A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle

Carine Colombani Colombani, Pascal Croiseau, S. S. Fritz, F. F. Guillaume, Andres Legarra, Vincent Ducrocq, Christèle Robert-Granié

J. Dairy Sci. 95:2120–2131
http://dx.doi.org/10.3168/jds.2011-4647

A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle

C. Colombani,*[1] P. Croiseau,† S. Fritz,‡ F. Guillaume,§ A. Legarra,* V. Ducrocq,† and C. Robert-Granié*
*INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
†INRA, UMR1313-GABI, 78352 Jouy en Josas, France
‡UNCEIA, 149 rue de Bercy, 75595 Paris, France
§Institut de l'Élevage, 149 rue de Bercy, 75595 Paris, France

ABSTRACT

Genomic selection involves computing a prediction equation from the estimated effects of a large number of DNA markers based on a limited number of genotyped animals with phenotypes. The number of observations is much smaller than the number of independent variables, and the challenge is to find methods that perform well in this context. Partial least squares regression (PLS) and sparse PLS were used with a reference population of 3,940 genotyped and phenotyped French Holstein bulls and 39,738 polymorphic single nucleotide polymorphism markers. Partial least squares regression reduces the number of variables by projecting independent variables onto latent structures. Sparse PLS combines variable selection and modeling in a one-step procedure. Correlations between observed phenotypes and phenotypes predicted by PLS and sparse PLS were similar, but sparse PLS highlighted some genome regions more clearly. Both PLS and sparse PLS were more accurate than pedigree-based BLUP and generally provided lower correlations between observed and predicted phenotypes than did genomic BLUP. Furthermore, PLS and sparse PLS required similar computing time to genomic BLUP for the study of 6 traits.
Key words: partial least squares regression, sparse partial least squares, genomic selection, French dairy cattle

INTRODUCTION

Genomic selection relies on computing genomic estimated breeding values (**GEBV**) using high-density SNP marker data. Meuwissen et al. (2001) suggested a 2-step approach to calculate GEBV. First, the effects of SNP are estimated to obtain a prediction equation using a reference population in which the animals are genotyped and phenotyped. Then, GEBV are predicted for the genotyped animals (without phenotypes) from this equation.

In the past few years, the accuracy of GEBV provided by genomic selection has been assessed using different methods in dairy cattle populations in the United States, New Zealand, Australia, the Netherlands, and France, among others. A simple BLUP, as described in Meuwissen et al. (2001) and known as genomic BLUP (**GBLUP**) in subsequent literature, was used as the reference method. The simple BLUP assumes that all SNP have an effect sampled from the same normal distribution. Hayes et al. (2009) treated Australian Holstein-Friesian bull data using a method derived from BayesA, which exploits the prior knowledge that many SNP have small individual effects on the trait and only a few have moderate to large effects. The Bayesian method was shown to be slightly more reliable (+0.02 to +0.07 compared with the reliability of BLUP) for most traits. Using New Zealand dairy cattle, Harris et al. (2009) also compared the BLUP approach with Bayesian methods (BayesA and BayesB), in which some SNP may have zero effect (Meuwissen et al., 2001). Bayesian methods slightly improved reliability (2%), whereas the use of regression methods such as least angle regression (Efron et al., 2004) did not lead to any improvement. VanRaden et al. (2009) compared the reliability of GEBV in US and Canadian young bulls, using a method similar to GBLUP that fits the allelic effects of each SNP as random effects with a normal distribution with known variance (VanRaden, 2008), and a similar method to BayesA with a heavier tail distribution for the SNP effects. As in the Australian and New Zealand results, the Bayesian approach slightly increased reliability (1% compared with the reliability of GBLUP).

Moser et al. (2009) compared 5 methods on dairy bull data including regression methods (least squares regression), shrinkage methods [Bayes regression (**Bayes-R**) similar to BayesA, and random regression BLUP (**RR-BLUP**), comparable to GBLUP], sup-

Received June 22, 2011.
Accepted December 9, 2011.
[1]Corresponding author: carine.colombani@toulouse.inra.fr

port vector machine learning methods (nonparametric support vector regression), and dimension reduction methods such as partial least squares (**PLS**) regression. The accuracy of Bayes-R, RR-BLUP, PLS, and support vector regression was very similar for the 2 traits studied by these authors. However, PLS and RR-BLUP required substantially less computation time than the Bayesian method.

Using simulated data, Coster et al. (2010) demonstrated the superiority of PLS over Bayesian methods with regard to the stability of results according to the number of QTL or the distribution of QTL variance. They also showed that the computation time for the PLS method required to fit, cross validate, and evaluate the models was less than that for the Bayesian method. However, the Bayesian method was more accurate. Solberg et al. (2009) also used simulated data to compare PLS and principal component regression with BayesB. They obtained the same results: BayesB was more accurate than other methods but PLS and principal component regression were computationally faster and simpler.

The PLS regression (Wold et al., 2001) appears to be an efficient method to deal with genomic selection data, both in its capacity to handle large data sets and its prediction ability. This approach is particularly suitable when the matrix of predictors has more variables than observations, and when multicollinearity exists among variables. The sparse PLS regression (**sPLS**, Lê Cao et al., 2008) is a recent approach that combines variable selection and modeling in a one-step procedure. Dimension reduction methods and variable selection approaches may be an attractive way to deal with the increasing number of markers used in genomic selection in dairy cattle by limiting computing time (Coster et al., 2010). Furthermore, even though PLS has already been studied in a genomic evaluation context, the authors used simulated data and did not compare PLS accuracy with current genomic selection methods such as BLUP and GBLUP (Solberg et al., 2009; Coster et al., 2010). Long et al. (2011) introduced sparsity in PLS and tested the predictive ability of sPLS versus principal component regression and PLS methods but did not apply other current genomic selection methods on their real data. They showed that combining dimension reduction and variable selection for accurate prediction of genomic breeding values was promising.

The aim of the present study was to compare PLS and sPLS on a real data set with other methods currently used in the evaluation of dairy cattle such as pedigree-based BLUP and GBLUP. Both PLS and sPLS regressions were compared based on their predictive abilities and then with pedigree-based BLUP and GBLUP results to evaluate their accuracy.

MATERIALS AND METHODS

Data

A data set of genotyped French Holstein bulls was split into a training data set and a validation data set using a cut-off birth date defined so that the validation set included the youngest 25% genotyped bulls. First, the prediction equation was estimated with the training data set, which comprised 2,976 genotyped and phenotyped Holstein bulls born before June 2002. Then, phenotypes were predicted for the bulls in the validation data set, which comprised 964 bulls (born between June 2002 and 2004).

Genotypes for 39,738 polymorphic SNP were used as independent variables. The selected SNP, provided by the Illumina Bovine SNP50 Beadchip (Illumina, San Diego, CA), had minor allele frequencies >3%. Mendelian segregation was checked. Missing genotypes were inferred from large family information with a low error rate using DualPHASE software (Druet and Georges, 2009).

