



HAL
open science

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

Pascal Croiseau, Andres Legarra, François F. Guillaume, Sébastien S. Fritz,
Aurélia A. Baur, Carine Colombani Colombani, Christèle Robert-Granié,
Didier Boichard, Vincent Ducrocq

► To cite this version:

Pascal Croiseau, Andres Legarra, François F. Guillaume, Sébastien S. Fritz, Aurélia A. Baur, et al.. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genetics Research*, 2011, 93 (6), pp.409-417. 10.1017/S0016672311000358 . hal-01000269

HAL Id: hal-01000269

<https://hal.science/hal-01000269>

Submitted on 29 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

2
3 PASCAL CROISEAU^{1*}, ANDRÉS LEGARRA², FRANÇOIS GUILLAUME³,
4 SÉBASTIEN FRITZ⁴, AURÉLIA BAUR⁴, CARINE COLOMBANI²,
5 CHRISTÈLE ROBERT-GRANIÉ², DIDIER BOICHARD¹ AND VINCENT DUCROCQ¹

6 ¹INRA, UMR1313 – Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France

7 ²INRA, UR 631, Station d'Amélioration Génétique des Animaux, F-31320 Castanet-Tolosan, France

8 ³Institut de l'élevage, 149 rue de Bercy, 75595 Paris, France

9 ⁴UNCEIA, 149 rue de Bercy, 75595 Paris, France

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

11 Summary

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

27 1. Introduction

28 The availability of dense single-nucleotide poly-
29 morphism (SNP) arrays has considerably changed the
30 landscape of dairy cattle selection worldwide. With
31 such chips, it is now possible to retrieve information
32 about quantitative trait locus (QTL) all over the ge-
33 nome. Genomic estimated breeding values (GEBV),
34 which correspond to a combination of the sum of
35 the effects of genetic markers (direct genomic value
36 (DGV)) and estimated breeding value (EBV), can be
37 used instead of the classical pedigree-based genetic
38 evaluations in selection programmes. Meuwissen *et al.*
39 (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 simula-
tions, 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 gener-
ation 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, 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

57 others) implemented genomic selection for their
58 national evaluations (Hayes *et al.*, 2009; Van Raden
59 *et al.*, 2009; Boichard *et al.*, 2010; Harris & Johnson,
60 2010; Liu *et al.*, 2010).

61 Numerous methods have been proposed to per-
62 form genomic evaluations with variable resulting
63 accuracy depending on the underlying genetic as-
64 sumptions, on the trait, breed and reference popu-
65 lation size. For instance, Habier *et al.* (2010a) tested
66 a large panel of Bayesian approaches on a data set
67 from the Holstein breed and even though Bayes
68 A appeared to be a nearly optimal choice in their
69 study, they recommended determining the best
70 method for each quantitative trait separately. Indeed,
71 in another study on Australian Holstein Friesian
72 dairy cattle, Bayes A provided the lowest correla-
73 tion between predicted GEBV and breeding values
74 among the set of tested methods (Verbyla *et al.*, 2009).
75 On French data that Legarra *et al.* (2011) conducted
76 for production traits, better predictions were ob-
77 tained for Bayesian LASSO than for genomic
78 BLUP (gBLUP). For other traits like fertility, it
79 was shown that gBLUP performed slightly better
80 than Bayesian LASSO (Hayes *et al.*, 2009; Van Raden
81 *et al.*, 2009).

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

85 In a genomic evaluation procedure where the com-
86 plete set of SNP is used, the statistical challenge is
87 to evaluate effects attached to each of the available
88 SNPs. In a context where the number of SNPs (p) is
89 much higher than the number of bulls (n), this may
90 lead to a poor estimation of the SNP effects even
91 though the sum of genotypes time effects may be ad-
92 equate on this reference population. In a routine
93 evaluation with new animals, the best way to be con-
94 fident in DGV or GEBV is to attach an effect to SNP
95 in linkage disequilibrium with a QTL which reflects
96 the effect of the QTL and an effect regressed towards
97 zero to the others.

