Iteratively reweighted LASSO for mapping multiple quantitative trait loci

Yongxin Liu*, Tianfu Yang*, Hongwang Li and Runqing Yang

Submitted: 10th July 2012; Received (in revised form): 24th August 2012

Abstract

The iteratively reweighted least square (IRLS) method is mostly identical to maximum likelihood (ML) method in terms of parameter estimation and power of quantitative trait locus (QTL) detection. But the IRLS is greatly superior to ML in terms of computing speed and the robustness of parameter estimation. In conjunction with the priors of parameters, ML can analyze multiple QTL model based on Bayesian theory, whereas under a single QTL model, IRLS has very limited statistical power to detect multiple QTLs. In this study, we proposed the iteratively reweighted least absolute shrinkage and selection operator (IRLASSO) for extending IRLS to simultaneously map multiple QTLs. The LASSO with coordinate descent step is employed to efficiently estimate non-zero genetic effect of each locus scanned over entire genome. Simulations demonstrate that IRLASSO has a higher precision of parameter estimation and power to detect QTL than IRLS, and is able to estimate residual variance more accurately than the unweighted LASSO based on LS. Especially, IRLASSO is very fast, usually taking less than five iterations to converge. The barley dataset from the North American Barley Genome Mapping Project is reanalyzed by our proposed method.

Keywords: QTL mapping; IRLS; LASSO; multiple QTL model

INTRODUCTION

In earlier quantitative trait loci (QTL) mapping based on a single QTL model, least square (LS) method developed by Haley and Knott [1] was generally considered to be simpler and more robust than maximum likelihood method [2, 3]. However, LS method would result in the inflation of residual variance after replacing QTL genotype indicator by its expectation conditional on the flanking markers [1]. By quantifying the inflated residual variance, iteratively reweighted least square (IRLS) was therefore proposed by Xu [4–6]. Simulations demonstrated that for normal traits, the performance of IRLS is almost identical to that of maximum likelihood (ML), but the computational speed of IRLS is many times faster than that of ML [5]. In addition, IRLS inherits the robustness of LS. The estimates of IRLS are essentially suboptimal when the weighted value was treated as a constant. As a result, Han and Xu [7] introduced the Fisher scoring algorithm for finding the maximum likelihood estimates of QTL parameters. Within the framework of generalized linear model, subsequently, the IRLS was extended to analyzing discrete traits, such as binary, ordinal and Poisson traits, by taking into account the heterogeneity of individual residual variances [8]. Meanwhile, this method is more appropriate to identify large QTL that locates in large marker intervals than the ML [7, 8].

As most of quantitative traits are controlled by multiple genetic and environmental factors, aforementioned interval mapping methods are not optimal without considering other linked QTLs. Thus, some...
mapping methodologies based on multiple QTL model were developed to improve the power of QTL detection. Theoretically, the multiple QTL mapping can be regarded as the statistical question of model search and selection, for either non-Bayesian model selection methods [9–16] or Bayesian model selection methods [17–22]. Compared with interval mapping, these advanced statistical methods tend to be more sensitive to the experimental designs, the data structures, the initial values of the parameters and other factors such as assumptions of the models. Specifically, the results of Bayesian mapping are hard to be interpreted due to lack of significance test. Although the Bayes factor was used to infer the significance of QTL in Bayesian model selection mapping by Yi et al. [20, 21], its distribution also depends on population structure and marker density. Che and Xu [23] have proposed significance test in Bayesian shrinkage analysis by using permutation test, but computation burden of the approach is heavy.

Currently, the Bayesian shrinkage methods increasingly become a promising and active research area of QTL mapping and genomic selection [15, 24–28], in which, all estimated effects are forced to shrink toward zero. The fully Bayesian shrinkage has heavy computing burden and thus has not been widely applied to mapping QTL. In contrast, empirical Bayes methods [24, 29, 30] can produce results similar to those of the fully Bayesian approaches but could be quickly computed. LASSO is another shrinkage method that has been widely used in regression analysis for large models [31]. It can solve large equations extremely fast using coordinate descent algorithm [32]. If a double-exponential prior is assigned to regression coefficient, the Bayesian estimates are equivalent to the LASSO estimates [33]. Many improved LASSO [16, 26, 29, 34] have been developed for QTL mapping. By choosing various priors for QTL effects, those mapping methods may enhance shrinking efficiency to a certain extent, but their computing speeds are far inferior to regular LASSO with a double-exponential prior [31]. Moreover, these studies only analyze known markers and do not provide an efficient and convincing method for statistically inferring QTL.

In this study, we extend Xu’s IRLS to simultaneously mapping multiple QTLs. Fast LASSO [32, 35] is employed to estimate genetic effect of each locus scanned over entire genome. This mapping method is known as iteratively reweighted LASSO. It can, not only estimate QTL effects, but also test the significance of QTLs, like regular interval mapping. The performance of our proposed method is demonstrated via computer simulation. Finally, we apply this method to reanalyzing the barley dataset from the North American Barley Genome Mapping Project.

