Genomic selection in the French Lacaune dairy sheep breed

Sandrine Duchemin, Carine Colombani Colombani, Andres Legarra, Guillaume G. Baloche, Helene H. Larroque, Jean-Michel Astruc, Francis F. Barillet, Christèle Robert-Granié, Eduardo Manfredi

J. Dairy Sci. 95:2723–2733
http://dx.doi.org/10.3168/jds.2011-4980

Genomic selection in the French Lacaune dairy sheep breed

S. I. Duchemin,* C. Colombani,* A. Legarra,* G. Baloche,* H. Larroque,* J.-M. Astruc,† F. Barillet,* C. Robert-Granié,* and E. Manfredi*[1]
*Institut National de la Recherche Agronomique (INRA), UR631, Station d'Amélioration Génétique des Animaux (SAGA), BP52627, 31326 Castanet-Tolosan, France
†Institut de l'Elevage, 149 rue de Bercy, 75595 Paris, France

ABSTRACT

Genomic selection aims to increase accuracy and to decrease generation intervals, thus increasing genetic gains in animal breeding. Using real data of the French Lacaune dairy sheep breed, the purpose of this study was to compare the observed accuracies of genomic estimated breeding values using different models (infinitesimal only, markers only, and joint estimation of infinitesimal and marker effects) and methods [BLUP, Bayes Cπ, partial least squares (PLS), and sparse PLS]. The training data set included results of progeny tests of 1,886 rams born from 1998 to 2006, whereas the validation set had results of 681 rams born in 2007 and 2008. The 3 lactation traits studied (milk yield, fat content, and somatic cell scores) had heritabilities varying from 0.14 to 0.41. The inclusion of molecular information, as compared with traditional schemes, increased accuracies of estimated breeding values of young males at birth from 18 up to 25%, according to the trait. Accuracies of genomic methods varied from 0.4 to 0.6, according to the traits, with minor differences among genomic approaches. In Bayes Cπ, the joint estimation of marker and infinitesimal effects had a slightly favorable effect on the accuracies of genomic estimated breeding values, and were especially beneficial for somatic cell counts, the less heritable trait. Inclusion of infinitesimal effects also improved slopes of predictive regression equations. Methods that select markers implicitly (Bayes Cπ and sparse PLS) were advantageous for some models and traits, and are of interest for further quantitative trait loci studies.
Key words: dairy sheep, genomic selection, Bayes Cπ, partial least squares

INTRODUCTION

Genomic selection (**GS**) aims at the improvement of accuracy of genetic indexes for young animals, thus allowing their early selection, a concomitant decrease in generation intervals, and new selection steps with potential favorable effect on selection intensities. Outperforming traditional genetic evaluation of dairy species is one of the main challenges of GS. Traditional evaluation warrants high selection accuracy for males and reliable EBV for traits with varying heritability, via progeny testing. However, progeny testing generates long generation intervals at high testing costs.

In dairy sheep, progeny testing is inserted into a pyramidal management of the population (Barillet, 1997, 2007; Carta et al., 2009) with the breeders of the nucleus flock at the top, benefiting from pedigree and official milk recording, AI, and complementary natural mating. The genetic progress is then transferred to the commercial population, either through AI or natural-mating rams born from AI sires (Barillet, 1997; Carta et al., 2009). The delay observed between the genetic gain in the nucleus and the commercial population in the Lacaune breed is around 5 to 7 yr (Barillet, 1997; Barillet et al., 2001).

In the 1980s, sheep schemes were applied to improve milk composition and milk yield using EBV based on a linear combination of fat and protein yields combined with fat and protein contents (Barillet et al., 2001). By 2006, genetic gains per year in the Lacaune breed were close to 6 L for milk yield and more than 0.1 g/L for fat and protein content. Progeny testing was a key element with more than 400 rams tested per year in about 400 flocks (Barillet, 2007). In recent years, more interest has been given to functional traits, such as milkability, udder traits, reproduction, and genetic disease resistance (mastitis and scrapie resistance), to improve quality of products, animal health and welfare, and to limit costs of the dairy sheep industry and the recording system altogether.

By implementing GS, breeding organizations may create new opportunities: cost reductions of selection schemes with a substantial reduction on the genetic gains gap between the nucleus and the commercial flocks (Schaeffer, 2006), an increase of present rates of genetic progress, the improvement of the monitoring

Received September 27, 2011.
Accepted January 5, 2012.
[1]Corresponding author: Eduardo.Manfredi@toulouse.inra.fr

and control of inbreeding rates, and new possibilities to include new traits as selection criteria (Colleau et al., 2009). The dramatic decrease in the cost of whole-genome genotyping and sequencing of animals has made it possible to scan the entire genome of thousands of animals with high-density markers. Using marker information allows capturing parts of total genetic variance unexplained by traditional genetic models, on hard-to-measure, low-heritable, sex-limited, and postmortem traits (Goddard and Hayes, 2007; Hayes et al., 2009).

Currently, in dairy sheep, molecular information is being used in selection for scrapie resistance (genotypes of the *PrP* gene) and parentage testing (microsatellites and SNP). Quantitative trait loci fine mapping is being conducted after the detection of several QTL affecting functional and productive traits (milk traits, SCC, nematode resistance, FA content in milk fat, and udder traits; Barillet, 2007; Gutiérrez-Gil et al., 2008; Carta et al., 2009).

The recently created International Sheep Genomic Consortium (ISGC; http://www.sheephapmap.org), a partnership among 20 countries, has developed genomic tools, such as the 60K SNP chip, thus bringing new perspectives for GS implementation by sheep breeding organizations. Astruc et al. (2010) points out that the French Lacaune dairy sheep breed has a large training population available to warrant accurate genomic EBV (**GEBV**). Simulation studies show that accuracies of GEBV are influenced by density of markers and reference population size, among other factors (e.g., Meuwissen et al., 2001). Studies in real populations yielded smaller accuracies than those reported in simulation studies (e.g., Legarra et al., 2008, in experimental populations; de Roos, 2011, in commercial populations). To help clarify the effect of GS on real data, the purpose of this study was to compare several statistical models and methods by assessing the best accuracies of prediction of daughter yield deviations (**DYD**) weighted by their effective daughter contribution (**EDC**) in the French Lacaune dairy sheep breed.

