



ORIGINAL ARTICLE

Supervised independent component analysis as an alternative method for genomic selection in pigs

C.F. Azevedo¹, F.F. Silva^{1,2}, M.D.V. de Resende^{1,3}, M.S. Lopes⁴, N. Duijvesteijn⁴, S.E.F. Guimarães², P.S. Lopes², M.J. Kelly⁵, J.M.S. Viana⁶ & E.F. Knol⁴

1 Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Brazil

2 Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Brazil

3 Embrapa Florestas, Departamento de Engenharia Florestal, Universidade Federal de Viçosa, Viçosa, Brazil

4 TOPIGS Research Center IPG, Beuningen, the Netherlands

5 Queensland Alliance for Agriculture & Food Innovation, The University of Queensland, St Lucia, QLD, Australia

6 Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, Brazil

Keywords

Accuracy; animal breeding; genomic selection; SNP.

Correspondence

F.F. Silva, Departamento de Zootecnia,
Universidade Federal de Viçosa, 36570-000,
Viçosa, Minas Gerais, Brazil.
Tel: +55 31 3899-3321;
Fax: +55 31 3899-2275;
E-mail: fabyanofonseca@ufv.br

Received: 29 January 2014;

accepted: 5 June 2014

Summary

The objective of this work was to evaluate the efficiency of the supervised independent component regression (SICR) method for the estimation of genomic values and the SNP marker effects for boar taint and carcass traits in pigs. The methods were evaluated via the agreement between the predicted genetic values and the corrected phenotypes observed by cross-validation. These values were also compared with other methods generally used for the same purposes, such as RR-BLUP, SPCR, SPLS, ICR, PCR and PLS. The SICR method was found to have the most accurate prediction values.

Introduction

In recent years, a large amount of genomic information has become available regarding animal production. As breeding animals are genotyped with thousands of single nucleotide polymorphism (SNP) markers, their individual genetic merit can be estimated in the context of genomewide selection (GWS). However, the practical application of this genomic information is challenging. The appropriate use of functional models that estimate the effect of each SNP in the phenotype is typically not possible because the number of markers is generally much higher than the number of genotyped and phenotyped animals. Therefore, a key point for the success of GWS is the appropriate choice of methodologies; the development of computational tools and comparison of these methodologies is an important line of research in the current framework of animal breeding.

To date, the random regression best linear unbiased predictor (RR-BLUP and G-BLUP) and Bayesian methods (Bayes A and B and LASSO Bayesian) have been most widely used methods for GWS. However, other methods have been successfully applied to GWS, including the partial least squares (PLS) and principal component regression (PCR) methods (Long *et al.* 2011). These two methods can be categorized as dimensionality reduction methods, which are widely applicable and relatively simple in theory compared with the RR-BLUP and Bayesian methods.

According to Solberg *et al.* (2009), the PCR and PLS methods allow large amounts of data to be rapidly analysed to obtain estimates of the genomic breeding values (GEBVs) of individuals, assuming only additive marker effects. Boulesteix & Strimmer (2006) have shown that some positive aspects of PLS are its high flexibility, versatility, statistical efficiency and computational speed. Another dimensionality reduction

method is the independent component regression (ICR) method, which was initially applied to GWS by Azevedo *et al.* (2013). These authors demonstrated that the ICR method more accurately and efficiently predicted phenotypic values compared with RR-BLUP, PLS and PCR.

However, traditional methods for dimension reduction do not consider covariate selection. Thus, new strategies, such as the sparse partial least squares (SPLS) (Colombani *et al.* 2012) and supervised principal component regression (SPCR) methods, have been applied (Long *et al.* 2011). The combination of dimensionality reduction and covariate selection can be effective and accurate for the prediction of phenotypic values. Moreover, the proposed method, the so-called supervised independent components regression (SICR) method, also fits in this context. However, it has not yet been applied to genomic selection.

Boar taint and carcass traits have been used in genomewide association studies (GWASs) and can be considered specialized phenotypes. Boar taint is the undesirable smell and taste of pork derived from some uncastrated male pigs, and its main causes are related to the compounds androstenone and skatole (Gregeresen *et al.* 2012). Duijvesteijn *et al.* (2010) and Ramos *et al.* (2011) have performed association studies aiming to identify SNPs associated with androstenone and skatole levels in pig carcasses. Luo *et al.* (2012) conducted a GWAS for meat quality traits, and the results effectively narrowed the associated regions compared with previous QTL studies and candidate genes. Although several GWASs have been conducted regarding boar taint and carcass traits, GWS studies are still scarce. Thus, it is necessary to compare GWS methodologies for these phenotypes, as dimensionality reduction methods are viable alternatives. In the light of this research need, this study aimed to conduct a comparative analysis between PCR, ICR, SPLS, SPCR, SICR and RR-BLUP in terms of their efficiency in the estimation of GBV and the effects of SNPs on boar taint and carcass traits in pigs.

Materials and methods

Phenotypic data

In this study, 622 boars from different farms in the Netherlands were phenotyped for the following traits: concentration of androstenone and skatole, backfat thickness (HGP backfat) and loin depth (HGP loin). The field experiment was conducted according to the Dutch law for the protection of animals.

A Hennessy Grading Probe (HGP) was used to measure backfat thickness and loin depth. The generated profiles were scanned to identify tissue interfaces, from which phenotypic measurements were produced according to the site (<http://www.hennessy-technology.com>). Samples were taken from the fat of the neck on the left side of the animal carcass. Samples were stored under vacuum at -20°C until the date that concentrations of androstenone and skatole were measured. Additional information about the collection and phenotype processing is found in Duijvesteijn *et al.* (2010).