**METHOD**

**Genetic model**

Our mapping method is developed for improving linkage analysis. It requires that on $n$ individuals, phenotypes are observed and $m$ co-dominant markers with known linkage map are genotyped in a genetically designed population. Prior to mapping QTL, all positions on whole genome may be candidates of QTLs. The difference among present mapping methods lies mainly in the way to search QTLs on the genome. Traditional interval mapping identifies QTL by genome-wide scanning, whereas Bayesian mapping by randomly sampling QTL positions. For mapping multiple QTLs, we partition the entire genome into $k$ loci that include not only the marker positions but also points between markers. The points between markers are equally spaced by the length of scanning step or tuning value used in sampling QTL position. The linear model of phenotypic value ($y_i$, for $i = 1, 2, \ldots, n$) on QTL effects of $k$ loci is then expressed as

$$
y_i = \sum_{j=1}^{k} x_{ij} \beta_j + \sum_{j=1}^{k} z_{ij} \gamma_j + \epsilon_i
$$

where $\beta_j$ is the $j$th systematic environmental effect that is irrelevant to QTL effects, $x_{ij}$ is the corresponding design matrix on the $i$th individual, $\gamma_j$ is the $j$th locus genetic effect, $z_{ij}$ is the corresponding incidence matrix determined by the genotypes of locus $k$, $\epsilon_i$ is the residual error, which is assumed to be a normal distribution $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma^2$ being residual variance.

In the Equation (1), $z_{ij}$ is missing value unless the locus is just a fully informative marker. Following LS method by Haley and Knott [1], we can simply estimate unknown $z_{ij}$ by the expectation of $z_{ij}$ conditional on flanking markers genotype. Thus, Equation (1) becomes

$$
y_i = \sum_{j=1}^{k} x_{ij} \beta_j + \sum_{j=1}^{k} u_{ij} \gamma_j + \epsilon_i
$$

where $u_{ij} = E(z_{ij})$ is the conditional expectation of $z_{ij}$, which depends on the genetically designed population (see the Appendix for details).
Considering the heterogeneity of residual variance after replacing \( z_j \) by its conditional expectation \( u_{ij} \), we exactly write Equation (1) by

\[
\gamma_i = \sum_{j=1}^k x_{ij} \beta_j + \sum_{j=1}^k u_{ij} \gamma_j + \epsilon_i
\]

Let

\[
e_i = \sum_{j=1}^k (z_{ij} - u_{ij}) \gamma_j + \epsilon_i
\]

be the new residual error, which combines unobservable \( z_{ij} \) with the residual error \( \epsilon_i \). Then, Equation (2) can be rewritten as

\[
\gamma_i = \sum_{j=1}^k x_{ij} \beta_j + \sum_{j=1}^k u_{ij} \gamma_j + \epsilon_i
\]

The expectation of \( \epsilon_i \) is

\[
E(\epsilon_i) = [E(Z_i) - U_i] \gamma + E(\epsilon_i) = 0
\]

and its variance is

\[
\text{var}(\epsilon_i) = \sigma_i^2 = \gamma^T \Sigma_i \gamma + \sigma^2
\]

with \( \Sigma_i = \text{Cov}(Z_i) = \left\{ (1 - 2r_{ij}) \sqrt{\text{var}(z_{ij})} \sqrt{\text{var}(z_{ij'})} \right\} \), for \( j \) and \( j' \) = 1, 2, ..., \( k \) being the covariance matrix of indicator variances. Different from what Xu [6] defined earlier, \( \Sigma_i \) is derived from multiple QTLs and considers the linkages among multiple QTLs, where \( r_{ij} \) is recombination coefficient between loci \( j \) and \( j' \), \( r_{ij} = 0 \) if \( j = j' \), \( \text{var}(z_{ij}) \) and \( \text{var}(z_{ij'}) \) are defined as conditional variances at loci \( j \) and \( j' \) for the \( i \)th individual, which have explicit form:

\[
\text{var}(z_{ij}) = E(z_{ij}^2) - [E(z_{ij})]^2 \quad \text{for } l = j, j'
\]

The equation has different form due to genetically designed population, as shown in the Appendix. If these loci overlap with completely informative markers, \( \text{var}(z_{ij}) = 0.5 \) under the assumption of no segregation distortion.

### Parameter estimation

The \( s \) in Equation (5) can be relatively small, especially in genetically designed population with a single family, but \( k \) may vary large and sometimes may be even larger than the sample size \( n \). Thus, we decide to adopt a modified LASSO to estimate the effects of these loci.

Transform residual variance \( \sigma_i^2 \) into \( \frac{1}{w_i} = \sigma^2 \) with \( \sigma_i^2 = \frac{1}{w_i}, \sigma^2 = \left\{ \frac{1}{w_i} \Sigma_i \gamma + 1 \right\} \sigma_i^2 \) being weighted value, and let \( y'_i = y_i \sqrt{w_i} \), \( x'_i = x_{ij} \sqrt{w_i} \) and \( u'_i = u_{ij} \sqrt{w_i} \). By minimizing \( Q = \sum_{i=1}^n \left( y'_i - \sum_{j=1}^k x'_i \beta_j - \sum_{j=1}^k u'_i \gamma_j \right)^2 + \lambda \sum_{j=1}^k |\gamma_j| \), the regression coefficients in Equation (5) can be solved using extremely fast LASSO with a coordinate descent step [35]: to partially optimize with respect to \( \gamma_j (l = 1, 2, ..., k) \), the gradient at \( \gamma_j = \hat{\gamma}_j \) is computed by

\[
\frac{\partial Q}{\partial \gamma_j} \bigg|_{\gamma = \hat{\gamma}} = -\frac{1}{n} \sum_{i=1}^n u'_i \left( y'_i - \sum_{j=1}^k x'_i \beta_j - \sum_{j=1}^k u'_i \gamma_j \right) + \lambda \gamma_j
\]

