Theor Appl Genet (2008) 117:683–690
DOI 10.1007/s00122-008-0809-0

ORIGINAL PAPER

Multiple quantitative trait loci Haseman–Elston regression using all markers on the entire genome

Yuan-Ming Zhang · Hai-Yan Lü · Li-Li Yao

Received: 28 September 2007 / Accepted: 17 May 2008 / Published online: 19 June 2008

Abstract The Haseman–Elston (HE) regression, developed in the 1970s, remains in common use to detect genetic linkage between a quantitative trait and a genetic marker. Although the technique has been improved in a number of ways, it predicts a high rate of false positive quantitative trait locus (QTL) because it is based on a single-QTL model. We have extended the origin HE regression to multi-QTL HE (MQHE) regression, so that all markers across the entire genome can be exploited simultaneously. The parameters have been estimated by the penalized maximum likelihood method, and several response variables for phenotypic difference have been compared in order to optimize the procedure. The method has been tested by simulation in a pedigree population of maize inbred lines of known ancestry. These simulations show that the trait product is the optimal response variable for phenotypic difference. The false positive rate produced by the MQHE regression is substantially lower than that generated by either variance component analysis or the origin HE regression. The MQHE regression, with the trait product as the response variable, represents a significant improvement on existing methods for QTL mapping in a set of inbred lines (or cultivars) of known ancestry.

Communicated by J.-L. Jannink.

Y.-M. Zhang (✉) · H.-Y. Lü · L.-L. Yao
Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, China
e-mail: soyzhang@njau.edu.cn

Introduction

Most quantitative trait loci (QTL) mapping approaches rely on segregating populations derived from controlled crosses (Lander and Botstein 1989; Jansen 1993; Zeng 1993; Li et al. 2007; Kao et al. 1999; Xu 2003; Wang et al. 2005; Zhang and Xu 2005; Xu and Jia 2007). Often, however, it can be impossible, difficult or even unethical to perform such crosses (Liu 1998). An alternative resource is represented by natural populations, such as sibling pairs and breeding pedigrees. Note that the sib-pair-based Haseman–Elston (HE) regression (Haseman and Elston 1972) is probably the oldest QTL mapping approach still in common use. The method is rather limited by its focus on sibling pairs, but its strength lies in its computational simplicity and the robustness of the regression framework (Feingold 2001). Therefore, its extension to other populations, and in particular to a breeding pedigree of crop cultivars (or inbred lines) of known pedigree (Buckler and Thornsberry 2002; Flint-Garcia et al. 2003; Zhang et al. 2005; Yu and Buckler 2006; McClurg et al. 2007) is appealing.

The HE regression is based on the single locus model which includes only one marker at a time, and makes the critical assumption that each linkage group contains a maximum of one QTL. This limitation is problematical (Zhang 2006) since, most seriously, only the effects of one putative QTL in a given map region can be included in the model, while all other QTL effects have to be ignored. As a result, similar to interval mapping, a bias in the estimates of both the size of the effect and the position of the QTL occurs whenever more than one QTL is in fact present on a given linkage group (Zeng 1994). The result of this bias is an increase in the QTL false positive rate (FPR). To deal with the multi-QTL problems, composite interval mapping (Zeng 1993; Jansen 1993; Li et al. 2007) and multiple QTL mapping (Kao et al. 1999; Xu 2003; Zhang and Xu 2005; Wang et al. 2005; Xu and Jia 2007) have been successively proposed. However, these QTL mapping approaches are typically focused on segregating populations derived from controlled crosses, rather than on natural populations. Thus, our current priority was to incorporate multi-QTL mapping into the HE regression framework in a form which allows it be applied to analyze the real dataset for crop cultivars of known pedigree.

As pointed out by Wright (1997), the squared trait difference (y_k^{D}) in the HE regression discards some useful information, and so some benefit has been seen in using the trait values of both members of a sib-pair. In effect, the squared difference (y_k^{D}) and the trait sum (y_k^{S}) together contain exactly the same information as the original two trait values, and critically, these values are independent of one another. Drigalenko (1998) developed the idea further, by estimating a regression coefficient using a simple mean of the estimates from two regressions for response variables y_k^{D} and y_k^{S}. This average estimate is equivalent to the fitting of a single regression of the trait product (y_k^{P}) on identity-by-descent (IBD). These have been exploited as the basis for the extended HE regression described in this article.