Next, all phenotypes were precorrected for the fixed effects of hot carcass weight (as a linear covariate) and contemporary groups (month and year of slaughter). Furthermore, the concentrations of androstenone and skatole were also precorrected for the covariate age at slaughtering. This precorrection is not required in the application of PCR, ICR and RR-BLUP because the multiple regression models assume fixed effects in addition to components (PCR and ICR), and the original marker (RR-BLUP) regressors can be easily implemented from theoretical and computational perspectives. In contrast, when the components from the PLS depend directly on the dependent variable considered in the analysis, fixed effects can affect the generation of components and thus the prediction ability of this method if not removed. Thus, this precorrection was adopted to preserve the comparability between the methods and their direct interpretation. Moreover, the cross-validation analysis postulates that the GEBV predictions must be correlated with adjusted phenotype values. Therefore, there is no actual correlation between observed and predicted genetic values if these values are influenced by fixed effects.

Genotypic data

A panel of 2500 SNPs previously identified as an optimal set of markers for commercial pig lines (Lopes *et al.* 2013) was used. The SNPs were distributed throughout the chromosomes, with an average of 131 SNPs per chromosome and an average distance between the SNPs of 1038 kb.

Model

This study used the following model, which was proposed by Meuwissen *et al.* (2001):

$$\mathbf{y} = \mathbf{1}'\mu + \mathbf{X}\mathbf{m} + \mathbf{e},$$

where \mathbf{y} is the column vector of phenotypes, $\mathbf{1}'$ is an row vector of phenotypes, μ is the general mean, \mathbf{m} is

the column vector of markers with incidence matrix \mathbf{X} (values of -1 , 0 and 1 for the number of an allele in the SNP), \mathbf{e} is the vector residual with

$$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$$

and σ_e^2 is the error variance.

The methods used in the analysis are detailed in Appendix S2 and described briefly below.

Random regression best linear unbiased predictor

The RR-BLUP method uses BLUP-type predictors and assumes that the SNP marker effects are covariates of the random effects. Prediction by RR-BLUP is based on the following mixed model equation:

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} + \mathbf{I} \frac{\sigma_g^2}{(\sigma_g^2/n_Q)} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{m}}_{\text{rr-blup}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix} \quad 1$$

where \mathbf{b} is the fixed-effects vector, $\mathbf{m}_{\text{rr-blup}}$ is the vector of markers' random effects with incidence matrix \mathbf{X} , σ_g^2 is the genetic variance, $n_Q = \sum_{j=1}^j 2p_j(1-p_j)$, and p_j is the allelic frequency of marker j .

Aside from enabling the regularization in the estimation process, all dimensionality reduction methods guarantee the removal of multicollinearity present in the data once the correlation between any pair of components (linear combinations of SNPs) is equal to zero. In the dimensionality reduction methods, matrix \mathbf{X} is defined as the matrix of SNP markers, and \mathbf{y} is defined as the vector of phenotypes corrected for fixed effects.

Principal components regression

PCR reduces the dimensionality without resulting in a significant loss of information present in the data (Otto 1999). In this method, the components Z_v , $v = 1, \dots, n_{\text{PCR}}$ are linear combinations of the explanatory variables X_1, \dots, X_j . Thus, the following equation holds:

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{P}} \quad 2$$

where $\hat{\mathbf{P}}$ is the matrix of the n_{PCR} first eigenvectors of the covariance matrix of \mathbf{X} and \mathbf{Z} is the component matrix. Multiple linear regression is used to establish the relation between \mathbf{y} and Z_v , obtaining the following prediction equation:

$$\hat{\mathbf{y}} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\mathbf{Z}}_1 + \hat{\alpha}_2 \hat{\mathbf{Z}}_2 + \dots + \hat{\alpha}_{n_{\text{PCR}}} \hat{\mathbf{Z}}_{n_{\text{PCR}}}, \quad 3$$

where $\hat{\alpha}_v$'s are the estimated regression coefficients, which have no biological interpretation. However, the coefficients associated with the original variables

(SNPs) can be estimated by combining (3) and (2) as follows:

$$\hat{\mathbf{m}}_{\text{PCR}} = \hat{\mathbf{P}}\hat{\alpha}. \quad 4$$

Supervised principal components regression

In situations with a large number of markers, Colombani *et al.* (2012) confirmed that SPLS and SPCR enabled relevant variables to be identified more easily than PLS and PCR, respectively. Supervised principal components regression (SPCR) is a dimensionality reduction and covariate selection method, which was first applied to GWS by Long *et al.* (2011). This method consists of two steps. The first step is based on a full PCR model involving all SNPs. After obtaining the full model coefficients (SNP effect), the magnitudes of these coefficients are used to rank all SNPs. Next, the specified number of top-ranked SNPs is selected. In the second step, the PCR method is applied again using only the selected SNPs.

Partial least squares

PLS is considered an appropriate method for data containing more covariates than observations (Hoskuldsson 1998), as in GWS. This methodology consists of simultaneously decomposing the \mathbf{X} and \mathbf{y} as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{L}' + \mathbf{E}_1 \quad 5$$

$$\mathbf{y} = \mathbf{U}\mathbf{q}' + \mathbf{e}_2 \quad 6$$

where \mathbf{T} and \mathbf{U} are matrices of components, \mathbf{L} and \mathbf{q} are the matrix and vector of the loads, respectively, and \mathbf{E}_1 and \mathbf{e}_2 are the matrix and vector of the residuals, respectively. The decomposition of \mathbf{X} and \mathbf{y} is not independent but is carried out simultaneously, which allows for the establishment of a relation between the X and Y components such that the following relation is obtained for each factor:

$$\hat{u}_l = \hat{b}_l t_l \quad 7$$

Where u_l and t_l are the vectors of the components and

$$\hat{b}_l = (u_l' t_l) / (t_l' t_l)$$

is the regression coefficient between factors, which are grouped in a diagonal matrix \mathbf{B} . Thus, the following prediction equation is obtained:

$$\hat{\mathbf{y}} = \hat{\mathbf{T}}\hat{\mathbf{B}}\hat{\mathbf{q}}'. \quad 8$$