The equation is required to treat separately by \( \hat{\gamma}_j > 0, \hat{\gamma}_j < 0 \) or \( \hat{\gamma}_j = 0 \). Simple calculus for the coordinate-wise update [36, 37] is implemented in the following way

\[
\hat{\gamma}_j = \frac{1}{1 + \lambda} \left( \frac{1}{n} \sum_{i=1}^n u'_i (y'_i - \hat{\gamma}_j^{(0)}), \lambda \right)
\]

where, \( \hat{\gamma}_j^{(0)} = \sum_{i=1}^n x'_i \beta_j + \sum_{i=1,j \neq l}^k u'_i \gamma_j \) is the fitted value excluding the genetic effect for the \( l \)th loci, and \( \frac{1}{n} \sum_{i=1}^n u'_i (y'_i - \hat{\gamma}_j^{(0)}) \) denoted by \( z \) is the simple least-squares coefficient when fitting this partial residual to the \( l \)th loci. \( S(z, \lambda) \) is the soft-thresholding operator with value

\[
\text{sign}(z)(|z| - \lambda)_+ = \begin{cases} 
  z - \lambda & \text{if } z > 0 \text{ and } \lambda < |z| \\
  z + \lambda & \text{if } z < 0 \text{ and } \lambda < |z| \\
  0 & \text{if } \lambda \geq |z|
\end{cases}
\]

As a tuning parameter, \( \lambda \) can be chosen by cross-validation for LASSO. In our computation, however, path-wise coordinate descent proposed by Friedman et al. [32] is adopted to choose lambda efficiently.

Because \( w_i \) is a function of estimated parameters, iterations are required. The iteration process is recommended as follow:

(i) Initialize \( w_i = 1 \)

(ii) Estimate \( \beta_j, \gamma_j \) and \( \sigma^2 \) using unweighted LASSO
(iii) Update $w_i$ using $\sigma^2$ and non-zero $\beta_j$ and $\gamma_j$
(iv) Estimate $\beta_j$ and $\gamma_j$ using the weighted LASSO
(v) Repeat step 3 and 4 until $|\sigma^{2(t+1)} - \sigma^{2(t)}|/\sigma^{2(t)} \leq 10^{-6}$ or $10^{-8}$, where $t$ is iterating time.

Herein, we call the iteration process as iteratively reweighted LASSO.

**Significance test**

A subset of loci with non-zero genetic effects are retained after the shrinkage of IRLASSO. Once the number of non-zero effects is greater than the sample size, we can drop those linked loci by assuming that there is only one QTL within a marker interval. For the retained loci with non-zero effects, the iteratively weighted least square estimates of the parameters are

$$
\begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = 
\begin{bmatrix}
\sum_{i=1}^{n} X_i^T W_i X_i & \sum_{i=1}^{n} X_i^T W_i U_i \\
\sum_{i=1}^{n} U_i^T W_i X_i & \sum_{i=1}^{n} U_i^T W_i U_i
\end{bmatrix}^{-1}
\begin{bmatrix}
\sum_{i=1}^{n} X_i^T W_i y_i \\
\sum_{i=1}^{n} U_i^T W_i y_i
\end{bmatrix}
$$

(10)

where $W_i = diag(w_1, w_2, \ldots, w_n)$.

and

$$
\hat{\sigma}^2 = \frac{1}{n-p-s} \sum_{i=1}^{n} W_i (y_i - X_i \hat{\beta} - U_i \hat{\gamma})^2
$$

(11)

The variance–covariance matrix of the estimated parameter is calculated as

$$
\text{var} \left[ \begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} \right] = 
\begin{bmatrix}
\sum_{i=1}^{n} X_i^T W_i X_i & \sum_{i=1}^{n} X_i^T W_i U_i \\
\sum_{i=1}^{n} U_i^T W_i X_i & \sum_{i=1}^{n} U_i^T W_i U_i
\end{bmatrix}^{-1} \hat{\sigma}^2
$$

(12)

From Equations (10), (11) and (12), we can form $t$-test statistic to test the significance of the retained loci with non-zero effects.

Considering that $t$-test statistic is influenced by population structure and marker density, we use the critical threshold generated by permutation tests [38]. By repeatedly shuffling the relationships between marker genotypes and phenotypes, a series of the maximum $t$-statistic values are calculated, from the distribution of which, the critical threshold is obtained.

**Simulations**

Simulations are conducted to investigate statistical behavior of our proposed method. In a backcross population, we simulate six chromosomes, each with a length of 100 cM and evenly placed co-dominant markers. The 2, 2, 1, 2, 1 and 2 QTLs are respectively put on six chromosomes, whose positions and genetic effects are listed in Tables 1 and 2. To access the influence of sample size and marker density on mapping results, sample size is designed as the three levels: 150, 200 and 300. For sample sizes of 150 and 300, 11 markers are distributed on each chromosome. For sample sizes of 200, there are 21 markers on each of the first 2 chromosomes, 11 on the middle 2 chromosomes and 5 on the rest two chromosomes. Only one QTL is put on each chromosome, whose genetic effects are classified into two levels of 2.5 and 0.8 on the chromosome with same marker density.

