Heredity (2005) 94, 119–128
© 2005 Nature Publishing Group All rights reserved 0018-067X/05 $30.00

www.nature.com/hdy

An EM algorithm for mapping quantitative resistance loci

C Xu, Y-M Zhang and S Xu

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Many disease resistance traits in plants have a polygenic background and the disease phenotypes are modified by environmental factors. As a consequence, the phenotypic values usually show a quantitative variation. The phenotypes of such disease traits, however, are often measured in discrete but ordered categories. These traits are called ordinal traits. In terms of disease resistance, they are called quantitative resistance traits, as opposed to qualitative resistance traits, and are controlled by the quantitative resistance loci (QRL). Classical quantitative trait locus mapping methods are not optimal for ordinal trait analysis because the assumption of normal distribution is violated. Methods for mapping binary trait loci are not suitable either because there are more than two categories in ordinal traits. We developed a maximum likelihood method to map these QRL. The method is implemented via a multicycle expectation-conditional-maximization (ECM) algorithm under the threshold model, where we can estimate both the QRL effects and the thresholds that link the disease liability and the categorical phenotype. The method is verified in simulated data under various combinations of the parameters. An SAS program is available to implement the multicycle ECM algorithm. The program can be downloaded from our website at www.statgen.ucr.edu.

Heredity (2005) **94,** 119–128. doi:10.1038/sj.hdy.6800583
Published online 15 September 2004

Keywords: ECM-algorithm; maximum likelihood; ordinal trait; quantitative resistance locus

Introduction

Many traits of biological and economic importance in plants, animals and human populations are measured in a discrete manner. For example, most disease resistance traits in plants, such as sheath blight resistance in rice (Zou *et al*, 2000), clubroot resistance in brassica napus (Manzanares-Dauleux *et al*, 2000) and cucumber mosaic virus resistance in pepper (Caranta *et al*, 2002), are all scored in several ordered categories, based on the magnitude of disease symptom. Similarly, there are many characters in animals and humans, such as scores for calving difficulty, expression of congenital malformations, numbers of reproductive events and so on, which are expressed as binary or ordinal traits. Although the expression of some discrete traits is a consequence of the expression of a single segregating factor, multiple loci are often involved (Lynch and Walsh, 1998). Naturally, we may postulate that a number of different genes along with a number of environmental variables act jointly as risk and protective factors for the trait development. When enough risk factors accumulate and greatly outweigh the protective factors, the trait phenotype develops. As many factors contribute to the trait variation, the liability or predisposition towards the trait is really a continuous and quantitative trait. Once the liability passes a certain critical point or threshold, the trait phenotype emerges. Attributes that are categorical on an outward (observed) scale but believed to be continuous on an underlying (unobserved) scale are called the threshold or quasi-continuous characters (Lynch and Walsh, 1998).

Rice sheath blight, caused by *Rhizoctonia solani* Kühn, is one of the three major diseases of rice and severely impairs both rice yields and quality. Resistance to sheath blight in the rice shows a quantitative nature, that is, different rice varieties show different degrees of resistance and the disease phenotypes usually overlap (Zou *et al*, 2000). The phenotypic value of sheath blight resistance is measured in grade, ranging from 0 (complete resistance) to 9 (complete susceptible) (Rush *et al*, 1976). However, the distribution of the grade severely deviates from normality. Therefore, classical quantitative genetics analysis for normal traits is not optimal for this type of ordinal traits. Binary trait analysis techniques are not suitable either because they cannot handle multiple categories. Therefore, new statistical methods are required to map such quantitative resistance loci (QRL).

A number of statistical methods are now available to map quantitative trait locus (QTL) for continuous traits (Lander and Botstein, 1989; Haley and Knott, 1992; Jansen, 1993; Zeng, 1994; Kao *et al*, 1999), but relatively little work has been carried out on mapping ordinal traits (Xu and Atchley, 1996; Visscher *et al*, 1996; Galecki *et al*, 2001; Xu *et al*, 2003), especially for multiple ordinal traits (Hackett and Weller, 1995; Rao and Xu, 1998; Rao and Li, 2001). Genetic analysis for ordinal categorical traits is difficult because the observed phenotype (category) cannot be described by a straightforward linear model. Hackett and Weller (1995) developed an approximate logistic regression method using the threshold model to map QTL for such traits in backcross (BC) population.