Six traits with different heritability were used as dependent variables: milk yield, fat yield, and protein yield ($h^2 = 0.3$), fat content and protein content ($h^2 = 0.5$), and conception rate ($h^2 = 0.02$; Boichard and Manfredi, 1994). The bulls' phenotypes used in this study were daughter yield deviations (**DYD**, VanRaden and Wiggans, 1991; Mrode and Swanson, 2004) from a French national evaluation October 2009; that is, the average performance of the daughters of a sire, adjusted for fixed and nongenetic random effects and for the additive genetic value of their dam. For each DYD, a weighting was added in the form of the effective daughter contribution (**EDC**; VanRaden and Wiggans, 1991; Fikse and Banos, 2001). To be included in the analysis, each observation required an EDC >20.

BLUP and GBLUP

The general statistical model in BLUP and GBLUP is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes (DYD), μ is the mean, $\mathbf{Z}$ is a design matrix allocating observations to breeding values, $\mathbf{g}$ is a random vector of additive genetic values, and $\mathbf{e}$ is a vector of random normal errors. In BLUP, $\mathrm{Var}(\mathbf{g}) = \mathbf{A}\sigma_g^2$, where $\mathbf{A}$ is the pedigree-based relationship matrix, and σ_g^2 is the additive genetic variance. In GBLUP, $\mathrm{Var}(\mathbf{g}) = \mathbf{G}\sigma_g^2$, and $\mathbf{G}$ is the genomic relationship matrix as defined by VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{WW}'}{2\sum_{j=1}^{p} q_j\left(1-q_j\right)},$$

where p is the number of loci considered, q_j is the frequency of an allele of the marker j, and $\mathbf{W}$ is a centered incidence matrix of SNP genotypes. The SNP marker effects are assumed to have a prior normal distribution and mixed model equations are used with the genomic relationship matrix (Cole et al., 2009; VanRaden et al., 2009).

PLS and sPLS Regression

PLS. The PLS regression introduced by Wold (1966) is a data analysis method that generalizes and combines principal component analysis and multiple regression. The method was mainly developed for industrial applications (petroleum and food processing industries) and the social sciences. The PLS method was designed to deal with the "$p \gg n$ problem"; that is, when the number of independent variables (p) is much larger than the number of observations (n). Partial least squares regression is very useful to predict dependent variables from a very large number of predictors that might be highly correlated.

In its general form, the PLS regression replaces the initial independent variable space ($\mathbf{X}$) and the initial response variable space ($\mathbf{Y}$) by smaller spaces that rely on a reduced number of variables named latent variables, which are included one by one in an iterative process. These factors will be the new variables of a usual linear regression. The main idea is to perform successive regressions by projections onto latent structures to reveal hidden or latent underlying biological effects (Wold et al., 2004; Lê Cao et al., 2008).

Using the same notation as in Lê Cao et al. (2008), the PLS regression looks for a decomposition of centered data matrices $\mathbf{X}$ and $\mathbf{Y}$ in terms of component scores, called latent variables: ($\xi_1, \ldots, \xi_h, \ldots, \xi_H$) and ($\omega_1, \ldots, \omega_h, \ldots, \omega_H$), which are linear combinations of the columns of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and associated loading vectors: ($\mathbf{u}_1, \ldots, \mathbf{u}_h, \ldots, \mathbf{u}_H$) and ($\mathbf{v}_1, \ldots, \mathbf{v}_h, \ldots, \mathbf{v}_H$), where H is the number of latent variables retained in the final model. However, the regression coefficients that define these components are not linear, as they are solved via successive local regressions on the latent variables. The loading vectors are estimated to solve the following optimization problem:

$$\max_{\mathbf{u_h}=1,\mathbf{v_h}=1} \operatorname{cov}\left(\mathbf{X}_{h-1}\mathbf{u_h}, \mathbf{Y}\mathbf{v_h}\right),$$

where $\mathbf{X}_{h-1}$ is the residual $\mathbf{X}$ matrix in the regression of $\mathbf{Y}$ on ($\xi_1, \ldots, \xi_{h-1}$) for each dimension $h = 1, \ldots, H$, and the associated latent variables are denoted $\xi_h = \mathbf{X}_{h-1}\,\mathbf{u}_h$ and $\boldsymbol{\omega}_h = \mathbf{Y}\mathbf{v}_h$.

As in principal component analysis, the loading vectors and the latent variables are directly interpretable. The loading vectors $\mathbf{u}_h$ and $\mathbf{v}_h$ indicate how the x_j and y_i variables explain the relationship between $\mathbf{X}$ and $\mathbf{Y}$. The latent variables contain information regarding similarities or dissimilarities between individuals (Wold et al., 2004).

sPLS. The sPLS regression (Lê Cao et al., 2008) aims at combining variable selection and modeling in a one-step procedure. It was first proposed to handle transcriptomic data and was adapted for genomic data in this study. To understand the sPLS approach, it is helpful to describe first the principle of the PLS-singular value decomposition (Lorber et al., 1987) that solves PLS problems efficiently by decomposing the $\mathbf{X}'\mathbf{Y}$ matrix into singular values and vectors.

For a real matrix $\mathbf{M}$ ($p \times q$) of rank r, the singular value decomposition of $\mathbf{M}$ can be obtained as follows:

$$\mathbf{M} = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Theta}'$$

where $\boldsymbol{\Gamma}$ ($p \times r$) and $\boldsymbol{\Theta}$ ($q \times r$) are orthonormal and $\boldsymbol{\Delta}$ ($r \times r$) is a diagonal matrix with singular values δ_k ($k = 1 \ldots r$).

The loading vectors $\mathbf{u_1}$ and $\mathbf{v_1}$ of $\mathbf{X}$ and $\mathbf{Y}$, respectively, correspond to the first singular vectors γ_1 and θ_1 if $\mathbf{M} = \mathbf{X}'\mathbf{Y}$. Then, for $h = 2, \ldots, H$, $\mathbf{M}_h$ is directly deflated by its rank-one approximation, as explained in Lê Cao et al. (2008): $\mathbf{M_h} = \mathbf{M_{h-1}} - \delta_h \mathbf{u_h} \mathbf{v_h}'$.

Sparsity of the loading vectors is introduced iteratively by penalizing both $\mathbf{u_h}$ and $\mathbf{v_h}$ with a soft-thresholding penalization, as for sparse principal component analysis (Shen and Huang, 2008). The optimization problem becomes

$$\min_{\mathbf{u,v}} \mathbf{M} - \mathbf{u}\mathbf{v}'^2_F + g_{\lambda_1}\left(\mathbf{u}\right) + g_{\lambda_2}\left(\mathbf{v}\right),$$

where $g_{\lambda_1}\left(x\right) = sign\left(x\right)\left(\mid x \mid - \lambda_1\right)_+$ is the soft-thresholding penalty function.