MATERIALS AND METHODS

Genotypes

All genotyped rams in the current study were from the Agence Nationale de la Recherche (ANR)-SheepSNPQTL and Fonds Unique Interministériel (FUI)-Roquefort'in projects, both conducted at the UR631 Station d'Amélioration Génétique des Animaux (SAGA), Institut National de la Recherche Agronomique (INRA)-Toulouse, France, in close cooperation with the Lacaune associations of breeders responsible for the management of the breeding schemes. A total of 54,582 SNP were available on 2,812 animals. Controls on markers were performed to check the quality of the SNP and the coherence between the sample identification and the pedigree information (Robert-Granié et al., 2011). The call rate procedure of 97% resulted in the exclusion of 3,106 SNP. Autosomal SNP were checked for Hardy-Weinberg equilibrium, resulting in the exclusion of 2,630 SNP that did not meet the expected allele frequencies. Imposed minor allele frequency requirements of 1% resulted in the additional exclusion of 4,841 SNP. Parent and progeny conflicts led to the exclusion of additional 76 SNP. After these quality controls, 43,929 SNP were available for the statistical analyses. Genotypes for individual SNP were coded as 0, 1, and 2, representing allele counts at each locus and assigned the number 5 when genotypes were missing. Missing values accounted for 0.18% of the total SNP available.

Animals, Phenotypes, and Pedigree Information

During quality controls, 166 animals were excluded by the call rate procedure, together with 79 animals that were excluded due to parent-progeny conflicts. Phenotypes on 2,567 rams were DYD weighted by their EDC and computed from total lactations of ewes (standardized to 165 d of lactation; Barillet et al., 2001; Baloche et al., 2011). Daughter yield deviations are calculated from daughter averages corrected for environmental effects and the merit of their dams (VanRaden et al., 2009). The EDC takes into account the unequal distribution of phenotypes across herds and parities (Fikse and Banos, 2001). Three traits were considered, with varying heritabilities (Barillet, 2007): milk yield ($h^2 = 0.32$), fat content ($h^2 = 0.41$), and SCS ($h^2 = 0.14$). The average numbers of daughters per ram were 70, 76, and 74 for milk yield, fat content, and SCS, respectively, and in accordance with the Lacaune breeding scheme. The pedigree file of 52,152 animals accounted for 10 generations of ancestors.

Rams for this study belong to the AI progeny testing scheme and were born from 1998 to 2008. Data were split into 2 populations: a training population composed by 1,886 rams born between 1998 and 2006, and a validation population, comprising 681 rams born between 2007 and 2008. This sampling approach is well adapted to practical situations, although Amer and Banos (2010) and Olson et al. (2011) warned about the lack of complete independence between training and validation data.

Statistical Models

Three models were used: marker effects only, infinitesimal genetic effects only (here, infinitesimal is

preferred over polygenic, which is sometimes used to represent genome-wide marker effects), and marker and infinitesimal effects estimated jointly. The idea was to assess whether or not a jointly estimated model would perform as well as a model considering marker or infinitesimal effects separately. Scenarios in each method also considered the restriction or not on the total number of markers having a potential effect on phenotypes.

Methods

Infinitesimal BLUP and Genomic BLUP. Breeding values were estimated by both Infinitesimal BLUP (**I-BLUP**; Henderson, 1973) and Genomic BLUP (**G-BLUP**; VanRaden, 2008; Goddard, 2009) methods, with the BLUPF90 software from Misztal et al. (2002), updated in 2010 to account for the genomic relationship matrix between markers and animals. In this software, it is possible to compute breeding values for genotyped and ungenotyped animals, and the relationships among genotyped animals are computed as

$$\mathbf{G}_w = w\mathbf{G} + (1 - w)\ \mathbf{A}_{22}, \qquad [1]$$

where $\mathbf{G}_w$ is a matrix of relationships combining pedigree and marker data, w is a weighting coefficient, and $\mathbf{A}_{22}$ is the matrix of pedigree relationships among genotyped animals, as obtained from the whole pedigree (Van Raden, 2008; Aguilar et al., 2011). The genomic matrix ($\mathbf{G}$) was $\mathbf{G} = \mathbf{MM}'/f$, where $\mathbf{M}$ is the incidence matrix of marker effects, corrected by the expected genotype frequencies, and $f = 2\sum_k p_k\left(1 - p_k\right)$ is a function of allele frequencies at the SNP loci, where p_k is the allele frequency of the kth SNP (Van Raden, 2008; Aguilar et al., 2011).

In I-BLUP predictions, the $\mathbf{G}$ matrix in equation [1] had a weight w of 1%, to reproduce the relationship matrix $\mathbf{A}$ based on the pedigree information. In G-BLUP predictions, the weight for the $\mathbf{G}$ matrix in equation [1] was set to 99% to create the genomic relationship matrix based on the marker information. Missing values are taken into account by the software which sets these missing marker genotypes to the average of the population, thus not impairing genomic predictions.

Bayes Cπ. Habier et al. (2011) and Sun et al. (2011) proposed the Bayes Cπ method, where the phenotypes $\mathbf{y}$ are a function of marker effects $\boldsymbol{\tau}$: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zd}\boldsymbol{\tau} + \mathbf{e}$, where μ is an overall mean, $\mathbf{Z}$ is an incidence matrix, $\mathbf{d}$ is a vector of indicator variables for implicit selection of markers, and $\mathbf{e}$ is a vector of uncorrelated residuals normally distributed, whose variance is inversely proportional to the EDC of each DYD. The distribution for $\mathbf{d}$ is a multivariate Bernoulli such that $\Pr(d_i = 1|\pi) = \pi$, where d_i is the indicator variable of the ith SNP, π is the probability that the marker has an effect on the phenotype. The prior distribution of marker effects was normal with a common variance following an inverted chi-square distribution. Therefore, the marker variance $\left(\sigma^2_\tau\right)$ is common to all loci in contrast to the locus-specific variance assumed in the Bayes B approach by Meuwissen et al. (2001).

Genomic EBV of genotyped rams were obtained through GS3 software developed by Legarra et al. (2010). Posterior distribution of variances were computed using a full Monte Carlo Markov chain of 100,000 iterations, with a burn-in of 20,000 iterations. When π was estimated by the algorithm, its prior distribution was considered to be uniform.

Four different scenarios were considered in Bayes Cπ: 1) a model with marker effects only and estimated π (**NOPEDPIFREE**); 2) a model with marker effects only and π = 10%, meaning that 10% of the total SNP would effectively explain all genetic variance (**NOPEDPI10%**); 3) a model with marker and infinitesimal effects and estimated π (**PEDPIFREE**); and 4) a model with marker and infinitesimal effects and π = 10% (**PEDPI10%**).