The simulated datasets are analyzed by using IRLASSO, unweighted LASSO (UWLASSO) and IRLS. The critical values of the test statistic for declaring statistical significance may be different due to different methods and situations. Under the null model with zero genetic effects, we simulate 500 additional samples to determine the critical values (Table 3). The 400 replicated simulations are used to estimate QTL parameters and to access the statistical power of QTL detection. Statistical power of QTL detection is calculated by each locus as the percentage of the number of those simulations that statistic value exceeds critical value at the locus. Also, false positive rate is assessed with the 400 replicated simulations under the null model with zero genetic effects.

For sample sizes of 150 and 300, the statistical powers to detect QTLs with three mapping methods are given in Table 4, along with false positive rate and exactly identifying proportion for all true QTLs. As can be seen, although the statistical powers at each locus between IRLASSO and UWLASSO are similar, statistical powers of IRLS are greatly lower than those of IRLASSO and UWLASSO at those loci with smaller genetic effects. False positive rates are <10% for the two sample sizes and three mapping methods, but IRLASSO performs lower false positive rates than the others. Exactly identifying proportions for all true QTLs are generally low due to the difficulty to detect the fifth and ninth QTLs with low genetic effects, although IRLASSO is capable of identifying all true QTLs simultaneously with higher statistical power than UWLASSO and IRLS. The two LASSO methods can estimate QTL effects more accurately than IRLS (Table 2). Generally, the statistical performance of each
mapping method increases as sample size and genetic effect increase. Compared with UWLASSO, especially, the IRLASSO is able to precisely evaluate residual variance. With IRLASSO and UWLASSO, the estimates (standard deviations) are 1.09 (0.21) and 1.88 (0.24), respectively, for sample size of 150, whereas 1.07 (0.17) and 1.91 (0.20), respectively, for sample size of 300.

Table 5 shows the statistical powers of QTL detection and estimates of QTL parameters obtained with three mapping methods for sample size of 200. No influence of sparse markers on convergence to the correct solutions has been found. As expected, the statistical power and precision of parameter estimation increase with the increased marker density and QTL effect. The residual variances (standard deviations) are estimated as 1.28 (0.23) and 2.52 (0.29) by IRLASSO and UWLASSO methods, respectively, again demonstrating the capability of more accuracy estimating residual variance by IRLASSO than UWLASSO method.

Examples
The barley dataset from the North American Barley Genome Mapping Project (http://wheat.pw.usda.gov/ggpages/SxM/) is used to demonstrate our proposed methods. This mapping population consists of 150 doubled haploids (DHs) derived from the cross Steptoe × Morex. A total of eight traits are observed under multiple different environments and 223 markers covering a genome of length ~1500 cM are genotyped for each DH line.

Based on the above dataset, we adopt IRLASSO, UWLASSO and IRLS methods, respectively, to locate QTLs for five traits, including alpha amylase (U), grain protein (%), grain yield (MT/ha), height (cm) and lodging (%) in barley. The phenotypic values of these traits are measured by

---

Table 1: Mean estimates and standard deviations (in parentheses) of QTL positions detected with three mapping methods for the simulated datasets with sample sizes 150 and 300

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>QTL no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>IRLASSO</td>
<td>25.1 (2.6)</td>
<td>58.5 (2.9)</td>
<td>151.5 (3.7)</td>
<td>204.3 (12.3)</td>
<td>264.8 (8.0)</td>
<td>335.7 (4.3)</td>
<td>403.9 (12.3)</td>
<td>474.8 (11.5)</td>
<td>524.9 (5.3)</td>
<td>580.5 (4.6)</td>
<td></td>
</tr>
<tr>
<td>UWLASSO</td>
<td>25.0 (2.8)</td>
<td>58.4 (2.9)</td>
<td>151.2 (4.0)</td>
<td>204.8 (12.5)</td>
<td>264.2 (8.1)</td>
<td>335.6 (4.4)</td>
<td>403.9 (13.8)</td>
<td>474.2 (11.7)</td>
<td>524.5 (5.6)</td>
<td>580.5 (4.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>44.6 (8.3)</td>
<td>48.3 (9.3)</td>
<td>166.1 (3.1)</td>
<td>186.7 (10.8)</td>
<td>267.1 (20.1)</td>
<td>322.4 (15.3)</td>
<td>386.7 (5.2)</td>
<td>480.3 (8.3)</td>
<td>533.5 (7.5)</td>
<td>580.1 (13.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>IRLASSO</td>
<td>25.4 (1.9)</td>
<td>58.8 (2.1)</td>
<td>151.6 (3.0)</td>
<td>196.6 (2.3)</td>
<td>269.5 (5.0)</td>
<td>337.1 (2.9)</td>
<td>395.3 (3.2)</td>
<td>482.4 (2.1)</td>
<td>526.2 (4.7)</td>
<td>580.6 (3.1)</td>
<td></td>
</tr>
<tr>
<td>UWLASSO</td>
<td>25.3 (2.2)</td>
<td>58.5 (2.6)</td>
<td>151.4 (3.2)</td>
<td>196.7 (2.3)</td>
<td>269.4 (5.0)</td>
<td>336.9 (3.3)</td>
<td>395.0 (3.4)</td>
<td>482.2 (2.5)</td>
<td>525.8 (3.2)</td>
<td>580.3 (3.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>45.0 (7.7)</td>
<td>48.2 (8.8)</td>
<td>169.1 (11.2)</td>
<td>188.8 (9.2)</td>
<td>265.9 (20.3)</td>
<td>332.4 (13.7)</td>
<td>388.8 (13.8)</td>
<td>481.2 (5.8)</td>
<td>537.8 (7.1)</td>
<td>582.3 (10.6)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Table 2: Mean estimates and standard deviations (in parentheses) of QTL effects obtained with three mapping methods for the simulated datasets with sample sizes 150 and 300

