Improved lasso for genomic selection

Andres Legarra, Christèle Robert-Granié, Pascal Croiseau, François F. Guillaume, Sébastien S. Fritz

Genet. Res., Camb. (2011), **93**, *pp.* 77–87. © Cambridge University Press 2010
doi:10.1017/S0016672310000534

Improved Lasso for genomic selection

ANDRÉS LEGARRA[1]*, CHRISTÈLE ROBERT-GRANIÉ[1], PASCAL CROISEAU[2], FRANÇOIS GUILLAUME[3] AND SÉBASTIEN FRITZ[4]
[1] *INRA, UR 631 SAGA, F-31326 Castanet-Tolosan, France*
[2] *INRA, UMR1313 GABI, F-78352 Jouy en Josas, France*
[3] *Institut de l'Elevage, F-75595 Paris, France*
[4] *UNCEIA, F-75595 Paris, France*

(*Received 26 April 2010 and in revised form 26 October 2010; first published online 14 December 2010*)

Summary

Empirical experience with genomic selection in dairy cattle suggests that the distribution of the effects of single nucleotide polymorphisms (SNPs) might be far from normality for some traits. An alternative, avoiding the use of arbitrary prior information, is the Bayesian Lasso (BL). Regular BL uses a common variance parameter for residual and SNP effects (BL1Var). We propose here a BL with different residual and SNP effect variances (BL2Var), equivalent to the original Lasso formulation. The λ parameter in Lasso is related to genetic variation in the population. We also suggest precomputing individual variances of SNP effects by BL2Var, to be later used in a linear mixed model (HetVar-GBLUP). Models were tested in a cross-validation design including 1756 Holstein and 678 Montbéliarde French bulls, with 1216 and 451 bulls used as training data; 51 325 and 49 625 polymorphic SNP were used. Milk production traits were tested. Other methods tested included linear mixed models using variances inferred from pedigree estimates or integrated out from the data. Estimates of genetic variation in the population were close to pedigree estimates in BL2Var but not in BL1Var. BL1Var shrank breeding values too little because of the common variance. BL2Var was the most accurate method for prediction and accommodated well major genes, in particular for fat percentage. BL1Var was the least accurate. HetVar-GBLUP was almost as accurate as BL2Var and allows for simple computations and extensions.

1. Introduction

Genome-wide strategies for genetic evaluation can be roughly divided into BLUP-like methods (postulating normal distribution of single nucleotide polymorphism (SNP) effects) and variable selection methods using more sophisticated distributions. The seminal paper of Meuwissen *et al.* (2001) already made this distinction, by creating BLUP and Bayes (A, B) methods. In the first group, marker effects are posited normal distributions with zero mean and identical variance for all markers. This results in nice properties, like simplicity of computations and, in particular, an equivalent model using a 'genomic' relationship matrix (Van Raden, 2008). The latter can be meshed with additive relationship matrices and extended to the whole pedigree (Legarra *et al.*, 2009). Further, under mild assumptions, equivalences exist between genetic variances in an additive relationship model and marker variances (Gianola *et al.*, 2009).

However, at least for some traits, it has been shown that departures of SNP effects from normality exist. This results in (and can be seen by) higher accuracy of methods with more sophisticated *a priori* distributions of the marker effects, like BayesA or non-linear regression (Hayes *et al.*, 2009; Van Raden *et al.*, 2009*b*). These methods are sometimes called 'Bayesian methods' (Lund *et al.*, 2009). This is inappropriate, because BLUP is also a Bayesian method, and also because they have frequentist counterparts (e.g. Usai *et al.*, 2009). Thus, we will call them 'variable selection methods' because most of them assume values of most SNP effects to be zero or close to zero. Another

* Corresponding author. INRA, UR 631 SAGA, BP52627, F-31326 Castanet Tolosan, France. Tel: +33561285182. Fax: +33561285353. e-mail: andres.legarra@toulouse.inra.fr

property of variable selection methods, shown in simulations, is that these methods have better properties in the long run, that is, estimates of SNP effects are stable after several generations (Habier *et al*., 2007). In addition, small (and possibly much cheaper) subsets of markers chosen by variable selection methods have been shown to be of acceptable accuracy (Weigel *et al*., 2009). Thus, variable selection methods are being heavily used in simulations (Meuwissen *et al*., 2001; Calus *et al*., 2008; Kizilkaya *et al*., 2010) and in real data analysis (Hayes *et al*., 2009; Van Raden *et al*., 2009*b*).

Most variable-selection methods nevertheless require *a priori* distributions or tuning parameters. These include the number of SNPs *a priori* in the model and its variance (Meuwissen *et al*., 2001; Verbyla *et al*., 2009); or the ratio of variances of SNPs 'in' or 'out' (Calus *et al*., 2008; Verbyla *et al*., 2009); or the *a priori* variance of SNP effects (Kizilkaya *et al*., 2010). No clear clue, based on biological knowledge, exists about these *a priori* distributions. This complicates their practical application.

The Lasso (least absolute shrinkage and selection operator; Tibshirani, 1996) combines variable selection and shrinkage. Its Bayesian counterpart, the Bayesian Lasso (Park & Casella, 2008) provides a more natural interpretation in terms of *a priori* distributions. It is well known that, generally, conditional expectations are optimal for selection (Gianola & Fernando, 1986). These can be obtained through the Bayesian Lasso but not the regular Lasso. Also, in particular, Bayesian Lasso provides a fully parametric model with a simple Gibbs sampler implementation, as well as an EM algorithm for the estimation of the 'sharpness' parameter λ, needing little (or no) prior information. Thus, Bayesian Lasso is an attractive candidate for genomic selection because of its simplicity, computational ease and little (or no) need to postulate prior information. Further, the exponential distribution of the Lasso is thought to reflect reasonably well the nature of quantitative trait locus (QTL) effects (Goddard, 2008).

