Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm

PASCAL CROISEAU1, ANDRÉS LEGARRA2, FRANÇOIS GUILLAUME3, SÉBASTIEN FRITZ4, AURÉLIA BAUR4, CARINE COLOMBANI2, CHRISTÉLE ROBERT-GRANIÉ2, DIDIER BOICHARD1 AND VINCENT DUCROCQ1

1 INRA, UMR1313 – Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France
2 INRA, UR 631, Station d’Amélioration Génétique des Animaux, F-31320 Castanet-Tolosan, France
3 Institut de l’élevage, 149 rue de Bercy, 75595 Paris, France
4 UNCEIA, 149 rue de Bercy, 75595 Paris, France

(Received 21 December 2010; revised 27 October 2011; accepted 27 October 2011)

Summary
For genomic selection methods, the statistical challenge is to estimate the effect of each of the available single-nucleotide polymorphism (SNP). In a context where the number of SNPs ($p$) is much higher than the number of bulls ($n$), this task may lead to a poor estimation of these SNP effects if, as for genomic BLUP (gBLUP), all SNPs have a non-null effect. An alternative is to use approaches that have been developed specifically to solve the ‘$p >> n$’ problem. This is the case of variable selection methods and among them, we focus on the Elastic-Net (EN) algorithm that is a penalized regression approach. Performances of EN, gBLUP and pedigree-based BLUP were compared with data from three French dairy cattle breeds, giving very encouraging results for EN. We tried to push further the idea of improving SNP effect estimates by considering fewer of them. This variable selection strategy was considered both in the case of gBLUP and EN by adding an SNP pre-selection step based on quantitative trait locus (QTL) detection. Similar results were observed with or without a pre-selection step, in terms of correlations between direct genomic value (DGV) and observed daughter yield deviation in a validation data set. However, when applied to the EN algorithm, this strategy led to a substantial reduction of the number of SNPs included in the prediction equation. In a context where the number of genotyped animals and the number of SNPs gets larger and larger, SNP pre-selection strongly alleviates computing requirements and ensures that national evaluations can be completed within a reasonable time frame.

1. Introduction
The availability of dense single-nucleotide polymorphism (SNP) arrays has considerably changed the landscape of dairy cattle selection worldwide. With such chips, it is now possible to retrieve information about quantitative trait locus (QTL) all over the genome. Genomic estimated breeding values (GEBV), which correspond to a combination of the sum of the effects of genetic markers (direct genomic value (DGV)) and estimated breeding value (EBV), can be used instead of the classical pedigree-based evaluations in selection programmes. Meuwissen et al. (2001) envisioned the consequences on the estimation of breeding values of a high-density marker map covering the whole genome (see also Haley & Visscher, 1998; Andersson & Georges, 2004). Through simulations, they showed that the use of GEBV can greatly improve accuracy of genetic evaluation of animals with no recorded performances hence leading to higher genetic gain, particularly by shortening generation intervals in dairy cattle. In dairy cattle, the use of GEBV is a promising alternative to the long and costly progeny test. Since 2007, the potential interest of genomic selection in dairy cattle has been clearly demonstrated in terms of accuracy of breeding values (van Raden et al., 2009; Habier et al., 2011) and in terms of design of breeding programmes (Goddard & Hayes, 2007; Wensch-Dorendorf et al., 2011). Recently, several countries (Australia, France, Germany, the Netherlands, New Zealand, USA and

* Corresponding author: INRA, UMR1313 – Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France. E-mail: pascal.croiseau@jouy.inra.fr
others) implemented genomic selection for their national evaluations (Hayes et al., 2009; Van Raden et al., 2009; Boichard et al., 2010; Harris & Johnson, 2010; Liu et al., 2010).

Numerous methods have been proposed to perform genomic evaluations with variable resulting accuracy depending on the underlying genetic assumptions, on the trait, breed and reference population size. For instance, Habier et al. (2010a) tested a large panel of Bayesian approaches on a data set from the Holstein breed and even though Bayes A appeared to be a nearly optimal choice in their study, they recommended determining the best method for each quantitative trait separately. Indeed, in another study on Australian Holstein Friesian dairy cattle, Bayes A provided the lowest correlation between predicted GEBV and breeding values among the set of tested methods (Verbyla et al., 2009). On French data that Legarra et al. (2011) conducted for production traits, better predictions were obtained for Bayesian LASSO than for genomic BLUP (gBLUP). For other traits like fertility, it was shown that gBLUP performed slightly better than Bayesian LASSO (Hayes et al., 2009; Van Raden et al., 2010).

Hence, it is still difficult to rank the large panel of available genomic evaluation methods according to their accuracy.