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>QTL no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>IRLASSO</td>
<td>1.52 (0.19)</td>
<td>1.98 (0.19)</td>
<td>0.91 (0.28)</td>
<td>0.74 (0.54)</td>
<td>−0.21 (0.17)</td>
<td>0.61 (0.19)</td>
<td>−0.40 (0.41)</td>
<td>1.12 (0.24)</td>
<td>0.35 (0.16)</td>
<td>−0.77 (0.15)</td>
<td></td>
</tr>
<tr>
<td>UWLASSO</td>
<td>1.52 (0.19)</td>
<td>1.98 (0.19)</td>
<td>0.91 (0.28)</td>
<td>0.71 (0.56)</td>
<td>−0.21 (0.18)</td>
<td>0.61 (0.20)</td>
<td>−0.39 (0.42)</td>
<td>1.10 (0.26)</td>
<td>0.34 (0.16)</td>
<td>−0.77 (0.16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>2.92 (0.22)</td>
<td>2.95 (0.22)</td>
<td>1.43 (0.29)</td>
<td>1.52 (0.30)</td>
<td>−0.28 (0.48)</td>
<td>0.50 (0.45)</td>
<td>−0.34 (0.52)</td>
<td>1.29 (0.28)</td>
<td>−0.08 (0.52)</td>
<td>−0.72 (0.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>IRLASSO</td>
<td>1.55 (0.11)</td>
<td>1.95 (0.12)</td>
<td>0.72 (0.09)</td>
<td>1.09 (0.10)</td>
<td>−0.23 (0.07)</td>
<td>0.70 (0.10)</td>
<td>−0.65 (0.10)</td>
<td>1.24 (0.07)</td>
<td>0.35 (0.10)</td>
<td>−0.79 (0.10)</td>
<td></td>
</tr>
<tr>
<td>UWLASSO</td>
<td>1.55 (0.14)</td>
<td>1.94 (0.14)</td>
<td>0.71 (0.11)</td>
<td>1.09 (0.11)</td>
<td>−0.22 (0.09)</td>
<td>0.69 (0.11)</td>
<td>−0.65 (0.12)</td>
<td>1.24 (0.08)</td>
<td>0.35 (0.11)</td>
<td>−0.78 (0.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>2.88 (0.18)</td>
<td>2.90 (0.18)</td>
<td>1.35 (0.22)</td>
<td>1.46 (0.21)</td>
<td>−0.26 (0.32)</td>
<td>0.49 (0.33)</td>
<td>−0.39 (0.40)</td>
<td>1.27 (0.21)</td>
<td>−0.15 (0.42)</td>
<td>−0.72 (0.26)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Table 3: The empirical critical thresholds obtained by three mapping methods under different sample sizes

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample size</th>
<th>150</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRLASSO</td>
<td>2.39</td>
<td>2.39</td>
<td>2.54</td>
<td></td>
</tr>
<tr>
<td>UWLASSO</td>
<td>2.38</td>
<td>2.36</td>
<td>2.55</td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>2.90</td>
<td>2.83</td>
<td>2.98</td>
<td></td>
</tr>
</tbody>
</table>
using the averages of observations across multiple environments.

For convenience to compare three mapping methods, we transform statistic $t$ to $-\log(P)$, where $P$ is the probability of greater than statistic value $t$. The genome-wide empirical critical thresholds for five traits are obtained by using 1000 permutation tests. Under the same mapping method, obviously, the critical thresholds are very near among the traits analyzed, as pointed by horizontal reference line in Figure 1. Figure 1 plots the profiles of $-\log(P)$ statistics over the genome for the five traits obtained with IRLASSO method. As shown, there are some peaks exceeding corresponding critical thresholds for all traits analyzed, providing the evidence for existence of QTLs controlling alpha amylase, grain protein, grain yield, height and lodging. Specifically, many QTLs are identified on markers: three markers for alpha amylase, five markers for grain protein, two markers for grain yield, three markers for height and two markers for lodging. UWLASSO identifies the totally same QTLs as those detected with IRLASSO, but more QTLs than those detected with IRLS. Other three traits have also been analyzed, which support the advantage of our proposed method (Results not shown).

Table 6 tabulates parameter estimates of QTLs detected with three mapping methods. With IRLASSO, most of the QTLs are identified to be positive effects, suggesting that the alleles of parent ‘Steptoe’ contribute to increase in all traits except for malt extract. The proportions of phenotypic variation explained by the detected QTLs ranged from 2% to 54%. The largest heritability is 0.54 of the QTL for height on chromosome 2, followed by 0.44 of the QTL for grain yield on chromosome 3 and the smallest one is 0.02 of the QTL for height on chromosome 7.

