ORIGINAL ARTICLE

Mixed linear model approach for mapping quantitative trait loci underlying crop seed traits

T Qi¹, B Jiang¹, Z Zhu¹, C Wei¹, Y Gao², S Zhu¹, H Xu¹ and X Lou³

The crop seed is a complex organ that may be composed of the diploid embryo, the triploid endosperm and the diploid maternal tissues. According to the genetic features of seed characters, two genetic models for mapping quantitative trait loci (QTLs) of crop seed traits are proposed, with inclusion of maternal effects, embryo or endosperm effects of QTL, environmental effects and QTL-by-environment (QE) interactions. The mapping population can be generated either from double back-cross of immortalized F_2 (IF_2) to the two parents, from random-cross of IF_2 or from selfing of IF_2 population. Candidate marker intervals potentially harboring QTLs are first selected through one-dimensional scanning across the whole genome. The selected candidate marker intervals are then included in the model as cofactors to control background genetic effects on the putative QTL(s). Finally, a QTL full model is constructed and model selection is conducted to eliminate false positive QTLs. The genetic main effects of QTLs, QE interaction effects and the corresponding P -values are computed by Markov chain Monte Carlo algorithm for Gaussian mixed linear model via Gibbs sampling. Monte Carlo simulations were performed to investigate the reliability and efficiency of the proposed method. The simulation results showed that the proposed method had higher power to accurately detect simulated QTLs and properly estimated effect of these QTLs. To demonstrate the usefulness, the proposed method was used to identify the QTLs underlying fiber percentage in an upland cotton IF_2 population. A computer software, QTLNetwork-Seed, was developed for QTL analysis of seed traits.

Heredity (2014) **113**, 224–232; doi:10.1038/hdy.2014.17; published online 12 March 2014

INTRODUCTION

Given the importance of cereal grain seeds as the staple food and nutrition resources for humans and animals, and raw materials for food industry, understanding the genetic architecture underlying the development of crop seed traits becomes increasingly demanding in crop breeding program (Benner *et al.*, 1989; Mazur *et al.*, 1999; van der Meer *et al.*, 2001). Seed development starts at double fertilization of the embryo sac in which two sperm cells fuse with a female gametophyte, the egg and central cell, respectively, giving rise to the diploid (2n) embryo and the triploid (3n) endosperm. One sperm cell fuses with the egg cell to produce a zygote and then divide asymmetrically to form embryo and its suspensor, respectively. The other sperm cell merges with the central cell to form a triploid endosperm nucleus that possibly develops into endosperm in most flowering plants such as rice and wheat through two steps: a coenocytic stage followed by a cellularized and differentiated stage (Olsen, 1998; Sundaresan, 2005). Although the diploid embryo and the triploid endosperm are enclosed in the maternal integument that is a part of seed, development of seed depends on the allocation of nutrient and other physiological active substances from the mother plant. Proper development of a seed trait requires coordinated communications among the involved components, embryo, endosperm and maternal plant. Thus, in order to thoroughly understand how seed traits are genetically

determined, it is provident to consider the integrity of these components in the genetic models (Bazzaz, 2001).

It is well evidenced that additive and/or epistatic maternal effects, embryo effects, endosperm effects and gene–environment interaction effects are involved in the genetic development and evolution of crop seed (Zhang *et al.*, 1999; Shi *et al.*, 2000; Cui *et al.*, 2004; Cui and Wu, 2005a,b; Cui *et al.*, 2006). A considerable body of genetic models and statistical methods have been developed and applied to the analysis of real data. On the basis of genetic features of seed characters, Mo (1995) advocated a statistical genetic model that focuses on partitioning the phenotypic variance of the endosperm traits into various genetic and environmental factors. Zhu and Weir (1994) further proposed the mixed linear model approach to analyze the maternal, embryo, endosperm and cytoplasm effects, and their interactions with environment underlying seed traits collected from a diallel cross experiment. The above approaches can only dissect the genetic variation of seed traits into several cumulative components, but fail to provide detailed information at the individual gene level such as the positions and effects of quantitative trait loci (QTLs) as all the genes controlling the seed traits were analyzed as a whole. Advancements in molecular markers and statistical methods enable us to partition the total genetic variation into the effects of individual QTLs, based on co-segregation between molecular markers and the putative QTL (Lander and Botstein, 1989; Zeng, 1994). A number of

¹Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, PR China; ²Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, PR China and ³Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

Correspondence: Dr H Xu, Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, PR China.

E-mail: hmxu@zju.edu.cn or Dr X Lou, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA.

E-mail: xylou@uab.edu

Received 22 October 2013; accepted 25 November 2013; published online 12 March 2014

researchers applied traditional diploid genetic model to map QTLs underlying endosperm traits (Sourdille *et al.*, 1996; Parker *et al.*, 1998; Araki *et al.*, 1999; Sene *et al.*, 2001; Tan *et al.*, 2001), until tailored QTL mapping methods were developed for seed traits. Genetically, endosperm represents the next generation developed in maternal plants, with a more complex segregation patterns and genetic mechanism than diploid tissues. For example, a biallelic locus (Q-q) has four possible genotypes, namely, QQQ, QQq, Qqq and qqq, and three kinds of endosperm genetic effects can be involved, namely additive effect (a), the first dominance effect (d_1) and the second dominance effect (d_2). Based on such genetic basis, Kao (2004) proposed a statistical method using multiple-interval mapping that took triploid nature of endosperm into account. It can estimate all three kinds of endosperm effect. However, one more important feature ignored in this method and others is that endosperm is a tissue developed in its maternal plant, implying that its phenotypes may be affected not only by its own triploid endosperm genotypes, but also by the diploid maternal genotypes via supplies of nutrition and other physiological active substances. This point of view has been supported by several studies on model species such as *Arabidopsis* and maize (Letchworth and Lambert, 1998; Garcia *et al.*, 2005; Ohto *et al.*, 2005). To take maternal effects into account, it is theoretically desirable to integrate both maternal genome and offspring genome into one model. Hu and Xu (2005) proposed a mapping method to characterize genetic effects of maternal genome on offspring traits that incorporates both the quantitative genetic model for diploid maternal traits and that for triploid endosperm traits into a unified QTL mapping framework. Wen and Wu (2007) further proposed a method of interval mapping of endosperm traits based on a two-stage hierarchical design that considers both the endosperm effects and the maternal effects. In addition, Cui and Wu (2005a) and Cui *et al.* (2006) proposed an epistasis model for mapping endosperm QTLs in a double backcross population. The above approaches do not include genetic effects of all QTLs in the model and can only give the information about individual or one-paired QTLs, such as position, the maternal additive/dominance effect, the endosperm additive effect and the first and the second dominance effect. Therefore, these methods still possibly suffer from limited power and precision in crop seed QTL mapping, because only one or one-paired putative QTLs are tested at a time, and the genetic background control has not been considered. If there is more than one QTL on a chromosome, the testing of one QTL at a position will be distorted by all other QTLs and thus the positions and effects of the QTL are likely to be biased. Moreover, QTL-by-environment (QE) interactions, another important factor for complex traits, are ignored in these methods, potentially leading to a biased estimation of QTL parameters.

In the present study, we propose a new method for systematically mapping QTLs underlying crop seed traits. We integrate maternal and offspring genome as well as the QE interaction effects into one unified QTL model. The Monte Carlo simulations were conducted to investigate the effectiveness, the stability and the efficiency of our model and methods. A case application from a cotton breeding experiment was demonstrated for usefulness of the proposed method. A computer software, QTLNetwork-Seed (<http://ibi.zju.edu.cn/software/>), was developed for mapping QTL of seed traits.

