Theor Appl Genet (2008) 116:243–260
DOI 10.1007/s00122-007-0663-5

ORIGINAL PAPER

Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations

Huihui Li · Jean-Marcel Ribaut · Zhonglai Li · Jiankang Wang

Received: 23 May 2007 / Accepted: 2 October 2007 / Published online: 6 November 2007

Abstract It has long been recognized that epistasis or interactions between non-allelic genes plays an important role in the genetic control and evolution of quantitative traits. However, the detection of epistasis and estimation of epistatic effects are difficult due to the complexity of epistatic patterns, insufficient sample size of mapping populations and lack of efficient statistical methods. Under the assumption of additivity of QTL effects on the phenotype of a trait in interest, the additive effect of a QTL can be completely absorbed by the flanking marker variables, and the epistatic effect between two QTL can be completely absorbed by the four marker-pair multiplication variables between the two pairs of flanking markers. Based on this property, we proposed an inclusive composite interval mapping (ICIM) by simultaneously considering marker variables and marker-pair multiplications in a linear model. Stepwise regression was applied to identify the most significant markers and marker-pair multiplications. Then a two-dimensional scanning (or interval mapping) was conducted to identify QTL with significant digenic epistasis using adjusted phenotypic values based on the best multiple regression model. The adjusted values retain the information of QTL on the two current mapping intervals but exclude the influence of QTL on other intervals and chromosomes. Epistatic QTL can be identified by ICIM, no matter whether the two interacting QTL have any additive effects. Simulated populations and one barley doubled haploids (DH) population were used to demonstrate the efficiency of ICIM in mapping both additive QTL and digenic interactions.

Communicated by M. Sillanpää.

H. Li · Z. Li
School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China

J.-M. Ribaut
Generation Challenge Programme, c/o CIMMYT, Apdo. Postal 6-641, Mexico, D.F. 06600, Mexico

H. Li · J. Wang (✉)
The National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Science and CIMMYT China Office, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: wangjk@caas.net.cn; jkwang@cgiar.org

Introduction

Epistasis or interactions between non-allelic genes makes a substantial contribution to the genetic control and evolution of quantitative traits (Frankel and Schork 1996; Lynch and Walsh 1998; Wade 2002; Kroymann and Mitchell-Olds 2005; Carlborg et al. 2003, 2006; Malmberg et al. 2005; Zeng 2005). The pattern of epistasis for a trait can be very complex, and the genetic model with epistatic effects potentially contains a large number of model effects. Therefore, it is more difficult to identify the epistatic QTL and estimate the epistatic effects (Frankel and Schork 1996; Mackay 2001). Our knowledge of how interacting genes influence the phenotype of quantitatively inherited traits remains incomplete. Statistical methodology for epistatic mapping is still under development.

Some mapping methods based on frequentist statistics, such as interval mapping (Lander and Botstein 1989) and regression interval mapping (Haley and Knott 1992; Whittaker et al. 1996; Feenstra et al. 2006) may be extended for mapping epistasis, but the mapping power was low as the background genetic variation was not well

controlled. Multiple interval mapping (MIM) proposed by Kao et al. (1999) fits multiple putative QTL effects and associated epistatic effects simultaneously in one model to facilitate the search, test and estimation of positions, effects and interactions of multiple QTL. However, it requires deciding the number of model terms (main effect and epistasis) in the model. As this is usually unknown, various models of different complexities have to be tested (Doerge 2002). Different MIM model selection methods implemented in the popular software of QTL Cartographer (Wang et al. 2005) give different, sometimes controversial mapping results, and the nature of the preferred model selection method is not clear (Li et al. 2007). Jannink and Jansen (2001) and Boer et al. (2002) proposed a statistical method to map epistatic QTL by identifying loci of high QTL by genetic background interaction through one-dimensional scanning. In their methods, either large mapping populations derived from multiple related inbred-line crosses are required, or the effective dimension of the epistatic effects needs to be specified by users. Assuming that QTL are at marker positions, multiple regression using modified Schwarz Bayesian information criterion has been proposed by Bogdan et al. (2004) and Baierl et al. (2006) to map digenic interactive QTL.

The use of Bayesian models in QTL mapping has been widely studied in recent years (Satagopan et al. 1996; Uimari et al. 1996; Uimari and Hoeschele 1997; Sen and Churchill 2001; Bogdan et al. 2004; Yi 2004; Yi et al. 2003, 2005). Earlier Bayesian models estimated the locations and the effect parameters for a pre-specified number of QTL (Satagopan et al. 1996; Uimari et al. 1996), which is normally unknown before mapping. To solve this problem, Bayesian methods implemented via the reversible jump Markov chain Monte Carlo (MCMC) algorithm have been proposed (Uimari and Hoeschele 1997; Sillanpää and Arjas 1999). Yi et al. (2003) extended the reversible MCMC Bayesian models to epistasis mapping. Considering that the complexity of the reversible jump steps increases computational demand and may prohibit further improvements of the algorithm, Yi et al. (2005) extended the composite model space approach proposed in Yi (2004) to include epistatic effects. However, Bayesian models have not been widely accepted due to the difficulty and arbitrary in choosing priors, and the intensive computing requirements. As pointed by Xu and Jia (2007), most Bayesian models failed for a barley population consisting of 145 doubled haploid (DH) lines and 127 markers, and therefore they proposed an empirical Bayesian model.

Recently, Li et al. (2007) proposed an inclusive composite interval mapping (ICIM) to improve the traditional composite interval mapping (CIM; Zeng 1994) for QTL with additive effects. In ICIM, marker selection was conducted only once through stepwise regression by considering all marker information simultaneously. The phenotypic values were adjusted by all markers retained in the regression equation except the two markers flanking the current mapping interval. The adjusted phenotypic values were then used in interval mapping. This strategy effectively separates the cofactor selection from the interval mapping using ML method. Simulations showed that ICIM is computationally less intensive, and has increased detection power, reduced false discovery rate, and less biased estimates of QTL effects (Li et al. 2007).

In this paper, we extend ICIM to map digenic interacting QTL. The efficiency of the proposed method is demonstrated through extensive simulations and one real population in barley.

Material and methods

The basic linear model and its properties in mapping digenic epistasis

For simplicity of theoretical derivation, we assume that two inbred parents P_1 and P_2 differ in m QTL, which are distributed in m intervals flanked by $m + 1$ markers on one chromosome. Intervals where no QTL are located are viewed as having QTL with effects of zero. Multiple QTL located in one marker interval are not considered here. The parental QTL genotype is assumed to be $Q_1Q_1Q_2Q_2\ldots Q_mQ_m$ for P_1, and $q_1q_1q_2q_2\ldots q_mq_m$ for P_2. Suppose that we have a sample of n individuals from a backcross population where P_1 is used as the recurrent parent. For an individual in a backcross population $\mathbf{X} = (x_1, x_2, \ldots, x_m, x_{m+1})$ represents known marker variables which are equal to 1 and -1, standing for the two marker types (homozygote and heterozygote), respectively, and $\mathbf{G} = (g_1, g_2,\ldots, g_m)$ represents the unknown QTL variables which are equal to 1 and -1, standing for the two QTL genotypes (homozygote and heterozygote), respectively. Additive effects of QTL are represented by $a_1, a_2, \ldots, a_m$, respectively, and the epistatic effect between QTL j and k is denoted by aa_{jk} ($j, k = 1, \ldots, m$ and $j < k$). Under the assumption of additivity of QTL effects on phenotype, the genetic value G of an individual under additive and epistasis genetic model can be written in the following form:

$$G = \sum_{j=1}^{m} a_j g_j + \sum_{j<k} aa_{jk} g_j g_k. \quad (1)$$

The expectation of QTL genotype g_j is dependent of the position of the jth QTL on the chromosomal interval flanked by the jth and $(j + 1)$th markers, and the length of this interval (Zeng 1994; Whittaker et al. 1996), i.e.,

$$E(g_j|\mathbf{X}) = \lambda_j x_j + \rho_j x_{j+1}, \quad (2)$$

where $\lambda_j = \frac{(r_{j,j+1}-r_{j,q_j})(1-r_{j,j+1}-r_{j,q_j})}{r_{j,j+1}(1-r_{j,j+1})(1-2r_{j,q_j})}$, $\rho_j = \frac{r_{j,q_j}(1-r_{j,q_j})(1-2r_{j,j+1})}{r_{j,j+1}(1-r_{j,j+1})(1-2r_{j,q_j})}$, r_{j,q_j} is the recombination fraction between the jth marker and the jth QTL, and $r_{j,j+1}$ is the recombination fraction between the jth and $(j+1)$th markers. The expectation of QTL genotype $g_j g_k$ conditional on marker type $\mathbf{X}$ can be proved as,

$$\begin{aligned} E(g_j g_k|\mathbf{X}) &= \lambda_j\lambda_k x_j x_k + \lambda_j\rho_k x_j x_{k+1} + \rho_j\lambda_k x_{j+1}x_k + \rho_j\rho_k x_{j+1}x_{k+1} \\ &= E(g_j|\mathbf{X})E(g_k|\mathbf{X}). \end{aligned} \quad (3)$$

Therefore, the expectation of the genotypic value G in model (1) conditional on known marker types can be written as a linear function of all the marker variables and their multiplications, i.e.,

$$\begin{aligned} E(G|\mathbf{X}) &= \sum_{j=1}^{m} a_j(\lambda_j x_j + \rho_j x_{j+1}) \\ &\quad + \sum_{j<k} aa_{jk}(\lambda_j\lambda_k x_j x_k + \lambda_j\rho_k x_j x_{k+1} + \rho_j\lambda_k x_{j+1}x_k \\ &\quad + \rho_j\rho_k x_{j+1}x_{k+1}) \\ &\hat{=} \sum_{j=1}^{m+1} b_j x_j + \sum_{j<k} b_{jk} x_j x_k, \end{aligned} \quad (4)$$

where

$$\begin{aligned} &b_1 = \lambda_1 a_1, \\ &b_j = \rho_{j-1}a_{j-1} + \lambda_j a_j \quad \text{if } j = 2,\ldots,m, \\ &b_{m+1} = \rho_m a_m, \\ &b_{12} = \lambda_1\lambda_2 aa_{12}, \\ &b_{1k} = \lambda_1\rho_{k-1}aa_{1,k-1} + \lambda_1\lambda_k aa_{1k} \quad \text{if } k = 3,\ldots,m, \\ &b_{1,m+1} = \lambda_1\rho_m aa_{1m}, \\ &b_{j,j+1} = \rho_{j-1}\rho_j aa_{j-1,j} + \rho_{j-1}\lambda_{j+1}aa_{j-1,j+1} \\ &\quad + \lambda_j\lambda_{j+1}aa_{j,j+1} \quad \text{if } j = 2,\ldots,m-1, \\ &b_{jk} = \rho_{j-1}\rho_{k-1}aa_{j-1,k-1} + \rho_{j-1}\lambda_k aa_{j-1,k} + \lambda_j\rho_{k-1}aa_{j,k-1} \\ &\quad + \lambda_j\lambda_k aa_{jk} \quad \text{if } j \neq 1, k \neq m+1 \text{ and } j < k-1, \\ &b_{j,m+1} = \rho_{j-1}\rho_m aa_{j-1,m} + \lambda_j\rho_m aa_{jm} \quad \text{if } j = 2,\ldots,m-1, \\ &\quad \text{and } b_{m,m+1} = \rho_{m-1}\rho_m aa_{m-1,m}. \end{aligned}$$

Thus, the epistatic effect between QTL j and k only contributes to b_{jk}, $b_{j+1,k}$, $b_{j,k+1}$ and $b_{j+1,k+1}$. If there is at least one empty interval between the two current intervals $(j, j+1)$ and $(k, k+1)$, and no QTL are located in their neighboring intervals, i.e., $(j-1, j)$, $(j+1, j+2)$, $(k-1, k)$, and $(k+1, k+2)$, $aa_{j-1,k-1}$, $aa_{j-1,k}$, $aa_{j-1,k+1}$, $aa_{j,k-1}$, $aa_{j,k+1}$, $aa_{j+1,k-1}$, $aa_{j+1,k}$ and $aa_{j+1,k+1}$ are equal to zero. In this case, $b_{jk} = \lambda_j\,\lambda_k\,aa_{jk}$, $b_{j+1,k} = \rho_j\,\lambda_k\,aa_{jk}$, $b_{j,k+1} = \lambda_j\,\rho_k\,aa_{jk}$ and $b_{j+1,k+1} = \rho_j\,\rho_k\,aa_{jk}$, and they contain all the position and effect information of the epistasis between the jth and kth QTL. These properties provide the theoretical basis for mapping epistasis in ICIM.

