine learning.* New York (NY): Springer.

Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN. 2012. Genomic selection in French dairy cattle. *Anim Prod Sci.* 52:115–120.

Fernando RL, Grossman M. 1989. Marker assisted selection using best linear unbiased prediction. *Genet Sel Evol.* 21:467–477.

Hidalgo AM, Lopes OS, Paixão DM, Silva FF, Bastiaansen JW, Paiva SR, Faria DA, Guimarães SEF. 2013. Fine mapping and single nucleotide polymorphism effects estimation on pig chromosomes 1, 4, 7, 8, 17 and X. *Genet Mol Res.* 36:511–519.

Jiao S, Maltecca C, Gray K, Cassady P. 2014. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: II. Genomewide association. *J Anim Sci.* 92:2846–2860.

Lopez BM, Kang HS, Kim TH, Viterbo VS, Kim HS, Na CS, Seo KS. 2016. Optimization of Swine Breeding Programs using genomic selection with ZPLAN+. *Asian Australas J Anim Sci.* 29:640–645.

Luan T, Woolliams JA, Ødegård J, Dolezal M, Roman-Ponce SI, Bagnato A, Meuwissen TH. 2012. The importance of identity-by-state information for the accuracy of genomic selection. *Genet Sel Evol.* 44:28–15.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* 157:1819–1829.

Nagamine Y. 2005. Transformation of QTL genotypic effects to allelic effects. *Genet Sel Evol.* 37:579–584.

Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics.* 198:483–495.

Pérez-Enciso M, Misztal I. 2011. Qxpak5: Old mixed model solutions for new genomics problems. *BMC Bioinformatics.* 12:202–206.

Rothschild MF, Ruvinsky A. 2010. *The genetics of pigs.* Cambridge: CABI.

Silva KM, Knol EF, Merks JWM, Guimarães SEF, Bastiaansen JWM, van Arendonk JAM, Lopes PS. 2011. Meta-analysis of results from quantitative trait loci mapping studies on pig chromosome 4. *Anim Genet.* 42:280–292.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit. *J R Stat Soc.* 64:583–639.

Usai MG, Gaspa G, Macciotta NP, Carta A, Casu S. 2014. XVI<sup>th</sup> QTLMAS: simulated data set and comparative analysis of submitted results for QTL mapping and genomic evaluation. *BMC Proceedings.* 8:5.

Verardo L, Silva FF, Varona L, Resende MDV, Bastiaansen JW, Lopes OS, Guimarães SEF. 2015. Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs. *J Appl Genet.* 56:123–132.

Wientjes YCJ, Veerkamp RF, Calus MPL. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics.* 193:621–631.

Zhang Z, Ren J, Ren D, Ma J, Guo Y, Huang L. 2009. Mapping quantitative trait loci for feed consumption and feeding behaviors in a White Duroc x Chinese Erhualian resource population. *J Anim Sci.* 87:3458–3463.