MATERIALS AND METHODS

Genetic model

For a specific mapping population derived from two pure lines (P_1 and P_2), the genetic experiment is conducted in t different environments each with

b blocks. Suppose the surveyed quantitative trait of seed is controlled by the genetic effects of maternal and embryo of s QTLs ($Q_1, Q_2, \dots, Q_s$), the phenotype (y_{hij}) of the i -th strain in the j -th block of the h -th environment can be expressed by the following QTL model,

$$y_{hij} = \mu + \sum_{k=1}^s (x_{ik}^{Am} a_k^m + x_{ik}^{Dm} d_k^m + x_{ik}^{Ao} a_k^o + x_{ik}^{Do} d_k^o) + e_h + \sum_{k=1}^s (x_{ik}^{Am} ae_{kh}^m + x_{ik}^{Dm} de_{kh}^m + x_{ik}^{Ao} ae_{kh}^o + x_{ik}^{Do} de_{kh}^o) + B_{j(h)} + \varepsilon_{hij} \quad (1)$$

where μ is the population mean; a_k^m and d_k^m are the maternal additive and dominance effects of Q_k , with coefficients x_{ik}^{Am} and x_{ik}^{Dm} , respectively; a_k^o and d_k^o are the embryo additive and dominance effects of Q_k , with coefficients x_{ik}^{Ao} and x_{ik}^{Do} , respectively; e_h is the random effect of the h -th environment, $e_h \sim (0, \sigma_e^2)$; ae_{kh}^m and de_{kh}^m are the interaction effects of the additive and the dominance of maternal QTL with environments, respectively, $ae_{kh}^m \sim (0, \sigma_{AmE}^2)$ and $de_{kh}^m \sim (0, \sigma_{DmE}^2)$; ae_{kh}^o and de_{kh}^o are the interaction effects of the additive and the dominance of embryo QTL with environments, respectively, $ae_{kh}^o \sim (0, \sigma_{AoE}^2)$ and $de_{kh}^o \sim (0, \sigma_{DoE}^2)$; $B_{j(h)}$ is the j -th block effect within the h -th environment; and ε_{hij} is the random residual effect, $\varepsilon_{hij} \sim (0, \sigma_e^2)$.

If the trait is controlled simultaneously by the genetic effects of maternal and endosperm QTLs, the model can be represented as follows,

$$y_{hij} = \mu + \sum_{k=1}^s (x_{ik}^{Am} a_k^m + x_{ik}^{Dm} d_k^m + x_{ik}^{Ae} a_k^e + x_{ik}^{De_1} d_{e_1k}^e + x_{ik}^{De_2} d_{e_2k}^e) + e_h + \sum_{k=1}^s (x_{ik}^{Am} ae_{kh}^m + x_{ik}^{Dm} de_{kh}^m + x_{ik}^{Ae} ae_{kh}^e + x_{ik}^{De_1} de_{1kh}^e + x_{ik}^{De_2} de_{2kh}^e) + B_{j(h)} + \varepsilon_{hij} \quad (2)$$

where a_k^e , $d_{e_1k}^e$ and $d_{e_2k}^e$ are additive effect, first-order dominance effect (interaction between Q_k and q_k) and second-order dominance effect (between Q_k and $q_k q_k$) of endosperm QTL, respectively, with coefficients x_{ik}^{Ae} , $x_{ik}^{De_1}$ and $x_{ik}^{De_2}$; ae_{kh}^e , de_{1kh}^e and de_{2kh}^e are the interaction effects of the additive, the first- and the second-order dominance effects with environments, respectively, $ae_{kh}^e \sim (0, \sigma_{AeE}^2)$, $de_{1kh}^e \sim (0, \sigma_{De_1E}^2)$, and $de_{2kh}^e \sim (0, \sigma_{De_2E}^2)$; the other terms have the same definitions as those in model (1). It should be pointed out that when a QTL is involved in both maternal and endosperm effects, $d_{e_1k}^e$ and $d_{e_2k}^e$ cannot be distinguished in this full model, and thus we need to reparameterize them with the constraint of $d_{e_1k}^e$ equal to $d_{e_2k}^e$, as well as their interaction effects with environment with the constraint of de_{1kh}^e equal to de_{2kh}^e .

In the above models, all coefficients of QTLs will be determined by the QTL genotype. If QTL genotypes are QQ, Qq and qq, the coefficients are 1, 0 and -1, respectively, for maternal or embryo additive effects, and 0, 0.5 and 0 for maternal or embryo dominance effects. The coefficients of endosperm additive, the first and the second dominance effects are 1.5, 0 and 0 for the endosperm QTL genotype QQQ, 0.5, 1 and 0 for the QQq, -0.5, 0 and 1 for the Qqq, -1.5, 0 and 0 for the qqq, respectively. In practice, the QTL genotype is usually unknown; their probabilities can be inferred based on the flanking marker genotype.

Mating designs

Several mapping populations are applicable to dissect various effect components in seed traits. We suggest to use the populations derived from an immortalized F_2 (IF_2) of which genetically identical individuals can be regenerated for replicates and be phenotyped in different environments for detection of gene-environment interactions.

Double back-cross of IF_2 . In this design, IF_2 lines, as maternal parent, are crossed with their two homozygous parents (P_1 and P_2) as paternal parents. Marker information of IF_2 and its offspring as well as the phenotype of offspring are employed to perform QTL mapping. There are eight types of QTL genotype combinations in maternal and in embryo (endosperm) and its offspring plants: QQ:QQ(QQQ), Qq:QQ(QQQ), Qq:Qq(Qqq), qq:Qq(Qqq), QQ:Qq(QQq), Qq:Qq(QQq), Qq:qq(qqq) and qq:qq(qqq); the first four and the last four combinations derive from back-cross of IF_2 line with P_1 , and with P_2 ,

respectively. We denote $p_{k|ij}$ as the conditional probability of the k -th QTL genotype combination on the i -th marker type in maternal plant and the j -th marker type in offspring generation and the expectations of QTL effect coefficients can be calculated based on the formula in Table 1, where k indexes the maternal and offspring QTL genotype combination ($k = 1, 2, \dots, 8$), i indexes the maternal genotype at the flanking markers ($i = 1, 2, \dots, 9$) and j indexes the offspring genotype at the flanking markers ($j = 1, 2, \dots, 9$). The conditional probabilities of eight QTL genotype combinations, $p_{k|ij}$, are summarized in Table 2 for the IF₂ from random mating of double haploid (DH) lines. If the IF₂ is from random mating of recombinant inbred lines (RILs), instead of the recombination rates, the proportions of recombinant zygote should be used in Table 2. As shown in Table 2, $p_{1|ij} + p_{2|ij} + p_{3|ij} + p_{4|ij} = 1$ and $p_{5|ij} = p_{6|ij} = p_{7|ij} = p_{8|ij} = 0$ in BC₁ (IF₂ × P₁) lines and $p_{5|ij} + p_{6|ij} + p_{7|ij} + p_{8|ij} = 1$ and $p_{1|ij} = p_{2|ij} = p_{3|ij} = p_{4|ij} = 0$ in BC₂ (IF₂ × P₂) lines.

Selfing of IF₂. In this design, we use marker information of two generations, parental generation (IF₂) and its offspring (F₃), and the phenotype of seed (F₃) to perform QTL mapping. There are totally six types of combination for QTL genotypes in maternal genome (IF₂), embryo genome and endosperm genome (F₃): QQ:QQ(QQQ), Qq:QQ(QQQ), Qq:Qq(QQq), Qq:Qq(Qq), Qq:qq(qqq) and qq:qq(qqq). According to the formula for calculating conditional probabilities by Wen (Wen et al., 2007), we denote $p_{k|ij}$ as the conditional probability of QTL genotype combination on the two flanking marker genotypes in IF₂ and F₃ plants, in which i stands for the index of marker genotype in IF₂ plant, j stands for the index of marker genotype in F₃ plant ($i, j = 1, 2, \dots, 9$) and k stands for the index of QTL genotype combination ($k = 1, 2, 3, 6, 7, 8$); then, the expectations of coefficients for QTL effects in models (1) and (2) could be obtained by using Table 1.

