Heredity (2015) 114, 107–115

www.nature.com/hdy

ORIGINAL ARTICLE

Empirical Bayesian elastic net for multiple quantitative trait locus mapping

A Huang[1], S Xu[2] and X Cai[1]

In multiple quantitative trait locus (QTL) mapping, a high-dimensional sparse regression model is usually employed to account for possible multiple linked QTLs. The QTL model may include closely linked and thus highly correlated genetic markers, especially when high-density marker maps are used in QTL mapping because of the advancement in sequencing technology. Although existing algorithms, such as Lasso, empirical Bayesian Lasso (EBlasso) and elastic net (EN) are available to infer such QTL models, more powerful methods are highly desirable to detect more QTLs in the presence of correlated QTLs. We developed a novel empirical Bayesian EN (EBEN) algorithm for multiple QTL mapping that inherits the efficiency of our previously developed EBlasso algorithm. Simulation results demonstrated that EBEN provided higher power of detection and almost the same false discovery rate compared with EN and EBlasso. Particularly, EBEN can identify correlated QTLs that the other two algorithms may fail to identify. When analyzing a real dataset, EBEN detected more effects than EN and EBlasso. EBEN provides a useful tool for inferring high-dimensional sparse model in multiple QTL mapping and other applications. An R software package 'EBEN' implementing the EBEN algorithm is available on the Comprehensive R Archive Network (CRAN).
Heredity (2015) **114,** 107–115; doi:10.1038/hdy.2014.79; published online 10 September 2014

INTRODUCTION

Quantitative traits are usually controlled by multiple quantitative trait loci (QTLs) and environmental factors. Because of the physical linkage of multiple QTLs, gene–gene interactions (epistasis) and gene–environment interactions, it is highly desirable to analyze a large number of loci and environmental factors simultaneously in a single QTL model. As technology advancement in molecular genotyping has made high-density genomic markers available, including all markers in a single QTL model leads to a large number of model variables, typically much larger than the sample size. Two techniques often used in the inference of such high dimensional QTL models are variable selection and shrinkage operator.

Variable selection typically employs a stepwise search method in conjunction with a selection criterion such as the Bayesian information criterion (Schwarz, 1978) to identify a subset of all possible genetic effects that best explain the phenotypic variation (Bogdan *et al.*, 2008; Li *et al.*, 2009; Yu *et al.*, 2009). On the other hand, shrinkage methods such as Lasso (Tibshirani, 1996) and Bayesian Lasso (Park and Casella, 2008; Yi and Xu, 2008) include all variables in the model but use a penalty function of the variables or appropriate prior distributions for the variables to shrink most variables toward zero. Especially, the Bayesian shrinkage approach (O'Hara and Sillanpaa, 2009) has received considerable attention recently and been applied to multiple QTL mapping (Xu, 2003; Wang *et al.*, 2005; Hoti and Sillanpää, 2006; Huang *et al.*, 2007; Yi and Xu, 2008). All these Bayesian methods rely on the Markov Chain Monte Carlo (MCMC) simulation to fit the Bayesian model, which is computationally intensive and time consuming when a large number of effects are considered in the model.

Recently, we developed two efficient empirical Bayesian Lasso (EBlasso) algorithms using a two-level hierarchical model with normal and exponential priors (EBlasso-NE) or a three-level hierarchical model with normal, exponential and Gamma priors (EBlasso-NEG) for multiple QTL mapping (Cai *et al.*, 2011; Huang *et al.*, 2013), which was shown to outperform other shrinkage methods including Lasso and MCMC-based Bayesian shrinkage methods in terms of power of detection and false discovery rate (FDR). Similar to Lasso, our EBlasso and other Bayesian shrinkage methods typically selects one variable out of a group of highly correlated variables. When QTLs are located closely, these shrinkage methods may not select all QTLs. Recently, the elastic net (EN) (Zou and Hastie, 2005) was developed to handle the issue of correlated variables in high-dimensional sparse models where only a relatively small number of variables are nonzero. An MCMC-based Bayesian EN method was also proposed (Li and Lin, 2010).

In this paper, capitalizing on the idea of EN, we propose a Bayesian EN (BEN) model for multiple QTL mapping, and then develop a novel empirical Bayesian EN (EBEN) algorithm to infer the BEN model. The EBEN algorithm is very efficient because of a coordinate ascent strategy and other algorithmic techniques used. Simulation studies demonstrate that our EBEN algorithm outperforms EN and EBlasso. Real data analysis demonstrates the utility of our EBEN algorithm.

[1]Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, USA and [2]Department of Botany and Plant Sciences, University of California, Riverside, CA, USA
Correspondence: Professor X Cai, Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Dr, Coral Gables, FL 33146, USA.
E-mail: x.cai@miami.edu
Received 22 November 2013; revised 27 June 2014; accepted 4 July 2014; published online 10 September 2014

MATERIALS AND METHODS

Linear model of multiple QTLs

Let y_i be the value of a quantitative trait of the ith individual in a mapping population. Suppose we observe y_i, $i=1, \cdots, n$, of n individuals and collect them into a vector $\boldsymbol{y}=[y_1, y_2, \cdots, y_n]^T$. In these n individuals, suppose there are p environmental covariates observed and m genetic markers genotyped. Let covariate l and genotype of marker j of individual i be x_{Eil} and x_{Gij}, respectively. Let us define $\mathbf{x}_{Ei}=[x_{Ei1}, x_{Ei2}, \cdots, x_{Eip}]^T$ and $\mathbf{x}_{Gi}=[x_{Gi1}, x_{Gi2}, \cdots, x_{Gim}]^T$. Then we have the following linear regression model for $\boldsymbol{y}$:

$$\boldsymbol{y}=\mu+\mathbf{X}_E\boldsymbol{\beta}_E+\mathbf{X}_G\boldsymbol{\beta}_G+\boldsymbol{\varepsilon}, \tag{1}$$

where μ is the population mean, vectors $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_G$ represent the environmental effects and the genetic effects of all markers, respectively; matrices $\mathbf{X}_E=[\mathbf{x}_{E1}, \mathbf{x}_{E2}, \cdots, \mathbf{x}_{En}]^T$ and $\mathbf{X}_G=[\mathbf{x}_{G1}, \mathbf{x}_{G2}, \cdots, \mathbf{x}_{Gn}]^T$ are the corresponding design matrices of different effects; and $\boldsymbol{\varepsilon}$ is the residual error that follows a normal distribution with zero-mean and covariance $\sigma_0^2\mathbf{I}$.

The design matrix $\mathbf{X}_G$ depends on a specific genetic model. We adopt the widely used Cockerham genetic model (Cockerham, 1954), which defines the values of a marker effect as -0.5 and 0.5 for two genotypes in a back cross design, and -1, 0 and 1 for three genotypes having additive effect, and -0.5 and 0.5 for homozygotes and heterozygotes having dominance effect in an intercross (F_2) design. For simplicity, we only consider additive effects in (1), although the method developed in this paper is also applicable to the model with dominance effects. Epistatic effects can also be incorporated into (1) as carried out in (Xu, 2007; Cai *et al.*, 2011), and the EBEN algorithm developed in this paper is applicable to the model with epistatic effects. However, for the ease of presentation, we will use model (1) throughout the paper.

Defining $\boldsymbol{\beta}=[\boldsymbol{\beta}_E^T, \boldsymbol{\beta}_G^T]^T$, and $\mathbf{X}=[\mathbf{X}_E, \mathbf{X}_G]$, we can write (1) in a more compact form:

$$\boldsymbol{y}=\mu+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \tag{2}$$

Given p environmental covariates and m markers with additive effects, the size of matrix $\mathbf{X}$ is $n\times k$ where $k=p+m$. Our goal is to identify all possible environmental effects on and QTLs for $\boldsymbol{y}$ manifested as the nonzero elements of the regression coefficients $\boldsymbol{\beta}$. When the number of environmental factors and number of markers are large, $\boldsymbol{\beta}$ contains a large number of unknowns, which makes model inference a challenging problem. However, we would expect that a small portion of markers are QTLs and a small portion of environmental factors influence the trait, which implies that $\boldsymbol{\beta}$ is a sparse vector meaning that most elements of $\boldsymbol{\beta}$ are zero.

We have developed an efficient EBlasso algorithm to infer sparse $\boldsymbol{\beta}$ from (2). However, we observed that similar to Lasso (Tibshirani, 1996), EBlasso typically outputs at most one nonzero regression coefficient for a group of several highly correlated variables. If several QTLs are relatively close, their correlation is high. For example, if two QTLs have a distance $d=5$ centi-Morgan (cM), their correlation $R=e^{-2d}=0.9$ assuming that the distance follows the Haldane map function (Wu *et al.*, 2007). EBlasso apparently cannot identify such highly correlated QTLs simultaneously. Borrowing the idea of EN (Zou and Hastie, 2005), we will apply a two-level hierarchical prior to $\boldsymbol{\beta}$ in (2) that will yield an equivalent EN prior for $\boldsymbol{\beta}$. Then we will develop the EBEN algorithm that can handle correlated QTLs, and will be shown to outperform both EN and EBlasso.

