ORIGINAL ARTICLE

Multiple loci *in silico* mapping in inbred lines

H-Y Lü, M Li, G-J Li, L-L Yao, F Lin and Y-M Zhang

Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Department of Crop Genetics and Breeding, Nanjing Agricultural University, Nanjing, China

The *in silico* mapping (ISM) technique and its extension represent major advances for novel gene discovery in germplasm resources of inbred lines. However, the techniques suffer from a relatively high false-positive rate (FPR) and they do not consider the effect of linkage disequilibrium (LD) markers around the identified quantitative trait locus (QTL). In addition, it has not yet been established whether it is optimal to use absolute trait differences as the response variable. To address these problems, this article presents the multiple loci ISM (MLISM) approach, which uses all markers on the entire genome, along with a penalized maximum likelihood. The method proposed here was verified by a

series of simulation experiments with a maize pedigree population of inbred lines of known ancestry. Results from the simulated studies show that the best response variable is the trait product. The MLISM FPR is substantially decreased and the proportion of the number of false QTL to the number of LD markers around the identified QTL is adequately reduced. The MLISM method, with the trait product as the response variable, is an improvement on the existing methods for novel QTL mapping in germplasm resources of inbred lines.

Heredity (2009) **103**, 346–354; doi:10.1038/hdy.2009.66; published online 3 June 2009

Keywords: *in silico* mapping; linkage disequilibrium; penalized maximum likelihood; identity-by-state; false-positive rate

Introduction

Novel genes in crop cultivars or animal inbred strains are a critical resource for plant and animal improvement. Therefore it is relevant to ask whether it is also possible to map a quantitative trait locus (QTL) within the crop cultivars or animal inbred strains. Grupe *et al.* (2001) proposed the use of the mean phenotypic values of inbred strains to map the likely genomic locations of QTL '*in silico*,' which theoretically would represent a major advance. Chesler *et al.* (2001) and Darvasi (2001) have questioned the validity of this so-called *in silico* mapping (ISM) on the basis that it is associated with a relatively high false-positive rate (FPR). Recently, Peltz's group further extended the computational method for mapping phenotypic traits that vary among inbred strains onto haplotypic blocks, known as the haplotype-based ISM (HISM) approach (Liao *et al.*, 2004). This method predicted the genetic basis for strain-specific differences in several biologically important traits of mice (Liao *et al.*, 2004; Guo *et al.*, 2006, 2007; Liang *et al.*, 2006a,b), but it only performs a single-locus analysis. Nevertheless, the advantages of detecting QTL from inbred strains without the need for conventional QTL analysis are strong enough to justify seeking improvements to the ISM methodology.

The ISM and its extension assume the presence of a single QTL per linkage group. This assumption is problematic (Kao *et al.*, 1999; Zhang, 2006), primarily because only the effects of the putative QTL at any

particular marker position can be included in the model and all other QTL effects have to be ignored. As a result, estimates of the effects and the positions of QTL will be biased whenever there is more than one QTL present on a given linkage group. On the other hand, when several markers are in strong linkage disequilibrium (LD), the method can fail to reject any of these markers because of strong linkage. This inevitably leads to a relatively high FPR. Several approaches have been proposed to ameliorate this situation. The early approaches to this problem applied composite interval mapping on a large sample size (Jansen, 1993; Zeng, 1993). More recently, multi-QTL mapping has been developed (Kao *et al.*, 1999; Xu, 2003, 2007; Zhang and Xu, 2005; Zhang, 2006; Xu and Jia, 2007). However, all these approaches are focused on segregating populations from controlled crosses rather than from the germplasm resource of inbred lines. Therefore, a current priority is to incorporate the notion of multi-QTL mapping into the extended ISM, which can be used in a natural population.

Another major concern is whether the choice of absolute trait differences as the response variable is optimal. In Haseman–Elston (H–E) regression (Haseman and Elston, 1972), the earliest suggestion was to apply the squared trait difference as the phenotypic difference. As pointed out by Wright (1997), this approach discards some useful information, and some benefit has been seen in using the trait values of both members of a sib pair. In effect, the squared difference and the trait sum together contain exactly the same information as the original two trait values. Drigalenko (1998) developed the idea further and suggested using the trait product as the phenotypic difference. The result has been further confirmed in multi-QTL H–E regression (Zhang *et al.*, 2008), and these methods were exploited as the basis for the extended ISM described in this article.

Correspondence: Dr Y-M Zhang, College of Agriculture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, PR China.