98 An alternative is to use approaches that have
99 been developed especially to solve the $p \gg n$ problem.
100 This is the case of variable selection methods and,
101 among them, we focused on the Elastic-Net (EN) al-
102 gorithm (Zou and Hastie, 2005) and we chose to
103 compare it to gBLUP, which is currently the most
104 used approach in practice. Secondly, a two-step ap-
105 proach was tested by adding an initial preparation
106 step consisting of an SNP pre-selection based on re-
107 sults from a QTL detection analysis. The second
108 step implements gBLUP or EN on this preselected
109 set of SNP with the hope that individual estimates
110 of effects of the retuned SNP would be more
111 accurate. To compare benefits and drawbacks of these
112 situations, a pedigree-based BLUP was used as the
113 reference.

Table 1. Number of animals genotyped per data set
for the three breeds studied

	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 Bov-
ine 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, Dual-
PHASE 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 be-
tween 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 pro-
portional to the sire's effective daughters' contri-
bution (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 mor-
phological traits, somatic cell counts and milking
speed. Initially, only six traits were chosen to com-
pare the different approaches and for fine tuning of
different parameters. These six traits were the five

159 production traits (Milk yield, Fat yield, Fat content,
160 Protein yield and Protein content) and cow concep-
161 tion rate (Boichard & Manfredi, 1994). Mean results
162 over the 25 traits will also be shown.

163 (ii) *Methods*

164 The first method used was gBLUP (Van Raden *et al.*,
165 2008) which uses the genomic relationship matrix,
166 \mathbf{G} (Habier *et al.*, 2007; Van Raden, 2008), instead of
167 the pedigree-based relationship matrix

$$\mathbf{G} = \mathbf{Z}\mathbf{Z}'/2 \sum_{i=1}^m p_i(1-p_i),$$

168 where m corresponds to the number of loci con-
169 sidered, p_i is the frequency of an allele of the locus
170 i and \mathbf{Z} is the incidence matrix of SNP (genotype
171 scores) on individuals, coded as in Van Raden (2008).
172 The model is therefore: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{g} + \mathbf{e}$, where \mathbf{g} is a
173 vector of breeding values whose covariance matrix is
174 described by $\mathbf{G}\sigma_u^2$, where σ_u^2 is the polygenic variance.

175 Van Raden (2008) and Goddard (2009) showed
176 that this model is equivalent to a mixed model fitting
177 the effect of the genotype score of each SNP, all SNPs
178 having *a priori* the same variance equal to $\sigma_a^2 =$
179 $\sigma_u^2/2\sum p_i(1-p_i)$, where σ_u^2 is the polygenic variance
180 used in regular genetic evaluation and p_i is the fre-
181 quency of an allele of the locus i (Gianola *et al.*, 2009).

182 The EN algorithm (Zou & Hastie, 2005; Croiseau
183 *et al.*, 2009) corresponds to a combination of the ridge
184 regression (RR) and LASSO procedures. The differ-
185 ence between RR $\hat{\beta}_{RR} = \arg \min \{ \sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 +$
186 $\lambda \sum_j \beta_j^2 \}$ and LASSO $\hat{\beta}_{LASSO} = \arg \min \{ \sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 +$
187 $\lambda \sum_j |\beta_j| \}$ estimates lies in the form of the penalty term.
188 In both equations, $\boldsymbol{\beta}$ is the vector of SNP effects β_j , y_i
189 is the phenotype of animal i and \mathbf{x}_i is its vector of
190 genotypes. The λ parameter corresponds to the in-
191 tensity of the penalty. In the EN algorithm, a second
192 parameter α , taking a value in $[0, 1]$ is used to weight
193 the RR and LASSO penalties.