The (Bayesian or not) Lasso has been used in an animal breeding context (de los Campos *et al*., 2009; Usai *et al*., 2009; Weigel *et al*., 2009), albeit a broad comparison with related methods using several traits and a real, large data set has not yet been published. In addition, we find that the particular case of Park and Casella's Bayesian Lasso includes a common variance term for modelling both residual terms and effects in the model, instead of two different variances. We find that this parameterization is not optimal. The purpose of this paper is thus manifold. First, to propose and compare a different, more general, model for the Bayesian Lasso, which in fact is equivalent to Tibshirani's (1996) original Lasso. This model implies different variances for residual terms and for SNP effects. Second, an alternative linear model for genomic prediction will be presented and tested empirically; in this model individual SNP variances are inferred via the Bayesian Lasso first and then used in a BLUP-like estimator. Third, we compare the performance of these models with a more standard 'genomic BLUP' (GBLUP) either fixing the variance for the marker effects from pedigree estimates applying a rough equivalence (Gianola *et al*., 2009), or inferring and integrating it out from the data via the Gibbs Sampler.

2. Parameterization of the Bayesian Lasso

The base of the Lasso is a typical linear model of the form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}; \quad \mathbf{e}|\sigma^2 \sim \mathrm{MVN}(\mathbf{0}, \mathrm{I}\sigma^2),$$

where **b** are fixed effects (e.g. an average mean), **a** are, in this work, SNP effects and MVN stands for multivariate normal. Originality of Lasso is in modelling effects **a**. In the Bayesian Lasso, the distribution of (a single) SNP effect a is modelled as

$$p(a|\sigma^2, \lambda) = \frac{\lambda}{2\sigma} \exp\left(\frac{-\lambda|a|}{\sigma}\right).$$

In the classical Lasso (Tibshirani, 1996) this distribution is actually $p(a|\sigma^2, \lambda) = (\lambda/2)\exp(-\lambda|a|)$; however, Tibshirani (1996) assumes that incidence matrix **Z** has been standardized, which is not assumed here or in the Bayesian Lasso.

Finally, in Bayesian Lasso the variance of a is $\mathrm{Var}(a) = 2\sigma^2/\lambda^2$.

Intriguingly, as shown in the expressions above, in Bayesian Lasso applications in genomic selection (e.g. de los Campos *et al*., 2009; Weigel *et al*., 2009) the variance σ^2 has been used at the same time to model the residual term as well as the distribution of the SNP effects. However, we do expect the distribution of SNP effects not to be related to unobservable, unaccounted (residual) effects that can, for example, vary from site to site for the same individuals. Assume, for instance, a crop trial design in which some varieties are tested. Each variety can be tested 1 or 100 times. If the phenotype to be analysed is the average yield of the variety, everything else being equal, it is expected that the residual variation is divided by 10 in the second option, but not the variation across SNP effects. Another example is as follows. Assume that a set of dairy bulls is tested in two different locations, the second with less frequent milk recording. The second location will show higher residual variation for milk yield, whereas genetic variation in the bulls will be the same.

The implementation of the Bayesian Lasso in Park & Casella (2008) does not take this into account.

A more general implementation would split the sources of variation in purely residual (σ^2_e) and variation due to SNPs (σ^2_a), by rewriting the model as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e};$$
$$\mathbf{a}|\lambda, \sigma^2_a \sim \prod_i \frac{\lambda}{2\sigma_a} \exp\left(\frac{-\lambda|a_i|}{\sigma_a}\right); \quad \mathbf{e}|\sigma^2_e \sim \text{MVN}(\mathbf{0}, \text{I}\sigma^2_e).$$

However, this is clearly equivalent to

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e};$$
$$\mathbf{a}|\lambda \sim \prod_i \frac{\lambda}{2} \exp(-\lambda|a_i|); \quad \mathbf{e}|\sigma^2_e \sim \text{MVN}(\mathbf{0}, \text{I}\sigma^2_e),$$

which is the original form of Tibshirani's (1996) original Lasso, because only the ratio λ/σ_a is used and thus they cannot be estimated separately. Equivalently, the model could be written in terms of σ^2_a by dropping λ.

In the original Lasso, cross-validation is used for the estimation of λ (Usai *et al*., 2009). Park & Casella (2008) proposed a fully parametric implementation by computing a posterior distribution (using the Gibbs sampler) or an empirical Bayes estimation by marginal maximum likelihood by a Monte Carlo Expectation–Maximization (MCEM) algorithm. The latter avoids the problem of choosing a hyperprior for λ, pointed out by both Park & Casella (2008) and de los Campos *et al*. (2009).

The hierarchical formulation of Lasso shown above includes explicitly two sources of variation and is thus akin to classical models in quantitative genetics and genetic evaluation (Henderson, 1984; Falconer & Mackay, 1996) where variation is split into environmental and genetic variances. The shape of the distribution of SNP effects is determined by λ, which effectively determines the variance of SNP effects by using $\text{Var}(a) = 2/\lambda^2$. Thus, λ plays the same role as the inverse of a standard deviation in normal models. This does not seem to have been recognized by previous scholars (de los Campos *et al*., 2009; Usai *et al*., 2009).

Applying the same logic as in Gianola *et al*. (2009), and in ideal conditions, it is possible to establish a rough equivalence between genetic variance in a population (σ^2_u; usually estimated by an additive, relationship-based model) and the variance of SNP effects:

$$\text{Var}(a) = \frac{2}{\lambda^2} = \frac{\sigma^2_u}{2\sum_i p_i(1-p_i)},$$

where p_i is the allelic frequency at the ith marker.

3. Estimation and cross-validation study

(i) *Data*

Two sets of bulls from French dairy cattle populations have been analysed from, respectively, Holstein (1756 bulls) and Montbéliarde (678 bulls) breeds. Bulls were genotyped with the Illumina Bovine SNP50 BeadChip. Markers were discarded based on low call rate, lack of positioning in the genome, or very high Mendelian inconsistency rate. No minor allele frequency threshold was imposed. Finally, 51 325 and 49 625 polymorphic SNP were, respectively, used in each breed. A cross-validation approach was used where 1216 and 451 bulls were taken as the training data and the rest as validation data. Bulls in the validation data set were, roughly, bulls being tested in 2004 and 2005, and younger than the training bulls. All parameter estimation in this work was carried out on the training population.

