RESEARCH ARTICLE

Simultaneous estimation of QTL effects and positions when using genotype data with errors

LIANG TONG[1,2], WEIJUN MA[1], HAIDONG LIU[1], CHAOFENG YUAN[1] and YING ZHOU[1]*

[1]*School of Mathematical Sciences, Heilongjiang University, Harbin 150080, People's Republic of China*
[2]*School of Information Engineering, Suihua University, Suihua 152061, People's Republic of China*

Abstract

Accurate genetic data are important prerequisite of performing genetic linkage test or association test. Currently, most analytical methods assume that the observed genotypes are correct. However, due to the constraint at the technical level, most of the genetic data that people used so far contain errors. In this paper, we considered the problem of QTL mapping based on biological data with genotyping errors. By analysing all possible genotypes of each individual in framework of multiple-interval mapping, we proposed an algorithm of inferring all model parameters through the expectation-maximization (EM) algorithm and discussed the hypothesis testing of the existence of QTL. We carried out extensive simulation studies to assess the proposed method. Simulation results showed that the new method outperforms the method that does not take the genotyping errors into account, and therefore it can decrease the impact of genotyping errors on QTL mapping. The proposed method was also applied to analyse a real barley dataset.

[Tong L., Ma W., Liu H., Yuan C. and Zhou Y. 2015 Simultaneous estimation of QTL effects and positions when using genotype data with errors. *J. Genet.* **94**, 27–34]

Introduction

Gene mapping is very important for genetic studies that can map genes of disease to some position of the chromosome, provide necessary genetic information that can make some genetic diseases diagnosed, and help to clone the pathogenesis of these diseases, and so on.

With the rapid development of molecular marker technique, it has been widely used in the gene mapping of animal and plant populations. Lander and Botstein (1989) proposed interval mapping based on the concept of molecular marker technique, which can be used to provide a good estimation of additive or dominant effect. By combining multiple regression with interval mapping, Zeng (1994) proposed a composite interval mapping method, which can be used to control the background effects by fitting QTL located outside a tested interval in the statistical model. Wang *et al.* (1999) established a mixed linear model based on composite interval mapping method (MCIM) to analyse both epistasis and QTL–environment (Q×E) interaction in a double haploid population. The MCIM method provide unbiased estimation for both position and effect of QTL, as well as unbiased predicted values for Q×E interactions, which makes it a good prospect for application. Kao *et al.* (1999) proposed multiple interval mapping (MIM) method for mapping QTL which uses multiple marker intervals simultaneously to fit multiple QTL directly in the model. With the MIM method, not only the precision and power for mapping QTL could be improved, but also the epistasis among QTL and the heritability of quantitative traits can also be estimated and analysed.

All the above methods depend on certain genotyping technologies. The genotyping technologies, TaqMan and OLA, are based on the fluorescence intensity value, and MassARRAY depends on other signal intensity value. By using genotying scoring softwares to cluster the signal intensity, one can get the genotypes of SNPs. Currently, the next-generation sequencing technology allows to obtain tens of thousands of SNPs across the genome in a fast and cost-effective way. However, due to the constraints in genotyping scoring softwares and biochemical anomalies, most of the data that people have used contain certain errors, where the errors refer to the random genotyping errors or the wrong codes for markers that an experimenter made. Unfortunately, even a small number of genotyping errors can have significant impact on the study of genetic analysis, such as linkage studies

*For correspondence. E-mail: yzhou@aliyun.com.

Keywords. backcross model; EM algorithm; genotyping errors; maximum likelihood estimation; QTL mapping.

(Buetow 1991; Douglas *et al.* 2000; Sobel *et al.* 2002), genetic distance estimation (Goldstein *et al.* 1997) and linkage disequilibrium (LD) estimation (Akey *et al.* 2001). Abecasis *et al.* (2001) showed that the impact of genotyping errors on family-based analysis of quantitative traits. That is to say that the errors of genetic data can distort gene mapping. Sobel *et al.* (2002) proposed that such errors in statistical analysis cannot be neglected, given their importance in gene mapping. Thus the identification of genotyping errors is important for gene mapping, and how to map genes using genetic data with genotyping errors is an issue of concern to researchers. Cartwright *et al.* (2007) extended the traditional likelihood model used for genetic mapping to include the possibility of genotyping errors, and their aim was to reconstruct the order of the markers on the chromosomes and estimate the genetic distances between them. Lebrec *et al.* (2008) studied the impact of genotyping errors on linkage mapping of complex traits. Hou (2011) analysed a special case that QTL are located at the markers probably with genotyping errors, but he did not do further research on the case that QTL belong to marker intervals.

By considering all possible genotypes of each individual based on the data with genotyping errors, in this paper we developed an effective method for mapping QTL by means of the classical EM algorithm (Dempster *et al.* 1977) to simultaneously estimate all genetic parameters in the framework of multiple-interval mapping. Simulation studies show that the new method has advantage over the method which does not take the genotyping errors into account, and that it can infer the impact induced by genotyping errors. We also analysed a real dataset to demonstrate its practicality.