In this article, we show how the origin HE regression can be extended to the multi- QTL HE (MQHE) regression which includes all markers across the entire genome. In the current version of the MQHE regression, the parameters have been estimated by applying the penalized maximum likelihood (PML) method, and several response variables for phenotypic difference have been compared in order to optimize the procedure. The method we propose has been contrasted with variance component analysis (VCA, see Appendix A) and the original HE regression, and used to detect novel QTL present in a set of inbred lines of known ancestry.

The new method here was tested by simulation. The purposes of the simulation were: (1) to select the best response variable for phenotypic difference; (2) to test whether the MQHE regression was more efficient than the HE regression and the VCA method; (3) to investigate the effect of sample size, the number of alleles, allelic frequency and QTL heritability on the performance of the MQHE regression, respectively.

Statistical methods

Materials

The number of inbred lines within the maize pedigree described by Zhang et al. (2005) was 404 (n) (Fig 1). Of these, n_0 (=103) were base (founder) lines or land races, while n_1 (=301) non-founder lines were bred via repeated self-pollination of a hybrid between two inbred lines. Thus, each non-founder line represents a recombinant inbred line with respect to a pair of known parents. The mapping population consisted of all the non-founder lines.

Genetic model

Let the kth inbred-line pair have trait values ($z_{k,1},\ z_{k,2}$), $\bar{z}$ be the mean value of $z_{k,1}$ and $z_{k,2}$ over all sib pairs, the squared difference be $y_k^{\mathrm{D}} = (z_{k,1} - z_{k,2})^2$, and the IBD at a locus between the two inbred lines of each pair be π. Based on the original HE regression, the expected value of $y_k{}^{\mathrm{D}}$ conditional on π is described by

$$E(y_k^{\mathrm{D}}|\pi) = b_0 + b\pi \tag{1}$$

where b_0 is regression intercept, and b is regression slope (Haseman and Elston 1972). Provided that each marker locus on the genome can be linked to putative QTL, the model (1) can be extended to the MQHE regression:

$$y_k^{\mathrm{D}} = b_0 + \sum_{i=1}^{p} b_i \pi_{ik} + e_k \tag{2}$$

where b_i is the regression coefficient for the ith QTL; p the number of all markers on the entire genome; π_{ik} the IBD of the kth inbred-line pair at the ith marker locus; e_k the residual error with an assumed $N(0,\ \sigma^2)$ distribution; and $\theta = (b_0,\ b_1,\ldots,\ b_{\mathrm{p}},\ \sigma^2)$. The trait sum is given by $y_k^{\mathrm{S}} = [(z_{k,1} - \bar{z}) + (z_{k,2} - \bar{z})]^2$, the trait product by $y_k^{\mathrm{P}} = (z_{k,1} - \bar{z})(z_{k,2} - \bar{z})$, and the absolute trait difference by $y_k^{\mathrm{A}} = |z_{k,1} - z_{k,2}|$. Similarly, the regressions of $y_k{}^{\mathrm{S}}$, y_k^{P} and y_k^{A} on the IBD can be established. In what follows, response variables are denoted by y.

IBD calculation

The method of Zhang et al. (2005) was used to calculate the IBD in model (2). The method is briefly re-capitulated in Appendix B.

Parameter estimation

The PML method (Zhang and Xu 2005) was used to estimate the parameters in model (2). The method is briefly re-capitulated here. In the PML method, the penalized likelihood function is the product of likelihood function $L(\theta|\mathbf{Y},\mathbf{M})$ and penalty function $P(\theta,\ \xi)$. The former is given by

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

where $m = n_1(n_1 - 1)/2$, $\alpha_k = b_0 + \sum_{i=1}^{p} b_i \pi_{ik}$, $\mathbf{Y} = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}$, $\mathbf{M}$ represents marker information, and