Data for training (**y** in the model) were daughter yield deviations (DYDs; Van Raden & Wiggans, 1991) as computed with data available in 2004; data for validation were DYDs from data available in 2009. Thus, the validation mimics well a real scenario. To account for different accuracies in the estimation of DYDs, these were weighted by their prediction error variances (in terms of number of equivalent daughters) as estimated from regular genetic evaluation. This will be explained in more detail later.

Traits analysed were milk, fat and protein yields (MY, FY and PY) and fat and protein percentages (FP and PP). Several models were used. The estimation was mostly made by Bayesian methods using Markov Chain Monte Carlo (MCMC) as well as, for certain cases, a marginal maximum likelihood by an MCEM algorithm, as suggested by Park & Casella (2008) to avoid the use of a hyperprior for λ. An example of marginal maximum likelihood in the genetics literature is the REML estimator of variance components (Patterson & Thompson, 1971). The models used to analyse the data sets are described next.

(ii) *Bayesian Lasso with genetic and residual variances* (*Bayesian lasso with two variances; BL2Var*)

The model is as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Za} + \mathbf{e};$$
$$\mathbf{a}|\lambda \sim \prod_i \frac{\lambda}{2} \exp(-\lambda|a_i|); \quad \mathbf{e}|\sigma^2_e \sim \text{MVN}(\mathbf{0}, \mathbf{F}\sigma^2_e),$$

where **y** contains twice the DYDs for each bull, μ is a general mean, **Z** is an incidence matrix of SNP effects **a**, **e** is a vector of residuals and **F** is a diagonal matrix that contains, in the diagonal, the inverse of the number of equivalent daughters for each DYD. The parameterization of SNP effects is as in Van Raden (2008): $-2p_i$, $1-2p_i$, and $2-2p_i$ for the genotypes 00, 01 and 11, where p_i is the allelic frequency of '1'. In this way, assuming Hardy–Weinberg equilibrium,

SNP genotypic effects are substitution effects with average effect of 0 in the population (Falconer & Mackay, 1996), which is one of the conditions for the expression of the genetic variation in the population as $2\sum_i p_i(1-p_i)\mathrm{Var}(a)$ (Gianola *et al.*, 2009). This parameterization also results in slightly better predictive abilities compared to other ones such as -1, 0, 1 for the 00, 01 and 11 genotypes (data not shown).

The prior distribution for σ^2_e was an inverted chi-square distributions with 4 degrees of freedom and expectations equal to the value used in regular genetic evaluation for σ^2_e. Prior for λ was deliberately vague, being uniform between 0 and 1 000 000.

In practice, the model above was transformed in an equivalent model, yielding the same solutions, as follows:

$$\mathbf{y}^* = \mathbf{F}^{-1/2}\mathbf{y}; \quad \mathbf{x} = \mathbf{F}^{-1/2}\mathbf{1}; \quad \mathbf{Z}^* = \mathbf{F}^{-1/2}\mathbf{Z};$$
$$\mathbf{e}^* = \mathbf{F}^{-1/2}\mathbf{e},$$

which amounts to multiply each row of $\mathbf{1}$ and $\mathbf{Z}$ by the square root of the number of equivalent daughters, so that

$$\mathbf{y}^* = \mathbf{x}\mu + \mathbf{Z}^*\mathbf{a} + \mathbf{e}^*,$$
$$\mathbf{a}|\lambda \sim \prod_i \frac{\lambda}{2}\exp(-\lambda|a_i|); \quad \mathbf{e}^*|\sigma^2_e \sim \mathrm{MVN}(\mathbf{0}, \mathrm{I}\sigma^2_e),$$

which simplifies the computations.

A Gibbs sampler was implemented as in Park & Casella (2008) or de los Campos *et al.* (2009), via the introduction of additional (augmented) variables τ^2_i, which can be seen as variance components for each SNP effect. The Gibbs sampler with residual update (Legarra & Misztal, 2008) was used to speed up sampling of location parameters μ and $\mathbf{a}$. The full conditional posterior distributions are as follows (the symbol $\tilde{b}$ indicates the current state of variable b):

$$\mu|\text{else} \propto N(\mathbf{x}'(\mathbf{y}^* - \mathbf{Z}^*\tilde{\mathbf{a}})/\mathbf{x}'\mathbf{x}, 1/\mathbf{x}'\mathbf{x}\tilde{\sigma}_e^{-2}),$$

or

$$\mu|\text{else} \propto N(\mathbf{x}'(\tilde{\mathbf{e}} + \mathbf{x}\tilde{\mu})/\mathbf{x}'\mathbf{x}, 1/\mathbf{x}'\mathbf{x}\sigma_e^{-2}),$$

$$a_i|\text{else} \propto N(\mathbf{z}_i^{*\prime}(\mathbf{y}^* - \mathbf{x}\tilde{\mu} - \mathbf{Z}^*\tilde{\mathbf{a}}_{-i})\tilde{\sigma}_e^{-2}/\text{lhs}_i, 1/\text{lhs}_i),$$

or

$$a_i|\text{else} \propto N(\mathbf{z}_i^{*\prime}(\tilde{\mathbf{e}} + z_i^*\tilde{a}_i)\tilde{\sigma}_e^{-2}/\text{lhs}_i, 1/\text{lhs}_i),$$

where $\text{lhs}_i = \mathbf{z}_i^{*\prime}\mathbf{z}_i^*\tilde{\sigma}_e^{-2} + \tilde{\tau}_i^{-2}$, z_i^* is the row of $\mathbf{Z}^*$ corresponding to the ith effect and $\mathbf{a}_{-i}$ indicates all $\mathbf{a}$ variables except for a_i. Further,

$$\tau_i^{-2}|\text{else} \propto \mathrm{IG}\left(\sqrt{\frac{\tilde{\lambda}^2}{\tilde{a}_i^2}}, \lambda^2\right),$$

where IG stands for inverted Gaussian,

$$\lambda^2|\text{else} \propto G(s = nsnp, \text{sc} = 2/\textstyle\sum\tilde{\tau}_i^2)$$

bounded between 0 and 1 000 000, and where G is a gamma distribution with shape 's' and scale 'sc' and '$nsnp$' is the number of $\mathbf{a}$ effects. Finally,

$$\sigma^2_e|\text{else} \propto \chi^{-2}(\tilde{\mathbf{e}}^{*\prime}\tilde{\mathbf{e}}^* + S^2_e, 4 + n\text{data}),$$

where S^2_e is the scale of the *a priori* distribution of the residual variance and $n$data is the number of records in $\mathbf{y}$. For the inverted Gaussian distribution, we used the algorithm of Michael *et al.* (1976) with a minor modification: extracting the largest root of the quadratic to avoid numerical cancellation.