Bayesian EN prior

The unknown parameters in model (2) are μ, σ_0^2 and $\boldsymbol{\beta}$. Although our main concern is $\boldsymbol{\beta}$, parameters μ and σ_0^2 need to be estimated so that we can infer $\boldsymbol{\beta}$. To this end, we assign a noninformative uniform prior to μ and σ_0^2, that is, $p(\mu)\propto 1$ and $p(\sigma_0^2)\propto 1$. Then we assume a two-level hierarchical model for $\boldsymbol{\beta}$. Let us denote the elements of $\boldsymbol{\beta}$ as β_j, $j=1,2,\cdots,k$. At the first level, β_j, $j=1,2,\cdots,k$, follow independent normal distributions with mean zero and unknown variance σ_j^2 : $\beta_j \sim N(0,\sigma_j^2)$. Let us define $\alpha_j=1/\sigma_j^2$, $j=1, 2, \cdots, k$, as precision of the normal prior distribution, and let $\boldsymbol{\sigma}^2=[\sigma_1^2,\sigma_2^2,\cdots,\sigma_k^2]^T$ and $\boldsymbol{\alpha}=[\alpha_1,\alpha_2, \cdots,\alpha_k]^T$. It turns out to be more convenient to estimate $\boldsymbol{\alpha}$ than $\boldsymbol{\sigma}^2$. At the second level, we decompose α_j as $\alpha_j=\lambda_1+\tilde{\alpha}_j$, $j=1, 2, \cdots, k$, where $\lambda_1\geqslant 0$ is a constant and $\tilde{\alpha}_j>0$ is a random variable whose distribution will be specified as follows. Defining $\tilde{\sigma}_j^2=1/\tilde{\alpha}_j$, we assign a generalized Gamma distribution to $\tilde{\sigma}_j^2$:

$$f(\tilde{\sigma}_j^2)=c\left(\lambda_1\tilde{\sigma}_j^2+1\right)^{-1/2}\exp\left(-\lambda_2\tilde{\sigma}_j^2\right),\ j=1,2,\cdots,k, \tag{3}$$

where c is a normalization constant, and $\lambda_2\geqslant 0$ is another constant.

The prior distribution (3) has two important properties. First, for $\lambda_1=0$, it becomes an exponential distribution $f(\tilde{\sigma}_j^2)=c\exp(-\lambda_2\tilde{\sigma}_j^2)$ with $c=\lambda_2$, and the distribution of β_j can be found to be the Laplace distribution $p(\beta_j)\propto\exp(-\sqrt{2\lambda_2}\mid\beta_j\mid)$, yielding the penalty used by Lasso (Tibshirani, 1996). Second, for given $\lambda_1>0$, $\lambda_2\geqslant 0$ and $c=\sqrt{\lambda_1\lambda_2/\pi}\exp(-\lambda_2/\lambda_1)$, $f(\tilde{\sigma}_j^2)$ becomes a shifted Gamma distribution (Pal *et al.*, 2005):

$$f(\tilde{\sigma}_j^2\mid a,b,\gamma)=\frac{b^a}{\Gamma(a)}(\tilde{\sigma}_j^2-\gamma)^{a-1}\exp\left(-b(\tilde{\sigma}_j^2-\gamma)\right),\ j=1,2,\cdots,k, \tag{4}$$

where $a=1/2$, $b=\lambda_2$ and $\gamma=-1/\lambda_1$, and the distribution of β_j can be found to be $p(\beta_j)\propto\exp(-\frac{\lambda_1}{2}\beta_j^2-\sqrt{2\lambda_2}\mid\beta_j\mid)$ (see Appendix A for the proof), yielding the penalty used by EN (Zou and Hastie, 2005). Throughout the paper, we will refer to the regression model (2) with the two-level hierarchical prior as the BEN model. Note that when $\lambda_1=0$, the prior distribution is the same as the normal-exponential (NE) prior of the EBlasso (EBlasso-NE) (Huang *et al.*, 2013), thus the EBlasso-NE model is a special case of the BEN model.

Let us define $\tilde{\boldsymbol{\sigma}}^2=[\tilde{\sigma}_1^2,\tilde{\sigma}_2^2,\cdots,\tilde{\sigma}_k^2]^T$ and collect all parameters that need to be estimated as $\boldsymbol{\theta}=(\mu,\sigma_0^2,\boldsymbol{\beta},\tilde{\boldsymbol{\sigma}}^2)$. The joint posterior distribution of $\boldsymbol{\theta}$ can be easily found and MCMC simulation can be employed to draw samples from the posterior distribution for each parameter (Robert and Casella, 2004). However, the fully Bayesian approach-based on MCMC sampling requires a prohibitive computational cost when the number of parameters $2k+4$ becomes relatively large. Here, we adopt the same strategy used by EBlasso to develop an efficient empirical Bayesian algorithm to infer the BEN model.

Maximum *a posteriori* estimation of variance components

We will show that $\tilde{\boldsymbol{\sigma}}^2$ can be estimated in closed-form, which will result in the efficient EBEN algorithm. The posterior distribution of parameters $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta}\mid\boldsymbol{y})\propto p(\boldsymbol{y}\mid\mu,\boldsymbol{\beta},\sigma_0^2)p(\mu)p(\sigma_0^2)p(\boldsymbol{\beta}\mid\tilde{\boldsymbol{\sigma}}^2)p(\tilde{\boldsymbol{\sigma}}^2\mid\lambda_1,\lambda_2) \tag{5}$$

The marginal posterior distribution of μ, $\tilde{\boldsymbol{\sigma}}^2$ and σ_0^2 can be written as $p(\mu,\tilde{\boldsymbol{\sigma}}^2,\sigma_0^2\mid\boldsymbol{y})=\int p(\boldsymbol{\theta}\mid\boldsymbol{y})d\boldsymbol{\beta}$, from which the log marginal posterior distribution of $\tilde{\boldsymbol{\alpha}}$ is derived as follows:

$$L(\tilde{\boldsymbol{\alpha}})=-\frac{1}{2}\left[\log|\mathbf{C}|+(\boldsymbol{y}-\mu)^T\mathbf{C}^{-1}(\boldsymbol{y}-\mu)\right]-\frac{1}{2}\sum_j^k\log\left(\frac{\lambda_1}{\tilde{\alpha}_j}+1\right)-\sum_j^k\frac{\lambda_2}{\tilde{\alpha}_j}+\text{constant}, \tag{6}$$

where $\mathbf{C}=\sigma_0^2\mathbf{I}+\Sigma_{j=1}^k(\lambda_1+\tilde{\alpha}_j)^{-1}\mathbf{x}_j\mathbf{x}_j^T$ is the covariance matrix of $\boldsymbol{y}$ with a given $\tilde{\boldsymbol{\alpha}}$.

Let us define $\mathbf{C}_{-j}=\mathbf{C}-(\lambda_1+\tilde{\alpha}_j)^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^T$. Then we can write $L(\tilde{\boldsymbol{\alpha}})$ in (6) as $L(\tilde{\boldsymbol{\alpha}})=L(\tilde{\boldsymbol{\alpha}}_{-j})+L(\tilde{\alpha}_j)$, where $L(\tilde{\boldsymbol{\alpha}}_{-j})$ does not depend on $\tilde{\alpha}_j$ and $L(\tilde{\alpha}_j)$ is given by

$$L(\tilde{\alpha}_j)=\frac{1}{2}\left[\log\frac{\tilde{\alpha}_j}{\tilde{\alpha}_j+\lambda_1+s_j}+\frac{q_j^2}{\tilde{\alpha}_j+\lambda_1+s_j}\right]-\frac{\lambda_2}{\tilde{\alpha}_j}, \tag{7}$$

with $s_j=\boldsymbol{x}_j^T\mathbf{C}_{-j}^{-1}\boldsymbol{x}_j$ and $q_j=\boldsymbol{x}_j^T\mathbf{C}_{-j}^{-1}(\boldsymbol{y}-\mu)$. It is seen that (7) is similar to $L(\alpha_j)$ of EBlasso-NE (Huang, *et al.*, 2013, Equation (11)) except that λ_1 appears in the denominators of the first two terms. Therefore, as shown in Appendix B, $L(\tilde{\alpha}_j)$ has a unique global maximum and the optimal $\tilde{\alpha}_j$ maximizing $L(\tilde{\alpha}_j)$ is given by

$$\tilde{\alpha}_j^*=\begin{cases} r, \text{if } q_j^2-s_j>\lambda_1+2\lambda_2 \\ \infty, \text{otherwise,}\end{cases} \tag{8}$$

where $r=\frac{-(s_j+\lambda_1+4\lambda_2)-\sqrt{\Delta}}{2(s_j-q_j^2+\lambda_1+2\lambda_2)}\cdot(s_j+\lambda_1)$, and $\Delta=(s_j+\lambda_1)^2+8\lambda_2q_j^2$. Of note, if $\tilde{\alpha}_j=\infty$, then $\sigma_j^2=0$, which is equivalent to $\beta_j=0$.