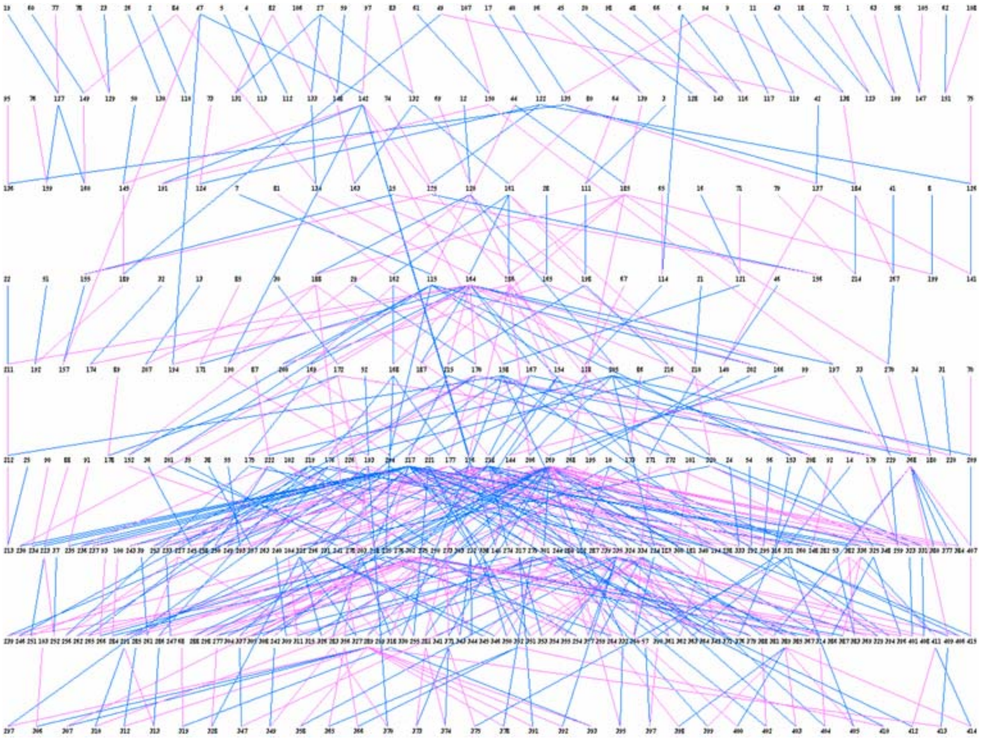

Fig. 1 The familial relationships between the 404 maize inbred lines used in the all simulation experiments and derived from Zhang et al. (2005)

$\varphi(y;\ \alpha,\ \sigma^2)$ is a normal density function with mean α and variance σ^2. Similarly, the latter function is:

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

where $\xi=\left(\mu_1,\ldots,\mu_p,\sigma_1^2,\ldots,\sigma_p^2\right)$ is the vector of hyperparameters, and $\eta>0$ is prior sample size for assessing μ_i. Note that $p(\sigma_i^2)\propto 1$ (Zhang and Xu 2005) for the response variable y_k^{P} , and $p(\sigma_i^2)\sim\mathrm{inv}-\chi^2(v,s_i^2)$ with $s_i^2=0$ $(i = 1,\ 2,\ldots,\ p)$ for the other response variables. So the penalized likelihood function is

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

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

$$b_0=\frac{1}{m}\sum_{k=1}^{m}\left(y_k-\sum_{i=1}^{p}\pi_{ik}b_i\right) \tag{6}$$

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

$$\sigma^2=\frac{1}{m}\sum_{k=1}^{m}\left(y_k-b_0-\sum_{i=1}^{p}\pi_{ik}b_i\right)^2 \tag{8}$$

$$\mu_i=b_i/(\eta+1) \tag{9}$$

$$\sigma_i^2=\begin{cases}\frac{1}{2}\left[(b_i-\mu_i)^2+\eta\mu_i^2\right], & \text{for } y^{\mathrm{P}}\\ \frac{1}{v+4}\left[(b_i-\mu_i)^2+\eta\mu_i^2+vs_i^2\right], & \text{otherwise}\end{cases} \tag{10}$$

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

The slope (b_i) depends on the genetic relationship between individuals (such as full sib, half sib, grandparent–grandchild, etc.) and the recombination fraction c (Lynch and Walsh 1998). If $c\approx 0$, the slope approximates $-2\sigma_{\mathrm{A}}^2$ for y_k^{D} so that the estimates of b_i for y_k^{D} in all the simulation experiments can be transferred into those of σ_{A}^2, where σ_{A}^2 is the additive variance associated with the chromosomal region of interest. This implies that all types of family structure can be analyzed together, as long as the marker density is sufficient (≈5 cM). For this reason, all the inbred lines depicted in Fig 1 can be analyzed together.

Likelihood ratio test

It is now possible to test the null hypothesis $H_0{:}b_i = 0$ that there is no QTL at a given location λ, by using the likelihood-ratio (LR) test statistic:

$$\mathrm{LR}_i = -2[L(\theta_{-(i+1)}) - L(\theta)] \qquad (11)$$

where $\theta_{-(i+1)} = \{b_0, b_1,\ldots, b_{i-1}, b_{i+1},\ldots, b_p, \sigma^2\}$ is the vector of parameters which excludes b_i. Once the data set was analyzed by VCA, the critical values of the test statistic used to declare statistical significance at the 5% experiment-wise type I error rate were calculated from the quick method suggested by Piepho (2001). In other cases, the conventional QTL significance criterion (LOD $\geq$ 3) was applied.