For the MCEM estimation of λ (BL2Var-EM), the iterations proceed as above but sampling of λ is substituted by an updated estimate

$$\hat{\lambda}^2 = \frac{2\,nsnp}{\sum_i E(\tilde{\tau}_i^2|\tilde{\mathbf{y}}^*)},$$

where $E(\tilde{\tau}_i^2|\tilde{\mathbf{y}}^*)$ are obtained by MonteCarlo using the previous estimate of λ. In our case and after experimentation with one trait, the number of iterations to get $E(\tilde{\tau}_i^2|\tilde{\mathbf{y}}^*)$ was reduced to just one. This seems to be possible because the very large number (51 325) of $\tilde{\tau}_i^2$ variables included provides a reasonable estimate. At convergence, the last 100 samples of λ were averaged to obtain a MonteCarlo error-free estimate (as suggested by Park & Casella, 2008).

(iii) *Bayesian Lasso with one variance* (*BL1Var*)

The model by Park & Casella (2008) and de los Campos *et al.* (2009) postulates a one-variance component linked to *a priori* variation in both residual and SNP effects, and thus:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \mathbf{e};$$
$$\mathbf{a}|\lambda, \sigma^2_e \sim \prod_i \frac{\lambda}{2\sigma_e}\exp\left(\frac{-\lambda|a_i|}{\sigma_e}\right); \quad \mathbf{e}|\sigma^2_e \sim \mathrm{MVN}(\mathbf{0}, \mathbf{F}\sigma^2_e)$$

The conditional distributions are as above, with the following modifications:

$$lhs_i = \mathbf{z}_i^{*\prime}\mathbf{z}_i^*\tilde{\sigma}_e^{-2} + \tilde{\tau}_i^{-2}\tilde{\sigma}_e^{-2}$$

$$\tau_i^{-2}|\text{else} \propto \mathrm{IG}\left(\sqrt{\frac{\tilde{\lambda}^2\sigma^2_e}{\tilde{a}_i^2}}, \lambda^2\right)$$

and

$$\sigma^2_e|\text{else} \propto \chi^{-2}(\tilde{\mathbf{a}}'\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{a}} + \tilde{\mathbf{e}}^{*\prime}\tilde{\mathbf{e}}^* + S^2_e, 4 + nsnp + n\text{data}),$$

where $\mathbf{D}$ is a diagonal matrix with $\tau_i^2\sigma^2_e$ in the (i,i) position. This conditional distribution shows well

that SNP effects are in practice considered as pseudo-residuals in the one-variance Bayesian Lasso.

(iv) *Bayesian mixed model with unknown genetic and residual variances* (*MCMC-GBLUP*)

This model is similar to the 'BLUP' model in Meuwissen *et al*. (2001), although the variance components are not fixed *a priori*. Instead, they are estimated in the model as in Legarra *et al*. (2008):

$$\mathbf{y}=\mathbf{1}\mu+\mathbf{Z}\mathbf{a}+\mathbf{e},\quad \mathbf{a}|\sigma_a^2\sim \mathrm{MVN}\left(0,\mathbf{I}\sigma_a^2\right);\quad \mathbf{e}|\sigma_e^2\sim \mathrm{MVN}\left(\mathbf{0},\mathbf{F}\sigma_e^2\right).$$

The prior distribution for σ_e^2 is as in the Bayesian lasso with two variances; the prior distribution for σ_a^2 was a chi-squared distribution with 4 degrees of freedom and expectation equal to $\sigma_u^2/2\sum_i p_i(1-p_i)$; σ_u^2 being the genetic variance component used in genetic evaluation. The Gibbs sampler for this model has been extensively described (e.g. Sorensen & Gianola, 2002).

(v) *Bayesian mixed model with known genetic and residual variances* (*GBLUP*)

This model is as the previous one, except that variance components were assumed to be known with certainty and inferred from values used in current genetic evaluation, as for the priors in MCMC-GBLUP. To estimate solutions for μ and $\mathbf{a}$, Henderson's (1984) mixed model equations were used, which were solved by preconditioned conjugated gradients as described in Legarra & Misztal (2008).

(vi) *Bayesian mixed model with heterogeneous genetic variances* (*Het-GBLUP*)

This model assumes that components of overall genetic variation (λ and σ_e^2) are known with certainty but allows for heterogeneous variances of SNP effects, which are τ_i^2 for the ith SNP. In order to accommodate heterogeneous variances in a linear estimator, these have to be previously known. Thus, we followed a three-step procedure. First, λ and σ_e^2 were estimated as in the Bayesian Lasso with two variances. Second, estimates $\hat{\tau}_i^2$ of τ_i^2 were computed by a Gibbs sampler (as the one in the Bayesian Lasso with two variances) with λ and σ_e^2 fixed to their estimated values. Finally, a diagonal matrix $\mathbf{D}$ was formed to describe the heterogeneous variance, with $\hat{\tau}_i^2$ in the (i,i) position. Thus, the model becomes

$$\mathbf{y}=\mathbf{1}\mu+\mathbf{Z}\mathbf{a}+\mathbf{e},\quad \mathbf{a}|\lambda,\tau\sim \mathrm{MVN}(0,\mathbf{D});\quad \mathbf{e}|\sigma_e^2\sim \mathrm{MVN}(\mathbf{0},\mathbf{F}\sigma_e^2),$$

which is solvable by Henderson's mixed model equations as above.

All models above were fit to the five traits for the Holstein breed; for the Montbéliarde, only the Bayesian Lasso with two variances and GBLUP were fit. The MCEM was run for 50 000 iterations, with final convergence; the MCMC were run for 50 000 iterations with 25 000 of burn-in after which solutions for all unknowns were estimated by their posterior means. Self-made programs were written in Fortran95.