E-mail: soyzhang@njau.edu.cn

Received 11 February 2009; revised 18 April 2009; accepted 30 April 2009; published online 3 June 2009

In this paper, we show that the ISM can be extended to multi-loci ISM (MLISM) using all markers of a given genome. In the current version of MLISM, the parameters were estimated by the penalized maximum likelihood (PML) method; several response variables for phenotypic difference were compared with each other in order to optimize the procedure. The effect of LD markers around the identified QTL was also investigated.

The new method proposed in this paper was tested by simulation. The purposes of the simulation were three-fold: (1) to select the best response variable, (2) to identify whether the MLISM FPR was substantially decreased and (3) to understand the effect of LD markers around detected QTL.

Materials and methods

Simulation study design

We conducted seven simulation experiments in this paper. In the first, the simulated pedigree was the maize pedigree described by Zhang *et al.* (2008) (Figure 1). The number of inbred lines within the maize pedigree was 404(n). Of these, n_0 (=103) were base (founder) lines, which were in linkage equilibrium so that the genotypes for markers and QTL with four alleles could be simulated. Non-founders (n_1 =301) were bred by repeated self-pollination of a hybrid between two inbred lines. Thus, each non-founder line represents a recombi-

nant inbred line with respect to a pair of known parents. The genotype of all non-founders could be generated from the genotypes of their corresponding parents, analogous to simulating the genotypes of recombinant inbred lines from their two parents. All of the non-founder lines can be used to detect QTL. Sixty-one equally spaced markers were simulated on a single-chromosome segment 600 cM long. A single QTL was located at position 200 cM and overlapped with markers. The environmental variance was calculated as

$$\sigma_e^2 = (1 - h^2)\sigma_g^2/h^2$$

where σ_g^2 is the genetic variance and h^2 is heritability. Allelic effects were calculated by relating the genetic variance of the QTL to both the allelic frequencies and the allelic number. The phenotypic value of each line was the sum of corresponding QTL genotypic values and the residual error, with an assumed $N(0, \sigma_e^2)$ distribution. Each simulation run consisted of 200 replicates. The other simulation experiments were carried out similarly. All simulated parameters are given in Table 1.

Genetic model

Let the k th inbred-line pair among all non-founders have trait phenotypic values ($z_{k,1}$, $z_{k,2}$), and the average value of $z_{k,1}$ and $z_{k,2}$ over all pairs be $\bar{z}$. The absolute trait difference is $y_k^a = |z_{k,1} - z_{k,2}|$, and the identity-by-state (IBS) sharing at a marker locus for the two lines is x . On

Figure 1 The familial relationships between the 404 inbred maize lines used in the all simulation experiments except for the last one.

Table 1 Simulated parameters in all the simulation experiments

Simulated experiment	Pedigree						Allele for QTL and marker		QTL		Reference
	Name	No. of founders	No. of non-founders	No. of chromo-some	Marker density (cM)	Genome length (cM)	Number	Distribution	Position (cM)	Heritability	
1	Maize	101	301	1	Equal, 10	600	4	1:1:1:1	200	0.20	Zhang <i>et al.</i> (2008)
2	Maize	101	301	1	Equal, 10	600	2	1:1	100, 300, 500	0.05, 0.10, 0.15	Zhang <i>et al.</i> (2008)
3	Maize	101	100, 200, 300	2	Equal, 10	100 × 2	3	1:1:1	20, 80	0.10, 0.20	Zhang <i>et al.</i> (2008)
4	Maize	101	301	2	Equal, 10	100 × 2	2, 3, 4	1:1, 1:1:1, 1:1:1:1	20, 80	0.10, 0.20	Zhang <i>et al.</i> (2008)
5	Maize	101	301	2	Equal, 10	100 × 2	3	1:1:1; 1:2:3	20, 80	0.10, 0.20	Zhang <i>et al.</i> (2008)
6	Maize	101	301	12	Unequal	2121.15	2	1:1	0.00, 60.32, 123.52, 942.78, 1672.05, 1723.00	0.025, 0.05 × 3, 0.075, 0.15	Zhu <i>et al.</i> (2007)
7	Simulated	10	100	10	Equal, 2	200 × 10	2	1:1	220, 270, 320, 370, 650, 1050, 1150, 1450, 1850, 1950	0.025 × 3, 0.05 × 4, 0.075, 0.10, 0.15	—

Abbreviation: QTL, quantitative trait locus.