Simulation studies

We conducted five simulation experiments to evaluate the performance of the new method. In all the simulation experiments, the pedigrees used were the maize pedigree described in the section of materials. The founders were in linkage equilibrium so that their genotypes for markers and QTL could be simulated, while the number of the alleles was set at 4 except for that in the third simulation experiment and the allelic frequencies were equal except for that in the fourth simulation experiment. The genotypes of all non-founder lines could be generated from the genotypes of their corresponding parents, like the way of simulating the genotypes of recombinant inbred lines from their two parents. In the first simulation experiment, 61 equally-spaced markers were placed on a 600 cM chromosome segment, and a single QTL with a 0.20 heritability was located at 200 cM. Although only one QTL was simulated in most simulation experiments, multiple QTL were considered simultaneously in the genetic model of equation (2). The environmental variance was calculated as $\sigma_{\mathrm{e}}^2 = (1-h^2)\sigma_{\mathrm{g}}^2 \big/ h^2$. Allelic effects were calculated by relating the genetic variance of the QTL to the allelic frequencies. The phenotypic value of each line was the sum of its QTL genotypic value and the residual error, with an assumed $N(0, \sigma_{\mathrm{e}}^2)$ distribution. Each simulation run consisted of 200 replicates. For each QTL simulated, the sample for which the LOD exceeded the threshold was counted. A QTL detected within 20 cM of the simulated QTL position was considered as true. The ratio of the number of such true QTL to the total number of replicates (200) represented the empirical power. The FPR was calculated as the ratio of the number of false QTL to the total number of zero effects in the genetic model considered. Note that linked false positives were only counted once.

To demonstrate the first objective of the simulation experiments, the absolute (squared) trait difference y_k^{A} (y_k^{D}) and the trait product (sum) y_k^{P} (y_k^{S}) were compared. Each data set was analyzed four times by the MQHE regression with each response variable in turn (Table 1). The analysis showed that the choice of y_k^{P} minimized both FPR and the standard deviation for the estimates of QTL position, and almost maximized the QTL detection power. Thus, the trait product appears to be the optimal response variable in the new method.

To demonstrate the second objective of the simulation experiments, each data set in the first simulation experiment was analyzed three times, once by the HE regression, VCA and the MQHE regression (Table 1). This experiment demonstrated that the FPR achieved by the MQHE regression was substantially less than that generated by either of the two other methods, and that the standard error

Table 1 Comparison of MQHE regression with variance component analysis (VCA) and HE regression (200 replicates)

Method	Response variable	Power (%)	Position (cM)	σ_{Q}^2 or RC	σ_{p}^2 or Intercept	R_e^2	FPR (%)
True value			200.00	1.2500	–	0.8000	–
VCA	Phenotypic value	55.5	200.05 (2.35)	2.5377 (2.1250)	0.2567 (0.4601)	0.8351 (0.0054)	44[a]
HE regression	Squared difference	95.0	200.42 (6.96)	1.5892 (0.6440)	12.5491 (1.2605)	0.9960 (0.0032)	14.00
MQHE regression	Absolute difference	83.5	199.58 (4.30)	−0.5321 (0.2922)	2.8135 (0.1426)	0.9872 (0.0065)	6.13
	Squared difference	77.5	199.94 (4.19)	2.2657 (1.1960)	12.4103 (1.2323)	0.9873 (0.0066)	5.05
	Trait product	79.0	200.06 (1.78)	1.7290 (1.1229)	−0.1435 (0.0689)	0.9938 (0.0075)	1.10
	Trait sum	54.0	199.17 (7.12)	3.7089 (2.5878)	11.9978 (1.1522)	0.9872 (0.0067)	7.84

The standard deviations obtained from 200 replicates are given in parentheses

HE Haseman–Elston, *MQHE* multi-QTL HE, *RC* regression coefficient, σ_{Q}^2 and σ_{p}^2 are the variances of the QTL and the polygenes, respectively, R_e^2 the ratio of residual variance to phenotypic variance, *FPR* false positive rate

[a] The number of false QTL identified in 200 replicates. The same is true for the later tables

for QTL position was least for the MQHE regression. Thus, using the trait product in a MQHE regression represents a significant improvement on existing methods for QTL mapping in a set of inbred lines of known ancestry. In the following simulation experiments, only outcomes using y_k^{P} are reported.

In the second simulation experiment, we just pruned the maize pedigree to have the right number of non-founders of 100, 200 and 300, respectively, so the effect of sample size on the performance of the new method was evaluated. We simulated a single chromosome of 200 cM long, covered by 21 evenly spaced markers. A single QTL with a 0.20 heritability was located at 50 cM and overlapped with marker. As expected, the power and the levels of accuracy and precision increased as the number of non-founder lines increased (Table 2).

