

# Extended Bayesian LASSO for Multiple Quantitative Trait Loci Mapping and Unobserved Phenotype Prediction

Crispin M. Mutshinda\* and Mikko J. Sillanpää\*<sup>†,1</sup>

\*Department of Mathematics and Statistics, University of Helsinki, Helsinki FIN-00014, Finland and

<sup>†</sup>Department of Agricultural Sciences, University of Helsinki, Helsinki FIN-00014, Finland

Manuscript received June 4, 2010

Accepted for publication August 27, 2010

## ABSTRACT

The Bayesian LASSO (BL) has been pointed out to be an effective approach to sparse model representation and successfully applied to quantitative trait loci (QTL) mapping and genomic breeding value (GBV) estimation using genome-wide dense sets of markers. However, the BL relies on a single parameter known as the regularization parameter to simultaneously control the overall model sparsity and the shrinkage of individual covariate effects. This may be idealistic when dealing with a large number of predictors whose effect sizes may differ by orders of magnitude. Here we propose the extended Bayesian LASSO (EBL) for QTL mapping and unobserved phenotype prediction, which introduces an additional level to the hierarchical specification of the BL to explicitly separate out these two model features. Compared to the adaptiveness of the BL, the EBL is “doubly adaptive” and thus, more robust to tuning. In simulations, the EBL outperformed the BL in regard to the accuracy of both effect size estimates and phenotypic value predictions, with comparable computational time. Moreover, the EBL proved to be less sensitive to tuning than the related Bayesian adaptive LASSO (BAL), which introduces locus-specific regularization parameters as well, but involves no mechanism for distinguishing between model sparsity and parameter shrinkage. Consequently, the EBL seems to point to a new direction for QTL mapping, phenotype prediction, and GBV estimation.

**R**EGULARIZATION or shrinkage methods are gaining increasing recognition as a valuable alternative to variable selection techniques in dealing with oversaturated or otherwise ill-defined regression problems in both the classical and Bayesian frameworks (*e.g.*, O'HARA and SILLANPÄÄ 2009). Many studies (*e.g.*, XU 2003; WANG *et al.* 2005; ZHANG and XU 2005; DE LOS CAMPOS *et al.* 2009; USAI *et al.* 2009; WU *et al.* 2009; XU *et al.* 2009) have documented the potential of shrinkage methods for quantitative trait locus (QTL) mapping and genomic breeding value (GBV) estimation using genome-wide dense sets of markers. LEE *et al.* (2008) make a clear connection between phenotype prediction and GBV estimation, suggesting that methods developed for one are also applicable to the other. We thus use the two concepts interchangeably throughout this article.

Regularized regression methods, such as ridge regression (HOERL and KENNARD 1970) or the least absolute shrinkage and selection operator (LASSO) (TIBSHIRANI 1996), are essentially penalized likelihood procedures, where suitable penalty functions are added to the negative log-likelihood to automatically shrink spurious effects (effects of redundant covariates) to-

ward zero, while allowing relevant effects to take values farther from zero.

It has been pointed out that these non-Bayesian shrinkage methods are not suitable for oversaturated models. ZOU AND HASTIE (2005) and PARK AND CASELLA (2008) noted that the LASSO cannot select a number of nonzero effects exceeding the sample size. XU (2003) found that for ridge regression to work, the number of model effects should be in the same order as the number of observations. This is impractical for genomic selection, which capitalizes on the variation due to small-marker effects, the number of which can exceed the sample size, by contrast to QTL mapping where interest lies mostly in a small subset of loci with large effects on the focal phenotype. In connection with the LASSO, the Bayesian LASSO (BL) (PARK AND CASELLA 2008; YI AND XU 2008) has been proposed to overcome this limitation by imposing a selective shrinkage across regression parameters. XU (2003) also proposed a Bayesian shrinkage method for QTL mapping, which extends ridge regression in a similar fashion.

Although the BL has been successfully applied to QTL mapping (*e.g.*, YI AND XU 2008) and to GBV estimation (*e.g.*, DE LOS CAMPOS *et al.* 2009), it relies on a single parameter known as the regularization parameter to simultaneously regulate the overall model sparsity and the extent to which individual regression coefficients are

<sup>1</sup>Corresponding author: Department of Mathematics and Statistics, P.O. Box 68, Gustaf Hallstrominkatu 2b, University of Helsinki, Helsinki FIN-00014, Finland. E-mail: mjs@rolf.helsinki.fi

shrunk. However, this is unrealistic when dealing with a large number of predictors whose effect sizes may differ by orders of magnitude. It is therefore natural to ask whether this practice can be relaxed and how such an attempt may impinge on the model performance (*e.g.*, SUN *et al.* 2010).

Here we propose an extension to the Bayesian LASSO for QTL mapping and unobserved phenotype prediction. Our method, the extended Bayesian LASSO (EBL), introduces locus-specific regularization parameters and utilizes a parameterization that clearly separates the overall model sparsity from the degree of shrinkage of individual regression parameters. We use simulated data to investigate the performance of the EBL relative to the Bayesian LASSO in mapping QTL and in predicting unobserved phenotypes. We also compare the performance of the EBL to the Bayesian adaptive LASSO (BAL) recently proposed by SUN *et al.* (2010), which also assumes locus-specific regularization parameters.

## METHODS

**Statistical model:** Let  $y_i$  ( $i = 1, \dots, n$ ) and  $x_{ij}$  denote, respectively, the value of the phenotypic trait of interest for the  $i$ th individual and the genotype code of the  $i$ th individual at locus  $j$  ( $j = 1, \dots, p$ ). The model is developed in the context of experimental crosses derived from two inbred lines such as backcross (BC) or double haploid (DH) progenies with only two possible genotypes at any locus. The dummy variables  $x_{ij}$  are coded as  $x_{ij} = 0$  for one genotype and  $x_{ij} = 1$  for the other. We assume an additive model that regresses the phenotypic values of the  $n$  individuals on their genotypes at  $p$  putative loci, which may be markers or alternatively pseudomarkers (SEN and CHURCHILL 2001; SERVIN and STEPHENS 2007). That is,