When no sparsity is required, the same results are obtained as in classical PLS. The details of this algorithm are presented in Lê Cao et al. (2008). Although PLS and sPLS can perform multi-trait analyses, each trait in this study was considered independently with $\mathbf{X}$ the matrix of SNP and $\mathbf{y}$ the vector of phenotypes of one trait. In this case, the sPLS used here is similar to the sPLS introduced by Chun and Keles (2010) and used by Long et al. (2011), in a genomic selection con-

text. When one trait is studied, PLS and sPLS are easy to implement and not time consuming because they are based on successive regressions and do not require matrix inversion.

Parameter Tuning

In both PLS and sPLS, the optimal number H of dimensions has to be determined. The parameter H can be tuned by cross-validation as in the original PLS and as proposed by Chun and Keles (2010). Coster et al. (2010) also proposed to use cross-validation to find the number of dimensions that minimized the prediction error. In this study, the root mean squared error of prediction (**RMSEP**) was minimized with 10-fold cross-validation in the training data set and for each given dimension h (Mevik and Cederkvist, 2004):

$$RMSEP = \sqrt{\frac{1}{10}\sum_{k=1}^{10}\left(\hat{\mathbf{y}}_k - \mathbf{y}_k\right)},$$

where $\hat{\mathbf{y}}_k$ is the vector of predicted values for the sample k. Solberg et al. (2009) suggested keeping the number of dimensions leading to the highest correlation between predicted values and observed values in the validation data set. This approach was also tested in this study and the results of the 2 ways used to fix H are discussed in the Results and Discussion section.

In sPLS regression, in addition to H (that was selected as above), the number of variables selected in each dimension of the model has to be fixed. Long et al. (2011) tested different values of H and different values of the number of variables selected in each dimension, based on results of their PCR study, to maximize the cross-validation correlation. In our study, we chose to minimize the RMSEP with 10-fold cross-validation in the training data set, for the previously fixed number of dimensions H. In practice, for each trait, several sPLS were performed depending on the number of selected SNP in each latent variable or dimension (assumed to be constant), as a percentage of the number of SNP in the whole data set. By construction, the same SNP could be selected in several dimensions. Ten sPLS regressions (keeping 0.2 to 10% of all SNP for each dimension considered) and the PLS regression were tested using dimensions 1 to 100.

Importance of SNP Effects

To enable better interpretation of the models, coefficients that represent the power of x_j to explain $\mathbf{y}$ has to be defined. The "variable importance in projection" (**VIP**) coefficients measure the contribution of x_j to the construction of $\mathbf{y}$ through latent variables $\boldsymbol{\xi}_\mathbf{h}$, ($h = 1,\ldots, H$) and is defined by

$$VIP_{Hj} = \sqrt{\frac{p}{\sum_{h=1}^{H} cor^2\left(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h}\right)}\sum_{h=1}^{H} cor^2\left(\mathbf{y}, \boldsymbol{\xi}_\mathbf{h}\right)\omega_{hj}^2},$$

with

$$\sum_{j=1}^{p} VIP_{Hj}^2 = p.$$

The sum of squares of the VIP coefficients of all the SNP in one dimension of the PLS models is equal to the number of independent variables. Thus, the VIP coefficient of a SNP is related to the number of SNP that have a nonzero effect in the model.

The contribution of x_j to the construction of $\xi_\mathbf{h}$ is measured by its weight ω_{hj}, provided by PLS or sPLS. Although the weight ω_{hj} of x_j is interpretable, it does not account for the contribution of the latent variable $\xi_\mathbf{h}$. The VIP coefficients are able to classify the variables x_j according to their weight in each latent variable and the weight of each latent variable in the construction of $\mathbf{y}$. So they could be considered as an evaluation of the effects of SNP on the prediction of $\mathbf{y}$. Both PLS and sPLS were performed using the R package named "mixOmics" (Lê Cao et al., 2009).

Comparison of Methods and EDC

Prediction Equation. In this study, we compared 2 currently used methods for the evaluation of dairy cattle, BLUP and GBLUP, with PLS and sPLS regressions. The application of the different methods followed the same pattern, regardless of the method. The prediction equation was estimated using the training data set. The $\mathbf{y}$ phenotypes were DYD. One EDC was associated with each DYD, reflecting its uncertainty. This generates heterogeneity of variances, so that the ith DYD has a (pseudo-residual) variance σ_e^2/EDC_i (VanRaden and Wiggans, 1991). An equivalent model was constructed in all cases (BLUP, GBLUP, PLS, and sPLS), multiplying y_i and the ith row of the incidence matrix by the square root of the EDC to obtain homogeneous variances.

Accuracy of the Methods. Two criteria were used to test the accuracy of the different methods and to compare their predictive ability: the correlation (ρ) between observed and predicted values and the regression slope (b) of observed to predicted values (a value of 1 is expected; Henderson, 1963). The bulls in the validation data set were predicted by the prediction

equations provided by the different methods. Bulls in the validation set were progeny tested, so that observed DYD and associated EDC were also available for this population. These weights were taken into account in the calculation of the correlation, using

$$\rho_{xy} = \frac{\sum_i w_i \left(x_i - \overline{x}_w\right)\left(y_i - \overline{y}_w\right)}{\sqrt{\sum_i w_i \left(x_i - \overline{x}_w\right)^2 \sum_i w_i \left(y_i - \overline{y}_w\right)^2}},$$

where

$$\overline{x}_w = \frac{\sum_i w_i x_i}{\sum_i w_i} \text{ and } \overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i},$$

and w_i are the weights (EDC) for each observation. In the regression of observed DYD onto predicted DYD, EDC were introduced in the same way as for the production of prediction equations.

The Hotelling-Williams procedure was used to test for differences between the correlations obtained from the different methods. It tests the null hypothesis of equality between 2 dependent correlations that share a variable (Steiger, 1980; VanSickle, 2003). Under the null hypothesis, the statistical test is distributed as t with $n - 3$ degrees of freedom. All the correlations discussed in this study were compared with one another using the Hotelling-Williams test with a 5% threshold.

RESULTS AND DISCUSSION

Parameter Tuning

Figure 1 presents RMSEP obtained by cross-validation in the training data set by PLS, according to the number of dimensions. For milk yield, fat yield, protein yield, and protein content, the pattern of the different curves became very stable after only 10 dimensions. For fat content and conception rate, RMSEP stabilized after about 30 dimensions. The minimum RMSEP was reached after around 20 dimensions for milk yield, protein yield, fat content, and conception rate and after around 30 dimensions for fat yield and protein content. However, the differences in RMSEP between 2 values for the number of dimensions were very small (the minimum of the curves was not accentuated), so cross-validation did not appear to be the best criterion to choose the number of dimensions. The same conclusions were reached using the sparse PLS approach, so the number of dimensions that led to the highest correlation between phenotypes and predicted values in the validation data set was kept. This was also done for practical reasons. In fact, creating pseudo-training and pseudo-validation data sets within the training population to calibrate H was difficult, especially if a time structure (old vs. young) had to be used. Partial least squares regression was tested up to dimension 100 but the correlations obtained after more than 50 dimensions no longer increased for most traits. Figure 2 shows the correlations between observed phenotypes and predicted values in the validation data set obtained by PLS as a function of the number of latent variables built for each trait. The pattern of the different curves was the same whatever the trait: the correlation continued to increase with the number of dimensions until a plateau was reached at around dimension 30 for conception rate, 40 for milk yield, fat yield, and protein yield, and 80 for fat content and protein content. The number of dimensions in the final model was fixed at these minimum values to avoid overfitting the data (Abdi, 2010). As can be seen in Figure 2, the number of dimensions is not critical, and therefore the choice of H using the validation data set did not overestimate the predictive ability of PLS.