### Table 4: Statistical powers of QTL detection, false positive rate and exactly identifying proportion for all true QTL obtained with three mapping methods for the simulated datasets with sample sizes 150 and 300

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>Statistical power of each QTL</th>
<th>FPR</th>
<th>EIP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>150</td>
<td>IRLASSO</td>
<td>100</td>
<td>100</td>
<td>96.5</td>
</tr>
<tr>
<td></td>
<td>UWLASSO</td>
<td>100</td>
<td>100</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>IRLS</td>
<td>100</td>
<td>100</td>
<td>88</td>
</tr>
<tr>
<td>300</td>
<td>IRLASSO</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>UWLASSO</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>IRLS</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
</tbody>
</table>

FPR, false positive rate; EIP, exactly identifying proportion.

### Table 5: Statistical powers of QTL detection and parameter estimates (standard deviations) obtained with three mapping methods for the simulated dataset with sample size 200

<table>
<thead>
<tr>
<th>Method</th>
<th>QTL no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Position</td>
<td>56</td>
<td>157</td>
<td>258</td>
<td>359</td>
<td>460</td>
<td>561</td>
</tr>
<tr>
<td></td>
<td>Effect</td>
<td>2.5</td>
<td>0.8</td>
<td>2.5</td>
<td>0.8</td>
<td>2.5</td>
<td>0.8</td>
</tr>
<tr>
<td>IRLASSO</td>
<td>Power</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>99.3</td>
<td>100</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>Position</td>
<td>58.09 (2.98)</td>
<td>461.81 (4.90)</td>
<td>260.96 (2.47)</td>
<td>361.74 (4.09)</td>
<td>159.34 (2.18)</td>
<td>563.60 (4.72)</td>
</tr>
<tr>
<td></td>
<td>Effect</td>
<td>2.47 (0.13)</td>
<td>0.80 (0.11)</td>
<td>2.47 (0.11)</td>
<td>0.79 (0.10)</td>
<td>2.49 (0.11)</td>
<td>0.80 (0.11)</td>
</tr>
<tr>
<td>UWLASSO</td>
<td>Power</td>
<td>100</td>
<td>98</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>98.5</td>
</tr>
<tr>
<td></td>
<td>Position</td>
<td>5783 (3.23)</td>
<td>462.10 (5.19)</td>
<td>260.91 (2.42)</td>
<td>361.56 (4.05)</td>
<td>158.80 (3.18)</td>
<td>563.45 (5.03)</td>
</tr>
<tr>
<td></td>
<td>Effect</td>
<td>2.46 (0.17)</td>
<td>0.81 (0.13)</td>
<td>2.46 (0.12)</td>
<td>0.80 (0.13)</td>
<td>2.47 (0.16)</td>
<td>0.80 (0.13)</td>
</tr>
<tr>
<td>IRLS</td>
<td>Power</td>
<td>100</td>
<td>21</td>
<td>100</td>
<td>29</td>
<td>100</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>Position</td>
<td>5708 (2.35)</td>
<td>458.64 (5.71)</td>
<td>25947 (3.91)</td>
<td>358.20 (4.46)</td>
<td>15794 (3.04)</td>
<td>560.17 (13.51)</td>
</tr>
<tr>
<td></td>
<td>Effect</td>
<td>2.54 (0.27)</td>
<td>0.88 (0.40)</td>
<td>2.49 (0.20)</td>
<td>0.89 (0.29)</td>
<td>2.55 (0.26)</td>
<td>0.91 (0.33)</td>
</tr>
</tbody>
</table>
Figure 1: The profile of $-\log(P)$ test statistics obtained with three mapping methods for the analyzed five traits in barley: (i) alpha amylase, (ii) grain protein, (iii) grain yield, (iv) height and (v) lodging. In each plot, the genome-wide threshold value is given as a horizontal reference line. Chromosomes are separated by the vertical dotted lines and marker positions are indicated by the ticks on the horizontal axis.
For the same trait, the residual variance estimated by UWLASSO is consistently greater than that by IRLASSO, so that the UWLASSO generally underestimates the heritabilities of QTLs detected.

DISCUSSION

In this study, we extend Xu’s IRLS for a single QTL model to simultaneously analyzing multiple QTLs, greatly improving the power to detect QTL. So-called IRLASSO method inherits quick convergence of Xu’s IRLS, and uses extremely fast LASSO with coordinate descent step to estimate non-zero QTL effects. Thus, the mapping method can identify multiple QTLs in manner of high computing efficiency, as compared with existing LASSO for QTL mapping. Through multiple regression analysis for few loci with non-zero effects, the IRLASSO method is capable of statistically inferring the significance of QTL detected. Like regular interval mapping, this significance test includes construction of test statistic and permutation test for determining critical value. It should be clarified that if the marker density is sufficiently high, there is no need for the IRLS because LS, ML and IRLS, all would be the same. High dense marker maps are not always available for many species. Therefore, our proposed method remains useful as long as marker density is low or there are missing genotypes or both. Missing genotypes will always happen no matter how dense a marker map is.