In a genomic evaluation procedure where the complete set of SNP is used, the statistical challenge is to evaluate effects attached to each of the available SNPs. In a context where the number of SNPs \( p \) is much higher than the number of bulls \( n \), this may lead to a poor estimation of the SNP effects even though the sum of genotypes time effects may be adequate on this reference population. In a routine evaluation with new animals, the best way to be confident in DGV or GEBV is to attach an effect to SNP in linkage disequilibrium with a QTL which reflects the effect of the QTL and an effect regressed towards zero to the others.

An alternative is to use approaches that have been developed especially to solve the \( p >> n \) problem. This is the case of variable selection methods and, among them, we focused on the Elastic-Net (EN) algorithm (Zou & Hastie, 2005) and we chose to compare it to gBLUP, which is currently the most used approach in practice. Secondly, a two-step approach was tested by adding an initial preparation step consisting of an SNP pre-selection based on results from a QTL detection analysis. The second step implements gBLUP or EN on this preselected set of SNP with the hope that individual estimates of effects of the retuned SNP would be more accurate. To compare benefits and drawbacks of these situations, a pedigree-based BLUP was used as the reference.

| Breed              | Montbéliarde | Normande | Holstein |
|--------------------|--------------|----------|----------|
| Training data set  | 950          | 970      | 2976     |
| Validation data set| 222          | 248      | 964      |
| Total              | 1172         | 1218     | 3940     |

2. Materials and methods

(i) Data

The data sets consisted of 1172 Montbéliarde, 1218 Normande and 3940 Holstein bulls, which were all progeny tested and genotyped with the Illumina Bovine SNP50 BeadChip®. With a minimum minor allele frequency of 3%, 38,460 SNPs were retained for the Montbéliarde breed, 38,534 SNPs for the Normande breed and 39,738 SNPs for the Holstein breed. Mendelian segregation was checked. The SNP pre-selection chosen in this study uses a QTL detection method based on haplotypes which requires phased data. To infer missing genotypes and phases, DualPHASE software was used (Druet & Georges, 2009).

The data set was divided into a training data set to derive prediction equations and a validation data set where predictions were compared with observed phenotypes. Table 1 shows the size of training and validation data sets for the three breeds. To define the training and validation data sets, a cut-off date for the bulls’ birth date was introduced so that 25% of the youngest genotyped bulls were included in the validation dataset. Bulls without genotyped sire in the training dataset were excluded. Animals from the training data set were born before June 2002, while animals from the validation data set were born between June 2002 and 2004. This cross-validation design corresponds to the one used in studies of the EuroGenomics Consortium (Lund et al., 2010).

Phenotypes used for this study were daughter yield deviations (DYD) corresponding to the average performance of a sire’s daughters, adjusted for fixed and non-genetic random effects and for the additive genetic value of their dam (Mrode & Swanson, 2004). To account for the varying accuracy of the DYD, they were weighted by their error variance, which is proportional to the sire’s effective daughters’ contribution (EDC) (Fikse & Banos, 2001). DYD were included in the analysis if EDC exceeded 20.

For the three breeds, 25 traits were available: five production traits, two conception rate traits, 16 morphological traits, somatic cell counts and milking speed. Initially, only six traits were chosen to compare the different approaches and for fine tuning of different parameters. These six traits were the five
production traits (Milk yield, Fat yield, Fat content, Protein yield and Protein content) and cow conception rate (Boichard & Manfredi, 1994). Mean results over the 25 traits will also be shown.

(ii) Methods

The first method used was gBLUP (Van Raden et al., 2008) which uses the genomic relationship matrix, $G$ (Habier et al., 2007; Van Raden, 2008), instead of the pedigree-based relationship matrix

$$G = ZZ^T/2 \sum_{i=1}^{m} p_i(1-p_i),$$

where $m$ corresponds to the number of loci considered, $p_i$ is the frequency of an allele of the locus $i$ and $Z$ is the incidence matrix of SNP (genotype scores) on individuals, coded as in Van Raden (2008). The model is therefore: $y = Xb + g + e$, where $g$ is a vector of breeding values whose covariance matrix is described by $G\sigma_g^2$, where $\sigma_g^2$ is the polygenic variance.

Van Raden (2008) and Goddard (2009) showed that this model is equivalent to a mixed model fitting the effect of the genotype score of each SNP, all SNPs having a priori the same variance equal to $\sigma_g^2/2\sum_i p_i(1-p_i)$, where $\sigma_g^2$ is the polygenic variance used in regular genetic evaluation and $p_i$ is the frequency of an allele of the locus $i$ (Gianola et al., 2009).