Sparse PLS required 2 parameters: the number of latent variables (H) and the proportion of SNP selected in each dimension. The same criterion as in PLS was used to fix the number of dimensions kept in the sparse PLS models. The pattern was the same as in PLS (results not shown): a plateau was reached at around dimension 25 for conception rate, 40 for milk, fat, and protein yields, and 50 for fat and protein contents.

The proportion of SNP selected by sPLS was tuned by cross-validation within the training data set. Table 1 shows the RMSEP provided by PLS and the different sPLS (according to the proportion of SNP selected in each dimension) for each trait and for the previously fixed number of dimensions. The selected proportion of SNP for sPLS was the one that minimized RMSEP (Table 1, in bold, based on 3 decimal places). The minimum RMSEP obtained from sPLS was close to that obtained from PLS. For example, for fat yield, the RMSEP obtained with sPLS with 2% of the total number of SNP for each dimension was the smallest (0.18) and was the same as the error of prediction obtained with PLS. The heritability of the trait played a role in the magnitude of the RMSEP: traits with the same heritability led to similar prediction errors. Milk, fat, and protein yield ($h^2 = 0.3$) obtained an RMSEP of around 0.15. Fat and protein content ($h^2 = 0.5$) gave an RMSEP of 0.56. Conception rate had the highest error (0.79) but this result was the same as the error with the PLS model (0.78). As DYD were pseudo-phenotypes calculated from the performance of the bull's daughters, low heritability traits were difficult to predict.

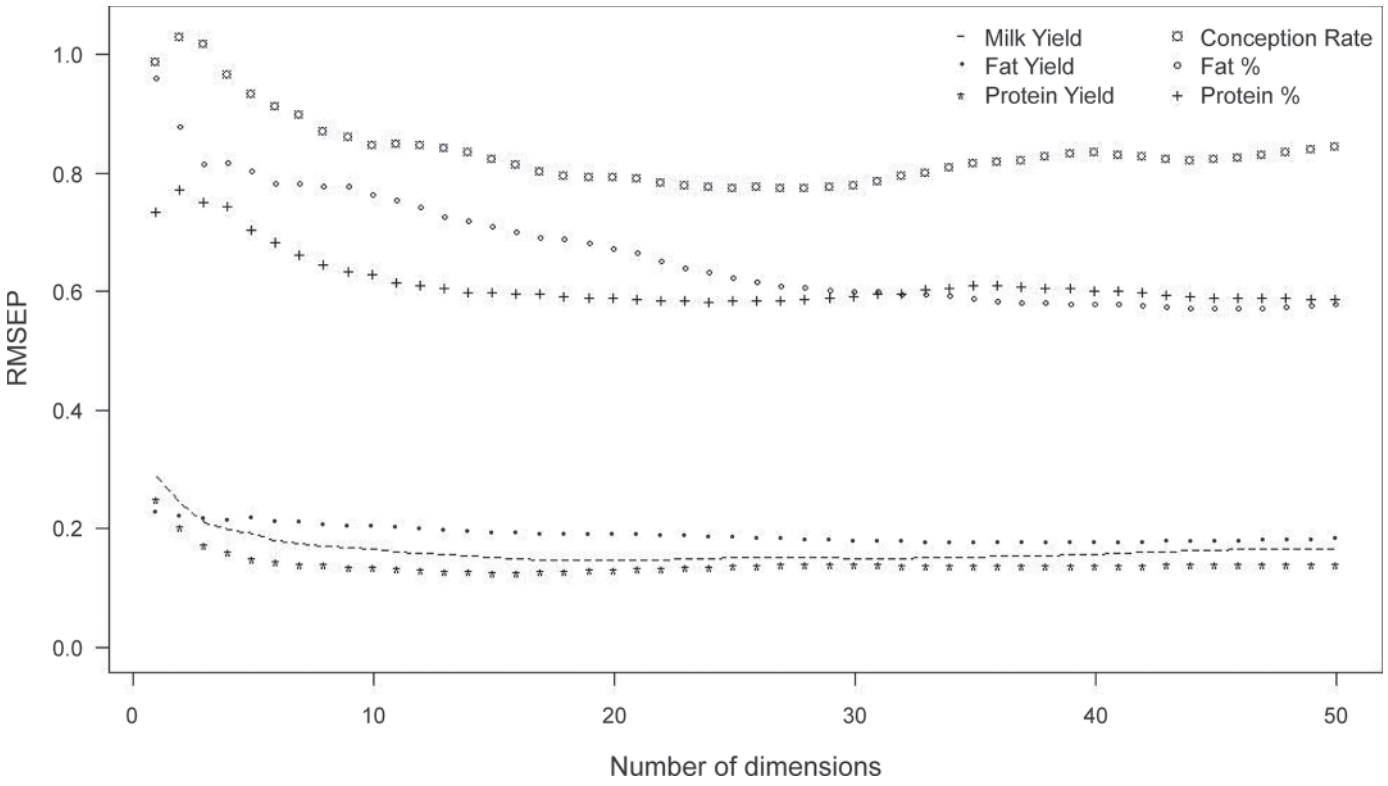

Figure 1. Root mean square error of prediction (RMSEP) for the 6 traits studied plotted against the number of dimensions for partial least squares regression.