Random mating of IF₂. In QTL mapping for seed traits, random mating design is also a common choice. For the random mating of IF₂ lines, we could also apply mixed linear model approach to analyze the QTL effect, in which those coefficients are more complicated than those in the above systematic or planned mating designs. Suppose a cross between two lines, sampled at random from the IF₂ with the i -th marker genotype in maternal genome and the j -th marker genotype in paternal genome for one marker interval (Table 3; $i, j = 1, 2, \dots, 9$), respectively, we denote $p_{k|i}$ as the conditional probability of the k -th QTL genotype ($k = 1, 2, 3$) on the i -th marker genotype of the maternal genome and $p_{l|j}$ as the conditional probability of the l -th QTL genotype ($l = 1, 2, 3$) on the j -th marker genotype of the paternal genome. Let $p_{k|i, l|j} = p_{k|i} p_{l|j}$, then the expectations of the coefficients of QTL effects could be calculated as follows: ($p_{1|i} - p_{3|i}$) is the expectation of coefficient for maternal additive effect; $p_{2|i}$ for the maternal dominance effect; ($p_{1|i, 1|j} + \frac{1}{2}p_{1|i, 2|j} + \frac{1}{2}p_{2|i, 1|j} - \frac{1}{2}p_{2|i, 3|j} - \frac{1}{2}p_{3|i, 2|j} - p_{3|i, 3|j}$) for the embryo additive effect; ($\frac{1}{2}p_{1|i, 2|j} + p_{1|i, 3|j} + \frac{1}{2}p_{2|i, 1|j} + \frac{1}{2}p_{2|i, 2|j} + \frac{1}{2}p_{2|i, 3|j} + p_{3|i, 1|j} + \frac{1}{2}p_{3|i, 2|j}$) for the embryo dominance effect; ($\frac{2}{3}p_{1|i, 1|j} + p_{1|i, 2|j} + \frac{1}{3}p_{1|i, 3|j} + \frac{2}{3}p_{2|i, 1|j} - \frac{1}{2}p_{2|i, 3|j} - \frac{1}{2}p_{3|i, 1|j} - p_{3|i, 2|j} - \frac{2}{3}p_{3|i, 3|j}$) for the endosperm additive effect; ($\frac{1}{2}p_{1|i, 2|j} + p_{1|i, 3|j} + \frac{1}{4}p_{2|i, 2|j} + \frac{1}{2}p_{2|i, 3|j}$) for the endosperm first-order dominance effect; ($\frac{1}{2}p_{2|i, 1|j} + \frac{1}{4}p_{2|i, 2|j} + p_{3|i, 1|j} + \frac{1}{2}p_{3|i, 2|j}$) for the second-order dominance effect.

Mapping strategy

In QTL mapping, the QTL genotype at a given locus is not observed; we could not determine the exact coefficients for the QTL effects in the above models.

One reasonable method is to substitute them with their expectations based on the flanking marker genotypes. On the other hand, we still need to know the number of QTLs associated with the seed trait of study and their positions on the chromosome(s). Therefore, we first conduct QTL position scanning across the whole genome using a walking step of 1 cM with prescreened significant marker intervals as cofactors, and then construct the QTL full model to analyze QTL effects. The procedure is outlined as follows, and for more details please refer to the literature of mapping QTL of complex traits (Yang et al., 2007).

QTL position scanning

Similar to the composite interval mapping (Zeng, 1994), we perform scanning on the QTL position using marker information. First, we treat each interval consisting of paired adjacent makers as an integrated effect in the regression model and conduct regression analysis of phenotype on each marker interval. F-statistic was calculated and used to screen potential candidate marker intervals by the permutation method. Based on the screened potential candidate intervals, multivariate regression was performed again and significant candidate marker intervals harboring QTLs were finally screened via the conventional stepwise method (Yang et al., 2007). Second, with the selected candidate intervals as cofactors, QTL position scanning is performed with a walking step such as 1 cM across the whole genome. When one putative QTL is tested, all other predetected significant QTLs will be included in the model to control genetic noise from background QTLs. Significant threshold is specified by permutation testing. Without loss of generality, we use the model (2) to illustrate the analysis. The following model is used to identify the significant marker intervals on the whole genome,

$$y_{hij} = \mu_h + \zeta_{it}^{+Am} a_{th}^{+m} + \zeta_{it}^{+Dm} d_{th}^{+m} + \zeta_{it}^{+Ae} a_{th}^{+e} + \zeta_{it}^{+De1} d_{1th}^{+e} + \zeta_{it}^{+De2} d_{2th}^{+e} + \zeta_{it}^{-Am} a_{th}^{-m} + \zeta_{it}^{-Dm} d_{th}^{-m} + \zeta_{it}^{-Ae} a_{th}^{-e} + \zeta_{it}^{-De1} d_{1th}^{-e} + \zeta_{it}^{-De2} d_{2th}^{-e} + B_{j(h)} + e_{hij} \quad (3)$$

where y_{hij} is the phenotype of the i -th line in the j -th block under the h -th environment; μ_h is the population mean under the h -th environment; t ($t = 1, \dots, T$) stands for the t -th marker interval to be tested in T intervals; a_{th}^{+m} (a_{th}^{-m}) and d_{th}^{+m} (d_{th}^{-m}) stand for the maternal additive and dominance effects for the right marker (left marker) in h -th environment, with coefficients ζ_{it}^{+Am} (ζ_{it}^{-Am}) and ζ_{it}^{+Dm} (ζ_{it}^{-Dm}) respectively; a_{th}^{+e} (a_{th}^{-e}), d_{1th}^{+e} (d_{1th}^{-e}) and d_{2th}^{+e} (d_{2th}^{-e}) stand for the endosperm additive effect, first-order dominance effect and second-order dominance effect for the right marker (left marker) in the h -th environment with coefficients ζ_{it}^{+Ae} (ζ_{it}^{-Ae}), ζ_{it}^{+De1} (ζ_{it}^{-De1}) and ζ_{it}^{+De2} (ζ_{it}^{-De2}) respectively; the other parameters have the same definition as model (2). These ζ s corresponding to different marker effects could be determined by the same way for the coefficients of QTL effects after replacing Q and q alleles in QTL genotype with M and m marker alleles. F-statistic based on Henderson III method (Henderson, 1953; Searle, 1968) is calculated to test the significance of marker interval effects, and permutation method is employed to specify the threshold of significance.

Within each marker interval, we scan QTL with a step size of 1 cM along the chromosome. The selected marker intervals are added into the mapping model to control the background genetic effects when scanning QTL position. To test QTL k with all s significant markers considered, we apply the

Table 1 Coefficients of QTL effects based on the conditional probabilities of QTL combinations on two flanking marker genotype for selfing and double back-cross populations of IF₂

Design	x^{Am}	x^{Dm}	x^{Ao}	x^{Do}	x^{Ae}	x^{De1}	x^{De2}
Selfing of IF ₂	$p_{1 ij} - p_{8 ij}$	$1 - p_{1 ij} - p_{8 ij}$	$p_{1 ij} + p_{2 ij} - p_{7 ij} - p_{8 ij}$	$p_{3 ij} + p_{6 ij}$	$3/2(p_{1 ij} + p_{2 ij}) + 1/2(p_{6 ij} - p_{3 ij}) - 3/2(p_{7 ij} + p_{8 ij})$	$p_{6 ij}$	$p_{3 ij}$
IF ₂ × P ₁	$p_{1 ij} - p_{4 ij}$	$p_{2 ij} + p_{3 ij}$	$p_{1 ij} + p_{2 ij}$	$p_{3 ij} + p_{4 ij}$	$3/2(p_{1 ij} + p_{2 ij}) - 1/2(p_{3 ij} + p_{4 ij})$	0	$p_{3 ij} + p_{4 ij}$
IF ₂ × P ₂	$p_{5 ij} - p_{8 ij}$	$p_{6 ij} + p_{7 ij}$	$-p_{7 ij} - p_{8 ij}$	$p_{5 ij} + p_{6 ij}$	$1/2(p_{5 ij} + p_{6 ij}) - 3/2(p_{7 ij} + p_{8 ij})$	$p_{5 ij} + p_{6 ij}$	0

Abbreviations: IF₂, immortalized F₂; QTL, quantitative trait locus.

Note that $p_{k|ij}$ is the conditional probability of the k -th QTL genotype combination on the i -th two-locus marker genotype in maternal plant and the j -th two-locus marker genotype in offspring plant. $k = 1, 2, \dots, 8$ for QQ:QQ(QQQ), Qq:QQ(QQQ), Qq:Qq(QQq), qq:Qq(Qq), Qq:Qq(QQq), Qq:Qq(QQq), Qq:qq(qqq) and qq:qq(qqq), respectively.