EBEN algorithm and statistical significance test

Similar to the EBlasso algorithm, EBEN employs a coordinate ascent method to estimate unknown parameters $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_k$, μ and σ_0^2. After these parameters are estimated, the posterior distribution of $\boldsymbol{\beta}$, which is a Gaussian distribution, can be found. Specifically, in each cycle of the coordinate ascent method, $\tilde{\alpha}_j$ is estimated from (8) with all other parameters fixed, and μ and σ_0^2 are estimated using Equations (15) and (16) in Cai *et al.* (2011), respectively. In the initial cycle, only one appropriately selected $\tilde{\alpha}_j$ is finite (Cai *et al.*, 2011), which corresponds to a model with only one variable $\boldsymbol{x}_j$. In the following cycles, a variable $\boldsymbol{x}_l$ is added to the model if $\tilde{\alpha}_l$ is finite, or is removed from the model if $\tilde{\alpha}_l$ is infinite. The iterative process continues until convergence criterion is satisfied. Specifically, the following convergence criteria are applied: (i) no effect can be added to or deleted from the model, (ii) the change of $\tilde{\alpha}_j$ between two consecutive iterations, $|\Delta\tilde{\alpha}_j|$, is smaller than a pre-specified small value, and (iii) the Euclidean norm of the change of $\tilde{\boldsymbol{\alpha}}$ between two consecutive iterations, $\|\Delta\tilde{\boldsymbol{\alpha}}\|_2$, is less than a pre-specified value. During the iteration, many $\tilde{\alpha}_j$s will be infinite, and the corresponding β_js are zero.

The EBEN algorithm can be obtained from the EBlasso algorithm (Cai *et al.*, 2011) with the following two modifications: (i) replace α_j with $\alpha_j = \lambda_1 + \tilde{\alpha}_j$ and estimate $\tilde{\alpha}_j$ from (8), and (ii) replace hyperparameters (a, b) in EBlasso with λ_1 and λ_2 and use cross validation (CV) to determine λ_1 and λ_2. A step-by-step description for the EBlasso algorithm is given in (Cai *et al.*, 2011). The EBEN algorithm is provided in Appendix C.

The EBEN algorithm will select k' (typically $k' \ll k$) nonzero elements of $\boldsymbol{\beta}$, which is denoted as a $k' \times 1$ vector $\boldsymbol{\beta}'$, that corresponds to finite $\tilde{\alpha}_j$s. Let $\hat{\boldsymbol{\alpha}}$ be a $k' \times 1$ vector contain all finite $(\lambda_1 + \tilde{\alpha}_j)$s. Given $\hat{\boldsymbol{\alpha}}$, it is not difficult to show that the posterior distribution of $\boldsymbol{\beta}'$ is a Gaussian distribution with mean $\hat{\boldsymbol{\beta}}' = \sigma_0^2 \hat{\boldsymbol{\Sigma}} \tilde{\boldsymbol{X}} (\mathbf{y} - \mu)$ and covariance $\hat{\boldsymbol{\Sigma}} = (\mathbf{A} + \sigma_0^{-2} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1}$, where $\tilde{\mathbf{X}}$ is an $n \times k'$ matrix that contains the columns of $\mathbf{X}$ corresponding to $\boldsymbol{\beta}'$, and $\mathbf{A}$ is a diagonal matrix with $\hat{\boldsymbol{\alpha}}$ on its diagonal. Note that given $\mathbf{A}$, $\boldsymbol{\beta}'$ is equivalent to the best linear unbiased prediction of $\boldsymbol{\beta}'$ in the linear model with k' random effects. For the jth element of $\boldsymbol{\beta}'$,β_j', the Bayesian approach needs to calculate the Bayesian factor to determine the significance of hypothesis H_1: $\beta_j' \neq 0$ against hypothesis H_0: $\beta_j' = 0$. However, the Bayesian factor is not easy to calculate. One way to overcome this problem is to employ the EBEN algorithm to select variables and then use the multisplit method (Meinshausen and Bühlmann, 2010) to determine the statistical significance of selected regression coefficients. However, the multisplit method is computationally demanding and its conservative approach to calculating *P*-values may reduce the power of detection. In this paper, we will use the following *t*-test to determine the significance of β_j' and compare its performance with that of the multisplit method. Because the standard deviation of β_j' in the posterior distribution is $s_j = (\hat{\Sigma}_{jj})^{1/2}$, where $\hat{\Sigma}_{jj}$ is the jth diagonal element of $\hat{\boldsymbol{\Sigma}}$, we will use the *t*-statistics $\hat{\beta}_j'/s_j$ to test if $\hat{\beta}_j' \neq 0$ at 0.05 significance level. Essentially, we assume that the posterior distribution of β_j' follows Student's *t*-distribution and use the 0.95 credible interval to determine if $\hat{\beta}_j' \neq 0$.

Cross validation

Two hyperparameters $\lambda_1 \geqslant 0$ and $\lambda_2 \geqslant 0$ need to be determined with CV. To facilitate CV, we define λ_1 and λ_2 in terms of other two parameters $\lambda > 0$ and $\upsilon \in [0,1]$: $\lambda_1 = (1-\upsilon)\lambda$ and $\lambda_2 = \upsilon\lambda$. Note that when $\upsilon = 1$, EBEN is equivalent to EBLasso-NE for a given pair of λ and υ. We perform fivefold CV and calculate the prediction error (Tibshirani, 1996) $PE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the predicted phenotype. We calculate $\lambda_{\max} = \arg_j \max |\boldsymbol{x}_j^T(\boldsymbol{y} - \mu)|$, and chose a set of values for λ decreasing from $\lambda_{\max}$ to 0.001 $\lambda_{\max}$ in 20 even steps on the logarithmic scale. We vary υ from 1 to 0 at a step size of 0.05, and for each υ we repeat CV for all values of λ from $\lambda_{\max}$ to 0.001 $\lambda_{\max}$. The pair of (υ, λ) that yields the smallest prediction error is chosen to be the optimal parameters, which are then used by EBEN to infer the model.

Simulation setup and real data analysis

We simulated a population of an F_2 family derived from the cross of two inbred lines with $m = 481$ genetic markers which were evenly spaced on a large chromosome of 2400 cM (interval $d = 5$ cM). The dummy variable for the three genotypes, A_1A_1, A_1A_2 and A_2A_2 of individual i at marker j was defined as $x_{ij} = 1,\ 0,\ -1$, respectively. We assumed that QTLs were coincided with markers. If QTLs were not on markers, they may still be detected because correlation between a QTL and a nearby marker was high, although a slightly larger sample size may be needed to give the same power of detection.

We performed two sets of simulations based on the F_2 population, each with 50 QTLs, whose effect sizes were randomly generated from a normal distribution with mean zero and variance equal to four. Environmental effects were not simulated. The true population mean was $\mu = 100$ and the residual variance was $\sigma_0^2 = 10$. In the first set of simulations (*SimI*), 10 groups of two adjacent markers were selected randomly as QTLs; the minimum distance between any two groups of QTLs was 65 cM. The remaining 30 QTLs were selected randomly from the remaining markers. In the second set of simulations (*SimII*), 10 groups of five consecutive markers were randomly selected as 50 QTLs; the minimum distance between any two groups was 25 cM. For each set of simulations, 100 replicates were generated with sample sizes 200, 400, 600, 800 and 1000, and analyzed using EBEN, EBLasso-NEG (Cai *et al.*, 2011) and EN (Zou and Hastie, 2005). Power of detection, FDR and power of detecting groups of QTLs of three methods were compared.

We used a barley double haploid population published by Hayes *et al.* (1993) as an example to test our method. The dataset consisted of $n = 150$ double haploid derived from the cross of two spring barley varieties, Morex and Steptoe. The total number of markers was $q = 495$ distributed along seven pairs of chromosomes of the barley genome. The traits included three agronomic traits, grain yield, heading date and height, and five malting quality traits, lodging, grain protein, alpha amylase, diastatic power and malt extract. The marker intervals ranged from 0.6 to 23.3 cM, with median interval size 1.4 cM. With such high-density markers, correlations among markers were high. Genotype of the markers were encoded as $+1$ for genotype A (the Steptoe parent), -1 for genotype B (the Morex parent), and 0 for missing genotype. The total missing genotypes account for about 4.2% of all the genotypes.