The third simulation experiment was designed to investigate the effect of the number of alleles on the performance of the new method by letting the number of the alleles (both marker and QTL) be set at 2, 4 and 6, while a single QTL was simulated with a 0.1 heritability and a 50 cM position. The simulated chromosome was the same as that in the second simulation experiment. The QTL detection power decreased with the increase of the number of alleles (Table 3).

In the fourth simulation experiment, the effect of the allelic frequency on the performance of the new method was assessed by letting the frequency ratio of the four alleles for a simulated QTL be set as 1:1:1:1 and 1:1:3:3. The simulations were performed as described in the third simulation experiment except that the number of the alleles for markers and QTL was fixed at 4, and the heritability of the QTL was fixed at 0.15. The skewed distribution decreased the statistical power (Table 4).

Finally, we implemented the new method to map multiple QTL. A 1,000 cM simulated chromosome was populated with 101 equally spaced markers. Three QTL were simulated with heritabilities of 0.05, 0.10 and 0.15

Table 2 Effect of the number of non-founder lines on the results of QTL mapping in a pedigree of inbred lines (200 replicates)

Method	Sample size	Power (%)	Position (cM)	σ_{Q}^2 or RC	Intercept	R_e^2	FPR (%)
HE regression	100	63.0	50.16 (9.76)	2.6980 (0.9527)	13.1167 (2.1007)	0.9923 (0.0041)	7.58
	200	90.0	50.22 (7.98)	2.1119 (0.9048)	12.8693 (1.6605)	0.9943 (0.0042)	14.10
	300	96.5	50.16 (6.41)	2.0241 (0.9524)	12.7462 (1.3023)	0.9947 (0.0041)	15.50
MQHE regression trait product)	100	46.0	49.78 (5.34)	2.3442 (1.1935)	−0.1907 (0.1376)	0.9935 (0.0085)	1.15
	200	73.5	49.93 (3.61)	1.6509 (1.0950)	−0.1638 (0.1041)	0.9945 (0.0071)	3.00
	300	92.5	50.00 (2.55)	1.6215 (1.1405)	−0.1542 (0.0809)	0.9937 (0.0075)	4.68

Table 3 Effect of the number of alleles on the results of QTL mapping in a pedigree of inbred lines (200 replicates)

Method	No of alleles	Power (%)	Position (cM)	σ_{Q}^2 or RC	Intercept	R_e^2	FPR (%)
HE regression	2	84.5	51.07 (8.66)	1.6141 (0.7271)	11.4057 (1.0512)	0.9972 (0.0025)	16.98
	4	79.0	50.32 (8.99)	1.1522 (0.4637)	11.4294 (0.9660)	0.9979 (0.0015)	17.70
	6	76.0	51.12 (10.39)	1.0738 (0.4651)	11.3683 (1.0141)	0.9980 (0.0017)	17.20
MQHE regression	2	85.5	49.94 (2.30)	1.0892 (0.6327)	−0.0967 (0.0417)	0.9978 (0.0023)	2.43
(trait product)	4	71.0	50.14 (3.15)	0.8429 (0.4650)	−0.0912 (0.0540)	0.9982 (0.0022)	2.35
	6	63.0	50.24 (2.96)	0.7583 (0.4660)	−0.0771 (0.0486)	0.9985 (0.0021)	1.80

Table 4 Effect of allelic distribution on the results of QTL mapping in a pedigree of inbred lines (200 replicates)

Method	Allelic distribution	Power (%)	Position (cM)	σ_{Q}^2 or RC	Intercept	R_e^2	FPR (%)
True value			50.00	1.2500		–	
HE regression	1:1:1:1	89.0	49.10 (7.97)	1.5504 (0.6568)	12.1314 (1.1311)	0.9966 (0.0026)	16.13
	1:1:3:3	77.0	49.42 (9.85)	1.2726 (0.5688)	11.4181 (1.0043)	0.9976 (0.0020)	16.63
MQHE regression (trait product)	1:1:1:1	80.0	50.06 (4.12)	1.1166 (0.8709)	−0.1141 (0.0680)	0.9966 (0.0052)	3.33
	1:1:3:3	56.5	50.18 (3.27)	1.0201 (0.6667)	−0.0776 (0.0590)	0.9979 (0.0037)	1.90

and locations at marker position 100, 300 and 500 cM, respectively. So the effect of QTL heritability on the performance of the new method was studied. The general trend was that the statistical power increased as the heritability increased (Table 5).