Genomic Predictions with PLS and sPLS

Table 2 shows the accuracy of PLS and sPLS regressions for the different traits, with the number of SNP in sPLS that led to the minimum RMSEP, and the number of dimensions that led to the maximum correlation. The PLS regression gave significantly higher correlations whatever the trait with an average increase of 0.04 compared with sparse PLS. The best correlations were obtained for fat and protein content (0.70 and 0.71 in PLS, 0.66 and 0.65 in sPLS, respectively). Using the Hotelling-Williams test with a 5% threshold, production traits with a higher heritability, such as milk, fat, and protein yield ($h^2 = 0.3$), and fat and protein content ($h^2 = 0.5$) were predicted more accurately (from 0.53 in PLS and 0.48 in sPLS for milk yield to 0.71 in PLS for protein content and 0.66 in sPLS for fat content) than traits with lower heritability, such as conception rate ($h^2 = 0.02$) with an accuracy of 0.33 in PLS and 0.29 in sPLS. Thus, accuracy of prediction and heritability of the trait are closely related. Moser et al. (2010) processed data from 2,144 Holstein-Friesian bulls and compared accuracy between traits of different heritability. Production traits such as protein content, fat content, and milk yield, which have high heritability (0.56, 0.52, and 0.28, respectively), achieved higher accuracy than survival ($h^2 = 0.03$), which showed similar heritability to conception rate in the present study. For conception rate, a larger reference population is required to achieve the same level of accuracy as for production traits (Hayes et al., 2009). However, unlike the data sets used by Hayes et al. (2009), which contained only 332 Australian Holstein bulls for fertility and 798 for the other traits, the data sets of this study for conception rate used a similar number of bulls with as many daughters to evaluate DYD as the number of bulls and daughters used for production traits. Therefore, for conception rate, the power of the analysis was not reduced by a smaller data set but by low heritability. Regarding the regression slope b, both PLS and sPLS gave values below 1, with PLS values closer to 1 (from 0.60 for conception rate to 0.83 for protein content) than those for sPLS (from 0.53 for milk yield to 0.76 for protein content). Furthermore, the relationship between the heritability of the trait and slope was less clear than between heritability and the correlation. For example, the sparse PLS slopes for milk yield ($b = 0.53$) and for conception rate ($b = 0.54$) were similar but milk yield is a trait with a moderate heritability ($h^2 = 0.3$), whereas conception rate is a low heritability trait (h^2

Table 1. Root mean square error of prediction (RMSEP) in partial least squares (PLS) and each sparse PLS tested as a function of the percentage of SNP selected in each latent variable [minimum RMSEP (3 decimal places) in bold]

	Sparse PLS (% of the SNP data set selected)										PLS
Variable	0.2	0.4	0.6	0.8	1	2	3	4	5	10	100
Milk yield	0.17	0.16	0.17	0.16	0.16	0.16	0.15	**0.15**	0.15	0.15	0.15
Fat yield	0.18	0.20	0.19	0.18	0.21	**0.18**	0.18	0.19	0.18	0.18	0.18
Protein yield	0.14	0.15	0.14	0.14	0.14	0.14	0.13	0.13	0.13	**0.13**	0.13
Fat content	0.60	0.67	0.58	**0.56**	0.57	0.58	0.57	0.59	0.63	0.60	0.57
Protein content	0.64	0.62	0.59	0.60	0.58	0.57	0.56	0.59	**0.56**	0.56	0.58
Conception rate	0.94	0.83	0.81	0.83	0.86	0.81	0.87	**0.79**	0.83	0.82	0.78

= 0.02). Two traits with 2 different heritabilities would lead to 2 different correlations but not necessarily to 2 different regression slopes.

Sparse PLS gave significantly less accurate predictions than PLS. However, sPLS performed a variable selection by allowing the number of variables in the final model to be reduced by 50% (Table 2). The number of selected SNP was reduced to 9,832 for fat content. One explanation could be the presence of *DGAT1* (Grisart et al., 2004), a gene on bovine chromosome 14 that leads to a mutation that has a major effect on fat content in milk in Holstein dairy cattle. Therefore, a small number of SNP, of which many were located around this QTL, was sufficient to obtain accurate predictions. However, the number of selected SNP could have been smaller, but the large number of latent components (50) used in the model led to a high number of selected SNP, irrespective of the presence of a large QTL.

Indeed, the number of SNP in the final model was directly related to the percentage of SNP kept for each latent variable and to the number of latent variables. The large number of SNP selected for protein yield was the consequence of the large proportion of SNP (10%) selected for each of the 38 latent variables. Fifty latent variables were used for fat content but the number of SNP was reduced because the sPLS that led to the

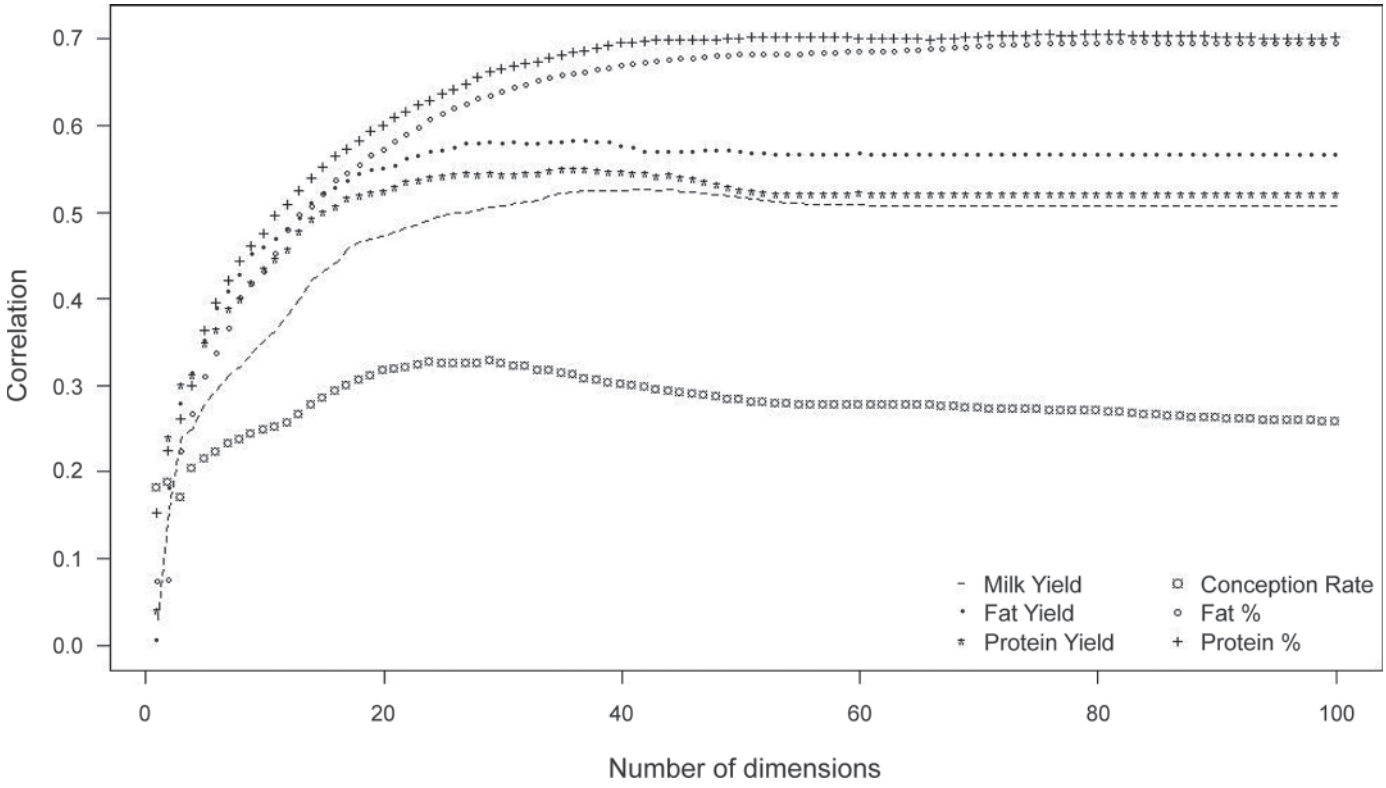

Figure 2. Correlation between observed and predicted daughter yield deviations for the 6 traits studied plotted against the number of dimensions for partial least squares regression.