RESULTS

Estimated effects for one replicate in *SimI*

For a replicate of *SimI* shown in Table 1, we obtained the total phenotypic variance for the trait by

$$\sigma_\gamma^2 = \sigma_0^2 + \sum_{j=1}^{m}\sum_{j'=1}^{m} \beta_j \beta_{j'} \mathrm{cov}(\boldsymbol{x}_j, \boldsymbol{x}_{j'}), \tag{9}$$

where $\mathrm{cov}(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$ is the covariance between $\boldsymbol{x}_j$ and $\boldsymbol{x}_{j'}$ if $j \neq j'$ or the variance of $\boldsymbol{x}_{j'}$ if $j = j'$, which can be estimated from the data. For this example, $n = 1000$ samples were used. The total phenotypic variance was calculated from (9) to be $\sigma_\gamma^2 = 102.44$ and the total genetic variance contributed by the main effects of the markers was calculated as 92.44. If we ignore the contributions from the covariance terms that are relatively small, the proportions of the phenotypic variance explained by a particular QTL effect j can be approximated by

$$h_j^2 = \frac{\beta_j^2 \mathrm{var}(\boldsymbol{x}_j)}{\sigma_\gamma^2} \tag{10}$$

where $\mathrm{var}(\boldsymbol{x}_j)$ is the variance of X_j. In the simulated data, the proportion of contribution from an individual QTL varied from 0.01% to 6.61% as shown in Table 1.

The data were analyzed in R on a personal computer using EBEN, EN (Zou and Hastie, 2005) and EBLasso-NEG (Cai *et al.*, 2011). We obtained the program *glmnet* (Friedman *et al.*, 2010) that implements EN. CV for EBEN determined the optimal υ and λ as $\upsilon = 0.95$ and $\lambda = 0.0072$. With the two values, EBEN identified 54 effects with a *P*-value $\leqslant 0.05$. We counted multiple identified effects that were within 20 cM distance to a true QTL as a true effect, and all effects with more than 20 cM distance to any true QTLs as false effects, resulted in 47 true effects and 3 false effects (Table 1). Simulated effects and QTLs identified by EBEN are visualized in Figure 1 (top).

Table 1 True and estimated effects for the simulated data

Locus[a]	*cM*	*True* $\hat{\beta}(h^2)$	*EBEN* $\hat{\beta}$[b]	*EN* $\hat{\beta}$[c]	*EBlasso* $\hat{\beta}$
1	0	0.73 (0.0027)	0.65	−0.54[d]	0.94
39	190	0.52 (0.0013)	0.60	—	—
40	195	0.52 (0.0013)	0.49	—	0.77
67	330	0.22 (0.0002)	0.21[d]	—	0.27[d]
96	475	0.37 (0.0007)	0.46	0.45	0.48
128	635	2.32 (0.0271)	2.18	2.37	2.20
146	725	0.79 (0.0029)	0.79	0.87	1.18
147	730	0.70 (0.0023)	0.67	—	—
148	735	1.29 (0.0079)	1.30	1.23	1.74
156	775	0.37 (0.0007)	3.26	—	—
157	780	3.48 (0.0593)	0.53[d]	3.26	3.87
168	835	−3.31 (0.0509)	−3.42	−3.54	−3.64
172	855	−0.87 (0.0036)	−0.56	—	−0.80[d]
184	915	−0.63 (0.0019)	−0.51[d]	−0.89[d]	−0.51[d]
196	975	1.11 (0.0060)	0.68	—	0.84
197	980	2.24 (0.0251)	2.32	2.44	2.46
217	1080	2.04 (0.0203)	1.66	1.59	1.73
218	1085	2.72 (0.0357)	2.87	3.15	3.72
245	1220	2.35 (0.0272)	2.17	2.19	2.49
261	1300	1.27 (0.0076)	1.21	0.98	1.37
268	1335	1.47 (0.0108)	1.21	1.21	1.37
283	1410	−0.95 (0.0044)	−1.19	−1.22	−1.16
284	1415	−1.05 (0.0052)	−0.83	−0.98	−1.03
288	1435	1.44 (0.0102)	1.25	1.65	1.59
290	1445	−1.38 (0.0097)	−1.08	−1.19	−1.42
294	1465	−2.13 (0.0222)	−1.87	−1.74	−2.19
304	1515	−3.59 (0.0661)	−3.12	−2.94	−3.53
315	1570	1.09 (0.0057)	0.79	0.81	1.08
351	1750	−0.95 (0.0041)	−1.07	−1.02	−1.51
352	1755	−1.16 (0.0063)	−0.71	—	—
353	1760	−1.13 (0.0062)	−1.39	−1.48	−1.85
358	1785	1.67 (0.0133)	1.21	1.28	1.69
361	1800	−3.25 (0.0514)	−2.74	−2.55	−3.10
370	1845	−3.43 (0.0571)	−4.01	−3.60	−3.77
371	1850	−1.36 (0.0090)	−1.08	−1.73	−1.43
375	1870	0.96 (0.0044)	0.83	1.23	0.96
382	1905	−2.93 (0.0415)	−2.83	−3.06	−2.99
392	1955	1.20 (0.0073)	0.99	1.15	1.17
397	1980	−2.06 (0.0212)	−1.77	−1.86	−1.91
408	2035	−1.15 (0.0064)	−0.96	−0.92	−1.14
409	2040	−0.53 (0.0013)	−0.94[d]	−0.96[d]	−1.10[d]
411	2050	−1.67 (0.0141)	−0.89	—	−1.22
426	2125	0.22 (0.0002)	—	—	—
436	2175	0.10 (0.0001)	—	—	—
442	2205	1.37 (0.0090)	1.20	1.30	1.27
457	2280	−1.72 (0.0138)	−1.90	−1.81	−2.08
462	2305	−0.42 (0.0008)	—	—	—
476	2375	1.29 (0.0079)	0.67	—	0.64
477	2380	1.57 (0.0118)	1.75	1.64	1.66
478	2385	1.67 (0.0137)	1.99	2.01	2.09
True/False positive			47/3	37/4	43/2

Abbreviations: cM, centi-Morgan; EBEN, empirical Bayesian elastic net; EBlasso, empirical Bayesian Lasso; EN, elastic net.
[a]Groups of neighboring QTLs are highlighted.
[b]The estimated effect is denoted by $\hat{\beta}$.
[c]EN selected variables were refitted to a linear regression model to obtain $\hat{\beta}$.
[d]The estimated marker effect was obtained from a neighboring marker ($\leqslant$20 cM) rather than from the true QTL.

EN has the same pair of parameters υ and λ as EBEN, and for each υ, λ is chosen from $\lambda_{\max}$ to 0.001 $\lambda_{\max}$ in 100 even steps on the logarithmic scale. CV gave the optimal values $(\upsilon, \lambda) = (0.95, 0.0734)$. Using these optimal values, EN identified 116 markers with nonzero regression coefficients. However, EN does not give a *P*-value for each estimated coefficient. If we regarded all 116 effects as QTLs, we would get a large number of false QTLs. To avoid this problem, we refitted an ordinary linear regression model with the 116 markers and calculated a *P*-value for each marker. Among those markers with a *P*-value $\leqslant 0.05$ in the refitted model, 37 markers corresponded to true-positive effects and 4 corresponded to false-positive effects were identified. The estimated sizes of 37 true effects and their standard deviations are listed along with all 50 true effects in Table 1, and QTLs identified are depicted in Figure 1 (bottom).

EBLasso-NEG (Cai *et al.*, 2011) has two hyperparameters a and b controlling the degree of shrinkage, and CV chose the optimal values $(a, b) = (-0.9, 1)$. Using the values, EBlasso-NEG identified 43 true- and 2 false-positive effects with a *P*-value $\leqslant 0.05$. The estimated sizes of true effects and their standard deviations are listed in Table 1, and QTLs identified are plotted in Figure 1 (middle).

Comparing the results of three algorithms, we observe that EBEN detected the most number of true effects, whereas three methods yielded similar number of false-positive effects. To see if the three algorithms can detect correlated effects, we highlight 10 groups that include neighboring markers in Table 1. Because the genetic markers were simulated with Haldane map function, the correlation between two neighboring markers is 0.9048, and the correlation between every other neighboring marker is 0.8187. It is seen that EBEN missed only markers 157 and 409 but detected them from other nearby markers. However, EN and EBlasso-NEG missed at least one QTL of 7 and 6 groups, respectively.

Results for *SimI* and *SimII*

The power of detection, FDR and power of detecting groups obtained from *Sim I* using EBEN, EBLasso and EN were plotted in Figure 2. As described in the Materials and Methods section, 10 groups of highly correlated QTLs are present in *Sim I*. When computing the power of detecting groups, a group was detected if all effects in the group were detected. From Figure 2, we observed that EBEN offered the highest power of detection and all three methods provided similar FDR; EBEN also had the highest power of detecting groups of QTLs as expected. Both EBEN and EBlasso-NEG outperformed EN. Taking sample size $n = 400$ as an example, we see that the power of detection, FDR and power of detecting groups are 0.82, 0.11 and 0.64, respectively, for EBEN, 0.76, 0.11 and 0.48, respectively, for EBlasso-NEG, and 0.53 0.15, and 0.26, respectively, for EN.

