OPEN

Heredity (2017) 118, 517–524
Official journal of the Genetics Society
www.nature.com/hdy

ORIGINAL ARTICLE

pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies

J Zhang[1,4], J-Y Feng[1,4], Y-L Ni[1], Y-J Wen[1], Y Niu[1], CL Tamba[1], C Yue[1], Q Song[2] and Y-M Zhang[1,3]

Multilocus genome-wide association studies (GWAS) have become the state-of-the-art procedure to identify quantitative trait nucleotides (QTNs) associated with complex traits. However, implementation of multilocus model in GWAS is still difficult. In this study, we integrated least angle regression with empirical Bayes to perform multilocus GWAS under polygenic background control. We used an algorithm of model transformation that whitened the covariance matrix of the polygenic matrix K and environmental noise. Markers on one chromosome were included simultaneously in a multilocus model and least angle regression was used to select the most potentially associated single-nucleotide polymorphisms (SNPs), whereas the markers on the other chromosomes were used to calculate kinship matrix as polygenic background control. The selected SNPs in multilocus model were further detected for their association with the trait by empirical Bayes and likelihood ratio test. We herein refer to this method as the pLARmEB (polygenic-background-control-based least angle regression plus empirical Bayes). Results from simulation studies showed that pLARmEB was more powerful in QTN detection and more accurate in QTN effect estimation, had less false positive rate and required less computing time than Bayesian hierarchical generalized linear model, efficient mixed model association (EMMA) and least angle regression plus empirical Bayes. pLARmEB, multilocus random-SNP-effect mixed linear model and fast multilocus random-SNP-effect EMMA methods had almost equal power of QTN detection in simulation experiments. However, only pLARmEB identified 48 previously reported genes for 7 flowering time-related traits in *Arabidopsis thaliana*.
Heredity (2017) **118,** 517–524; doi:10.1038/hdy.2017.8; published online 15 March 2017

INTRODUCTION

Most complex traits in human, plant and animal genetics are quantitative traits and these traits are controlled by multiple quantitative trait loci (QTLs). The identification of these loci is usually performed by QTL mapping or genome-wide association study (GWAS). A large number of single-nucleotide polymorphisms (SNPs) can be easily obtained for the genotypes by the rapid development of sequencing and genotyping technologies. If all the SNPs are included in a genetic model, the number of SNPs will be much larger than the sample size. The commonly used methods are infeasible for such an oversaturated model.

Many approaches have been proposed to estimate the parameters in the oversaturated model and these approaches include ridge regression (Hoerl and Kennard, 1970), stochastic search variable selection (George and McCulloch, 1993; Yi *et al.*, 2003), Bayesian shrinkage estimation (Meuwissen *et al.*, 2001; Wang *et al.*, 2005), penalized maximum likelihood (Zhang and Xu, 2005; Hoggart *et al.*, 2008; Zhang *et al.*, 2012), empirical Bayes (Xu, 2010) and Bayesian-LASSO (Bayesian-least absolute shrinkage and selection operator; Park and Casella, 2008; Yi and Xu, 2008). However, these methods are mainly proposed for linkage analysis in biparental segregation populations, rather than for GWAS in natural population.

GWAS has been used to dissect the genetic foundation of quantitative traits (Zhang *et al.*, 2005, 2010; Yu *et al.*, 2006; Kang *et al.*, 2008; Zhou and Stephens, 2012; Wang *et al.*, 2016). The widely used approach, such as efficient mixed model association (EMMA; Kang *et al.*, 2008; Zhou and Stephens, 2012), was proposed for single-marker analysis under the population structure and polygenic background controls. However, this method has relatively low power in detecting small-effect QTLs. To overcome these problems, therefore, multilocus model methods have been suggested (Fridley *et al.*, 2010; Lü *et al.*, 2011), for example, a Bayesian-inspired penalized maximum likelihood approach (Zhang and Xu, 2005; Hoggart *et al.*, 2008) and PUMA (Penalized Unified Multiple-locus Association; Hoffman *et al.*, 2013). These methods can be used if the number of variables in the multilocus model is not too large. Recent strategies for high-dimensional modeling have focused on reducing the dimension of a large matrix and then selecting the most potentially associated SNPs by using shrinkage methods such as the LASSO and SCAD (smoothly clipped absolute deviation) penalty (Fan and Lv, 2008; Wu *et al.*,

[1]State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China; [2]Soybean Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD, USA and [3]Statistical Genomics Lab, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China
Correspondence: Dr Y-M Zhang, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China or College of Agriculture, Nanjing Agricultural University, Nanjing 210095, China.
E-mail: soyzhang@mail.hzau.edu.cn or soyzhang@njau.edu.cn
[4]These authors contributed equally to this work.
Received 5 October 2016; revised 14 January 2017; accepted 20 January 2017; published online 15 March 2017

2009). Although other multilocus approaches have also been proposed by Segura *et al.* (2012), Moser *et al.* (2015), Liu *et al.* (2016), Wang *et al.* (2016) and Wen *et al.* (2017), now further refinement and studies are still needed.

In this study, we integrated least angle regression (LARS) algorithm with empirical Bayes to perform multilocus GWAS for quantitative traits, as the LARS algorithm makes LASSO (Tibshirani, 1996) efficient and acceptable (Efron *et al.*, 2004). To control polygenic background, we adopted the model transformation of Wen *et al.* (2017) that whitens the covariance matrix of the polygenic matrix K and residual noise. The LARS algorithm was implemented on the transformed model to select SNPs that are most potentially associated with the trait, empirical Bayes was used to estimate the effects of all the selected SNPs and all the nonzero effects were further examined by likelihood ratio test so as to confirm true quantitative trait nucleotides (QTNs). We refer to this method as the pLARmEB (polygene-background-control-based least angle regression plus empirical Bayes). pLARmEB was validated by analysis of the data sets from a series of Monte Carlo simulation experiments and seven *Arabidopsis* flowering time traits. We also discussed the possibility of applying pLARmEB for linkage analysis.