Table 2. Effective daughter contribution-weighted correlations (ρ) and regression slopes (*b*) provided by partial least squares regression (PLS) and sparse PLS (sPLS)

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
PLS						
ρ	0.53	0.58	0.55	0.70	0.71	0.33
b	0.65	0.83	0.67	0.80	0.83	0.60
Dim[1]	42	37	36	83	75	29
sPLS						
ρ	0.48	0.54	0.51	0.66	0.65	0.29
b	0.53	0.70	0.60	0.69	0.76	0.54
Dim	44	43	38	50	51	27
No. of SNP[2]	22,948	16,296	32,578	9,832	26,034	20,150

[1]Number of latent variables or dimensions included in the final model.

[2]Number of SNP selected by sPLS.

minimum RMSEP kept only 0.8% of SNP at each dimension. Finally, for both milk yield and conception rate, 4% of SNP on each dimension were required and similar numbers of SNP were kept (22,948 and 20,150, respectively) but with a much larger number of latent variables for milk yield (44 vs. 27 for conception rate) because the same SNP can be selected for several latent variables.

The number of latent variables required by PLS and sPLS was very high (between 27 and 83 dimensions depending on the trait). Long et al. (2011) evaluated PLS, sPLS developed by Chun and Keles (2010), and principal component regression in Holstein bulls. They showed that to predict milk yield in Holsteins by PLS, only 15 latent components were sufficient to obtain the largest predictive correlation, suggesting a strong predictive power of the latent variables. The lack of predictive power of the additional latent components in the present study seems to be due to the presence of highly EDC-weighted bulls in the training data set, which had a major effect on the distribution of the phenotypes.

Effect of EDC on PLS and sPLS

The distribution of the phenotypes in the training data set was normal, but applying EDC disturbed the normal distribution of the phenotypes (results not shown). Furthermore, the correlation between observed DYD and observed weighted DYD was surprisingly small: about 0.20 for milk, fat, and protein yield, about 0.45 for fat and protein content, and about 0.35 for conception rate. Consequently, we investigated the effect of the use of EDC on the training population.

Figure 3 shows the distribution of EDC in the training data set for the 6 traits studied. The computation of EDC relies on the number of daughters per bull and trait parameters (heritability and repeatability). The EDC were the same for milk yield, fat yield, and protein yield and fat content and protein content. In order for the information content in conception rate to be consistent with that of production traits, a large number of daughters per bull was assumed both for production traits and conception rate. The graphs show that some bulls differed from others in their very high EDC. These bulls were generally older than the other bulls in the training population and did not obtain stronger DYD. The significant difference in weights between bulls resulted in a bias and had a major effect on the number of latent components introduced in the PLS and sPLS models. To test this hypothesis, the same study was performed without considering EDC either in the weighting of DYD or in the calculation of accuracy and of the regression slope.

Table 3 shows the accuracy of PLS and sPLS regressions with respect to the different traits without EDC. For production traits, with PLS, the correlations were very similar to the results of PLS with EDC (Table 2). With sPLS, accuracy was significantly better in the study without EDC (fat yield $\rho = 0.59$ and protein content $\rho = 0.72$, for example) than in the study including EDC (fat yield $\rho = 0.54$ and protein content $\rho = 0.65$). For conception rate, using EDC reduced significantly the accuracy of both PLS and sPLS. Sparse PLS and PLS gave no significantly different correlations without EDC, whereas sPLS was shown to be significantly less accurate than PLS with EDC. The regression slopes were differently affected by the use or nonuse of EDC, with regression slopes lower than or close to those in the previous study with PLS, and regression slopes greater than or equal to those in the previous study with sPLS. Irrespective of the trait, the number of dimensions was considerably reduced with both methods, which led to a stronger variable selection in sPLS and a restricted number of SNP in the prediction equation. Only 10 or fewer latent variables were needed to obtain the best correlations for fat yield, protein yield, and conception rate with both PLS and sPLS. These results are in

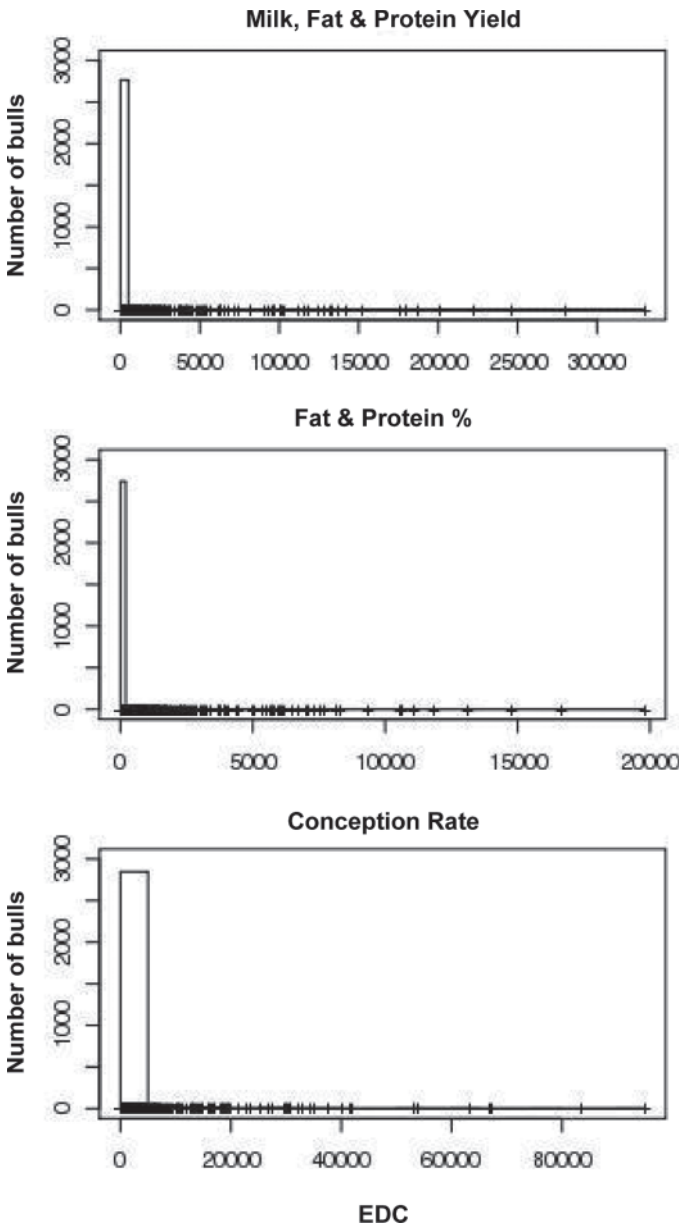

Figure 3. Distribution of effective daughter contribution (EDC) in the training data set.

agreement with the number of dimensions obtained by Long et al. (2011). For protein yield and protein content, the number of SNP remained high because the sPLS, with the minimum RMSEP by cross-validation, kept 10% and 5% of the SNP in each dimension. Therefore, introducing EDC did not have a major effect on the predictive ability of PLS but did affect the number of latent variables of the model. With PLS methods, it is probably wiser to have a more homogeneous distribution of the weights to favor the dimension-reducing ability of the PLS variants and hence to reduce computation time. In the remainder of the study, EDC were not included.