Partial Least Squares and Sparse Partial Least Squares. These methods are especially useful when the number of independent variables $\mathbf{X}$ (in our context, the incidence matrix for SNP) is larger than the number of observations and there is multicollinearity among $\mathbf{X}$. Partial least squares (**PLS**) regression (Wold, 1966) combines features from and generalizes principal component analysis (**PCA**) and multiple linear regressions. Its goal is to predict a dependent variable from a set of independent variables or predictors ($\mathbf{X}$), which is achieved by extracting from the predictors a set of orthogonal factors called latent variables (Solberg et al., 2009; Colombani et al., 2010). To maximize the covariance between phenotypes and genotypes, several successive regressions are performed by projections onto latent variables to highlight biological effects. Sparse PLS (**sPLS**), developed by Lê Cao et al. (2009), performs simultaneous variable selection between genotypes by introducing a lasso penalization on the pair of PLS loading vectors, which is the total number of SNP retained per latent variable (or per dimension). The mixOmics (previously called IntegrOmics) package for R, developed by Lê Cao et al. (2009), was used to obtain PLS and sPLS predicted phenotypes for genotyped rams. Missing values in the training data are normally taken into account by mixOmics. However, the subroutine for prediction of lamb phenotypes required complete genotypes. So, each individual SNP missing

Table 1. Correlations (ρ) between observed daughter yield deviations (DYD) and predicted DYD computed in the validation population of 681 rams using methods infinitesimal BLUP (I-BLUP), genomic BLUP (G-BLUP), Bayes Cπ, partial least squares (PLS), and sparse PLS (sPLS) for milk yield (MY), fat content (FC), and SCS

	Method[1]										
	I-BLUP	G-BLUP	Bayes Cπ				PLS		sPLS		
Trait	ρ	ρ	NOPEDPIFREE	NOPEDPI10%	PEDPIFREE	PEDPI10%	ρ	Dim	ρ	Dim	N_{SNP}
MY	0.37	0.42	0.43	0.44	0.44	0.44	0.41	7	0.42	5	10,201
FC	0.46	0.56	0.57	0.57	0.57	0.57	0.56	12	0.56	11	35,014
SCS	0.39	0.44	0.46	0.46	0.47	0.46	0.43	6	0.43	6	28,954

[1]NOPEDPIFREE = marker effects only and estimated π; NOPEDPI10% = marker effects only and π = 10%, meaning that 10% of the total SNP would effectively explain all genetic variance; PEDPIFREE = marker and infinitesimal effects and estimated π; PEDPI10% = marker and infinitesimal effects and π = 10%; Dim = dimensions retained in the final model; N_{SNP} = number of SNP selected in the final model.

value was replaced by the average value of the SNP codes. The imputation method did not have a strong effect on the predictions, as missing values in our data represented only 0.18% of the total SNP available (results not shown from a test on a subset of complete data, where some data were voluntarily deleted).

Criteria for Method Comparison

The predictive ability of methods, as suggested by Mäntysaari et al. (2010), was assessed in the validation population by 1) EDC weighted correlation between observed DYD and predicted DYD and 2) EDC weighted regression slopes of observed DYD on predicted DYD.

Marker Contributions

Assessment of marker effects for Bayes Cπ was done by expressing the effects of the markers in genetic standard deviation units for each trait. For sPLS, it was done through variable importance in projection (**VIP**), which allows classification of SNP variables according to their relative importance in predicting the phenotypes (Lê Cao et al., 2008). All VIP higher than 1 are considered as significant (Tenenhaus, 1998).

RESULTS AND DISCUSSION

Weighted correlations and regression slopes for all methods are shown in Tables 1 and 2, respectively.

Comparison Between Genomic Methods and Conventional Pedigree Method

All genomic methods had better predictive ability than I-BLUP (Table 1). Recall that we are computing EBV of lambs at birth, when only ancestor and collateral phenotypes are available for I-BLUP. In this situation, it is expected that methods like G-BLUP improve accuracy because they also use ancestor and collateral

Table 2. Regression slopes of observed on predicted daughter yield deviations(DYD) computed in the validation population of 681 rams using methods infinitesimal BLUP (I-BLUP), genomic BLUP (G-BLUP), Bayes Cπ, partial least squares (PLS), and sparse PLS (sPLS) for milk yield (MY), fat content (FC), and SCS

	Trait[2]								
	MY			FC			SCS		
Method[1]	b	SE	Interval	b	SE	Interval	b	SE	Interval
I-BLUP	0.93	0.09	[0.75; 1.11]	0.88	0.07	[0.75; 1.01]	0.93	0.08	[0.76; 1.09]
G-BLUP	0.85	0.07	[0.71; 0.99]	0.86	0.05	[0.76; 0.96]	0.85	0.07	[0.72; 0.99]
Bayes Cπ									
NOPEDPIFREE	0.93	0.07	[0.78; 1.08]	0.90	0.05	[0.80; 1.00]	0.88	0.06	[0.76; 1.01]
NOPEDPI10%	0.94	0.07	[0.80; 1.09]	0.89	0.05	[0.79; 0.99]	0.88	0.07	[0.75; 1.01]
PEDPIFREE	0.99	0.08	[0.83; 1.14]	0.93	0.05	[0.83; 1.03]	0.94	0.07	[0.80; 1.07]
PEDPI10%	1.00	0.08	[0.84; 1.16]	0.92	0.05	[0.82; 1.02]	0.92	0.07	[0.79; 1.06]
PLS	0.90	0.08	[0.75; 1.06]	0.84	0.05	[0.75; 0.94]	0.83	0.07	[0.70; 0.97]
sPLS	0.91	0.08	[0.76; 1.06]	0.81	0.05	[0.72; 0.90]	0.82	0.07	[0.68; 0.95]

[1]NOPEDPIFREE = marker effects only and estimated π; NOPEDPI10% = marker effects only and π = 10%, meaning that 10% of the total SNP would effectively explain all genetic variance; PEDPIFREE = marker and infinitesimal effects and estimated π; PEDPI10% = marker and infinitesimal effects and π = 10%.

[2]b = regression slope; interval = b ± 2 SE.

information, plus the additional SNP information used to compute the actual proportion of DNA shared by animals instead of the expected average relationship used in I-BLUP.

Interestingly, methods that do not use pedigree information, like PLS, sPLS, and NOPED implementations of Bayes Cπ, were also better than I-BLUP. Partial least squares and sPLS methods outperformed I-BLUP in our study and these results are similar to those in French dairy cattle studies (Colombani et al., 2010). However, superiorities of PLS and sPLS over the infinitesimal approach reported here for sheep are higher (up to +21.7% for fat content in Table 1) than those found in French dairy cattle.