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i, \quad (1)$$

where  $\beta_0$  is the intercept. Under the genotype coding assumed here,  $\beta_j$  represents the genetic effect associated with locus  $j$  ( $j = 1, \dots, p$ ) or more precisely, the difference (contrast) between the genetic effects associated with the two genotypes for locus  $j$ . The residual errors  $\varepsilon_i$  are assumed to be independent and normally distributed with mean zero and common variance  $\sigma^2$ . The model is applicable to other experimental designs, such as F2 intercross progeny under appropriate genotype coding and to population-based samples. For breeding purposes or analysis of family data, model (1) can be extended to include an infinitesimal polygenic term that accounts for relatedness between individuals (see, *e.g.*, DE LOS CAMPOS *et al.* 2009 or PIKKUHOOKANA and SILLANPÄÄ 2009 for more details). In matrix notation, model (1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

where  $\mathbf{y} = (y_1, \dots, y_n)^\top$ ,  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ , and  $\mathbf{X}$  is the  $n \times (p + 1)$  design matrix comprising the genotype profiles of the  $p$  loci, with all elements of the first column (which corresponds to the intercept) set to 1. The likelihood for this model is given by

$$L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{1/2} \exp(-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / 2\sigma^2), \quad (3)$$

where  $\boldsymbol{\Sigma} = \sigma^{-2}\mathbf{I}_n$  is the precision matrix associated with the residual errors,  $\mathbf{I}_n$  is the  $n \times n$  identity matrix, and  $\|\cdot\|_2$  denotes the  $L_2$  norm.

Maximum likelihood estimates (MLEs) of  $\boldsymbol{\beta}$  and  $\sigma^2$  are obtained by maximizing the likelihood function  $L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$  or alternatively, by minimizing the negative likelihood function with respect to  $\boldsymbol{\beta}$  and  $\sigma^2$ . However, as already pointed out, when dealing with genome-wide dense sets of markers, the number,  $p$ , of predictors can be larger than the sample size  $n$ , with most of the markers having weak or no effect on the focal phenotype (*e.g.*, XU 2003; ZHANG and XU 2005). In such circumstances, regularization methods offer an interesting approach to inducing a sparse model representation by automatically shrinking the effects of redundant covariates (spurious effects) toward zero, while effectively estimating the relevant ones.

**Regularization methods and the LASSO:** The rationale of regularization methods is to penalize the log-likelihood function  $-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / 2\sigma^2$  with a suitable nondecreasing function,  $g(\boldsymbol{\beta}) > 0$ , and estimate  $\boldsymbol{\beta}$  by  $\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} (-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / 2\sigma^2 - g(\boldsymbol{\beta}))$ , or equivalently,

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / 2\sigma^2 + g(\boldsymbol{\beta})). \quad (4)$$

Several penalty functions have been proposed, including  $g(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2$  ( $\lambda > 0$ ), which leads to ridge regression (HOERL and KENNARD 1970; WHITTAKER *et al.* 2000; MALO *et al.* 2008), and  $g(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_j |\beta_j|$ , which leads to the LASSO estimate

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / 2\sigma^2 + \lambda\|\boldsymbol{\beta}\|_1). \quad (5)$$

(TIBSHIRANI 1996; USAI *et al.* 2009; WU *et al.* 2009). The smoothing or regularization parameter,  $\lambda$ , determines the extent of the penalization, with larger values implying a stronger penalization on the  $L_w$  norm ( $w = 1$  for LASSO and  $w = 2$  for ridge regression), which results in more regression coefficients being shrunk toward zero. Henceforth, our attention is focused on the LASSO owing to its desirable feature of setting some of the redundant effects to be exactly zero (TIBSHIRANI 1996). Note that if  $\beta_j$  is shrunk to zero, then the

LASSO objective function,  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2\sigma^2 + \lambda\|\boldsymbol{\beta}\|_1$ , in Equation 5 becomes independent of  $\beta_j$  so that  $X_j$  is pruned (discarded) from the model. As  $\lambda$  approaches zero,  $\hat{\boldsymbol{\beta}}$  tends to the ordinary least-squares estimate (no shrinkage). The intercept,  $\beta_0$ , is not subject to shrinkage since we consider that it is always included into the model. The term  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$  in Equation 5 expresses the least-squares approximation, whereas the penalty function  $g(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$  is intended to promote sparseness. The LASSO estimate  $\hat{\boldsymbol{\beta}}_{\text{LASSO}}$  can be interpreted from a Bayesian perspective as the *maximum a posteriori* (MAP) or posterior mode of the parameter vector  $\boldsymbol{\beta}$  under the prior specification  $\beta_j \sim N(0, \sigma_j^2)$  and  $\sigma_j^2 \sim \text{Exp}(\lambda^2/2)$ , independently for  $j = 1, \dots, p$  (TIBSHIRANI 1996; PARK and CASELLA 2008; YI and XU 2008). Equivalently, this corresponds to assigning to the  $\beta_j$  independent Laplacian or double exponential (DE) priors with mean zero and variance  $2/\lambda^2$ . This follows from the fact that, for any  $x \in \mathfrak{R}$ , and  $\lambda > 0$ ,

$$\begin{aligned} \text{DE}(x|0, \lambda/2) &= \frac{\lambda}{2} \exp(-\lambda|x|) \\ &= \int_0^\infty (1/\sqrt{2\pi s}) \exp(-x^2/2s) \frac{\lambda^2}{2} \exp(-\lambda^2 s/2) ds, \end{aligned} \quad (6)$$

where  $\text{DE}(x|0, \lambda/2)$  denotes the probability density function of the DE distribution with mean 0 and variance  $2/\lambda^2$ , evaluated at  $x$  (PARK and CASELLA 2008).

The DE distribution is sharply peaked at the mode and has heavier than Gaussian tails. Setting the prior mode at zero encodes a prior belief of no effect for most of the regression covariates. In addition, the strength of this belief, and hence the degree of model sparsity, is determined by the variance of the DE prior assumed for the regression coefficients, which in turn depends solely on the regularization parameter  $\lambda$ . The LASSO estimate corresponds to the MAP of  $\boldsymbol{\beta}$  when all  $\beta_j$  ( $j = 1, \dots, p$ ) are independently assigned  $\text{DE}(0, \lambda/2)$  priors with  $\lambda$  set to a fixed nonnegative constant. But as pointed out earlier, the number of nonzero effects that the LASSO can select is bounded from above by the sample size. A Bayesian formulation of the LASSO known as the Bayesian LASSO has been proposed to overcome this limitation.