Background and statistical model

Consider N backcross individuals, and M marker intervals divided by $M+1$ genetic marker loci. M_j and m_j, respectively denote the two alleles of the jth locus. Let

$$\boldsymbol{Y} = (Y_1, \cdots, Y_N)^T, X_i = (X_{i1}, \cdots, X_{i(M+1)})^T,$$
$$\tilde{X}_i = (\tilde{X}_{i1}, \cdots, \tilde{X}_{i(M+1)})^T, X_i^* = (X_{i1}^*, \cdots, X_{i(M)}^*)^T,$$

where Y_i $(i = 1, \cdots, N)$ denotes the phenotype value of the ith individual, X_{ij} $(i = 1, \cdots, N, j = 1, \cdots, M+1)$ and $\tilde{X}_{ij}$ $(i = 1, \cdots, N, j = 1, \cdots, M+1)$, respectively denote the true genotype and genotype probably with error of the jth marker of the ith individual, and X_{ij}^* $(i = 1, \cdots, N, j = 1, \cdots, M)$ denotes the genotype of the latent QTL in the jth marker interval of the ith individual. We define the following genotype values:

$$X_{ij} = \begin{cases} 1, & \textit{if marker genotype is } M_jM_j, \\ 0, & \textit{otherwise}, \end{cases}$$

$$X_{ij}^* = \begin{cases} 1, & \textit{if QTL genotype is } Q_jQ_j, \\ 0, & \textit{otherwise}. \end{cases}$$

Let γ_j and γ_{j1}, respectively denote recombination rate of the jth marker interval which is known and the recombination rate between the jth marker and the latent QTL in the interval. At the most one QTL in a marker interval is assumed. Let $p(X_{ij}^*|X_{ij}^M)$ denote the conditional probability of the QTL genotype X_{ij}^* given the genotype combination X_{ij}^M of the jth marker interval of the ith individual. For backcross families, X_{ij}^M has four possible values (M_jM_{j+1}/M_jM_{j+1}, M_jM_{j+1}/M_jm_{j+1}, M_jM_{j+1}/m_jM_{j+1}, and M_jM_{j+1}/m_jm_{j+1}, which are coded as 1, 2, 3 and 4, respectively). The conditional probabilities $p(X_{ij}^*|X_{ij}^M)$ are presented in table 1.

Table 1. The conditional probabilities of QTL genotypes given the flanking marker genotypes.

	Genotype of	Genotype of QTL	
Code	marker	QQ	Qq
1	M_jM_{j+1}/M_jM_{j+1}	1	0
2	M_jM_{j+1}/M_jm_{j+1}	$\frac{\gamma_j-\gamma_{j1}}{\gamma_j}$	$\frac{\gamma_{j1}}{\gamma_j}$
3	M_jM_{j+1}/m_jM_{j+1}	$\frac{\gamma_{j1}}{\gamma_j}$	$\frac{\gamma_j-\gamma_{j1}}{\gamma_j}$
4	M_jM_{j+1}/m_jm_{j+1}	0	1

Without loss of generality, we assume that the error rate of each marker is equal (we will discuss the case of unequal error rates later). Here we let $\theta = P(\tilde{X}_{ij} = k|X_{ij} = 1-k)$ denote the genotyping error rate $(k = 0, 1)$, and further let φ^i denote the joint error rate of the ith individual. When the true genotype X_i is assigned a detailed value in each step of our iterative algorithm and compared with genotype $\tilde{X}_i$, the number k_i of the incorrect genotype codes of the $M+1$ marker loci can be calculated. Assume whether one marker genotype has genotyping error is independent of the others, thus we obtain that:

$$\varphi^i = P(\tilde{X}_i|X_i) = \prod_{j=1}^{M+1} P(\tilde{X}_{ij}|X_{ij}) = \theta^{k_i}(1-\theta)^{(M+1)-k_i}.$$

In this study, we consider the follow additive statistical model:

$$Y_i = \alpha + \sum_{j=1}^{M} X_{ij}^* \beta_j + \epsilon_i, \quad i = 1, \cdots, N,$$

where α is the total mean, β_j is the genotype effect of latent QTL in the jth marker interval, $\epsilon_i \sim N(0, \sigma^2)$, X_{ij}^* and ϵ_i are mutually independent.

Method

Here, we present a simultaneous multiple-interval mapping method for QTL while using data with genotying errors. Let parameter vector $\Omega = (\alpha, \beta_1, \ldots, \beta_M, \gamma, \theta, \sigma^2)$, where $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{M1})$. When the observed data are obtained, we utilize the method of weights to estimate the parameter vector Ω by considering the unobserved true genotypes as the missing data, and the EM algorithm is implemented to numerically compute these estimates. First, we augment the observed data $\{(\tilde{X}_i, Y_i),\ i = 1 \ldots N\}$ by the unobserved QTL genotype and the true marker genotypes. Then $\{(\tilde{X}_i, Y_i, X_i, X_i^*),\ i = 1 \ldots N\}$ are obtained as the complete data. Since the distribution of

X_i has no relationship with Ω, based on the conditional independence properties among variables $\tilde{X}_i, Y_i, X_i$ and X_i^*, the complete likelihood function for the ith individual is given by:

$$\begin{aligned} L_c^i(\Omega) &= P(\tilde{X}_i, Y_i, X_i, X_i^*|\Omega) \\ &\propto P(Y_i|X_i^*, \Omega) \cdot P(X_i^*|X_i, \Omega) \cdot P(\tilde{X}_i|X_i, \Omega), \end{aligned}$$

thus the complete log-likelihood function is:

$$\begin{aligned} l_c(\Omega) &= \ln \prod_{i=1}^{N} P(Y_i|X_i^*, \Omega) \cdot P(X_i^*|X_i, \Omega) \cdot P(\tilde{X}_i|X_i, \Omega) \\ &= \sum_{i=1}^{N} [\ln P(Y_i|X_i^*, \Omega) + \ln P(X_i^*|X_i, \Omega) + \ln P(\tilde{X}_i|X_i, \Omega)]. \end{aligned}$$

The detail of the method for inferring the parameter vector Ω in the simultaneous multiple-interval mapping can be described as follows:

E-step: given $\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}$, compute the conditional expectation of $l_c(\Omega)$.