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

194 With $\alpha = 1$, a LASSO model is defined, whereas with
195 $\alpha = 0$, a full RR model is chosen. Zou and Hastie
196 (2003, 2005) showed that in the presence of correlated
197 explanatory variables (e.g. effects corresponding to
198 SNP in linkage disequilibrium in our case), RR re-
199 tains all predictors and their corresponding coeffi-
200 cients tend to be equal and no variable selection is
201 performed. On the other hand, LASSO retains only
202 one predictor and removes the others (Zou & Hastie,
203 2003, 2005). Hence, by including RR and LASSO as
204 extreme cases, the EN algorithm provides a more
205 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 im-
plementation of EN using cyclical 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 as-
sumption 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 dis-
equilibrium and linkage analysis (LDLA) (Druet
et al., 2008; Meuwissen & Goddard, 2001). First, the
existence of a single QTL was tested in the training
data set at all positions along the chromosomes de-
fined by haplotypes of six SNPs, with a sliding win-
dow 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
(± 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 cor-
relation coefficient is calculated as (Peers, 1996):

$$r_{(x,y)} = \frac{\sum w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum w_i(x_i - \bar{x}_w)^2 \sum w_i(y_i - \bar{y}_w)^2}}$$

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

Table 2. Optimal α and λ 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

	Montbéliarde			Normande			Holstein		
	α	λ	SNP	α	λ	SNP	α	λ	SNP
Milk yield	0	267.17	24 037	0.09	25.25	1529	0.25	15.18	1355
Protein yield	0	12.31	23 044	0.37	0.24	866	0.01	5.79	5648
Fat yield	0.01	6.71	5444	0.13	0.58	1474	0.25	0.65	1271
Protein content	0.13	0.01	1776	1	0.005	737	0.25	0.01	2297
Fat content	1	0.01	723	0.59	0.06	403	0.65	0.02	1351
Conception rate	0	120.41	8215	0.02	4.33	2879	0	17.49	20 904

255 The aim was to measure the accuracy of the
256 different methods to predict DYD using genomic in-
257 formation (DGV). Since GEBV combine the infor-
258 mation available from DGV and EBV, it is not
259 possible to know if an observed gain in accuracy is
260 due to the prediction equation or to a good combi-
261 nation of DGV and EBV. This is why the correlation
262 between DGV and observed DYD was preferred in
263 this study (see e.g. Guillaume *et al.*, 2008).

264 (v) Parameters used for the different methods

265 For the pedigree-based BLUP, genetic parameters
AQ2 266 were estimated using an AI-REML approach (Jensen
267 *et al.*, 1996). For the LDLA, it is necessary to incor-
AQ3 268 porate IBD matrices among QTL allelic effects.
269 Software of Misztal *et al.* (2002) was modified ac-
270 cordingly. Heritability estimates used in pedigree-
271 based BLUP and gBLUP were those used in routine
272 genetic evaluations.

273 For EN, values for the α and λ penalization para-
274 meters needed to be chosen and there is currently no
275 way to predict which range of values is the most ap-
276 propriate for each parameter. Consequently, a large
277 range of combinations of α and λ was tested by grid
278 search to find the optimal values. The search aimed at
279 finding the maximum correlation between DGV and
280 observed DYD in the validation data set. The vali-
281 dation data set is consequently used to identify the
282 optimal set of parameters. This can be an advantage
283 in comparison with other methods with respect to the
284 accuracy of GEBV if this set of parameters is specific
285 to this training and validation data sets. However, by
286 looking at reference populations of increasing sizes,
287 we found that these parameters were breed- and trait-
288 specific with a rather large range of combinations
289 giving similar results (data not shown). The EN ap-
290 proach appears robust to moderate departures from
291 the optimal combination of parameters. To define the
292 optimal α parameter, a dichotomous search was per-
293 formed on the [0, 1] interval. Initially, α values of 0, 1
294 and 0.5 were tested. If $\alpha=0$ provided the best corre-
295 lation, at the second iteration, the interval was re-
296 duced 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 α was lower than 0.02.
The dichotomous approach requires a unimodal dis-
tribution for these correlations which is not guaran-
teed. Nevertheless, after testing a large panel of α
values for some traits (data not shown), this unimodal
distribution seems to be the rule.

For each α , 500 values of the penalty intensity λ
were tested in the interval [0–max(β)], where max(β)
corresponds to the absolute value of the highest esti-
mate when no penalization is applied.

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

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 re-
sults for Milk and Protein yield in the Montbéliarde
breed, while optimal α 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 α and λ
on the number of SNPs included in the regression
model. When α is near a complete LASSO procedure
($\alpha=1$), there were many fewer SNPs retained com-
pared with a complete RR procedure ($\alpha=0$). Also, for
a given α , high values of λ 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 be-
fore, this SNP pre-selection relied on two criteria: a
given LRT threshold and a given window size. Table 3
reports the effect of both criteria on the number of
LRT peaks identified in the case of milk yield.