Different parameters were estimated. In addition to σ_e^2 and λ, a rough equivalent of the classical, pedigree-based genetic variance (σ_u^2) was estimated for each model (except for GBLUP where it is supposed fixed). For the Bayesian Lasso with two variances, this is $\sigma_u^2=2\sum_i p_i(1-p_i)\mathrm{Var}(a)=2\sum_i p_i(1-p_i)(2/\lambda^2)$. For the Bayesian Lasso with one variance, this is $\sigma_u^2=2\sum_i p_i(1-p_i)(2\sigma_e^2/\lambda^2)$. For the MCMC-GBLUP, this is $\sigma_u^2=2\sum_i p_i(1-p_i)\sigma_a^2$. A pedigree-based estimate of σ_u^2 was obtained by REML for the Holstein breed using REMLF90 (Misztal *et al*., 2002).

(vii) *Cross-validation*

Predictions (genomic estimated breeding values (GEBVs)) for the validation data test were computed as $\hat{\mathbf{u}}=\mathbf{Z}\hat{\mathbf{a}}$ for the different models, and compared with 2009 progeny test-based DYDs. Predictive ability was measured as the correlation between both. The correlation was weighted by the number of equivalent daughters in 2009 DYD data. The formula for the weighted Pearson product moment correlation coefficient is as follows (e.g. Peers, 1996):

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

where $\bar{x}_w=\sum w_i x_i/\sum w_i$ and $\bar{y}_w=\sum w_i y_i/\sum w_i$ and w are the weights for each data point. Cross-validation results for BL2Var-EM approach were essentially the same as BL2Var (correlation among EBVs was higher than 0·99) and are therefore not shown.

4. Results

(i) *Estimates of parameters*

Tables 1 (for Holstein) and 2 (for Montbéliarde) show estimates of λ parameters. Estimates are generally accurate as shown by their standard errors. Estimates from BL1Var are very similar across traits, which does not occur for BL2Var. On the other hand, estimates by full Bayesian inference (BL2Var) or marginal maximum likelihood (BL2Var – EM) are virtually identical in both Holstein (Table 1) and Montbéliarde (Table 2). Also, estimates in Holstein and in Montbéliarde are quite similar for the same traits.

Estimates of genetic variation in the population (σ_u^2) are shown in Tables 2 and 3. Estimates from

Table 1. *Estimates of 'sharpness' parameter* λ (± *SE*) *in Holstein*

Trait	BL1Var	BL2Var	BL2Var-EM
MY	17·06 ± 0·05	0·26 ± 0·01	0·26
FY	20·60 ± 0·05	6·56 ± 0·20	6·51
PY	19·92 ± 0·05	8·43 ± 0·23	8·41
FP	15·06 ± 0·05	57·20 ± 1·64	55·61
PP	16·32 ± 0·06	135·82 ± 4·40	134·01

MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percentage; PP, protein percentage.

BL2Var and MCMC-GBLUP (which are remarkably close) are similar to pedigree-based estimates by REML or to values currently used, so that they make clear biological sense and are understandable in terms of genetic variation. Differences of the REML estimate from current values can be explained by the incompleteness of the DYD data; differences among pedigree-based and SNP-based models can be due to this reason, but also to the different assumptions of both models, which make them not fully comparable, as discussed below. Estimates from BL1Var are clearly different from any other estimate and are not reliable as estimates of the genetic variation in the population.

(ii) *Empirical accuracies*

Tables 4 (for Holstein) and 2 (for Montbéliarde) show empirical accuracies of predicted GEBVs. Possibly due to the relatively small data set, accuracies are lower than some previously reported estimates (Hayes *et al*., 2009; Van Raden *et al*., 2009*b*) although comparable to accuracies in a small study of similar size (Luan *et al*., 2009).

Accuracies are almost systematically highest for BL2Var, in particular for traits controlled by major genes as DGAT1 (Grisart *et al*., 2002) (FP). GBLUP and MCMC-GBLUP perform similarly (with 10% less accuracy than BL2Var for FP). The two-step approach HetVar-GBLUP reaches similar accuracies to BL2Var, with only 2% less accuracy for FP. As for BL1Var, its performance is the poorest almost systematically, although the difference is minimal for FY and PY. Thus, results from previous users of Bayesian Lasso (e.g. Weigel *et al*., 2009) might underestimate the true potential of Lasso. This is possibly trait dependent.

Results in Montbéliarde show no major difference between BL2Var and GBLUP. This is partly due to the fact that DGAT1 has an extremely low minor allele frequency (4%) in Montbéliarde and causes almost no genetic variation in the population (Gautier *et al*., 2007).

In general, models with the same accuracy are almost identical (correlations among EBVs higher than 0·99), as shown in Table 6 for MY and FP. This implies that errors in estimation of EBVs are very similar across methods. Models with different accuracies show, obviously, lower correlations. For example, correlations of EBVs estimated with BL2Var with those estimated by BL1Var and MCMC-GBLUP for FP are, respectively, 0·73 and 0·92.

Table 5 shows regression coefficients of 2DYDs on GEBVs. These regression coefficients should ideally be 1, implying that the predicted has the same variance as the true value. This is relevant for the comparison of estimated breeding values across generations. Methods give often inflated variances of GEBVs ($b < 1$) for yields. For contents, they oscillate around 1; the reason is that most genetic variation is well captured due to large QTL effects. However, BL1Var results in inflation for all traits because it does not shrink estimators enough. Even in the absence of genomic information, predictions of young bulls by parent average are known to be biased (Van Raden *et al*., 2009*a*). One explanation for the generally low value for b is pre-selection of validation bulls (Mäntysaari *et al*., 2010); according to our information, this is actually not the case in the French industry. Another more likely explanation is lack of enough information, because in this work dams 'information is not added to the genomic predictions. A combined index was suggested by Van Raden *et al*. (2009*b*).