Similar to PCR, the coefficients $\hat{\mathbf{B}}\hat{\mathbf{q}}'$ have no biological interpretation. However, the original coefficients can be obtained by combining Model (6) and

Equation (8) to obtain the following relation:

$$\hat{\mathbf{m}}_{\text{pls}} = \hat{\mathbf{L}} \hat{\mathbf{B}} \hat{\mathbf{q}}'. \quad 9$$

Sparse partial least squares

SPLS aims to provide the sparsity of the original variables \mathbf{X} . Thus, the product given by $\mathbf{M} = \mathbf{X}'\mathbf{Y}$ is decomposed into singular values and vectors. The singular value decomposition of \mathbf{M} can be obtained as follows:

$$\mathbf{M} = \Gamma \Delta \Theta' \quad 10$$

where Γ and Θ are orthonormal vectors and Δ is a diagonal matrix with singular values $\delta_k (k = 1, \dots, n_{\text{spls}})$. The \mathbf{M}_k component is obtained by an approximation using the previous step $k-1$, as explained in Lê Cao *et al.* (2008):

$$\mathbf{M}_k = \mathbf{M}_{k-1} - \delta_k \mathbf{L} \mathbf{q}', \quad 11$$

where \mathbf{L} and \mathbf{q} correspond to vectors γ_{k-1} and θ_{k-1} , respectively. The penalties in loading vectors that are determined by optimization become

$$\min_{L,q} M - L q_F^2 + g_{\gamma_1}(L) + g_{\gamma_2}(q), \quad 12$$

where $g_{\gamma_i}(x) = \text{sign}(x)(|x| - g_{\gamma_i})$ is the penalty function. After obtaining the prediction equation of SPLS, it is possible to obtain the original coefficients of markers, some of which will be zero due to the covariate selection made in the method.

Independent component regression

ICR is the decomposition of matrix \mathbf{X} into linear combinations of completely independent components in terms of both linear and nonlinear relations. One advantage of ICR compared with PCR is the possibility of completely removing any relationship of dependence between covariates. For this purpose, each independent component is built using the most representative SNPs chosen from a group of correlated SNPs.

Such an analysis is also suitable to any distribution of the indicator variable in matrix \mathbf{X} given that it can be a non-Gaussian distribution. Thus, ICR is suitable for GWS because the matrix of markers \mathbf{X} is parameterized with values the -1 , 0 and 1 (non-Gaussian distribution). Accordingly, the decomposition is as follows:

$$\mathbf{X}' = \mathbf{A}' \mathbf{S}', \quad 13$$

where \mathbf{S} is the matrix of independent components and \mathbf{A} is the matrix of mixtures.

Special algorithms attempt to find an orthogonal matrix \mathbf{R} that maximizes the statistical independence

of the columns of matrix \mathbf{S} using a quantitative measure of independence, which is a function of contrasts. The iterative algorithm developed by Hyvärinen (1998) is based on the maximum entropy $J(r)$ concept, assuming that the variable r is standardized. According to this algorithm, the following approximation is obtained:

$$J(r) \propto [E\{G_i(r)\} - E\{G_i(v)\}]^2, \quad 14$$

where v is a standardized variable and

$$G_1(v) = -\exp(-v^2/2).$$

After the iterative process, the following component matrix is obtained:

$$\hat{\mathbf{S}}' = \mathbf{X} \mathbf{K} \mathbf{R}, \quad 15$$

where \mathbf{K} is an orthogonalization matrix and $\mathbf{K} \mathbf{R}$ is an approximation of \mathbf{A}' . Thus, the equation of prediction is obtained based on ICR as follows:

$$\hat{\mathbf{y}} = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{\mathbf{S}}_1 + \hat{\gamma}_2 \hat{\mathbf{S}}_2 + \dots + \hat{\gamma}_{n_{\text{icr}}} \hat{\mathbf{S}}_{n_{\text{icr}}}, \quad 16$$

where the γ_k coefficients are determined by the OLS method, $k = 1, \dots, n_{\text{icr}}$. Similar to other dimensionality reduction methods, the marker effects can be obtained by combining Equations (13) and (16), resulting in the following estimates:

$$\mathbf{m}_{\text{icr}} = \mathbf{K} \mathbf{R} \hat{\mathbf{y}}. \quad 17$$

Supervised independent components regression

Applying the method of Long *et al.* (2011) in the context of ICR yields SICR, which is applied to the area of genomics for the first time in this study. SICR also consists of two steps. The first step is based on a full ICR model involving all SNPs. After obtaining the full model coefficients (SNP effects), the magnitudes of these coefficients are used to rank all SNPs and the specified number of top-ranked SNPs are selected. Then, in the second step, the ICR method is applied using only the selected SNPs.

Determination of the number of components and comparison of methods

An important step of dimensional reduction methods is the choice of the optimum number of latent variables to be inserted into the model. There is no formal rule to determine this number; thus, we used the minimum number of components that stabilizes the predictive ability of the method.