the basis of the ISM of Grupe *et al.* (2001), the regression of y^A on x is described by:

$$y^A = b_0 + bx + e \quad (1)$$

where b_0 is the regression intercept and b is the regression slope (Grupe *et al.*, 2001). Provided that each marker locus on the entire genome can be linked to putative QTL, the model (1) can be extended to the MLISM:

$$y_k^A = b_0 + \sum_{i=1}^p b_i x_{ki} + e_k \quad (2)$$

where b_i is the regression coefficient related to the i th putative QTL; p is the number of all markers on the entire genome; x_{ki} is the IBS of the k th inbred-line pair at the i th marker locus; and e_k is the residual error with an assumed $N(0, \sigma^2)$ distribution. The squared trait difference is given by $y_k^p = (z_{k,1} - z_{k,2})^2$, the trait product by $y_k^p = (z_{k,1} - \bar{z})(z_{k,2} - \bar{z})$, and the trait sum by $y_k^{\bar{z}} = [(z_{k,1} - \bar{z}) + (z_{k,2} - \bar{z})]^2$. The regressions of y_k^p , y_k^p and $y_k^{\bar{z}}$ on the IBS can be established similarly. Response variables are denoted by y .

Parameter estimation

There are several methods to estimate the parameters in model (2), that is, the PML (Zhang and Xu, 2005) and Bayesian LASSO (Park and Casella, 2008; Yi and Xu, 2008). If residual variance is heterozygous, the method proposed by Yi and Banerjee (2009) is available as well. We here adopt the PML method. Briefly, in the PML method, the penalized likelihood function is the product of a likelihood function $L(\theta | \mathbf{Y}, \mathbf{M})$ and penalty function $P(\theta, \xi)$. The likelihood function is calculated by

$$L(\theta | \mathbf{Y}, \mathbf{M}) = \prod_{k=1}^m \varphi(y_k; \alpha_k, \sigma^2) \quad (3)$$

where $\theta = (b_0, b_1, \dots, b_p, c, \sigma^2)$, $m = n_1(n_1 - 1)/2$, $\alpha_k = b_0 + \sum_{i=1}^p b_i x_{ki}$, $\mathbf{Y} = (y_1, y_2, \dots, y_m)^T$, $\mathbf{M}$ represents marker information, and $\varphi(y; \alpha, \sigma^2)$ is a normal density function with mean α and variance σ^2 . The penalty

function is:

$$P(\theta, \xi) = \prod_{i=1}^p [\varphi(b_i; \mu_i, \sigma_i^2) \varphi(\mu_i; 0, \sigma_i^2/\eta) p(\sigma_i^2)] \quad (4)$$

where $\xi = (\mu_1, \dots, \mu_p, \sigma_1^2, \dots, \sigma_p^2)$ is the vector of hyperparameters, and $\eta > 0$ is the prior sample size for assessing μ_i . Note that $p(\sigma_i^2) \propto 1$ for the response variable y_k^p and $p(\sigma_i^2) \sim \text{inv-}\chi^2(v, s_i^2)$, with $s_i^2 = 0$ ($i = 1, 2, \dots, p$), for the other response variables (Zhang *et al.*, 2008). The penalized likelihood function is

$$\psi(\theta, \xi) = L(\theta | \mathbf{Y}, \mathbf{M}) P(\theta, \xi) \quad (5)$$

Thus, the PML estimates for both model parameters and hyperparameters are

$$b_0 = \frac{1}{m} \sum_{k=1}^m (y_k - \sum_{i=1}^p x_{ki} b_i) \quad (6)$$

$$b_i = \left(\sum_{k=1}^m x_{ki}^2 + \sigma^2/\sigma_i^2 \right)^{-1} \times \left[\sum_{k=1}^m x_{ki} (y_k - b_0 - \sum_{t \neq i}^p x_{kt} b_t) + \mu_i \sigma^2/\sigma_i^2 \right] \quad (7)$$

$$\sigma^2 = \frac{1}{m} \sum_{k=1}^m (y_k - b_0 - \sum_{i=1}^p x_{ki} b_i)^2 \quad (8)$$

$$\mu_i = b_i/(\eta + 1) \quad (9)$$

$$\sigma_i^2 = \begin{cases} \frac{1}{2} [(b_i - \mu_i)^2 + \eta \mu_i^2], & \text{for } y^p \\ \frac{1}{v+4} [(b_i - \mu_i)^2 + \eta \mu_i^2 + v s_i^2], & \text{otherwise} \end{cases} \quad (10)$$

The iterative steps for parameter estimation are identical to those given by Zhang and Xu (2005) and He and Zhang (2008). The convergence criterion was $\Sigma |\theta^{(t+1)} - \theta^{(t)}| < 10^{-6}$. In equation (10), the value of v depends on the response variable when $s_i^2 = 0$ ($i = 1, \dots, p$). From a wide range of values, we have empirically determined that v should be set to 6 for y_k^A and 7 for y_k^P (data not shown).