Superiority of genomic methods over I-BLUP depended on traits. For instance, G-BLUP outperformed I-BLUP, with correlations being 15% higher for milk yield, 21% higher for fat content, and 12% higher for SCS. For all traits, the extra accuracy provided by genomic data was not observed by chance, as shown by a significant ($P < 0.01$) reduction of mean square errors when G-BLUP was added as a second covariate to the reference I-BLUP prediction model. These improvements are consistent with validation results of GS in cattle breeds (Hayes et al., 2009; Croiseau et al., 2010; Fritz et al., 2010). In these applications, there is not a clear relative advantage of genomic indexes for low-heritability traits. However, the extra accuracy provided by genomic methods allows selecting young animals for traits with low heritability.

Comparison Among Genomic Methods

In terms of correlations, G-BLUP, PLS, and sPLS methods did not show important differences. This is a rather interesting result, as G-BLUP considers total marker effects as sums of individual SNP effects, whereas in contrast, PLS finds linear combinations among SNP, thus building multiple SNP latent variables whose correlation with phenotypes is maximized. It is important to mention that latent variables incorporate information regarding the similarities and dissimilarities between individuals or original variables, although PLS and sPLS methods do not consider the distribution of marker variances explicitly.

The maximum correlation between phenotypes and genotypes was rapidly reached by PLS and sPLS methods, with few latent variables built (7, 12, and 6 dimensions for milk yield, fat content, and SCS, respectively; Figure 1). The low number of dimensions retained illustrates the good ability of both methods to capture the variability included in the genotypic information. The fact that correlations for G-BLUP, PLS, and sPLS methods are not far apart highlights the robustness of PLS approaches.

Bayes Cπ (all scenarios considered), as compared with G-BLUP, showed, on average, gains in correlations of 3.3% for milk yield, 2.3% for fat content, and 5.2% for SCS. Scenarios including infinitesimal and marker effects allowed estimation of infinitesimal and marker variances, whereas our G-BLUP application is solved with a fixed weight w of 99% to compute the relationship matrix using equation [1]. These results suggest that G-BLUP might be improved by the use of intermediate weights in equation [1]. Comparisons between Bayes Cπ, PLS, and sPLS methods gave similar results for fat content, whereas for milk yield and SCS, PLS and sPLS had a slightly lower performance.

Overall, genomic methods yielded correlations comparable to previous results, which depended on the amount of data available. Hayes et al. (2009) found a correlation of 0.60 using 4,369 significant SNP on protein content in dairy cattle milk and the Bayes A method with a validation population of 637 animals. This result is consistent with the correlation of 0.57 found for fat content in the current study, with 681 rams used as validation population, and with marker effects estimated by Bayes Cπ NOPEDPI10%, which limits the number of markers having effects to 10% of total SNP (4,392 SNP; Table 1).

Inclusion of Infinitesimal Effects

Minor changes were perceived across Bayes Cπ scenarios (i.e., models that jointly estimated marker and infinitesimal effects did not perform clearly better than models that only considered marker effects, in terms of correlations; Table 1). Including infinitesimal effects into a marker model improved slightly the correlations in low-heritable traits such as SCS: gains in correlations of PEDPIFREE over NOPEDPIFREE were 1% for milk yield and 1.4% for SCS. A similar advantage was observed in scenarios with π fixed at 10%, where gains in correlations for PEDPI10% reached 1.1% for SCS over NOPEDPI10%. Results suggest that for low-heritable traits such as SCS, the inclusion of pedigree and marker information might bring additional contribution on predictions, which are neglected when considering only 1 source of information.

Olson et al. (2011) report results for milk yield and SCS on Holstein cattle using GEBV that combine parent averages with individual genomic indexes in a context of national genetic evaluations. The number of data in their study was substantially higher than ours (8,022 Holstein bulls vs. 1,896 rams, for training; 2,653 Holsteins bulls vs. 681 rams, for validation) and

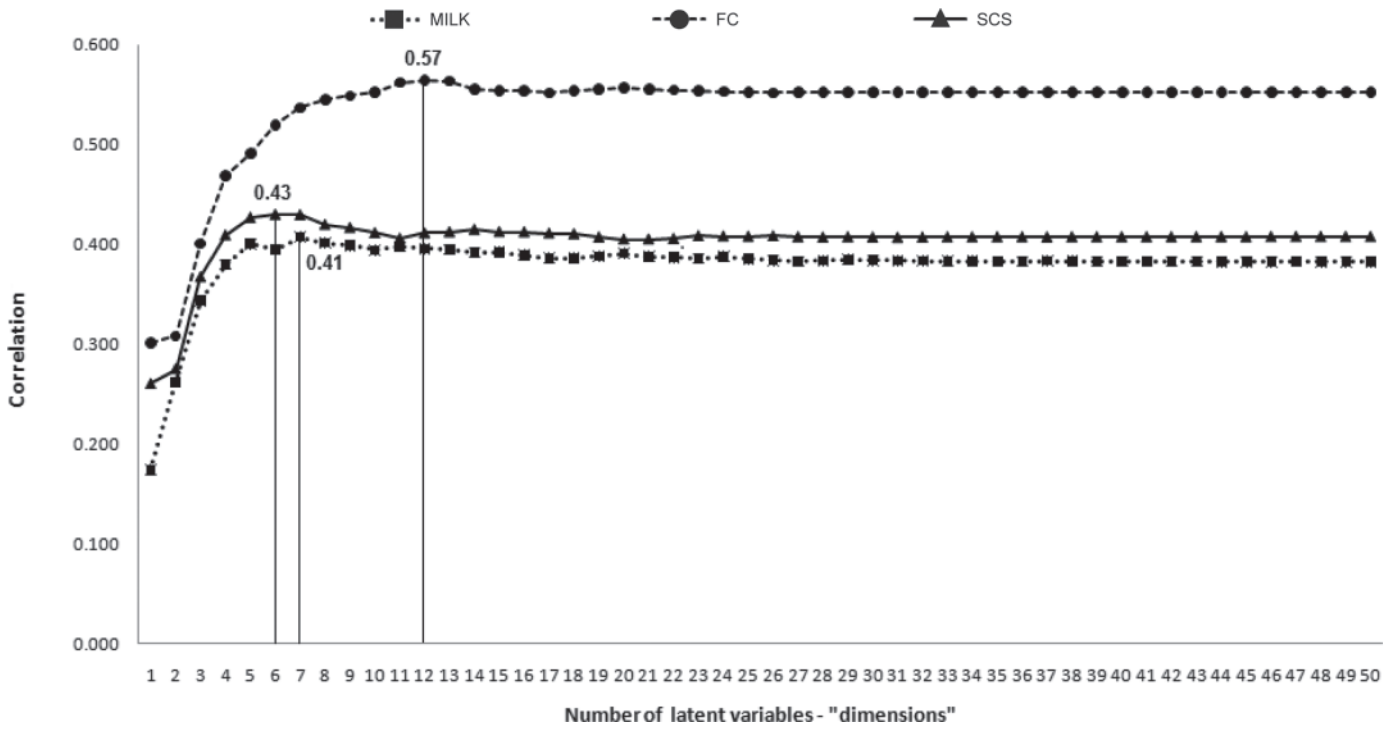

Figure 1. Changes in correlations yielded by the partial least squares (PLS) method according to the number of latent variables considered in the final model. Values of correlations are represented as squares (milk yield), triangles (SCS), and dots (fat content, FC). Correlations between observed daughter yield deviations (DYD) and predicted DYD were weighted by effective daughter contributions (EDC).

the model they used could be roughly compared with our Bayes Cπ PEDPIFREE model. Thus differences in accuracies were expected and ours were lower for milk yield (0.44 vs. 0.63) and for SCS (0.47 vs. 0.53). Inclusion of parent average in further estimations of GEBV for Lacaune dairy sheep might be a good option and might improve our results for statistical approaches not benefiting from pedigree information, such as PLS or sPLS.