The dimensional reduction methods and RR-BLUP were compared using the results from the validation

based on Jackknife (Resende *et al.* 2012). For each method and each phenotype, the original data set with 622 animals was divided into 622 training data sets (D_{-i}) of 621 individuals, D_{-1} , D_{-2} , ..., D_{-622} , each of which contained the marker and phenotype information of all animals except animal i . In these analyses, the predicted genomic breeding value of animal i for each trait was calculated by $\hat{u}_i^* = \mathbf{X}_i \hat{\mathbf{m}} - i$, where \mathbf{X}_i denotes the SNP genotype vector of animal i and $\hat{\mathbf{m}} - i$ denotes the estimated marker effect vector from the analysis that considered all animals except animal i . The process was performed separately for each trait. Thus, the predictive ability was obtained by correlation between the GEBV and corrected phenotypes. The estimation bias was obtained by the regression coefficient between the two aforementioned elements. The relative efficiency of each dimensional reduction method compared with RR-BLUP was calculated using the ratio between the predictive ability of the method and that of RR-BLUP.

Once the best method for each trait was determined, the heritability of each trait was obtained by the expression $h^2 = \sum_{j=1}^J 2p_j(1 - p_j)m_j^2/V_f$, where V_f represents the phenotypic variance. Moreover, the absolute values of the marker effects were estimated and standardized using Equations (1), (4), (9) and (14) for RR-BLUP, PLS and SPLS, PCR and SPCR, and ICR and SICR, respectively. This information was used to construct a Manhattan plot, where each point represents a SNP marker, the X axis indicates location in the chromosome, and the Y axis indicates the magnitude of the effect.

All computational routines were implemented using the R software (R Development Core Team 2011) using the packages rrBLUP (RR-BLUP), pls (PLS, PCR, SPCR), spls (SPLS) and caret (ICR, SICR) and the functions mixed.solve (RR-BLUP), plsr (PLS, PCR, SPCR), spls (SPLS) and icr (ICR, SICR), which are shown in the Appendix S1.

Results and discussion

An average of 63 components was needed to stabilize the predictive ability, providing a reduction of 97.48% in the original data, as 63 components correspond to 2.52% of the total number of original variables (2500 SNPs). The covariate selection was made only by the SPCR, SICR and SPLS methods. The first two methods selected only 20% of the markers (500 SNPs), which stabilized the predictive ability. In contrast, the SPLS method did not select any covariate, and only a few of the coefficients were close to zero.

The predictive abilities of the dimensional reduction methods and RR-BLUP for each trait are presented in Table 1. Considering these results, SICR was the most efficient for all traits, with a predictive ability of 0.68, 0.68, 0.61 and 0.47 for Andro, SKA, HFA and HLO, respectively. RR-BLUP, PLS, PCR, ICR and SPLS yielded predictive abilities in the range of 0.01 and 0.41, which are considerably lower than that of SICR. Although SPCR outperformed the other methods, it was worse than SCIR, with a predictive ability of 0.62, 0.58, 0.57 and 0.44 for Andro, SKA, HFA and HLO, respectively. PCR and ICR yielded similar results but were poorer than their supervised versions, SICR and SPCR, respectively. In addition to its efficiency in prediction, the other advantage of SICR compared with the other dimensional reduction methods is the fact that it considers the complete independence between components, guaranteeing the absence of both linear and nonlinear relations between latent variables. SICR also provides for the selection of covariates.

In contrast, PLS displayed the lowest predictive ability of values for all traits (0.34, 0.14, 0.18 and 0.01 for Andro, SKA, HFA and HLO, respectively), possibly because this method does not consider the dependence between the explicative variables, which are SNPs in this case.

The use of dimensional reduction methods has not yet been reported for genomic selection in pigs. However, it has been already applied to other species, and such results can be used as a reference. Moser *et al.* (2009) performed a study comparing five methods for dairy cattle data, including PLS and RR-BLUP, which displayed similar predictive abilities. This finding differs from the results obtained in the present study. In contrast, Solberg *et al.* (2009) performed a study comparing PLS and PCR and observed similar predictive ability values (0.47 and 0.45, respectively), which disagree with the values obtained in our study, where PCR outperformed PLS considerably. However, the predictive ability values found by Macciotta *et al.* (2010) for PCR were similar (0.28–0.46) to those obtained in this study. For the covariate selection methods, SPLS and SPCR, Long *et al.* (2011) found predictive ability values similar to those found in this study.

The relative efficiencies of the dimensionality reduction methods compared with RR-BLUP are presented in Table 1. This measure is calculated as the ratio between the predictive ability of the dimensionality reduction methods and RR-BLUP. For the traits Andro, Ska and Hfa, the SPCR and SICR methods displayed efficiencies varying from 1.51 to 2.26, which are 51–126% greater than the efficiencies of

Table 1 Predictive ability (correlation coefficients and its confidence intervals) of the methods, relative efficiency^a of the dimensionality reduction methods over the RR-BLUP, regression coefficients (bias) between phenotypic values and GEBVs and heritability estimates, respectively, for each trait