MATERIALS AND METHODS

Genetic model

Let y_i ($i=1,\cdots,n$) be the phenotypic value of the ith individual in a sample of size n from a natural population. The genetic model is expressed by

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{W}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}=(y_1, \cdots,y_n)^T$; $\mathbf{1}$ is a $n\times 1$ vector of 1 and μ is total average; $\boldsymbol{\alpha}$ is population structure effect as fixed; $\boldsymbol{\gamma}\sim \mathrm{MVN}_m(\mathbf{0},\Sigma_\gamma)$ are QTN effects as random, $\Sigma_\gamma = diag\{\sigma_1^2, \cdots, \sigma_m^2\}$ and m is the number of putative QTNs; $\mathbf{W}$ and $\mathbf{Z}$ are the corresponding designed matrices for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$; polygenic effects $\mathbf{u}\sim \mathrm{MVN}_n(\mathbf{0}, \sigma_g^2\mathbf{K})$ is a $n\times 1$ random vector and $\mathbf{K}$ is a known $n\times n$ relatedness matrix; and $\boldsymbol{\epsilon}$ is residual error with an assumed $\mathrm{MVN}_n(\mathbf{0},\sigma^2\mathbf{I}_n)$ distribution, σ^2 is residual error variance and $\mathbf{I}_n$ is an $n\times n$ identity matrix.

As $\boldsymbol{\gamma}$ is treated as being random, the variance of $\mathbf{y}$ in the model (1) is

$$\begin{aligned}\mathrm{var}(\mathbf{y}) &= \mathbf{Z}\Sigma_\gamma\mathbf{Z}^T + \sigma_g^2\mathbf{K} + \sigma^2\mathbf{I}_n = \sum_{k=1}^{m}\sigma_k^2\mathbf{Z}_k\mathbf{Z}_k^T + \sigma_g^2\mathbf{K} + \sigma^2\mathbf{I}_n \\ &= \sigma^2\left(\sum_{k=1}^{m}\lambda_k\mathbf{Z}_k\mathbf{Z}_k^T + \lambda_g\mathbf{K} + \mathbf{I}_n\right) = \sigma^2\mathbf{H}\end{aligned} \tag{2}$$

where $\lambda_k = \sigma_k^2/\sigma^2$ ($k=1,\cdots,m$), $\lambda_g = \sigma_g^2/\sigma^2$ and $\mathbf{H}=\mathbf{Z}diag\{\lambda_1,\cdots,\lambda_m\}\mathbf{Z}^T+\lambda_g\mathbf{K}+\mathbf{I}_n$

Using EMMA, we can obtain the estimate of λ_g, denoted by $\hat{\lambda}_g$. Let $\mathbf{B} = \hat{\lambda}_g\mathbf{K} + \mathbf{I}_n$, an eigen (or spectral) decomposition of the positive semidefinite matrix $\mathbf{B}$ was

$$\begin{aligned}\mathbf{B} = \mathbf{Q}_B\boldsymbol{\Lambda}_B\mathbf{Q}_B^T &= \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{\Lambda}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_1^T\\ \mathbf{Q}_2^T\end{pmatrix} \\ &= \left(\mathbf{Q}_1\boldsymbol{\Lambda}_r^{\frac{1}{2}}\mathbf{Q}_1^T\right)\left(\mathbf{Q}_1\boldsymbol{\Lambda}_r^{\frac{1}{2}}\mathbf{Q}_1^T\right)\end{aligned} \tag{3}$$

where $\mathbf{Q}_B$ is orthogonal, $\boldsymbol{\Lambda}_r$ is a diagonal matrix with positive eigenvalues, $r=Rank(\mathbf{B})$, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the $n\times r$ and $n\times(n-r)$ block matrices of $\mathbf{Q}_B$, respectively, and $\mathbf{0}$ is the corresponding block zero matrix (Wen *et al.*, 2017).

Let $\mathbf{C} = \mathbf{Q}_1\boldsymbol{\Lambda}_r^{-\frac{1}{2}}\mathbf{Q}_1^T$, the model (1) is changed to

$$\mathbf{y}_c = \mathbf{1}_c\mu + \mathbf{W}_c\boldsymbol{\alpha} + \mathbf{Z}_c\boldsymbol{\gamma} + \boldsymbol{\epsilon}_c \tag{4}$$

where $\mathbf{y}_c=\mathbf{C}\mathbf{y}$, $\mathbf{1}_c=\mathbf{C}\mathbf{1}$, $\mathbf{W}_c=\mathbf{C}\mathbf{W}$, $\mathbf{Z}_c=\mathbf{C}\mathbf{Z}$ and $\boldsymbol{\epsilon}_c=\mathbf{C}\mathbf{u}+\mathbf{C}\boldsymbol{\epsilon}\sim \mathrm{MVN}_n(\mathbf{0},\sigma^2\mathbf{I}_n)$ (Wen *et al.*, 2017).

In the above model (4), let $\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\gamma}\end{pmatrix}$, $\mathbf{Y}=\mathbf{y}_c-\mathbf{1}_c\,\mu$ with a zero mean, and standardizing each column in matrix $(\mathbf{W}_c\ \mathbf{Z}_c)$ produces a new matrix $\mathbf{X}$ with $\sum_{i=1}^n x_{ij}=0$ and $\sum_{i=1}^n x_{ij}^2=1$ ($j=1,\cdots,m$). Therefore, the model (4) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5}$$

Parameter estimation

LARS for the full model. LARS is a flexible method for variable selection that has been described previously (Efron *et al.*, 2004). We used the LARS algorithm to select the $n-1$ variables that are most likely associated with quantitative trait of interest.

First, let $\hat{\boldsymbol{\beta}}_0 = 0$, so $\hat{\boldsymbol{\mu}}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}_0 = 0$

Then, suppose that $\hat{\boldsymbol{\mu}}_\mathrm{F}$ is the current LARS estimate and that

$$\hat{\mathbf{c}} = \mathbf{X}^T(\mathbf{Y} - \hat{\boldsymbol{\mu}}_\mathrm{F}) \tag{6}$$

is the vector of current correlations. The active set $\in$ is the set of indices corresponding to covariates with the greatest absolute current correlations,

$$\hat{C} = \max_j\left\{\left|\hat{c}_j\right|\right\} \text{ and } \mathrm{F} = \left\{j : \left|\hat{c}_j\right| = \hat{C}\right\}$$

Let $s_j = sign\left\{\hat{c}_j\right\}$ for $j\in \mathrm{F}$. We can calculate $\mathbf{X}_\mathrm{F}=(\cdots s_j\mathbf{x}_j\cdots)_{j\in \mathrm{F}}$, $\mathbf{u}_\mathrm{F}=\mathbf{X}_\mathrm{F}\omega_\mathrm{F}$, $\mathbf{G}_\mathrm{F} = \mathbf{X}_\mathrm{F}^T\mathbf{X}_\mathrm{F}$, $\mathbf{F}_\mathrm{F} = \left(\mathbf{1}_\mathrm{F}^T\mathbf{G}_\mathrm{F}^{-1}\mathbf{1}_\mathrm{F}\right)^{-1/2}$, where $\omega_\mathrm{F} = \mathbf{F}_\mathrm{F}\mathbf{G}_\mathrm{F}^{-1}\mathbf{1}_\mathrm{F}$, and $\mathbf{1}_\mathrm{F}$ being a vector of 1 with the length of equaling |F|.