Table 2 Conditional probabilities of QTL genotype combination on the genotypes of flanking markers in both IF₂ and its double-cross offspring plants

IF ₂	BC ₁	QQ:QQ(QQQ)	Qq:QQ(QQQ)	Qq:Qq(Qqq)	qq:Qq(Qqq)
M ₁ M ₁ M ₂ M ₂	M ₁ M ₁ M ₂ M ₂	$s_1^2 s_2^2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$r_1^2 r_2^2 / s^2$
M ₁ M ₁ M ₂ m ₂	M ₁ M ₁ M ₂ M ₂	$s_1^2 s_2 r_2 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$r_1^2 r_2 s_2 / rs$
M ₁ M ₁ M ₂ m ₂	M ₁ M ₁ M ₂ m ₂	$s_1^2 s_2 r_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 t_2 / rs$	$r_1^2 r_2 s_2 / rs$
M ₁ M ₁ m ₂ M ₂	M ₁ M ₁ M ₂ m ₂	$s_1^2 r_2^2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$r_1^2 s_2^2 / r^2$
M ₁ m ₁ M ₂ M ₂	M ₁ M ₁ M ₂ M ₂	$s_1 r_1 s_2^2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 r_2^2 / rs$
M ₁ m ₁ M ₂ M ₂	M ₁ m ₁ M ₂ M ₂	$s_1 r_1 s_2^2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 r_1 r_2^2 / rs$
M ₁ m ₁ M ₂ m ₂	M ₁ M ₁ M ₂ M ₂	$s_1 s_2 r_1 r_2 / t$	$t_1 t_2 / t$	$s_1 s_2 r_1 r_2 / t$	$s_1 s_2 r_1 r_2 / t$
M ₁ m ₁ M ₂ m ₂	M ₁ M ₁ M ₂ m ₂	$s_1 s_2 r_1 r_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$
M ₁ m ₁ M ₂ m ₂	M ₁ m ₁ M ₂ M ₂	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 t_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 s_2 r_1 r_2 / rs$
M ₁ m ₁ m ₂ M ₂	M ₁ m ₁ M ₂ m ₂	$s_1 r_1 r_2^2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 s_2^2 / rs$
M ₁ m ₁ m ₂ m ₂	M ₁ m ₁ M ₂ m ₂	$s_1 r_1 r_2^2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 r_1 s_2^2 / rs$
m ₁ M ₁ M ₂ M ₂	M ₁ m ₁ M ₂ M ₂	$r_1^2 s_2^2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$s_1^2 r_2^2 / r^2$
m ₁ M ₁ M ₂ m ₂	M ₁ m ₁ M ₂ M ₂	$r_1^2 r_2 s_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 t_2 / rs$	$s_1^2 s_2 r_2 / rs$
m ₁ M ₁ M ₂ m ₂	M ₁ m ₁ M ₂ M ₂	$r_1^2 r_2 s_2 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1^2 s_2 r_2 / rs$
m ₁ m ₁ M ₂ M ₂	M ₁ m ₁ M ₂ M ₂	$r_1^2 r_2^2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$s_1^2 s_2^2 / s^2$
IF ₂	BC ₂	QQ:Qq(QQq)	Qq:Qq(QQq)	Qq:qq(Qqq)	qq:qq(Qqq)
M ₁ M ₁ M ₂ M ₂	M ₁ m ₁ M ₂ m ₂	$s_1^2 s_2^2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$r_1^2 r_2^2 / s^2$
M ₁ M ₁ M ₂ m ₂	M ₁ m ₁ M ₂ m ₂	$s_1^2 s_2 r_2 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$r_1^2 r_2 s_2 / rs$
M ₁ M ₁ M ₂ m ₂	M ₁ m ₁ M ₂ m ₂	$s_1^2 s_2 r_2 / rs$	$s_1 s_2 r_1 r_2 u_2 / rs$	$s_1 r_1 t_2 / rs$	$r_1^2 r_2 s_2 / rs$
M ₁ M ₁ m ₂ M ₂	M ₁ m ₁ m ₂ m ₂	$s_1^2 r_2^2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$r_1^2 s_2^2 / r^2$
M ₁ m ₁ M ₂ M ₂	M ₁ m ₁ M ₂ M ₂	$s_1 r_1 s_2^2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 r_2^2 / rs$
M ₁ m ₁ M ₂ M ₂	m ₁ M ₁ M ₂ m ₂	$s_1 r_1 s_2^2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 r_1 r_2^2 / rs$
M ₁ m ₁ M ₂ m ₂	M ₁ m ₁ M ₂ m ₂	$s_1 s_2 r_1 r_2 / t$	$t_1 t_2 / t$	$s_1 s_2 r_1 r_2 / t$	$s_1 s_2 r_1 r_2 / t$
M ₁ m ₁ M ₂ m ₂	M ₁ m ₁ m ₂ m ₂	$s_1 s_2 r_1 r_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$
M ₁ m ₁ M ₂ m ₂	m ₁ M ₁ M ₂ m ₂	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 t_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 s_2 r_1 r_2 / rs$
M ₁ m ₁ M ₂ m ₂	m ₁ m ₁ M ₂ m ₂	$s_1 s_2 r_1 r_2 / t$	$t_1 t_2 / t$	$t_1 t_2 / t$	$s_1 s_2 r_1 r_2 / t$
M ₁ m ₁ m ₂ M ₂	M ₁ m ₁ m ₂ m ₂	$s_1 r_1 r_2^2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1 r_1 s_2^2 / rs$
M ₁ m ₁ m ₂ M ₂	m ₁ m ₁ m ₂ m ₂	$s_1 r_1 r_2^2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_2 r_2 t_1 / rs$	$s_1 r_1 s_2^2 / rs$
m ₁ M ₁ M ₂ M ₂	M ₁ m ₁ M ₂ M ₂	$r_1^2 s_2^2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$s_1 s_2 r_1 r_2 / r^2$	$s_1^2 r_2^2 / r^2$
m ₁ M ₁ M ₂ m ₂	m ₁ M ₁ M ₂ m ₂	$r_1^2 r_2 s_2 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1^2 s_2 r_2 / rs$
m ₁ M ₁ M ₂ m ₂	m ₁ m ₁ M ₂ m ₂	$r_1^2 r_2 s_2 / rs$	$s_1 r_1 t_2 / rs$	$s_1 s_2 r_1 r_2 / rs$	$s_1^2 s_2 r_2 / rs$
m ₁ m ₁ M ₂ M ₂	M ₁ m ₁ M ₂ M ₂	$r_1^2 r_2^2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$s_1 s_2 r_1 r_2 / s^2$	$s_1^2 s_2^2 / s^2$

Abbreviations: IF₂, immortalized F₂; QTL, quantitative trait locus.

The IF₂ indicates the population from random mating of double haploid (DH) lines; BC₁ (BC₂) indicates the back-cross population of IF₂ with P₁ (P₂); r_1 , r_2 and r are the recombination rates between marker 1 and QTL, between QTL and marker 2 and between the two flanking markers, $s = 1 - r$, $t = s^2 + r^2$, and similarly for s_1 , s_2 , t_1 and t_2 . If the IF₂ is from random mating of recombinant inbred lines (RILs), instead of the recombination rates, the proportions of recombinant zygote should be used.

following model,

$$\begin{aligned}
 y_{hij} = & \mu_h + x_{ik}^m a_{kh}^m + x_{ik}^{Dm} d_{kh}^m + x_{ik}^{Ae} a_{kh}^e + x_{ik}^{De_1} d_{kh}^e + x_{ik}^{De_2} d_{kh}^e \\
 & + \sum_{t=1}^s (\zeta_{it}^{+Am} a_{th}^{+m} + \zeta_{it}^{+Dm} d_{th}^{+m} + \zeta_{it}^{+Ae} a_{th}^{+e} + \zeta_{it}^{+De_1} d_{th}^{+e} + \zeta_{it}^{+De_2} d_{th}^{+e}) \\
 & + \sum_{t=1}^s (\zeta_{it}^{-Am} a_{th}^{-m} + \zeta_{it}^{-Dm} d_{th}^{-m} + \zeta_{it}^{-Ae} a_{th}^{-e} + \zeta_{it}^{-De_1} d_{th}^{-e} + \zeta_{it}^{-De_2} d_{th}^{-e}) \\
 & + B_{j(h)} + e_{hij}
 \end{aligned} \quad (4)$$

where a_{kh}^m and d_{kh}^m are the maternal additive and dominance effects of the QTL k in the h -th environment; a_{kh}^e , d_{kh}^e and d_{kh}^e are the endosperm additive, the first-order dominance and the second-order dominance effects, respectively; all the other parameters have the same definition in model (2) and model (3). Similar to scanning marker intervals, F-statistic is calculated for each putative QTL position, the F-statistic profile along chromosome will be obtained and the peaks, which reach the criteria of significance by permutation, are selected as candidate QTLs. Finally, we construct a QTL full model that consists of the detected candidate QTLs, and then conduct model selection to eliminate the false positive QTLs. The number of QTLs and their positions will be determined in the final model for estimation of QTL effects.