	Methods	ANDRO	SKA	HFA	HLO
Predictive ability	RR-BLUP	0.41 _[0.34;0.47]	0.29 _[0.20;0.35]	0.27 _[0.20;0.35]	0.01 _[-0.08;0.08]
	PLS	0.34 _[0.27;0.41]	0.15 _[0.07;0.23]	0.18 _[0.11;0.26]	0.01 _[-0.08;0.08]
	PCR	0.30 _[0.23;0.38]	0.28 _[0.21;0.35]	0.24 _[0.16;0.31]	0.08 _[0.00;0.16]
	ICR	0.31 _[0.24;0.38]	0.29 _[0.22;0.36]	0.24 _[0.16;0.31]	0.07 _[-0.01;0.15]
	SPLS	0.34 _[0.27;0.41]	0.15 _[0.08;0.23]	0.18 _[0.10;0.25]	0.01 _[-0.08;0.08]
	SPCR	0.62 _[0.58;0.67]	0.58 _[0.53;0.63]	0.57 _[0.52;0.64]	0.44 _[0.38;0.50]
Relative efficiency	SICR	0.68 _[0.63;0.72]	0.68 _[0.61;0.75]	0.61 _[0.57;0.65]	0.47 _[0.41;0.53]
	PLS	0.83	0.52	0.67	1.00
	PCR	0.73	0.96	0.89	8.00
	ICR	0.76	1.00	0.89	7.00
	SPLS	0.83	0.52	0.67	1.00
	SPCR	1.51	2.00	2.11	44.00
Regression coefficients (bias)	SICR	1.66	2.34	2.26	47.00
	RR-BLUP	0.98	0.98	0.96	0.02
	PLS	0.38	0.15	0.20	0.01
	PCR	0.81	0.68	0.68	0.21
	ICR	0.80	1.00	1.00	1.00
	SPLS	0.56	0.35	0.34	0.02
Heritability	SPCR	1.02	1.96	1.01	1.90
	SICR	1.02	1.00	1.00	1.00
	RR-BLUP	0.40	0.15	0.16	0.01
	PLS	0.40	0.37	0.37	0.04
	PCR	0.18	0.14	0.13	0.12
	ICR	0.17	0.14	0.13	0.12
	SPLS	0.40	0.37	0.37	0.04
	SPCR	0.38	0.28	0.33	0.21
	SICR	0.44	0.38	0.36	0.24

^aThe relative efficiency is calculated by the ratio between the predictive ability of the dimensionality reduction methods and that of the RR-BLUP. ANDRO, concentration of androstenone; SKA, concentration of skatole; HFA, backfat thickness (HGP backfat); HLO, backfat loin depth (HGP loin).

RR-BLUP, respectively. The PCR and ICR methods were inferior to RR-BLUP. Conversely, PLS and SPLS exhibited poorer results compared with the other methods.

The regression coefficients between observed and predicted phenotypes are presented in Table 1. The only method in which all regression coefficient estimates were close to unity was SICR, indicating that the genetic evaluations are not biased and are effective in predicting the actual magnitudes of differences between individuals in evaluation (Resende *et al.* 2010). This result highlights the superiority of SICR compared with the other methods. The fact that the coefficients of other methods were greater or less than unity indicates that the GEBVs were under- and over-predicted, respectively.

The SICR method is superior to SPCR because ICR uses single value decomposition (SVD) of the marker incidence matrix, whereas PCR uses spectral decomposition (SD). Independence between components is achieved when using the estimated whitening matrix

from SVD, whereas only the orthogonality between them is achieved when using SD. Moreover, the success of the ICR approach comes from adequately handling the non-Gaussian distributions of covariates (Bishop 2006). The ICR method outperforms the other dimensionality reduction methods because the marker genotypes follow a binomial distribution. In contrast, the independence between components in PCR is only achieved under the assumption that the covariates are normally distributed, which is not verified for SNP genotypes.

Table 1 presents the estimates of trait heritability obtained by all methods. According Sellier *et al.* (2000), the heritability of androstenone concentration ranges from 0.25 to 0.88, which is in agreement with the results found in this study (values of 0.38–0.44), except for the PCR and ICR methods (0.18 and 0.17, respectively). For skatole concentration, the heritability found in this study of the supervised methods and PLS (values of 0.28–0.38) is in agreement with the results of Tajet *et al.* (2006), who reported heritability

for the same trait ranging from 0.19 to 0.55. Regarding backfat thickness, studies using the traditional restricted maximum likelihood (REML) method observed a heritability of 0.45 (van Wijk *et al.* 2005). However, this study found lower heritability values for this trait (0.13–0.37). For loin depth, the heritability values obtained by PCR and ICR were close to the value of 0.13 obtained by van Wijk *et al.* (2005).

The selection of markers increased the predictive ability and heritability estimate of the dimensionality reduction methods, except for the PLS and SPLS methods. This behaviour can also be observed in the RR-BLUP_B method presented by Resende *et al.* (2012). The robustness of the PLS method is due to the fact that its development intrinsically includes the selection of SNPs in the construction of the components.

The estimates of the correlations are associated with sampling errors and are thus associated with a confidence interval. The expected correlation between the GEBV estimated by the SPCR and RR-BLUP methods for HLO is less than that found in this paper (0.59) due to the values obtained for the predictive ability of the methods. However, this estimate may be associated with a large confidence interval. When the accuracy tends to 0.0, other accuracies should not be

expressed in relation to that one in terms of the number of the times the efficiency is greater, as the relative efficiency would go to infinity in this case. Instead, one should infer that the accuracy is null and the others are moderate to low. In addition, ICR allows for a large increase in reliability when h^2 is near zero (as in HLO).

The correlation estimates of GEBVs obtained by the various methods (Table 2) indicate that the highest correlation occurs between ICR and PCR. Moreover, the PLS method yields the lowest correlations between the other methods. The correlations between other methods were high, ranging from 0.63 to 0.99.

The percentages of selected individuals (top 10%) that are coincident between SICR and RR-BLUP were calculated for each trait and are presented below. These percentages of agreement for androstenone concentration, backfat thickness, loin depth and skatole were 54, 48, 35 and 52%, respectively.