Correspondence: S Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
E-mail: xu@genetics.ucr.edu

Received 11 August 2003; accepted 26 July 2004; published online 15 September 2004

Rao and Xu (1998) also proposed a similar method for QTL mapping in four-way crosses. Rao and Li (2001) further extended the methods to map QTL using independent multiple families. All the above methods were implemented using either a different statistical model from traditional QTL mapping or a different optimization algorithm from the commonly used EM algorithm. We found that the models and the optimization algorithms for mapping ordinal traits and quantitative traits can be formulated in the same framework. Thus, QTL mapping and ordinal trait mapping can be unified under the same statistical framework. The same linear model and EM algorithm can be used for both types of traits. We also found that the multicycle expectation-conditional-maximization (ECM) algorithm developed by Meng and Rubin (1993), an extended EM algorithm, is more intuitive and easy to understand to the QTL mapping community. In addition, the multicycle ECM algorithm can be easily programmed in computers. Therefore, the objective of this study is to introduce such an ECM algorithm for mapping ordinal traits.

Mapping populations that can be handled in most statistical methods involve only two inbred lines. The drawback of these designs is that the statistical inference space is quite narrow (within the two inbred lines), and thus results from one cross cannot be generalized to other crosses derived from different inbred lines. Xu (1996, 1998) proposed the four-way cross design of QTL mapping, intended to increase the statistical inference space and the opportunity for detecting more QTL. We found that the different designs of line cross can be incorporated into a unified QTL mapping strategy that is aimed to handle a four-way cross family but treat commonly used mapping populations, such as F_2 and BC, as special cases. In this study, we will discuss how to implement this strategy.

Theory and methods

Statistical model for ordinal traits

Consider n individuals in the mapping population and denote the observed ordered category of individual j by w_j where $j=1, 2, \ldots, n$. For C categories, the phenotype of an individual can be defined as $w_j=c$ if individual j belongs to class c, for $c=1,\ldots, C$. A set of fixed thresholds, $t_1, t_2,\ldots, t_{C-1}$, on the underlying scale define the observed categories on an ordinal scale $1, 2, \ldots, C$. Further define y_j as the underlying variable for individual j. We thus have model

$$t_{c-1}<y_j\leq t_c \Leftrightarrow w_j=c; \quad t_0=-\infty; \quad t_C=\infty \tag{1}$$

Here we have $C+1$ thresholds but only $C-1$ thresholds, $\mathbf{t}=\{t_1, t_2,\ldots, t_{C-1}\}$, which are parameters that are subject to estimation.

Although the natural choice for the distribution of y would be the normal distribution, Hackett and Weller (1995), Rao and Xu (1998) and Rao and Li (2001) all used logistic distribution to approximate the normal distribution for the purpose of computational simplicity. In contrast to the above methods, we directly use the normal distribution. The underlying variable y is assumed to be a continuous variable similar to the phenotypic value of a common quantitative trait. The only difference is that y_j is not observable but inferred from the observed phenotype of individual j. As a quantitative trait, y_j can be described by the linear model

$$y_j=\mathbf{X}_j\mathbf{b}+\mathbf{Z}_j\mathbf{u}+e_j \tag{2}$$

where $\mathbf{b}$ is a vector of nongenetic effects, for example, block and year effects in plant or sex and age effects in animals, $\mathbf{X}_j$ is a known design matrix for the nongenetic effects, $\mathbf{u}$ is a vector of genetic effects, $\mathbf{Z}_j$ is the design matrix for the genetic (QTL) effects, and e_j is a random environmental effect defined as a standardized normal variable.

Under this assumption, the probability that individual j is classfied into the cth category is

$$\begin{aligned}\Pr(w_j=c|\mathbf{Z}_j,\mathbf{t},\mathbf{b},\mathbf{u}) &= \Pr(t_{c-1}<y_j\leq t_c|\mathbf{Z}_j,\mathbf{t},\mathbf{b},\mathbf{u}) \\ &= \Phi(t_c-\mathbf{X}_j\mathbf{b}-\mathbf{Z}_j\mathbf{u}) \\ &\quad -\Phi(t_{c-1}-\mathbf{X}_j\mathbf{b}-\mathbf{Z}_j\mathbf{u})\end{aligned} \tag{3}$$

where $\Phi(t_c-\mathbf{X}_j\mathbf{b}-\mathbf{Z}_j\mathbf{u})$ is the standard normal cumulative distribution function. The above multiple thresholds model for ordinal trait provides a link between w_j and y_j. If we know the thresholds, mapping QTL for ordinal categorical trait has been formulated as a problem of mapping QTL for regular quantitative trait. The difficulty, however, is that these thresholds are unknown and must be estimated simultaneously along with the QTL effects.