Third, update $\hat{\boldsymbol{\mu}}_\mathrm{F}$ in the LARS algorithm:

$$\hat{\boldsymbol{\mu}}_{F+} = \hat{\boldsymbol{\mu}}_\mathrm{F} + \hat{\gamma}\mathbf{u}_\mathrm{F}$$

where $\hat{\gamma} = \min_{j\in \mathrm{F}^c}{}^{+}\left\{\frac{\hat{C}-\hat{c}_j}{\mathbf{F}_\mathrm{F}-a_j}, \frac{\hat{C}+\hat{c}_j}{\mathbf{F}_\mathrm{F}+a_j}\right\}$, $\min^+$ indicates that the minimum is taken over only positive components within each choice of j in the formula of $\hat{\gamma}$, and $\mathbf{a}\equiv\mathbf{X}^T\mathbf{u}_\mathrm{F}$.

Repeat step 2 to step 3 until a criterion of convergence is satisfied. The above algorithm was conducted by lars package (http://cran.r-project.org/web/packages/lars/) in R language.

Usually, if all the marker effects are included in one genetic model, the parameters cannot be estimated under the situation of $m\gg n$, where n is sample size and m is the number of variables. As most markers are not likely associated with the trait of interest, once the markers with zero effects are deleted from the full model, marker effects of the reduced model is estimable. In each LARS variable selection, the $n-1$ SNPs that are most potentially associated with the trait are selected to construct the reduced model.

Empirical Bayes estimation in the reduced model. In the reduced model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{7}$$

where $\mathbf{y}$ is the same as that in the model (1); $\boldsymbol{\beta}$ is a vector of fixed effect, $\boldsymbol{\gamma}$ is a vector of random effect of the selected markers and $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. All the parameters in the model (7) were estimated by empirical Bayes proposed by Xu (2010).

The fixed effect $\boldsymbol{\beta}$ and residual variance σ^2 were estimated by

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}) \tag{8}$$

$$\sigma^2 = \frac{1}{n}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\left[\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{k=1}^{p}\mathbf{Z}_k\mathrm{E}(\gamma_k)\right] \tag{9}$$

where $\mathbf{V} = \sigma^2\mathbf{I} + \sum_{k=1}^{p}\mathbf{Z}_k\mathbf{Z}_k^T\sigma_k^2 = \mathbf{I}\sigma^2 + \mathbf{Z}diag\left\{\sigma_1^2, \cdots, \sigma_p^2\right\}\mathbf{Z}^T$. The random effect γ_k of each marker and its prediction error $\mathrm{var}(\gamma_k)$ were predicted by best linear unbiased prediction:

$$E(\gamma_k) = \sigma_k^2\mathbf{Z}_k^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{10}$$

$$\mathrm{var}(\gamma_k) = \sigma_k^2\mathbf{I} - \sigma_k^2\mathbf{Z}_k^T\mathbf{V}^{-1}\mathbf{Z}_k\sigma_k^2 \tag{11}$$

where $\sigma_k^2 = \frac{E(\gamma_k^T\gamma_k)+\omega}{\tau+2+m_k}$, $\omega=\tau=0$, and m_k is the number of genotypes at locus k. The method requires inverse of matrix $\mathbf{V}$. If the sample size is large, that is,

$n>p$, binomial inverse theorem (Henderson and Searle, 1980) can be used:

$$\mathbf{V}^{-1} = (\mathbf{I}\sigma^2)^{-1} - (\mathbf{I}\sigma^2)^{-1}\mathbf{Z}\boldsymbol{\Sigma}\left[\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{Z}^T(\mathbf{I}\sigma^2)^{-1}\mathbf{Z}\boldsymbol{\Sigma}\right]^{-1}\boldsymbol{\Sigma}\mathbf{Z}^T(\mathbf{I}\sigma^2)^{-1} \quad (12)$$

where $\boldsymbol{\Sigma} = diag\left\{\sigma_1^2, \cdots, \sigma_p^2\right\}$

Based on our experiences, empirical Bayes is feasible when the number of variables is less than 40 times of the sample size. However, this condition is not frequently met in GWAS. If the LARS algorithm is used to select the variables that are most potentially associated with the trait under polygenic background control, the effects of the selected markers can be estimated by empirical Bayes.

Likelihood ratio (*LR*) test

Based on the estimate of marker effect γ_k in the reduced model, markers with $|\hat{\gamma}_k| < 10^{-4}$ are considered not to be associated with the trait; however, the association of the chosen markers with the trait and the effects $\theta = \{\gamma_{(1)}, \cdots, \gamma_{(q)}\}$ needs to be tested, where q is the number of SNPs in the reduced model. To test the null hypothesis H_0:$\gamma_{(i)} = 0$, that is, no QTL linked to the marker, we conducted an *LR* test by

$$LR_i = -2[L(\theta_{-i}) - L(\theta)] \quad (13)$$

where $\theta_{-i} = \left\{\gamma_{(1)}, \cdots, \gamma_{(i-1)}, \gamma_{(i+1)}, \cdots, \gamma_{(q)}\right\}^T$, $L(\theta) = \sum_{i=1}^{n} \ln\phi\left(y_i; \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\right)$ is a log-likelihood function, $\phi(y_i;\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma},\sigma^2)$ is a normal density function with mean $\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}$ and variance σ^2 and $LOD = LR/4.605$. The critical value for significance was set at $LOD = 2.0$ (Bu *et al.*, 2015).