Likelihood ratio test

As stated by Zhang and Xu (2005) and Zhang et al. (2008), it is now possible to test the null hypothesis $H_0: b_i = 0$ that there is no QTL linked to the i th marker locus by using the LR test statistic:

$$LR_i = -2[L(\theta_{-i}) - L(\theta')] \quad (11)$$

where $\theta' = \{b_0, b_{(1)}, \dots, b_{(q)}, \sigma_{-i}^2\}$ with $|b_{(k)}| > 10^{-6}$, $k = 1, 2, \dots, q$, $\theta_{-i} = \{b_0, b_{(1)}, \dots, b_{(i-1)}, b_{(i+1)}, \dots, b_{(q)}, \sigma_{-i}^2\}$ is the vector of parameters that excludes $b_{(i)}$, σ_{-i}^2 is the residual variance of the reduced model under H_0 , and $L(\theta)$ is the log-likelihood function. For simplicity, the usual QTL significance criterion ($\text{LOD}_i \geq 3$) was applied. For each QTL simulated, the samples for which LOD or $-\lg P(H_0)$ exceeded the threshold of 3.0 were counted. The ratio of the number of actual QTL to the total number of replicates (200) represents the empirical power of the method. The FPR was given by the ratio of the number of false-positive effects to the total number of zero effects. Note that linked false positives were only counted once.

Monte Carlo simulation studies

The choice of the best response variable for phenotypic difference

To demonstrate the first objective, four response variables, including the absolute (squared) trait difference $y_k^A(y_k^P)$ and the trait product (sum) $y_k^A(y_k^P)$, were compared in the first and second simulation experiments with the maize pedigree in Figure 1. In the first simulation experiment, the number of non-founders in the maize pedigree was 301. Sixty-one equally spaced markers, each with four alleles, were simulated on a single-chromosome segment of length 600 cM, and a single QTL with 0.20 heritability and four alleles was located at 200 cM. Each of the 200 simulated data sets was analyzed

four times by MLISM with each response variable, in turn (Table 2). This analysis showed that the minimum FPR is achieved using y_k^A , and the QTL detection power and the standard deviation for the estimates of QTL position were highest and lowest, respectively, using y_k^A . In the second simulation experiment, three QTL, each with two alleles, were simulated with heritabilities of 0.05, 0.10 and 0.15 and located at marker positions 100, 300 and 500 cM, respectively. The number of alleles for markers was set at 2, and other parameters were the same as those in the first simulation experiment. The results were consistent with the above trend (Table 3). This implies that the trait product is the optimal response variable in the MLISM. It should be noted that the estimate of σ^2 for the MLISM is different from the simulated value. This is because that the MLISM uses pairwise distances analysis rather than origin response variable analysis. In addition, under the situation of low heritability, the power for QTL detection is less for the HISM than for the MLISM. In the following simulation experiments, only outcomes using y_k^A are reported.

The evaluation of FPR for multiple loci ISM

To demonstrate whether the FPR for the MLISM was significantly reduced, the FPR for five simulation experiments was calculated. Each data set was analyzed by the ISM, the HISM and the MLISM, for a total of three analyses.

In the third simulation experiment, we pruned the maize pedigrees to have the correct number of non-founders of 100, 200 and 300, respectively. Eleven equally spaced markers, each with three alleles, were placed on each of two 100 cM chromosome segments; two QTL, each with three alleles, were simulated with heritabilities of 0.10 and 0.20 and located at marker positions 20 and 80 cM, respectively. The results are shown in Figure 2a (Supplementary Table S1). As expected, the FPRs in the MLISM were substantially decreased under various sample sizes as compared with those in the ISM. The power of QTL detection is less for the HISM than for the MLISM with small sample sizes. In addition, the FPR in the MLISM increased as sample size increased because the QTL detection power increases as sample size increases.

The fourth simulation experiment was designed to investigate the effects of the number of alleles (both

Table 2 Comparisons of *in silico* mapping (ISM), haplotype-based ISM (HISM) and variance component analysis (VCA) with multiple loci ISM (MLISM) under a single-quantitative trait locus (QTL) with a 0.20 heritability with a maize pedigree of 301 inbred lines (200 replicates)