Genetic effect estimation

With all positions of QTLs identified, we can estimate the maternal QTL effects and endosperm QTL effects or the maternal QTL effects and embryo QTL effects, as well as their corresponding QE interaction effects based on the final QTL full model. To fit the mixed linear model, we first obtain a set of initial estimations of QTL parameters including the variances of random effects by the minimum norm quadric unbiased estimation (Rao, 1971), the fixed effects by the ordinary least square estimation and the random effects by the adjusted unbiased prediction, and then input them as prior values of parameters into the process of the Markov chain Monte Carlo for the mixed linear model by applying Gibbs sampling (Smith and Roberts, 1993). After we get the distribution of each parameter in the mixed linear model, the mean value is set as the final estimation and its significance is tested by t -statistic.

Simulation scenarios

To investigate the efficiency and accuracy of the proposed methods, we performed Monte Carlo simulations to verify the unbiasedness and robustness of our model. Here, double back-cross of IF₂ was simulated with 400 genotypes and two environments. The genetic map is set as 5 chromosomes, with 11 markers evenly distributed on each chromosome. The distance

between two consecutive markers is set to 10 cM. Six QTLs (Q_1 , Q_2 , Q_3 , Q_4 , Q_5 and Q_6) controlling a seed quantitative trait are scattered on the five chromosomes, in which Q_5 and Q_6 are on the same chromosome. One QTL (Q_4) is set with only maternal effects; two QTLs (Q_1 and Q_5) are set with only endosperm effects; three QTLs (Q_2 , Q_3 and Q_6) are set with both maternal and endosperm effects. Different heritabilities and population structures (that is, the ratio of the number of IF_2 line and the number of offspring each line) were used in simulations to investigate the influence of heritability and population structure on estimation of QTL parameter. Each simulation scenario was repeated 500 times.

A worked example

To demonstrate the use of the proposed method, we analyzed a set of real data from a cotton (*Gossypium hirsutum* L.) breeding experiment. A set of 188 recombinant inbred lines were developed from an intraspecific hybrid cross using two upland cotton germplasms HS46 and WARKCBUCAG8US-1-88 as the parents that have wide genetic differences in yield, fiber quality, disease resistance and seed quality traits. In this study, every two lines among the 188 recombinant inbred lines were randomly crossed during flowering to produce 376 IF_2 lines in 2009 and 2010 that were used for QTL analysis. Seeds of the IF_2 population and two parents were manually harvested at maturity. All the seeds were measured for fiber percentage. The genetic map includes 388 molecular

markers resided on 30 linkage groups. It covers a total length of 1946.22 cM, accounting for 41.55% of the whole genome, with an average distance of 5.03 cM between adjacent markers.

As the cotton seed consists of no triploid endosperm tissue, the cotton fiber percentage is mainly governed by the maternal effects and the embryo effects. Model (1) with inclusion of QTL maternal additive effect, embryo additive and dominance effects and their interaction effects with environments and its reduced model with exclusion of QE interaction effects were used in statistical analysis. The critical F-values were calculated to prescreen candidate intervals by permutation testing at the genome-wide significance level of 0.05. The full model was fit by the Markov chain Monte Carlo method.

RESULTS

Estimation of QTL parameter

A summary of results is presented in Tables 4–7. Under the heritability of 25% and the population structure of 200:2 (200 IF_2 lines and 2 offspring each IF_2 line), it is clear that all QTL positions could be unbiasedly estimated, the largest bias estimated value – Parameter value (Est. – Par.) is only 0.3 cM, and all the six ratios of bias to true value were <1% (Table 4). The majority of powers of QTL detection reached 100%, and even the lowest still reached 94% for Q_4 . For the estimation of QTL effects, most estimates were close to the true values of parameter and the estimation accuracies of QTL effects are reasonably good, especially for the QTL main effects (Tables 4 and 6). Most ratios of biases to true values were within 5% and all were <10%. We also found that the estimation accuracies of QE interaction effects appeared to be a little bit worse than those of QTL main effects (Table 6). The interaction effects of maternal and endosperm additive with environment (ae^m and ae^e) exhibited relatively higher estimation accuracy than maternal and endosperm dominance interaction with environment (de^m and de^e); all the ratios of bias to true value were <15% for ae^m and ae^e , but >15% for de^m and de^e .

Influence of heritability and population structure

The simulations, under other three heritabilities of 75%, 50 and 10% with population structure of 200:2 (200 IF_2 lines and 2 offspring each line), showed that the estimates of QTL positions and QTL effects were very close to each other among the three levels of heritabilities (Table 6). All the estimates of QTL position were unbiased. The estimates of maternal additive and endosperm additive effects were also robust for different heritabilities. The s.d. of parameter estimation became a little bit larger, and the power decreased when the heritability decreased from 75 to 10%; especially for the Q_4 , its power decreased to 76.8% at 10% heritability.

Table 3 Conditional probability of QTL genotype given a pair of flanking marker genotypes in the populations of IF_2 from random mating of DH or RIL lines

	IF_2		
	QQ	Qq	qq
$M_1M_1M_2M_2$	p_1^2	$2p_1q_1$	q_1^2
$M_1M_1M_2m_2$	p_1p_2	$p_1q_2 + p_2q_1$	q_1q_2
$M_1M_1m_2m_2$	p_2^2	$2p_2q_2$	q_2^2
$M_1m_1M_2M_2$	p_1p_3	$p_1q_3 + q_1p_3$	q_1q_3
$M_1m_1M_2m_2$	$\alpha p_1p_4 + \beta p_2p_3$	$\alpha(p_1q_4 + q_1p_4) + \beta(p_2q_3 + q_2p_3)$	$\alpha q_1q_4 + \beta q_2q_3$
$M_1m_1m_2m_2$	p_2p_4	$p_2q_4 + q_2p_4$	q_2q_4
$m_1m_1M_2M_2$	p_3^2	$2p_3q_3$	q_3^2
$m_1m_1M_2m_2$	p_3p_4	$p_3q_4 + q_3p_4$	q_3q_4
$m_1m_1m_2m_2$	p_4^2	$2p_4q_4$	q_4^2

Abbreviations: DH, double haploid; IF_2 , immortalized F_2 ; QTL, quantitative trait locus; RIL, recombinant inbred lines.

Note that r_1 and r_2 are the recombination rates between QTL and two flanking markers (M_1 and M_2), respectively, and r is the recombination rate between M_1 and M_2 ; $s_1 = 1 - r_1$, $s_2 = 1 - r_2$, $s = 1 - r$, $p_1 = s_1s_2/s$, $p_2 = s_1r_2/r$, $p_3 = r_1s_2/r$, $p_4 = r_1r_2/s$, $\alpha = s^2/(r^2 + s^2)$ for IF_2 from random mating of DH lines, and $p_1 = 1/(1 + 4r_1r_2)$, $p_2 = r_2/(r_1 + r_2)$, $p_3 = r_1/(r_1 + r_2)$, $p_4 = 4r_1r_2/(1 + 4r_1r_2)$, $\alpha = 1/(1 + 4r^2)$ for IF_2 from random mating of RIL lines; $q_i = 1 - p_i$, $i = 1, 2, 3, 4$, $\beta = 1 - \alpha$.

Table 4 Mapping power and estimates of QTL positions and main genetic effects

QTL	Chr.	Pos.		a^m		d^m		a^e		d^e		Power (%)
		Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	
Q_1	1	75	74.91 (2.47)	0	0.11 (0.68)	0	−0.01 (0.86)	3.77	3.54 (0.56)	2.61	2.55 (0.60)	100
Q_2	2	34	34.05 (2.92)	3.5	3.16 (0.88)	2.6	2.36 (0.84)	−3.5	−3.16 (0.63)	−2.6	−2.41 (0.67)	100
Q_3	3	44	44.30 (1.39)	3.3	3.39 (0.73)	2.5	2.48 (0.78)	4.1	3.88 (0.49)	1.9	1.87 (0.56)	100
Q_4	4	53	52.96 (3.69)	2	1.99 (0.71)	3.1	2.92 (0.86)	0	0.01 (0.47)	0	−0.03 (0.59)	94
Q_5	5	6	6.06 (2.20)	0	−0.03 (0.73)	0	0.01 (0.82)	−3.3	−3.08 (0.56)	−1.8	−1.68 (0.63)	100
Q_6	5	77	77.04 (1.69)	3.2	3.16 (0.72)	2.3	2.17 (0.78)	3.3	3.12 (0.50)	2.1	2.01 (0.60)	100

Abbreviations: Chr., the number of simulated chromosome; Est., the estimate of parameter; Par., the true parameter value in simulations; Pos., position (the distance (cM) between the QTL and the first marker on the same chromosome); Power, the percentage of the QTL that is detected correctly to be significant at 0.05 and/or within the 95% confidence interval; QTL, quantitative trait locus.