Suppose that we have a sample of n individuals from a backcross population with observations on a quantitative trait of interest and $m+1$ ordered markers. The following linear regression model based on equation (4) can be used in QTL mapping,

$$y_i = b_0 + \sum_{j=1}^{m+1} b_j x_{ij} + \sum_{j<k} b_{jk} x_{ij} x_{ik} + e_i, \quad (5)$$

where y_i is the trait phenotypic value of the ith individual in the mapping population; b_0 is the overall mean of the linear model; x_{ij} is a dummy variable for the genotype of the ith individual at the jth marker, taking value 1 for homozygote of marker type, and -1 for heterozygote; b_j is the partial regression coefficient of the phenotype on the jth marker variable; b_{jk} is the partial regression coefficient of the phenotype on the multiplication variable of the jth and kth markers; and e_i is the residual random error which is assumed to be normally distributed.

Stepwise regression for coefficient estimation of markers and marker-pairs

The number of additive and epistatic QTL detectable for a trait of interest using a moderate size of mapping population, say 200 individuals, is less likely to be more than 20, which is much lower than the number of marker and marker pair variables. Thus, to identify the markers flanking these QTL is to correctly select the best model among all possible models, which is an issue of model selection (Broman and Speed 2002; Sillanpää and Corander 2002). A number of statistical methods are available to search through the space of models and various criteria can be used to select the best model (Miller 1990; Piepho and Gauch 2001). However, there is no universally best model selection method for all situations (Miller 1990). Here we consider using stepwise regression, but we do not exclude the possibility that other model selection methods may achieve similar performance in model selection and parameter estimation of model (5).

A two-stage stepwise regression strategy was adopted to determine the parameters in model (5). Significant marker variables in model (5) were selected in the first stage, which is similar to ICIM for additive mapping (Li et al. 2007). Then stepwise regression was applied to the residuals from the first stage to select significant marker pairs and estimate their effects in model (5). Stricter probability levels were applied in the second stage to avoid over-fitting since the number of regression variables is much larger.

Two-dimensional interval mapping or scanning for epistasis

When conducting two-dimensional scanning for epistatic QTL, there are two current testing intervals represented by $(j, j+1)$ and $(k, k+1)$, where $j < k$. The observation values in model (5) were adjusted by

$$\Delta y_i = y_i - \sum_{r\neq j,j+1,k,k+1} \hat{b}_r x_{ir} - \sum_{\substack{r\neq j,j+1 \\ s\neq k,k+1}} \hat{b}_{rs} x_{ir} x_{is}. \quad (6)$$

where $\hat{b}_r$ and $\hat{b}_{rs}$ are the estimates of b_r and b_{rs} in model (5), respectively. The adjusted phenotype Δy_i thus obtained contains the information of QTL in the two testing intervals, which includes two positions and two additive effects of individual QTL, and one epistatic effect between the two QTL, and at the mean time, the additive and epistatic effects of QTL located on other intervals and chromosomes are completely controlled. The adjusted observation Δy_i does not change until either of the two testing positions moves into a new interval.

Individuals in the mapping population can be classified into sixteen groups based on their marker types (Table 1). If there are two QTL (with the two alleles denoted as Q_j and q_j, and Q_k and q_k) at the two testing positions, Δy_i follows a mixture distribution consisting of four QTL genotypes: $Q_jQ_jQ_kQ_k$, $Q_jQ_jQ_kq_k$, $Q_jq_jQ_kQ_k$, and $Q_jq_jQ_kq_k$ (Table 1). The proportions of the four QTL genotypes for each marker type group can be defined from recombination frequencies (Table 1). Therefore, QTL at the current two mapping positions can be tested by the following hypotheses:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs.
H_A: at least two of μ_1, μ_2, μ_3 and μ_4 are not equal.

Then the log-likelihood function under the alternative hypothesis H_A is,

$$L_A = \sum_{j=1}^{16} \sum_{i\in S_j} \log \left[\sum_{k=1}^{4} f_{jk} f(\Delta y_i; \mu_k, \sigma^2) \right] \quad (7)$$

where S_j denotes the jth marker type group ($j = 1, \ldots, 16$), f_{jk} ($k = 1, \ldots, 4$) is the proportion of the kth QTL genotypes in the jth group (Table 1), and $f(\cdot;\mu_k, \sigma^2)$ represents the density of the kth normal distribution $N(\mu_k, \sigma^2)$.

The expectation and conditional maximization (ECM) algorithm (Dempster et al. 1977; Meng and Rubin 1993) was used to estimate the four means and one variance in equation (7). Since most individuals in groups 1, 4, 13, and 16 have QTL types $Q_jQ_jQ_kQ_k$, $Q_jQ_jQ_kq_k$, $Q_jq_jQ_kQ_k$, and $Q_jq_jQ_kq_k$, respectively (Table 1), the initial values of the five unknown parameters can be defined from these groups, i.e.,

$$\mu_1^{(0)} = \frac{1}{n_1} \sum_{i=1}^{n_1} \Delta y_i, \mu_2^{(0)} = \frac{1}{n_4} \sum_{i=\mathrm{sum}(n_1:n_3)+1}^{\mathrm{sum}(n_1:n_4)} \Delta y_i,$$

$$\mu_3^{(0)} = \frac{1}{n_{13}} \sum_{i=\mathrm{sum}(n_1:n_{12})+1}^{\mathrm{sum}(n_1:n_{13})} \Delta y_i, \mu_4^{(0)} = \frac{1}{n_{16}} \sum_{i=\mathrm{sum}(n_1:n_{15})+1}^{n} \Delta y_i,$$

$$\sigma^{2(0)} = \frac{1}{n_1+n_4+n_{13}+n_{16}} \left[\sum_{i=1}^{n_1} (\Delta y_i - \mu_1^{(0)})^2 + \sum_{i=\mathrm{sum}(n_1:n_3)+1}^{\mathrm{sum}(n_1:n_4)} (\Delta y_i - \mu_2^{(0)})^2 + \sum_{i=\mathrm{sum}(n_1:n_{12})+1}^{\mathrm{sum}(n_1:n_{13})} (\Delta y_i - \mu_3^{(0)})^2 + \sum_{i=\mathrm{sum}(n_1:n_{15})+1}^{n} (\Delta y_i - \mu_4^{(0)})^2 \right],$$

where $\mathrm{sum}(n_1{:}n_4)$ denotes the summation from n_1 to n_4, and so on. In the E-step, the posterior probability of the ith individual ($i = 1, \ldots, n$) belonging to the kth ($k = 1, \ldots, 4$) QTL genotype was calculated as,

$$w_{ik}^{(0)} = f_{jk} f\left(\Delta y_i; \mu_k^{(0)}, \sigma^{2(0)}\right) \Big/ \sum_{h=1}^{4} f_{jh} f\left(\Delta y_i; \mu_h^{(0)}, \sigma^{2(0)}\right),$$

where j denotes the marker type group into which the ith individual is classified. In the M-step, the five parameters were updated as,

$$\mu_k^{(1)} = \sum_{i=1}^{n} w_{ik}^{(0)} \Delta y_i \Big/ \sum_{i=1}^{n} w_{ik}^{(0)} \text{ for } k = 1, \cdots, 4, \quad \text{and}$$

$$\sigma^{2(1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{4} w_{ik}^{(0)} (\Delta y_i - \mu_k^{(1)})^2.$$

The EM algorithm continues until the difference in the likelihood between two consecutive iterations reaches a pre-assigned precision, say 10^{-6}. The ML estimates thus obtained are represented as $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4$ and $\hat{\sigma}^2$, from which two additive effects (a_j and a_k) of the putative QTL and their epistatic effect (aa_{jk}) can be estimated as follows,

$$a_j = \frac{1}{4}(\hat{\mu}_1 + \hat{\mu}_2 - \hat{\mu}_3 - \hat{\mu}_4),$$
$$a_k = \frac{1}{4}(\hat{\mu}_1 - \hat{\mu}_2 + \hat{\mu}_3 - \hat{\mu}_4), \text{ and}$$
$$aa_{jk} = \frac{1}{4}(\hat{\mu}_1 - \hat{\mu}_2 - \hat{\mu}_3 + \hat{\mu}_4)$$

Under the null hypothesis, H_0, all Δy_i follow the normal distribution of $N(\mu_0, \sigma_0^2)$. The mean and variance of this distribution can be estimated as,

$$\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^{n} \Delta y_i \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^{n} (\Delta y_i - \hat{\mu}_0)^2.$$

Thus, the log-likelihood function under the null hypothesis H_0 is,

Table 1 Marker types and their frequencies on the two current mapping intervals and the frequencies of QTL genotypes within each marker type