**The Bayesian LASSO:** In the BL (PARK and CASELLA 2008; YI and XU 2008; DE LOS CAMPOS *et al.* 2009), each regression parameter,  $\beta_j$  ( $j = 1, \dots, p$ ), is *a priori* assumed to be normally distributed around zero with its own variance  $\sigma_j^2$ , *i.e.*,  $\beta_j \sim N(0, \sigma_j^2)$ . Consequently, the degree of shrinkage is made locus specific through differences in the variances  $\sigma_j^2$  across loci. These variance parameters are independently assigned  $\text{Exp}(\lambda^2/2)$  priors. With this hierarchical prior specification, the marginal prior of the regression parameters,  $\beta_j$ , is  $\text{DE}(0, \lambda/2)$ , independently for  $j = 1, \dots, p$ , by virtue of Equation 6. Differences in locus-specific variances induce a selective shrinkage with smaller variances implying stronger shrinkage toward zero, as expected for

spurious effects. It is clear, however, that the variance  $4/\lambda^4$  of  $\sigma_j^2$ , and hence the amount of shrinkage of all regression parameters is largely influenced by the regularization parameter  $\lambda$ . In the BL, the hyperparameter  $\lambda$  is meant to be estimated alongside the other model parameters and as such, requires a prior distribution *e.g.*,  $\lambda \sim \Gamma(a, b)$  for some suitably selected values of  $a$  and  $b$ . PARK and CASELLA (2008) and YI and XU (2008) recognized that the performance of the BL depends critically on the tuning of  $\lambda$ . The hierarchical specification of the priors of the regression coefficients in the BL is graphically depicted in Figure 1a. As can be seen from Figure 1a, all the variances  $\sigma_j^2$  for ( $j = 1, \dots, p$ ) are controlled by the hyperparameter  $\lambda$ . This may be unrealistic when dealing with a large number of predictors in which effect sizes may differ greatly, as frequently happens in QTL mapping and in phenotypic value prediction using genome-wide dense sets of markers. Our new method, the extended Bayesian LASSO, relaxes this assumption.

**The extended Bayesian LASSO:** In the vein of the BL, the EBL proceeds by assigning to each regression parameter a Gaussian prior with its own variance, independently. That is,  $\beta_j \sim N(0, \sigma_j^2)$ , independently for  $j = 1, \dots, p$ . Each locus-specific variance,  $\sigma_j^2$ , is further independently assigned an  $\text{Exp}(\lambda_j^2/2)$  prior. The EBL departs from the BL in that the regularization parameters are locus specific. Crucially for us here,  $\lambda_j$  is defined in terms of a measure,  $\delta$ , of model sparsity, which is common to all loci, and a locus-specific deviation,  $\eta_j$ , from  $\delta$ , which represents the strength of shrinkage specific to locus  $j$ . More specifically, we let

$$\lambda_j = \delta \eta_j. \quad (7)$$

The common factor  $\delta$  can, just like  $\lambda$  in the BL, be used to adjust the model to the desired level of sparsity, whereas the locus-specific random deviations,  $\eta_j$ , from  $\delta$  enforce the difference in the extent of shrinkage across loci. This makes the EBL “doubly adaptive” compared to the adaptiveness of the BL. Figure 1b shows the hierarchical specification for the priors of the regression coefficients in the EBL. The extra layer of the EBL relative to the BL is apparent from Figure 1b. The hyperparameters  $\delta$  and  $\eta_j$  are also assigned prior distributions and estimated along with the other model parameters. A nice feature of the EBL is that tuning is not critical like in the BL. We discuss this point in more detail further.

## APPLICATION

**Simulated data analysis:** The performance of the EBL was evaluated on simulated data. The data-generation process used marker data from a well-known data set from the North American Barley Genome Mapping project (TINKER *et al.* 1996), which has been widely

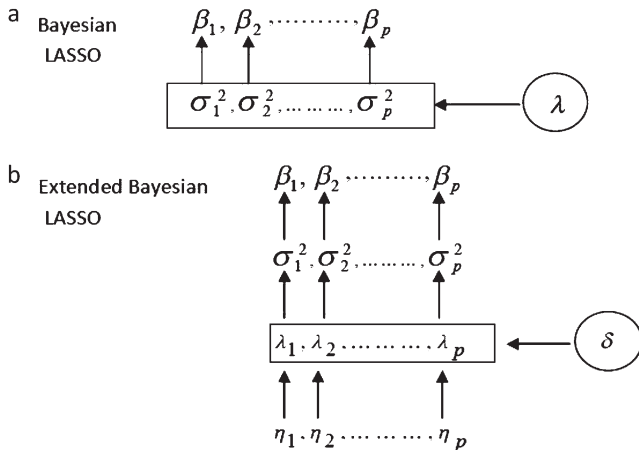


FIGURE 1.—Graphical representation of the hierarchical specification of the priors of the locus-specific effects in the BL (a) and the EBL (b), illustrating the roles and scopes of each hyperparameters:  $\lambda$  in the BL and  $\delta$  and  $\eta_j$  in the EBL.

analyzed in mapping studies (*e.g.*, XU 2003, 2007; XU and JA 2007; YI and XU 2008). The Barley data, initially intended for analyzing economically important traits in two-row barley (*Hordeum vulgare* L.), comprise 127 biallelic markers spanning seven chromosomes for 150 DH lines grown in 25 different environments. The markers cover 1270 cM, with an average distance between consecutive markers of 10.5 cM. The original data involved 150 individuals, but five individuals with missing phenotypes (days to heading) were left out. The incentive for using the Barley marker data is to draw close to a real-world situation. The data set also includes some missing genotypes. At the outset, the missing genotypes were filled-in with random draws from *Bernoulli* (0.5). For a DH population, an individual can take only one of two possible genotypes. The dummy variables  $x_{ij}$  were coded as  $x_{ij} = 0$  for one genotype and  $x_{ij} = 1$  for the other. We used the model in (1) to generate phenotypic values assuming a few QTL, namely at loci 4, 25, 50, and 65 with respective effects set to 2.5,  $-2.5$ , 4, and  $-4$ . We used different values of the residual variance to induce different heritability levels in the data. Heritability,  $\hat{h}^2 = (\hat{\sigma}_y^2 - \sigma^2) / \hat{\sigma}_y^2$ , refers to the proportion of phenotypic variance attributable to genetic factors. Here  $\sigma^2$  is the known residual variance and  $\hat{\sigma}_y^2$  is the empirical phenotypic variance estimated from the data.