$$\begin{aligned} Q(\Omega|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}) &= \sum_{i=1}^{N} E\big[\ln P(Y_i|X_i^*, \Omega) + \ln P(X_i^*|X_i, \Omega) \\ &\quad + \ln P(\tilde{X}_i|X_i, \Omega)|\tilde{X}, \boldsymbol{Y}, \Omega^{(k)}\big] \\ &= \sum_{i=1}^{N} E_{X_i^*}[\ln P(Y_i|X_i^*, \Omega)|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}] \\ &\quad + \sum_{i=1}^{N} E_{X_i^*, X_i}[\ln P(X_i^*|X_i, \Omega)|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}] \\ &\quad + \sum_{i=1}^{N} E_{X_i}[\ln P(\tilde{X}_i|X_i, \Omega)|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}], \end{aligned}$$

where $\Omega^{(k)}$ represents the current estimate of Ω and $\tilde{\mathbf{X}} = (\tilde{X}_1 \ldots \tilde{X}_N)$ denotes the marker genotypes probably with errors of N individuals. Let $P(X_i = x_i, X_i^* = x_i^*|\tilde{X}_i, Y_i, \Omega^{(k)}) = \omega_{x_i x_i^*}^{(k)}$, then the marginal probabilities are:

$$P\left(X_i^* = x_i^*|\tilde{X}_i, Y_i, \Omega^{(k)}\right) = \sum_{x_i} \omega_{x_i x_i^*}^{(k)} = \omega_{x_i^*}^{(k)},$$

$$P\left(X_i = x_i|\tilde{X}_i, Y_i, \Omega^{(k)}\right) = \sum_{x_i^*} \omega_{x_i x_i^*}^{(k)} = \omega_{x_i}^{(k)}.$$

Here $\omega_{x_i x_i^*}^{(k)}, \omega_{x_i^*}^{(k)}, \omega_{x_i}^{(k)}$ are weights, in detail,

$$\begin{aligned} \omega_{x_i x_i^*}^{(k)} &= P\left(X_i = x_i, X_i^* = x_i^*|\tilde{X}_i, Y_i, \Omega^{(k)}\right) \\ &= \frac{P\left(x_i^*, x_i, Y_i, \tilde{X}_i|\Omega^{(k)}\right)}{P(Y_i, \tilde{X}_i|\Omega^{(k)})} \\ &= \frac{P\left(Y_i|x_i^*, \Omega^{(k)}\right) P\left(x_i^*|x_i, \Omega^{(k)}\right) P\left(\tilde{X}_i|x_i, \Omega^{(k)}\right)}{\sum\limits_{x_i} \sum\limits_{x_i^*} P(Y_i|x_i^*, \Omega^{(k)}) P(x_i^*|x_i, \Omega^{(k)}) P(\tilde{X}_i|x_i, \Omega^{(k)})} \\ &= \frac{\varphi\left(Y_i, \alpha^{(k)}, \beta^{(k)}, (\sigma^2)^{(k)}\right) \cdot\ P(x_i^*|x_i, \Omega^{(k)}) \cdot (\theta^{(k)})^{k(x_i)} \cdot (1-\theta^{(k)})^{(M+1)-k(x_i)}}{\sum\limits_{x_i} \sum\limits_{x_i^*} \varphi\left(Y_i, \alpha^{(k)}, \beta^{(k)}, (\sigma^2)^{(k)}\right) \cdot\ P(x_i^*|x_i, \Omega^{(k)}) \cdot (\theta^{(k)})^{k(x_i)} \cdot (1-\theta^{(k)})^{(M+1)-k(x_i)}}, \end{aligned}$$

where $\varphi(Y_i, \alpha^{(k)}, \beta^{(k)}, (\sigma^2)^{(k)}) = \frac{1}{\sqrt{2\pi}\sigma^{(k)}} \exp\{-\frac{(Y_i - \alpha^{(k)} - \sum_{j=1}^{M} \beta_j^{(k)} x_{ij}^*)^2}{2(\sigma^2)^{(k)}}\}$; $\sum\limits_{x_i} = \sum\limits_{x_{i1}=0}^{1} \cdots \sum\limits_{x_{i(M+1)}=0}^{1}$; $\sum\limits_{x_i^*} = \sum\limits_{x_{i1}^*=0}^{1} \cdots \sum\limits_{x_{i(M)}^*=0}^{1}$; $k(x_i)$ represents the number of incorrect codes when the true genotype of ith individual is x_i. Therefore the Q-function

$$\begin{aligned} Q(\Omega|\tilde{\mathbf{X}}, \boldsymbol{Y}, \Omega^{(k)}) &= \sum_{i=1}^{N} \sum_{x_i^*} \omega_{x_i^*}^{(k)} \ln P(Y_i|x_i^*, \Omega) \\ &\quad + \sum_{i=1}^{N} \sum_{x_i} \sum_{x_i^*} \omega_{x_i x_i^*}^{(k)} \ln P(x_i^*|x_i, \Omega) \\ &\quad + \sum_{i=1}^{N} \sum_{x_i} \omega_{x_i}^{(k)} \ln P(\tilde{x}_i|x_i, \Omega). \qquad (1) \end{aligned}$$

Note that the first term of eq. (1) depends on parameters $\alpha, \beta_1, \ldots, \beta_M, \sigma^2$, the second term depends only on γ and the last term depends only on θ.

M-step: maximize the conditional expected log likelihood $Q(\Omega|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)})$ to obtain $\Omega^{(k+1)}$.