Marker type group	Frequency	Two pairs of flanking markers				QTL genotype			
		M_j x_j	M_{j+1} x_{j+1}	M_k x_k	M_{k+1} x_{k+1}	$Q_jQ_jQ_kQ_k$ $a_j + a_k + aa_{jk}$	$Q_jQ_j\ Q_kq_k$ $a_j - a_k - aa_{jk}$	$Q_jq_jQ_kQ_k$ $-a_j + a_k - aa_{jk}$	$Q_jq_jQ_kq_k$ $-a_j - a_k + aa_{jk}$
1	$\frac{1}{2}(1-r_{j,j+1})(1-r_{j+1,k})(1-r_{k,k+1})$	1	1	1	1	$p_1\ p_3$	$p_1\ (1-p_3)$	$(1-p_1)\ p_3$	$(1-p_1)(1-p_3)$
2	$\frac{1}{2}(1-r_{j,j+1})(1-r_{j+1,k})r_{k,k+1}$	1	1	1	−1	$p_1\ p_4$	$p_1\ (1-p_4)$	$(1-p_1)\ p_4$	$(1-p_1)(1-p_4)$
3	$\frac{1}{2}(1-r_{j,j+1})r_{j+1,k}r_{k,k+1}$	1	1	−1	1	$p_1(1-p_4)$	$p_1\ p_4$	$(1-p_1)(1-p_4)$	$(1-p_1)\ p_4$
4	$\frac{1}{2}(1-r_{j,j+1})r_{j+1,k}(1-r_{k,k+1})$	1	1	−1	−1	$p_1(1-p_3)$	$p_1\ p_3$	$(1-p_1)(1-p_3)$	$(1-p_1)\ p_3$
5	$\frac{1}{2}r_{j,j+1}r_{j+1,k}(1-r_{k,k+1})$	1	−1	1	1	$p_2\ p_3$	$p_2\ (1-p_3)$	$(1-p_2)\ p_3$	$(1-p_2)(1-p_3)$
6	$\frac{1}{2}r_{j,j+1}r_{j+1,k}r_{k,k+1}$	1	−1	1	−1	$p_2\ p_4$	$p_2\ (1-p_4)$	$(1-p_2)\ p_4$	$(1-p_2)(1-p_4)$
7	$\frac{1}{2}r_{j,j+1}(1-r_{j+1,k})r_{k,k+1}$	1	−1	−1	1	$p_2(1-p_4)$	$p_2\ p_4$	$(1-p_2)(1-p_4)$	$(1-p_2)\ p_4$
8	$\frac{1}{2}r_{j,j+1}(1-r_{j+1,k})(1-r_{k,k+1})$	1	−1	−1	−1	$p_2(1-p_3)$	$p_2\ p_3$	$(1-p_2)(1-p_3)$	$(1-p_2)\ p_3$
9	$\frac{1}{2}r_{j,j+1}(1-r_{j+1,k})(1-r_{k,k+1})$	−1	1	1	1	$(1-p_2)\ p_3$	$(1-p_2)(1-p_3)$	$p_2\ p_3$	$p_2(1-p_3)$
10	$\frac{1}{2}r_{j,j+1}(1-r_{j+1,k})r_{k,k+1}$	−1	1	1	−1	$(1-p_2)\ p_4$	$(1-p_2)(1-p_4)$	$p_2\ p_4$	$p_2(1-p_4)$
11	$\frac{1}{2}r_{j,j+1}r_{j+1,k}r_{k,k+1}$	−1	1	−1	1	$(1-p_2)\ (1-p_4)$	$(1-p_2)\ p_4$	$p_2\ (1-p_4)$	$p_2\ p_4$
12	$\frac{1}{2}r_{j,j+1}r_{j+1,k}(1-r_{k,k+1})$	−1	1	−1	−1	$(1-p_2)\ (1-p_3)$	$(1-p_2)\ p_3$	$p_2\ (1-p_3)$	$p_2\ p_3$
13	$\frac{1}{2}(1-r_{j,j+1})r_{j+1,k}(1-r_{k,k+1})$	−1	−1	1	1	$(1-p_1)\ p_3$	$(1-p_1)(1-p_3)$	$p_1\ p_3$	$p_1\ (1-p_3)$
14	$\frac{1}{2}(1-r_{j,j+1})r_{j+1,k}r_{k,k+1}$	−1	−1	1	−1	$(1-p_1)\ p_4$	$(1-p_1)(1-p_4)$	$p_1\ p_4$	$p_1\ (1-p_4)$
15	$\frac{1}{2}(1-r_{j,j+1})(1-r_{j+1,k})r_{k,k+1}$	−1	−1	−1	1	$(1-p_1)\ (1-p_4)$	$(1-p_1)\ p_4$	$p_1\ (1-p_4)$	$p_1\ p_4$
16	$\frac{1}{2}(1-r_{j,j+1})(1-r_{j+1,k})(1-r_{k,k+1})$	−1	−1	−1	−1	$(1-p_1)\ (1-p_3)$	$(1-p_1)\ p_3$	$p_1\ (1-p_3)$	$p_1\ p_3$

Notes: "1" and " −1" denote homozygote and heterozygote marker genotypes, respectively. $p_1 = (1 - r_{j,q_j})(1 - r_{q_j,j+1})/(1 - r_{j,j+1})$, $p_2 = (1 - r_{j,q_j})r_{q_j,j+1}\ /r_{j,j+1}$, $p_3 = (1 - r_{k,q_k})(1 - r_{q_k,k+1})/(1 - r_{k,k+1})$, and $p_4 = (1 - r_{k,q_k})r_{q_k,k+1}\ /r_{k,k+1}$. $r_{\cdot,\cdot}$ is the recombination frequency between two markers or between one marker and one QTL indicated by subscribes. $r_{\cdot,\cdot} = 0.5$ if the two markers are located on two chromosomes

$$L_0 = \sum_{i=1}^{n} \log[f(\Delta y_i; \hat{\mu}_0, \hat{\sigma}_0^2)].$$

The LOD score (denoted by LOD_A) at the current testing positions can be calculated from the log-likelihoods under the two hypotheses, i.e., $L_A - L_0$. Therefore, LOD_A can be used to test whether there is a significant difference among the four QTL genotypes. As Δy_i contains the information of QTL positions, additive and epistatic effects in the two testing intervals, both additive and epistatic effects affect LOD_A. In order to test the presence of epistasis, the influence of additive QTL on LOD score needs to be removed, and another alternative hypothesis H_{AA} is needed for this purpose, i.e.,

H_{AA}: $\mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$, or $aa_{jk} = 0$.

The difference in ML between H_{AA} and H_A represents the net contribution from epistatic effect. The ML estimates under H_{AA} was calculated by the conditional maximum of L_A. Let $L_{AA} = L_A - \lambda(\mu_1 - \mu_2 - \mu_3 + \mu_4)$, where λ is the Lagrange multiplier. In the EM algorithm, the calculation of posterior probabilities was the same as previous one. In the M-step, the five parameters were updated as follows,

$$(\lambda\sigma^2)^{(0)} = \left[\sum_{i=1}^{n} w_{i1}^{(0)} \Delta yi \Big/ \sum_{i=1}^{n} w_{i1}^{(0)} - \sum_{i=1}^{n} w_{i2}^{(0)} \Delta yi \Big/ \sum_{i=1}^{n} w_{i2}^{(0)} - \sum_{i=1}^{n} w_{i3}^{(0)} \Delta yi \Big/ \sum_{i=1}^{n} w_{i3}^{(0)} + \sum_{i=1}^{n} w_{i4}^{(0)} \Delta yi \Big/ \sum_{i=1}^{n} w_{i4}^{(0)}\right] \Big/ \sum_{k=1}^{4}\left[1 \Big/ \sum_{i=1}^{n} w_{ik}^{(0)}\right],$$

$$\mu_1^{(1)} = \left[\sum_{i=1}^{n} w_{i1}^{(0)} \Delta y_i - (\lambda\sigma^2)^{(0)}\right] / \sum_{i=1}^{n} w_{i1}^{(0)},$$

$$\mu_2^{(1)} = \left[\sum_{i=1}^{n} w_{i2}^{(0)} \Delta y_i + (\lambda\sigma^2)^{(0)}\right] / \sum_{i=1}^{n} w_{i2}^{(0)},$$

$$\mu_3^{(1)} = \left[\sum_{i=1}^{n} w_{i3}^{(0)} \Delta y_i + (\lambda\sigma^2)^{(0)}\right] / \sum_{i=1}^{n} w_{i3}^{(0)}, \text{ and}$$

$$\mu_4^{(1)} = \left[\sum_{i=1}^{n} w_{i4}^{(0)} \Delta y_i - (\lambda\sigma^2)^{(0)}\right] / \sum_{i=1}^{n} w_{i4}^{(0)}.$$

The LOD score (denoted by LOD_{AA}) calculated by $L_A - L_{AA}$ indicates whether there is a significant interaction at the two testing positions. It is worth noting that the EM algorithms described above have a fast convergence speed. The precision approaches 10^{-6} within at most 10 iterations for any testing positions.

Calculation of genetic variance under linkage and digenic epistasis

The theoretical additive variance of the genetic value G in model (1) is,

$$V_A = \mathrm{Var}\left(\sum_{j=1}^{m} a_j g_j\right) = \sum_{j,k=1}^{m} \mathrm{Cov}(g_j, g_k) a_j a_k = \sum_{j,k=1}^{m} (1 - 2r_{jk}) a_j a_k, \quad (8)$$

where r_{jk} is the recombinant frequency between the jth and kth QTL. The theoretical epistatic variance in model (1) is,

$$V_I = \mathrm{Var}(\sum_{j<k} aa_{jk} g_j g_k) = \sum_{j<k,l<m} \mathrm{Cov}(g_j g_k, g_l g_m) aa_{jk} aa_{lm} = \sum_{j<k,l<m} \left[(1 - 2r_{jl})(1 - 2r_{km}) - (1 - 2r_{jk})(1 - 2r_{lm})\right] \times aa_{jk} aa_{lm}, \quad (9)$$

where r_{jk}, r_{jl}, r_{km} and r_{lm} are the recombinant frequencies between the jth and kth QTL, between the jth and lth QTL, between the kth and mth QTL, and between the lth and mth QTL, respectively. It can be proved that $\mathrm{Cov}(g_j\ g_k,\ g_l\ g_m) = 0$ in equation (9) if $l \geq k$ and $m \leq j$. Equations (8) and (9) can be used to evaluate the relative importance of epistatic variance for any defined genetic models containing the additive effects and digenic epistasis or after a QTL mapping study.

Genetic models used in simulation studies

Three hypothetical genomes were used in our simulation studies. For consistency, the additive effect is defined as half of the difference between two QTL genotypes in this study, i.e., QQ and Qq in B_1 ($F_1 \times P_1$), Qq and qq in B_2 ($F_1 \times P_2$), and QQ and qq in DH or recombination inbred lines (RIL).

The first genome consisted of six chromosomes, each of 150 cM in length and with 16 evenly distributed markers. Ten predefined QTL similar to those in Zeng (1994) (represented by QZ1–QZ10; Table 2) were assumed to contribute to the trait of interest. Three QTL were located on each of the first three chromosomes, and one QTL on the fourth chromosome. There was no QTL on chromosomes 5 and 6. The locations and genetic effects (additive and epistatic effects) of the ten simulated QTL are shown in Table 2. Under this QTL distribution, the theoretical additive and epistatic variances were both equal to 4.67, calculated using equations (8) and (9). Heritability in the

Table 2 Additive, and additive by additive epistatic effects of ten simulated QTL in genome 1

	Chromosome	1	1	1	2	2	2	3	3	3	4
	Position (cM)	16	48	108	3	43	77	33	68	129	26
QTL		QZ1	QZ2	QZ3	QZ4	QZ5	QZ6	QZ7	QZ8	QZ9	QZ10
QZ1		0.00									
QZ2			0.51								
QZ3				0.40							
QZ4					0.70						
QZ5			−0.56			−0.84					
QZ6							−0.86				
QZ7		−0.90						0.00			
QZ8									1.10		
QZ9					−0.77			1.27		0.60	
QZ10		0.98				−0.64					0.53

Both additive variance (V_A) and interaction variance (V_I) were equal to 4.67. The error variance (V_e) was calculated by $V_e = (V_A + V_I)(1 - H)/H$, where H is heritability in the broad sense. H was set to 0.8 in our simulation study, so the error variance was 2.34. There are totally six digenic epistasis, i.e., QZ1 × QZ7, QZ1 × QZ10, QZ2 × QZ5, QZ4 × QZ9, QZ5 × QZ10 and QZ7 × QZ9

broad sense for the trait of interest was set at 0.8. One hundred backcross populations each of 200 individuals were simulated.