It is well known that correlations among covariates may greatly affect the performance of variable selection methods. To verify how well our method would perform in the presence of dense marker data, we considered, in addition to the moderately dense Barley marker data, a dense marker data set simulated through the WinQTL Cartographer program (WANG *et al.* 2006), and used by SILLANPÄÄ and NOYKOVA (2008). This data set involves 200 backcross progeny and 102 marker points spanning three chromosomes, with 34 evenly spaced markers on

each chromosome. The length of each chromosome was 99 cM, implying a distance of 3 cM between consecutive markers. The methodology described here is straightforwardly applicable to this data set since a backcross can have only two possible genotypes (initially coded as  $-1$  and  $1$ ) at each locus. We changed the initial coding to 0 and 1. As before, we assumed only four QTL at loci 25, 60, 70, and 75 with QTL effects of 2.5,  $-2.5$ , 4, and  $-4$ , respectively. Note that the last two QTL are linked with 15 cM between them and in repulsion (*i.e.*, having effects of opposite signs).

For both the moderately dense Barley marker data and the simulated dense marker data, we generated 50 synthetic data sets of approximately same heritability with the design matrix held fixed and the intercept set to zero without loss of generality. For the reported results, the residual variance was set to 2, yielding an average heritability of 0.8 for the Barley marker data and 0.72 for the simulated dense marker data.

**Model fitting and performance evaluation:** We fitted the Bayesian LASSO and extended Bayesian LASSO to the simulated data replicates by Markov chain Monte Carlo (MCMC) (GILKS *et al.* 1996) simulation through the Bayesian freeware OpenBUGS (THOMAS *et al.* 2006). The BUGS code is available upon request from the authors. The response variable was considered on the original scale to facilitate the comparison between the true effect sizes and their estimates. We used the following, essentially noninformative, priors:  $\beta_0 \sim N(0, 100)$ ;  $\beta_j \sim N(0, 100)$  for  $j = 1, \dots, p$ ;  $\sigma^2 \sim \text{Inv} - \Gamma(0.1, 0.1)$ . To mitigate the effect of prior specification on our results, the reported results are based on the noninformative priors  $\lambda \sim \text{Uni}(0, 100)$  in the BL, and  $\delta \sim \text{Uni}(0, 100)$ ,  $\eta_j \sim \text{Uni}(0, 100)$  in the EBL. These priors are diffuse (flat). This implies no tuning, which is important if we are to compare the performances of the two models.

Initially, we ran three separate Markov chains for 10,000 iterations each, to assess how fast the chain would converge and how well they would mix (mixing refers to the ease with which the Markov chains explore the full parameter space). For the simulations based on the Barley marker data, 10,000 iterations took 21,483 sec for the BL and 22,978 sec for the EBL (*i.e.*, comparable computation time) on an AMD with 1.33 GHz and 768 MB of RAM. For both methods, the chains seemed to reach the stationary distribution after about 1500 iterations. The convergence was assessed through visual inspection of the MCMC traceplots. The sensitivity of the results to the prior specification was assessed by varying the range of the uniform priors for hyperparameters  $\lambda$  in the BL, and  $\delta$  and  $\eta_j$  in the EBL, but the results were robust to these changes.

After the initial evaluation of the two models, we used simulated data replicates to evaluate their performances in estimating the regression parameters with regard to both the true signals (*i.e.*, QTL) and false signals (*i.e.*,



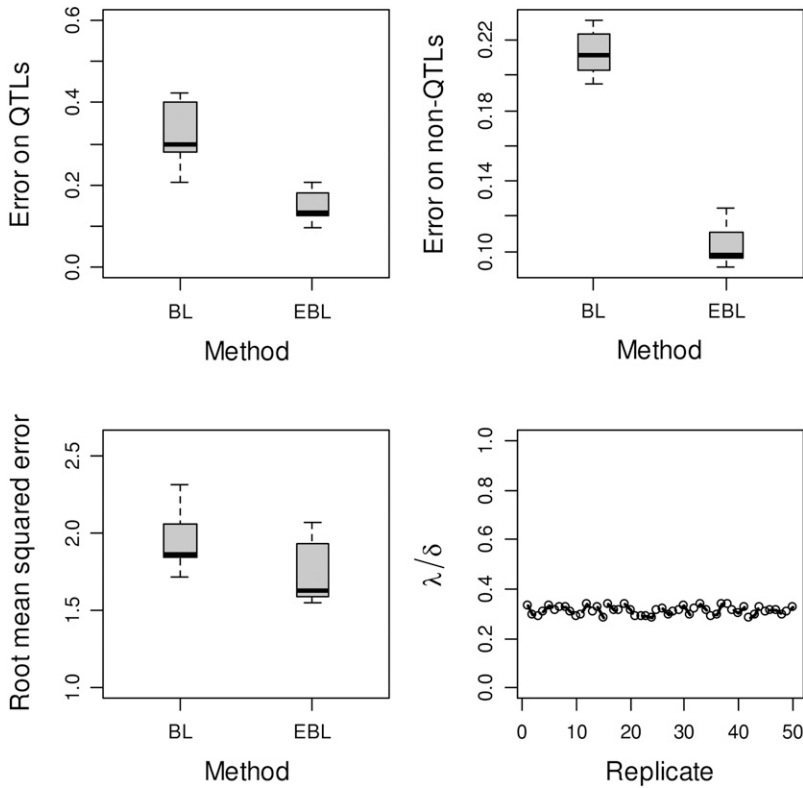


FIGURE 2.—Performance evaluation of the EBL and the BL over 50 replicated data sets using the Barley marker data. The results are summarized by the posterior means over 50 replicated data sets. Top left: the estimation errors (differences between the posterior means and known true values) with regard to true signals *i.e.*, QTL. Top right: the estimation errors with regard to false signals *i.e.*, non-QTL. Bottom left: predictive root mean square errors for the BL and the EBL. Bottom right: the ratios of posterior means of  $\lambda$  (in the BL) and  $\delta$  (in the EBL) over 50 replicated data sets, illustrating the correspondence between the roles of these two locus-independent parameters ( $\lambda$  and  $\delta$ ) in the two methods.

non-QTL), and in predicting unobserved phenotypes. For each replicated data set, we ran 10,000 MCMC iterations and discarded the first 2000 samples as burn-in.