The EN algorithm (Zou & Hastie, 2005; Croiseau et al., 2009) corresponds to a combination of the ridge regression (RR) and LASSO procedures. The difference between RR $\hat{\beta}_{\text{RR}} = \arg \min \{\sum_{i=1}^{n}(y_i-x_i\beta)^2 + \lambda \sum_j \beta_j^2\}$ and LASSO $\hat{\beta}_{\text{LASSO}} = \arg \min \{\sum_{i=1}^{n}(y_i-x_i\beta^2 + \lambda \sum_j \beta_j^2\}$ estimates lies in the form of the penalty term. In both equations, $\beta$ is the vector of SNP effects $\beta_j$, $y_i$ is the phenotype of animal $i$ and $x_i$ is its vector of genotypes. The $\lambda$ parameter corresponds to the intensity of the penalty. In the EN algorithm, a second parameter $\alpha$, taking a value in $[0, 1]$ is used to weight the RR and LASSO penalties.

$$\hat{\beta}_{\text{EN}} = \arg \min \left\{\sum_{i=1}^{n}(y_i-x_i\beta)^2 + \lambda \left(1-\alpha\right) \sum_j \beta_j^2 + \alpha \sum_j |\beta_j|\right\}$$

With $\alpha = 1$, a LASSO model is defined, whereas with $\alpha = 0$, a full RR model is chosen. Zou & Hastie (2003, 2005) showed that in the presence of correlated explanatory variables (e.g. effects corresponding to SNP in linkage disequilibrium in our case), RR retains all predictors and their corresponding coefficients tend to be equal and no variable selection is performed. On the other hand, LASSO retains only one predictor and removes the others (Zou & Hastie, 2003, 2005). Hence, by including RR and LASSO as extreme cases, the EN algorithm provides a more flexible tool.

In this study, EN procedures were used using an R package named ‘glmnet’ (http://cran.r-project.org/web/packages/glmnet/index.html) implemented by Friedman et al. (2008). They proposed a fast implementation of EN using cyclic coordinate descent, computed along a regularization path.

(iii) Pre-selection of the SNP

For most traits, not all SNPs on the SNP chip are likely to be close to a QTL. In other words, the assumption that effects attached to each of the SNPs are non null is unrealistic. Consequently, our conjecture is that whatever the genomic evaluation method used, a pre-selection of the SNP with an attached non-null effect may help to improve the quality of genomic prediction. This was tested in the situation where pre-selection is based on QTL detection. QTL detection was performed using a combined linkage disequilibrium and linkage analysis (LDLA) (Meuwissen & Goddard, 2001; Druet et al., 2008). First, the existence of a single QTL was tested in the training data set at all positions along the chromosomes defined by haplotypes of six SNPs, with a sliding window of two SNPs. From this LDLA, a value of the likelihood ratio test (LRT) was obtained for each haplotype. Positions where a potential QTL is located were defined as haplotypes each time an LRT peak higher than a threshold value of 3 or 5 was found. These values were quite arbitrary at this stage and low enough to catch any potential QTL that can be identified through this analysis. An LRT peak was defined as the position where the highest LRT value was found within a window of 25 or 50 SNP upstream and downstream of the current haplotype.

Then, the 50 SNPs around each detected LRT peak ($\pm 25$) were included in a pre-selected set of SNPs used for genomic evaluation using either a gBLUP or EN approach. The choice of the number of SNPs to retain was based on a preliminary study where this value of 50 gave the best results (data not shown).

(iv) Quality assessment of the genomic prediction

To measure the quality of prediction equations (derived from the training set), the equations were applied to the animals of the validation data set to get DGVs. Then, the weighted correlation between DGV and observed DYD was computed using EDC as weights. The weighted Pearson product moment correlation coefficient is calculated as (Peers, 1996):

$$r_{(x,y)} = \frac{\sum_w x_w(y_w - \bar{x}_w)(y_w - \bar{y}_w)}{\sqrt{\sum_w (x_w - \bar{x}_w)^2 \sum_w (y_w - \bar{y}_w)^2}}$$

where $\bar{x}_w = \frac{\sum_w x_w}{\sum_w}$, $\bar{y}_w = \frac{\sum_w y_w}{\sum_w}$ and $w_i$ is the EDC weight of $y_i$. 

The aim was to measure the accuracy of the different methods to predict DYD using genomic information (DGV). Since GEBV combine the information available from DGV and EBV, it is not possible to know if an observed gain in accuracy is due to the prediction equation or to a good combination of DGV and EBV. This is why the correlation between DGV and observed DYD was preferred in this study (see e.g. Guillaume et al., 2008).

(v) Parameters used for the different methods

For the pedigree-based BLUP, genetic parameters were estimated using an Average Information-Restricted Expectation Maximization Likelihood (AI-REML) approach (Jensen et al., 1996). For the LDLA, it is necessary to incorporate Identical by Descent (IBD) matrices among QTL allelic effects. Software of Misztal et al. (2002) was modified accordingly. Heritability estimates used in pedigree-based BLUP and gBLUP were those used in routine genetic evaluations.