Figure 1 shows the estimates by HetVar-GBLUP of SNP effects for FP in chromosomes 13 (representative of the rest of the genome) and 14 in a log-10 scale. At the beginning of chromosome 14, the effect of DGAT1 can be appreciated, presenting a sharp peak even in the logarithmic scale of the plot. Peaks of this size cannot be observed elsewhere in the genome. Also, it can be observed that whereas most of the effect of the markers range between 10^{-2} and 10^{-4}, a few have very low values of about 10^{-8}; this corresponds to markers whose effects are 'almost nullified' by the estimate. As pointed out by Usai *et al*. (2009), in the original Lasso a joint mode is estimated and most markers are expected to have values of exactly zero; whereas in Bayesian Lasso, posterior means are estimated, possibly with small values but not zero. Posterior means are optimal for selection (Gianola & Fernando, 1986; Goddard, 2008).

5. Discussion

(i) *Sense of hyperparameters in Bayesian Lasso and genetic variation in the population*

Little has been discussed on estimates of λ in Bayesian Lasso for genomic selection. In BL1Var, values are

Table 2. *Results in Montbéliarde: estimates* ($\pm$ *SE*) *of* '*sharpness*' *parameter* λ, *of population genetic variance* σ^2_u *and accuracies r* (*correlations between GEBVs and 2DYDs in the validation data set*)

	BL2Var-EM		BL2Var		
Trait	λ	λ	σ^2_u	r	GBLUP r
MY[a]	0·26	0·26 $\pm$ 0·01	412 $\pm$ 36	0·36	0·35
FY	6·79	6·80 $\pm$ 0·30	618 $\pm$ 57	0·46	0·46
PY	8·39	8·43 $\pm$ 0·38	402 $\pm$ 35	0·41	0·41
FP	74·73	74·11 $\pm$ 2·59	5·18 $\pm$ 0·36	0·35	0·34
PP	144·68	143·63 $\pm$ 5·75	1·40 $\pm$ 0·11	0·40	0·41

[a] Divided by 1000.
MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percentage; PP, protein percentage.

Table 3. *Estimates of population genetic variance* σ^2_u ($\pm$ *SE*) *in Holstein*

Trait	BL1Var	BL2Var	MCMC-GBLUP	Pedigree REML	Current values[a]
MY[b]	1260 $\pm$ 50	448 $\pm$ 27	451 $\pm$ 26	570	635
FY	1876 $\pm$ 84	710 $\pm$ 44	710 $\pm$ 39	893	973
PY	1127 $\pm$ 50	429 $\pm$ 24	428 $\pm$ 20	473	520
FP	27·6 $\pm$ 1·09	9·32 $\pm$ 0·54	11·60 $\pm$ 0·60	14·90	8·80
PP	5·51 $\pm$ 0·03	1·66 $\pm$ 0·10	1·60 $\pm$ 0·12	2·56	2·19

[a] As used in regular genetic evaluation.
[b] Divided by 1000.
MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percentage; PP, protein percentage.

Table 4. *Accuracies: correlations between GEBVs and 2DYDs in the validation data set, in Holstein*

Trait	BL1Var	BL2Var	GBLUP	MCMC-GBLUP	HetVar-GBLUP
MY	0·28	0·41	0·42	0·40	0·41
FY	0·35	0·37	0·34	0·37	0·36
PY	0·27	0·30	0·31	0·30	0·30
FP	0·53	0·73	0·59	0·61	0·71
PP	0·36	0·48	0·44	0·46	0·47

MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percentage; PP, protein percentage.

Table 5. *Regression coefficients b of 2DYDs on GEBVs in the validation data set, in Holstein*

Trait	BL1Var	BL2Var	GBLUP	MCMC-GBLUP	HetVar-GBLUP
MY	0·25	0·67	0·59	0·66	0·67
FY	0·35	0·80	0·65	0·78	0·77
PY	0·17	0·42	0·41	0·43	0·43
FP	0·50	1·18	0·97	1·11	1·13
PP	0·35	1·10	0·83	1·10	0·99

MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percentage; PP, protein percentage.

Table 6. *Correlation among GEBVs in the validation data set predicted by various methods for milk yield (above diagonal) and fat percentage (below diagonal), in Holstein*

Trait	BL1Var	BL2Var	BL2Var-EM	GBLUP	MCMC-GBLUP	HetVar-GBLUP
BL1Var		0·61	0·61	0·69	0·62	0·56
BL2Var	0·73		1·00	0·96	1·00	0·96
BL2Var-EM	0·73	1·00		0·96	1·00	0·96
GBLUP	0·70	0·90	0·89		0·96	0·92
MCMC-GBLUP	0·70	0·92	0·91	0·98		0·96
HetVar-GBLUP	0·74	1·00	1·00	0·92	0·94	

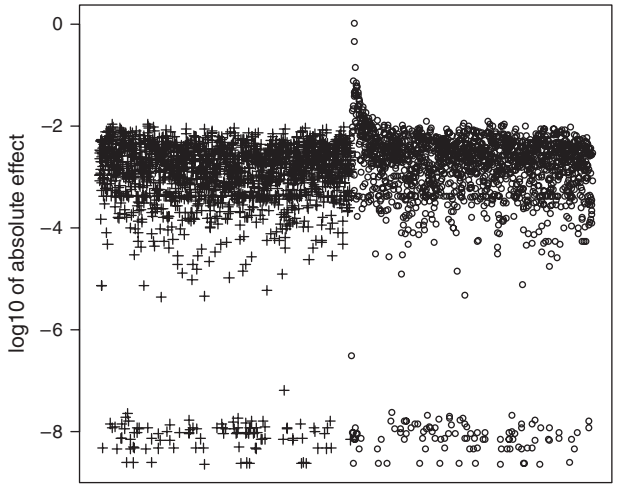

Fig. 1. Estimated effects of SNP loci for FP in Holstein by the HetVar-GBLUP method, for chromosomes 13 (crosses) and 14 (rounds).

similar across traits. This is possibly a by-product of modelling residual and SNP effects with a common σ^2 parameter. However, in BL2Var, λ differs from trait to trait. Further, we provide an interpretation of the meaning of λ in a quantitative genetics context: it is a measure of the genetic variation, as in, for example,

$$\text{BL2Var: } \sigma^2_u = 2\sum_i p_i(1-p_i)\frac{2}{\lambda^2}.$$

This estimate of genetic variation in the population is appropriate for an ideal population in Hardy–Weinberg and linkage equilibrium (Gianola *et al.*, 2009). Yet, in spite of this assumption, estimates of σ^2_u are close to pedigree-based estimates and currently used estimates in the overall population for both BL2Var and MCMC-GBLUP, whereas estimates of σ^2_u in BL1Var do not agree well and make no biological sense. Thus, BL2Var has the advantage of providing biologically reasonable parameters. We recall that pedigree estimates of genetic variation are also ideal, assuming, for instance, unrelated base individuals; thus, both cannot be exactly compared.