The genetic correlations across the GEBVs of the traits considering SICR, the most efficient method for the prediction of phenotypic values, are presented in Table 3. In this study, the genetic correlation between backfat thickness and loin depth is low, in contrast to the values of -0.40 and -0.60 reported by Tomiyama

Table 2 Estimates of correlation coefficients and its confidence intervals of Genomic Breeding Values (GEBVs) between pairs of methods for each trait

Traits	Methods	PLS	PCR	ICR	SPLS	SPCR	SICR	RR-BLUP
ANDRO	PLS	1	0.44 _[0.38;0.50]	0.45 _[0.39;0.51]	0.98 _[0.97;0.98]	0.64 _[0.59;0.68]	0.66 _[0.62;0.71]	0.79 _[0.75;0.81]
	PCR		1	0.98 _[0.97;0.98]	0.45 _[0.38;0.51]	0.70 _[0.65;0.73]	0.63 _[0.58;0.69]	0.80 _[0.77;0.83]
	ICR			1	0.45 _[0.39;0.51]	0.69 _[0.64;0.73]	0.64 _[0.59;0.68]	0.81 _[0.78;0.83]
	SPLS				1	0.65 _[0.60;0.69]	0.67 _[0.62;0.71]	0.79 _[0.76;0.82]
	SPCR					1	0.89 _[0.87;0.90]	0.86 _[0.83;0.88]
	SICR						1	0.84 _[0.82;0.86]
HFA	PLS	1	0.39 _[0.33;0.46]	0.40 _[0.33;0.46]	0.99 _[0.98;0.99]	0.56 _[0.52;0.60]	0.59 _[0.56;0.62]	0.67 _[0.62;0.71]
	PCR		1	0.99 _[0.98;0.99]	0.40 _[0.34;0.47]	0.71 _[0.68;0.74]	0.63 _[0.60;0.66]	0.86 _[0.84;0.88]
	ICR			1	0.40 _[0.34;0.47]	0.71 _[0.69;0.73]	0.63 _[0.61;0.65]	0.86 _[0.84;0.88]
	SPLS				1	0.58 _[0.52;0.65]	0.59 _[0.53;0.65]	0.68 _[0.64;0.72]
	SPCR					1	0.90 _[0.88;0.92]	0.82 _[0.79;0.85]
	SICR						1	0.78 _[0.74;0.82]
HLO	PLS	1	0.40 _[0.33;0.46]	0.40 _[0.33;0.46]	0.97 _[0.97;0.98]	0.46 _[0.39;0.52]	0.47 _[0.40;0.53]	0.47 _[0.41;0.53]
	PCR		1	0.98 _[0.97;0.98]	0.40 _[0.33;0.46]	0.69 _[0.65;0.73]	0.63 _[0.58;0.68]	0.80 _[0.77;0.83]
	ICR			1	0.39 _[0.32;0.46]	0.68 _[0.63;0.72]	0.63 _[0.59;0.68]	0.80 _[0.77;0.82]
	SPLS				1	0.45 _[0.38;0.51]	0.46 _[0.40;0.52]	0.46 _[0.40;0.52]
	SPCR					1	0.83 _[0.80;0.85]	0.63 _[0.58;0.68]
	SICR						1	0.59 _[0.52;0.66]
SKA	PLS	1	0.42 _[0.35;0.48]	0.42 _[0.36;0.49]	0.99 _[0.98;0.99]	0.59 _[0.56;0.62]	0.58 _[0.53;0.63]	0.66 _[0.61;0.70]
	PCR		1	0.97 _[0.96;0.97]	0.44 _[0.38;0.50]	0.67 _[0.64;0.70]	0.64 _[0.59;0.68]	0.88 _[0.87;0.90]
	ICR			1	0.44 _[0.38;0.50]	0.67 _[0.63;0.71]	0.65 _[0.61;0.70]	0.89 _[0.87;0.90]
	SPLS				1	0.59 _[0.55;0.63]	0.59 _[0.54;0.64]	0.67 _[0.63;0.72]
	SPCR					1	0.91 _[0.90;0.92]	0.83 _[0.80;0.86]
	SICR						1	0.76 _[0.73;0.80]

ANDRO, concentration of androstenone; SKA, concentration of skatole; HFA, backfat thickness (HGP backfat); HLO, backfat loin depth (HGP loin).

Table 3 Heritability and genetic correlation estimates for traits obtained by the SICR method

Traits	Andro	HFA	HLO	SKA
ANDRO	0.44	−0.04	−0.05	0.34
HFA		0.36	0.07	0.02
HLO			0.24	−0.08
SKA				0.38