Genetic model of a four-way cross

The genetic model is developed based on a four-way cross design because backcross and F_2 designs are shown to be special cases of such a general design. The genetic model for a four-way cross has been proposed by Xu (1996, 1998). In order for the paper to be self-contained, these models are summarized and described here. Let L_1 and L_2 be the two inbred lines initiating the first cross and L_3 and L_4 be the inbred lines intiating the second cross. Denote the QTL genotypes of L_1 and L_2 by $Q_1^mQ_1^m$ and $Q_2^mQ_2^m$, respectively, and the genotypes of L_3 and L_4 by $Q_1^fQ_1^f$ and $Q_2^fQ_2^f$, respectively. The genetic constitution of the four-way cross population will consist of four genotypes: $Q_1^mQ_1^f$, $Q_1^mQ_2^f$, $Q_2^mQ_1^f$ and $Q_2^mQ_2^f$, with equal frequency. Let G_{ab} be the value of genotype $Q_\mathrm{a}^mQ_\mathrm{b}^f$ where $a,b=1,2$, and it can be expressed by the following linear model:

$$\mathbf{G}=\mathbf{H}\mathbf{u} \tag{4}$$

where

$$\mathbf{G}=\begin{bmatrix}G_{11}\\G_{12}\\G_{21}\\G_{22}\end{bmatrix},\ \mathbf{H}=\begin{bmatrix}1&1&1\\1&-1&-1\\-1&1&-1\\-1&-1&1\end{bmatrix} \text{ and } \mathbf{u}=\begin{bmatrix}a^m\\a^f\\d\end{bmatrix}.$$

The three elements of vector $\mathbf{u}$ are defined as the additive effect for the maternal parent, the additive effect for the paternal parent and the dominance effect of the QTL, respectively. Let $\mathbf{H}_g$ be the gth row of matrix $\mathbf{H}$, then $G_{11}=\mathbf{H}_1\mathbf{u}$, $G_{12}=\mathbf{H}_2\mathbf{u}$, $G_{21}=\mathbf{H}_3\mathbf{u}$ and $G_{22}=\mathbf{H}_4\mathbf{u}$.

We now connect the threshold model and the genetic model in the four-way cross. In model (2), $\mathbf{Z}_j=\mathbf{H}_1$ if individual j takes the first genotype $Q_1^mQ_1^f$ and $\mathbf{Z}_j=\mathbf{H}_2$ if j takes the second genotype $Q_1^mQ_2^f$ and so on. Model (2) is a general linear model (GLM) with missing value in $\mathbf{Z}_j$ because the genotype of j is not observable.

The next step of the GLM analysis with missing value is to infer the probabilities of QTL genotypes conditional on marker information, denoted by $p_{jg}^{(0)}=\Pr(\mathbf{Z}_j=\mathbf{H}_g\,|\,\mathbf{I}_M)$

for $g = 1, \ldots, 4$ where $\mathbf{I}_M$ represents marker information. Multipoint method (Rao and Xu, 1998) can be used to infer the conditional probabilities of QTL genotypes. This method is the same as Jiang and Zeng (1997) in dealing with missing or partially informative markers and can be implemented in a simple way.

Maximum likelihood estimation (MLE)

Let us denote the parameters by a vector $\boldsymbol{\theta} = \{\mathbf{t}, \mathbf{b}, \mathbf{u}\}$. The probability of phenotype for the jth individual conditional on $\mathbf{Z}_j$ is

$$\begin{aligned}\Pr(w_j|\mathbf{Z}_j, \boldsymbol{\theta}) = &\Phi(t_c - \mathbf{X}_j\mathbf{b} - \mathbf{Z}_j\mathbf{u}) \\ &- \Phi(t_{c-1} - \mathbf{X}_j\mathbf{b} - \mathbf{Z}_j\mathbf{u}) \quad \text{for } w_j = c\end{aligned} \tag{5}$$

Since $\mathbf{Z}_j$ is missing and only $p_{jg}^{(0)}$ can be calculated, the actual likelihood function for the jth individual is

$$\begin{aligned}\Pr(w_j|\boldsymbol{\theta}) = &\sum_{g=1}^{4} p_{jg}[\Phi(t_c - \mathbf{X}_j\mathbf{b} - \mathbf{H}_g\mathbf{u}) \\ &- \Phi(t_{c-1} - \mathbf{X}_j\mathbf{b} - \mathbf{H}_g\mathbf{u})]\end{aligned} \tag{6}$$

The overall log likelihood for the entire mapping population is

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log[\Pr(w_j|\boldsymbol{\theta})] \tag{7}$$

Solving the above log likelihood function is tedious. We now introduce a multicycle ECM algorithm (Meng and Rubin, 1993) to find the solution. The multicycle ECM algorithm is to perform one E step before each CM step or a few selected CM steps. A cycle is defined as one E step followed by one CM step. The proposed multicycle ECM solution takes advantage of the simplicity of the original linear model with both y_j and $\mathbf{Z}_j$ being treated as missing values.