Method	Response variable	Power (%)	Position (cM)	Regression coefficient or QTL variance	Intercept or polygenic variance	σ^2	FPR
	True values		200.00	—	—	5.00	
VCA		61.0	200.13 (2.54)	2.14 (1.67)	0.23 (0.41)	5.02 (0.58)	28 ^a
ISM		100.0	199.95 (1.87)	-0.39 (0.16)	2.91 (0.18)	4.30 (0.40)	19.86
HISM		99.0	200.00 (0.00)	—	—	—	2.48
MLISM	$ z_{k,1} - z_{k,2} $	97.0	199.90 (1.44)	-0.39 (0.17)	2.90 (0.19)	4.27 (0.40)	5.73
	$(z_{k,1} - z_{k,2})^2$	97.5	200.10 (1.01)	-3.15 (1.43)	13.08 (1.61)	280.80 (55.48)	5.38
	$(z_{k,1} - \bar{z})(z_{k,2} - \bar{z})$	98.0	200.00 (0.00)	1.24 (0.73)	-0.51 (0.26)	35.50 (6.58)	1.98
	$[(z_{k,1} - \bar{z}) + (z_{k,2} - \bar{z})]^2$	71.5	200.28 (6.27)	2.49 (1.77)	11.82 (1.51)	277.87 (54.95)	8.04

Abbreviation: FPR, false-positive ratio.

^aTwenty-eight false QTL are detected in 100 replicates.

Table 3 Comparisons of *in silico* mapping (ISM) and haplotype-based ISM (HISM) with multiple loci ISM (MLISM) under multiple quantitative trait locus (QTL) simulated with a maize pedigree of 301 inbred lines (200 replicates)

Method	Response variable	QTL1 ($h^2 = 0.05$)			QTL2 ($h^2 = 0.10$)			QTL3 ($h^2 = 0.15$)			Intercept	σ^2	FPR
		Power (%)	Position (cM)	Regression coefficient	Power (%)	Position (cM)	Regression coefficient	Power (%)	Position (cM)	Regression coefficient			
ISM	True value	85.0	100.00	—	96.0	300.00	—	98.5	500.00	—	NL	5.00	18.28
		99.76 (9.67)	99.76 (9.67)	—0.24 (0.12)	300.05 (5.46)	300.05 (5.46)	—0.36 (0.17)	500.10 (2.26)	500.10 (2.26)	—0.49 (0.19)	NL	NL	18.28
		64.5	100.08 (2.38)	—	91.5	300.00 (0.00)	—	98.0	500.00 (0.00)	—	—	—	2.72
		49.5	98.28 (8.09)	—0.25 (0.13)	71.5	300.00 (1.68)	—0.39 (0.15)	92.5	500.00 (0.00)	—0.50 (0.18)	3.35 (0.37)	4.77 (0.39)	4.97
HISM		52.5	99.14 (7.22)	—2.24 (1.08)	80.0	299.75 (2.74)	—3.26 (1.40)	94.0	500.05 (0.73)	—4.27 (1.62)	17.82 (3.27)	354.16 (64.80)	4.60
		72.5	99.86 (2.88)	0.70 (0.39)	91.5	300.00 (0.00)	1.12 (0.62)	97.5	500.00 (0.00)	1.63 (0.67)	—1.90 (0.69)	44.65 (7.39)	2.23
		49.5	99.59 (8.49)	1.70 (0.73)	59.5	300.17 (4.71)	2.50 (1.21)	73.5	500.00 (4.22)	3.14 (1.51)	12.84 (2.79)	352.96 (65.47)	7.39

Abbreviation: NL, not listed.

marker and QTL) on the FPR of the new method, so the number of alleles was set at 2, 3 and 4. The number of non-founders in the maize pedigree was 301, and other parameters were the same as those in the third simulation experiment. The results are given in Figure 2b (Supplementary Table S2). Compared with the FPR obtained by using the ISM, the FPR from the MLISM was substantially decreased with various conditions of allele numbers. The QTL detection power decreased as the number of alleles increased. In simulations with a small amount of alleles, the power for QTL detection is slightly less for the HISM than for the MLISM.

In the fifth simulation experiment, the effect of the allelic frequency on the FPR of the new method was assessed by letting the frequency ratio of the three alleles (for the two simulated QTL and all markers) be set as either 1:1:1 (uniform distribution) or 1:2:3 (skewed distribution). Other parameters were the same as those in the fourth simulation experiment. The results are given in Figure 2c (Supplementary Table S3). The trend for the FPR obtained by both distributions was the same. In addition, the skewed distribution decreased the statistical power.

The sixth simulation experiment needed to mimic the actual data, which did not have equally spaced markers. The corrected genetic linkage map of Zhu *et al.* (2007) was adopted, and this genome consisted of 12 chromosomes. We genotyped 103 founders and 301 non-founders from the maize pedigree for 167 markers, each with two alleles. These markers covered ~2121.15 cM of the genome with an average marker interval of 13.68 cM. Six QTL with heritabilities of 0.05, 0.05, 0.025, 0.15, 0.075 and 0.05 and locations at marker positions 0.00, 60.32, 123.52, 942.78, 1672.05 and 1723.00 cM, respectively, were simulated. The results are shown in Figure 2d (Supplementary Table S4) and the FPR for the MLISM was significantly reduced as well.