The second genome consisted of three chromosomes, each with 100 cM in length and 11 evenly distributed markers. Three QTL (represented by QB1–QB3; for details see Boer et al. 2002) contributed to the expression of a quantitative trait of interest. QB1 is located at 35 cM on chromosome 1, QB2 at 53 cM on chromosome 2, and QB3 at 22 cM on chromosome 3. For this genome, we considered two genetic models corresponding to Set I ($V_A = 0.375$, $V_I = 0.375$ and $H = 0.6$) and Set III ($V_A = 0$, $V_I = 0.375$ and $H = 0.3$) of Boer et al. (2002), respectively. Considering that the indicator variable has value 0.5 for homozygous marker type, and −0.5 for heterozygous marker type in Boer et al. (2002), the additive and epistatic effects used in our simulation study were respectively half and quarter of those in Boer et al. (2002), so as to achieve the same additive and epistatic variances. One hundred backcross populations each of 200 individuals were simulated for this genome.

The third genome consisted of four 100 cM chromosomes. Eleven markers on each chromosome were positioned as shown at the numerical labels of the horizontal axis of Fig. 3 in Yi et al. (2003). Seven QTL (represented by QY1–QY7) with epistatic patterns controlled the expression of a quantitative trait of interest, for details see Table 1 in Yi et al. (2003). Similar to Yi et al. (2003), the residual variance σ_e^2 was adjusted to 1, and one hundred backcross populations each of 300 individuals were simulated.

For each simulated populations, the number of total markers is far less than its population size. As a result, in the first stage of stepwise regression the largest P value for entering variables (PIN_1) was set at 0.05 and the smallest P value for removing variables ($POUT_1$) was twice of PIN_1, which is normally used in most stepwise regressions. In the second stage, considering the increasing speed of regression variables PIN was set as the square of PIN_1, i.e., $PIN_2 = PIN_1^2 = 0.0025$, and $POUT_2$ was twice of PIN_2.

LOD score and effect estimation in simulation

For calculating LOD score and QTL effect in a simulation study, a confidence interval for each predefined QTL is normally specified, and then simulation runs that have significant higher peaks along the LOD profile in the confidence interval are counted. In this case, the QTL effects are normally over-estimated. Unbiased estimation can be achieved if all runs with peaks in the confidence interval are counted, no matter whether the peaks are higher than the LOD threshold (Zeng 1994; Li et al. 2007). Similar methodology can be adopted for epistasis mapping, i.e., the power analysis is conducted based on the predefined two dimensional confidence interval for each interacting QTL. For simplicity, the LOD score and QTL effect were calculated for each scanned chromosomal position by averaging the 100 simulation runs. It is expected that the QTL effect is under-estimated since estimates from non-significant LOD scores are also counted.

The barley DH population

One real population was used, which was derived from a two-row barley (*Hordeum vulgare* L.) cross, Harrington × TR306, and consists of 145 random DH lines (Tinker et al. 1996). A subset of 127 markers was used to build a

base map with relatively uniform coverage. Data for seven agronomic traits were collected in 1992 and/or 1993 at 17 locations, and the average kernel weight (KWT) across 25 environments was used as the phenotypic data for QTL mapping in this study. The average KWT was 38.7 mg for Harrington, and 45.0 mg for TR306. The minimum, mean and maximum KWT of the 146 DH were 35.8, 42.0, and 48.1 mg, respectively.

Results

Simulation results from genome 1

As previously indicated, both additive and epistasis information were contained in LOD_A. Clear signs can be seen in the LOD contour profile when additive or epistatic effect or both effects were present (Fig. 1A). Genome regions with additive QTL showed clear light bands on both axes in Fig. 1A. The LOD score shown in Fig. 1B has excluded the influence of additive effect, and clearly indicates the six predefined epistatic QTL. The LOD scores at the other positions are close to 0. For QTL with additive effect, the LOD score was significantly higher at the QTL position, such as QZ8, QZ6, QZ5, QZ9, QZ4 and QZ10. The additive effects of QZ2 and QZ3 cannot be clearly seen due to their relatively small effects (Table 2, Fig. 1A). QZ1 and QZ7 do not have any additive effects, and therefore the LOD scores at these two positions were close to zero (Fig. 1A).

QZ6 and QZ8 showed significant additive effects in both one- and two-dimensional LOD profiles (Fig. 1A). But there was no clear signs in Fig. 1B indicating their significant interactions with other QTL. Therefore, QZ6 and QZ8 can be viewed as QTL with significant additive effects but no significant epistatic effects with other QTL. The interactions of QZ5 × QZ10 and QZ4 × QZ9 can be clearly seen in Fig. 1B. The high LOD scores at these QTL positions indicated they have significant additive effects (Fig. 1A). Thus, QZ4, QZ5, QZ9, and QZ10 have both additive and epistatic effects. The two-dimensional LOD profile in Fig. 1B showed the presence of interaction QZ2 × QZ5, but the LOD profile in Fig. 1A only indicated the presence of additive effect of QZ5, explaining 6.05% of the phenotypic variance (PVE; Table 3). No additive effect was evident for QZ2 due to its relatively small effect (PVE=2.23%; Table 3). There were clearly signs in Fig. 1B for interactions of QZ1 × QZ10 and QZ7 × QZ9. But in both one- and two-dimensional LOD profiles (Fig. 1A), clear peaks only appeared at chromosomal regions around QZ9 and QZ10, since the additive effects of QZ1 and QZ7 were set at zero. In these two interactions, only one QTL has additive effect. For interaction of QZ1 × QZ7, both QTL have no additive effects (Table 2). The LOD profile in Fig. 1B demonstrated the existence of this interaction (LOD_{AA}=6.16), but no additive effects at the chromosomal regions around QZ1 and QZ7 were detected from one-dimensional scanning (Fig. 1A).

Fig. 1 Two-dimensional average LOD contour profiles testing the significance of additive and epistasis (**a**), and epistasis only (**b**) under genome 1 (Table 2). The number of simulation runs is 100. The one-dimensional profile on each axis is the average LOD score testing the significance of additive effects. The size and direction of each arrow approximately represent the effect size and direction of the pointed QTL, respectively. QTL without arrows have no additive effects. Predefined digenic epistasis were indicated by *text boxes*. LOD score testing the significance of either additive and epistasis (**a**), or epistasis only (**b**) at the position of each network was shown in each box ▶

There is a trend that QTL with larger additive effect results in higher LOD score (Table 3). However, this was not always the case, especially when multiple QTL were linked. For example, QZ4 has a larger effect than QZ9 but lower LOD score than QZ9 (Table 3). QZ4 is linked with QZ5 in the repulsion phase, while QZ9 is linked with QZ8 in the coupling phase (Table 2). Therefore, QTL linked in coupling with the target QTL may increase the detection efficiency of the target QTL. Similar trend can be seen for epistasis (Table 3). QZ7 × QZ9 explains 13.82% of the phenotypic variation and the LOD score at this position reaches 13.90. QZ2 × QZ5 explains 2.69% of the phenotypic variation and the LOD score at this position reaches 2.50. Thus, it is expected that QZ7 × QZ9 can be more easily detected than QZ2 × QZ5.

The estimation of epistatic effects were close to zero in most genomic regions in the effect profile except around the six predefined interactions (Fig. 2).The epistatic effects were consistently under-estimated since they were estimated by the mean effect across all simulation runs (Fig. 2). Additive effects were also under-estimated. For example, the additive effect of QZ2 was estimated as 0.34 in one-dimensional additive effect profile in Fig. 2, while the true effect was 0.51 (Table 2). The true effects of QZ6 and QZ8 were −0.86 and 1.10, while the mean effects across 100 simulation runs were −0.82 and 1.02, respectively. Unbiased estimation could be achieved for additive QTL effects if all peaks in a predefined confidence interval were counted (Li et al. 2007). We expect that unbiased estimation of epistatic QTL effects can be also achieved in the same manner. It should be noted that the directions of all epistatic effects were correctly detected (Fig. 2).

Simulation results from genome 2

The ICIM achieved satisfying results for both additive and epistasis mapping under genome 2. Three peaks appeared around the predefined interacting QTL for Set I (Fig. 3A). The mean LOD_{AA} were high around the three interactions,

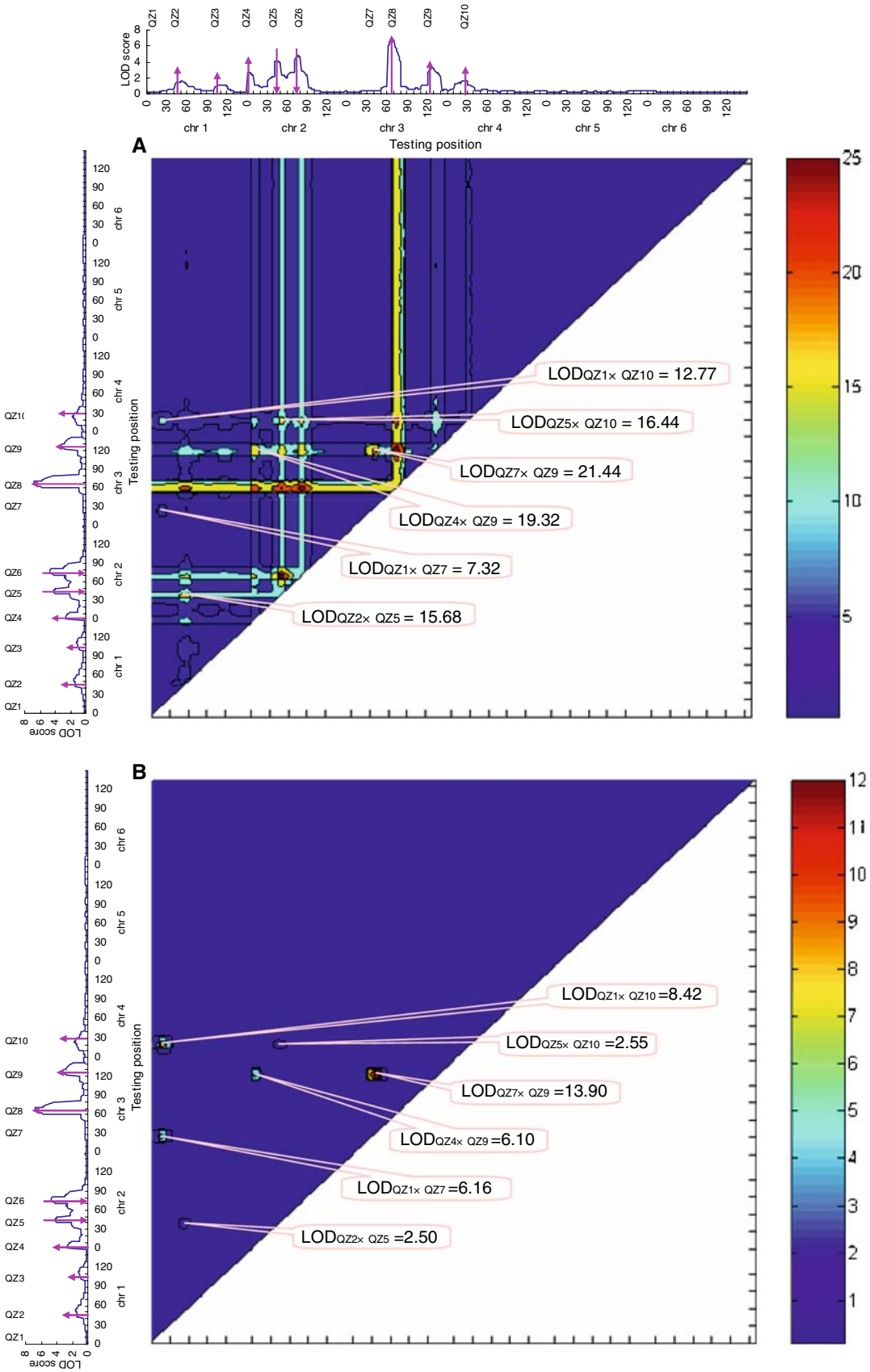
QZ1
QZ2
QZ3
QZ4
QZ5
QZ6
QZ7
QZ8
QZ9
QZ10
LOD score
chr 1
chr 2
chr 3
chr 4
chr 5
chr 6
Testing position
A
LODQZ1× QZ10 = 12.77
LODQZ5× QZ10 = 16.44
LODQZ7× QZ9 = 21.44
LODQZ4× QZ9 = 19.32
LODQZ1× QZ7 = 7.32
LODQZ2× QZ5 = 15.68
B
LODQZ1× QZ10 =8.42
LODQZ5× QZ10 =2.55
LODQZ7× QZ9 =13.90
LODQZ4× QZ9 =6.10
LODQZ1× QZ7 =6.16
LODQZ2× QZ5 =2.50