In *SimI*, there were several groups with three effects because of random selection of QTL locations. However, more than 25 out of the 50 QTLs were not in any group, which means that none of its neighboring markers were also QTL (see Table 1 for an example). In *SimII*, all effects were within one of the groups. The power of detection, FDR and power of detecting groups of the three methods were plotted in Figure 3, which shows that EBEN performed much better than the other two methods in terms of power of detection whereas three methods yielded similar FDR. Again, taking sample size $n = 400$ as an example, we see that the power of detection, FDR and power of detecting groups are 0.81, 0.10 and 0.35, respectively, for EBEN, 0.65, 0.12 and 0.14, respectively, for EBlasso-NEG, and 0.52, 0.16 and 0.05, respectively, for EN. Comparing the results of *SimI* and *SimII*, we observed that the performance of EBlasso and EN were degraded when the degree of grouping increased, whereas EBEN offered relatively stable power of detection and FDR.

As described in the Materials and Methods section, the multisplit method (Meinshausen *et al.*, 2009) can be another choice for testing

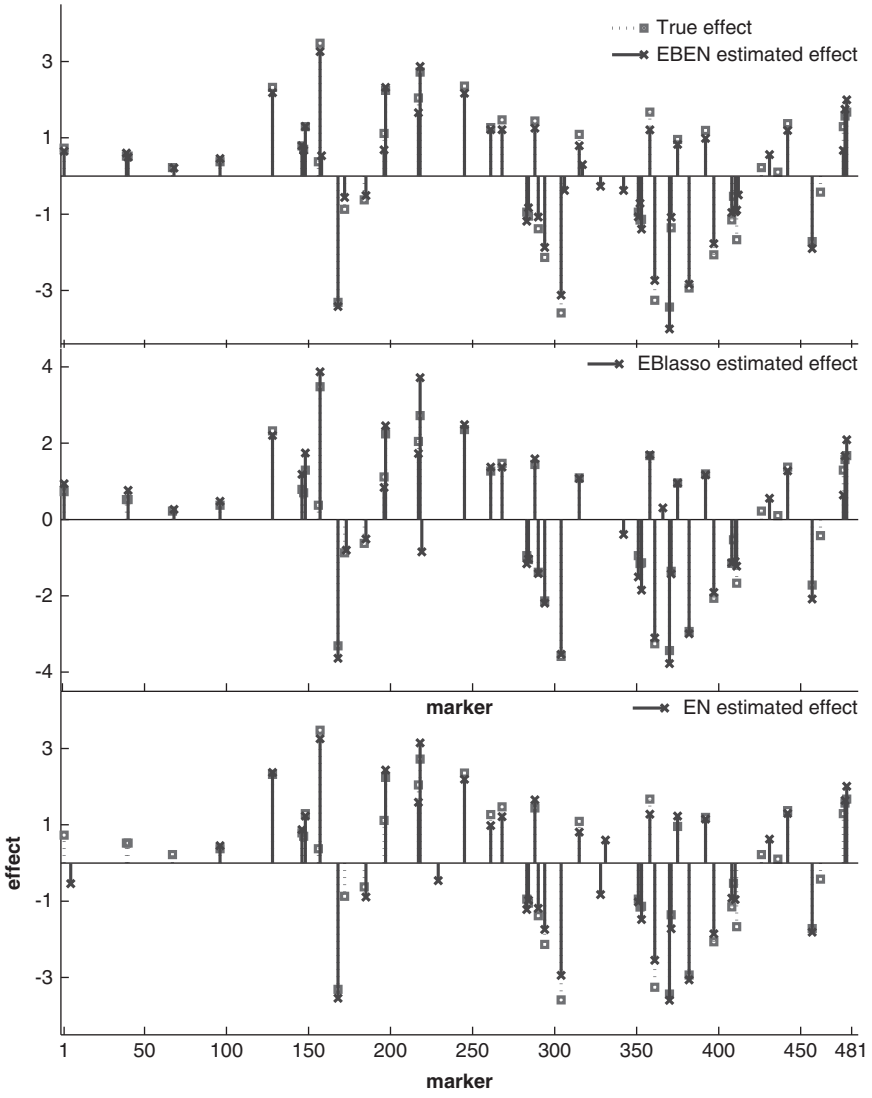

Figure 1 Effects estimated with EBEN, EBlasso and EN for the simulated data.

the significance of nonzero regression coefficients (Li and Sillanpää, 2012). To see the performance of the mulisplit method, we applied it to *SimI* and *SimII* with two sample sizes $N = 400$ and $N = 600$ and compared the power of detection, FDR and power of detecting groups of the multisplit method and the *t*-test, at a family-wise error rate of 0.05. The results are shown in Supplementary Figures S1 and S2 in the Supplementary Information. From Supplementary Figure S1 and S2, we observe that the *t*-test offered significantly higher power of detection than the multisplit method; its FDR was higher than the zero FDR of the multisplit method, but was still very low, less than 0.04.

Real data analysis

This dataset was used as an example for the application of EBEN in QTL mapping with high-density markers. We analyzed all eight traits but only presented results for three agronomic traits while leaving results for the five malting quality traits in the Supplementary Information.

In QTL mapping for grain yield, fivefold CV chose $(\upsilon, \lambda) = (0.35, 0.1710)$ for EBEN. With these values, EBEN identified 12 QTLs with a *P*-value $\leqslant 0.05$, which explained 57.93% of the total phenotypic variance (Table 2). Fivefold CV determined optimal parameters $(\upsilon, \lambda) = (0.65, 0.0354)$ for EN. Using the optimal values, EN identified 42 nonzero effects; after refitting the phenotype to the ordinary linear regression model with these 42 markers, six QTLs with a *P*-value $\leqslant 0.05$ were identified. The total phenotypic variance explained by six QTLs was 23.72%. Fivefold CV selected $(a, b) = (0.01, 0.05)$ for EBlasso-NEG, with which six effects with a *P*-value $\leqslant 0.05$ were identified. These six effects explained 51.08% of the total phenotypic variance. The identified QTLs using the three methods are listed in Table 2.

In QTL mapping for heading date, EBEN identified 14 QTLs having a *P*-value $\leqslant 0.05$ with CV selected optimal parameters, and 93.87% of the total phenotypic variance were explained (Table 3). EN identified 59 nonzero effects; after refitting the phenotype to the ordinary linear regression model with these 59 markers, 11 QTLs with a *P*-value $\leqslant 0.05$ were identified. The total phenotypic variance explained by six QTLs was 70.28%. EBlasso-NEG identified eight effects with a *P*-value $\leqslant 0.05$, which explained 91.13% of the total phenotypic variance. The identified QTLs using the three methods are listed in Table 3.

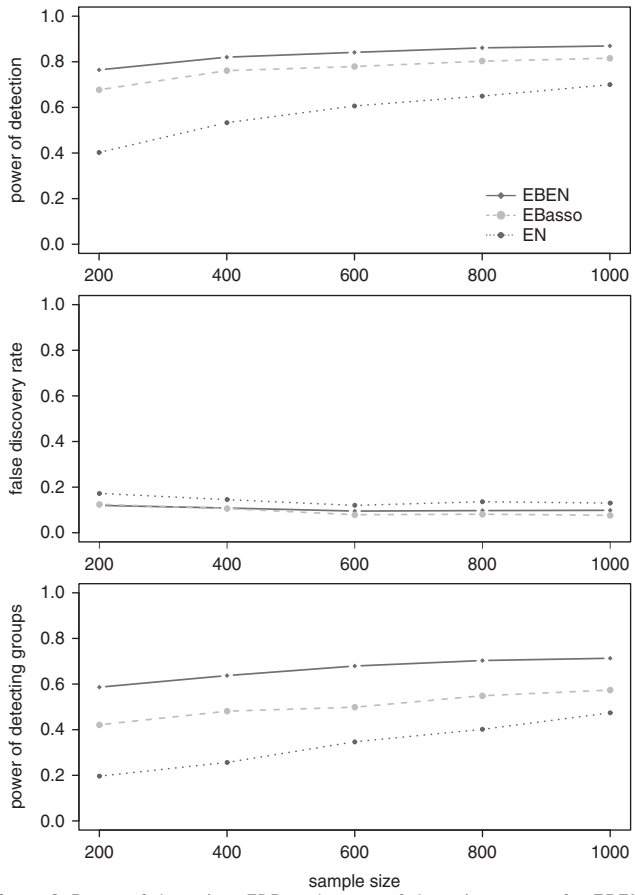

Figure 2 Power of detection, FDR and power of detecting groups for EBEN, EBlasso and EN in *SimI*.