Another simulation was performed to evaluate the effect of a large genome with a high marker density on the FPR of the new method. The simulated genome consisted of 10 chromosomes, each with 101 evenly spaced markers at a 2 cM per marker interval. For 1010 markers (two alleles) along the genome, we genotyped 10 founders and 100 non-founders from a simulated pedigree. Ten QTL with heritabilities of 0.025, 0.05, 0.025, 0.05, 0.15, 0.075, 0.025, 0.10, 0.05 and 0.05 were placed at marker positions 220, 270, 320, 370, 650, 1050, 1150, 1450, 1850 and 1950 cM, respectively. The simulation was repeated 100 times. The results are shown in Figure 2e (Supplementary Table S5) and the same trend in Figure 2d was observed.

The relationship between QTL detection and LD

To demonstrate the last objective, the last simulation experiment was used to investigate the relationship between QTL detection and LD. When one QTL at a marker position was identified, markers could be observed around the identified QTL. Of these markers, we needed to analyze not only whether there was LD between the identified QTL and adjacent markers, but also whether there were false QTL residing on these LD markers. Thus, the proportion of the number of false QTL to the number of LD markers could be calculated.

Figure 2 Box plot of false-positive rate for three methods, including *in silico* mapping (ISM), haplotype-based ISM (HISM) and multiple loci ISM (MLISM). Results labeled 'A' under the x axis are from the entire genome and those labeled 'N' from the chromosomes without quantitative trait locus. (a) The number of non-founders (sample size); (b) the number of alleles for markers and QTL; (c) 1:2:3 for the three alleles, denoted by 'S'; and 1:1:1 for the three alleles, denoted by 'U'; (d) a mimic genetic linkage map; (e) a high-density genetic linkage map.

Table 4 The relationship between quantitative trait locus (QTL) detection and linkage disequilibrium (LD) for *in silico* mapping (ISM), haplotype-based ISM (HISM) and multiple loci ISM (MLISM) approaches with a simulated pedigree of 100 non-founders and a high-density genetic linkage map (100 replicates)

Chromosome	True QTL		ISM			HISM			MLISM		
	Heritability	Position (cM)	No. of false QTL	No. of LD markers	%	No. of false QTL	No. of LD markers	%	No. of false QTL	No. of LD markers	%
2	0.025	220.00	213	1098	19.40	68	1115	6.10	25	1100	2.27
	0.050	270.00	283	1041	27.19	70	755	9.27	55	1115	4.93
	0.025	320.00	216	1019	21.20	78	977	7.98	22	1108	1.99
4	0.050	370.00	248	1100	22.55	55	835	6.59	39	1117	3.49
	0.150	650.00	374	1103	33.91	164	902	18.18	105	1105	9.50
6	0.075	1050.00	279	1074	25.98	136	1093	12.44	49	1116	4.39
	0.025	1150.00	204	1051	19.41	78	976	7.99	21	1101	1.91
8	0.100	1450.00	296	1128	26.24	141	1068	13.20	57	1110	5.14
10	0.050	1850.00	219	1085	20.18	88	1016	8.66	36	1118	3.22
	0.050	1950.00	224	1025	21.85	98	1002	9.78	33	1070	3.08

The results are listed in Table 4. The results showed that the false QTL proportion for the new method was generally lower than those for the ISM and the HISM. This implies that the new method has almost controlled for the effect of strong LD markers mentioned above.

Discussion

We used a maize pedigree population of inbred lines of known ancestry as an example to demonstrate the new method of MLISM. This method can be directly applied to inbred strains of mice. When haplotype data are available, the new method can further be extended to haplotype-based MLISM (Liao *et al.*, 2004; Guo *et al.*, 2006, 2007; Liang *et al.*, 2006a,b). When single-nucleotide polymorphism molecular data are available, the new method can be used to improve the precision of QTL mapping and to select the candidate genes. When microarray gene expression marker data are available, it will be possible to elucidate the functional relationships among genes. This study indicates that the MLISM method, with the trait product as the response variable, represents an improvement on the existing methods for QTL mapping in a population of inbred lines.