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

	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)

	Pedigree-based BLUP	gBLUP		EN	
		54 K	PS	54 K	PS
Montbéliarde					
Milk yield	0.28	0.44	0.43	0.45	0.42
Fat yield	0.40	0.50	0.50	0.50	0.51
Protein yield	0.27	0.46	0.47	0.46	0.47
Fat content	0.40	0.51	0.56	0.59	0.59
Protein content	0.25	0.44	0.42	0.44	0.42
Conception rate	0.43	0.43	0.42	0.47	0.48
Normande					
Milk yield	0.30	0.34	0.38	0.41	0.42
Fat yield	0.27	0.39	0.38	0.41	0.41
Protein yield	0.23	0.31	0.33	0.37	0.40
Fat content	0.58	0.61	0.63	0.71	0.75
Protein content	0.33	0.50	0.55	0.54	0.53
Conception rate	0.24	0.27	0.30	0.31	0.31
Holstein					
Milk yield	0.38	0.56	0.56	0.57	0.57
Fat yield	0.40	0.59	0.59	0.63	0.63
Protein yield	0.44	0.55	0.54	0.57	0.57
Fat content	0.44	0.72	0.74	0.80	0.79
Protein content	0.47	0.73	0.73	0.75	0.73
Conception rate	0.29	0.35	0.33	0.33	0.33

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)

	Pedigree-based BLUP	gBLUP		EN	
		54 K	PS	54 K	PS
Holstein					
Milk yield	0.80	0.68	0.68	0.80	0.80
Fat yield	0.96	0.80	0.61	1.06	1.05
Protein yield	0.86	0.65	0.76	0.80	0.78
Fat content	0.98	0.87	0.89	0.95	0.98
Protein content	0.94	0.90	0.83	0.93	0.92
Conception rate	0.80	0.78	0.69	0.84	0.84

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.

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, 0.11 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 number of SNPs retained was dependent on the genetic architecture of the trait. Traits such as Fat

339 Table 4 presents for the three breeds the results
340 obtained with the classical pedigree-based BLUP and
341 the two genomic selection methods (gBLUP and EN)
342 when either the whole set of SNP which passed the
343 quality control was used or after a pre-selection of the
344 SNP based on the LDLA approach.

345 All genomic methods improved the correlation
346 between DGV and observed DYD compared with
347 pedigree-based BLUP and the genetic architecture of
348 the trait seemed to play an important role on the gain
349 in correlation: for traits where some QTLs explain a
350 large part of the variance, such as protein content and
351 fat content (where DGAT1 gene is present), a mean

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	Holstein				
	54 K		PS		
	Correlation	Number of SNPs	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	20904	0.34	9677	0.01
Mean over the 6 traits	-	5471	-	2691	-0.01
Men over 25 traits	-	16334	-	10059	-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

		54 K	PS	2500 SNPs	1500 SNPs	1000 SNPs
		Milk yield	Correlation	0.569	0.573	0.573
	SNP	1328	2752	2422	1328	955
Fat yield	Correlation	0.631	0.626	0.631	0.631	0.624
	SNP	1273	1126	1273	1273	991
Protein yield	Correlation	0.573	0.568	0.568	0.565	0.561
	SNP	21716	2390	2120	1448	959
Fat content	Correlation	0.795	0.791	0.795	0.795	0.791
	SNP	1364	1068	1364	1364	985
Protein content	Correlation	0.748	0.731	0.748	0.696	0.694
	SNP	2368	3684	2368	1419	996
Conception rate	Correlation	0.335	0.328	0.320	0.307	0.301
	SNP	20853	9144	2379	1141	850

394 content where DGAT1 explains a very high part of
 395 the variance required fewer SNPs than conception
 396 rate. The mean number of SNPs over the 25 traits
 397 illustrates the impact of pre-selection on the number
 398 of retained SNPs.