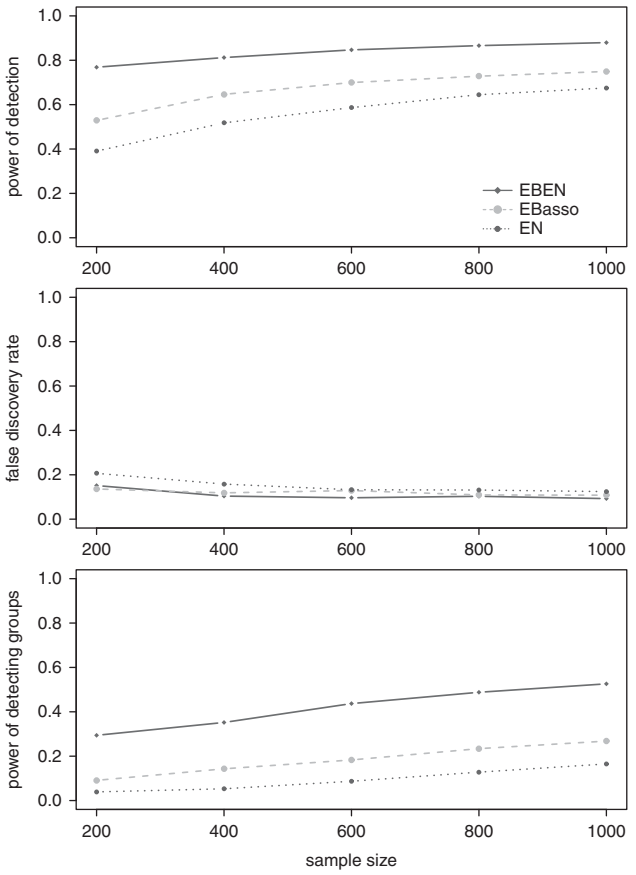

Figure 3 Power of detection, FDR and power of detecting groups for EBEN, EBlasso and EN in *SimII*.

In QTL mapping for the height of barley, using CV-selected optimal parameter values, EBEN identified 16 QTLs with a *P*-value $\leqslant 0.05$, which explained 93.29% of the total phenotypic variance (Table 4); EN identified 52 nonzero effects, with which 9 QTLs with a *P*-value $\leqslant 0.05$ were identified by refitting the phenotype to the ordinary linear regression model, and 44.30% of the total phenotypic variance were explained; EBlasso-NEG identified 9 effects with a *P*-value $\leqslant 0.05$, which explained 87.67% of the total phenotypic variance. The identified QTLs using the three methods are listed in Table 4.

Apparently, EBEN detected more effects than EN and EBlasso-NEG, although it also missed some of the effects detected by EN and EBlasso-NEG. Moreover, effects detected by EBEN explained more phenotypic variance than those detected by EN or EBlasso-NEG. Particularly, EBEN detected markers 403 and 406 for grain yield, 96, 97 and 98 for heading date, 74 and 75 for height, which were 4.1, 0.7 and 0.8 cM apart, respectively, and were highly correlated, but both EN and EBlasso-NEG were able to detect only one effect for each group. Results for other five traits are listed in Supplementary Tables S1–S5, which also shows that EBEN detected more effects, and these effects explained more phenotypic variance for all five traits compared with EN and EBlasso-NEG.

In computer simulations, it was observed that the computational time for the three methods was mainly determined by the number of nonzero markers in the inferred QTL model, and that EBEN had a speed similar to EN and EBlasso-NEG. In the analysis of the grain yield with the optimal hyperparameters chosen by CV, the

Table 2 QTLs and their effects obtained with EBEN, EN and EBlasso-NEG for the grain yield of barley

Marker IDs	*Position (Chr,cM)*	*EBEN* $\hat{\beta}$[a]	*EN* $\hat{\beta}$[b]	*EBlasso-NEG* $\hat{\beta}$
10	(1, 29.1)	−0.04	−0.09	—
40	(1, 97.8)	0.03	—	—
49	(1, 127.8)	—	—	0.06
57	(1, 139.6)	0.05	—	—
65	(2, 15.8)	−0.04	—	—
145	(3, 9.8)	—	—	−0.03
161	(3, 55.7)	0.29	—	0.25
193	(3, 118.1)	0.03	—	—
262	(4, 90.9)	—	—	0.03
294	(5, 45.4)	—	0.09	—
314	(5, 70.8)	−0.06	−0.12	—
347	(5, 154)	0.07	0.10	0.04
403	(6, 66.7)	−0.04	−0.11	−0.06
406	(6, 70.8)	−0.04	—	—
429	(7, 0.0)	−0.07	−0.12	—
492	(7, 134.8)	0.03	—	—
	Parameters[c]	(0.35, 0.1710)	(0.65, 0.0354)	(0.01, 0.05)
	Number of QTLs	12	6	6
	h^2	0.58	0.24	0.51

Abbreviations: cM, centi-Morgan; EBEN, empirical Bayesian elastic net; EBlasso, empirical Bayesian Lasso; EBlasso-NEG, empirical Bayesian Lasso-normal, exponential and Gamma; EN, elastic net; QTL, quantitative trait locus.
[a]The estimated marker effect is denoted by $\hat{\beta}$.
[b]Markers selected by EN were refitted to a linear regression model to obtain $\hat{\beta}$.
[c]Parameters are (ν, λ) for EBEN and EN, and (a, b) for EBlasso-NEG.

Table 3 QTLs and their effects obtained with EBEN, EN and EBlasso-NEG for the heading date of barley

Marker IDs	*Position (Chr,cM)*	*EBEN* $\hat{\beta}$[a]	*EN* $\hat{\beta}$[b]	*EBlasso-NEG* $\hat{\beta}$
11	(1,34.7)	—	0.48	—
51	(1,129.9)	0.32	—	0.28
74	(2,35.5)	−3.60	−2.61	−3.55
76	(2,37.8)	—	−0.55	—
96	(2,66.9)	0.22	—	—
97	(2,67.6)	0.57	1.22	1.39
98	(2,68.3)	0.34	—	—
129	(2,148.4)	—	−0.40	—
132	(2,155.3)	—	0.38	—
141	(3,3.6)	—	−0.38	—
164	(3,61.3)	—	—	0.24
176	(3,80.2)	0.14	—	—
244	(4,54.3)	0.22	—	—
252	(4,68.2)	0.25	—	0.47
271	(4,107.5)	0.24	0.24	0.25
278	(4,139)	0.17	—	—
352	(5,166.4)	0.15	—	0.1
375	(6,5.9)	−0.26	—	—
430	(7,1.3)	—	0.29	—
440	(7,13.7)	—	−0.68	—
445	(7,22.2)	—	—	−0.40
449	(7,29.2)	−0.19	—	—
452	(7,34)	—	−0.46	—
482	(7,101.8)	0.17	—	—
Parameters[c]		(0.70, 0.2139)	(0.85, 0.0652)	(1, 1)
Number of QTLs		14	11	8
h^2		0.94	0.70	0.91

Abbreviations: cM, centi-Morgan; EBEN, empirical Bayesian elastic net; EBlasso, empirical Bayesian Lasso; EBlasso-NEG, empirical Bayesian Lasso-normal, exponential and Gamma; EN, elastic net; QTL, quantitative trait locus.
[a]The estimated marker effect is denoted by $\hat{\beta}$.
[b]Markers selected by EN were refitted to a linear regression model to obtain $\hat{\beta}$.
[c]Parameters are (v, λ) for EBEN and EN, and (a, b) for EBlasso-NEG.

Table 4 QTLs and their effects obtained with EBEN, EN and EBlasso-NEG for the height of barley

Marker IDs	*Position (Chr,cM)*	*EBEN* $\hat{\beta}$[a]	*EN* $\hat{\beta}$[b]	*EBlasso-NEG* $\hat{\beta}$
74	(2,35.5)	−4.10	−2.88	−5.61
75	(2,36.3)	−0.74	—	—
97	(2,67.6)	0.66	—	1.29
128	(2,146.3)	−1.17	−1.06	−1.31
151	(3,32.5)	−0.58	—	—
159	(3,54.4)	−2.62	−2.08	−2.99
169	(3,68.2)	−0.34	−0.94	—
216	(3,157.7)	−0.36	—	—
240	(4,50.8)	0.84	—	1.08
321	(5,80.5)	−1.90	−2.67	−2.60
339	(5,139.3)	0.80	1.00	0.94
368	(5,200.2)	—	−0.50	—
375	(6,5.9)	−0.51	−1.01	—
415	(6,81.1)	0.50	0.92	—
474	(7,76.1)	0.77	—	0.89
481	(7,96.2)	0.34	—	—
487	(7,113.1)	—	—	1.11
488	(7,114.4)	0.50	—	—
Parameters[c]		(1.00, 0. 0657)	(0.85, 0.2341)	(−0.6, 0.1)
Number of QTLs		16	9	9
h^2		0.93	0.44	0.88

Abbreviations: cM, centi-Morgan; EBEN, empirical Bayesian elastic net; EBlasso, empirical Bayesian Lasso; EBlasso-NEG, empirical Bayesian Lasso-normal, exponential and Gamma; EN, elastic net; QTL, quantitative trait locus.
[a]The estimated marker effect is denoted by $\hat{\beta}$.
[b]Markers selected by EN were refitted to a linear regression model to obtain $\hat{\beta}$.
[c]Parameters are (v, λ) for EBEN and EN, and (a, b) for EBlasso-NEG.

computational time was 0.10 s for EBEN, 0.05 s for EN and 0.06 s for EBlasso-NEG. All computations were performed on a personal computer with a 2.6 GHz Intel Core 2 CPU and 4 Gb memory running Windows7.