AIC and *BIC* for testing goodness of fit of models

The goodness of fit for a statistical model can be measured by

$$AIC = -2\ln(L) + 2k \quad (14)$$

$$BIC = -2\ln(L) + k\ln(n) \quad (15)$$

where L is the likelihood function value and k is the number of independent variables, and n is sample size. Smaller *Akaike information criterion* (*AIC*) or *Bayesian information criterion* (*BIC*) value indicates a good fit.

pLARmEB has been implemented in R and its software can be downloaded from https://cran.r-project.org/web/packages/mrMLM/index.html.

Data sets for analyses

One *Arabidopsis* data set and four Monte Carlo simulated data sets were used to validate pLARmEB. Each data set contained phenotypic observations for quantitative traits and genotypic values for molecular markers.

The Arabidopsis data set. The data set downloaded from http://www.arabidopsis.org/ includes 199 diverse inbred lines each with 216 130 SNPs and 107 traits (Atwell *et al.*, 2010). Among these traits, seven are related to flowering time, including days to flowering under long days, days to flowering under long days with vernalization, days to flowering under short days, days to flowering under short days with vernalization, days to flowering at 10 °C, days to flowering at 16 °C and days to flowering at 22 °C. We analyzed these traits using pLARmEB, EMMA, multilocus random-SNP-effect mixed linear model (mrMLM) and fast multilocus random-SNP-effect EMMA (FASTmrEMMA) methods. The population structure Q matrix and kinship coefficient matrix K between all the pairs of lines were used to control population structure and polygenic background. We also deleted the SNPs with minor allele frequency <10%. When all the markers on one chromosome were in one genetic model, the markers on other chromosomes were used to calculate K matrix as polygenic background control (Rincent *et al.*, 2014; Yang *et al.*, 2014; Wei and Xu, 2016). Here 50 SNPs most potentially associated with the trait are selected to construct the reduced model. This number may vary across different data sets.

Data sets from Monte Carlo simulation in natural population. Three Monte Carlo simulation experiments were conducted to validate pLARmEB. The three data sets are the same as those in Wang *et al.* (2016). In the first experiment, all the SNP genotypes were derived from 216 130 SNPs reported by Atwell *et al.* (2010) and 2000 SNPs were randomly sampled from each chromosome (Chr.). The positions of these SNPs in the genome were between 11 226 256 and 12 038 776 bp on Chr. 1, between 5 045 828 and 6 412 875 bp on Chr. 2, between 1 916 588 and 3 196 442 bp on Chr. 3, between 2 232 796 and 3 143 893 bp on Chr. 4 and between 19 999 868 and 21 039 406 bp on Chr. 5 (Wang *et al.*, 2016). The sample size was 199, and this was the number of lines in Atwell *et al.* (2010). Six QTNs were simulated and placed on the SNPs with rare allelic frequency of 0.30. The heritabilities of the QTNs were set as 0.10, 0.05, 0.05, 0.15, 0.05 and 0.05, respectively; their positions and effects are listed in Supplementary Table S1. The total average was set at 10.0 and residual variance was set at 10.0. For each simulated QTN, we counted the number of samples in which the LOD (logarithm (base 10) of odds) exceeded 2.0 (Bu *et al.*, 2015). A detected QTN within 2 kb of the simulated QTN was considered a true QTN. The ratio of the number of such samples to the total number of replicates (1000) represented the empirical power of this QTN. False positive rate (FPR) was calculated as the ratio of the number of false positive effects to the total number of zero effects considered in the full model. To measure the bias of gene effect estimate, mean squared error (MSE)

$$\mathrm{MSE}_k = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\gamma}_{k(i)} - \gamma_k)^2 \quad (16)$$

was calculated, where $\hat{\gamma}_{k(i)}$ is the estimate of effect γ_k in the ith sample.

We investigated the effect of polygenic background on pLARmEB in the second experiment by adding polygenic effects from a multivariate normal distribution $\mathrm{MVN}_n(\mathbf{0}, \sigma_{pg}^2\mathbf{K})$, where σ_{pg}^2 is polygenic variance and $\mathbf{K}$ is a pairwise kinship coefficient matrix among individuals. Here $\sigma_{pg}^2 = 2$, so $h_{pg}^2 = 0.092$. The QTN size (h^2), total average, residual variance and other parameter values were the same as those in the first experiment, and all the parameters are listed in Supplementary Table S2.

In the third experiment, we investigated the effect of epistatic background on pLARmEB. Three epistatic QTNs were added. The related parameters for the simulated three epistatic QTNs have been described in Wang *et al.* (2016). The QTN sizes (h^2), total average, residual variance and other parameter values were also the same as those in the first experiment (Supplementary Table S3).

Monte Carlo simulation experiments in backcross. To test whether pLARmEB can be used in biparental population, we conducted another simulation experiment. In this experiment, 200 individuals each with 10 001 evenly spaced markers on the entire genome of 100 000 cM length were simulated in backcross population. Eight main-effect QTLs were simulated and placed at marker positions. The sizes and locations of these QTLs are listed in Supplementary Table S4. The population mean (b_0) and residual error variance (σ^2) were set at 10 and 10, respectively. The number of replicates was set at 200.

RESULTS

Monte Carlo simulation studies

Statistical power for QTN detection. To validate pLARmEB, three simulation experiments were conducted. In the first experiment, each simulated sample was analyzed by pLARmEB, least angle regression plus empirical Bayes (LARmEB), EMMA, FASTmrEMMA, mrMLM and Bayesian hierarchical generalized linear model (BhGLM). Among the 1000 samples, the first 100 were further analyzed using the BhGLM method. As shown in Supplementary Table S1 and Figure 1a, the average power for the above 6 methods was 77.1, 68.9, 46.0, 70.7, 68.6 and 54.5%, respectively. The method in which polygenic background was controlled had the highest average power among the six methods (Figure 1a). To further confirm the effectiveness of pLARmEB, polygenic effect simulated from multivariate normal distribution ($r^2 = 9.2\%$) was added to each phenotype in the second experiment and three epistatic QTNs ($r^2 = 15\%$) were added in the third simulation experiment. The average powers based on pLARmEB, LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM were 78.3, 69.6, 42.5, 75.0, 67.6 and 60.7%, respectively, in the second experiment (Supplementary Table S2); and 74.4, 57.5, 39.1, 59.2, 58.9 and