The estimation error on the true and the false signals were respectively evaluated through the statistics

$$R_T = \frac{1}{N_t} \sum_t |\Delta\beta_t| / |\beta_t| \quad \text{and} \quad R_F = \frac{1}{N_f} \sum_f |\Delta\beta_f|,$$

where  $\Delta\beta_t = (\beta_t - \hat{\beta}_t)$ ,  $\beta_t$  and  $\hat{\beta}_t$  being the true and the estimated effects of locus  $t$ , respectively. The summation in the definition of  $R_T$  is taken over the true signals, and  $N_t$  is the number of true signals.  $R_F$  is defined in a similar manner to  $R_T$ , but  $\beta_f = 0$ , and the formula for  $R_F$  involves no scaling by the absolute value of the true effect size, which is zero. The number,  $N_f$ , of false signals, is given by  $N_f = (p - N_t)$ .

The predictive performances of the two models were evaluated through posterior predictive cross-validation. The basic idea of cross-validation (*e.g.*, PICARD and COOK 1984) is to fit the model to a subset of the data (the training set) and utilize the remaining data (the test set) to evaluate the model predictive performance. A cross-validation loss function (usually the mean square error or the root mean square error) is required for comparison, and the model associated with the smallest loss is chosen as the best. Here we left out (*i.e.*, we omitted from the training set) two randomly selected phenotypic values in each replicated data set and monitored their posterior predictive means estimated from the MCMC samples. For each replicated data set, we used the

posterior predictive means of the omitted data points to compute the root mean square error (RMSE)

$$\text{RMSE} = \sqrt{(1/2) \sum_j (y_j^{\text{pred}} - y_j)^2},$$

where  $j$  runs over the indices of the discarded phenotypes,  $y_j^{\text{pred}}$  is the posterior predictive mean of the  $j$ th discarded phenotype, its actual value being  $y_j$ .

To have significance thresholds for distinguishing QTL from non-QTL, we randomly shuffled the phenotypes (permutation without replacement) to artificially destroy the association between the markers and the phenotypes. We fitted the model to each of the ensuing data sets using MCMC simulation. We ran 4000 MCMC iterations, discarding the first 2000 samples as burn-in. For each data set, we monitored the largest absolute posterior mean effect size used as our test statistic. This provided us with an empirical distribution  $F$  of the test statistic, and we used the  $100 \times (1 - \alpha)$  percentile of as the critical value (CHURCHILL and DOERGE 1994). The significance thresholds based on 100 phenotype permutations with  $\alpha = 0.10$  were found to be 0.64 for the BL and 0.67 for the EBL.

## RESULTS

The results for the simulations based on the Barley marker data were broadly similar to those based the simulated dense marker data set. In this section, we report only on the performance of the BL and the EBL

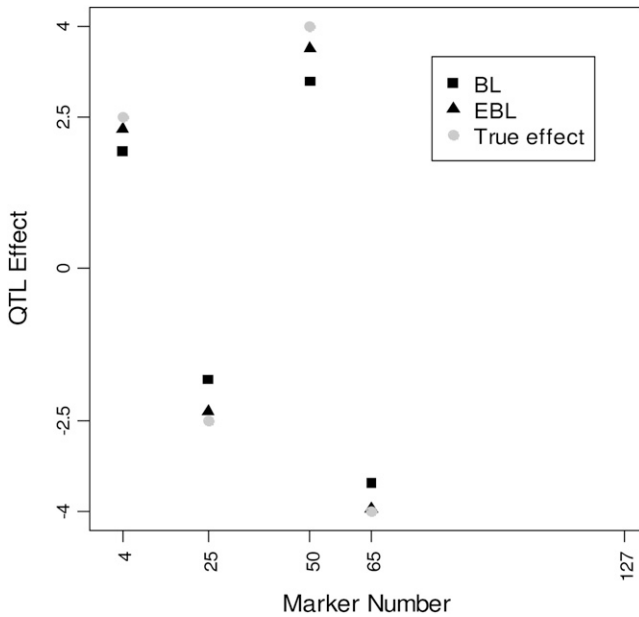


FIGURE 3.—Posterior means of estimated QTL effects for the BL (solid squares) and the EBL (solid triangles) at QTL positions, along with their true effects (shaded circles). The plotted values represent the posterior means averaged over 50 replicated data sets based on the Barley marker data.

based on the Barley marker data (an extensive comparison of the EBL to the contending BAL, including for the dense marker data, is provided in the DISCUSSION). Figure 2 summarizes the performance of the EBL relative to the BL with regard to the accuracy of effect size estimates on both true and false signals and the accuracy of phenotypic value predictions. Figure 2, bottom right, displays the ratios  $\lambda/\delta$  of the locus-independent variance hyperparameters in the BL and the EBL, which is roughly constant over the 50 replicated data sets.

Figure 3 shows the posterior means of QTL effects under the BL and the EBL, along with the true values. The results displayed in Figure 3 suggest that the EBL estimates track the true effects more efficiently than their BL counterparts.

Figure 4 shows the posterior means of all loci effects averaged over the 50 replicated data sets for the BL and the EBL. For each method, the significance thresholds for declaring QTL are indicated by the horizontal dashed lines.

For a significance level of 10%, the detection probabilities for QTL were roughly 0.87 for the BL and 0.91 for the EBL.

## DISCUSSION

Our discussion is structured in subsections relating to different aspects on which the BL and the EBL are being evaluated. We also compare the performance of the EBL to the BAL recently proposed by SUN *et al.* (2010), which also introduces locus-specific regularization parameters.