(ii) *Predictive ability and use of prior information*

Predictive ability is optimal for BL2Var almost systematically. This agrees with Van Raden *et al.* (2009*b*) who found better predictions for non-linear than for linear equations (GBLUP) for these traits. However, Luan *et al.* (2009) found similar or better results for GBLUP than for non-linear (Mixture and BayesB) methods. For other traits (e.g. fertility) both Van Raden *et al.* (2009*b*) and Hayes *et al.* (2009) found that GBLUP performed better than non-linear methods. In general, for large data sets, the difference seems to be negligible for most traits in dairy cattle except for FP and PP (Van Raden *et al.*, 2009*b*).

A possible explanation for the superiority of BLVar2 over GBLUP is that we did use very little prior information or tuning parameters, extracting most information from data. This is a very important issue in genomic selection nowadays. As extensively shown by simulations (e.g., Meuwissen *et al.*, 2001), the use of the correct – biological – *a priori* information results in better predictive abilities. However, current state of knowledge on QTL action and location does not allow the construction of this prior information, which is replaced by somehow controversial (e.g. Gianola *et al.*, 2009; Hill, 2010) figures or deductions from population genetics theory. This prior information includes the number of QTLs for non-linear regression or BayesB (Meuwissen *et al.*, 2001; Van Raden, 2008); or the *a priori* variance of 'true' respect to 'false' SNP effects (Calus *et al.*, 2008); or, yet, the *a priori* variance of SNP effects in BayesA and BayesB (Meuwissen *et al.*, 2001). Some of this information does not vanish asymptotically as data cumulate (Gianola *et al.*, 2009).

Another option is the 'trial and error' of several 'priors' (i.e. Usai *et al.*, 2009) to find the best predictive ability. This is not real Bayesian (or parametric) inference. In fact, this is an estimation of parameters by trial and error, like the cross-validation methods used in non-parametric inference. This is indeed a legitimate strategy, albeit in practice it is hard to ascertain if all the parametric space of the prior has been covered or how to conceive the different priors. Another problem is that inference depends on the

constitution of the validation data set, a difficult problem in an animal breeding context where data are correlated and selected.

In hierarchical Bayesian models, prior is established in high-order hyperparameters (e.g. variance components or λ), so that the influence of the prior vanishes asymptotically. This is used for example in MCMC-GBLUP (in this work) or in Bayesian Lasso (Park & Casella, 2008; de los Campos *et al.*, 2009; Weigel *et al.*, 2009). This is routinely done for the estimation of genetic parameters via the Gibbs Sampler; the influence of the prior information is negligible for reasonably large data sets (Van Tassell *et al.*, 1996). There is little experience, though, on the practical influence of the prior information on genomic selection. de los Campos (2009) found this influence to exist in estimates of λ, but not really on estimates of SNP effects. Indeed, priors for λ were difficult to conceive, because no natural interpretation on this parameter was recognized. From this work, a reasonable (but not exact) guess of λ is $\lambda^2 = 2\sum_i p_i(1-p_i)(2/\sigma^2_u)$. At any rate, if priors are not sought, and as suggested by Park & Casella (2008), the marginal maximum likelihood estimate can be used. We have shown the feasibility of this estimate by MC-EM, finding values similar to the use of low informative priors for λ. Thus, Bayesian Lasso is rather unique among parametric variable selection methods in genomic selection because it is readily estimable using fully parametric methods, either fully Bayesian or using marginal maximum likelihoods.

(iii) *Shape of SNP effects*

Figure 2 shows clearly that BL2Var is able to accommodate SNPs of large effect (i.e. around DGAT1) and also of small, almost nil, effects. Because of the nature of shrinkage caused by models positing *a priori* normal distributions, both features are generally difficult to attain, unless large amounts of information exist. It is unknown if these kinds of distributions (e.g. similar to double exponential) for SNP effects are frequent in nature or not, but the case of DGAT1 shows its importance in practice, at least for the dairy cattle industry. Because of this ability, BL2Var results in optimal predictive abilities.

Figure 2 shows the theoretical distribution of SNP effects for FP in Holstein according to the distributions for a described in Methods and estimates for λ (for BL2Var), λ and σ^2_e (for BL1Var) and σ^2_a (for MCMC-GBLUP). FP is the trait with more differences observed for predictive ability in the cross-validation approach, and partially controlled by DGAT1. For the BL1Var approach, the peak is not very sharp and is indeed lower than for BL2Var. This produces a distribution for SNP effects with little shrinkage for BL1Var, which is also reflected in Table 6. Thus, BL1Var does not seem to shrink enough in the appropriate regions. This is reflected in its poor predictive ability. On the other hand, BL2Var shrinks more than the normal distribution for most of the domain except for effects of more than about 3 standard units. This illustrates well that in order to account for the distribution of SNP effects both the sharpness and the variance have to be considered.

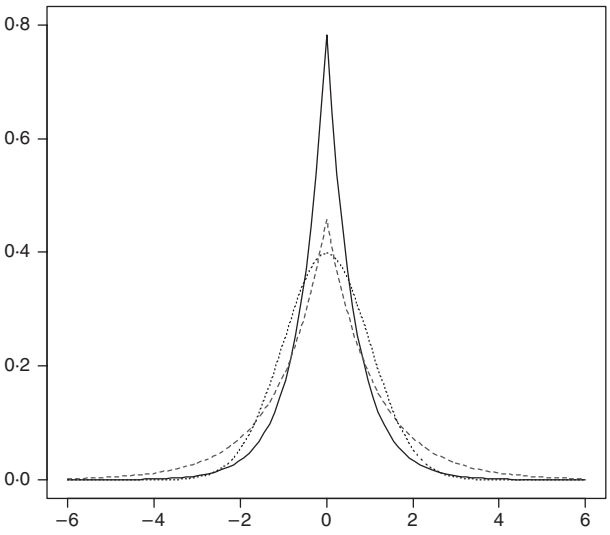

Fig. 2. Theoretical distribution of SNP effects for fat content according to estimates of σ^2_e, σ^2_a and λ in BL2Var (continuous line), BL1Var (grey dashed line) and MCMC-GBLUP normal model (dotted black line). The figure has been scaled so that the normal distribution has a variance of 1.