For EN, values for the $\alpha$ and $\lambda$ penalization parameters needed to be chosen and there is currently no way to predict which range of values is the most appropriate for each parameter. Consequently, a large range of combinations of $\alpha$ and $\lambda$ was tested by grid search to find the optimal values. The search aimed at finding the maximum correlation between DGV and observed DYD in the validation data set. The validation data set is consequently used to identify the optimal set of parameters. This can be an advantage in comparison with other methods with respect to the accuracy of GEBV if this set of parameters is specific to this training and validation data sets. However, by looking at reference populations of increasing sizes, we found that these parameters were breed- and trait-specific with a rather large range of combinations giving similar results (data not shown). The EN approach appears robust to moderate departures from the optimal combination of parameters. To define the optimal $\alpha$ parameter, a dichotomous search was performed on the $[0, 1]$ interval. Initially, $\alpha$ values of 0, 1 and 0.5 were tested. If $\alpha=0$ provided the best correlation, at the second iteration, the interval was reduced to $[0, 0.5]$. If the best correlation was found with $\alpha=1$, the new interval was $[0.5, 1]$. If the best correlation was found with $\alpha=0.5$, the new interval was $[0.25, 0.75]$. We applied this method until the difference between two tested $\alpha$ was lower than 0.02. The dichotomous approach requires a unimodal distribution for these correlations which is not guaranteed. Nevertheless, after testing a large panel of $\alpha$ values for some traits (data not shown), this unimodal distribution seems to be the rule.

For each $\alpha$, 500 values of the penalty intensity $\lambda$ were tested in the interval $[0–\max(\beta)]$, where $\max(\beta)$ corresponds to the absolute value of the highest estimate when no penalization is applied.

This research of optimal values for $\alpha$ and $\lambda$ was performed separately for the pre-selected and the full data sets. The search for the optimal $\alpha$ parameter is the most time-consuming step of the glmnet package and takes around 2 CPU minutes in Holstein for each tested $\alpha$.

3. Results

Table 2 shows the optimal set of EN parameters for the six traits initially studied. Depending on the trait and breed, the optimal set of parameters differed. For instance, a complete RR approach gave the best results for Milk and Protein yield in the Montbéliarde breed, while optimal $\alpha$ values of 0.25 for Milk yield in Holstein and of 0.37 for Protein yield in Normande were found, which correspond to a general EN model. Moreover, there was a strong impact of both $\alpha$ and $\lambda$ on the number of SNPs included in the regression model. When $\alpha$ is near a complete LASSO procedure ($\alpha=1$), there were many fewer SNPs retained compared with a complete RR procedure ($\alpha=0$). Also, for a given $\alpha$, high values of $\lambda$ led to a high intensity of penalization and consequently to a lower number of SNP (results not shown).

In the second analysis, the SNP pre-selection based on QTL detection was performed. As indicated before, this SNP pre-selection relied on two criteria: a given LRT threshold and a given window size. Table 3

---

Table 2. Optimal $\alpha$ and $\lambda$, parameters and corresponding number of SNPs with non-null effect for the six traits studied and for the three breeds using the EN procedure on the complete set of SNPs

| Trait               | Montbéliarde | Normande | Holstein |
|---------------------|--------------|----------|----------|
|                     | $\alpha$     | $\lambda$| SNP      |
| Milk yield          | 0            | 267·17   | 24037    |
| Protein yield       | 0            | 12·31    | 23044    |
| Fat yield           | 0.01         | 6·71     | 5444     |
| Protein content     | 0.13         | 0·01     | 1776     |
| Fat content         | 1            | 0·01     | 723      |
| Conception rate     | 0            | 120·41   | 8215     |

---

The parameters used for the different methods are as follows:

- Milk yield: $\alpha=0$, $\lambda=0.09$, 24037
- Protein yield: $\alpha=0.37$, $\lambda=0.24$, 23044
- Fat yield: $\alpha=0.13$, $\lambda=0.58$, 5444
- Protein content: $\alpha=0.13$, $\lambda=0.005$, 1776
- Fat content: $\alpha=1$, $\lambda=0.06$, 723
- Conception rate: $\alpha=0$, $\lambda=4·33$, 8215

---

The optimal set of parameters for each trait and breed was determined, with the exception of Conception rate. For Montbéliarde, the optimal set was $\alpha=0$, $\lambda=0.09$. For Normande, it was $\alpha=0.37$, $\lambda=0.24$, and for Holstein, $\alpha=0.13$, $\lambda=0.58$. The number of SNPs with a non-null effect varied from 24037 to 723, depending on the trait and breed.
Fine tuning genomic evaluations in dairy cattle