Discussion

We used a maize pedigree population of inbred lines (or cultivars) of known ancestry as an example to demonstrate the MQHE regression. The method can be directly applied to sibling pairs and a general pedigree. The new method is indeed multiple marker analysis that potentially assumes one QTL residing on each marker position. When marker density is too high, choosing one marker from the cluster of markers avoids a high degree of multicollinearity (Zhang and Xu 2005). When the marker is too sparse, a virtual marker may be inserted. By making use of Zhang et al.'s (2005) method, it is not difficult to calculate the values of the IBD at the virtual marker positions. Once one QTL at position λ is detected, the interval $[\lambda - d, \lambda + d]$, with d being about 5 cM, can be scanned in order to optimize the position of the QTL detected according to the idea of optimizing QTL position in the multiple interval mapping of Windows QTL Cartographer 2.5 software (Wang et al. 2007). This is an extension to QTL.

The MQHE regression differs from HE regression in several ways. First, it extends the analysis from a single- to a multi-QTL model. Second, the response variable used in the HE regression is replaced by the trait product. The outcome further confirms Drigalenko's (1998) improvement for the HE regression. Note that the absolute trait difference is also valuable. This is because that it gives a maximum power. Third, as the parameters are estimated by the PML method rather than by the least squares method, it is able to estimate the parameters in an over-saturated genetic model (Zhang and Xu 2005; He and Zhang 2008). Finally, the new method reduces the FPR at a high cost in power when the number of non-founders is small. However, it does almost at the same power level while the number is large, i.e., more than 300.

The MQHE regression differs from other methods published to date (Grupe et al. 2001; Crepieux et al. 2005; Zhang et al. 2005; Yu et al. 2006; Iwata et al. 2007; McClurg et al. 2007). Along with Iwata et al.'s (2007) method, it is based on a multi-QTL genetic model, while the others rely on a single-QTL model. Although Xu and Jia (2007) also developed an IBD-based multi-QTL method, this was focused on the analysis of a mapping population derived from controlled cross. Along with "in silico" mapping (Grupe et al. 2001), MQHE regression is based on regression analysis, while Iwata et al.'s (2007)

Table 5 Effect of multiple QTL and various mapping methods on the results of QTL mapping in a pedigree of inbred lines (200 replicates)

Method	QTL_1 ($h^2 = 0.05$)			QTL_2 ($h^2 = 0.10$)			QTL_3 ($h^2 = 0.15$)			Intercept	R_e^2	FPR (%)
	Power (%)	Position (cM)	σ_Q^2 or RC	Power (%)	Position (cM)	σ_Q^2 or RC	Power (%)	Position (cM)	σ_Q^2 or RC			
True value	–	100.00	0.3571	–	300.00	0.7143	–	500.00	1.0714	–	0.7000	
Variance component analysis	6.0	104.08 (6.39)	2.5109 (2.0846)	21.5	299.84 (4.46)	2.4885 (1.6985)	42.0	499.85 (2.67)	2.4778 (1.8071)	–	0.7839 (0.0904) 0.7555 (0.0884) 0.7581 (0.0931)	41[a]
HE regression	86.0	99.59 (12.30)	1.5450 (0.6948)	90.5	300.88 (11.02)	1.7408 (0.7420)	98.0	499.59 (8.22)	2.1192 (0.9112)	14.2352 (1.4591) 14.2449 (1.4963) 14.3668 (1.8223)	0.9976 (0.0021) 0.9969 (0.0026) 0.9956 (0.0031)	14.90
MQHE regression (trait product)	39.0	100.26 (7.48)	0.8452 (0.4530)	59.0	298.07 (5.08)	1.2291 (0.9438)	80.0	500.19 (2.09)	1.6516 (1.0429)	−0.2064 (0.1.38)	0.9899 (0.0116)	5.06

σ_Q^2 variance explained by QTL. *RC* regression coefficient

[a] The number of false QTL identified in 200 replicates

method uses a Bayesian analysis, McClurg et al. (2007) an analysis of variance, and the remainder a VCA. Moreover, the targeted populations differ. In the case of Crepieux et al.'s (2005) analysis, this is a set of F_6 lines derived from multiple crosses, while the rest consider a set of inbred lines.