Note that a^m , d^m , a^e and d^e are maternal additive and dominance effects, endosperm additive and dominance effects, respectively. The s.d. indicates the standard deviation for the estimate of the true parameter value.

Table 5 Estimates of QE interaction effects

QTL	Chr.	ae ₁ ^m		ae ₂ ^m		de ₁ ^m		de ₂ ^m		ae ₁ ^e		ae ₂ ^e		de ₁ ^e		de ₂ ^e	
		Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)	Par.	Est. (s.d.)
Q ₁	1	0	0.01 (0.13)	0	-0.01 (0.13)	0	-0.00 (0.18)	0	0.00 (0.18)	0	0.01 (0.09)	0	-0.01 (0.09)	0	-0.01 (0.16)	0	0.01 (0.16)
Q ₂	2	0	-0.01 (0.14)	0	0.01 (0.14)	0	-0.01 (0.18)	0	0.01 (0.18)	0	0.01 (0.09)	0	-0.00 (0.09)	0	0.00 (0.15)	0	-0.00 (0.15)
Q ₃	3	-1.4	-1.39 (0.50)	1.4	1.42 (0.52)	-1.6	-1.33 (0.58)	1.6	1.32 (0.58)	-1	-0.88 (0.36)	1	0.88 (0.35)	1	0.64 (0.56)	-1	-0.62 (0.56)
Q ₄	4	1.2	1.03 (0.42)	-1.2	-1.05 (0.43)	-1.2	-0.77 (0.64)	1.2	0.78 (0.64)	0	0.05 (0.15)	0	-0.05 (0.15)	0	0.00 (0.17)	0	-0.01 (0.17)
Q ₅	5	0	0.07 (0.23)	0	-0.07 (0.23)	0	0.02 (0.20)	0	-0.01 (0.20)	1.1	0.99 (0.28)	-1.1	-1.02 (0.27)	1.2	0.96 (0.57)	-1.2	-0.95 (0.57)
Q ₆	5	0	0.05 (0.21)	0	-0.05 (0.21)	0	0.01 (0.16)	0	-0.01 (0.16)	1.3	1.16 (0.25)	-1.3	-1.11 (0.26)	1.1	0.79 (0.54)	-1.1	-0.80 (0.54)

Abbreviations: Chr., the number of simulated chromosome; Est., the estimate of parameter; Par., the true parameter value in simulations; QE, QTL-by-environment; QTL, quantitative trait locus. Note that ae₁^m, ae₂^m, de₁^m, de₂^m are maternal additive × environment interaction effects, maternal dominance × environment interaction effects in the first and the second environment, respectively; ae₁^e, ae₂^e, de₁^e and de₂^e are endosperm additive × environment interaction effects, endosperm dominance × environment interaction effects in the first and second environment, respectively. The s.d. indicates the standard deviation for the estimate of the true parameter value.

Table 6 Estimates of QTL positions and effects under three different heritabilities

QTL	Chr.	Pos.			a ^m			a ^e			Power (%)					
		Par.	Est. (s.d.)	Par.	I	II	III	Par.	Est. (s.d.)	Par.	I	II	III			
Q ₁	1	75	74.95 (2.19)	74.95 (2.25)	74.93 (2.89)	0	0.11 (0.62)	0.11 (0.64)	0.14 (0.82)	3.77	3.55 (0.50)	3.55 (0.51)	3.52 (0.64)	100	100	99.2
Q ₂	2	34	34.05 (2.82)	34.06 (2.86)	34.19 (3.26)	3.5	3.14 (0.86)	3.13 (0.88)	3.19 (0.99)	-3.5	-3.14 (0.63)	-3.13 (0.65)	-3.18 (0.67)	100	100	98.6
Q ₃	3	44	44.29 (1.32)	44.30 (1.36)	44.28 (1.53)	3.3	3.40 (0.66)	3.40 (0.68)	3.40 (0.87)	4.1	3.88 (0.43)	3.88 (0.45)	3.87 (0.59)	100	100	100
Q ₄	4	53	53.09 (3.36)	53.11 (3.44)	52.93 (3.97)	2	1.96 (0.65)	1.98 (0.66)	2.12 (0.79)	0	0.01 (0.42)	0.01 (0.43)	-0.03 (0.59)	97.8	97.2	76.8
Q ₅	5	6	5.98 (2.03)	6.00 (2.09)	6.10 (2.49)	0	-0.02 (0.66)	-0.03 (0.67)	-0.05 (0.89)	-3.3	-3.10 (0.51)	-3.08 (0.52)	-3.06 (0.66)	100	100	100
Q ₆	5	77	77.00 (1.60)	77.03 (1.61)	77.08 (1.89)	3.2	3.16 (0.66)	3.16 (0.67)	3.17 (0.85)	3.3	3.12 (0.46)	3.12 (0.46)	3.11 (0.58)	100	100	100

Abbreviations: Chr., the number of simulated chromosome; Est., the estimate of parameter; Par., the true parameter value in simulations; Pos., position (the distance (cM) between the QTL and the first marker on the same chromosome); Power, the percentage of the QTL that is detected correctly to be significant at 0.05 and/or within the 95% confidence interval; QTL, quantitative trait locus. I, II and III stand for three different heritabilities of 75%, 50% and 10%, respectively, that are the proportions of the total phenotypic variation ascribed to QTL main effects and QE interaction effects; a^m and a^e are maternal additive and endosperm effects, respectively. The s.d. indicates the standard deviation for the estimate of the true parameter value.

Table 7 Estimates of QTL position and effects under three different population structures

QTL	Chr.	Pos.			a ^m			a ^e			Power (%)		
		Est. (s.d.)			Par.			Est. (s.d.)			I		
		I	II	III	I	II	III	I	II	III	I	II	III
Q ₁	1	74.95 (2.19)	75.16 (2.55)	75.15 (3.40)	0	0.11 (0.62)	0.23 (0.88)	0.11 (1.22)	3.77	3.55 (0.50)	3.39 (0.66)	100	100
Q ₂	2	34.05 (2.82)	33.94 (3.42)	34.19 (3.91)	3.5	3.14 (0.86)	3.01 (1.09)	2.93 (1.43)	-3.5	-3.14 (0.63)	-2.92 (0.84)	100	100
Q ₃	3	44.29 (1.32)	44.31 (1.94)	44.05 (2.76)	3.3	3.40 (0.66)	3.40 (0.81)	3.34 (1.17)	4.1	3.88 (0.43)	3.73 (0.60)	100	100
Q ₄	4	53.09 (3.36)	53.02 (4.24)	52.81 (4.55)	2	1.96 (0.65)	2.01 (0.76)	1.96 (1.19)	0	0.01 (0.42)	0.03 (0.40)	97.8	86.4
Q ₅	5	5.98 (2.03)	6.14 (2.60)	6.39 (2.91)	0	-0.02 (0.66)	-0.14 (0.84)	-0.00 (1.22)	-3.3	-3.10 (0.51)	-2.91 (0.62)	100	99.6
Q ₆	5	77.00 (1.60)	77.23 (2.00)	77.43 (2.92)	3.2	3.16 (0.66)	3.22 (0.84)	3.12 (1.12)	3.3	3.12 (0.46)	3.00 (0.54)	100	100

Abbreviations: Chr., the number of simulated chromosome; Est., the estimate of parameter; Par., the true parameter value in simulations; Pos., position (the distance (cM) between the QTL and the first marker on the same chromosome); Power, the percentage of the QTL that is detected correctly to be significant at 0.05 and/or within the 95% confidence interval; QTL, quantitative trait locus; I, II and III stand for three different population structures of 200:2, 100:4 and 50:8, respectively; the first number is the number of immortalized F₂ (IF₂) lines and the second number is the offspring number of each IF₂ line; heritability of 75% is employed in each case; a^m and a^e are maternal additive and endosperm effects, respectively. The s.d. indicates the standard deviation for the estimate of the true parameter value.