Table 3. Number of LRT peaks identified for milk yield as a function of LRT threshold and window size in the Montbéliarde, Normande and Holstein breeds

| Breed       | SNP window size | LRT threshold |
|-------------|----------------|---------------|
|             | 3              | 5             |
| Montbéliarde| 25             | 432           | 265           |
|             | 50             | 273           | 180           |
| Normande    | 25             | 363           | 197           |
|             | 50             | 219           | 142           |
| Holstein    | 25             | 481           | 350           |
|             | 50             | 268           | 204           |

Table 4. Weighted correlation between DGV and observed DYD for the three breeds obtained using pedigree-based BLUP, gBLUP and EN on the complete set of SNP (54 K) or after a pre-selection of the SNP (PS)

| Trait            | Pedigree-based BLUP | gBLUP PS | EN PS |
|------------------|---------------------|----------|-------|
| Milk yield       | 0.28                | 0.44     | 0.43  |
| Fat yield        | 0.40                | 0.50     | 0.50  |
| Protein yield    | 0.27                | 0.46     | 0.47  |
| Fat content      | 0.40                | 0.51     | 0.56  |
| Protein content  | 0.25                | 0.44     | 0.42  |
| Conception rate  | 0.43                | 0.43     | 0.47  |
| Normande         | Milk yield          | 0.30     | 0.34  |
|                  | Fat yield           | 0.27     | 0.39  |
|                  | Protein yield       | 0.23     | 0.31  |
|                  | Fat content         | 0.58     | 0.61  |
|                  | Protein content     | 0.33     | 0.50  |
|                  | Conception rate     | 0.24     | 0.27  |
| Holstein         | Milk yield          | 0.38     | 0.56  |
|                  | Fat yield           | 0.40     | 0.59  |
|                  | Protein yield       | 0.44     | 0.55  |
|                  | Fat content         | 0.44     | 0.72  |
|                  | Protein content     | 0.47     | 0.73  |
|                  | Conception rate     | 0.29     | 0.35  |

Table 5. Slope of the regression of observed DYD on DGV for the Holstein breed obtained using pedigree-based BLUP, gBLUP and EN on the complete set of SNP (54 K) or after a pre-selection of the SNP (PS)

| Trait            | Pedigree-based BLUP | gBLUP | EN |
|------------------|---------------------|-------|----|
| Milk yield       | 0.80                | 0.68  | 0.80|
| Fat yield        | 0.96                | 0.80  | 0.61|
| Protein yield    | 0.86                | 0.65  | 0.76|
| Fat content      | 0.98                | 0.87  | 0.89|
| Protein content  | 0.94                | 0.90  | 0.83|
| Conception rate  | 0.80                | 0.78  | 0.69|

When an SNP pre-selection was applied, the gain in correlation using gBLUP and EN was very similar to the one observed using the complete set of SNP. Again, among the two different genomic approaches, the best results were obtained with EN. Compared with the pedigree-based BLUP, the mean gains over the six traits were 0.14, 0.15 and 0.20 for Montbéliarde, Normande and Holstein, respectively, compared with 0.12, 0.08 and 0.18 with gBLUP.

Table 6 presents the number of SNPs with a non-null effect retained by EN algorithm without or with a pre-selection of SNP in the Holstein breed. Similar results were obtained in Montbéliarde and Normande (data not shown). The results for the six traits are given, as well as the average of the number of SNPs over the 25 traits available for the three breeds. The large part of the variance, such as protein content and fat content (where DGAT1 gene is present), a mean gain in correlation over the three breeds of +0.22 and +0.23, respectively, was observed. In contrast, when the trait background appears to be polygenic with many QTLs explaining only a small part of the variance each, as for conception rate, the observed mean gain in correlation was more limited (+0.06). Between the two genomic approaches, EN gave better results with a mean gain (compared with pedigree-based BLUP) over the six traits of 0.15, 0.13 and 0.20 for Montbéliarde, Normande and Holstein, respectively, compared with 0.12, 0.08 and 0.18 with gBLUP.

Table 5 shows the slope of the regression of observed DYD on DGV for Holstein. A value close to 1 is expected. In dairy cattle, genomic evaluations are validated by Interbull if the slope of regression is included between 0.8 and 1.2 (Interbull, 2011). Over the three tested methods, similar ranges of values were observed for pedigree-based BLUP and EN. The slope for gBLUP deviated more from 1 than for the two other methods (on average, 0.22 for gBLUP compared with 0.11 for pedigree-based BLUP and 0.12 for EN). The same analysis was performed for the approach with SNP pre-selection. For EN, the SNP pre-selection had no impact on the slope.