Table 3 Percentage of phenotypic variance explained (PVE) by the additive effect of individual QTL and the epistatic effect of digenic interacting QTL, and the corresponding mean LOD score across the 100 simulation runs

Additive QTL	QZ1	QZ2	QZ3	QZ4	QZ5	QZ6	QZ7	QZ8	QZ9	QZ10
PVE (%)	0.00	2.23	1.37	4.20	6.05	6.34	0.00	10.37	3.09	2.41
LOD in additive mapping	0.24	1.41	1.04	2.63	4.15	4.75	0.33	6.82	3.24	1.65

Interacting QTL	QZ1 × QZ7	QZ1 × QZ10	QZ2 × QZ5	QZ4 × QZ9	QZ5 × QZ10	QZ7 × QZ9
PVE (%)	6.94	8.23	2.69	5.08	3.51	13.82
LOD_{AA}	6.16	8.42	2.50	6.10	2.55	13.90

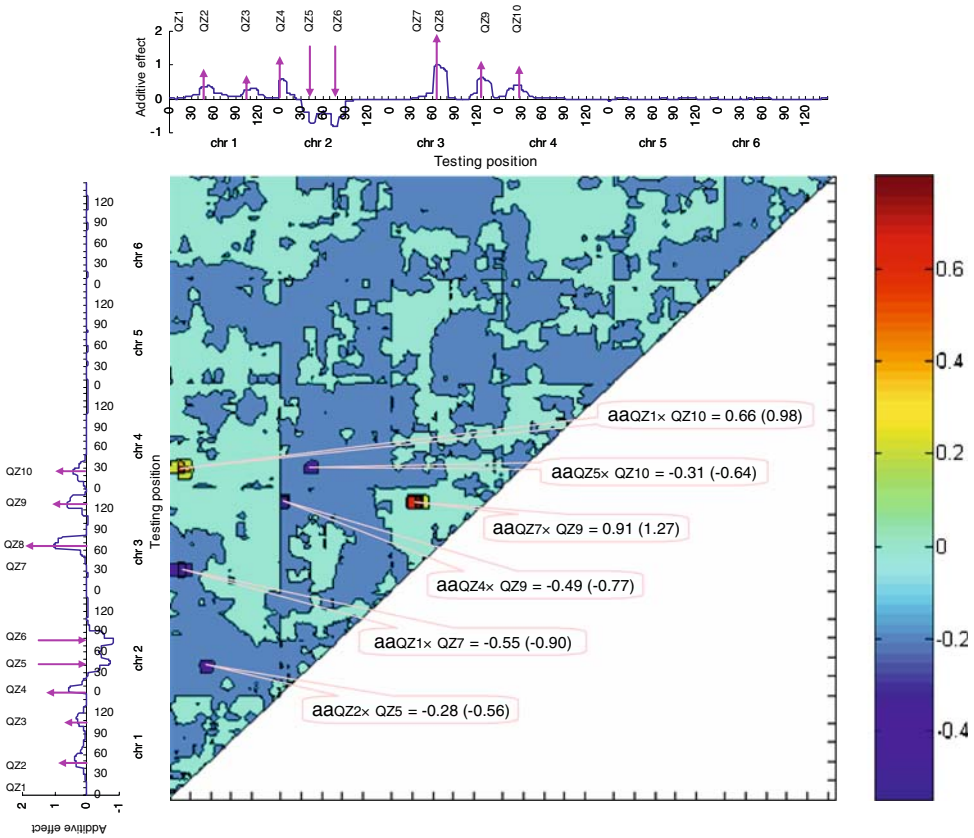

Fig. 2 Average epistatic effect profile for genome 1. The number of simulation runs is 100. The one-dimensional profile on each axis is the average additive effect. The size and direction of each arrow approximately represent the effect size and direction of the pointed QTL, respectively. QTL without arrows have no additive effects. Predefined digenic epistasis were indicated by *text boxes*. Estimated additive by additive epistatic effect was shown in each box. True epistatic effect was given in *parentheses*

but close to zero in other regions (Fig. 3B). In the two-dimensional LOD profile (Fig. 3A), high LOD score can be seen at the position of QB3 due to its large additive effect. Since the additive effects of QB1 and QB2 were only half of the QB3 effect, the LOD scores at the positions of QB1 and QB2 in Fig. 3A were low, which was consistent with the results from additive mapping (see the one-dimensional LOD profile on the axes of Fig. 3A, B). Since the three QTL in genome 2 are not linked, the additive effect was less under-estimated in genome 2 compared with genome 1 (Figs. 2, 3C). The epistatic effects were under-estimated, as expected (Fig. 3C).

For Set III, the three defined QTL have no additive effects. In this case, LOD_A was only affected by epistatic effect (Fig. 3D). Therefore, both Fig. 3D and E indicate the three defined interactions. The additive mapping results also indicate there is no significant additive QTL along the genome (see the one-dimensional LOD profile on the axes of Fig. 3D, E). These results suggest that epistasis has little influence on additive QTL mapping of ICIM, and vice versa. The mean epistatic effects for the three identified interactions were −0.43, −0.18 and 0.17 for QB1 × QB2, QB1 × QB3, and QB2 × QB3, respectively, which were all under-estimated (Fig. 3F).

Simulation results from genome 3

Similar results to the first two genomes were achieved for genome 3 (Fig. 4). Three pairs of epistatic QTL were

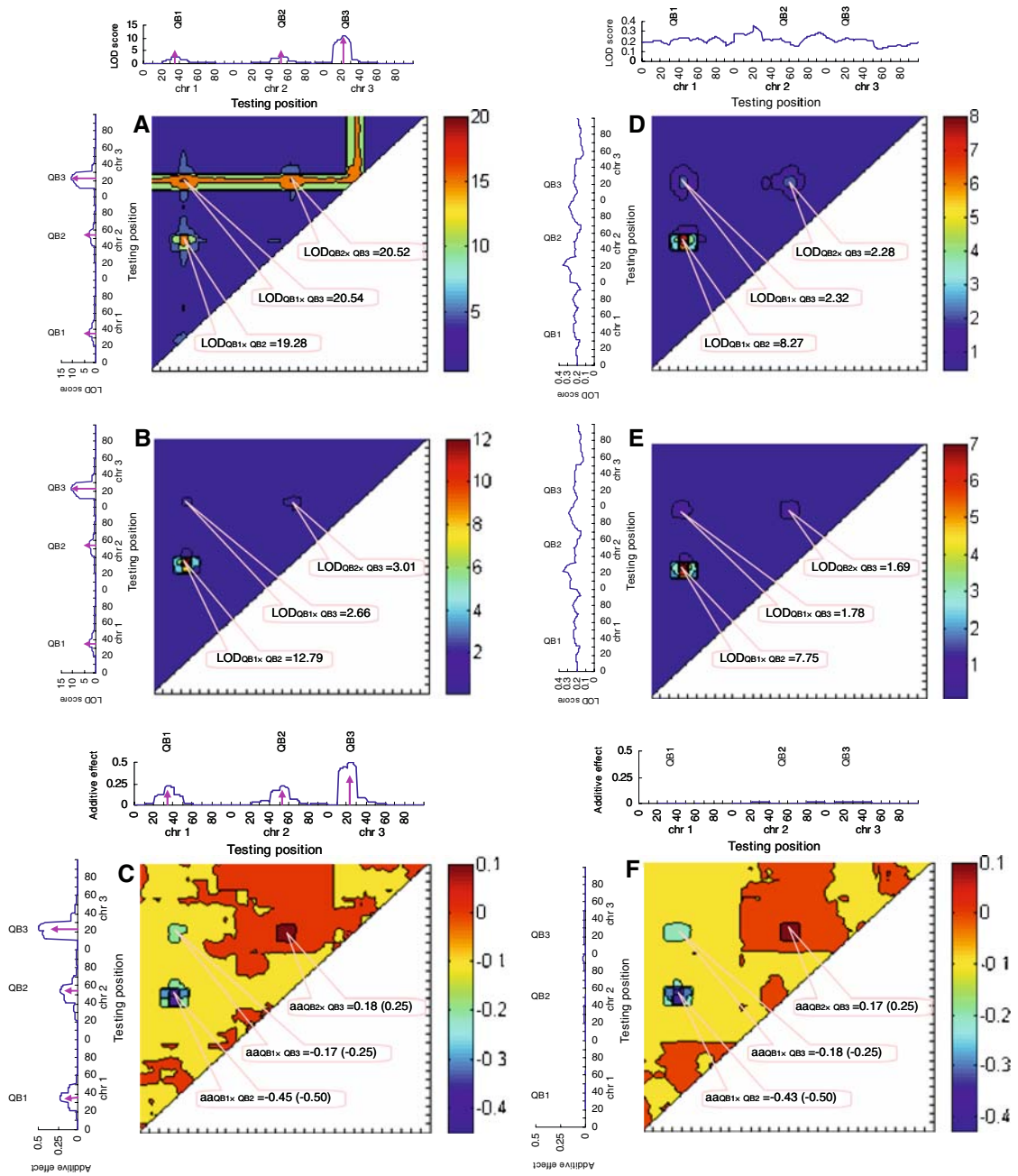

Fig. 3 Two-dimensional average LOD contour profiles testing the significance of additive and epistasis (**a** and **d**), and epistatic only (**b** and **e**), and average epistatic effect profile (**c** and **f**) for genome 2. The number of simulation runs is 100. On the coordinate axes of the two-dimensional average LOD contour profiles are the one-dimensional average LOD profiles testing the significance of additive effects. On the coordinate axes of the two-dimensional average epistatic effect profiles are the one-dimensional average additive effect profiles. The size and direction of each arrow approximately represent the effect size and direction of the pointed QTL, respectively. QTL without arrows have no additive effects. Predefined digenic epistasis were indicated by *text boxes*. LOD score testing the significance of either additive and epistasis (**a** and **d**) or epistasis (**b** and **e**), or estimated additive-by-additive epistatic effect (**c** and **f**) was shown in each box. True epistatic effect was given in *parentheses*

clearly identified through the two-dimensional scanning (Fig. 4B). Similar LOD scores were obtained for all the three pairs (Fig. 4B), since they explained the same amount of phenotypic variation. The high LOD_{AA} value for the identified interactions (11.23, 13.76, and 14.90 for QY1 × QY2, QY3 × QY4, and QY5 × QY6, respectively) indicated high detection power (Fig. 4B). The three interacting chromosomal regions showed different LOD scores in Fig. 4A due to their different additive effects. All the four additive QTL were clearly identified from additive