399 For the six presented traits, pre-selection led to a
 400 reduction of the number of SNPs needed in the pre-
 401 diction equation. Among these traits, conception rate
 402 is the one with the highest polygenic part as the
 403 number of SNPs included in the EN model shows.
 404 Production traits required between 1271 and 5648,
 405 which is much less than the 20904 SNPs required for
 406 conception rate. The highest reduction of the number
 407 of SNPs retained was for conception rate (from 20904
 408 to 9677 SNPs, which corresponds to a reduction of
 409 54%).

410 The impact of this SNP pre-selection on correla-
 411 tions was an absolute decrease limited to 1–2% and
 412 was relatively limited. For the 25 available traits,
 413 the average number of SNPs used in the prediction

equation derived from the EN algorithm applied on
 the whole set of SNPs was 16334. After pre-selection,
 this number declined to 10059. This important de-
 crease 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.
 The results presented in Table 6 correspond to the
 optimal α and λ values. During the EN procedure, a
 large number of parameter combinations were tested
 and some suboptimal combinations required an even
 smaller number of SNPs. Table 7 presents, for the
 Holstein breed and for the six initial traits, the highest
 correlations that were observed when the total num-
 ber of SNPs with non-null effect was limited to a value

Comment citer ce document :

Croiseau, P., Legarra, A., Guillaume, F., Fritz, S., Baur, A., Colombani, C., Robert Granie, C., Boichard, D., Ducrocq, V. (2011). Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genetics Research*, 93 (6), 409-417. DOI : 10.1017/S0016672311000358

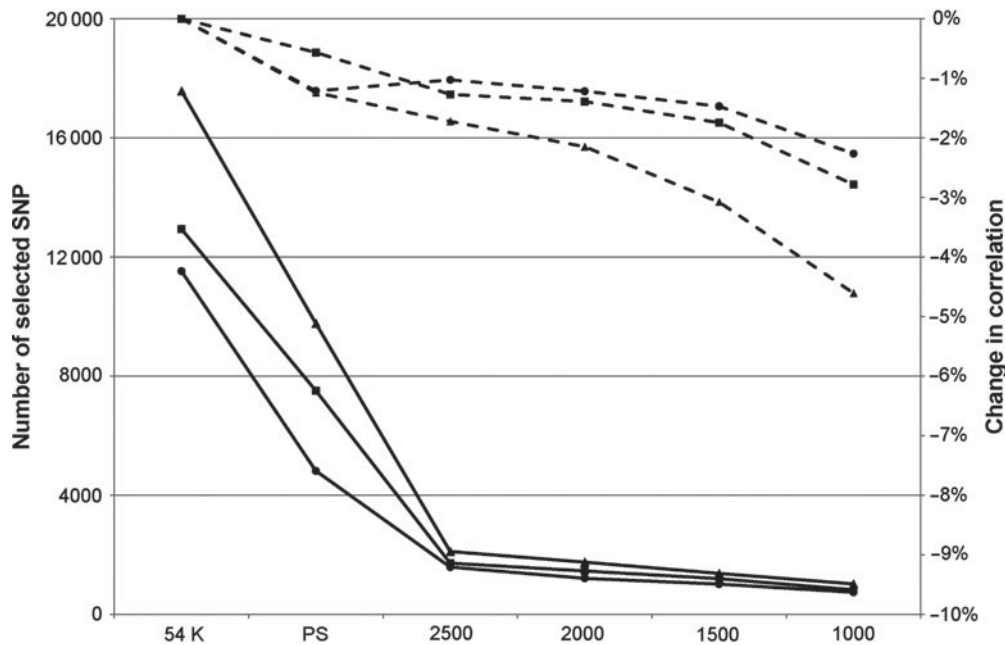


Fig. 1. Mean change in correlation (dashed lines) over the 25 traits for Montbéliarde (■), Normande (●) and Holstein (▲) when the maximum number of SNPs selected by EN is restricted to the value indicated on the x-axis. Continuous lines represent the actual number of selected SNPs.

434 between 2500 and 1000 SNPs. An option of the R
435 package *glmnet* allows the maximum number of
436 variables to be set. This option acts on the intensity of
437 the penalization to validate this constraint.