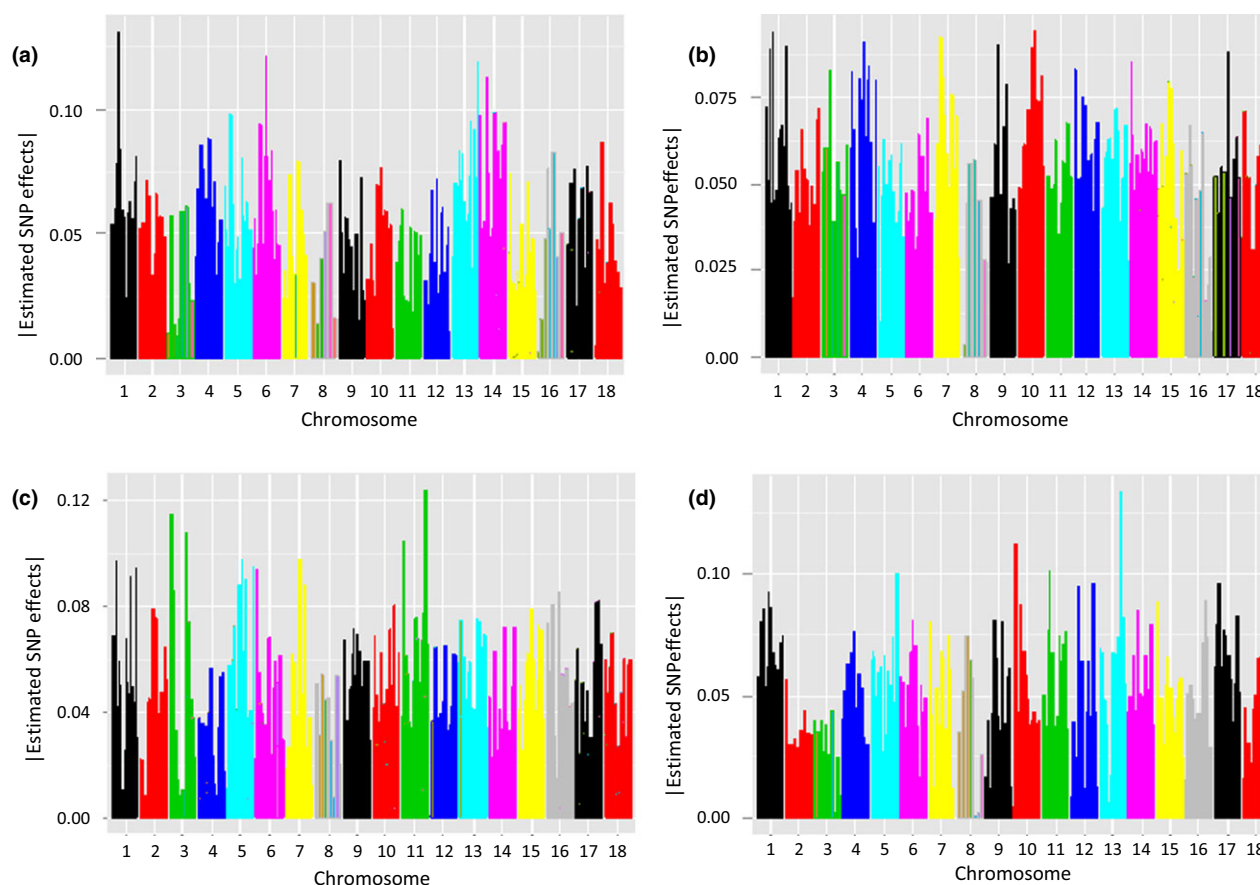
Heritability estimates are presented in diagonal; genetic correlation estimates are shown above the diagonal. ANDRO, concentration of androstenone; SKA, concentration of skatole; HFA, backfat thickness (HGP backfat); HLO, backfat loin depth (HGP loin).

et al. (2009) and van Wijk *et al.* (2005), respectively. Androstenone concentration and skatole displayed a genetic correlation of 0.34, which is similar to the results reported by Tajet *et al.* (2006) and Windig *et al.* (2012), who obtained values of 0.36 and 0.37, both in Landrace.

Identifying markers of large effects is critical for GWS because it allows for the determination of the

position of these SNPs to verify the existence of QTLs that affect the quantitative trait in these regions. To facilitate such identification, Manhattan graphs were created and are presented in Figure 1. Backfat thickness presented higher polygenic behaviour, as the effects were distributed uniformly along the chromosomes.

The SNPs with the largest effects for androstenone concentration were found in the final third of chromosome 6. This result agrees with those of Lee *et al.* (2005) and Gregersen *et al.* (2012), who detected significant QTLs in the same region for this trait in a cross-bred (Large White x Meishan) and Danish Duroc populations, respectively. The effect observed on chromosome 14 is also in agreement with the results of Gregersen *et al.* (2012) and Markljung *et al.* (2008), who worked with a cross between Finnish Landrace and Swedish Hampshire. Regarding backfat loin depth, the SNPs with the largest effects were found in the final third of chromosome 11 and the initial third of chromosome 5, corroborating with the

**Figure 1** Distribution and plot of markers absolute effects for: (a) concentration of androstenone, (b) backfat thickness, (c) loin depth, (d) concentration of skatole.

results of van Wijk *et al.* (2007). Some of the most significant SNPs for skatole concentration were found in the final third of chromosome 13 and the middle third of chromosome 14. These results are in agreement with those presented by Lee *et al.* (2005), who evaluated a population of pigs obtained from the cross between European Large White and Chinese Meishan.

With respect to computational efficiency, the dimensionality reduction methods required fewer computational resources than RR-BLUP, but this difference is most remarkable when compared with Bayesian methods, as reported by Long *et al.* (2011) and Colombani *et al.* (2012). For our data set, the average processing time using an Intel(R) i7-2600 (3.4 GHz) processor with 4 GB of RAM was 45 min, corresponding to approximately 0.08 s for the training data sets.

In general, the proposed SICR method presented the highest predictive ability values and was the most efficient for the prediction of phenotypic values, proving to be unbiased. In contrast, the PLS and SPLS methods presented the lowest predictive ability and were inefficient for prediction purposes; they also did not stand out for any of the traits considered. The RR-BLUP displayed lower predictive ability values than the dimensionality reduction methods, which include a selection of covariates (SPCR and SICR). However, RR-BLUP displayed better results than the methods that do not consider covariate selection (PCR, PLS and ICR). Most of the methods allowed for the identification of relevant SNPs associated with the traits evaluated. Moreover, these relevant SNPs are located in genomic regions previously reported as regions related to the presence of QTLs affecting the evaluated traits in this study.