Both PLS and sPLS seemed to fit well in the context of genomic selection, but sPLS led to slightly smaller correlations than PLS but with no significant differences when EDC were not considered. However, sPLS favored a variable selection that can highlight the most important SNP, if required. A secondary aim of the present study was to underline the explanatory power of PLS regarding biological processes. However, to interpret the model in a biological context, coefficients that represent the explanatory power of the variables in the construction of the response variable are needed.

VIP Coefficients of PLS and sPLS

Figure 4 shows the VIP coefficient computed for each variable according to the position of the SNP on the genome. All the SNP variables are shown in the graph. A large number of VIP coefficients were set to zero in sPLS, whereas all the coefficients differed from zero in PLS. Therefore, sPLS was able to select variables based on VIP coefficients. The scale of the y-axis was the same for all the traits except fat content, which had the highest VIP coefficients.

Excluding the use of EDC, the variable selection performed by sPLS highlighted areas of interest. For fat yield, some SNP on chromosomes 2 and 5 were already highlighted by PLS but were clearly weighted up in sPLS (VIP coefficients of around 2.2 and 2.5 in PLS and of 12 and 16 in sparse PLS, respectively). For fat content, the SNP located at the beginning of chromosome 14 were highlighted by both methods. This location corresponds to a region of the genome that hosts the QTL *DGAT1* (Grisart et al., 2004). Furthermore, chromosomes 5 and 20 were more clearly revealed by sPLS than by PLS. The same comments can be made for most traits. The differences between PLS and sPLS were obvious for all traits with higher VIP coefficients in sPLS due to the reduced number of SNP considered. Conception rate showed many regions of interest with both methods. The aim of this study was to highlight differences and similarities between PLS and sPLS. We are not yet able to affirm that the different genome areas localized by PLS or sPLS correspond to QTL locations. This is currently under study (Colombani et al., 2011).

Comparing PLS and sPLS with Current Methods

The aim of this study was to compare PLS variants with currently used methods in the evaluation of dairy cattle. Table 4 shows the correlation between observed and estimated DYD for all the traits, with 3 methods

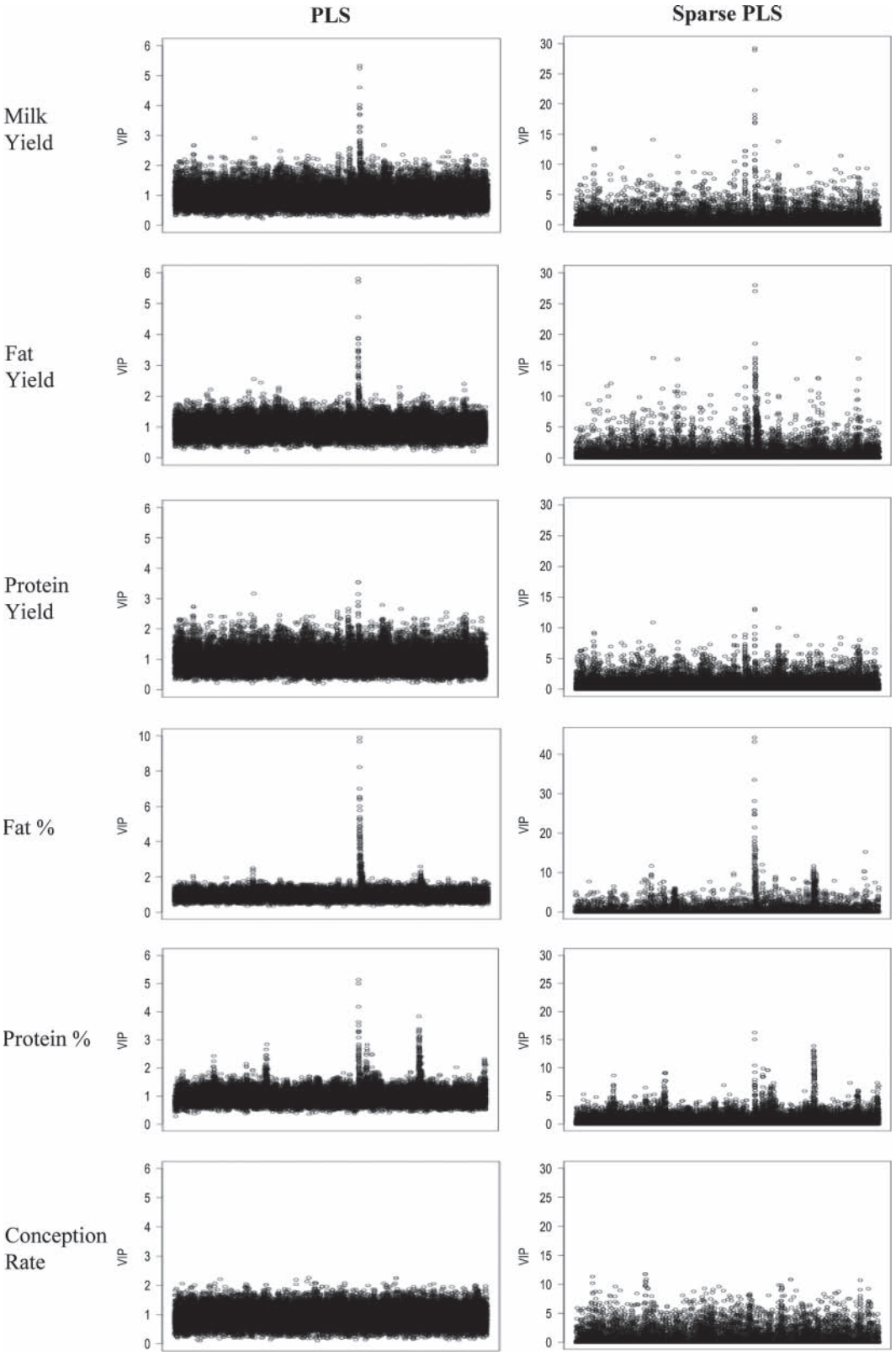

Figure 4. Variable importance in projection (VIP) coefficients from partial least squares regression (PLS) and sparse PLS for the 6 traits without effective daughter contribution (EDC).