If $\mathbf{Z}_j$ and y_j were observed for every individual, the estimates of the parameters $\mathbf{b}$ and $\mathbf{u}$ at the $(k+1)$th iteration could be found explicitly using the following iterative equations by the two conditional maximizatiom (CM) steps:

$$\begin{aligned}\mathbf{b}^{(k+1)} &= \left[\sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\right]^{-1}\left[\sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}(y_j - \mathbf{Z}_j\mathbf{u}^{(k)})\right] \\ \mathbf{u}^{(k+1)} &= \left[\sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right]^{-1}\left[\sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}(y_j - \mathbf{X}_j\mathbf{b}^{(k+1)})\right]\end{aligned} \tag{8}$$

In QTL mapping for continuous traits, $\mathbf{Z}_j$ is missing but the distribution of $\mathbf{Z}_j$ is given, the ECM algorithm can be adopted to take advantage of the above equations. The ECM equations simply replace all the terms related to $\mathbf{Z}_j$ by their expectations, that is,

$$\begin{aligned}\mathbf{b}^{(k+1)} &= \left[\sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\right]^{-1}\left[\sum_{j=1}^{n} E[\mathbf{X}_j^{\mathrm{T}}(y_j - \mathbf{Z}_j\mathbf{u}^{(k)})]\right] \\ \mathbf{u}^{(k+1)} &= \left[\sum_{j=1}^{n} E(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j)\right]^{-1}\left[\sum_{j=1}^{n} E[\mathbf{Z}_j^{\mathrm{T}}(y_j - \mathbf{X}_j\mathbf{b}^{(k+1)})]\right]\end{aligned} \tag{9}$$

The expectations are obtained conditional on both marker information and the value of liability y_j. The connection between the phenotype and the QTL genotype is through the parameter values, but the parameters are what we are trying to find. Therefore, we need iterations on equation (9) by providing some initial values of the parameters to start the iteration. This is the ECM algorithm. The E step is to find the expectations and the CM step is to invoke equation (9) for iterations.

Recall that the probability of $\mathbf{Z}_j$ conditional on marker information is denoted by $p_{jg}^{(0)}$. This probability may be called the prior probability. After incorporating the phenotypic value, we obtain the posterior probability at the $(k+1)$th iteration, denoted by

$$\begin{aligned}p_{jg}^{(k+1)} &= \Pr(\mathbf{Z}_j = \mathbf{H}_g|\mathbf{I}_M, y_j) \\ &= \frac{p_{jg}^{(0)}\phi(y_j - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)})}{\sum_{h=1}^{4} p_{jh}^{(0)}\phi(y_j - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_h\mathbf{u}^{(k)})}\end{aligned} \tag{10}$$

where

$$\begin{aligned}&\phi(y_j - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)}) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(y_j - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)})^2\right]\end{aligned}$$

is the standardized normal density. Note that the prior probability $p_{jg}^{(0)}$ in equation (10) is used for all iterations to calculate the posterior probability.

The expectations are actually obtained using the posterior probabilities rather than the prior probabilities. Therefore,

$$\begin{aligned}E[\mathbf{X}_j^{\mathrm{T}}(y_j - \mathbf{Z}_j\mathbf{u}^{(k)})] &= \sum_{g=1}^{4} p_{jg}^{(k+1)}\mathbf{X}_j^{\mathrm{T}}(y_j - \mathbf{H}_g\mathbf{u}^{(k)}) \\ E(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j) &= \sum_{g=1}^{4} p_{jg}^{(k+1)}\mathbf{H}_g^{\mathrm{T}}\mathbf{H}_g \\ E[\mathbf{Z}_j^{\mathrm{T}}(y_j - \mathbf{X}_j\mathbf{b}^{(k+1)})] &= \sum_{g=1}^{4} p_{jg}^{(k+1)}\mathbf{H}_g^{\mathrm{T}}(y_j - \mathbf{X}_j\mathbf{b}^{(k+1)})\end{aligned} \tag{11}$$

The problem here is that y_j is also missing for ordinal traits. Thus, we need to use $\hat{y}$, the expectation of y conditional on w, $\mathbf{Z}_j$ and $\boldsymbol{\theta}$, in place of y for the estimation of $\boldsymbol{\theta}$ before each CM step and this becomes the multicycle ECM.