Iterative formulae of α, $\beta_1, \ldots, \beta_M$, and σ^2

Parameters α, $\beta_1, \ldots, \beta_M$, and σ^2 are contained in the first term of eq. (1). For simplicity, we consider the forms of vector and matrix for all the variables and parameters (Chen 2005). Let

$$\boldsymbol{X}^* = (X_1^*, \cdots, X_N^*)^T, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^T, \boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_N)^T,$$

then the previous statistical model can be written as $\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$. Therefore $\boldsymbol{Y}|\boldsymbol{X}^* \sim N(\alpha\mathbf{1} + \boldsymbol{X}^*\boldsymbol{\beta}, \sigma^2 I)$, and

$$\begin{aligned} \ln P(\boldsymbol{Y}|\boldsymbol{X}^*) &= -\left(\frac{N}{2}\right) \ln(2\pi)\sigma^2 - \frac{1}{2\sigma^2}[(\boldsymbol{Y} - \alpha\mathbf{1})^T(\boldsymbol{Y} - \alpha\mathbf{1}) \\ &\quad -2(\boldsymbol{Y} - \alpha\mathbf{1})^T\boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^{*T}\boldsymbol{X}^*\boldsymbol{\beta}]. \end{aligned}$$

Let $E_1 = E_{X^*}(X^*|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}), E_2 = E_{X^*}(\boldsymbol{X}^{*T}\boldsymbol{X}^*|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)})$. Maximize $E_{X^*}[\ln P(\boldsymbol{Y}|\boldsymbol{X}^*)|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)})]$, we can get

$$\alpha^{(k+1)} = \bar{\boldsymbol{Y}} - \frac{1}{N}\mathbf{1}^T E_1 \boldsymbol{\beta}^{(k+1)}$$

$$\boldsymbol{\beta}^{(k+1)} = \Big(E_2 - \frac{1}{N}E_1{}^T\mathbf{1}\mathbf{1}^T E_1\Big)^{-1} E_1{}^T\Big(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T\Big)\boldsymbol{Y}$$

$$(\sigma^2)^{(k+1)} = \frac{1}{N}[(\boldsymbol{Y}-\mathbf{1}\alpha^{(k+1)})^T(\boldsymbol{Y}-\mathbf{1}\alpha^{(k+1)})-\boldsymbol{\beta}^{T(k+1)}E_2\boldsymbol{\beta}^{(k+1)}].$$

where the elements of E_1 can be obtained as follows:

$$\begin{aligned} E(X^*_{ij}|\tilde{X}_i, Y_i, \Omega^{(k)}) &= P(X^*_{ij} = 1|Y_i, \tilde{X}_i, \Omega^{(k)}) \\ &= \frac{\sum\limits_{X_i}\sum\limits_{X^*_{ik}, k\neq j} P(X^*_{ij} = 1, X^*_{i1}\ldots X^*_{ij-1}, X^*_{ij+1}\ldots X^*_{iM}, Y_i, X_i, \tilde{X}_i|\Omega^{(k)})}{\sum\limits_{X_i}\sum\limits_{X^*_i} P(X^*_i, X_i, Y_i, \tilde{X}_i|\Omega^{(k)})} \\ &= \frac{\sum\limits_{X_i}\sum\limits_{X^*_{ik}, k\neq j} P(Y_i|X^*_{ij} = 1, J^*, \Omega^{(k)})P(X^*_{ij} = 1, J^*|X_i, \Omega^{(k)})P(\tilde{X}_i|X_i, \Omega^{(k)})}{\sum\limits_{X_i}\sum\limits_{X^*_i} P(Y_i|X^*_i, \Omega^{(k)})P(X^*_i|X_i, \Omega^{(k)})P(\tilde{X}_i|X_i, \Omega^{(k)})}, \end{aligned}$$

where J^* denotes genotype $X^*_{i1}, \ldots, X^*_{ij-1}, X^*_{ij+1}, \ldots, X^*_{iM}$. Similarly, the elements of E_2 can be obtained as follows:

$$\begin{aligned} &E\Big[\sum_{c=1}^{N} X^*_{cs}X^*_{ct}|\tilde{\boldsymbol{X}}, \boldsymbol{Y}, \Omega^{(k)}\Big] \\ &= \sum_{c=1}^{N} \frac{P(X^*_{cs} = 1, X^*_{ct} = 1, Y_c, \tilde{X}_c|\Omega^{(k)})}{P(Y_c, \tilde{X}_c|\Omega^{(k)})} \\ &= \sum_{c=1}^{N} \frac{\sum\limits_{X_c}\sum\limits_{X^*_{ck}, k\neq s,t} P(Y_c|X^*_{cs} = 1, X^*_{ct} = 1, T^*, \Omega^{(k)})P(X^*_{cs} = 1, X^*_{ct} = 1, T^*|X_c, \Omega^{(k)})\tilde{P}}{\sum\limits_{X_c}\sum\limits_{X^*_c} P(Y_c|X^*_c, \Omega^{(k)})P(X^*_c|X_c, \Omega^{(k)})\tilde{P}}, \end{aligned}$$

where $\tilde{P} = P(\tilde{X}_c|X_c, \Omega^{(k)})$, and T^* denotes $X^*_{i1}, \ldots, X^*_{is-1}, X^*_{is+1}, \ldots, X^*_{it-1}, X^*_{it+1}, \ldots, X^*_{iM}, s, t{=}1, \ldots, M$.

Iterative formula of γ

Let

$$I^{ij}_{(st)} = \begin{cases} 1, \ X^*_{ij} = s, X^M_{ij} = t\ ; s = 0, 1\ ; t = 1, \ldots, 4, \\ 0, \ else, \end{cases}$$

where $i = 1, \ldots, N; j = 1, \ldots, M$. Thus we obtain that the second term of eq. (1)

$$\begin{aligned} \sum_{i=1}^{N}\sum_{x_i}\sum_{x^*_i} \omega^{(k)}_{x_i x^*_i} \ln P(x^*_i|x_i, \Omega) &= \sum_{i=1}^{N}\sum_{x_i}\sum_{x^*_i} \omega^{(k)}_{x_i x^*_i} \sum_{j=1}^{M} \ln P(x^*_{ij}|x^M_{ij}, \Omega) \\ &= \sum_{i=1}^{N}\sum_{x_i}\sum_{x^*_i} \omega^{(k)}_{x_i x^*_i} \sum_{j=1}^{M} \Big[\ln\frac{\gamma_j - \gamma_{j1}}{\gamma_j}\cdot I^{ij}_{(12)} + \ln\frac{\gamma_{j1}}{\gamma_j}\cdot I^{ij}_{(02)} + \ln\frac{\gamma_{j1}}{\gamma_j}\cdot I^{ij}_{(13)} + \ln\frac{\gamma_j - \gamma_{j1}}{\gamma_j}\cdot I^{ij}_{(03)}\Big]. \end{aligned} \tag{2}$$