Estimates of Marker and Infinitesimal Variances

The 4 Bayes Cπ scenarios provided posterior distributions for the individual marker variance $\left(\sigma_\tau^2\right)$ and for the proportion π of markers with effects on phenotypes, which allowed us to approximate the genetic variance due to markers as $2\pi\sum_k p_k\left(1-p_k\right)\sigma_\tau^2$, following Gianola et al. (2009) and assuming linkage equilibrium among marker genotypes. Variances of marker and infinitesimal effects estimated in the 4 Bayes Cπ scenarios are presented in Figure 2, where all estimated variances are expressed as percentage of the marker variance estimated in the NOPEDPIFREE scenario.

Models including the infinitesimal effects (Bayes Cπ PEDPIFREE and PEDPI10% scenarios) yielded higher variances of genetic origin than those estimated in simple marker models (+11 to +17% according to traits; Figure 2). The estimation of π had no effect on the results because the Bayes Cπ algorithm was able to compute total marker variance but it could not disentangle π and the variance of individual markers. Inspection of results of samples generated by the Bayes Cπ MCMC procedure revealed a negative correlation between π and the individual marker variances: when the generated π is high, the individual marker variance is low, and vice versa, such that the estimate of total marker variance remained stable in different samples. For instance, posterior means (and their coefficients of variation) in the NOPEDPIFREE model for fat content were 0.20 (106%) for π, 0.01 (64%) for σ_τ^2, and, much less variable, 14.86 (8%) for the total marker variance.

When models included the infinitesimal effect, the variance due to markers decreased by about 30%. It is unclear whether this is an advantage (control of spurious marker effects) or a disadvantage (redundancy between infinitesimal and marker effects) of these models. In any case, the effect in prediction accuracy was small.

In fact, the higher variances of genetic origin in the models including infinitesimal effects had a favorable effect on the correlations between observed DYD and predicted DYD in the training population (i.e., when infinitesimal effects were included, correlations for all

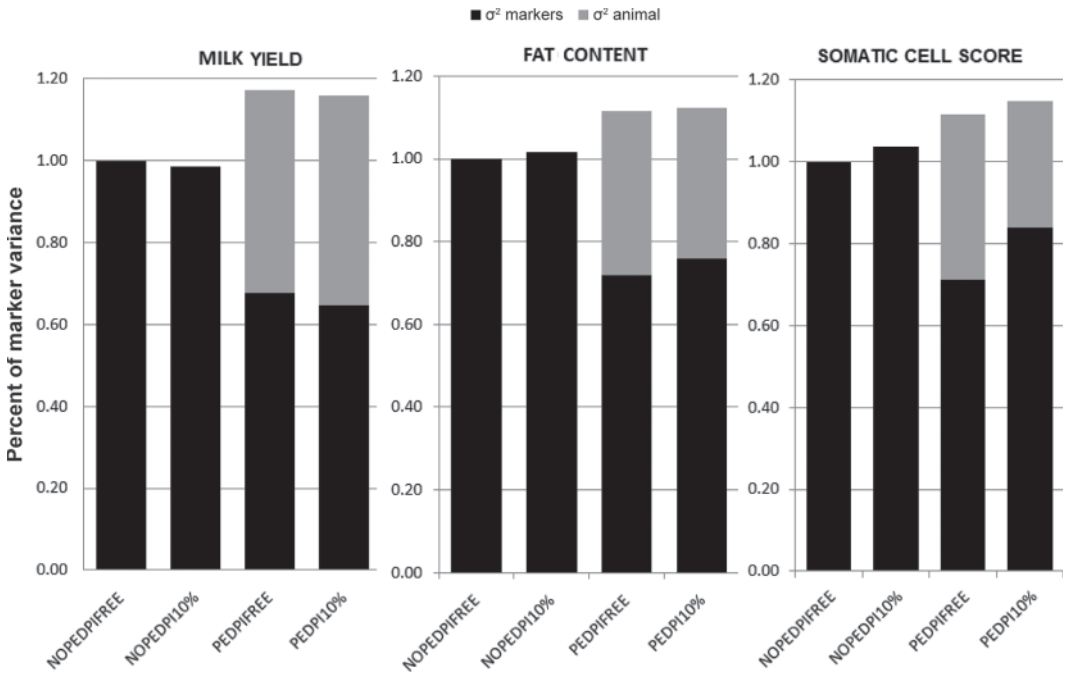

Figure 2. Variances of marker and infinitesimal effects in the 4 scenarios of Bayes Cπ expressed as the percentage of the marker variances estimated in the scenario with marker effects only and estimated π (NOPEDPIFREE). NOPEDPI10% = marker effects only and π = 10%, meaning that 10% of the total SNP would effectively explain all genetic variance; PEDPIFREE = marker and infinitesimal effects and estimated π; PEDPI10% = marker and infinitesimal effects and π = 10%.

traits were higher than 0.99) and always higher than correlations obtained via markers-only models. In the validation population, inclusion of infinitesimal effects improved the accuracy of GEBV for SCS and the slope for all traits (Tables 1 and 2).

Marker Selection

The methods Bayes Cπ and sPLS select markers during the estimation process. Marker selection could be helpful to explain, specifically for each trait, parts of the genetic variances not explained by infinitesimal models. In the current study, pre-selection of markers did not show a clear advantage (Tables 1 and 2), suggesting that the advantage of such approach may be trait dependent.

Partial least squares and sPLS performed very similarly across traits, although in sPLS, a restriction exists on the total number of SNP retained per dimension. Looking at sPLS results and keeping 5, 13, and 15% of all markers for milk yield, fat content, and SCS, respectively, to represent the total genetic variation explained by markers, revealed results that were very close to those of G-BLUP.