As $\mathbf{t}$ contains a set of parameters different from $\mathbf{u}$ and $\mathbf{b}$, we now build the equations as follows. The solution for $\mathbf{b}$ and $\mathbf{u}$ conditional on $\mathbf{t}$ at the $(k+1)$th iteration is

$$\begin{aligned}\mathbf{b}^{(k+1)} &= \left[\sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\right]^{-1}\left[\sum_{j=1}^{n} E[\mathbf{X}_j^{\mathrm{T}}(\hat{y}_j^{(k+1)} - \mathbf{Z}_j\mathbf{u}^{(k)})]\right] \\ \mathbf{u}^{(k+1)} &= \left[\sum_{j=1}^{n} E(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j)\right]^{-1}\left[\sum_{j=1}^{n} E[\mathbf{Z}_j^{\mathrm{T}}(\hat{y}_j^{(k+1)} - \mathbf{X}_j\mathbf{b}^{(k+1)})]\right]\end{aligned} \tag{12}$$

Before taking the above CM steps, we first need to calculate the corresponding expectations by

$$\begin{aligned} &E[\mathbf{X}_j^{\mathrm{T}}(\hat{y}_j^{(k+1)} - \mathbf{Z}_j\mathbf{u}^{(k)})] \\ &\quad= \sum_{g=1}^{4} p_{jg}^{(k+1)}[\mathbf{X}_j^{\mathrm{T}}(\hat{y}_{jg}^{(k+1)} - \mathbf{H}_g\mathbf{u}^{(k)})] \\ &E[\mathbf{Z}_j^{\mathrm{T}}(\hat{y}_j^{(k+1)} - \mathbf{X}_j\mathbf{b}^{(k+1)}) \\ &\quad= \sum_{g=1}^{4} p_{jg}^{(k+1)}[\mathbf{H}_g^{\mathrm{T}}(\hat{y}_{jg}^{(k+1)} - \mathbf{X}_j\mathbf{b}^{(k+1)})] \end{aligned} \qquad (13)$$

where

$$\begin{aligned} \hat{y}_{jg}^{(k+1)} &= E(y_j|w_j, \boldsymbol{\theta}^{(k)}, \mathbf{Z}_j = \mathbf{H}_g) \\ &= \mathbf{X}_j\mathbf{b}^{(k)} + \mathbf{H}_g\mathbf{u}^{(k)} + \sum\nolimits_{c=1}^{C} i_{(w_j=c)} \\ &\times \frac{\phi(t_{c-1}^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)}) - \phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)})}{\Phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)}) - \Phi(t_{c-1}^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)})} \end{aligned} \qquad (14)$$

The quantity $i_{(w_j=c)}$ in (14) is an indicator variable and defined as one for $w_j = c$ and zero otherwise. Formulae (12)–(14) consist of the first cycle in our multicycle ECM algorithm.

A closed form for the exact solution of $\mathbf{t}$ is hard to define. However, an explicit solution can be approximated. The solution of the cth element of $\mathbf{t}$, for $c = 1, \ldots, (C-1)$, conditional on $\mathbf{b}$ and $\mathbf{u}$ at the $(k+1)$th iteration is

$$t_c^{(k+1)} = -\frac{1}{n}\sum_{j=1}^{n} E\left[\hat{y}_j^{(k+1)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{Z}_j\mathbf{u}^{(k+1)}\right] \qquad (15)$$

where

$$\begin{aligned} &E\left[\hat{y}_j^{(k+1)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{Z}_j\mathbf{u}^{(k+1)}\right] \\ &\quad= \sum_{g=1}^{4} p_{jg}^{(k+1)}\left[\hat{y}_{jg}^{(k+1)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{H}_g\mathbf{u}^{(k+1)}\right] \end{aligned} \qquad (16)$$

and

$$\begin{aligned} \hat{y}_{jg}^{(k+1)} =& E(y_{jg}|w_j, \mathbf{b}^{(k+1)}, \mathbf{u}^{(k+1)}, \mathbf{t}^{(k)}, \mathbf{Z}_j = \mathbf{H}_g) \\ =& \mathbf{X}_j\mathbf{b}^{(k+1)} + \mathbf{H}_g\mathbf{u}^{(k+1)} - t_c^{(k)} \\ & - i_{w_j \le c}\frac{\phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{H}_g\mathbf{u}^{(k+1)})}{\Phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{H}_g\mathbf{u}^{(k+1)})} \\ & + i_{w_j > c}\frac{\phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{H}_g\mathbf{u}^{(k+1)})}{1 - \Phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k+1)} - \mathbf{H}_g\mathbf{u}^{(k+1)})} \end{aligned} \qquad (17)$$

The quantity $i_{(w_j \le c)}$ in (17) is also an indicator variable and defined as one for $w_j \le c$ and zero otherwise. Here we have (C−1) thresholds to estimate and thus have (C−1) ECM cycles. In each cycle, we first calculate the expectation using equations (16) and (17) and then estimate the threshold by using equation (15). Therefore, we have a total of C ECM cycles in one iteration for estimation of all the parameters.