Table 3. Correlations (ρ) and regression slopes (*b*) provided by partial least squares regression (PLS) and sparse PLS (sPLS) without effective daughter contribution

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
PLS						
ρ	0.52	0.58	0.55	0.71	0.71	0.25
b	0.60	0.77	0.64	0.80	0.82	0.62
Dim[1]	20	9	10	23	26	3
sPLS						
ρ	0.50	0.59	0.53	0.72	0.72	0.21
b	0.56	0.76	0.59	0.81	0.77	0.49
Dim	14	10	10	11	23	3
No. of SNP[2]	17,408	6,779	26,458	2,870	27,191	4,592

[1]Number of latent variables or dimensions included in the final model.

[2]Number of SNP selected by sPLS.

of genomic selection: PLS, sPLS, and GBLUP, compared with pedigree-based BLUP. On average, the correlations obtained by genomic selection methods were significantly higher than with pedigree-based BLUP for the 5 production traits concerned (0.426 for pedigree-based BLUP and 0.614, 0.612, and 0.630 for PLS, sPLS, and GBLUP, respectively). The differences between pedigree-based BLUP and genomic selection methods were not as clear for the conception rate trait, with a correlation of 0.28 for BLUP and 0.21 and 0.35 for sPLS and GBLUP. For conception rate, the BLUP correlation and the PLS correlation were not significantly different. Genomic BLUP (from 0.35 to 0.73), PLS (from 0.25 to 0.71), and sPLS (from 0.21 to 0.72) gave similar results for all traits concerned except for milk yield and conception rate, for which GBLUP gave significantly better results.

As expected, the genomic selection methods tested in this study were more efficient than pedigree-based BLUP. Genomic BLUP was accurate for use with French Holstein data with significantly higher accuracy for some traits. However, PLS and sPLS methods were comparable to GBLUP if we considered one trait at a time. Regarding computing time, GBLUP requires one inversion of the genomic relationship matrix for all traits, which took about 1 h. Then, once the genomic relationship matrix was inverted, computation was a matter of seconds. For each trait, PLS took about 10 min and sPLS took less than 10 min, depending on the number of SNP selected. The disadvantage of PLS and sPLS with respect to GBLUP for some traits could be overcome by building a multi-trait model.

CONCLUSIONS

Sparse PLS regression was compared with PLS and with 2 currently used methods in the evaluation of dairy cattle: pedigree-based BLUP and GBLUP. The results demonstrated that PLS variants were more efficient than pedigree-based BLUP but less accurate than GBLUP for 2 out of 5 traits. Furthermore, GBLUP provided a clear biologic interpretation by the use of a genomic relationship matrix that PLS and sPLS may lack, and PLS and sPLS do not provide reliabilities of GEBV, in contrast to GBLUP. Sparse PLS enabled easier identification of relevant variables than PLS, which are possibly associated with QTL regions. Currently, more and more markers are being genotyped, forcing the handling of increasing quantities of data and consequently a critical need exists for methods that perform well in this context. Sparse PLS could be used as a preliminary step in genomic selection to reduce the number of SNP used in the prediction equations provided by other genomic selection methods, such as Bayesian methods. Most importantly, the sPLS algorithm is fast to compute even with a large reference

Table 4. Correlations between observed daughter yield deviations (DYD) and predicted DYD provided by partial least squares regression (PLS), sparse PLS (sPLS), pedigree-based BLUP (BLUP), and genomic BLUP (GBLUP)

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
BLUP	0.38	0.40	0.44	0.44	0.47	0.28
PLS	0.52	0.58	0.55	0.71	0.71	0.25
sPLS	0.50	0.59	0.53	0.72	0.72	0.21
GBLUP	0.56	0.59	0.55	0.72	0.73	0.35

population and a large number of explanatory variables in a one-trait evaluation.

ACKNOWLEDGMENTS

This work was supported by the French project AMASGEN, financed by the French National Research Agency (ANR, Paris France) and ApisGene (Paris, France). Labogena (Jouy-en-Josas, France) is gratefully acknowledged for providing the genotypes. We thank the reviewers for their constructive suggestions that enabled major improvement of the manuscript

REFERENCES

Abdi, H. 2010. Partial least squares regression and projection on latent structure regression (PLS regression). Comput. Stat. 2:97–106.

Boichard, D., and E. Manfredi. 1994. Genetic analysis of conception rate in French Holstein cattle. Acta Agric. Scand. A Anim. Sci. 44:138–145.

Chun, H., and S. Keles. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Series B Stat. Methodol. 72:3–25.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946.

Colombani, C., P. Croiseau, C. Hozé, S. Fritz, F. Guillaume, D. Boichard, A. Legarra, V. Ducrocq, and C. Robert-Granié. 2011. Could genomic selection be efficient to detect QTL? 15th QTLMAS Workshop, Rennes, France. BMC Proceedings, London, UK.

Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. 2010. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet. Sel. Evol. 42:9.

Druet, T., and M. Georges. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. Ann. Stat. 32:407–451.

Fikse, W. F., and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J. Dairy Sci. 84:1759–1767.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101:2398–2403.

Harris, B. L., D. L. Johnson, and R. J. Spelman. 2009. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 325–330 in Proc. 36th ICAR Biennial Session: Identification, Breeding, Production, Health and Recording of Farm Animals. International Committee for Animal Recording (ICAR), Rome, Italy.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443.

Henderson, C. R. 1963. Selection index and expected genetic advance. Statistical genetics and plant breeding. Nat. Acad. Sci. Nat. Res. Counc. Pub. 983:141–163.

Lê Cao, K. A., I. Gonzalez, and S. Dejean. 2009. integrOmics: An R package to unravel relationships between two omics datasets. Bioinformatics 25:2855–2856.

Lê Cao, K. A., D. Rossouw, C. Robert-Granie, and P. Besse. 2008. A sparse PLS for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol. 7:35.

Long, N., D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2011. Dimension reduction and variable selection for genomic selection: Application to predicting milk yield in Holsteins. J. Anim. Breed. Genet. 128:247–257.

Lorber, A., L. E. Wangen, and B. R. Kowalski. 1987. A theoretical foundation for the PLS algorithm. J. Chemometr. 1:19–31.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Mevik, B. H., and H. R. Cederkvist. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J. Chemometr. 18:422–429.

Moser, G., M. Khatkar, B. Hayes, and H. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Mrode, R. A., and G. J. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livest. Prod. Sci. 86:253–260.

Shen, H. P., and J. H. Z. Huang. 2008. Sparse principal component analysis via regularized low rank matrix approximation. J. Multivariate Anal. 99:1015–1034.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:29.

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87:245–251.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal-model information. J. Dairy Sci. 74:2737–2746.

VanSickle, J. 2003. Analysing correlations between stream and watershed attributes. J. Am. Water Resour. Assoc. 39:717–726.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. Pages 391–420 in Multivariate Analysis. P. R. Krishnaiah, ed. Academic Press, New York, NY.

Wold, S., L. Eriksson, J. Trygg, and N. Kettaneh. 2004. The PLS method—Partial least squares projections to latent structures and its application in industrial RDP (research, development, and production). Technical report. Umea University, Sweden.

Wold, S., M. Sjostrom, and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58:109–130.