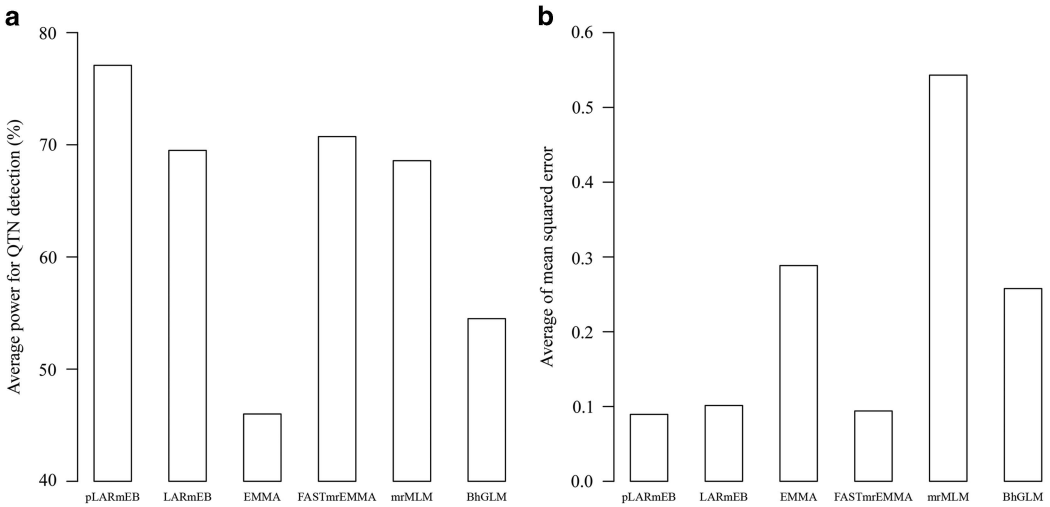

Figure 1 Average powers in the detection of QTNs (**a**) and average of mean squared errors in the estimation of QTN effects (**b**) across six simulated QTNs using pLARmEB, LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM.

56.3%, respectively, in the third experiment (Supplementary Table S3). The highest average power was observed when pLARmEB included polygenic background control.

Accuracies of estimated QTN effects. MSE measured accuracies of estimated QTN effects, and low MSE indicates high accuracy for parameter estimation. As shown in Figure 1b and Supplementary Tables S1–S3, the average MSEs based on pLARmEB, LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM were 0.0895, 0.1005, 0.5432, 0.2885, 0.0940 and 0.2577, respectively, in the first experiment (Figure 1b and Supplementary Table S1); 0.0917, 0.0997, 0.5680, 0.3227, 0.0852 and 1.3139, respectively, in the second experiment (Supplementary Table S2); and 0.0973, 0.1240, 0.5973, 0.3450, 0.1024 and 0.3934, respectively, in the third experiment (Supplementary Table S3). pLARmEB had the highest accuracy for estimating QTN effect among the six methods.

FPR and ROC curve. High FPR is a major concern in GWAS. To overcome this issue, a very high significance level was frequently adopted in genome-wide single marker scan. In our multilocus method, a less stringent significance level (LOD = 2.0) was recommended. We wanted to know whether this criterion produces high FPR. All the FPR results in the three simulation experiments are listed in Supplementary Tables S1–S3. Clearly, the FPRs based on pLARmEB, LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM were 0.0009, 0.0127, 0.0325, 0.0084, 0.0168 and 0.0115 (%), respectively, in the first experiment (Supplementary Table S1); 0.0025, 0.0010, 0.0166, 0.0081, 0.0210 and 0.0093%, respectively, in the second experiment (Supplementary Table S2); and 0.0089, 0.0031, 0.0253, 0.0148, 0.0265 and 0.0120%, respectively, in the third experiment (Supplementary Table S3). These results indicate that pLARmEB had a low FPR.

To compare various approaches for their efficiencies in the detection of significant QTNs, receiver operating characteristic (ROC) curve was plotted. ROC is a plot of average power against FPR. We calculated the corresponding average powers for the 41 thresholds between 10^{-6} and 10^{-2} in the first simulation experiment, and compared the ROC curves among the above 6 methods. Under

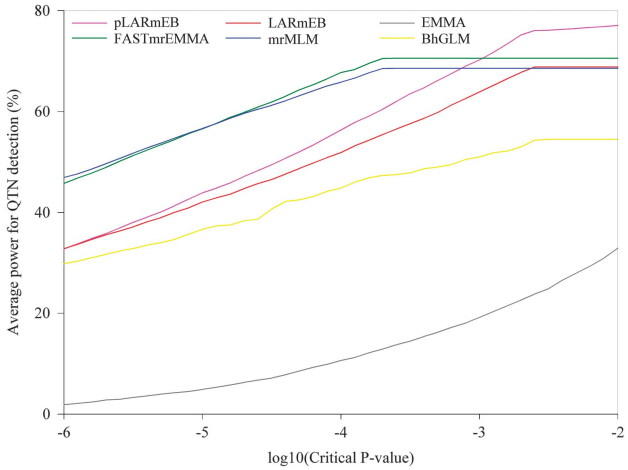

Figure 2 Statistical powers of six simulated QTNs in the first simulation experiment plotted against false positive rate (in a $\log_{10}$ scale) for pLARmEB, LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM.

the 0.01 to 0.001 significant levels, pLARmEB has the highest power to detect QTN among the six methods (Figure 2).

Computational efficiency. We scanned and identified SNPs that were associated with the trait on each chromosome using LARS. We then included all the potentially associated SNPs across the genome into one genetic model and estimated their effects by empirical Bayes (Xu, 2010). For the first simulation experiment, the above procedures took 4.20, 6.82, 68.77, 8.32, 13.29 and >100 h (Intel Core i5-4570 CPU 3.20 GHz, Memory 7.88G, Nanjing, China) for pLARmEB, LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM, respectively. pLARmEB took the least computing time among the six approaches. A similar trend was found in real data analyses (Supplementary Table S5).