Note that $p_{jg}^{(k+1)}$ used in equations (13) and (16) for ordinal traits are different from that used in equation (10) for quantitative traits. The $p_{jg}^{(k+1)}$ used here for ordinal traits is

$$\begin{aligned} p_{jg}^{(k+1)} &= \Pr(\mathbf{Z}_j = \mathbf{H}_g|\mathbf{I}_M, w_j) \\ &= \frac{p_{jg}^{(0)}\left(\sum_{c=1}^{C} i_{(w_j=c)}[\Phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)}) - \Phi(t_{c-1}^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_g\mathbf{u}^{(k)})]\right)}{\sum_{h=1}^{4} p_{jh}^{(0)}\left(\sum_{c=1}^{C} i_{(w_j=c)}[\Phi(t_c^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_h\mathbf{u}^{(k)}) - \Phi(t_{c-1}^{(k)} - \mathbf{X}_j\mathbf{b}^{(k)} - \mathbf{H}_h\mathbf{u}^{(k)})]\right)} \end{aligned} \qquad (18)$$

The calculation begins with some starting values for $\mathbf{b}^{(0)}$, $\mathbf{u}^{(0)}$, $\mathbf{t}^{(0)}$ and $p_{jg}^{(0)}$. Iterations are then made between (18), (14), (13), (12), (17), (16), and (15) and terminated until a predetermined convergence criterion is satisfied. The MLE of parameters are denoted as $\hat{\mathbf{b}}, \hat{\mathbf{u}}$ and $\hat{\mathbf{t}}$, which will then be used for the calculation of the maximum likelihood value for hypothesis testing.

Likelihood ratio test statistic

Define the log-likelihood value evaluated at the MLE of parameters as

$$L(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n} \log[\Pr(w_j|\hat{\boldsymbol{\theta}})] \qquad (19)$$

where

$$\Pr(w_j|\hat{\boldsymbol{\theta}}) = \sum_{g=1}^{4} p_{jg}^{(0)} \Pr(w_j|\mathbf{Z}_j, \hat{\boldsymbol{\theta}})$$

and

$$\begin{aligned} \Pr(w_j|\mathbf{Z}_j, \hat{\boldsymbol{\theta}}) = \sum_{c=1}^{C} i_{(w_j=c)}[&\Phi(t_c - \mathbf{X}_j\hat{\mathbf{b}} - \mathbf{Z}_j\hat{\mathbf{u}}) \\ &- \Phi(t_{c-1} - \mathbf{X}_j\hat{\mathbf{b}} - \mathbf{Z}_j\hat{\mathbf{u}})] \end{aligned}$$

This is also called the likelihood value under the full model. We need the likelihood values under various restricted models to test various hypotheses.

The overall null hypothesis is no effect of QTL at the locus of interest, denoted by H_0: $a^m = a^f = d = 0$ or H_0: $\mathbf{Lu} = \mathbf{0}$, where

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

If we solve the MLE of the parameters under the restriction of $\mathbf{Lu} = \mathbf{0}$ and evaluate the likelihood value at the solutions with this restriction, we have

$$L(\hat{\boldsymbol{\theta}}|\mathbf{Lu} = \mathbf{0}) \qquad (20)$$

The likelihood ratio test statistic is

$$\Lambda = -2[L(\hat{\boldsymbol{\theta}}|\mathbf{Lu} = \mathbf{0}) - L(\hat{\boldsymbol{\theta}})] \qquad (21)$$

Various other test statistics can be defined by redefining the $\mathbf{L}$ matrix. To test the hypothesis of H_1: $a^m = 0$, we define $\mathbf{L}_1 = [1\,0\,0]$. The likelihood ratio test statistic is $\Lambda = -2[L(\hat{\boldsymbol{\theta}}|\mathbf{L}_1\mathbf{u} = 0) - L(\hat{\boldsymbol{\theta}})]$. To test the hypothesis of H_2: $a^f = 0$, we define $\mathbf{L}_2 = [0\,1\,0]$ and use $\Lambda = -2[L(\hat{\boldsymbol{\theta}}|\mathbf{L}_2\mathbf{u} = 0) - L(\hat{\boldsymbol{\theta}})]$. Similarly, we use $\Lambda = -2[L(\hat{\boldsymbol{\theta}}|\mathbf{L}_3\mathbf{u} = 0) - L(\hat{\boldsymbol{\theta}})]$ to test the hypothesis of H_3: $d = 0$ where $\mathbf{L}_3 = [0\,0\,1]$.