Sparse PLS analysis can be compared with Bayes Cπ NOPEDPI10% and PEDPI10%, where π is fixed at 10%. For milk yield, whereas Bayes Cπ NOPEDPI10% yielded a correlation of 0.44 with 4,393 SNP, sPLS correlations reached 0.42 with 10,201 SNP (results shown in Table 1). For fat content and SCS, sPLS results yielded similar correlations to other methods but with a higher number of SNP retained per dimension in the final model. In this context, only a larger group of SNP, all with little effects, could explain part of the total genetic variance.

Marker Contributions

Marker effects along the genome are illustrated in Figure 3 (Bayes Cπ results; absolute values of effects in units of genetic standard deviation of each trait) and Figure 4 (sPLS results expressed as VIP). In Bayes Cπ, the 4 scenarios studied (π fixed or not and inclusion or not of infinitesimal effects; Figure 3) pointed to the same chromosome regions affecting each trait, but dispersion patterns of estimates of marker effects changed according to the model. This is clearly illustrated in the SCS analyses (plots in the right of Figure 3), where several chromosome regions have estimated effects over 0.03 genetic standard deviations in the PIFREE scenarios, whereas only 1 region reaches that threshold in both PI10% scenarios. Including infinitesimal effects decreased further the dispersion of marker effects (PEDPIFREE vs. NOPEDPIFREE and PEDPI10% vs.

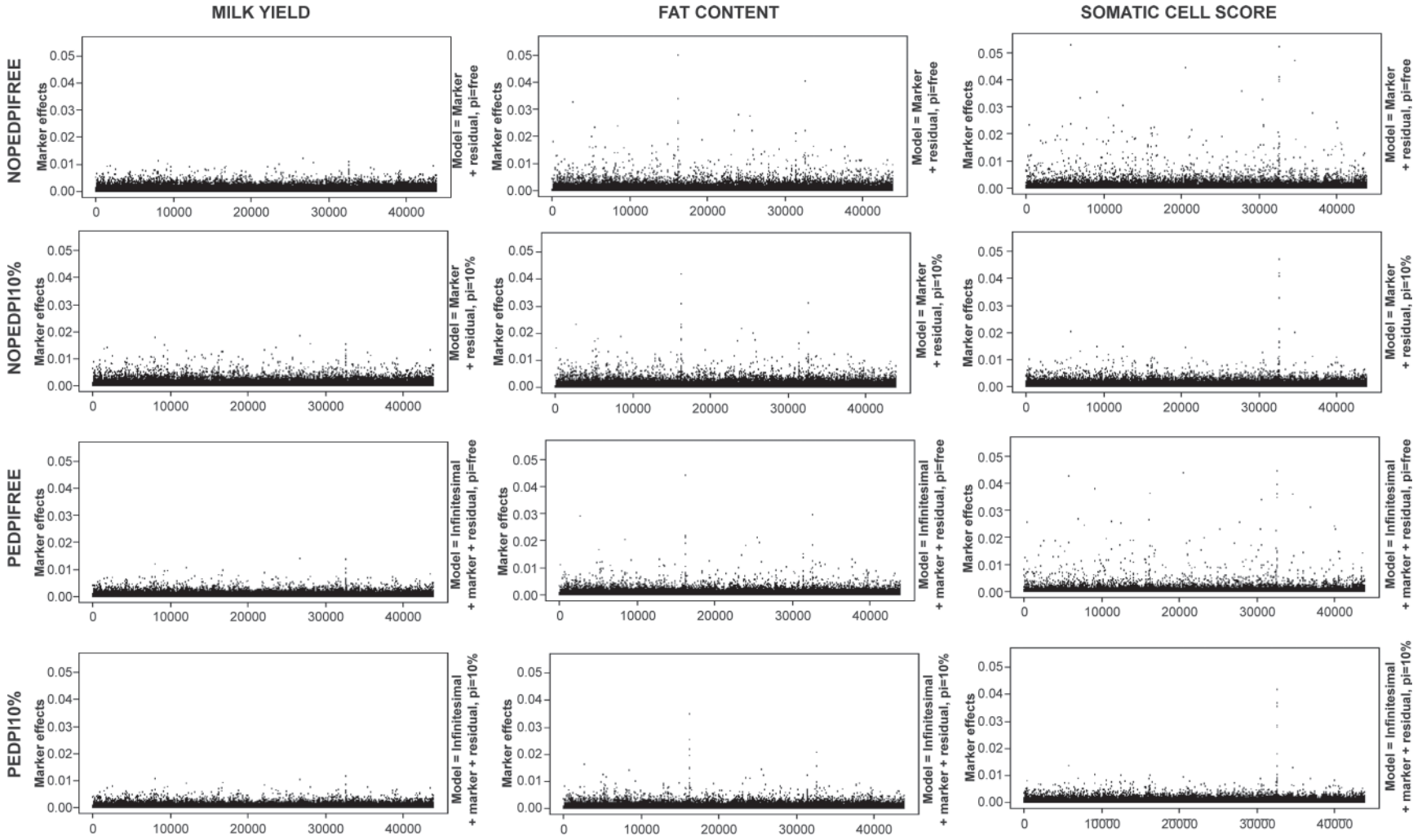

Figure 3. Absolute marker effects expressed as genetic standard deviations across all scenarios of Bayes Cπ for milk yield (on the left side), fat content (in the middle), and SCS (on the right side). NOPEDPIFREE = marker effects only and estimated π; NOPEDPI10% = marker effects only and π = 10%, meaning that 10% of the total SNP would effectively explain all genetic variance; PEDPIFREE = marker and infinitesimal effects and estimated π; PEDPI10% = marker and infinitesimal effects and π = 10%.