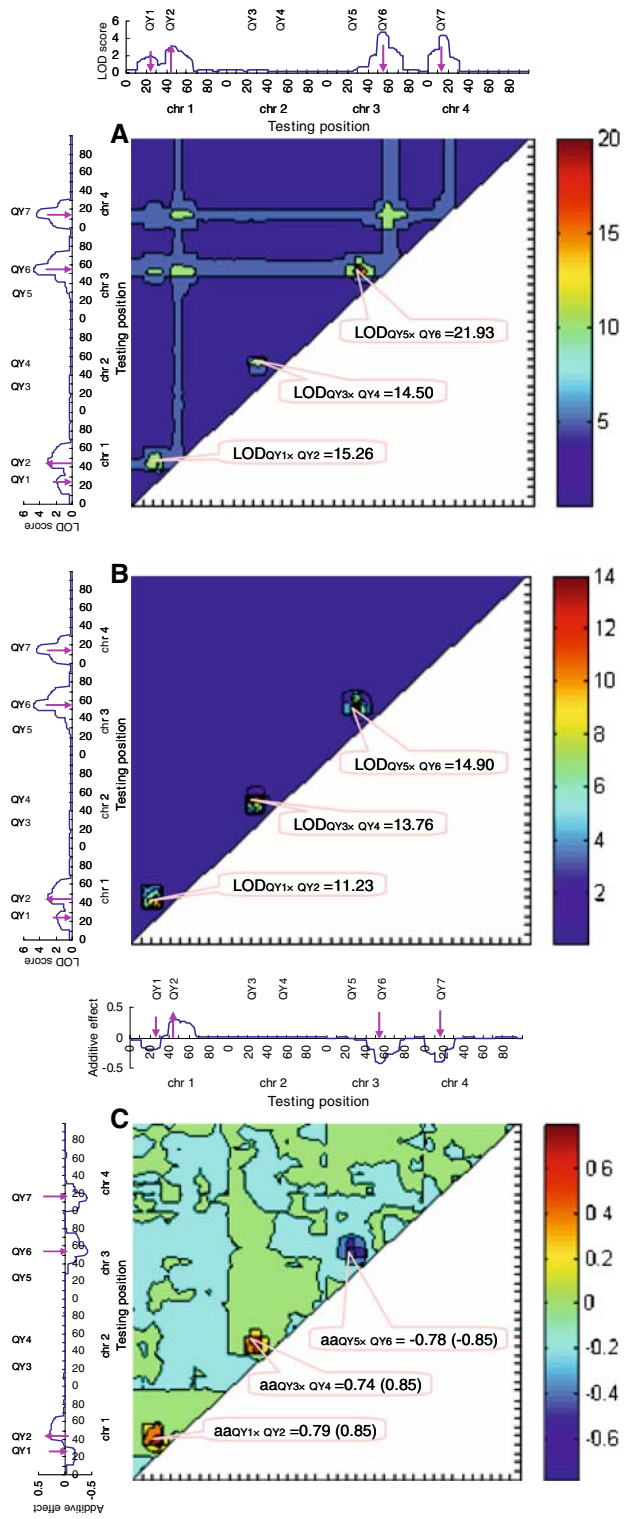

Fig. 4 Two-dimensional average LOD contour profiles testing the significance of additive and epistasis (**a**) and epistasis only (**b**), and average epistatic effect profile (**c**) for genome 3. The number of simulation runs is 100. On the coordinate axes of the two-dimensional average LOD contour profiles are the one-dimension average LOD profiles testing the significance of the additive effects. On the coordinate axes of the two-dimensional average epistatic effect profiles are the one-dimensional average additive effect profiles. The size and direction of each arrow approximately represent the effect size and direction of the pointed QTL, respectively. QTL without arrows have no additive effects. Predefined digenic epistasis are indicated by *text boxes*. LOD score testing the significance of either additive and epistasis (A) or epistasis (B), or estimated additive by additive epistatic effect (C) was shown in each box. True epistatic effect was given in *parentheses*

mapping and the estimated additive effects were almost identical to the predefined effects (see the one-dimensional additive effect profile on the axes of Fig. 4C). Similar to genomes 1 and 2, the effects of the three identified interactions were under-estimated (Fig. 4C).

Comparison of ICIM with MIM using simulated populations

To demonstrate the efficiency of ICIM in QTL mapping compared with other methods, we used one mapping population from the first simulation run in genome 1 as an example (Fig. 5). For this population, MIM detected one QTL on chromosome 1, two each on chromosomes 2 and 3, and one each on chromosomes 4 and 5 (Fig. 5A). LOD scores from MIM were much lower than those from ICIM. ICIM detected six QTL with LOD scores more than 2.0. In addition, the estimated additive effects from MIM were more biased than those from ICIM (Fig. 5B, D; Table 2). MIM reported one interaction at 80 cM of chromosome 1 and 20 cM of chromosome 4. In comparison, five interactions (i.e., epi1 for QZ1 × QZ10, epi2 for QZ1 × QZ7, epi3 for QZ2 × QZ5, epi4 for QZ5 × QZ10, and epi5 for QZ7 × QZ9) were identified with LOD scores more than 3.0 by ICIM (Fig. 5E, F). The positions of interacting QTL were biased in both ICIM and MIM (Fig. 5E; Table 2), but the epistatic effects were nearly unbiased in ICIM (Fig. 5F; Table 2). A couple of false interacting QTL appeared on the LOD contour profile of Fig. 5E, but they can be avoided by increasing the LOD threshold.

When the two QTL having no additive effects are interacting, epistasis can be significant (Carlborg and Haley 2004). Such epistasis is difficult to detect using MIM and other one-dimensional epistatic mapping methods. In comparison, ICIM can identify epistatic QTL no matter whether the two interacting QTL have any additive effects. For example, QZ1 and QZ7 defined in genome 1 have no additive effects, but the additive by additive epistatic effect is −0.90 (Table 2) which explains 6.94% of phenotypic variation (Table 3). In our two-dimensional scanning, the average LOD_{AA} corresponding to these two QTL reached 6.16 (Table 3), which indicates they can be easily identified. For the simulated population shown in Fig. 5, ICIM reported a LOD score of 9.7 at this interaction, but MIM failed to detect this epistasis.

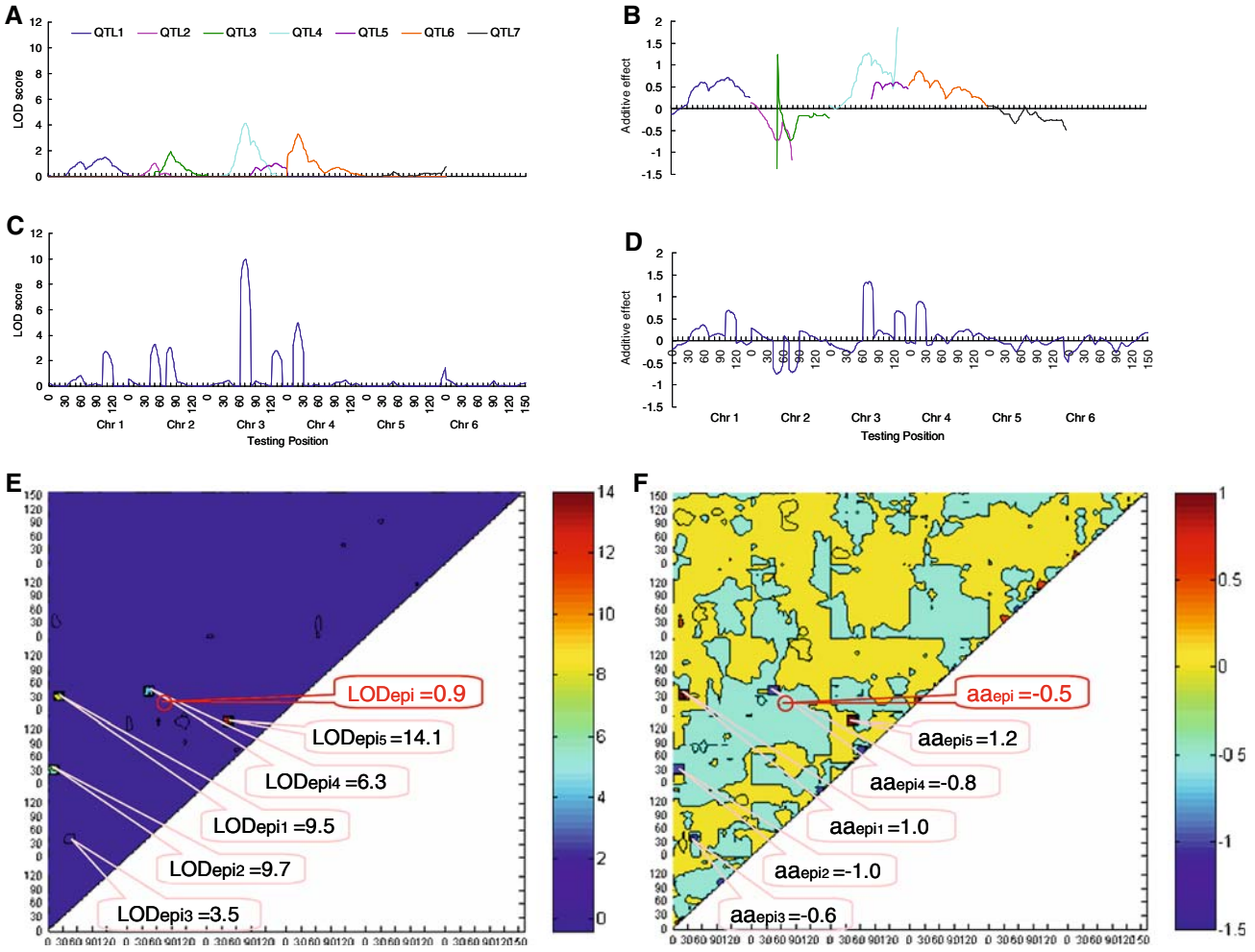

Fig. 5 Additive and epistatic QTL mapping of MIM and ICIM using one simulated backcross population of 200 individuals from the first genome. The LOD score and additive effect profiles from MIM were shown in **a** and **b**, and those from ICIM were shown in **c** and **d**. Two-dimensional LOD score and epistatic effect contour profiles of ICIM were shown in **e** and **f**. Significant QTL interactions were indicated by *text boxes* in **e** and **f**, and *red boxes* indicated the interactions detected by MIM, otherwise by ICIM

Using simulated populations and Set III from the second genome we found that MIM only detected some false positive QTL and failed to identify any interactions. In comparison, ICIM was able to detect these interacting QTL with reasonable powers (Fig. 3D, E). Theoretically, MIM should be the optimal method in terms of accuracy. ICIM only improves the computational efficiency (less computing time). The less efficiency in terms of accuracy of QTL detection for MIM may be due to convergence to a local optimum rather than global optimum, and the implementation algorithm used in MIM.

Mapping results for the barley DH population

Since the number of total markers and the size of the barley population are fairly comparable, stricter probabilities should be adopted to avoid over-fitting of marker variables in the stepwise regression. When PIN_1 was set at 0.01 and $POUT_1$ was twice of PIN_1, nine additive QTL for KWT (denoted by qKWT1 to qKWT9) were identified to be distributed on five of the seven barley chromosomes by ICIM under the LOD threshold of 2.5 (Table 4), eight of which could be seen from the LOD profile in Tinker et al. (1996). qKWT7 (explaining 38.37% of the phenotypic variance (PVE)) located at 5.0 cM on chromosome 5H and qKWT9 (PVE=17.20%) located at 95.0 cM on chromosome 7H are the two largest additive QTL (Table 4). Different probability levels for PIN_1 and $POUT_1$ had some effects on smaller QTL such as qKWT4 on 3H, and qKWT6 on 4H. But large effect QTL such as qKWT7, qKWT8, and qKWT9 were less affected by the probability levels (Fig. 6).