The influence of population structure was also investigated with three population structures 200:2, 100:4 and 50:8 under the heritability of 75%. Similar to the results for different heritabilities, the estimates of QTL positions and of effects still keep unbiased. The estimates of position or effects of QTL were also close to each other among three population constitutions, whereas the s.d. exhibited increasing tendency with changing of structure from I to III (Table 7). For the power, as the population structure changes, it seemed that the most unstable was the QTL with only maternal effect (Q₄). With the number of IF₂ lines decreasing from 200 to 50, the power of detecting Q₄ changes from 97.8 to 39.2%. This result is expected because the sample, to estimate maternal effects, decreases with small number of IF₂.

Application to the real data

The genetic main effects and environment interaction effects of all the QTLs are presented in Table 8. Two significant QTLs were located on chromosomes 19 and 21, respectively. Both the QTLs are sensitive to environment. The estimated relative contribution of Q₂ is 16.03%. In addition, we found that nearly 6.19% of the phenotypic variation is attributed to the maternal effects of QTL and ~10.83% to the QE interaction. Furthermore, we used a reduced model without QE interaction effect to reanalyze the data. Although the Q₂ was still detected by the reduced model, the corresponding estimated relative contribution is decreased to 8.19%, clearly showing the adequacy of the proposed analysis. Hence, in practical mapping, we suggest to take maternal effect and environment interactions into account when we have no knowledge available on the real genetic architecture of surveyed seed traits.

DISCUSSION

In the present study, we have proposed a new statistical method that takes multiple QTLs and QE interactions into account for mapping QTLs underlying seed traits. It has been well recognized that QE interaction is one important component of the total genetic effect that may cause the phenotype to vary across various environments. QE interactions have been considered in QTL mapping for diploid plant traits (Jansen, 1992; Jansen *et al.*, 1995; Jiang and Zeng, 1995; Cockerham and Zeng, 1996; SariGorla *et al.*, 1997; Yan *et al.*, 1998), but have not been considered for diploid (maternal plus embryo) or triploid (maternal plus endosperm) seed traits. When QE interactions do play a significant role for endosperm traits of interest, the methods that ignore such interactions may lead to biased estimation. The new method integrates the effects of QTLs that are expressed in maternal and/or offspring genome and QE interactions into one mapping framework, and the detected QE interactions can be further dissected into subfactors, such as interaction effects of maternal additive by environment, maternal dominance by environment and so on, that are valuable information for selecting or breeding environment-specific or environment-insensitive varieties by marker-assisted selection.

A major difficulty in developing a powerful statistical approach for mapping QTLs with QE interactions is to handle an increased number of parameters in the statistical model, especially for the random effects. To tackle this problem, we adopted a mixed linear model instead of more commonly used regression models. In spite of more burdensome computation, the mixed linear model approach is able to keep nondecreasing or even increasing statistical power even if the number of parameters is nearly doubled because of inclusion of QE interactions.

Table 8 Estimates of QTL position and effects for cotton fiber percentage in a IF₂ population from random mating of a set of RIL lines

QTL	Chromosome	Marker interval	Pos.	a ^m	a ^o	d ^o	ae ₁ ^m	ae ₂ ^m	ae ₁ ^o	ae ₂ ^o	de ₁ ^o	de ₂ ^o
Q1 ^a	19	NAU1023-MGHES76-B	21.3	1.62**	-0.81**	1.60**	1.28**	-1.25**	-0.68**	0.70**	1.28*	-1.28*
Q2 ^a	21	NAU1272-A-NAU3217	47.9	-0.85**	1.94**	0.80*	-0.46	0.45	1.55**	-1.157**	0.44	-0.45
Q2 ^b	21	NAU1272-A-NAU3217	47.9	-0.63**	1.68**	0.85*						

Abbreviations: IF₂, immortalized F₂; Pos., position (the distance (cM) between the QTL and the first marker on the same chromosome); QTL, quantitative trait locus; RIL, recombinant inbred lines. a^m, a^o, d^o, ae₁^m, ae₂^m, ae₁^o, ae₂^o, de₁^o and de₂^o are the maternal effect, embryo additive effect, embryo dominance effect of QTL and their interaction effects with the first and the second environment, respectively.

** And * indicate reaching significance at level of 0.01 and 0.05, respectively.

^aResult from the model with QTL effect and QTL-by-environment (QE) interaction effect.

^bResult from the model with exclusion of QE interaction effects.

Unlike the previous methods (Wu *et al.*, 2002; Xu *et al.*, 2003), most of which are based on one QTL model and thus may have a potential bias in parameter estimation when two or more linked QTLs exist in genome, the proposed method uses a QTL full model to estimate all possible effects of QTLs. False positive QTL(s) could be eliminated by model selection, while higher power and more accurate estimation of QTL effects could be achieved, benefiting from smaller and more accurate estimation of the residual effects because of background control. It should be noted that we have integrated the maternal genome into the model in addition to the offspring genome (Foolad and Jones, 1992; Cui *et al.*, 2004; Hu and Xu, 2005; Wen and Wu, 2007), based on the view that development of embryo/endosperm traits partially depends on its own maternal plant.

It is well documented that epistasis is extensively involved in genetic variation of complex traits. Epistatic effects and their interaction effects with environment have been included in the model for mapping QTL of general crop agronomic traits (Wang *et al.*, 1999; Yang *et al.*, 2007). For mapping epistatic effects of QTL for seed traits, Cui and Wu (2005a) and Cui *et al.* (2006) proposed a model including interaction effects of one paired QTL in the same genome (either of maternal genome, embryo genome or endosperm genome) or two different genomes based on the double back-cross design. Although our models (1) and (2) currently do not include the epistatic interactions among genes in the different genomes, under current framework of our method, it is in principle feasible to extend the proposed method to cover such epistatic effects and their interaction effects with environments. As such, an extension involves a more complicated QTL full model and also requires intensive work in the development of methodology and software for various experimental designs, and this will be our follow-up study.

Monte Carlo simulations demonstrate that the method could precisely locate the positions of QTLs, and well estimate the genetic effects, even if the environment or gene-environment interaction exists. A satisfactory power of QTL detection is still obtained, despite increased number of parameters in the model due to inclusion of QE interactions as compared with current available methods. According to the simulation results under different heritabilities (Table 6), even in the heritability of 10%, most QTLs (5 in 6) could be detected with power of >98%. As the population structure changes, the power of QTL except Q₄ are >99% (Table 7). For different population structures with the same sample size, we can easily conclude that increasing maternal plant number will improve the mapping precision in the estimation of QTL parameter. This is reasonable as variation of trait ascribed to maternal effect decreases with a reduction in the number of maternal plants.

Understanding the genetic architecture of seed development helps us to answer many fundamental questions in higher plants. There is a

straightforward application of our model in breeding practice or genetic studies. For example, if the estimated maternal effects surpass offspring effect, more attention should be paid to selection of maternal effects for improving seed traits; if the offspring effects dominate, it is better to directly select the seed for reproduction. But there are some other issues that should be noted. Our method is based on a QTL full model that can control the false positive rate and improve estimation precision of QTL parameter, but for the double backcross of IF₂ design, linear dependency will occur between endosperm genetic effects of any pair of QTLs for the model (2), because of $x^{Ae} - 2x_1^{De} + 2x_2^{De} = 1.5$ or -1.5 for any individual QTL. The simplest way to tackle this problem is to reparameterize, for example, to merge the first-order and the second-order endosperm dominance effects for each QTL, that has been employed in our simulation study. Another potential strategy is to keep the first-order and the second-order endosperm dominance effects of one QTL but to merge the two dominance effects as well as their interaction effects with environment for other endosperm QTL(s). However, the potential strategy will increase computational complexity as the genetic effects of all QTLs cannot be simultaneously estimated, and the QTL full model has to be adjusted for each QTL; on the other hand, the effectiveness of this strategy still need investigating in further study. For the selfing of IF₂ design, the $p_{3|ij}$ and $p_{4|ij}$ (Table 1) is compound because of inability to separate the $Qq:Qq(QQq)$ and $Qq:Qq(Qqq)$ based on marker information. Therefore, the first-order and the second-order dominance effects need to merge in the model (2).