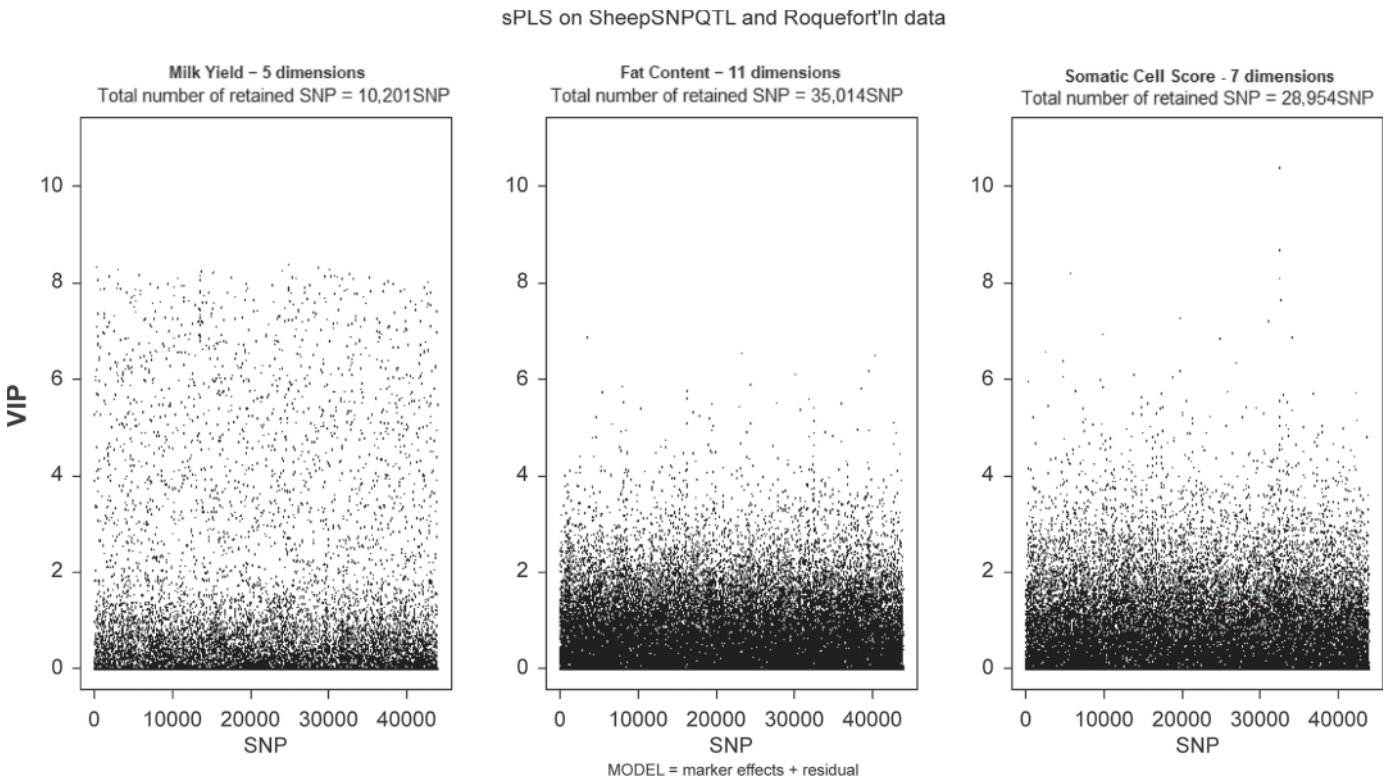

Figure 4. Variable importance in projection (VIP) coefficients yielded by the sparse partial least squares (sPLS) method for milk yield (on the left side), fat content (in the middle), and SCS (on the right side).

NOPEDPI10% scenarios for SCS in Figure 3). For the other traits, the choice of the model was less critical: for fat content, the effects in PIFREE were more variable than those yielded by the PI10% scenario, and for milk yield differences among scenarios were less visible.

Variable importance in projection coefficients using the sPLS method for milk yield show that by selecting 23% of the total SNP (10,201 pre-selected SNP) most SNP with important effects on the trait have been captured. These results also show that the group of SNP with strong effects on milk yield (VIP >2; Figure 4) contributes with similar weights and its distribution is homogeneous along the genome.

In contrast with milk yield results, the sPLS method retained almost all SNP in the analysis for fat content. Here, VIP coefficients show that the distribution of SNP effects along the genome is heterogeneous, clearly suggesting that small genes contribute by having effects on fat content.

In Figure 4, VIP coefficients for SCS show that 1) 1 SNP has a very strong effect (VIP >10), 2) some SNP have strong effects (VIP >2), and 3) most SNP have very small effects (VIP <2). The SNP effects are heterogeneously spread along the genome, the group of SNP with stronger effects all contributing with different weights and many small genes contributing to SCS phenotypes. This may explain why many SNP per dimension were retained in sPLS for SCS.

Although the scale of the graphs for each method is different and they are not directly comparable, in Bayes Cπ and in sPLS very similar results on marker effect plots can be seen (Figure 3 vs. Figure 4). The markers that showed strong effects were retrieved and will be validated in the protocol of QTL detection of SCS. This preliminary result, where effects of chromosome regions are estimated simultaneously, may help to decrease the number of candidate genomic regions to be studied in QTL detection.

Regression Slopes

Regression slopes are used to validate genomic evaluation by comparison between the observed and expected regression coefficients (Mäntysaari et al., 2010). An expected coefficient of 1 should be expected if individuals in the validation data are unselected. In our study, the validation data included 84% of all rams entering the AI centers before progeny testing, with EBV averages

for all traits not different from those of ungenotyped individuals. This was reflected in slopes around 1 yielded by some of the methods compared here (Table 2). In cattle studies, Olson et al. (2011) reported expected and realized slopes lower than those in Table 2 when genotyped bulls of the validation set were selected. In Table 2, the reference method I-BLUP yielded slopes close to 1, especially for milk yield and SCS. Within the genomic methods, upper bounds of confidence intervals were always less than 1 for G-BLUP and always larger than 1 for the Bayes Cπ method when infinitesimal and marker effects estimated jointly (PEDPIFREE and PEDPI10%).

CONCLUSIONS

This first study on genomic selection in the Lacaune dairy sheep shows that molecular markers can be effectively used to improve current selection methods. Accuracies of GEBV for males at birth can be improved from +18 to +25% according to traits. These results were obtained with a reference population of about 2,500 proven rams and about 44,000 SNP. Accuracies in future implementations should be higher due to an increase of the size of the reference population and the inclusion of all the historical information used in the present routine of genetic evaluation. Enough selection accuracy would lead to an early selection of males, with a concomitant reduction of generation intervals. Expected additional genetic gain and economic advantages in sheep will be lower than in dairy cattle due to smaller generation intervals and lower maintenance costs of dairy rams. Implementation of Bayes Cπ (all scenarios considered) yielded maximum accuracies of 0.44 for milk yield, 0.57 for fat content, and 0.46 for somatic cells. The other methods yielded comparable accuracies. Nonetheless, the Bayes Cπ method remains costly in computing time among the methods considered. The PLS and sPLS methods show robustness in genomic EBV, as maximization of covariances between phenotypes and markers could be reached with few latent variables. Inclusion of infinitesimal effects in the prediction model had little effect on accuracies and it was trait dependent, with favorable results for the computation of GEBV for SCS. Also, inclusion of infinitesimal effects led to better slopes of regressions of observed DYD on predicted DYD. Implicit selection of markers had little effect on accuracies and its advantages depended on the method used to choose the markers. Comparison of regions detected with the whole-genome approaches used in the present study and those found by QTL detection approaches will continue in the ongoing SheepSNPQTL and Roquefort'in projects.