DISCUSSION

We have developed a novel EBEN algorithm for multiple QTL mapping. Simulation results demonstrated that our EBEN outperformed two other algorithms EN (Zou and Hastie, 2005) and EBlasso-NEG (Cai *et al.*, 2011). Particularly, EBEN could detect more correlated effects than other two algorithms. When applied to a real barley dataset, EBEN was able to detect more QTLs and explain higher proportion of phenotypic variance than other two algorithms.

Our EBEN model essentially uses the same prior for regression coefficients as the one used by EN. For model inference, our EBEN first estimates the covariance of regression coefficients. During the estimation process, many coefficients are shrunk to zero if the corresponding variance is zero. After the covariance is obtained, the nonzero coefficients were estimated as a Gaussian random vector with an estimated mean and an estimated covariance. On the other hand, EN directly estimates the nonzero regression coefficients without estimating the covariance. Because our EBEN yields not only a point estimate of regression coefficients but also an estimate of their covariance, this gives more information than the point estimate of EN, which may help to improve performance. Our EBEN model and the Bayesian EN model in Li and Lin (2010) have some similarities and differences. The model of Li and Lin assumes the following prior: $\boldsymbol{\beta}\,|\,\boldsymbol{\tau}, \sigma_0^2 \sim \prod_{j=1}^{k} N(0, [\frac{\lambda_2}{\sigma_0^2}\frac{\tau_j}{\tau_j - 1}]^{-1})$, $\boldsymbol{\tau}\,|\,\sigma_0^2 \sim \prod_{j=1}^{k} TG\left(\frac{1}{2}, 0, \frac{8\lambda_2\sigma_0^2}{\lambda_1^2}, (1,\infty)\right)$, $\sigma_0^2 \sim 1/\sigma_0^2$, where TG(·) is a truncated Gamma distribution. Unlike the model of Li and Lin, where prior of β_j is conditioned on the noise variance σ_0^2, the prior of β_j in our BEN model is independent of σ_0^2, because only a point estimate of σ_0^2 is needed in our model inference. In the model of Li and Lin (2010), if we define $\tilde{\tau}_j = \tau_j - 1$, then $\sigma_0^2/\sigma_j^2 = \lambda_2 + \lambda_2/\tilde{\tau}_j$, this decomposition of σ_0^2/σ_j^2 is similar to that in our BEN model: $\alpha_j = \lambda_1 + \tilde{\alpha}_j$. Because τ_j follows a truncated Gamma distribution with a support of (1, ∞), $\tilde{\tau}_j$ obeys a shifted Gamma distribution similar to the prior for $\tilde{\sigma}_j^2$ in our BEN model. We assigned a uniform prior to σ_0^2, whereas Li and Lin (2010) used the Jeffrey's prior for σ_0^2. Li and Lin (2010) employed MCMC for model inference, which is computationally demanding, whereas our EBEN algorithm does not rely on MCMC and is more efficient.

Simulations demonstrated that our EBEN algorithm improves performance in terms of power of detection and FDR by taking into account the possible correlations among QTLs, which agrees with previous observations (Gianola *et al.*, 2003). Several methods for predicting genetic values incorporate the spatial correlations among markers. Yang and Tempelman (2012) included a first-order antedependence correlation structure for regression coefficients $\boldsymbol{\beta}$ into their Bayesian hierarchical mixed effects model so that β_j depends on β_{j-1}, $2 \leqslant j \leqslant k$, resulted in increased accuracy in predicting genetic values. Shen *et al.* (2011) incorporated a specific correlation structure in their smoothed double hierarchical generalized linear model, and a spatial correlation parameter was introduced to control correlation between two markers. Although our EBEN exploits the possible

correlations among QTLs, unlike those of Shen *et al.* (2011), Yang and Tempelman (2012), our EBEN does not specify a correlation structure for markers in the QTL model. Therefore, our EBEN is more robust, because a mis-specified correlation structure may significantly degrade performance. Our EBEN can shrink most variables in the QTL model to zero, yielding a sparse QTL model, which significantly decreases FDR without sacrificing the power of detection; whereas the method of Shen *et al.* (2011) does not employ the shrinkage technique, and it is not clear if the method of Yang and Tempelman (2012) can shrink variables in the QTL model to zero. Although performance of predicting genetic values may not degrade without using the shrinkage technique or an appropriate variable selection method to identify QTLs, shrinkage is very important to the performance of QTL mapping.

The EBEN algorithm inherits the efficiency of the EBlasso algorithm, because it is modified from the later, although the BEN model used by the EBEN algorithm is different from the Bayesian Lasso model used by the EBlasso algorithm. Our simulations (Cai *et al.*, 2011; Huang *et al.*, 2013) have shown that EBlasso outperformed a number of other competing algorithms in terms of detection power and FDR, and it offered a speed comparable to Lasso implemented with *glmnet* (Friedman *et al.*, 2010) but faster than other algorithms compared. EBEN improves the power of detection relative to EBlasso by detecting more correlated effects as shown in the simulation. However, in real data analysis, we observed that although EBEN detected more effects, it also missed several effects detected by EBlasso. One explanation is that EBEN outputs smaller estimates for the absolute amplitudes of correlated effects than EBlasso. A similar effect was observed for EN (Zou and Hastie, 2005) compared with Lasso. This may reduce the significance of the estimated effects. Therefore, when analyzing real data, we may apply both EBEN and EBlasso and take QTLs identified by either algorithm.

The EBEN algorithm was developed for quantitative traits, it can also be easily extended to QTL mapping with a logistic regression model for binary traits, following the derivations in (Huang *et al.*, 2013). Moreover, thanks to the regression model, it is straightforward to incorporate other covariates and maker interactions into the EBEN model. Recently, EBlasso has been applied to whole-genome QTL mapping (Huang *et al.*, 2014b) and pathway-based genome-wide association study (Huang *et al.*, 2014a), where linear regression models with millions of variables were inferred with EBLasso. Because EBEN inherits the computational efficiency of EBlasso, it can also be applied to both whole-genome QTL mapping and genome-wide association study. In conclusion, EBEN algorithm provides a useful tool for inference of high-dimensional sparse regression model in multiple QTL mapping and other applications.

DATA ARCHIVING

The genotype and phenotype data for simulation settings *SimI* and *SimII* are available at Dryad (DOI: 10.5061/dryad.jf142).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation [NSF CAREER Award no. 0746882 to XC] and by the Agriculture and Food Research Initiative (AFRI) of the USDA National Institute of Food and Agriculture under the Plant Genome, Genetics and Breeding Program [2007-35300-18285 to SX].

Andrews DF, Mallows CL (1974). Scale mixtures of normal distributions. *J R Stat Soc Series B Stat Methodol* **36**: 99–102.
Bogdan M, Frommlet F, Biecek P, Cheng R, Ghosh JK, Doerge RW (2008). Extending the modified Bayesian information criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics* **64**: 1162–1169.
Cai X, Huang A, Xu S (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC Bioinformatics* **12**: 211.
Cockerham CC (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**: 859–882.
Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**: 1–22.
Gianola D, Perez-Enciso M, Toro MA (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* **163**: 347–365.
Hayes P, Liu B, Knapp S, Chen F, Jones B, Blake T *et al.* (1993). Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theor Appl Genet* **87**: 392–401.
Hoti F, Sillanpää MJ (2006). Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity* **97**: 4–18.
Huang A, Martin E, Vance J, Cai X (2014a). Detecting genetic interactions in pathway-based genome-wide association studies. *Genet Epidemiol* **38**: 300–309.
Huang A, Xu S, Cai X (2013). Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. *BMC Genet* **14**: 5.
Huang A, Xu S, Cai X (2014b). Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice. *PLoS ONE* **9**: e87330.
Huang H, Eversley CD, Threadgill DW, Zou F (2007). Bayesian multiple quantitative trait loci mapping for complex traits using markers of the entire genome. *Genetics* **176**: 2529–2540.
Li Q, Lin N (2010). The Bayesian elastic net. *Bayesian Anal* **5**: 151–170.
Li S, Lu Q, Fu W, Romero R, Cui Y (2009). A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy. *Stat Appl Genet Mol Biol* **8**: 1–28.
Li Z, Sillanpää M (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor Appl Genet* **125**: 419–435.
Meinshausen N, Bühlmann P (2010). Stability selection. *J R Stat Soc Series B Stat Methodol* **72**: 417–473.
Meinshausen N, Meier L, Bühlmann P (2009). P-values for high-dimensional regression. *J Am Stat Assoc* **104**: 713–725.
O'Hara RB, Sillanpaa MJ (2009). A review of Bayesain variable selection methods: what, how and which. *Bayesian Anal* **4**: 8–118.
Pal N, Jin C, Lim WK (2005). *Handbook of Exponential and Related Distributions for Engineers and Scientists*. CRC Press: New York.
Park T, Casella G (2008). The Bayesian lasso. *J Am Stat Assoc* **103**: 681–686.
Robert CR, Casella G (2004). *Monte Carlo statistical methods*, 2 edn. Springer: New York.
Schwarz G (1978). Estimating the dimension of a model. *Ann Stat* **6**: 461–464.
Shen X, Ronnegard L, Carlborg O (2011). Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps. *BMC Proceedings* **5**:Suppl 3 S14.
Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* **58**: 267–288.
Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ *et al.* (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**: 465–480.
Wu R, Ma CX, Casella G (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*. Springer: New York.
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.
Yang W, Tempelman RJ (2012). A Bayesian antedependence model for whole genome prediction. *Genetics* **190**: 1491–1501.
Yi N, Xu S (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045–1055.
Yu J, Zhang Z, Zhu C, Tabanao DA, Pressoir G, Tuinstra MR *et al.* (2009). Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. *Plant Gen* **2**: 63–77.
Zou H, Hastie T (2005). Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* **67**: 301–320.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)