Extension to F_2 and BC populations

The four-way cross model is a general model from which the F_2 and BC models are considered as special cases. Let us first consider a BC population. The genotypes of the two parents of the BC family is defined as $Q_1^m Q_2^f \times Q_1^m Q_1^m$ or $Q_1^m Q_2^f \times Q_2^f Q_2^f$, depending on which inbred line is used as the tester. The constitution of genotypes of the mating pair may be called the mating type. Let us assume that $Q_1^m Q_2^f \times Q_2^f Q_2^f$ is the parental mating type for the BC family. A progeny from this mating type can take one of the four possible genotypes: $Q_1^m Q_2^f$, $Q_1^m Q_2^f$, $Q_2^f Q_2^f$ and $Q_2^f Q_2^f$. Note that the first and the second genotypes are not distinguishable, and neither are the third and the fourth. If we use the same notation as that of the four-way cross for the four genotypic values, we have $G_{11} = G_{12}$ and $G_{21} = G_{22}$. The genetic effects defined in the notation of a four-way cross are $a^m = G_{11} + G_{12} - G_{21} - G_{22}$, $a^f = G_{11} - G_{12} + G_{21} - G_{22} = 0$ and $d = G_{11} - G_{12} - G_{21} + G_{22} = 0$. Therefore, we can use the same four-way cross model for the BC mapping with the restriction of $a^f = d = 0$. This can be acomplished by searching for the MLE of the four-way cross model with the restriction of $\mathbf{Lu} = \mathbf{0}$, where

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

All marker genotypes are considered as either partially informative (when typed) or noninformative (when missing), and thus the same multipoint method can be used to infer the QTL genotype of a putative position using all markers.

Let us now consider an F_2 population. The genotypes of the two parents of the F_2 family can be defined as $Q_1^m Q_2^f \times Q_1^m Q_2^f$. A progeny from this parental mating type can take one of the four possible genotypes: $Q_1^m Q_1^m$, $Q_1^m Q_2^f$, $Q_2^f Q_1^m$ and $Q_2^f Q_2^f$. Note that the second and the third genotypes are not distinguishable. If we use the same notation as that of the four-way cross for the four genotypic values, we have $G_{12} = G_{21}$. The genetic effects defined in the four-way cross are $a^m = G_{11} + G_{12} - G_{21} - G_{22}$, $a^f = G_{11} - G_{12} + G_{21} - G_{22}$ and $d = G_{11} - G_{12} - G_{21} + G_{22}$. As $G_{12} = G_{21}$, we have $a^m = a^f$. Therefore, we can use the same four-way cross model for the F_2 mapping with the restriction of $a^m = a^f$. This can be acomplished by searching for the MLE of the four-way cross model with $\mathbf{Lu} = \mathbf{0}$, where $\mathbf{L} = [1\ -1\ 0]$. A marker genotype is considered as fully informative if it is homozygous. A heterozygous genotype is considered as partially informative because we cannot tell the difference between the second and the third genotypes. The same multipoint method can be used to infer the QTL genotype of a putative position.

Simulation studies

We designed a series of simulation experiments to verify the proposed multicycle ECM algorithm and the computer program. Since F_2 and BC populations are special cases of the four-way cross design, for the purpose of simplicity, we only simulated a BC population. We assumed that the liability of a BC population has a zero mean and unity residual variance. A single QTL was placed at position 25 cM (between markers 3 and 4) of a chromosome with 100 cM long covered by 11 evenly distributed markers. For the single QTL model, the QTL variance is defined as a^2, where a is the QTL effect (the difference of the allelic values of the segregating parent of the BC progeny). If the segregating parent is the female parent, $a = a^f$, otherwise, $a = a^m$ (see the notation in the previous paragraph). The QTL variance in the traditional BC analysis is $a^2/4$, which is different from what we defined here. This is because we defined the genotype indicator variable as 1 and -1 for the two alternative genotypes, whereas the genotype indicator variable is defined as 1 and 0 for the two alternative genotypes in the traditional BC analysis (Lynch and Walsh, 1998). The total variance of the liability is $\sigma_y^2 = a^2 + 1$ because the environmental variance of the liability is defined as 1. The proportion of the liability variance explained by the QTL is called the QTL heritability and is denoted by $h^2 = a^2/(a^2 + 1)$.