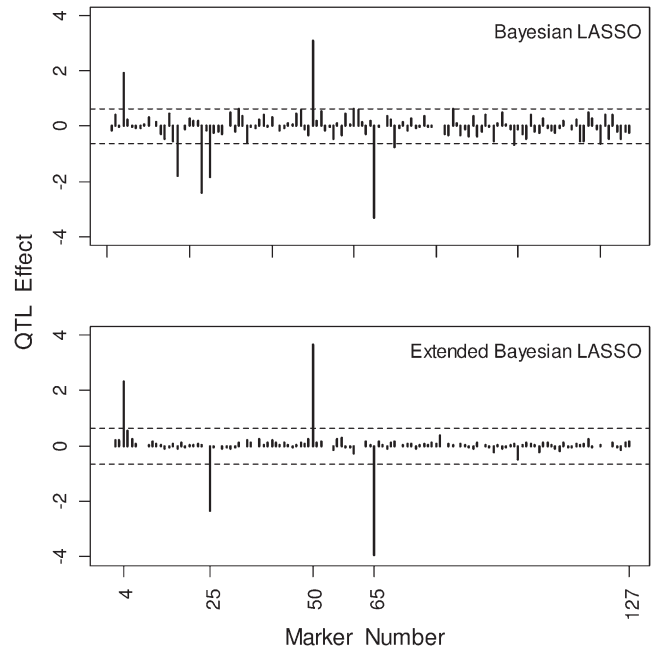


FIGURE 4.—Posterior means of all marker effects averaged over the 50 replicated data sets based on the Barley marker data for the BL and the EBL. The dashed lines indicate the permutation-based QTL significance thresholds.

**Signal detection:** The locus-effect estimates under the EBL were always closer to the true values than their BL counterparts (Figure 3), resulting in lower estimation errors with regard to the true signals (*i.e.*, QTL) and false signals (*i.e.*, non-QTL) (Figure 2). Nevertheless, both the BL and the EBL were effective at identifying the known QTL in our simulation settings, as exemplified by the plots in Figures 3 and 4.

**Predictive performance:** The EBL outperformed BL in predictive accuracy, with lower prediction errors (RMSEs displayed in Figure 2). This implies that the EBL is also better suited to breeding value prediction in genomic selection. Our specification of the regularization parameters provides the possibility of incorporating prior information regarding the level of model sparsity through the prior of the overall sparsity parameter,  $\delta$ , while letting the locus-specific hyperparameters,  $\eta_j$  ( $j = 1, \dots, p$ ), take care of keeping the nonzero effects “on” and adjusting their sizes according to their importance. This doubly adaptive feature is appealing in that it allows the model to effectively exert less shrinkage on the effects of important covariates required to be into the model, and more shrinkage on spurious effects. This results in more accurate estimates of marker effects and predictions of phenotypic values as implied by the results of our simulation study. The BL may miss the point as it relies on a single regularization parameter for model sparsity and parameters shrinkage.

**Sensitivity to prior specification and separation of QTL from non-QTL loci:** In the BL, the hyperparameters of the prior  $p(\lambda)$  require suitable tuning, with

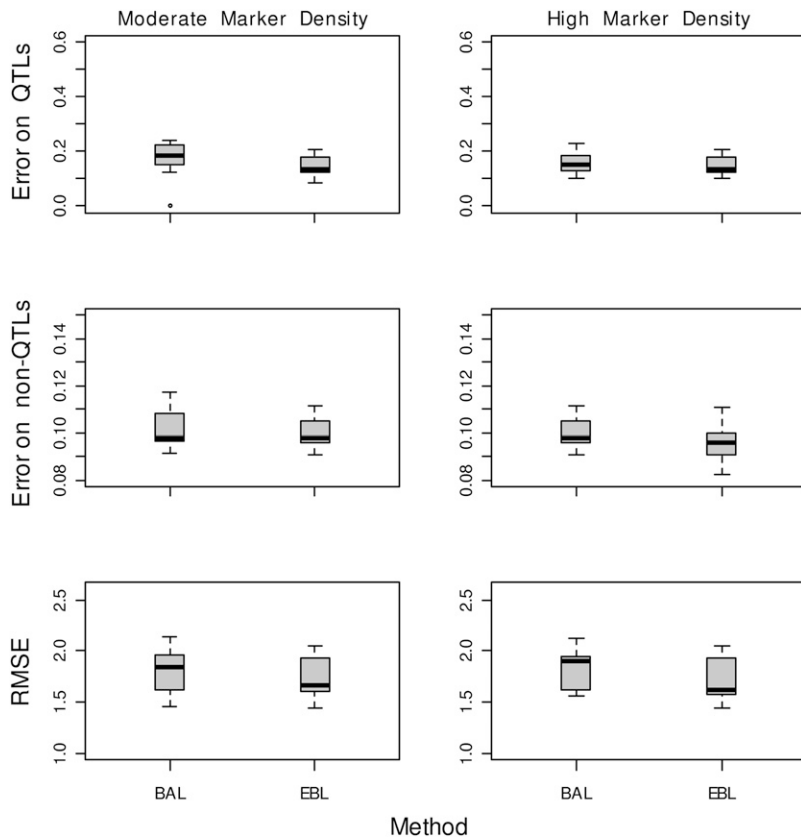


FIGURE 5.—Performance evaluation of the EBL and the BAL on the Barley marker data (left), and a simulated dense marker data set (right). The results are summarized by boxplots of posterior means over 50 replicated data sets. Top: the estimation errors (differences between the posterior means and known true values) for true signals *i.e.*, QTL. Middle: the estimation errors with regard to false signals *i.e.*, non-QTL. Bottom: predictive root mean square errors.

optimal values depending on the data and the number of variables included (YI and XU 2008), meaning that tuning is critical to the performance of the BL. The EBL estimates proved to be robust to tuning, owing presumably to the way the priors of the locus-specific hyperparameters  $\lambda_j$  are specified. In particular, the ratio of the posterior means of the parameters  $\lambda$  and  $\delta$  in the BL and the EBL, respectively, was found to be roughly constant (Figure 2, bottom right). This indicates that these two parameters equivalently regulate the level of model sparseness in the two models by “borrowing strength” across loci. A consequence of the reliance of the BL on a single smoothing parameter is that placing on this parameter a prior that strongly enforces sparsity will result in all regression parameters being indiscriminately shrunken. This is not a desirable feature for phenotype prediction, as most of the small but relevant parameters will be set to zero.