Through maximizing eq. (2) we obtain:

$$\gamma_{j1}^{(k+1)} = \frac{\gamma_j \cdot \sum\limits_{i=1}^{N}\sum\limits_{x_i}\sum\limits_{x^*_i} \omega^{(k)}_{x_i x^*_i}(I^{ij}_{(02)} + I^{ij}_{(13)})}{\sum\limits_{i=1}^{N}\sum\limits_{x_i}\sum\limits_{x^*_i} \omega^{(k)}_{x_i x^*_i}(I^{ij}_{(12)} + I^{ij}_{(02)} + I^{ij}_{(13)} + I^{ij}_{(03)})}, \quad j = 1, \cdots, M.$$

Iterative formula of θ

To find $\theta^{(k+1)}$, we maximize the above Q-function eq. (1) and obtain that

$$\theta^{(k+1)} = \frac{\sum\limits_{i=1}^{N}\sum\limits_{x_i} k(x_i)P(X_i = x_i|\tilde{X}_i, Y_i, \theta^{(k)})}{N(M+1)}.$$

In each step of our iterative algorithm, all parameters in Ω have closed-form solutions. Given the initial value $\Omega^{(0)}$ of Ω, we can get the MLE of parameter vector Ω when the above procedure is iteratively carried out until convergence. Besides estimating the effects and positions of QTL, the estimate of the error rate θ can also be obtained by the method at the same time.

Simulation study

Simulation design

We conduct simulation studies to evaluate the proposed method (PM) of simultaneous multiple-interval mapping in the presence of genotyping errors and compared our method with the ordinary method (OM) in which genotype errors are not considered.

For the sake of simplicity, we considered the situation that a biological trait is contributed by two QTL located within two marker intervals of equal length on a chromosome. The length of marker interval is 10 cM, which can be converted to the recombination rate 0.0906 (Zhou 2010). Two settings of sample size are designed in the simulation, with $N = 500$ and $N = 1000$, separately. To evaluate the influences of different factors for QTL mapping (e.g., genotyping errors, heritability and signs of β_1 and β_2) we, respectively simulate marker genotypes with error rates of 0, 0.01, 0.05 and 0.1, considering three scenarios of heritability (h^2 =0.05, 0.1, 0.2) by choosing different true values of parameters and design different sign combinations of β_1 and β_2. Besides, we also considered three cases of recombination rates $\gamma_{11} = 0.02, \gamma_{21} = 0.03$; $\gamma_{11} = 0.02, \gamma_{21} = 0.05$; $\gamma_{11} = 0.05, \gamma_{21} = 0.06$.

For demonstration purpose, we provide the generating process of simulation data for each set of parameters in detail: (i) according to the true values of recombination rates of two marker intervals, we randomly generated genotype vector X_i of markers for the two maker intervals. (ii) Based on genotype vector X_i, we generated genotypes X^*_{i1} and X^*_{i2} of two latent QTL according to the conditional probabilities given in table 1. (iii) Generate phenotype value of individual i from the model: $Y_i = \alpha + X^*_{i1}\beta_1 + X^*_{i2}\beta_2 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, $X^*_{ij} = 1$ if the genotype is homozygous, and 0 otherwise. X^*_{ij} and ϵ_i are independent. (iv) According to the true value of error rate θ, we randomly assign whether the genotype of a marker has error to obtain $\tilde{X}_i$. (v) Repeat steps (i) to (iv) for N times, then the observed data $\{(\tilde{X}_i, y_i),\ i = 1, \cdots, N\}$ can be obtained.

For each set of parameters, we compute the estimates by PM and OM in which genotype errors are not considered, and the whole processes was repeated 500 times. To evaluate the accuracy of estimates, the mean square errors (MSE) for each parameter was also calculated.

Simulation results

The estimates of parameter and MSE when $N = 500$, $h^2 = 0.2$, $\gamma_{11} = 0.02, \gamma_{21} = 0.03$ with different error rates θ are listed in table 2. When $\theta = 0$, we get the same results with the two methods, which is intuitive. This also shows that the OM can be seen as a special case of the new method. With the increase of θ, the deviation of each estimate from its true value becomes higher and higher for both methods, as well as the corresponding MSE. But the accuracy of the proposed method is higher than that of the OM all along except for α. Further, we consider the total MSE of all parameters (TM), which is the mean of MSEs of all parameters except θ. It can be seen from table 2 that each value of the TM of the proposed method (PM) is uniformly lower than the corresponding one of the OM. Thus the new method can reduce the influence of genotyping errors on QTL mapping. The same conclusion can be made for other cases when $h^2 = 0.05, 0.1$, and $\gamma_{11} = 0.02, \gamma_{21} = 0.05$ and $\gamma_{11} = 0.05, \gamma_{21} = 0.06$.

Table 2. Simulation results with different error rates when N=500, $h^2 = 0.2$.