(iv) *Precomputation of variance of SNP effects*

In HetVar-GBLUP, variances of SNPs are pre-computed via BL2Var. This results in good accuracies. This strategy presents several practical advantages. Computation of SNP solutions, once variances are known, is very fast following, for example, Legarra & Misztal (2008). Also, a genomic relationship matrix with the same results, can be constructed as $\mathbf{G} = \mathbf{ZDZ}'\sigma^2_a$ (Goddard, 2008; Van Raden, 2008), giving more importance to SNPs with large than small effect. Mixed model equations using $\mathbf{G}$ can be set up with several nice properties. Solving is quite simple (Van Raden, 2008) even for singular $\mathbf{G}$ (Henderson, 1984). Pseudo-reliabilities can be constructed from their inverse; extensions exist to include full pedigree in the relationships and all data available (Legarra *et al.*, 2009; Aguilar *et al.*, 2010; Christensen & Lund, 2010). Models including an additional polygenic term (i.e. Guillaume *et al.*, 2008) can easily be set up.

We suggest, based on these practical considerations, a two-step procedure to include large amounts of data (i.e. genotyped cows, genotyped individuals with no record, or all genotyped and ungenotyped

individuals). First, variances in **D** can be estimated in a small yet informative sample of data (e.g. bulls). Second, **D** can be used for genetic evaluation either for all genotyped animals or all animals in the population. Pre-computation is the strategy used for variance components in regular genetic evaluation. It is an open question whether this strategy is stable with time or across different strata in a population.

6. Conclusion

The Bayesian Lasso with different variances for residual or SNP effects (BL2Var), which is equivalent to the original Lasso (Tibshirani, 1996) is appropriate for genomic selection, with generally highest accuracies and less inflation of GEBVs than other methods included in this study. Park & Casella's (2008) original BL1Var cannot be recommended because of inappropriate constraints in the model. We have shown how to estimate parameter λ with little (or no) prior information, and its biological interpretation in relation to genetic variation in the population. The inclusion of specific SNP variances in linear models is feasible by pre-computing the variances with the BL2Var. These methods should be further explored in other data sets including different traits and species.

This work was financed by ANR project AMASGEN (2009–2011). Project partly supported by Toulouse Midi-Pyrénées bioinformatic platform. We thank G. de los Campos for carefully explaining to us the Bayesian Lasso and for his initial code in R. Suggestions of the reviewers are gratefully acknowledged.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S. & Lawlor, T. J. (2010). A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* **93**, 743–752.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W. & Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**, 553–561.

Christensen, O. F. & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**, 2.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**, 375–385.

Falconer, D. & T. Mackay (1996). *Introduction to Quantitative Genetics*. New York: Longman.

Gautier, M., Capitan, A., Fritz, S., Eggen, A., Boichard, D. & Druet, T. (2007). Characterization of the DGAT1K232A and variable number of tandem repeat polymorphisms in French dairy cattle. *Journal of Dairy Science* **90**, 2980–2988.

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

Gianola, D. & Fernando, R. L. (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science* **63**, 217–244.

Goddard, M. (2008). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M. & Snell, R. (2002). Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**, 222–231.

Guillaume, F., Fritz, S., Boichard, D.& Druet, T. (2008). Correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. *Journal of Dairy Science* **91**, 2520–2522.

Habier, D., Fernando, R. L. & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397.

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

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph.

Hill, W. G. (2010). Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B* **365**, 73–85.

Kizilkaya, K., Fernando, R. L. & Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of Animal Science* **88**, 544–551.

Legarra, A., Aguilar, I. & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**, 4656–4663.

Legarra, A. & Misztal, I. (2008). Technical note: Computing strategies in Genome-wide selection. *Journal of Dairy Science* **91**, 360–366.

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

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M. & Meuwissen, T. H. E. (2009). The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. *Genetics* **183**, 1119–1126.

Lund, M. S., Sahana, G., de Koning, D. J., Su, G. & Carlborg, O. (2009). Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. *BMC Proceedings* **3**, S1.

Mäntysaari, E., Liu, Z. & Van Raden, P. (2010). Interbull validation test for genomic evaluations. *Interbull Bulletin* **41**.

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

Michael, J., Schucany, W. & Haas, R. (1976). Generating random variates using transformations with multiple roots. *American Statistician* **30**, 88–90.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. & Lee, D. (2002). BLUPF90 and related programs (BGF90). In Seventh World Congress on Genetics

Applied to Livestock Production, 2002, CD-ROM Communication N° 28–07.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686.

Patterson, H. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

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

Sorensen, D. & D. Gianola (2002). *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York: Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Usai, M. G., Goddard, M. E. & Hayes, B. J. (2009). LASSO with cross-validation for genomic selection. *Genetics Research* **91**, 427–436.

Van Raden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414–4423.

Van Raden, P. M., Tooker, M. E. & Cole, J. B. (2009 *a*). Can you believe those genomic evaluations for young bulls? *Journal of Animal Science* **87**(E-Suppl. 2), 314(abstr. 279).

Van Raden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. & Schenkel, F. S. (2009 *b*). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16–24.

Van Raden, P. M. & Wiggans, G. R. (1991). Derivation, calculation, and use of national animal model information. *Journal of Dairy Science* **74**, 2737–2746.

Van Tassell, C. P. & Van Vleck, L. D. (1996). Multiple-trait Gibbs sampler for animal models: flexible programs for Bayesian and likelihood-based (co)variance component inference. *Journal of Animal Science* **74**, 2586–2597.

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

Weigel, K. A., de los Campos, G., González-Recio, O., Naya, H., Wu, X. L., Long, N., Rosa, G. J. M. & Gianola, D. (2009). Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of Dairy Science* **92**, 5248–5257.