This method was validated using simulated data. When applied to real data, most of the favorable properties will remain; however, some minor modifications are required. First, population structure generates spurious genotype/phenotype associations (Pritchard and Rosenberg, 1999; Yu *et al.*, 2006). As reported by Yu *et al.* (2006), false positives can be reduced by incorporating a Q matrix $\{p_{(k,1)h}\}$ obtained from the STRUCTURE software into model (2). The model appears to be

$$y_k^A = b_0 + \sum_{i=1}^p b_i x_{ki} + \sum_{h=1}^H (p_{(k,1)h} - p_{(k,2)h}) s_h + e_k \quad (12)$$

where H is the number of sub-populations; $p_{(k,1)h}$ and $p_{(k,2)h}$ are the posterior probabilities of $z_{k,1}$ and $z_{k,2}$, respectively, conditional on the h th sub-population; and s_h is the regression coefficient related to the h th sub-population. In the simulation studies, therefore, we have focused on evaluating the performance of the proposed method rather than addressing the effect of population

structure on the new method. Second, the polygenic effect is generally included in the mixed genetic models, such as Yu *et al.* (2006) and Zhang *et al.* (2005). However, there are no polygenic effects in the MLISM. The reason for this is that all QTL across the genome have been included in the present genetic model. For ISM with unrelated individuals, one would expect that a method with all markers fitted in the model would be a better choice. Thus, there would be no need to use the mixed model with a combined polygene control, which requires testing single loci one-by-one. Third, in reality, markers may not be evenly distributed along the genome. Although the MLISM does not depend on the uniformity of marker distribution, tightly linked markers may cause poor estimates of the marker effects because of a high degree of multicollinearity. Therefore, it is recommended to use only one marker from a cluster of markers.

The proposed method differs from the ISM and the HISM in several ways. First, it extends the analysis from a single QTL to a multi-QTL model, which significantly reduces the FPR. Second, the response variable used in ISM is replaced by the trait product. Third, because the parameters are estimated by PML rather than by least squares method or analysis of variance, the FPR is substantially decreased because PML associates extremely small estimates for loci that are only loosely linked to a QTL. In particular, the PML approach is able to estimate the parameters for an oversaturated genetic model (Zhang and Xu, 2005; He and Zhang, 2008). It should also be noted that this property of the PML method results in a lower power of the new method compared with the ISM, but this shortcoming can be overcome by increasing the sample size. In addition, when several markers are in strong LD, the LR test in the single-QTL analysis could fail to reject any of these markers because of strong linkage. Thus, the relationship between QTL detection and LD was analyzed in the last simulation experiment. Results in Table 4 showed that the ratio of the number of false QTL to the number of the LD markers for the new method is less than those for the ISM and the HISM. This indicates that the new method almost controls for the effect of strong LD, which is a significant advance. In the cases of small sample sizes, low QTL heritability, and a small number of alleles, the

power for QTL detection is less for the HISM than for the MLISM. Therefore, the new method is better than the ISM and its extension.

The new method differs from many QTL mapping methods proposed to date. The MLISM is designed for whole genome data sets collected from unrelated or very distantly related inbred lines, whereas other mapping methods are designed mostly for segregating populations from controlled crosses. This new method differs from genome-wide association mapping as well. The MLISM uses pairwise distances analysis, whereas genome-wide mapping directly fits the value of the trait with a linear model where the independent variables are simply the marker genotypes. Results from the Monte Carlo simulation showed that the former technique is better than the latter under the situations of both multiple alleles and small number of non-founders (data not shown), because the sample size is much more for the former than for the latter. On the other hand, the large sample size can result in slightly higher FPR, although their FPRs have been controlled under a low level.

Although the MLISM is similar to multi-QTL Haseman–Elston regression (Zhang *et al.*, 2008), it differs in several respects. The first difference concerns the IBS and the identity-by-descent in the multi-loci genetic model. Second, if the pedigree is not complete, that is, it is missing or mistaken, then the MLISM is a better choice than multi-QTL Haseman–Elston regression. Finally, the power of the detection of QTL is relatively higher for the MLISM than for multi-QTL Haseman–Elston regression (Zhang *et al.*, 2008).

In coarse QTL mapping, mapping precision is relatively low because of a combination of too few generations, limited marker density and too few recombinants. The availability of plentiful single-nucleotide polymorphism markers or haplotypes across a population of inbred lines can address this problem. Single-nucleotide polymorphism marker density can, in principle, be increased to as high a level as necessary. In addition, there are a number of advantages to using inbred lines instead of backcross and F₂ populations for QTL mapping. These include a greater allele number and a broader reference population, a much higher mapping resolution and a lesser investment in time (Buckler and Thornsberry, 2002; Flint-Garcia *et al.*, 2003; Zhang *et al.*, 2005; Yu and Buckler, 2006; Iwata *et al.*, 2007). In particular, where single-nucleotide polymorphism genotypes are available, the fine mapping of QTL in a population of inbred lines becomes possible without the earlier need to generate mapping populations.