Table 6 presents the number of SNPs with a non-null effect retained by EN algorithm without or with a pre-selection of SNP in the Holstein breed. Similar results were obtained in Montbéliarde and Normande (data not shown). The results for the six traits are given, as well as the average of the number of SNPs over the 25 traits available for the three breeds. The
The number of SNPs retained was dependent on the genetic architecture of the trait. Traits such as Fat content where DGAT1 explains a very high part of the variance required fewer SNPs than conception rate. The mean number of SNPs over the 25 traits illustrates the impact of pre-selection on the number of retained SNPs.

For the six presented traits, pre-selection led to a reduction of the number of SNPs needed in the prediction equation. Among these traits, conception rate is the one with the highest polygenic part as the number of SNPs included in the EN model shows. Production traits required between 1271 and 5648, which is much less than the 20 904 SNPs required for conception rate. The highest reduction of the number of SNPs retained was for conception rate (from 20 904 to 9677 SNPs, which corresponds to a reduction of 54%).

The impact of this SNP pre-selection on correlations was an absolute decrease limited to 1–2% and was relatively limited. For the 25 available traits, the average number of SNPs used in the prediction equation derived from the EN algorithm applied on the whole set of SNPs was 16 334. After pre-selection, this number declined to 10 059. This important decrease in the number of SNPs used was obtained while correlations remained relatively stable (loss of 1% on average). Surprisingly, for some traits, the number of SNPs retained by EN after pre-selection was higher than when EN was applied to the whole set of SNPs. This was the case for body depth, chest width and milking speed for Holstein. Nevertheless, this phenomenon was marginal and, for most traits, pre-selection allowed a large decrease in SNP numbers.

Table 6. Correlation and number of SNP used in the prediction equation using the EN algorithm on the whole set of SNP (54 K) or after a pre-selection of the SNP (PS) for the Holstein breed

| Traits            | 54 K Correlation | Number of SNPs | PS Correlation | Number of SNPs | Impact on Correlation |
|-------------------|------------------|----------------|----------------|----------------|-----------------------|
| Milk yield        | 0.57             | 1355           | 0.59           | 1329           | 0.02                  |
| Fat yield         | 0.63             | 1271           | 0.62           | 1211           | 0.01                  |
| Protein yield     | 0.57             | 5648           | 0.56           | 1098           | 0.02                  |
| Fat content       | 0.79             | 1351           | 0.78           | 1087           | 0.01                  |
| Protein content   | 0.75             | 2297           | 0.73           | 1742           | 0.02                  |
| Conception rate   | 0.33             | 20 904         | 0.34           | 9677           | 0.01                  |
| Mean over the 6 traits | –             | 5471           | –              | 2691           | –0.01                 |
| Men over 25 traits | –             | 16 334         | –              | 10 059         | –0.01                 |

Table 7. Highest correlation and corresponding number of selected SNPs when using the whole set of SNP (54 K), after a pre-selection of the SNP (PS) or when the number of selected SNPs is limited to 2500, 1500 or 1000 in the Holstein breed

| Traits            | 54 K Correlation | SNPs | PS Correlation | SNPs | 2500 SNPs Correlation | 2500 SNPs SNPs | 1500 SNPs Correlation | 1500 SNPs SNPs | 1000 SNPs Correlation | 1000 SNPs SNPs |
|-------------------|------------------|------|----------------|------|-----------------------|----------------|-----------------------|----------------|-----------------------|----------------|
| Milk yield        | 0.569            | 1328 | 0.573          | 2752 | 0.573                 | 2422           | 0.569                 | 1328           | 0.551                 | 955            |
| Fat yield         | 0.631            | 1273 | 0.626          | 1126 | 0.631                 | 1273           | 0.631                 | 1273           | 0.624                 | 991            |
| Protein yield     | 0.573            | 21176| 0.568          | 2390 | 0.568                 | 2120           | 0.565                 | 1448           | 0.561                 | 959            |
| Fat content       | 0.795            | 1364 | 0.791          | 1068 | 0.795                 | 1364           | 0.795                 | 1364           | 0.791                 | 985            |
| Protein content   | 0.748            | 2368 | 0.731          | 3684 | 0.748                 | 2368           | 0.696                 | 1419           | 0.694                 | 996            |
| Conception rate   | 0.335            | 20 853| 0.328          | 9144 | 0.320                 | 2379           | 0.307                 | 1141           | 0.301                 | 850            |
correlations that were observed when the total number of SNPs with non-null effect was limited to a value between 2500 and 1000 SNPs. An option of the R package glmnet allows the maximum number of variables to be set. This option acts on the intensity of the penalization to validate this constraint.