438 Obviously, this limitation in the number of SNPs
439 led to a decrease in correlation, but this was relatively
440 limited: between 0 and 3·4% depending on the trait
441 and the maximum number of SNPs defined. In comple-
442 ment to this table, Figure 1 presents the mean
443 change in correlation over the 25 traits for the three
444 breeds according to the number of selected SNPs.

445 The breed found to be the most sensitive to the
446 limitation of selected SNP in EN was the Holstein
447 breed, but this is also the breed in which, on average,
448 the largest number of SNPs without pre-selection are
449 retained (17 341 selected SNPs in this situation against
450 11 526 in Normande and 12 939 in Montbéliarde).
451 When the number of selected SNPs was limited to
452 2500, the average absolute loss in correlation over the
453 25 traits ranged from 1 and 1·7%. This average loss in
454 correlation changed to 2·3 and 4·5% with a limit to
455 1000 selected SNPs.

456 4. Discussion

457 As for many previous studies, genomic evaluations
458 with gBLUP and EN substantially improved the
459 quality of prediction of observed DYD in the vali-
460 dation data set compared with pedigree-based BLUP
461 (Hayes *et al.*, 2009; Wolc *et al.*, 2011). Between these
462 two genomic evaluations, gBLUP has the advantage
463 of being conceptually simpler in the sense that there is
464 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 re-
gressed 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, $C\pi$, ...)
limits the number of SNPs with non-null estimated
effects in the model. This property can be an advan-
tage because it alleviates the $p \gg n$ problems, in par-
ticular for smaller breeds. Limiting the number of
SNP effects to estimates becomes important for an
accurate prediction equation.

476 Since this study shows that EN provides better re-
477 sults than gBLUP for most traits in the three breeds
478 studied, we tried to push further the idea of variable
479 selection both in the case of gBLUP and EN by add-
480 ing an SNP pre-selection step based on QTL detec-
481 tion. The resulting correlations between DGV and
482 observed DYD and also the slopes of the regression of
483 observed DYD on DGV were similar to the ones ob-
484 tained using the complete set of SNPs. Moreover,
485 both the EN algorithm and the pre-selection of the
486 SNP led to a reduction of the number of SNPs in-
487 cluded in the prediction equation with a minor effect
488 on the quality of prediction. This procedure seems
489 particularly relevant in the genomic selection context
490 for two reasons:

- 492 • From a genetic standpoint, it is consistent with the
493 assumption that not all SNPs are required to ex-
494 plain the genetic architecture of a given trait. Some
495 of them, with non-significant effects, can still carry
496 genetic information and particularly on genetic

relationships (Habier *et al.*, 2007, 2010*b*). 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 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.
- 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. (2010*a*). Extension of the Bayesian alphabet for genomic

Comment citer ce document :

Croiseau, P., Legarra, A., Guillaume, F., Fritz, S., Baur, A., Colombani, C., Robert Granie, C., Boichard, D., Ducrocq, V. (2011). Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genetics Research*, 93 (6), 409-417. DOI : 10.1017/S0016672311000358