Acknowledgements

We are grateful to the Subject Editor and two anonymous referees for their constructive comments and suggestions that significantly improved the presentation of the manuscript. This study was supported by the National Basic Research Program of China Grant 2006CB101708; the National Natural Science Foundation of China Grant 30671333; Jiangsu Natural Science Foundation Grant BK2008335, NCET Grant NCET-05-0489; 863 program Grant 2006AA10Z1E5; and the 111 Project Grant B08025 to YMZ.

References

- Buckler ES, Thornsberry JM (2002). Plant molecular diversity and application to genomics. *Curr Opin Plant Biol* 5: 107–111.
- Chesler EJ, Rodriguez SL, Mogil JS (2001). *In silico* mapping of mouse quantitative trait loci. *Science* 294: 2423.
- Darvasi A (2001). *In silico* mapping of mouse quantitative trait loci. *Science* 294: 2423.
- Drigalenko E (1998). How sib pairs reveal linkage. *Am J Hum Genet* 63: 1242–1245.
- Flint-Garcia SA, Thornsberry JM, Buckler ES (2003). Structure of linkage disequilibrium in plants. *Ann Rev Plant Biol* 54: 357–374.
- Guo Y, Lu P, Farrell E, Zhang X, Weller P, Monshouwer M *et al.* (2007). *In silico* and *in vitro* pharmacogenetic analysis in mice. *Proc Natl Acad Sci USA* 104: 17735–17740.
- Guo Y, Weller P, Farrell E, Cheung P, Fitch B, Clark D *et al.* (2006). *In silico* pharmacogenetics of warfarin metabolism. *Nat Biotech* 24: 531–536.
- Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF *et al.* (2001). *In silico* mapping of complex disease-related traits in mice. *Science* 292: 1915–1918.
- Haseman JK, Elston RC (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2: 3–19.
- He XH, Zhang Y-M (2008). Mapping epistatic QTL underlying endosperm trait loci in random hybridization design. *Heredity* 101: 39–47.
- 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.
- Jansen RC (1993). Interval mapping of multiple quantitative trait loci. *Genetics* 135: 205–211.
- Kao C-H, Zeng ZB, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* 152: 1203–1216.
- Liao G, Wang J, Guo J, Allard J, Cheng J, Ng A *et al.* (2004). *In silico* genetics: identification of a functional element regulating *H2-Ea* gene expression. *Science* 306: 690–695.
- Liang DY, Liao G, Lighthall GK, Peltz G, Clark DJ (2006a). Genetic variants of the P-glycoprotein gene *Abcb1b* modulate opioid-induced hyperalgesia, tolerance and dependence. *Pharmacogenet Genomics* 16: 825–835.
- Liang DY, Liao G, Wang J, Usuka J, Guo Y, Peltz G *et al.* (2006b). A genetic analysis of opioid-induced hyperalgesia in mice. *Anesthesiology* 104: 1054–1062.
- Park T, Casella G (2008). The Bayesian Lasso. *J Am Stat Assoc* 103: 681–686.
- Pritchard JK, Rosenberg NA (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65: 220–228.
- Wright FA (1997). The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet* 60: 740–742.
- Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* 163: 789–801.
- Xu S (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 63: 513–521.
- Xu S, Jia Z (2007). Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics* 175: 1955–1963.
- Yi N, Banerjee S (2009). Hierarchical generalized linear models for QTL mapping. *Genetics* 181: 1101–1113.
- Yi N, Xu S (2008). Bayesian Lasso for quantitative trait loci mapping. *Genetics* 179: 1045–1055.
- Yu J, Buckler ES (2006). Genetic association mapping and genome organization of maize. *Curr Opin Biotech* 17: 155–160.
- Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF *et al.* (2006). A untied mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38: 203–208.

- Zeng Z-B (1993). Theoretical basis for separation of multiple linked gene effects in mapping of quantitative trait loci. *Proc Natl Acad Sci USA* **90**: 10972–10976.
- Zhang YM (2006). Advances on methods for mapping QTL in plant. *Chin Sci Bull* **51**: 2809–2818.
- Zhang YM, Lü HY, Yao LL (2008). Multiple quantitative trait loci Haseman-Elston regression using all markers on the entire genome. *Theor Appl Genet* **117**: 683–690.
- Zhang YM, Mao YC, Xie CQ, Smith H, Luo L, Xu S (2005). Mapping QTL 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.
- Zhu C, Wang F, Wang J, Li G, Zhang HS, Zhang Y-M (2007). Reconstruction of linkage maps in the distorted segregation populations of backcross, doubled haploid and recombinant inbred lines. *Chin Sci Bull* **52**: 1648–1653.

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