In the EBL, the locus-specific hyperparameters  $\eta_j$  play a key adaptive role, which makes tuning much less of an issue. It also is obvious that the parameters  $\delta$  and  $\eta_j$  in the EBL have meaningful interpretations as they operate on different scales: global for the former and local (*i.e.*, locus specific) for the latter. The graphical representation of the hierarchical specifications of the priors for locus-specific effects in the BL and the EBL shown in Figure 1 better illustrates the roles and scopes of the hyperparameters  $\lambda$  in the BL, and ( $\delta$  and  $\eta_j$ ) in the EBL.

It is worth emphasizing that, although the Laplacian distribution is mathematically equivalent to a mixture of an infinite number of Gaussians with exponentially distributed mixing variances as implied in Equation 6, it is computationally more convenient to not integrate out the mixing layer when the required computations can efficiently be carried, for example, through Markov chain Monte Carlo (GILKS *et al.* 1996). This recommendation seems to oppose the all-too-common practice of analytically integrating out intermediary layers of hierarchical priors in the quest for simplicity. O’HARA AND SILLANPÄÄ (2009) found that using the Laplace prior directly on the regression parameters in the Bayesian LASSO as obtained by the analytical integration in Equation 6 may result in poor mixing of the MCMC sampler and bad separation of the variables. Like YI AND XU (2008) and DE LOS CAMPOS *et al.* (2009), we did not face any mixing problems with the hierarchical specification of the Laplace distribution when implementing the BL and the EBL. The same applies to the Student *t*-distribution, which is mathematically equivalent to a mixture of an infinite number of Gaussians with inverse gamma-mixing variances.

**Statistical identifiability issues:** From a statistical model-fitting perspective, our parameterization of the locus-specific regularization parameters,  $\lambda_j$ , as a product of a factor,  $\delta$ , common to all loci and a locus-specific effect,  $\eta_j$ , is crucial for statistical identifiability. If these two components were both indexed by  $j$ , they would be confounded and, hence, statistically unidentifiable.

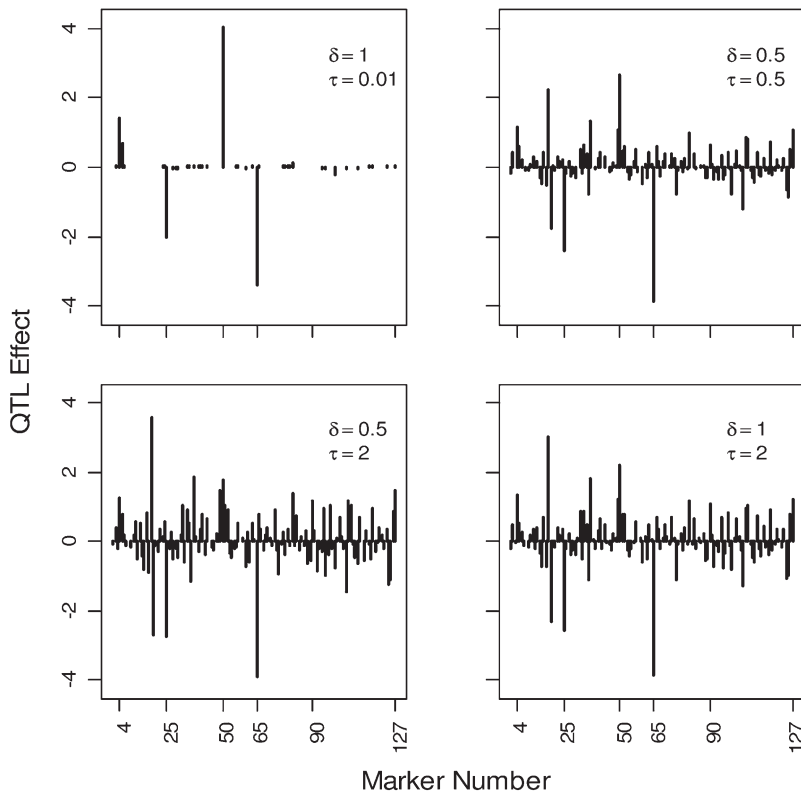


FIGURE 6.—Typical patterns of posterior means of marker effects for different hyperparameter values under our simulation setting based on the Barley marker data with QTL at loci 4, 25, 50, and 65 with respective effects 2,  $-2$ , 4, and  $-4$ . The results are based on 10,000 MCMC iterations, with the first 4000 samples discarded as burn-in.

**Comparison to the Bayesian adaptive LASSO:** SUN *et al.* (2010) also introduced locus-specific smoothing parameters in the Bayesian LASSO, but without a separation between model sparsity and parameter shrinkage hyperparameters. They proposed a fully Bayesian method, the BAL, along with an iterative method, the iterative adaptive LASSO, for fitting their model to the data. We compared the performance of the EBL to the BAL on simulated data on the basis of the Barley marker data and the simulated dense marker data set used before. We fitted the BAL to the simulated data replicates using the R library BPrimm (SUN *et al.* 2010). The BAL has two tuning hyperparameters  $\delta$  and  $\tau$ , with a default setting of  $\delta = 1$  and  $\tau = 0.01$ . We used a set of hyperparameter values. With the default setting, the BAL had comparable performance to the EBL on these particular data sets with a good separation between QTL and non-QTL effects as can be seen from Figure 5. The reported results are based on 10,000 MCMC iterations with 4000 burn-in iterations.

However, we found that the performance of the BAL may be sensitive to the tuning of the hyperparameters  $\delta$  and  $\tau$ , with larger values of  $\tau$  pointing to bad separation between QTL and non-QTL effects. This is apparent from the results displayed in Figure 6, which depicts typical patterns of posterior means of the marker effects for different hyperparameter values under our simulation setting based on the Barley marker data.