Obviously, this limitation in the number of SNPs led to a decrease in correlation, but this was relatively limited: between 0 and 3.4% depending on the trait and the maximum number of SNPs defined. In complement to this table, Figure 1 presents the mean change in correlation over the 25 traits for the three breeds according to the number of selected SNPs.

When the number of selected SNPs was limited to 2500, the average absolute loss in correlation over the 25 traits ranged from 1 and 1.7%. This average loss in correlation changed to 2.3 and 4.5% with a limit to 1000 selected SNPs.

4. Discussion

As for many previous studies, genomic evaluations with gBLUP and EN substantially improved the quality of prediction of observed DYD in the validation data set compared with pedigree-based BLUP (Hayes et al., 2009; Wolc et al., 2011). Between these two genomic evaluations, gBLUP has the advantage of being conceptually simpler in the sense that there is no extra parameter to define or to optimize. In theory, a method that estimates all SNP effects should ensure that false-positive or uninformative effects are regressed towards zero, but in practice, these false positive or uninformative effects are not strictly equal to zero. EN, which shares some variable selection properties with other methods (like Bayes B, Cp, …) limits the number of SNPs with non-null estimated effects in the model. This property can be an advantage because it alleviates the p >> n problems, in particular for smaller breeds. Limiting the number of SNP effects to estimates becomes important for an accurate prediction equation.

Since this study shows that EN provides better results than gBLUP for most traits in the three breeds studied, we tried to push further the idea of variable selection both in the case of gBLUP and EN by adding an SNP pre-selection step based on QTL detection. The resulting correlations between DGV and observed DYD and also the slopes of the regression of observed DYD on DGV were similar to the ones obtained using the complete set of SNPs. Moreover, both the EN algorithm and the pre-selection of the SNP led to a reduction of the number of SNPs included in the prediction equation with a minor effect on the quality of prediction. This procedure seems particularly relevant in the genomic selection context for two reasons:

- From a genetic standpoint, it is consistent with the assumption that not all SNPs are required to explain the genetic architecture of a given trait. Some
of them, with non-significant effects, can still carry genetic information and particularly on genetic relationships (Habier et al., 2007, 2010b). However, since very similar correlations were obtained using the complete set of SNPs or a fraction of them after pre-selection, it means that a subset of SNPs included in the model was not really informative for the trait and pre-selection avoids including in the prediction equation these uninformative SNPs.

- Furthermore, it is expected that in the near future the number of genotyped animals and the number of SNPs will get larger and larger. This will represent a major challenge for genomic evaluations from a computing point of view. The SNP pre-selection implemented here requires an LDLA approach and a detection of the LRT peak, which is based on two parameters (windows of SNP to consider and an LRT threshold). The LDLA approach requires phasing the data which, depending on the methodology used, could be computationally time consuming. However, the LDLA approach does not have to be performed at each genomic evaluation because animals that are added between two genomic evaluations are young and their performances have a very low weight compared with older ones. Moreover, as mentioned before, the time-consuming step is to phase the data. Actually, this step is not required for all the genomic selection methods used in national evaluation and consequently, constraints due to phasing data are not encountered. But if an additional imputation step is required to mix different versions of chips (Illumina Bovine SNP50 BeadChip® V1 and V2 for example) or different sizes of chips (3, 50 and 777 K), this phasing step is routinely needed anyway. Then, SNP pre-selection strongly alleviates computing requirements and consequently ensures that national evaluations can be completed within a reasonable time frame.

In this study, we focused on one variable selection method that is the EN and one pre-selection method that is LDLA. Obviously, other genomic selection methods (Bayesian methods for instance) and other pre-selection approaches (based on ‘pure’ association studies instead of LDLA for instance) should be also tested to complete this study. EN provided better results in our study and our model assumed that all genetic variation was explained by SNP. The latter may be true if all causal mutations are bi-allelic and if SNPs are in strong linkage disequilibrium with all causal mutations. If causal mutations are multi-allelic or if SNPs are in weak linkage disequilibrium with this causal mutation, model based on haplotypes could be more advantageous. The current French genomic evaluation (Boichard et al., 2010) combines Marker Assisted Selection (MAS) on QTL followed through haplotypes and genomic selection based on SNP detected with the EN algorithm. EN was used as a variable selection method and prediction equations were generated for the French genomic MAS.

In conclusion, the EN algorithm appears to be a very flexible and promising tool in the genomic selection framework that can be used for genomic evaluation or as a variable selection device to provide SNP of interest to a marker-assisted evaluation method.

This work was part of the AMASGEN project within the ‘UMT évaluation génétique’ financed by the French National Research Agency (ANR) and by ApisGene. Labogena is gratefully acknowledged for providing the genotypes.

Declaration of Interest

None.