A key issue for the detection of QTL in natural populations is to minimize the occurrence of false positives. In this article, several approaches have been adopted to this end. First, a multi-QTL model is used to reduce the FPR (Zhang 2006). Second, the PML method is designed to shrink the estimates towards zero by introducing a new prior on the variance of regression coefficient, so the FPR is low (Zhang and Xu 2005; He and Zhang 2008). Third, sample size $m = n_1(n_1 - 1)/2$ is so large that the size of FPR can be reduced as well. Finally, the sign of the regression coefficients becomes available to discriminate between true and false QTL, especially in the real data analysis and in the situation of large pedigree (more than 1,000 lines). To confirm the result, an accessional simulation was performed as described in the second simulation experiment except that a new pedigree randomly simulated and consisted of 1,000 inbred lines (non-founders) was fixed. The simulation showed that the FPR was 4.28% for the new method and 11.13% for the HE regression at the same QTL detection power level (100%) while the false QTL that is distinguishable from the sign of the estimates of the regression coefficients were eliminated. Therefore, the elimination of false QTL has a significant positive effect on reducing the FPR.

Is the number of markers in the new method limited? It is preferable to gather more samples or reduce the number of effects considered in the model (Zhang and Xu 2005; He and Zhang 2008). In Zhang and Xu (2005), the PML method can handle a model with a number of effects ten times larger than the sample size. Obviously, it is no problem to simultaneously include all markers across the entire genome. As for the convergence in the estimation of the parameters that most of them are zero, it has been confirmed in Zhang and Xu (2005) and He and Zhang (2008). Therefore, the new method is suitable to the genome-wide analysis and candidate gene analysis.

Acknowledgments We are grateful to the Associate Editor Dr Jannink, Jean-Luc, and two anonymous reviewers for his constructive comments and suggestions that significantly improve the presentation of the manuscript. The work was supported in part by 973 program (2006CB101708); the National Natural Science Foundation of China (30671333); NCET (NCET-05-0489); Jiangsu Natural Science Foundation (BK2005087); 863 program (2006AA10Z1E5); the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education and Personnel Ministry; and the 111 project (B08025) to YMZ.

Appendix A: Variance component analysis approach

The phenotypic values (**y**) of the inbred lines may be described by the following mixed model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wv} + \varepsilon \tag{12}$$

where $\mathbf{y} = \{y_j\}_{n_1\times 1}$; $\mathbf{u} = \{u_k\}_{n_0\times 1}$ and $\mathbf{v} = \{v_k\}_{n_0\times 1}$ are vectors for the QTL and polygenic effects of all founder lines, respectively; $\mathbf{X}$ is an incidence matrix for the fixed (non-genetic) effects, $\mathbf{b}$ is a vector of the fixed effects; and $\varepsilon = \{\varepsilon_j\}_{n_1\times 1}$ are the residual errors with an assumed $N(0, \sigma^2)$ distribution. The remaining symbols are defined as follows: $\mathbf{Z} = \{\mathbf{z}_j\}_{n_0\times 1}$, $\mathbf{z}_j$ is an incidence matrix for the QTL effects and defined as a $1 \times n_0$ vector with all but one element zero. The non-zero element is the value for unity, which occurs at the position corresponding to the founder, whose allele has been transmitted to the *j*th line. $\mathbf{W} = \{\mathbf{w}_j\}_{n_0\times 1}$, $\mathbf{w}_j$ is an incidence matrix for the polygenic effects and defined as a $1 \times n_0$ vector with the *k*th element the probability that the *k*th founder allele has been passed to the *j*th line.

A genome scan approach is taken to search for QTL linearly along the genome. To test $H_0 : \sigma_u^2 = 0$ at each putative position, we run the program twice, one to obtain the likelihood value under the full model,

$$L_1 = -\frac{1}{2}\left[\ln|\hat{\mathbf{V}}| + \ln|\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{X}| + \hat{\mathbf{r}}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\hat{\mathbf{r}} + (n-h)\ln(2\pi)\right] \tag{13}$$

where $\hat{\mathbf{V}} = \mathbf{\Pi_u}\hat{\sigma}_u^2 + \mathbf{\Pi_v}\hat{\sigma}_v^2 + \mathbf{I}\hat{\sigma}^2$; $\hat{\mathbf{r}} = y - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-\mathbf{1}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-\mathbf{1}}\mathbf{y}$; σ_u^2 and σ_v^2 are the variances of the QTL and the polygene, respectively; $\mathbf{\Pi}_u = \mathrm{E}(\mathbf{ZZ}^{\mathrm{T}})$ and $\mathbf{\Pi}_v = \mathrm{E}(\mathbf{WW}^{\mathrm{T}})$ are called the IBD and additive relationship matrices for the QTL and polygenes, respectively; and h is the rank of $\mathbf{X}$ and the other to obtain the likelihood value under the reduced model,