It is worth giving credit to SUN *et al.* (2010) for the recommendation of small values of  $\tau$ . However, too small a value for  $\tau$  may excessively shrink and eventually set to

zero small to moderate effects. As already pointed out above, this might not be desirable for phenotype prediction. The bottom line is that, while SUN *et al.*'s (2010) approach is an improvement over the Bayesian LASSO, tuning is still an issue therein. The EBL brings more flexibility by separating model sparsity and parameter shrinkage, thereby mitigating the impact of tuning on the model results. In our simulation study, the locus-specific hyperparameter,  $\lambda_j$ , was efficiently estimated from the data, even when uniform prior with large support such as  $(0, 100)$  were placed on  $\delta$  and  $\eta_j$ , and the results remained robust to changes in the range of these hyperpriors.

**Scalability:** It is important to point out that our focus here is on model structure. We are currently engaged in designing practical computational methods that may be needed for large-scale problems. For problems involving a couple of hundreds of predictors (loci), the MCMC-based full Bayesian approach using WinBUGS/OpenBUGS as described here can efficiently be applied. DE LOS CAMPOS *et al.* (2009) applied a MCMC-based fully Bayesian approach to fit the BL to a data set involving more than 10,000 predictors.

## CONCLUSION

In this article we have proposed the extended Bayesian LASSO method for QTL mapping and unobserved phenotype prediction. The EBL introduces locus-specific regularization parameters and is parameterized so as to separate the overall model sparsity to the



degree of shrinkage of each marker effect. The explicit separation of these two features allows the EBL to overcome the curse of tuning, which may affect the performance of a similar approach proposed by SUN *et al.* (2010), which, like the BL, does not include a mechanism for separating model sparsity from parameter shrinkage. We are unaware of any previous attempt to separate these two aspects.

Under our model parameterization, the parameter representing the overall model sparsity is effectively estimated by pooling information (*i.e.*, borrowing strength) across loci, whereas locus-specific effects meant to enforce differential shrinkage across loci are obtained as deviations from the overall parameter. XU (2003) also proposed the normal Jeffreys' Bayesian shrinkage method, which does not require tuning. However, unlike Xu's approach, the methodology proposed here entertains the possibility of incorporating information relating to the degree of model sparsity through the prior imposed on the corresponding parameter.

Simulations demonstrated the good performance of the EBL with regard to the accuracy of parameter estimates and phenotype predictions, suggesting the potential of this method in QTL/association mapping and phenotype prediction, as well as genomic breeding value estimation and prediction in animal and plant breeding programs.

We are grateful for the constructive comments of two anonymous reviewers. This work was supported by a research grant from the Academy of Finland and University of Helsinki's research funds.

#### LITERATURE CITED

- CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- DE LOS CAMPOS, G., H. NAYA, D. GIANOLA, J. CROSSA, A. LEGARRA *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**: 375–385.
- GILKS, W.R., S. RICHARDSON and D. J. SPIEGELHALTER, 1996 *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- HOERL, A. E., and R. W. KENNARD, 1970 Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.
- LEE, S. H., J. H. J. VAN DER WERF, B. J. HAYES, M. E. GODDARD and P. M. VISSCHER, 2008 Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* **4**: e1000231.
- MALO, N., O. LIBIGER and N. J. SCHORK, 2008 Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am. J. Hum. Genet.* **82**: 375–385.
- O'HARA, R. B., and M. J. SILLANPÄÄ, 2009 A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.* **4**: 85–118.
- PARK, T., and G. CASELLA, 2008 The Bayesian LASSO. *J. Am. Stat. Assoc.* **103**: 681–686.
- PICARD, R. R., and R. D. COOK, 1984 Cross-validation of regression models. *J. Am. Stat. Assoc.* **79**: 575–583.
- PIKKUHOOKANA, P., and M. J. SILLANPÄÄ, 2009 Correcting for relatedness in Bayesian models for genomic data association analysis. *Heredity* **103**: 223–237.
- SEN, S., and G. A. CHURCHILL, 2001 A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.
- SERVIN, B., and M. STEPHENS, 2007 Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**: e114.
- SILLANPÄÄ, M. J., and N. NOYKOVA, 2008 Hierarchical modeling of clinical and expression quantitative trait loci. *Heredity* **101**: 271–284.
- SUN, W., J. G. IBRAHIM and F. ZOU, 2010 Genome-wide multiple loci mapping in experimental crosses by the iterative penalized regression. *Genetics* **185**: 349–359.
- THOMAS, A., R.B. O'HARA, U. LIGGES and S. STURTZ, 2006 Making BUGS Open. *R News* **6**: 12–17.
- TIBSHIRANI, R., 1996 Regression shrinkage and selection via LASSO. *J. R. Stat. Soc. B* **58**: 267–288.
- TINKER, N. A., D. E. MATHER, B. G. ROSNAGEL, K. J. KASHA, and A. KLEINHOF, 1996 Regions of the genome that affect agronomic performance in two-row barley. *Crop Sci.* **36**: 1053–1062.
- USAI, M. G., D. M. E. GODDARD and B. J. HAYES, 2009 LASSO with cross-validation for genomic selection. *Genet. Res.* **91**: 427–436.
- WANG, H., Y-M. ZHANG, X. LI, G. L. MASINDE, S. MOHAN *et al.*, 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**: 465–480.
- WANG, S., C. J. BASTEN, and Z-B. ZENG, 2006 *Windows QTL Cartographer 2.5*. Department of Statistics, North Carolina State University, Raleigh, NC.
- WHITTAKER, J. C., R. THOMPSON and M. C. DENHAM, 2000 Marker-assisted selection using ridge regression. *Genet. Res.* **75**: 249–252.
- WU, T. T., F. C. YI, T. HASTIE, E. SOBEL and K. LANGE, 2009 Genome-wide association analysis by LASSO penalized logistic regression. *Bioinformatics* **25**: 714–721.
- XU, C., X. WANG, Z. LI and S. XU, 2009 Mapping QTL for multiple traits using Bayesian statistics. *Genet. Res.* **91**: 23–37.
- XU, S., 2003 Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.
- XU, S., 2007 An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**: 513–521.
- XU, S., and Z. JA, 2007 Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics* **175**: 1955–1963.
- YI, N., and S. XU, 2008 Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045–1055.
- ZHANG, Y-M., and S. XU, 2005 A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* **95**: 96–104.
- ZOU, H., and T. HASTIE, 2005 Regularization and variable selection via elastic net. *J. R. Stat. Soc. B* **67**: 301–320.

Communicating editor: I. HOESCHELE