Some QTL in Table 4 were located more closely to individual markers, such as qKWT3, qKWT4 and qKWT6, while some QTL were located in the middle of the two

Table 4 Nine additive QTL identified by ICIM (PIN = 0.01, POUT = 0.02) to control KWT in the barley DH population

Chromosome	Left marker ID (marker name) QTL name right marker ID (marker name)	Marker or QTL position (cM)	LOD score	Additive effect (mg)	PVE (%)
2H	2H19 (MWG520A)	74.3			
	qKWT1	83.0	4.60	0.39	3.13
	2H20 (Pox)	90.1			
2H	2H26 (ABC620)	130.9			
	qKWT2	140.0	7.23	−0.51	5.34
	2H27 (MWG882)	142.1			
2H	2H29 (ABG317)	195.4			
	qKWT3	201.0	5.59	0.43	3.77
	2H30 (ABG609A)	201.7			
3H	3H33 (ABC171)	0.0			
	qKWT4	1.0	4.39	−0.39	3.04
	3H34 (CDO395)	5.7			
3H	3H35 (ABG471)	17.2			
	qKWT5	22.0	7.41	0.51	5.33
	3H36 (Ugp2)	25.2			
4H	4H57 (MWG655C)	124.8			
	qKWT6	125.0	4.12	−0.37	2.73
	4H58 (ABG366)	140.1			
5H	5H62 (Act8B)	3.8			
	qKWT7	5.0	34.28	−1.37	38.37
	5H63 (MWG502)	7.0			
7H	7H108 (iPgd1A)	3.4			
	qKWT8	4.0	8.27	−0.55	6.07
	7H109 (BCD129)	7.6			
7H	7H118 (MWG626)	92.6			
	qKWT9	95.0	19.81	−0.92	17.20
	7H119 (VAtp57A)	97.7			
Total phenotypic variation explained by additive effects (%) (represented by R^2 in the regression of phenotype on markers)					80.76

Marker ID is represented by the barley chromosome name followed by a number from 1 to 127

flanking intervals, such as qKWT1, qKWT7 and qKWT9. In the later case, both flanking marker variables will have significant coefficients in the regression model. Assuming each significant marker variables corresponding to one QTL, therefore, it is understandable that Xu and Jia (2007) reported 13 additive QTL for KWT using the same population. Under PIN_1=0.01 and $POUT_1$ = 0.02, main effects of significant marker variables in the regression model can explain 80.76% of the phenotypic variance (Table 4), which was higher than the heritability of KWT, i.e. 0.71 (Tinker et al. 1996). Thus, in this population additive variation should be the major component of genetic variation.

When PIN_2 =0.001 and $POUT_2$ = 0.002 were used in the second stage stepwise regression, R^2 was almost equal to 1 indicating the over-fitting of regression variables. To demonstrate the effect of marker inclusion and exclusion criteria in stepwise regression on digenic epistasis detection, two stricter probability levels, i.e. PIN_2 =0.0005 and $POUT_2$ = 0.0010 (Fig. 7A, B), and PIN_2 =0.0001 and $POUT_2$ = 0.0002 (Fig. 7C, D), were considered. As both additive and epistatic effects contribute to LOD_A, the seven largest additive QTL (i.e., qKWT2–qKWT5 and qKWT7–qKWT9; Table 4) were clearly observed from the LOD_A contour profiles (Fig. 7A, C). When the additive effects were excluded from LOD_A, few interactions showed LOD_{AA} over 3.0 (Fig. 7B, D), confirming the less importance of epistasis for KWT in the barley population.

Discussion

Model selection in ICIM

Two steps are involved in ICIM. In the first step, the best regression model is selected, which properly identifies

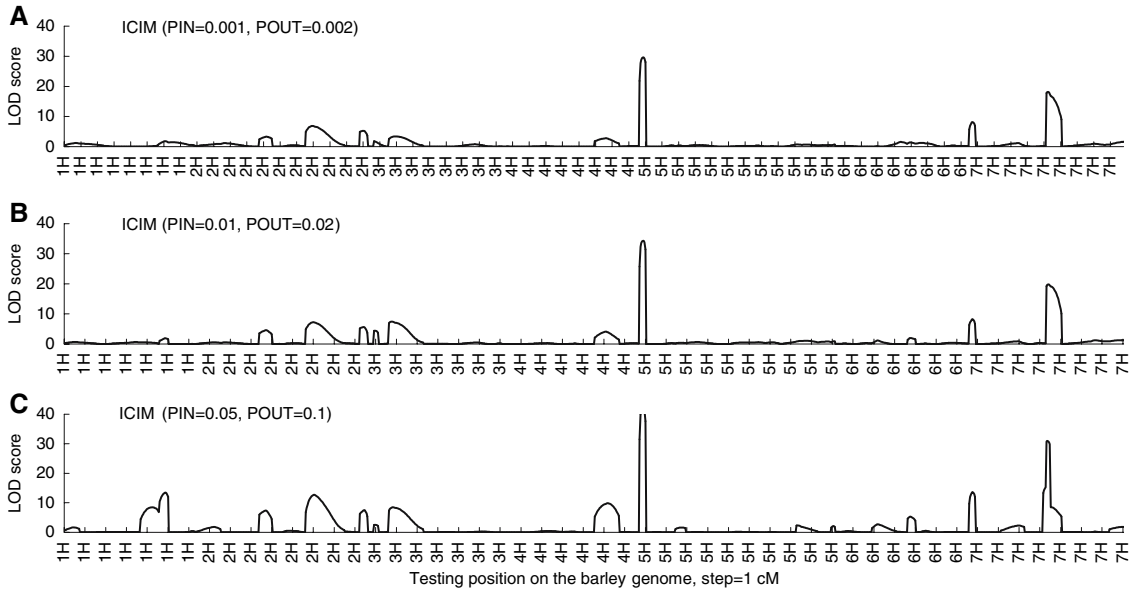

Fig. 6 Mapping results from ICIM for additive QTL affecting kernel weight (KWT) in the barley population consisting of 145 DH lines. Three probabilities for entering variables and removing variables were considered (i.e. PIN = 0.001, 0.01, 0.05 and POUT = 0.002, 0.02, 0.10, respectively). The scanning step is 1 cM and 1–7H represent the seven barley chromosomes

markers and marker pairs explaining additive and epistatic variations. In the second step, interval mapping approach is applied to the phenotypic values, adjusted by using the regression model selected in the first step, to locate QTL in marker intervals and estimate their effects. The use of adjusted phenotypic values for interval mapping can be viewed as using cofactors to control genetic background effect, an idea similar to the conventional CIM (Zeng 1994). The separation of cofactor selection using stepwise regression and the interval mapping using ML method effectively removes the requirement of deciding how many terms (main effects and epistasis) should be included in the model, which is the most difficult issue faced by the MIM proposed by Kao et al. (1999). It also avoids the choice of "the effective dimension" (i.e., number of QTL) for epistatic interactions, which is required by the one-dimensional genome searches approach developed by Boer et al. (2002). In addition, the computation speed is dramatically increased since cofactor is selected only once for the entire search. The interval mapping in ICIM is relatively straightforward, which involves the use of ECM to calculate LOD scores along the genome one- or two-dimensionally.

Clearly, the result of ICIM depends on the identification of an appropriate regression model in the first step. Choice of variables for multiple regression is a typical model selection issue. The number of possible models is huge due to the high number of markers and marker pairs. In this article the stepwise regression technique was applied and satisfactory results were obtained. Treating QTL mapping as model selection problem and the use of model selection criteria to identify the best model have been investigated by many authors (Piepho and Gauch 2001; Broman and Speed 2002; Bogdan et al. 2004; Baierl et al. 2006). The Schwarz Bayesian information criterion was modified by Bogdan et al. (2004) and Baierl et al. (2006) to suit the identification of main effect and interactive QTL using forward selection procedure for the backcross and intercross design, respectively. However, these studies made the assumption that QTL are sitting on the markers, which is not likely to be true when marker density is not very high. It will be worthy investigating whether the use of common selection criteria and their modified versions in the first step of ICIM can improve its performance significantly.

The influence of marker inclusion and exclusion criteria in stepwise regression

The largest P value for entering variables and the smallest P value for removing variables are required for the stepwise regression when using ICIM. This is where the subjectivity comes into play. We believe that this is much easier than cofactor selection in traditional CIM and model selection in MIM.

It can be proved that for any marker variables i, j and k $\mathrm{Cov}(x_i, x_j x_k) = 0$, indicating the effects of markers and marker-pair multiplications are independent under the assumption of large sample size. Therefore, we adopted a two-stage regression strategy to estimate the parameters in model (5). In the first stage, only individual markers were considered in stepwise regression. The largest P value for