Analysis of the *Arabidopsis* data set

To test the performance of pLARmEB, a data set containing 7 *Arabidopsis* flowering traits along with 216 130 SNPs in Atwell *et al.* (2010) were reanalyzed by pLARmEB, EMMA, FASTmrEMMA and

Table 1 *AIC* and *BIC* values for the regression of significantly associated SNPs on each *Arabidopsis* flowering time trait using pLARmEB, EMMA, FASTmrEMMA and mrMLM

Trait	*BIC*				*AIC*			
	pLARmEB	*EMMA*	*FASTmrEMMA*	*mrMLM*	*pLARmEB*	*EMMA*	*FASTmrEMMA*	*mrMLM*
LD	63.53	289.74	263.56	260.60	−26.90	286.62	201.20	195.12
LDV	−306.01	−104.50	−157.79	−142.31	−380.99	−113.87	−198.40	−176.67
SD	−118.34	118.17	48.55	31.26	−251.10	115.08	2.24	−42.84
SDV	−155.98	90.55	124.10	−96.31	−269.53	75.20	78.07	−148.49
FT10	−390.40	28.18	−99.08	−216.17	−514.58	24.92	−164.44	−281.52
FT16	−6.09	222.04	189.81	192.32	−84.40	218.78	144.13	127.06
FT22	182.71	332.36	283.04	235.13	120.72	329.10	230.84	160.09

Abbreviations: *AIC*, *Akaike information criterion*; *BIC*, *Bayesian information criterion*; EMMA, efficient mixed model association; FASTmrEMMA, fast multi-locus random-SNP-effect EMMA; FT10, FT16 and FT22, days to flowering at 10, 16 and 22 °C, respectively; LD, days to flowering under long days; LDV, days to flowering under long days with vernalization; mrMLM, multilocus random-SNP-effect mixed linear model; pLARmEB, polygenic-background-control-based least angle regression plus empirical Bayes; SD, days to flowering under short days; SDV, days to flowering under short days with vernalization; SNP, single-nucleotide polymorphism.

Table 2 The previously reported genes for seven flowering time traits in *Arabidopsis* that were detected only by pLARmEB

Trait[a]	*Gene*	*Chr.*	*SNP (bp)*	P-*value*	*Effect*	*LOD*	r^2 *(%)*	*Trait*[a]	*Gene*	*Chr.*	*SNP (bp)*	P-*value*	*Effect*	*LOD*	r^2 *(%)*
LD	*AT3G56960*	3	21 079 518	1.52E−03	−0.040	2.18	0.38	SDV	*AT2G19690* *AT2G19760*	2	8 516 520	6.65E−05	0.030	3.45	0.43
	AT5G11320	5	3594 757	7.54E−05	−0.028	3.40	0.23		*AT2G32700*	2	13 853 405	4.60E−08	−0.066	6.49	3.29
	AT5G64510	5	25 783 160	7.64E−05	0.021	3.40	0.12		*AT4G12920*	4	7 586 463	3.19E−10	−0.071	8.59	2.07
LDV	*AT1G68050* *AT1G68090* *AT1G68130*	1	25 525 403	8.71E−08	0.039	6.22	3.40		*AT5G01600*	5	239 433	8.39E−07	0.036	5.27	0.90
	AT2G19690 *AT2G19760*	2	8 516 520	7.37E−09	0.048	7.26	3.29		*AT5G16780*	5	5 526 925	4.30E−04	0.024	2.69	0.53
	AT3G07050	3	2 215 112	5.69E−06	0.029	4.47	1.99		*AT5G45890*	5	18 607 728	1.23E−03	−0.014	2.27	0.13
SD	*AT1G01510* *AT1G01530*	1	192 020	1.88E−06	0.029	4.93	0.52	FT10	*AT1G61290*	1	22 619 960	9.12E−06	0.012	4.28	0.56
	AT1G68090 *AT1G68130*	1	25 532 914	7.90E−13	0.036	11.14	1.18		*AT2G01200*	2	134 343	1.03E−05	−0.013	4.22	0.71
	AT2G07020 *AT2G07040* *AT2G07050*	2	2 910 430	3.63E−10	−0.036	8.53	1.71		*AT2G03500*	2	1 076 833	1.30E−05	0.006	4.13	0.15
	AT2G22540	2	9 588 685	1.00E−16	−0.072	16.74	5.41		*AT2G18790*	2	8 124 967	1.98E−04	0.019	3.01	0.81
	AT2G27990	2	11 931 686	4.95E−14	0.041	12.32	2.25		*AT3G47870*	3	17 653 089	9.62E−08	−0.015	6.18	0.46
	AT3G01780	3	286 197	1.29E−04	−0.017	3.18	0.32		*AT4G01220*	4	518 797	1.47E−07	−0.024	6.00	2.35
	AT3G28780	3	10 816 150	2.21E−08	−0.049	6.80	1.50		*AT4G33240*	4	16 017 869	6.61E−05	−0.006	3.46	0.08
	AT3G55200 *AT3G55220*	3	20 477 225	1.49E−03	0.011	2.19	0.16	FT16	*AT2G03060*	2	882 256	1.48E−03	−0.022	2.19	0.29
	AT4G00650	4	268 809	2.95E−06	0.013	4.74	0.14		*AT3G56960*	3	21 079 518	3.24E−04	−0.043	2.81	1.01
	AT4G03090 *AT4G03110*	4	1 371 766	1.32E−06	0.023	5.08	0.66		*AT4G01220*	4	500 090	5.46E−10	−0.090	8.36	3.17
	AT5G45890	5	18 611 542	1.00E−07	0.015	6.16	0.22	FT22	*AT1G52740*	1	19 629 918	3.00E−06	−0.087	4.74	2.24
	AT5G59570	5	24 008 772	8.91E−06	0.010	4.28	0.07		*AT1G71270*	1	26 869 825	8.33E−07	0.118	5.27	1.90
	AT5G63160	5	25 347 883	2.62E−04	0.002	2.89	0.002		*AT4G34040*	4	16 310 486	4.26E−04	0.064	2.70	0.76

Abbreviations: Chr., chromosome; LOD, logarithm (base 10) of odds; pLARmEB, polygenic-background-control-based least angle regression plus empirical Bayes; SNP, single-nucleotide polymorphism.
[a]Trait abbreviations are the same as those in Table 1.

mrMLM. All the significantly associated SNPs were used to fit the regression for each trait and model fitness was reflected by *AIC* and *BIC* values. The *AIC* values for all the seven traits based on pLARmEB were much lower than those based on EMMA, FASTmrEMMA and mrMLM (Table 1). Hence, FASTmrEMMA and mrMLM were better than EMMA and a similar result was also observed from the *BIC* values. The finding suggests that pLARmEB is better in model fit than EMMA, FASTmrEMMA and mrMLM.