Although the unweighted LASSO inflates the estimation of residual variance, as demonstrated in simulations, it is mostly consistent with IRLASSO method in terms of parameter estimation and power of QTL detection. For simplification of computation, we can omit the iteratively weighted step for LASSO and then choose few loci with non-zero effects obtained with UWLASSO. In this study, the iteratively weighted step is used in multiple regression analysis for few loci with non-zero effects to accuracy evaluate residual variance. To enhance shrinking efficiency, actually, the LASSO used in this study needs to be optimized by using cross-validation. Cross-validation of LASSO can obtain fewer loci with non-zero effects, but it will consume more computing time than iteratively weighted step for LASSO. Our simulation shows that those loci with

<table>
<thead>
<tr>
<th>Trait no.</th>
<th>QTL no.</th>
<th>Marker interval</th>
<th>Additive effect</th>
<th>Residual variance</th>
<th>Heritability</th>
<th>Trait no.</th>
<th>QTL no.</th>
<th>Marker interval</th>
<th>Additive effect</th>
<th>Residual variance</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Brz ~ ABC156D</td>
<td>1.20</td>
<td>1.69</td>
<td>0.22</td>
<td>1</td>
<td>Brz ~ ABC156D</td>
<td>1.20</td>
<td>1.69</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2-88</td>
<td>KFP91</td>
<td>1.00</td>
<td>0.16</td>
<td>5.63</td>
<td>2</td>
<td>2-88</td>
<td>1.00</td>
<td>0.16</td>
<td>5.63</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2-68</td>
<td>MWG557 ~ ABG316C</td>
<td>1.02</td>
<td>0.16</td>
<td>6.81</td>
<td>2</td>
<td>2-68</td>
<td>1.00</td>
<td>0.15</td>
<td>5.10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4-171</td>
<td>CDO69</td>
<td>0.72</td>
<td>0.08</td>
<td>4.37</td>
<td>3</td>
<td>4-171</td>
<td>0.70</td>
<td>0.07</td>
<td>4.20</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1-1018</td>
<td>WG364</td>
<td>0.86</td>
<td>0.12</td>
<td>6.31</td>
<td>4</td>
<td>1-1018</td>
<td>0.86</td>
<td>0.11</td>
<td>6.20</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6:** Estimated QTL parameters from the barley QTL mapping experiment

<table>
<thead>
<tr>
<th>Trait no.</th>
<th>QTL no.</th>
<th>Marker interval</th>
<th>Additive effect</th>
<th>Residual variance</th>
<th>Heritability</th>
<th>Trait no.</th>
<th>QTL no.</th>
<th>Marker interval</th>
<th>Additive effect</th>
<th>Residual variance</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2-76</td>
<td>ABC167B ~ bBE54D</td>
<td>0.38</td>
<td>0.09</td>
<td>0.27</td>
<td>6</td>
<td>2-76</td>
<td>0.38</td>
<td>0.09</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3-16</td>
<td>ABA307B ~ ABC171</td>
<td>0.14</td>
<td>0.04</td>
<td>4.26</td>
<td>7</td>
<td>3-16</td>
<td>0.14</td>
<td>0.04</td>
<td>4.26</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3-52.6</td>
<td>ABG399</td>
<td>0.18</td>
<td>0.06</td>
<td>5.45</td>
<td>8</td>
<td>3-52.6</td>
<td>0.18</td>
<td>0.05</td>
<td>5.40</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4-195</td>
<td>ABG500B</td>
<td>0.26</td>
<td>0.13</td>
<td>7.96</td>
<td>9</td>
<td>4-195</td>
<td>0.26</td>
<td>0.15</td>
<td>8.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6-6.8</td>
<td>MWG620</td>
<td>0.14</td>
<td>0.04</td>
<td>4.36</td>
<td>10</td>
<td>6-6.8</td>
<td>0.14</td>
<td>0.04</td>
<td>4.36</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6-93.5</td>
<td>Nar7</td>
<td>0.16</td>
<td>0.05</td>
<td>4.36</td>
<td>11</td>
<td>6-93.5</td>
<td>0.16</td>
<td>0.05</td>
<td>4.36</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7-52</td>
<td>Ubi2 ~ ksuA3A</td>
<td>0.33</td>
<td>0.20</td>
<td>14.33</td>
<td>12</td>
<td>7-52</td>
<td>0.33</td>
<td>0.20</td>
<td>14.33</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7-1378</td>
<td>MWG514B</td>
<td>0.17</td>
<td>0.05</td>
<td>4.60</td>
<td>13</td>
<td>7-1378</td>
<td>0.17</td>
<td>0.05</td>
<td>4.60</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3-52.6</td>
<td>ABG399</td>
<td>0.26</td>
<td>0.07</td>
<td>16.66</td>
<td>14</td>
<td>3-52.6</td>
<td>0.26</td>
<td>0.07</td>
<td>16.66</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6-675</td>
<td>CDO497</td>
<td>0.12</td>
<td>0.09</td>
<td>4.17</td>
<td>15</td>
<td>6-675</td>
<td>0.12</td>
<td>0.09</td>
<td>4.17</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2-1-20</td>
<td>MWG858 ~ ABG358</td>
<td>5.60</td>
<td>3.72</td>
<td>42.53</td>
<td>16</td>
<td>2-1-20</td>
<td>5.60</td>
<td>3.72</td>
<td>42.53</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2-71.8</td>
<td>ABC167B</td>
<td>2.10</td>
<td>0.07</td>
<td>9.30</td>
<td>17</td>
<td>2-71.8</td>
<td>2.10</td>
<td>0.07</td>
<td>9.30</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3-298</td>
<td>MWG584</td>
<td>1.25</td>
<td>0.03</td>
<td>5.01</td>
<td>18</td>
<td>3-298</td>
<td>1.25</td>
<td>0.03</td>
<td>5.01</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3-56</td>
<td>ABG399 ~ BCD828</td>
<td>2.86</td>
<td>0.14</td>
<td>18.69</td>
<td>19</td>
<td>3-56</td>
<td>2.86</td>
<td>0.14</td>
<td>18.69</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6-6</td>
<td>ABG062 ~ MWG620</td>
<td>0.99</td>
<td>0.02</td>
<td>4.62</td>
<td>20</td>
<td>6-6</td>
<td>0.99</td>
<td>0.02</td>
<td>4.62</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7-80</td>
<td>ABC302 ~ CDO57B</td>
<td>2.66</td>
<td>0.12</td>
<td>16.72</td>
<td>21</td>
<td>7-80</td>
<td>2.66</td>
<td>0.12</td>
<td>16.72</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7-1378</td>
<td>MWG514B</td>
<td>1.23</td>
<td>0.03</td>
<td>5.53</td>
<td>22</td>
<td>7-1378</td>
<td>1.23</td>
<td>0.03</td>
<td>5.53</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1-122.8</td>
<td>KFP910</td>
<td>3.79</td>
<td>3.97</td>
<td>3.79</td>
<td>23</td>
<td>1-122.8</td>
<td>3.79</td>
<td>3.97</td>
<td>3.79</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2-68.2</td>
<td>ABG316C</td>
<td>4.36</td>
<td>0.13</td>
<td>6.76</td>
<td>24</td>
<td>2-68.2</td>
<td>4.36</td>
<td>0.13</td>
<td>6.76</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3-168</td>
<td>ABG316 ~ Pcr1</td>
<td>3.75</td>
<td>0.09</td>
<td>6.40</td>
<td>25</td>
<td>3-168</td>
<td>3.75</td>
<td>0.09</td>
<td>6.40</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3-54</td>
<td>ABG399 ~ BCD828</td>
<td>8.09</td>
<td>0.43</td>
<td>13.43</td>
<td>26</td>
<td>3-54</td>
<td>8.09</td>
<td>0.43</td>
<td>13.43</td>
<td></td>
</tr>
</tbody>
</table>
non-zero effects solved from unweighted LASSO completely cover those from cross-validation of LASSO. Therefore, the cross-validation of LASSO is an option in our mapping method. Without the iteratively weighted step and cross-validation for LASSO, our proposed mapping method is faster than Xu’s IRLS. The program that implemented the mapping method is written in R language, which is freely available upon request from authors.