APPENDIX A
Proof of the EN prior distribution

The joint prior distribution for β_j and $\tilde{\sigma}_j^2$ can be written as:

$$p(\beta_j, \tilde{\sigma}_j^2|\lambda_1, \lambda_2) = p(\beta_j|\tilde{\sigma}_j^2)p(\tilde{\sigma}_j^2|\lambda_1, \lambda_2),\ j = 1, 2, \cdots, k,$$

where $p(\beta_j|\tilde{\sigma}_j^2)$ is a normal distribution: $\beta_j \sim N(0, \frac{\tilde{\sigma}_j^2}{\lambda_1\tilde{\sigma}_j^2+1})$ and $p(\tilde{\sigma}_j^2|\lambda_1, \lambda_2)$ is a generalized Gamma distribution: $p(\tilde{\sigma}_j^2|\lambda_1, \lambda_2) = C\left(\lambda_1\tilde{\sigma}_j^2+1\right)^{-1/2}\exp\left(-\lambda_2\tilde{\sigma}_j^2\right)$, with C being a normalization constant. The marginal prior distribution of β_j can be found as

$$\begin{aligned} p(\beta_j|\lambda_1, \lambda_2) &= C\int\sqrt{\frac{\lambda_1\tilde{\sigma}_j^2+1}{2\pi\tilde{\sigma}_j^2}}\exp\left(-\frac{\beta_j^2}{2\tilde{\sigma}_j^2/(\lambda_1\tilde{\sigma}_j^2+1)}\right) \\ &\quad \frac{1}{\sqrt{\lambda_1\tilde{\sigma}_j^2+1}}\exp\left(-\lambda_2\tilde{\sigma}_j^2\right)d\tilde{\sigma}_j^2 \\ &= C\exp\left(-\frac{\lambda_1\beta_j^2}{2}\right)\int\frac{1}{\sqrt{2\pi\tilde{\sigma}_j^2}}\exp\left(-\frac{\beta_j^2}{2\tilde{\sigma}_j^2}-\lambda_2\tilde{\sigma}_j^2\right)d\tilde{\sigma}_j^2. \end{aligned}$$

Using the result in Andrews and Mallows (1974), the integral can be found in a closed-form for $\lambda_2>0$, and $p(\beta_j|\lambda_1, \lambda_2)$ is simplified as $p(\beta_j|\lambda_1, \lambda_2) = C\exp(-\frac{\lambda_1\beta_j^2}{2}-\sqrt{2\lambda_2}\mid\beta_j\mid)$, which is the EN prior distribution.

APPENDIX B
Derivation of equation (8)

Note that $\lim_{\tilde{\alpha}_j\to 0}L(\tilde{\alpha}_j) = -\infty$, $\lim_{\tilde{\alpha}_j\to\infty}L(\tilde{\alpha}_j) = 0$, and the derivative of $L(\tilde{\alpha}_j)$is given by:

$$\frac{\partial L}{\partial\tilde{\alpha}_j} = \frac{(s_j+\lambda_1-q_j^2+2\lambda_2)\tilde{\alpha}_j^2+(s_j+\lambda_1+4\lambda_2)(s_j+\lambda_1)\tilde{\alpha}_j+2\lambda_2(s_j+\lambda_1)^2}{2\tilde{\alpha}_j^2(\tilde{\alpha}_j+s_j+\lambda_1)^2}.$$

Let us write the numerator of the derivative as $J(\tilde{\alpha}_j) = (s_j+\lambda_1-q_j^2+2\lambda_2)\tilde{\alpha}_j^2+(s_j+\lambda_1+4\lambda_2)(s_j+\lambda_1)\tilde{\alpha}_j+2\lambda_2(s_j+\lambda_1)^2$, and define $\Delta = (s_j+\lambda_1)^2+8\lambda_2q_j^2$. Because $\lambda_2>0$, we have $\Delta\geqslant 0$, which implies that $J(\tilde{\alpha}_j) = 0$ have two roots: $r_1 = \frac{-(s_j+\lambda_1+4\lambda_2)-\sqrt{\Delta}}{2(s_j+\lambda_1-q_j^2+2\lambda_2)}\cdot(s_j+\lambda_1)$ and $r_2 = \frac{-(s_j+\lambda_1+4\lambda_2)+\sqrt{\Delta}}{2(s_j+\lambda_1-q_j^2+2\lambda_2)}\cdot(s_j+\lambda_1)$. Next let us consider the following three cases:

Case 1: $s_j+\lambda_1-q_j^2+2\lambda_2>0$

We have $r_1<0$ and $r_2<0$ because $s_j>0$ and $\lambda_2>0$. Therefore, $\partial L/\partial\tilde{\alpha}_j>0$ for $\tilde{\alpha}_j>0$ and $L(\tilde{\alpha}_j)$ is an increasing function of $\tilde{\alpha}_j$. This implies that $L(\tilde{\alpha}_j)$ is maximized at $\tilde{\alpha}_j^* = \infty$.

Case 2: $s_j+\lambda_1-q_j^2+2\lambda_2 = 0$

In this case, we have $J(\tilde{\alpha}_j) = (s_j+\lambda_1+4\lambda_2)(s_j+\lambda_1)\tilde{\alpha}_j+2\lambda_2(s_j+\lambda_1)^2$. It is clear that $J(\tilde{\alpha}_j)>0$ for $\tilde{\alpha}_j>0$. Hence $\partial L/\partial\tilde{\alpha}_j>0$ and $L(\tilde{\alpha}_j)$ is an increasing function of $\tilde{\alpha}_j$. Then $L(\tilde{\alpha}_j)$ is maximized at $\tilde{\alpha}_j^* = \infty$.

Case 3: $s_j+\lambda_1-q_j^2+2\lambda_2<0$

We have $r_1>0$ and $r_2<0$. Therefore, $\partial L/\partial\tilde{\alpha}_j>0$ for $0<\tilde{\alpha}_j<r_1$, $\partial L/\partial\tilde{\alpha}_j = 0$ for $\tilde{\alpha}_j = r_1$, and $\partial L/\partial\tilde{\alpha}_j<0$ for $\tilde{\alpha}_j>r_1$. This implies that $L(\tilde{\alpha}_j)$ is maximized at $\tilde{\alpha}_j^* = r_1$ and $L(\tilde{\alpha}_j^*)>0$ because $\lim_{\tilde{\alpha}_j\to\infty}L(\tilde{\alpha}_j) = 0$.

Summarizing the results in three cases, we obtain $\tilde{\alpha}_j^*$ given in (8).

APPENDIX C
EBEN algorithm

1. Initialize parameters: choose $\upsilon\in[0,1]$ and $\lambda>0$, calculate $\mu = \mathbf{1}^T\mathbf{y}/n$, $\tilde{y} = \boldsymbol{y}-\mu$ and set σ_0^2 to be a small number, e.g., $0.1\times\tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{y}}/n$.
2. Initialize the model: Find $j = \arg_i\max\{|\boldsymbol{x}_i^T\tilde{\boldsymbol{y}}|, \forall i\}$, and calculate α_j from (8); set all other $\alpha_{j'}$, $j'\neq j$ to be ∞ and $\tilde{\mathbf{X}} = \boldsymbol{x}_j$.
3. Calculate $\boldsymbol{\Sigma}$, s_j and q_j, $\forall j$.
4. Update the model.
   Apply the EBlasso algorithm (Cai *et al*., 2011) to update $\mathbf{A}$ with $\hat{\boldsymbol{\alpha}}$ obtained from (8).
   If the global convergence criterion is not satisfied, go to step 4.
5. Output $\hat{\boldsymbol{\beta}}'$ and covariance $\hat{\boldsymbol{\Sigma}}$.