Within 20 kb of each SNP significantly associated with traits, we mined candidate genes for these traits. Among the genes identified in previous studies, pLARmEB, FASTmrEMMA and mrMLM identified more previously reported genes than EMMA (Supplementary Table S6). For example, pLARmEB, FASTmrEMMA and mrMLM identified more than three genes for long days with vernalization, whereas EMMA detected only one gene (*AT5G45890*). A similar trend was also observed for other traits (Supplementary Table S6). Among these previously

reported genes, 48 were identified only by pLARmEB (Table 2). Interestingly, genes *AT2G19690* and *AT2G19760* identified by pLARmEB were associated simultaneously with long days with vernalization and short days with vernalization SDV, and three genes (*AT2G07020*, *AT2G07040* and *AT2G07050*) adjacent to the SNP at 2 910 430 bp of chromosome 2 were found to be associated with short days.

DISCUSSION

Analysis of one random sample in the first Monte Carlo simulation experiment using LARS, empirical Bayes and pLARmEB showed that LARS identified many QTNs with small effects in addition to all the simulated QTNs, and thus its FPR was high (Figure 3a). The empirical Bayes was also able to identify simulated and small-effect QTNs although FPR was decreased (Figure 3b), and pLARmEB detected almost all the simulated QTNs and the effects of nonsimulated QTNs were almost close to zero (Figure 3c). More importantly, 48 previously reported genes in *Arabidopsis* were identified only by pLARmEB. Therefore, pLARmEB is a good alternative method for multilocus GWAS.

Although pLARmEB was proposed for GWAS, it is appropriate for mapping populations of backcross, doubled haploid and recombinant inbred lines. To illustrate the effectiveness of pLARmEB, pseudo-markers in every *d* cM were created genome-wide, and the fourth Monte Carlo simulation experiment with 200 simulated data sets was conducted and analyzed using pLARmEB and empirical Bayes. The higher power for QTL detection and less bias for the QTL-effect estimates were observed from pLARmEB than from empirical Bayes (Supplementary Table S4). pLARmEB is also suitable for a population consisting of chromosome segment substitution lines. However, we can only scan marker positions, because we cannot calculate

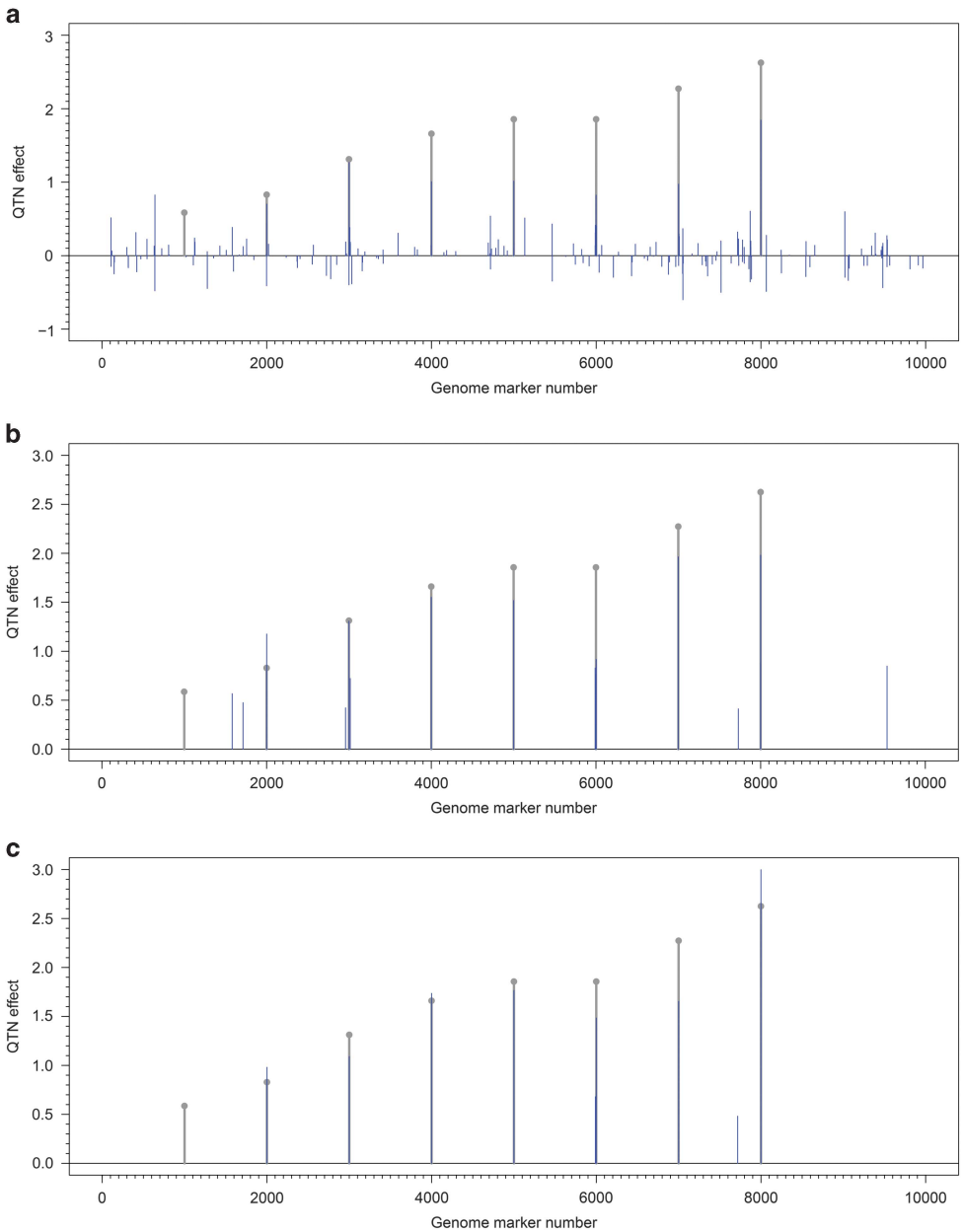

Figure 3 Comparison of least angle regression (**a**), empirical Bayes (**b**) and pLARmEB (**c**) in the estimation of QTN effects in one random sample of the first simulation experiment.

conditional probabilities of pseudo-marker positions. If the number of genotypes in a mapping population is more than two, for example, AA, Aa and aa in F_2, the current method requires some modifications.