The QTL genotypes are unobserved and need to be inferred from the flanking markers. In the traditional fixed models, there are two ways to infer the missing QTL genotypes: the mixture distribution method and the conditional expectation method used in interval mapping of Lander and Botstein (1989) and interval regression approach of Haley and Knott (1992), respectively. Both the methods are considerably effective for QTL parameter estimation. The comparisons between the two methods also support that there exists only a slight difference in most cases (Xu, 1995, 1996; Gessler and Xu, 1996; Kao, 2000). Different from those in a fixed model, the observations are no longer independent of each other in a mixed model that is considered in this study. Although feasible in theory, it is not easy to handle the joint mixture distribution of a set of correlated observations. Therefore, we choose the expectation method in inferring the coefficients associated with the QTL genotypes. Our estimation method is, in essence, an extension of the regression approach of Haley and Knott (1992).

Based on the models and methods proposed in the present study, the computer software named QTLNetwork-Seed was developed in C++ programming language. This software can be run on the most commonly used operation systems. Double back-cross of IF₂, selfing

of IF₂ and random mating for IF₂, double haploid or recombinant inbred line designs can be handled by this software for mapping QTLs underlying seed traits.

DATA ARCHIVING

Data are available from our own public Internet website <http://ibi.zju.edu.cn/software/> and from the Dryad Digital Repository doi:10.5061/dryad.8311n.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported in part by the National Basic Research Program of China (973 Program 2011CB109306, 2010CB126006), the National Natural Science Foundation 31271608 and the National Institutes of Health Grant DA025095, the Bill and Melinda Gates Foundation Project.

Araki E, Miura H, Sawada S (1999). Identification of genetic loci affecting amylose content and agronomic traits on chromosome 4A of wheat. *Theor Appl Genet* **98**: 977–984.

Bazzaz FA (2001). Plant biology in the future. *Proc Natl Acad Sci USA* **98**: 5441–5445.

Benner MS, Phillips RL, Kiriwara JA, Messing JW (1989). Genetic analysis of methionine rich storage protein accumulation in maize. *Theor Appl Genet* **78**: 761–767.

Cockerham CC, Zeng ZB (1996). Design III with marker loci. *Genetics* **143**: 1437–1456.

Cui Y, Casella G, Wu R (2004). Mapping quantitative trait loci interactions from the maternal and offspring genomes. *Genetics* **167**: 1017–1026.

Cui Y, Wu R (2005a). Mapping genome-genome epistasis: a high-dimensional model. *Bioinformatics* **21**: 2447–2455.

Cui YH, Wu JG, Shi CH, Littell RC, Wu RL (2006). Modelling epistatic effects of embryo and endosperm QTL on seed quality traits. *Genet Res* **87**: 61–71.

Cui YH, Wu RL (2005b). Statistical model for characterizing epistatic control of triploid endosperm triggered by maternal and offspring QTLs. *Genet Res* **86**: 65–75.

Foolad M, Jones R (1992). Models to estimate maternally controlled genetic variation in quantitative seed characters. *Theor Appl Genet* **83**: 360–366.

Garcia D, Fitz Gerald JN, Berger F (2005). Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in Arabidopsis. *Plant Cell* **17**: 52–60.

Gessler DD, Xu S (1996). Using the expectation or the distribution of the identity by descent for mapping quantitative trait loci under the random model. *Am J Hum Genet* **59**: 1382–1390.

Haley CS, Knott SA (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.

Henderson CR (1953). Estimation of variance and covariance components. *Biometrics* **9**: 226–252.

Hu ZQ, Xu CW (2005). A new statistical method for mapping QTLs underlying endosperm traits. *Chinese Sci Bull* **50**: 1470–1476.

Jansen RC (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor Appl Genet* **85**: 252–260.

Jansen RC, Vanooijen JW, Stam P, Lister C, Dean C (1995). Genotype-by-environment interaction in genetic-mapping of multiple quantitative trait loci. *Theor Appl Genet* **91**: 33–37.

Jiang CJ, Zeng ZB (1995). Multiple-trait analysis of genetic-mapping for quantitative trait loci. *Genetics* **140**: 1111–1127.

Kao CH (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**: 855–865.

Kao CH (2004). Multiple-interval mapping for quantitative trait loci controlling endosperm traits. *Genetics* **167**: 1987–2002.

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

Letchworth MB, Lambert RJ (1998). Pollen parent effects on oil, protein, and starch concentration in maize kernels. *Crop Sci* **38**: 363–367.

Mazur B, Krebbers E, Tingey S (1999). Gene discovery and product development for grain quality traits. *Science* **285**: 372–375.

Mo HD (1995). Identification of genetic control for endosperm traits in cereals. *Acta Genetica Sinica* **22**: 126–132.

Ohto M, Fischer RL, Goldberg RB, Nakamura K, Harada JJ (2005). Control of seed mass by APETALA2. *Proc Natl Acad Sci USA* **102**: 3123–3128.

Olsen OA (1998). Endosperm developments. *Plant Cell* **10**: 485–488.

Parker GD, Chalmers KJ, Rathjen AJ, Langridge P (1998). Mapping loci associated with flour colour in wheat (*Triticum aestivum* L.). *Theor Appl Genet* **97**: 238–245.

Rao CR (1971). Minimum variance quadratic unbiased estimation of variance components. *J Multivariate Anal* **1**: 445–456.

SariGorla M, Calinski T, Kaczmarek Z, Krajewski P (1997). Detection of QTL x environment interaction in maize by a least squares interval mapping method. *Heredity* **78**: 146–157.

Searle SR (1968). Another look at Henderson's methods of estimating variance components. *Biometrics* **24**: 749–787.

Sene M, Thevenot C, Hoffmann D, Benetrix F, Causse M, Prioul JL (2001). QTLs for grain dry milling properties, composition and vitreousness in maize recombinant inbred lines. *Theor Appl Genet* **102**: 591–599.

Shi CH, Zhu J, Wu JG, Fan LJ (2000). Genetic and genotype x environment interaction effects from embryo, endosperm, cytoplasm and maternal plant for rice grain shape traits of indica rice. *Field Crop Res* **68**: 191–198.

Smith AF, Roberts GO (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J Roy Stat Soc B* **55**: 3–23.

Sourdille P, Perretant MR, Charmet G, Leroy P, Gautier MF, Joudrier P *et al.* (1996). Linkage between RSLP markers and genes affecting kernel hardness in wheat. *Theor Appl Genet* **93**: 580–586.

Sundaresan V (2005). Control of seed size in plants. *Proc Natl Acad Sci USA* **102**: 17887–17888.

Tan YF, Sun M, Xing YZ, Hua JP, Sun XL, Zhang QF *et al.* (2001). Mapping quantitative trait loci for milling quality, protein content and color characteristics of rice using a recombinant inbred line population derived from an elite rice hybrid. *Theor Appl Genet* **103**: 1037–1045.

van der Meer IM, Bovy AG, Bosch D (2001). Plant-based raw material: improved food quality for better nutrition via plant genomics. *Curr Opin Biotech* **12**: 488–492.

Wang DL, Zhu J, Li ZK, Paterson AH (1999). Mapping QTLs with epistatic effects and QTL x environment interactions by mixed linear model approaches. *Theor Appl Genet* **99**: 1255–1264.

Wen Y, Wu W (2007). Interval mapping of quantitative trait loci underlying triploid endosperm traits using F3 seeds. *J Genet J Genet* **34**: 429–436.

Wu RL, Lou XY, Ma CX, Wang XL, Larkins BA, Casella G (2002). An improved genetic model generates high-resolution mapping of QTL for protein quality in maize endosperm. *Proc Natl Acad Sci USA* **99**: 11281–11286.

Xu C, He X, Xu S (2003). Mapping quantitative trait loci underlying triploid endosperm traits. *Heredity* **90**: 228–235.

Xu S (1995). A comment on the simple regression method for interval mapping. *Genetics* **141**: 1657–1659.

Xu S (1996). Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics* **144**: 1951–1960.

Yan JQ, Zhu J, He CX, Benmoussa M, Wu P (1998). Molecular dissection of developmental behavior of plant height in rice (*Oryza sativa* L.). *Genetics* **150**: 1257–1265.

Yang J, Zhu J, Williams RW (2007). Mapping the genetic architecture of complex traits in experimental populations. *Bioinformatics* **23**: 1527–1536.

Zeng ZB (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

Zhang AH, Xu CW, Mo HD (1999). Genetic expression of several quality traits in Indica-Japonica hybrids. *Acta Agronomica Sinica* **25**: 401–407.

Zhu J, Weir BS (1994). Analysis of cytoplasmic and maternal effects. 2. Genetic models for triploid endosperms. *Theor Appl Genet* **89**: 160–166.