References

- Azevedo C.F., Resende M. D. V., Silva F. F., Lopes P. S., Guimarães S. E. F. (2013) Independent component regression applied to genomic selection for carcass traits in pigs. *Pesqui. Agropecu. Bras.*, **48**, 619–626.
- Bishop C. (2006) Pattern Recognition and Machine Learning. Springer, Berlin. ISBN 0-387-31073-8.
- Boulesteix A.L., Strimmer K. (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.*, **8**, 32–44.
- Colombani C., Croiseau P., Fritz S., Guillaume F., Legarra A., Ducrocq V., Granié C.R. (2012) A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. *J. Dairy Sci.*, **95**, 2120–2131.
- Duijvesteijn N., Knol E., Merks J., Crooijmans R., Groenen M., Bovenhuis H., Harlizius B. (2010) A genome-wide association study on androstenedione levels in pigs reveals a cluster of candidate genes on chromosome 6. *BMC Genet.*, **20**, 11–42.
- Gregersen V.R., Conley L.N., Sonrensen K.K., Guldbrandtson B., Velandier I.H., Bendixen C. (2012) Genome-wide association scan and phased haplotype construction for quantitative trait loci affecting boar taint in three pig breeds. *BMC Genomics*, **13**, 1–22.
- Hoskuldsson P. (1998) PLS regression methods. *J. Chemom.*, **2**, 211–228.
- Hyvärinen A. (1998) New approximations of differential entropy for independent component analysis and projection pursuit. *Adv. Neural Inf. Process. Syst.*, **10**, 273–279.
- Lê Cao K.A., Rossouw D., Robert-Granié C., Besse P. (2008) A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 35.
- Lee G.J., Archibald A.L., Law A.S., Lloyd S., Wood J., Haley C.S. (2005) Detection of quantitative trait loci for androstenedione, skatole and boar taint in a cross between Large White and Meishan pigs. *Anim. Genet.*, **36**, 14–22.
- Long N., Gianola D., Rosa G.J.M., Weigel K.A. (2011) Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *J. Anim. Breed. Genet.*, **128**, 247–257.
- Lopes M.S., Silva F.F., Harlizius B., Duijvesteijn N., Lopes P.S., Guimarães S.E.F., Knol E.F. (2013) Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genet.*, **14**, 92.
- Luo W., Cheng D., Chen S., Wang L., Li Y., Ma X., Song X., Liu X., Li W., Liang J., Yan H., Zhao K., Wang C., Wang L., Zhang L. (2012) Genome-Wide Association Analysis of Meat Quality Traits in a Porcine Large White × Minzhu Intercross Population. *Int. J. Biol. Sci.*, **8**, 580–595.
- Macciotta A.N.P.P., Pintus M.A., Steri R., Pieramati C., Nicolazzi E.L., Santus E., Vicario D., Van Kaam J.T., Nardone A., Valentini A., Ajmone-Marsan P. (2010) Accuracies of direct genomic breeding values estimated in dairy cattle with a principal component approach. *J. Dairy Sci.*, **93**, 532–533.
- Markljung E., Braunschweig M.H., Karlsson-Mortensen P., Bruun C.S., Sawera M., Cho I.C., Hedebro V.E. (2008) Genome-wide identification of quantitative trait loci in a cross between Hampshire and Landrace II: meat quality traits. *BMC Genet.*, **9**, Article 22.
- 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.
- Moser G., Tier B., Crump R.E., Khatkar M.S., Raadsma H.W. (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.*, **41**, 41–53.
- Otto M. (1999) Chemometrics. Wiley, Weinheim.

- R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Ramos M.A., Duijvesteijn N., Knol E.F., Merks J.W.M., Bovenhuis H., Crooijmans R.P.M.A., Groenen M.A.M., Harlizius B. (2011) The distal end of porcine chromosome 6p is involved in the regulation of skatole levels in boars. *BMC Genet.*, **12**, Article 35.
- Resende M.D.V., Resende Junior M.F.R., Aguiar A.M., Abad J.I.M., Missiaglia A.A., Sansaloni C., Petroli C., Grattapalia D. (2010) Computação da seleção genômica ampla (GWS). Embrapa Florestas, Colombo.
- Resende M.F.R. Jr, Muñoz P., Resende M.D.V., Garrick D.J., Fernando R.L., Davis J.M., Jokela E.J., Martin A., Peter G.F., Kirst M. (2012) Accuracy of genomic selection methods in a standard data set of Loblolly Pine (*Pinus taeda* L.). *Genetics*, **190**, 1503–1510.
- Sellier P., Le Roy P., Fouilloux M., Gruand J., Bonneau M. (2000) Responses to restricted index selection and genetic parameters for fat androstenone level and sexual maturity status of young boars. *Livest. Prod. Sci.*, **63**, 265–274.
- Solberg T.R., Sonesson A.K., Woolliams J.A., Meuwissen T.H.E. (2009) Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol.*, **41**, 29.
- Tajet H., Andresen O., Meuwissen T. (2006) Estimation of genetic parameters of boar taint; skatole and androstenone and their correlations with sexual maturation. *Acta Vet. Scand.*, **48**(Suppl 1), S9.
- Tomiyama M., Oikawa T., Hoque M.A., Kanetani T., Mori H. (2009) Influence of early postweaning traits on genetic improvement of meat productivity in purebred Berkshire pigs. *J. Anim. Sci.*, **87**, 1613–1619.
- van Wijk H.J., Arts D.J.G., Matthews J.O., Webster M., Ducro B.J., Knol E.F. (2005) Genetic parameters for carcass composition and pork quality estimated in a commercial production chain. *J. Anim. Sci.*, **83**, 324–333.
- van Wijk H.J., Buschbell H., Dibbits B., Liefers S.C., Harlizius B., Heuven H.C.M., Knol E.F., Bovenhuis H., Groenen M.A.M. (2007) Variance component analysis of quantitative trait loci for pork carcass composition and meat quality on SSC4 and SSC11. *J. Ani. Sci.*, **85**, 22–30.
- Windig J.J., Mulder H.A., ten Napel J., Knol E.F., Mathur P.K., Crump R.E. (2012) Genetic parameters for androstenone, skatole, indole, and human nose scores as measures of boar taint and their relationship with finishing traits. *J. Anim. Sci.*, **90**, 2120–2129.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Computacional routines.

Appendix S2 Detailed description of the methods.