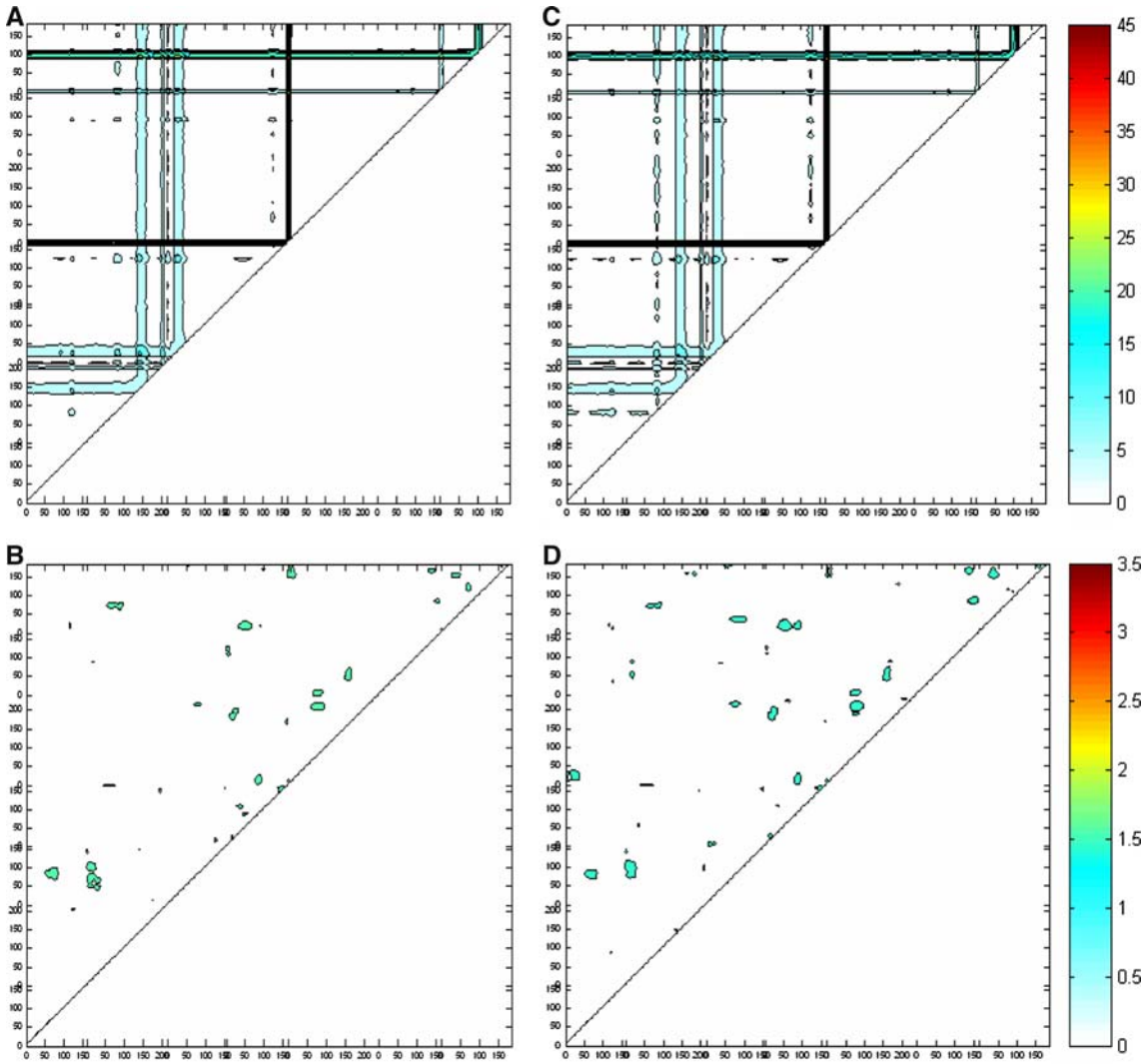

Fig. 7 Two-dimensional scanning from ICIM testing the significance of additive and epistasis (**a** and **c**), and epistasis only (**b** and **d**) affecting kernel weight (KWT) in the barley population consisting of 145 DH lines. Two probabilities for entering variables and removing variables in the second stage of stepwise regression were considered (PIN = 0.0005 and POUT = 0.0010 for **a** and **b**, and PIN = 0.0001 and POUT = 0.0002 for **c** and **d**)

entering variables (PIN_1) and the smallest P value for removing variables ($POUT_1$) are generally set to 0.05 and 0.10, respectively. More strict probability levels, such as PIN_1=0.01 and $POUT_1$=0.02 can be used to further reduce the false positives without sufficiently changing the detection power (Li et al. 2007). The regression residuals from the first stage were used to regress on all marker-pair multiplications in the second stage. Due to the large amount of variables in this stage, much stricter probability levels for entering variables (PIN_2) and removing variables ($POUT_2$) should be used to avoid over-fitting.

Trait heritability may be used to justify the used probability levels. From model (5), the R^2 after fitting all marker variables and marker pairs should approximate the proportion of phenotypic variation explained by additive and digenic interacting QTL, that is, the broad sense heritability. From genetic studies and breeding practice, the range of heritabilities of most quantitative traits is roughly known (Falconer and Mackay 1996). Therefore, an R^2 higher than heritability may suggest stricter probability levels should be applied. In comparison, an R^2 lower than heritability may suggest less stricter probability levels should be applied.

The LOD threshold for statistical inference in ICIM

Many factors affect the LOD score in QTL mapping, among which are population size, number and distribution of markers and putative QTL, QTL effects and error

variance, etc. Similar to ICIM for additive mapping (Li et al. 2007), permutation tests can also be conducted in the case of epistatic mapping to find the LOD score distribution. ICIM requires about 15 min in a personal computer (1.83 GHz CPU) to complete one run of additive and epistatic mapping (scanning step = 1 cM) for a backcross population from genome 1. Therefore, extensive computing time is required to conduct a large number of permutation tests, say 1,000 times, even if there is no theoretical restrict for conducting permutation tests in ICIM epsitatic mapping.

In practice, the choice of LOD threshold depends on research purpose, that is to say, what size of QTL need to be identified. QZ4, QZ5, QZ6, QZ8 and QZ9 explain more than 3.00% of phenotypic variation, and have average LOD scores 2.63, 4.15, 4.75, 6.82, and 3.24, respectively (Table 3). If the LOD threshold of 2.0 is applied, these QTL can be identified with relatively high powers. In comparison, if the LOD threshold of 3.0 is applied, the power for detecting QTL4 would be lower. The same is true for interacting QTL. Therefore, if one wants to detect additive or epistatic QTL with smaller effects, say explains less than 3% of phenotypic variation, a lower LOD threshold has to be used at the expense of a likely higher risk of false positives. From our experience, the normally accepted LOD threshold from 2.0 to 3.0 can be used for ICIM additive mapping. Similar or a little higher LOD threshold can be used for ICIM epistatic mapping, such as 2.5 to 3.5. These LOD thresholds will make sure the identified QTL are likely true QTL, but those explaining less than 3% of phenotypic variation may be ignored.

ICIM is an efficient mapping method for both additive and epistasis

The ICIM provides intuitive statistics for testing additive and epistasis, and can be used for experimental populations derived from two inbred parental lines. When mapping digenic epistasis, ICIM gives two LOD scores, i.e., LOD_A and LOD_{AA}. LOD_A contains the information of both additive and epistasis of QTL at the two testing positions, while LOD_{AA} contains the information of epistasis only. LOD_{AA} was the statistic excluding the influence of additive effects in LOD_A, so LOD_{AA} is lower than LOD_A (Figs. 1A, B, 3A, B, D, E, 4A, B, 7A, B, C, D). As previously shown, the additive mapping results from one-dimensional LOD profile (see profiles on the axes of Figs. 1A, B, 3A ,B, D, E and 4A ,B) and from two-dimensional LOD_A profile (Figs. 1A, 3A, D and 4A) are consistent. The predefined epistasis were well demonstrated in the two-dimensional LOD_{AA} profile (Figs. 1B, 3B, E, 4B). Therefore, we recommend that both of the one-dimensional LOD profile and the two-dimensional LOD_{AA} profile be used. The additive QTL can be deduced from the one-dimensional scanning, and the digenic epistatic QTL from the two-dimensional scanning.

Genetic variation due to the identified QTL with additive effects can be calculated from equation (8), which can be used to determine whether the two-dimensional scanning is necessary. Generally, if the identified additive QTL have explained most of the genotypic variation, i.e., the proportion of the phenotypic variation explained is close to heritability in the broad sense, epistasis is less important and the two-dimensional scanning for epistasis may not be necessary. Otherwise, the two-dimensional scanning should be conducted to detect epistatic QTL. Genetic variation due to the identified interacting QTL can be calculated from equation (9). If the identified additive QTL and digenic epsistasis failed to explain most of the genotypic variation, higher order of epistasis must exist. Higher order of epistasis is less likely to be identified from a standard mapping population such as backcross or recombination inbred lines, but may be detected in other populations such as chromosome segment substitute lines (Nadeau et al. 2000; Carlborg et al. 2006).

We have used mapping populations with two genotypes to illustrate our algorithm in this paper. The extension of ICIM to F_2 with three genotypes may need additional work, as more genetic parameters such as dominance effects and various epistatic effects involving dominance have to be added to model (1). But in theory, there is no limit to apply ICIM to F_2 populations. The software implementing ICIM additive and epistasis mapping called QTL IciMapping was written in Fortran 90/95, and is freely available from http://www.isbreeding.net. The simulated and real populations used in this study are also available from this website.

Acknowledgments This work was supported by the National 973 and 863 Programs of China (2006CB101700 and 2006AA10Z1B1), and the Generation Challenge Program of the Consultative Group for International Agricultural Research.

References

Baierl A, Bogdan M, Frommlet F, Futschik A (2006) On locating multiple interacting quantitative trait loci in intercross designs. Genetics 173:1693–1703

Boer MP, Ter Braak CJF, Jansen RC (2002) A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 162:951–960

Bogdan M, Ghosh JK, Doerge RW (2004) Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167:989–999

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Statist Soc B 64:641–656

Carlborg Ö, Haley C (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5:618–625

Carlborg Ö, Kerje S, Schütz K, Jacobsson L, Jensen P, Andersson L (2003) A global search reveals epistatic interaction between QTL for early growth in the chicken. Genome Res 13:413–421

Carlborg Ö, Jacobsson L, Ahgren P, Siegel P, Andersson L (2006) Epistasis and the release of genetic variation during long-term selection. Nat Genet 38:418–420

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experiment populations. Nat Rev Genet 3:43–52

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4 edn. Longman, Essenx

Feenstra B, Skovgaard IM, Broman KW (2006) Mapping quantitative trait loci by an extension of the Haley–Knott regression method using estimating equations. Genetics 173:2269–2282

Frankel WN, Schork NJ (1996) Who's afraid of epistasis. Nat Genet 14:371–373

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

Jannink J, Jansen R (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157:445–454

Kao C-H, Zeng Z-B, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1206

Kroymann J, Mitchell-Olds T (2005) Epistasis and balanced polymorphism influencing complex trait variation. Nature 435:95–98

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

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

Lynch M, Walsh B (1998) Genetic and analysis of quantitative Traits. Sinauer Associates, Sunderland

Mackay TFC (2001) Quantitative trait loci in *Drosophila*. Nat Rev Genet 2:11–20

Malmberg RL, Held S, Waits A, Mauricio R (2005) Epistasis for fitness-related quantitative traits in *Arabidopsis thaliana* grown in the field and in the greenhouse. Genetics 171:2013–2027

Meng X, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80:267–268

Miller AJ (1990) Subset selection in regression (Monographs on statistics and applied probability 40). Chapman and Hall, London

Nadeau JH, Singer JB, Martin A, Lander ES (2000) Analysis complex genetics traits with chromosome substitution strains. Nat Genet 24:221–225

Piepho H-P, Gauch HG (2001) Marker pair selection for mapping quantitative trait loci. Genetics 157:433–444

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805–816

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387

Sillanpää MJ, Arjas E (1999) Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics 151:1605–1619

Sillanpää MJ, Corander J (2002) Model choice in gene mapping: what and why. Trends Genet 18:302–307

Tinker NA, Mather DE, Rossnagel BG, Kasha KJ, Kleinhofs A, Hayes PM, Falk DE, Ferguson T, Shugar LP, Legge WG, Irvine RB, Choo TM, Briggs KG, Ullrich SE, Franckowiak JD, Blake TK, Graf RJ, Dofing SM, Saghai Maroof MA, Scoles GJ, Hoffman D, Dahleen LS, Kilian A, Chen F, Biyashev RM, Kudrna DA, and Steffenson BJ (1996) Regions of the genome that affect agronomic performance in two-row barley. Crop Sci 36:1053–1062

Uimari P, Hoeschele I (1997) Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics 146:735–743

Uimari P, Thaller G, Hoeschele I (1996) The use of multiple markers in a Bayesian method for mapping quantitative trait loci. Genetics 143:1831–1842

Wade MJ (2002) A gene's eye view of epistasis, selection and speciation. J Evol Biol 15:337–346

Wang S, Basten CJ, Zeng Z-B (2005) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC

Whittaker JC, Thompson R, Visscher PM (1996) On the mapping of QTL by regression of phenotype on marker-type. Heredity 77:23–32

Xu S, Jia Z (2007) Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics 175:1955–1963

Yi N (2004) A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics 167:967–975

Yi N, Xu S, Allison DB (2003) Bayesian model choice and search strategies for mapping interacting quantitative trait loci. Genetics 165:867–883

Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170:1333–1344

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

Zeng Z-B (2005) Modeling quantitative trait loci and interpretation of models. Genetics 169:1711–1725