		θ=0		θ=0.01		θ=0.05		θ=0.1	
Parameter	True value	PM[a]	OM[b]	PM	OM	PM	OM	PM	OM
α	1.0003	1.0055	1.0055	0.9963	1.0041	0.9949	0.9990	0.9085	0.9812
	—	(0.0377)[c]	(0.0377)	(0.0396)	(0.0391)	(0.0407)	(0.0390)	(0.0494)	(0.0458)
β_1	−0.7314	−0.7381	−0.7381	−0.7133	−0.6860	−0.5582	−0.5280	−0.4819	−0.4141
	—	(0.1914)	(0.1914)	(0.2323)	(0.2337)	(0.2552)	(0.2952)	(0.2950)	(0.3715)
β_2	1.5866	1.5857	1.5857	1.5675	1.5531	1.4683	1.4192	1.4418	1.3345
	—	(0.1801)	(0.1801)	(0.2128)	(0.2193)	(0.2458)	(0.2594)	(0.3189)	(0.3275)
σ^2	1	1.0024	1.0024	0.9893	1.0052	0.9820	1.0356	0.9681	1.0385
	—	(0.0477)	(0.0477)	(0.0483)	(0.0489)	(0.0493)	(0.0584)	(0.0564)	(0.0589)
γ_{11}	0.02	0.0207	0.0207	0.0259	0.0271	0.0270	0.0275	0.0278	0.0287
	—	(0.0142)	(0.0142)	(0.0160)	(0.0176)	(0.0165)	(0.0174)	(0.0168)	(0.0180)
γ_{21}	0.03	0.0289	0.0289	0.0320	0.0321	0.0332	0.0337	0.0333	0.0400
	—	(0.0084)	(0.0084)	(0.0087)	(0.0088)	(0.0121)	(0.0131)	(0.0133)	(0.0143)
TM[d]	—	0.0798	0.0798	0.0927	0.0946	0.1032	0.1138	0.1249	0.1394

PM[a], the proposed method; OM[b], the ordinary method; $(\cdot)$[c], MSE of estimate for each parameter; TM[d], the mean of MSEs of all parameters except θ.

Table 3. The simulation results with different heritabilities when N=500, $\theta = 0.01, \gamma_{11} = 0.02, \gamma_{21} = 0.03$.

			$\hat{\alpha}$		$\hat{\beta_1}$		$\hat{\beta_2}$		$\hat{\sigma^2}$		$\hat{\gamma_{11}}$		$\hat{\gamma_{21}}$	
h^2	β_1	β_2	PM[a]	OM[b]	PM	OM	PM	OM	PM	OM	PM	OM	PM	OM
0.05	−0.7314	0.3282	1.0088	0.9996	−0.7033	−0.6860	0.3466	0.3470	0.9865	0.9879	0.0279	0.0288	0.0319	0.0332
			(0.0395)[c]	(0.0392)	(0.2414)	(0.2420)	(0.2347)	(0.2353)	(0.0509)	(0.0519)	(0.0168)	(0.0171)	(0.0128)	(0.0131)
0.05	0.7314	0.3282	0.9943	1.0046	0.7748	0.7842	0.2908	0.2882	0.9870	0.9860	0.0291	0.0297	0.0327	0.0338
			(0.0414)	(0.0413)	(0.2337)	(0.2362)	(0.2249)	(0.2256)	(0.0508)	(0.0516)	(0.0170)	(0.0173)	(0.0130)	(0.0136)
0.1	−0.7314	0.9064	0.9894	1.0076	−0.7077	−0.7057	0.9196	0.8860	0.9878	1.0128	0.0267	0.0274	0.0320	0.0324
			(0.0389)	(0.0381)	(0.2360)	(0.2386)	(0.2115)	(0.2185)	(0.0484)	(0.0493)	(0.0165)	(0.0171)	(0.0110)	(0.0113)
0.1	0.7314	0.9064	0.9889	1.0102	0.7561	0.7653	0.9194	0.8859	0.9886	0.9871	0.0269	0.0275	0.0323	0.0325
			(0.0385)	(0.0383)	(0.2339)	(0.2347)	(0.2113)	(0.2166)	(0.0485)	(0.0490)	(0.0169)	(0.0170)	(0.0128)	(0.0133)
0.2	−0.7314	1.5866	0.9963	1.0041	−0.7133	−0.6860	1.5675	1.5531	0.9893	1.0052	0.0259	0.0271	0.0314	0.0315
			(0.0376)	(0.0371)	(0.2323)	(0.2337)	(0.2128)	(0.2193)	(0.0483)	(0.0489)	(0.0160)	(0.0176)	(0.0087)	(0.0088)
0.2	0.7314	1.5866	0.9922	0.9952	0.7501	0.7512	1.5693	1.5602	0.9893	1.0052	0.0261	0.0271	0.0323	0.0326
			(0.0379)	(0.0373)	(0.2204)	(0.2272)	(0.2108)	(0.2185)	(0.0483)	(0.0489)	(0.0166)	(0.0176)	(0.0114)	(0.0114)

See table 2 for explanation of PM[a], OM[b] and $(\cdot)$[c].

To study the influences of heritabilities as well as signs of QTL effects on parameter estimating, in table 3 we listed the simulation results when $N = 500$, $\theta = 0.01, \gamma_{11} = 0.02, \gamma_{21} = 0.03$. It can be seen that with the decrease of heritability, the deviations of estimates $\hat{\beta_1}$, $\hat{\beta_2}$, $\hat{\gamma_{11}}$, $\hat{\gamma_{21}}$ increase and the corresponding MSEs grows higher and higher, but the new method still outperforms the OM. When QTL effects β_1 and β_2 have opposite signs, the estimates of $\hat{\alpha}$, $\hat{\sigma^2}$, $\hat{\gamma_{11}}$ and $\hat{\gamma_{21}}$ are closer to their own true values and have smaller MSEs than the case that β_1 and β_2 have same signs. We get the same conclusion when $\gamma_{11} = 0.05, \gamma_{21} = 0.06$ and $\gamma_{11} = 0.02, \gamma_{21} = 0.05$. So we conclude that the signs of β_1 and β_2 have influences on the accuracy of parameter estimation.