Among the previously identified genes in *Arabidopsis* (Supplementary Table S6), a few were found commonly by several approaches and this is different from linkage analysis. The main reason is that GWAS mapping population has a complicated population structure. Although pLARmEB, FASTmrEMMA, and mrRMLM had similar powers of QTN detection in the simulation experiments, different previously reported genes were detected in real data analysis. For example, 48 previously reported genes were identified only by pLARmEB (Table 2). For this reason, we recommend pLARmEB as an alternative method for GWAS and also recommend the joint implementation of several methods in the GWAS analyses of one trait.

The *AIC* or *BIC* values of FASTmrEMMA in Wen *et al.* (2017) and mrRMLM in Wang *et al.* (2016) are different from the corresponding values in this study. In this study, we considered population structure in GWAS. With the inclusion of population structure in genetic model, some different SNPs are found to be significantly associated with the trait. The above two differences result in different *AIC* or *BIC* values for the same trait in different studies.

Multilocus GWAS has become the state-of-the-art GWAS procedure. Iwata *et al.* (2007, 2009) developed multilocus Bayesian GWAS approaches for quantitative and ordinal traits, although running time is a major concern. Segura *et al.* (2012) proposed a multilocus linear mixed model method that is simple, stepwise mixed model regression with forward inclusion and backward elimination. Wang *et al.* (2016) suggested mrMLM and Wen *et al.* (2017) proposed FASTmrEMMA. To make assumptions more suitable to a given data set, Zhou *et al.* (2013) and Moser *et al.* (2015) proposed a hybrid method of mixed linear model and sparse regression model, named Bayesian sparse linear mixed model. In this study, the integration of LARS with empirical Bayes under polygenic background control provides one simple and efficient way for multilocus GWAS. In *Arabidopsis* real data analysis, the number of SNPs was >1000 times larger than sample size and we were able to scan each chromosome by LARS and include all the associated SNPs across the genome in the multilocus model and estimate their effects by empirical Bayes, and thus pLARmEB is better than EMMA.

To obtain low FPR in GWAS, a relatively stringent significance criterion is widely adopted, such as Bonferroni correction. Even after using a less stringent significance criterion (such as LOD = 2.0), pLARmEB has less FPR and higher power than EMMA. We also conducted GEMMA (Zhou and Stephens, 2012) and its power is same as that of EMMA (results not shown). pLARmEB works better than all the other methods considered.

DATA ARCHIVING

All simulated data sets are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.sk652. The real data set can be retrieved from: http://www.arabidopsis.org/.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (31301229, 31571268), and Huazhong Agricultural University Scientific and Technological Self-innovation Foundation (Program No. 2014RC020).

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y *et al.* (2010). Genome-wide association study of 107 phenotypes in a common set of Arabidopsis thaliana inbred lines. *Nature* **465**: 627–631.

Bu SH, Zhao XW, Yi C, Wen J, Tu JX, Zhang YM (2015). Interacted QTL mapping in partial NCII design provides evidences for breeding by design. *PLoS One* **10**: e0121034.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). Least angle regression. *Ann Statist* **32**: 407–451.

Fan J, Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B* **70**: 849–911.

Fridley BL, Serie D, Jenkins G, White K, Bamlet W, Potter JD *et al.* (2010). Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. *Genet Epidemiol* **34**: 418–426.

George EI, McCulloch RE (1993). Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**: 881–889.

Henderson HV, Searle SR (1980). *On Deriving the Inverse of a Sum of Matrices*. Biometrics Unit, Cornell University: Ithaca, New York. Paper No. BU-647-M in the Biometrics Unit Series.

Hoerl AE, Kennard RW (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* **12**: 55–67.

Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet* **4**: e1000130.

Hoffman GE, Logsdon BA, Mezey JG (2013). PUMA: a unified framework for penalized multiple regression analysis of GWAS data. *PLoS Comput Biol* 9: e1003101.

Iwata H, Uga Y, Yoshioka Y, Ebana K, Hayashi T (2007). Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among Oryza sativa L. germplasms. *Theor Appl Genet* **114**: 1437–1449.

Iwata H, Ebana K, Fukuoka S, Jannink JL, Hayashi T (2009). Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among Oryza sativa L. germplasms. *Theor Appl Genet* **118**: 865–880.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al.* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**: 1709–1723.

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet* **12**: e1005767.

Lü HY, Liu XF, Wei SP, Zhang YM (2011). Epistatic association mapping in homozygous crop cultivars. *PLoS One* **6**: e17773.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet* **11**: e1004969.

Park T, Casella G (2008). The Bayesian Lasso. *J Am Stat Assoc Theor Methods* **103**: 681–686.

Rincent R, Moreau L, Monod H, Kuhn E, Melchinger AE, Malvar RA *et al.* (2014). Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics* **197**: 375–387.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q *et al.* (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**: 825–830.

Tibshirani R (1996). Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* **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.

Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ *et al.* (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep* **6**: 19444.

Wei JL, Xu S (2016). A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. *Genetics* **202**: 471–486.

Wen YJ, Zhang H, Ni YN, Huang B, Zhang J, Feng JY *et al.* (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinform*; e-pub ahead of print 1 February 2017; doi:10.1093/bib/bbw145.

Wu TT, Chen YT, Sobel E, Lange K (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**: 714–721.

Xu S (2010). An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* **105**: 483–494.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**: 100–106.

Yi N, Xu S (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045–1055.

Yi N, George V, Allison DB (2003). Stochastic search variable selection for identifying quantitative trait loci. *Genetics* **164**: 1129–1138.

Yu J, Pressoir G, Briggs WH, Vroh Bil, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.

Zhang J, Yue C, Zhang YM (2012). Bias correction for estimated QTL effects using the penalized maximum likelihood method. *Heredity* **108**: 396–402.

Zhang YM, Mao Y, Xie C, Smith H, Luo L, Xu S (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* **169**: 2267–2275.

Zhang YM, Xu S (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* **95**: 96–104.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA *et al.* (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**: 355–360.

Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**: 821–824.

Zhou X, Carbonetto P, Stephens M (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet* **9**: e1003264.

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