$$L_0 = -\frac{1}{2}\left[\ln|\hat{\mathbf{V}}_0| + \ln|X^{\mathrm{T}}\hat{\mathbf{V}}_0^{-1}X| + \hat{\mathbf{r}}_0^{\mathrm{T}}\hat{\mathbf{V}}_0^{-1}\hat{\mathbf{r}}_0 + (n-p)\ln(2\pi)\right] \tag{14}$$

where $\hat{\mathbf{V}}_0 = \mathbf{\Pi_v}\hat{\sigma}_v^2 + \mathbf{I}\hat{\sigma}^2$ and $\hat{\mathbf{r}}_0 = \mathbf{y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{y}$. The likelihood ratio test statistic is defined as

$$\lambda = -2(L_0 - L_1), \tag{15}$$

and is subsequently compared to a critical value for declaration of statistical significance.

Appendix B: IBD matrix of QTL and additive relationship matrix

The notations here are same as those in Appendix A. Let m and f be the male and female lines from which line j is derived, and let l_{m}, l_{f}, and l_j be the labels for the two parents and their recombinant inbred line j for $j = 1,\ldots, n$. If j is one of founders, say the kth founder, then $l_j = k$ for $k = 1,\ldots, n_0$. If j is not a founder, the parental lines of j must be known. Thus l_j takes l_m or l_f but not both. The recurrent relationship can be described by

$$l_j = z_j l_\mathrm{m} + (1 - z_j) l_\mathrm{f} \tag{16}$$

where z_j is an indicator variable defined as

$$z_j = \begin{cases} 1 \text{ if } j \text{ carries the allele from the male parent;} \\ 0 \text{ if } j \text{ carries the allele from the female parent.} \end{cases} \tag{17}$$

The value of z_j can be sampled from a Bernoulli distribution with probability $p(z_j = 1|\mathbf{M})$. With marker information ($\mathbf{M}$), the $p(z_j = 1|\mathbf{M})$ can be calculated using multi-point method. These sampled labels are used to reconstruct the $\mathbf{Z}$ matrix and thus the IBD matrix. The expected IBD matrix is then approximated by repeated simulations using

$$\mathbf{\Pi}_u \approx N^{-1} \sum_{i=1}^{N} \mathbf{Z}^{(i)} \mathbf{Z}^{(i)\mathrm{T}} \tag{18}$$

where N is the total number of repeated simulations and $\mathbf{Z}^{(i)}$ is the simulated Z matrix in the ith replicate.

The additive relationship matrix for polygene ($\mathbf{\Pi}_v$) is obtained similarly, under the situation of $p(z_j = 1) = p(z_j = 0) = 0.5$, except that the simulation does not depend on marker information.

References

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and application to genomics. Curr Opin Plant Biol 5:107–111

Crepieux S, Lebreton C, Flament P, Charmet G (2005) Application of a new IBD-based QTL mapping method to common wheat breed population: analysis of kernel hardness and dough strength. TAG 111:1409–1419

Drigalenko E (1998) How sib pairs reveal linkage. Am J Hum Genet 63:1242–1245

Feingold E (2001) Methods for linkage analysis of quantitative trait loci in human. Theor Popul Biol 60:167–180

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Grupe A, Germer S, Usuka J 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 YM (2008) Mapping epistatic quantitative trait loci underlying endosperm traits using all markers on the entire genome 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 Z-B, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Li HH, Ye GY, Wang JK (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374

Liu BH (1998) Statistical genomics: linkage, mapping, and QTL analysis. CRC Press, New York

McClurg P, Janes J, Wu C, Delano DL, Walker JR, Batalov S, Takahashi JS, Shimomura K, Kohsaka A, Bass J, Wiltshire T, Su AI (2007) Genome-wide association analysis in diverse inbred mice: power and population structure. Genetics 176:675–683

Piepho HP (2001) A quick method for computing approximately thresholds for quantitative trait loci detection. Genetics 157:425–432

Wang H, Zhang YM, Li XM et al (2005) Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 170:465–480

Wang S, Basten CJ, Zeng Z-B (2007) Windows QTL Cartographer 2.5. Department of Statistics. North Carolina State University, Raleigh, NC

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, Jia Z (2007) Genome-wide analysis of epistatic effects for quantitative traits in Barley. Genetics 175:1955–1963

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160

Yu JM, Pressoir G, Briggs WH et al (2006) A unified 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. PNAS 90:10972–10976

Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468

Zhang Y-M (2006) Advances on methods for mapping QTL in plant. Chin Sci Bull 51:2809–2818

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

Zhang Y-M, Mao YC, Xie CQ et al (2005) Mapping QTL using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). Genetics 169:2267–2275