The estimates of error rate θ and the MSEs for different scenarios are listed in table 4. We can see that the MSE of $\hat{\theta}$ becomes higher and higher with the increase of θ for each group of heritability and QTL effects, which means that the accuracy becomes lower and lower. When β_1 and β_2 have opposite signs, $\hat{\theta}$ is closer to the true value and has lower MSE than the case that β_1 and β_2 have same signs. For the same θ, the values of MSE become lower and lower with the decrease of heritability.

Table 4. Estimates of error rate θ with different heritabilities and QTL effects.

			True value of θ			
h^2	β_1	β_2	0	0.01	0.05	0.1
0.05	−0.7314	0.3282	0.0003	0.0097	0.0494	0.1006
			$(< 10^{-4})$[a]	(0.0027)	(0.0021)	(0.0032)
0.05	0.7314	0.3282	0.0005	0.0092	0.0512	0.0987
			$(< 10^{-4})$	(0.0048)	(0.0213)	(0.0216)
0.1	−0.7314	0.9064	0.0008	0.0111	0.0508	0.1050
			$(< 10^{-4})$	(0.0074)	(0.0074)	(0.0086)
0.1	0.7314	0.9064	0.0010	0.0115	0.0411	0.0600
			$(< 10^{-4})$	(0.0082)	(0.0223)	(0.0482)
0.2	−0.7314	1.5866	0.0012	0.0114	0.0422	0.0800
			$(< 10^{-4})$	(0.0077)	(0.0189)	(0.0304)
0.2	0.7314	1.5866	0.0015	0.0084	0.0399	0.0562
			$(< 10^{-4})$	(0.0086)	(0.0221)	(0.0493)

$(\cdot)$[a], see table 2 for the explanation.

In the PM, variance estimate for estimate of each parameter can be obtained by calculating the inverse of the observed Fisher information matrix (Louis 1982). Therefore, we also construct confidence intervals for each parameter in our simulations. As expected, for each parameter the frequency that the confidence intervals included the true value of the parameter is close to the considered nominal value 95% (e.g., in 500 simulation replicates, the frequency that the confidence intervals of β_1 included the true value is equal to 0.936).

In addition, the simulation results are better when sample size, $N = 1000$ than $N = 500$ for both methods, i.e., with the increase of N, the values of MSE all decrease correspondingly. So increase in sample size will improve the performance of QTL mapping. But the proposed method is however better than the OM. Totally, the above simulation results suggest that the PM is an efficient mapping method when genotype data exist errors.

Real example

Here we analysed a real barley dataset by the PM in this study to show its practicability when performing multiple-interval mapping in presence of genotyping errors. The barley dataset (DH population, which is completely similar to BC population) was from the North American Genome Mapping Project (Tinker *et al.* 1996). The DH population contained 145 lines and the 1500 cM genome consisted of seven linkage groups, which included M =127 markers. The phenotype of kernel weight across the environment was mainly analysed here, to detect the latent QTL. Xu (2007) and Ma *et al.* (2011) also investigated the dataset. Part of the genotypes in the dataset were missing (<5%), and the missing values were imputed by the corresponding modes or directly by 0 in existing methods. Here we considered the imputed missing values as data with genotyping errors, and therefore the PM can be applied to deal with the dataset.

Because of the large number of markers, there existed difficulty in performing multiple-interval mapping directly. To reduce the dimension, we performed multiple-interval mapping based on the six significant markers (2, 12, 21, 43, 75 and 102) which were detected by larger effects in Ma *et al.* (2011). During the analysis, we used the above six markers as well as their neighbouring markers to construct marker intervals, and then the estimates of positions and effects of the six detected QTL were obtained by the PM, respectively: (5.22, 0.404)@7; (90.41, 0.325)@1; (1.51, 0.205)@1; (20.08, 0.166)@3; (119.17, 0.155)@5; (173.47, 0.124)@1, where the notation for the estimates of positions and effects, e.g., (5.22, 0.404)@7 indicates chromosome 7, position 5.22 cM and effect 0.404.

The first three primary QTL that we detected by the PM are consistent with the reported results in the published paper (Tinker *et al.* 1996). In Xu (2007), the first three significant markers detected are markers 1, 11 and 101, but the further detection result of QTL is not provided. Compared with the results of Ma *et al.* (2011), the parameter estimate results of the first three primary QTL are closer, but there exists some difference in the effect estimates of the forth, fifth and sixth QTL. We hope the new results of estimate by the PM can provide valuable references on the further detection and filtration of trait loci. Besides estimating QTL effects and positions, the estimate of error rate of marker genotype is also obtained in our analysis, i.e., $\hat{\theta} = 0.0364$, which approximately coincides with the result estimated by the number of missing values in the original dataset. All these estimate results by the PM show that it is effective on QTL mapping when using genotype data with errors.

Discussion and conclusion

Currently there are numerous studies on QTL mapping, most of which assume that the observed genotypes are correct. However, most genotype datasets contain certain measure errors. In this study, we propose an algorithm of simultaneously estimating all model parameters using genetic data with genotyping errors. A computer program written in MATLAB is available upon request. Simulation results show that the estimates of genetic parameters are exactly affected by the genotyping errors and the new method performs better on QTL mapping than the method that does not consider genotyping errors sufficiently. Heritability has influences on the accuracy of estimates of QTL effects and positions, which can be improved with relatively high heritability. The results of QTL mapping are also affected by the population size. In practice, expanding population scale will improve the accuracy of QTL mapping, which is consistent with the conclusion of Jeon (1995).

The PM can also be applied to cases of intercross families and markers with different error rates, in which we only need to adjust the the presentation of φ^i and the conditional probabilities (table 1) when conducting QTL mapping. The process of estimating parameter vector Ω is completely analogous with backcross case.