References

Andersson, L. & Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics 5, 202–212.

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M.-N., Boscher, M.-Y., et al. (2010). Genomic selection in French dairy cattle. In 9th World Congress on Genetics Applied to Livestock Production. Germany: Leipzig.

Boichard, D. & Manfredi, E. (1994). Genetic analysis of conception rate in French Holstein cattle. Acta Agriculturae Scandinavica 44, 138–145.

Croiseau, P., Guillaume, F., Fritz, S. & Ducrocq, V. (2009). Use of the Elastic-Net algorithm for genomic selection in dairy cattle. In 60th Annual Meeting of the EAAP 2009. Spain: Barcelona.

Druet, T., Fritz, S., Boussaha, M., Ben-Jemaa, S., Guillaume, F., Derbala, D., et al. (2008). Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178, 2227–2235.

Druet, T. & Georges, M. (2009). A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and QTL fine mapping. Genetics 184, 789–798.

Fikse, W. F. & Banos, G. (2001). Weighting factors of sire daughter information in international genetic evaluations. Journal of Dairy Science 84, 1759–1767.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1).

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

Goddard, M. & Hayes, B. (2007). Genomic selection. Journal of Animal Breeding and Genetics 124, 323–330.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.

Guillaume, F., Fritz, S., Boichard, D. & Druet, T. (2008). Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genetics Selection Evolution 40, 91–102.
Habier, D., Fernando, R. L. & Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397.

Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. (2010a). Extension of the Bayesian alphabet for genomic selection. In *9th World Congress on Genetics Applied to Livestock Production*. Germany: Leipzig.

Habier, D., Tetens, J., Seefried, F. R., Lichtner, P. & Thaller, G. (2010b). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* **42**, 5.

Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **23**, 186.

Haley, C. S. & Visscher, P. M. (1998). Strategies to utilize marker-quantitative trait loci associations. *Journal of Dairy Science* **81**(Suppl. 2), 85–97.

Harris, B. & Johnson, D. (2010). Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *Journal of Dairy Science* **93**, 1243–1252.

Hayes, B., Bowman, P., Chamberlain, A. & Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* **92**, 433–443.

Interbull (2011). GEBV Test. http://www.interbull.org/index.php?option=com_content&view=article&id=80&Itemid=114

Jensen, J., Mantysaari, E. A., Madsen, P. & Thompson, R. (1996). Residual maximum likelihood estimation of (co-)variance components in multivariate mixed linear models using average information. *Journal of Indian Society for Agricultural Statistics* **49**, 215–236.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F. & Fritz, S. (2011). Improved Lasso for genomic selection. *Genetics Research (Cambridge)* **93**, 77–87.

Liu, Z., Seefried, F., Reinhardt, F., Thaller, G. & Reents, R. (2010). Dairy cattle genetic evaluation enhanced with genomic information. In *9th World Congress on Genetics Applied to Livestock Production*. Germany: Leipzig.

Lund, M. S., de Roos, A. P. W., de Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., *et al.* (2010). Improving genomic prediction by EuroGenomics collaboration. In *9th World Congress on Genetics Applied to Livestock Production*. Germany: Leipzig.

Meuwissen, T. & Goddard, M. (2001). Prediction of identity by descent probabilities from marker-haplotype. *Genetics Selection Evolution* **33**, 605–634.

Meuwissen, T., Hayes, B. & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Misztaš, I., Tsuruta, S., Stabell, A., Avrás, B., Druet, T. & Lee, D. H. (2002). BLUPF90 and related programs (BGF90). *7th World Congress on Genetics Applied to Livestock Production*. France: Montpellier.

Mrode, R. A. & Swanson, G. J. T. (2004). Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. *Livestock Production Science* **86**, 253–260.

Peers, I. (1996). *Statistical Analysis for Education and Psychology Researchers*. Washington, DC: Falmer Press.

Van Raden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414–4423.

Van Raden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., *et al.* (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16–24.

Verbyla, K., Hayes, B., Bowman, P. & Goddard, M. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research* **91**, 307–311.

Wensch-Dorendorf, M., Yin, T., Swalve, H. H. & König, S. (2011). Optimal strategies for the use of genomic selection in dairy cattle breeding programs. *Journal of Dairy Science* **94**, 4140–5151.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E. & O’Sullivan, N. P., *et al.* (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution* **43**, 5.

Zou, H. & Hastie, T. (2003). *Regression Shrinkage and Selection via the Elastic Net*, with Application to Microarrays. Stanford University: Department of Statistics.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Royal Statistical Society Series B* **67**, 301–320.

Downloaded from https://www.cambridge.org/core. IP address: 54.70.40.11, on 08 Aug 2021 at 22:09:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms . https://doi.org/10.1017/s0016672311000358