Only one kind of genetic effect is considered in our model for the simplification of description, which is suitable for backcross (BC), DH and recombinant inbreeding lines (RIL). Our method can be easily applied to analyze multiple genetic effects at each locus in other genetically designed population, such as intercross and outbred populations. The extension of continuous quantitative traits to discrete traits is being studied in our lab. It should be noted that the iteratively weighted step is not straightforward to incorporate in the LASSO for generalized linear model. In comparison, the extension of main effect model to epistatic model is not hard to map multiple interacting QTLs.

Key Points
- We developed the iteratively reweighted least absolute shrinkage and selection operator (IRLASSO) for extending IRLS to simultaneously map multiple QTLs.
- The LASSO with coordinate descent step is employed to efficiently estimate non-zero genetic effect of each locus scanned over entire genome.
- The method can, not only estimate QTL effects, but also test the significance of QTLs, like regular interval mapping.
- Without the iteratively weighted step and cross-validation for LASSO, our proposed mapping method is faster than Xu’s IRLS [4–6].

FUNDING
The National Natural Science Foundations of China (grants 30972077, 31172190 to R.Y. and 31201989 to Y.L.); the Applied National Science and Technology Support Program in China (grant 2012BAD26B01 to Y.L.).

References
APPENDIX
The expectation and variance of indicator variable.
For genetically designed populations with two genotypes at each locus, such as BC, DH and RIL, only additive effect is considered in genetic model. Define \( z = [1, -1]^T \) and let \( P_1 \) and \( P_2 \) be the conditional probabilities for the two genotypes given flanking marker information. The conditional expectation of indicator variable is
\[
E(z) = P_1 - P_2,
\]
and its variance is
\[
Var(z) = 4P_1P_2.
\]

For \( F_2 \) population, there are three genotypes, denoted by QQ, Qq and qq at each locus. Two genetic effects including additive and dominant effects are considered in genetic model. Define
\[
z_1 = \begin{cases} 
1 & \text{for QQ} \\
0 & \text{for Qq} \\
-1 & \text{for qq}
\end{cases}
\]
as indicator variables of additive and dominant effects, respectively. Let \( P_1, P_2 \) and \( P_3 \) be the conditional probabilities for the genotypes QQ, Qq and qq, given flanking marker information. The conditional expectations of indicator variables are
\[
E(z_1) = P_1 - P_3 \quad \text{and} \quad E(z_2) = P_2
\]
and their variance-covariance are
\[
Var(z_1) = P_1(1 - P_1) + P_3(1 - P_3) + 2P_1P_3,
\]
and
\[
Var(z_2) = P_2(1 - P_2).
\]
The covariance of indicator variables between additive and dominant effects is
\[
Cor(z_1, z_2) = 0.
\]