Here, we mainly consider the estimation problem of parameter vector Ω. Of course, after getting the estimate of Ω, we can further discuss whether the QTL significantly exist in the considered marker intervals. Let $H_0 : \beta_1 = \beta_2 = 0, H_1$: at least one interval has QTL. The test statistic is

$$\text{LOD} = Log_{10}[L(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma^2}, \hat{\gamma_{11}}, \hat{\gamma_{21}}, \hat{\theta}) / L(\tilde{\alpha}, 0, 0, \tilde{\sigma^2}, \tilde{\gamma_{11}}, \tilde{\gamma_{21}}, \tilde{\theta})].$$

A threshold needs to be chosen, which has significant influence on factual mapping. If the LOD score exceeds the threshold, it means that at least one QTL is declared to exist. While there are many factors that affect the threshold, such as sample size, genome size and density of markers, etc. (Lander and Botstein 1989). Thus the threshold should be chosen according to the practical situation. Permutation can be used to choose the threshold (Churchill and Doerge 1994), which can promise the accuracy of significant tests, although it is computation-intensive and time-consuming.

Indeed, the PM also has some shortcomings. Because of the genotyping errors and the unknown genotypes of latent QTL, our method may encounter a large amount of computation when the number of markers is large. To overcome this difficulty, we suggest using the idea of two-step method (Ma *et al.* 2011), i.e., the markers with larger effects are detected and retained in all markers at first, and then marker intervals are constructed by the selected markers to estimate all parameters simultaneously. Besides, the EM algorithm itself has limitation. For example, sometimes it has a slow convergence speed and the speed of convergence may depend on the initial values of parameters. We suggest that choosing different initial values in the estimating process and comparing each results to decrease the possible impact caused by initial values of parameters. Since QTL mapping plays an important role in the study of the genetic epidemiology, in our future work we will make further investigations and develop methods that are suitable for more markers.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (nos. 11201129 and 11371083), the Natural Science Foundation of Heilongjiang Province of China (A201207), the Scientific Research Foundation of Department of Education of Heilongjiang Province of China (no. 1253G044) and the Science and Technology Innovation Team in Higher Education Institutions of Heilongjiang Province (no. 2014TD005). This work was partly supported by the Science and Technology Programme of Suihua of China (no. KJZD20130112) and the Youth Fund Project in 2013 of Suihua University of China (KQ 1302004).

References

Abecasis G. R., Cherny S. S. and Cardon L. R. 2001 The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.* **9**, 130–134.

Akey J. M., Zhang K., Xiong M., Doris P. and Jin L. 2001 The effect that genotyping errors have on the robustness of common linkage disequilibrium measures. *Am. J. Hun. Genet.* **68**, 1447–1456.

Buetow K. H. 1991 Influence of aberrant observations on high resolution linkage analysis outcomes. *Am. J. Hum. Genet.* **49**, 985–994.

Cartwright D. A., Troggio M., Velasco R. and Gutin A. 2007 Genetic mapping in the presence of genotyping errors. *Genetics* **176**, 2521–2527.

Chen Z. H. 2005 The full EM algorithm for the MLEs of QTL effects and positions and their estimated variances in multiple-interval mapping. *Biometrics* **6**, 474–480.

Churchill G. A. and Doerge R. W. 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Dempster A. P., Laird N. M. and Rubin D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc., B* **39**, 1–38.

Douglas J. A., Boehnke M. and Lange K. 2000 A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am. J. Hum. Genet.* **66**, 1287–1297.

Goldstein D. R., Zhao H. and Speed T. P. 1997 The effects of genotyping errors and interference on estimation of genetics distance. *Hum. Hered.* **47**, 86–100.

Hou Y. J. 2011 Parameter estimation in quantitative traitloci mapping when using data with genotyping errors (*in Chinese*). M.Sc. dissertation. Heilongjiang University, Harbin, China.

Jeon G. J. 1995 The effects of population size and dominance of quatitative trait loci (QTL) on the detection of linkage between markers and QTL for livestock. *Asian Aust. J. Anim. Sci.* **8**, 651–655.

Kao C. H., Zeng Z. B. and Teasdale R. D. 1999 Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Lander E. S. and Botstein D. B. 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lebrec J. J., Putter H., Houwing-Duistermaat J. J. and van Houwelingen H. C. 2008 Influence of genotyping errors in linkage mapping for complex traits-an analytic study. *BMC Genet.* **9**, 57.

Louis T. A. 1982 Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc., B* **44**, 226–233.

Ma W. J., Zhou Y. and Zhang S. L. 2011 A two-step method for estimating QTL effects and positions in multi-marker analysis. *Genet. Res.* **93**, 115–124.

Sobel E., Papp J. C. and Lange K. 2002 Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* **70**, 496–508.

Tinker N. A., Mather D. E., Rossnagel B. G., Kasha K. J. and Kleinhofs A. 1996 Regions of the genome that affect agronomic performance in two-row barley. *Crop Sci.* **36**, 1053–1062.

Wang D. L., Zhu J., Li Z. K. L. and Paterson A. H. 1999 Mapping QTL with epistatic effects and QTL × environment interactions by mixed linear model approaches. *Theor. Appl. Genet.* **99**, 1255–1264.

Xu S. 2007 An empirical Bayes method for estimating eqistatic effects of quantitative trait loci. *Biometrics* **63**, 513–521.

Zeng Z. B. 1994 Precision mapping of quantitative trait loci. *Genetics* **135**, 1457–1468.

Zhou Y. 2010 Multiple interval mapping for quantitative trait loci via EM algorithm in the presence of crossover interference. *Commun. Stat.-Theor. Method* **39**, 3041–3057.

Received 28 April 2014, in revised form 3 August 2014; accepted 19 September 2014
Unedited version published online: 9 October 2014
Final version published online: 4 March 2015