- 616 selection. In '9th World Congress on Genetics Applied to
617 *Livestock Production*', Germany: Leibzig.
- 618 Habier, D., Tetens, J., Seefried, F. R., Lichtner, P. &
619 Thaller, G. (2010b). The impact of genetic re-
620 lationship information on genomic breeding values
621 in German Holstein cattle. *Genetics Selection Evolution*
622 **42**, 5.
- 623 Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J.
624 (2011). Extension of the Bayesian alphabet for genomic
625 selection. *BMC Bioinformatics* **23**, 186.
- 626 Haley, C. S. & Visscher, P. M. (1998). Strategies to utilize
627 marker-quantitative trait loci associations. *Journal of*
628 *Dairy Science* **81**(Suppl. 2), 85–97.
- 629 Harris, B. & Johnson, D. (2010). Genomic predictions for
630 New Zealand dairy bulls and integration with national
631 genetic evaluation. *Journal of Dairy Science* **93**,
632 1243–1252.
- 633 Hayes, B., Bowman, P., Chamberlain, A. & Goddard, M.
634 (2009). Invited review: Genomic selection in dairy cattle:
635 progress and challenges. *Journal of Dairy Science* **92**,
636 433–443.
- 637 Interbull (2011). GEBV Test. [http://www.interbull.org/
638 index.php?option=com_content&view=article&id=80&
639 Itemid=114](http://www.interbull.org/index.php?option=com_content&view=article&id=80&Itemid=114)
- 640 Jensen, J., Mantysaari, E. A., Madsen, P. & Thompson, R.
641 (1996). Residual maximum likelihood estimation of (co)-
642 variance components in multivariate mixed linear models
643 using average information. *Journal of Indian Society for*
644 *Agricultural Statistics* **49**, 215–236.
- 645 Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F.
646 & Fritz, S. (2011). Improved Lasso for genomic selection.
647 *Genetics Research (Cambridge)* **93**, 77–87.
- 648 Liu, Z., Seefried, F., Reinhardt, F., Thaller, G. & Reents, R.
649 (2010). Dairy cattle genetic evaluation enhanced
650 with genomic information. In '9th World Congress on
651 *Genetics Applied to Livestock Production*', Germany:
652 Leibzig.
- 653 Lund, M. S., de Roos, A. P. W., de Vries, A. G., Druet, T.,
654 Ducrocq, V., Fritz, S., *et al.* (2010). Improving genomic
655 prediction by EuroGenomics collaboration. In '9th
656 *World Congress on Genetics Applied to Livestock Pro-
657 duction*', Germany: Leipzig.
- Meuwissen, T. & Goddard, M. (2001). Prediction of identity
658 by descent probabilities from marker-haplotypes.
659 *Genetics Selection Evolution* **33**, 605–634. 660
- Meuwissen, T., Hayes, B. & Goddard, M. (2001). Prediction
661 of total genetic value using genome-wide dense marker
662 maps. *Genetics* **157**, 1819–1829. 663
- Misztal, I., Tsuruta, T., Strabel, T., Auvray, B., Druet, T. &
664 Lee, D. H. (2002). BLUPF90 and related programs
665 (BGF90). *7th World Congress on Genetics Applied to*
666 *Livestock Production*, France: Montpellier. 667
- Mrode, R. A. & Swanson, G. J. T. (2004). Calculating cow
668 and daughter yield deviations and partitioning of genetic
669 evaluations under a random regression model. *Livestock*
670 *Production Science* **86**, 253–260. 671
- Peers, I. (1996). *Statistical Analysis for Education and*
672 *Psychology Researchers*. Washington, DC. 673 AQ6
- Van Raden, P. (2008). Efficient methods to compute geno-
674 mic predictions. *Journal of Dairy Science* **91**, 4414–4423. 675
- Van Raden, P., Van Tassell, C., Wiggans, G., Sonstegard,
676 T., Schnabel, R., Taylor, J., *et al.* (2009). Invited review:
677 reliability of genomic predictions for North American
678 Holstein bulls. *Journal of Dairy Science* **92**, 16–24. 679
- Verbyla, K., Hayes, B., Bowman, P. & Goddard, M. (2009).
680 Accuracy of genomic selection using stochastic search
681 variable selection in Australian Holstein Friesian dairy
682 cattle. *Genetics Research* **91**, 307–311. 683
- Wensch-Dorendorf, M., Yin, T., Swalve, H. H. & König, S.
684 (2011). Optimal strategies for the use of genomic selection
685 in dairy cattle breeding programs. *Journal of Dairy*
686 *Science* **94**, 4140–5151. 687
- Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E. &
688 O'Sullivan, N. P., *et al.* (2011). Breeding value prediction
689 for production traits in layer chickens using pedigree
690 or genomic relationships in a reduced animal model.
691 *Genetics Selection Evolution* **43**, 5. 692
- Zou, H. & Hastie, T. (2003). *Regression Shrinkage and*
693 *Selection via the Elastic Net, with Application to*
694 *Microarrays*. Stanford University: Department of
695 Statistics. 696
- Zou, H. & Hastie, T. (2005). Regularization and variable
697 selection via the Elastic Net. *Royal Statistical Society*
698 *Series B* **67**, 301–320. 699