ACKNOWLEDGMENTS

This work benefitted from financial support of the Agence Nationale de la Recherche (ANR)-SheepSNPQTL, Apis Gene, and Fonds Unique Interministériel (FUI)-Roquefort'in projects. We thank the 5 breeder partners of Roquefort'in project, the genotyping platform LABOGENA (http://www.labogena.fr; Jouy-en-Josas, France), the bioinformatics support of SIGENAE (http://www.sigenae.org; Toulouse, France), the computing facilities of the Centre de Traitement de l'Information Génétique (CTIG; Jouy-en-Josas, France) and the bioinformatics platform Genotoul (http://bioinfo.genotoul.fr; Toulouse, France). The first author benefitted from academic and financial support of the Erasmus Mundus program, the European Master in Animal Breeding and Genetics (Wageningen, the Netherlands), the Koepon Foundation (Leusden, the Netherlands), and the Institut National de la Recherche Agronomique (INRA; Castanet-Tolosan, France).

REFERENCES

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

Amer, P. R., and G. Banos. 2010. Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. J. Dairy Sci. 93:3320–3330.

Astruc, J. M., G. Lagriffoul, H. Larroque, A. Legarra, C. Moreno, R. Rupp, and F. Barillet. 2010. Use of genomic data in French dairy sheep breeding programs: Results and prospects. In Proc. 37th International Committee for Animal Recording (ICAR) Annual Meeting, Riga, Latvia. ICAR, Rome, Italy.

Baloche, G., H. Larroque, J. M. Astruc, J. M. Babilliot, M. Y. Boscher, P. Boulenc, C. Chantry-Darmon, C. Boissieu, G. Frégeat, B. Giral-Viala, P. Guibert, G. Lagriffoul, C. Moreno, P. Panis, C. Robert-Granié, G. Sallé, A. Legarra, and F. Barillet. 2011. Work in progress on genomic evaluation using BGLUP in French Lacaune dairy sheep breed. Book of Abstracts of the 62nd Annual Meeting of the European Association for Animal Production (EAAP), Stavanger, Norway. Wageningen Academic Publishers, Wageningen, the Netherlands.

Barillet, F. 1997. Genetics of milk production. Pages 539–564 in The Genetics of Sheep. I. Piper and A. Ruvinsky, ed. CAB International, Wallingford, Oxfordshire, UK.

Barillet, F. 2007. Genetic improvement for dairy production in sheep and goats. Small Rumin. Res. 70:60–75.

Barillet, F., C. Marie, M. Jacquin, G. Lagriffoul, and J. M. Astruc. 2001. The French Lacaune dairy sheep breed: Use in France and abroad in the last 40 years. Livest. Prod. Sci. 71:17–29.

Carta, A., S. Casu, and S. Salaris. 2009. Invited review: Current state of genetic improvement in dairy sheep. J. Dairy Sci. 92:5814–5833.

Colleau, J. J., S. Fritz, F. Guillaume, A. Baur, D. Dupassieux, M. Y. Boscher, L. Journaux, A. Eggen, and D. Boichard. 2009. Simulating the potential of genomic selection in dairy cattle breeding. Rencontres Recherche Ruminants 16:419.

Colombani, C., A. Legarra, P. Croiseau, F. Guillaume, S. Fritz, V. Ducrocq, and C. Robert-Granié. 2010. Application of PLS and sparse PLS regression in genomic selection. In 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. German Society for Animal Science, Gießen, Germany.

Croiseau, P., C. Colombani, A. Legarra, F. Guillaume, S. Fritz, A. Baur, R. Dassonneville, C. Patry, C. Robert-Granié, and V. Ducrocq. 2010. Improving genomic evaluation strategies in dairy cattle through SNP pre-selection. In 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany. German Society for Animal Science, Gießen, Germany.

de Roos, A. P. W. 2011. Genomic selection in dairy cattle. PhD Thesis. Wageningen Univ., Wageningen, the Netherlands.

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

Fritz, S., F. Guillaume, P. Croiseau, A. Baur, C. Hoze, R. Dassonneville, M. Y. Boscher, L. Journaux, D. Boichard, and V. Ducrocq. 2010. Mise en place de la sélection génomique dans les trois principales races françaises de bovins laitiers. Rencontres Recherche Ruminants 17:455–458.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximization of long term response. Genetica 136:245–257.

Goddard, M. E., and B. J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323–330.

Gutiérrez-Gil, B., M. F. El-Zarei, L. Alvarez, Y. Bayón, L. de la Fuente, F. San Primitivo, and J. J. Arranz. 2008. Quantitative trait loci underlying udder morphology traits in dairy sheep. J. Dairy Sci. 91:3672–3681.

Habier, D., R. Fernando, K. Kizilkaya, and D. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

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. 1973. Sire evaluation and genetic trends. Pages 10–41 in Proc. Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush. Am. Soc. Anim. Sci. and Am. Dairy Sci. Assoc., Champaign, IL.

Lê Cao, K.-A., I. Gonzales, and S. Déjean. 2009. IntegrOmics: An R package to unravel relationships between two omics data sets. Bioinformatics 25:2855–2856.

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

Legarra, A., A. Ricard, and O. Filangi. 2010. GS3-Genomic selection, Gibbs Sampling, Gauss Seidel and Bayes Cπ. Accessed Feb. 10, 2011. http://snp.toulouse.inra.fr/~alegarra.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618.

Mäntysaari, E., Z. Liu, and P. VanRaden. 2010. Interbull validation test for genomic evaluations. Interbull Bull. 41:17–21.

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.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). Proc. 7th WCGALP, Montpellier, France. CD-ROM communication 28:07.

Olson, K. M., P. M. VanRaden, M. E. Tooker, and T. A. Cooper. 2011. Differences among methods to validate genomic evaluations for dairy cattle. J. Dairy Sci. 94:2613–2620.

Robert-Granié, C., S. Duchemin, H. Larroque, G. Baloche, J. M. Astruc, F. Barillet, C. Moreno, A. Legarra, and E. Manfredi. 2011. A comparison of various methods for the computation of genomic breeding values in French Lacaune dairy sheep breed. Book of Abstracts of the 62nd Annual Meeting of the European Federation of Animal Science (EAAP), Stavanger, Norway. Wageningen Academic Publishers, Wageningen, the Netherlands.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223.

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.

Sun, X., D. Habier, R. L. Fernando, J. D. Garrick, and J. C. M. Dekkers. 2011. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. BMC Proc. 5:S13.

Tenenhaus, M. 1998. La régression PLS: Théorie et pratique. Editions Technip, Paris, France.

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. K. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

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.