Comparison with logistic regression

In the first simulation experiment, we simulated five ordered categories ($C = 5$) with four threshold values. The four thresholds were chosen by trial and error so that the frequencies of the five categories occuring in the BC population have a ratio of 1:2:4:2:1. These threshold values depend on the genetic effects of the simulated QTL. The QTL effect was set at four levels, that is, $a = 0.2294$, 0.3333, 0.5000, 0.8165, so that the corresponding heritabilities at the four levels are $h^2 = 0.05$, 0.10, 0.20, 0.40, respectively. The simulated thresholds and the genetic effects are given in Table 1. The sample size of the BC population was $n = 300$. The simulation was replicated 100 times so that we can compare the empirical statistical powers, the mean estimated parameters and the standard errors of the estimates for different levels of the heritabilities. The critical values of the test statistic used to declare statistical significance at the 5% experiment-wise type I error rate were calculated from the approximate method of Piepho (2001). The empirical statistical power was calculated as the proportion of the simulated samples among the 100 replicates with the highest test statistical value along the genome greater than the approximate critical value.

Logistical regression analysis was the only existing method available for ordinal trait QTL mapping (Hackett and Weller, 1995; Rao and Xu, 1998). For each simulated sample, we also analyzed the data with the method of Rao and Xu (1998), which was implemented via the simplex algorithm (Nelder and Mead, 1965) for direct maximization of the likelihood function. In order to compare the estimated parameters of the logistic analysis with the probit model proposed here, the estimated QTL effect obtained from the logistic regression was multiplied by a constant $\sqrt{3}/\pi$ (Hackett and Weller, 1995). Results of both analyses are given in Table 1. The estimated parameters are close to the true values simulated for both methods. However, the estimated QTL effects and the threshold values from the logistic regression are slightly biased downwards due to the approximation of the constant factor $\sqrt{3}/\pi$. The statistical powers of the two methods are also comparable and both follow the expected trend that larger QTL tends to have a higher power to be detected. The estimated QTL positions for both methods are slightly biased and with a large estimation error when the QTL is small, which follows the usual expectation of QTL mapping studies.

Effect of the number of categories on QTL mapping

In the second simulation experiment, we evaluated the effect of the number of categories on the result of QTL

Table 1 Comparison of the new method of QTL mapping with the logistic regression analysis

Heritability (h^2)	*Parameter*			*Estimation with the new method*				*Estimation with the logistic regression*			
	Name		*True value*	*Mean*	*STD*[a]	*Power (%)*	*Position*[b]	*Mean*	*STD*[a]	*Power (%)*	*Position*[b]
0.05	Threshold	t_1	−1.3152	−1.3273	0.0954	74	28.77 (15.47)	−1.2583	0.1028	74	28.28 (15.83)
		t_2	−0.5383	−0.5298	0.0697			−0.4765	0.0649		
		t_3	0.5383	0.5478	0.0850			0.4933	0.0786		
		t_4	1.3152	1.3329	0.1068			1.2639	0.1142		
	QTL effect		0.2294	0.2355	0.0707			0.2224	0.0688		
0.10	Threshold	t_1	−1.3524	−1.3716	0.1105	97	25.34 (5.78)	−1.3018	0.1139	96	25.23 (5.60)
		t_2	−0.5542	−0.5557	0.0812			−0.5044	0.0748		
		t_3	0.5542	0.5555	0.0818			0.5052	0.0758		
		t_4	1.3524	1.3699	0.1140			1.3015	0.1206		
	QTL effect		0.3333	0.3381	0.0694			0.3216	0.0668		
0.20	Threshold	t_1	−1.4394	−1.4431	0.0958	100	25.14 (3.82)	−1.3670	0.0979	100	25.29 (3.84)
		t_2	−0.5932	−0.5953	0.0842			−0.5513	0.0797		
		t_3	0.5932	0.6121	0.0741			0.5672	0.0710		
		t_4	1.4394	1.4590	0.0989			1.3851	0.1037		
	QTL effect		0.5000	0.5080	0.0646			0.4842	0.0648		
0.40	Threshold	t_1	−1.6807	−1.6887	0.1112	100	25.04 (1.63)	−1.6202	0.1100	100	25.07 (1.68)
		t_2	−0.7212	−0.7237	0.0913			−0.7115	0.0860		
		t_3	0.7212	0.7411	0.0844			0.7272	0.0864		
		t_4	1.6807	1.7180	0.1084			1.6414	0.1154		
	QTL effect		0.8165	0.8330	0.0719			0.8144	0.0783		

[a] STD stands for the standard deviation of the estimated parameters obtained from 100 replicated simulations.
[b] The true QTL position is at 25 cM of the simulated chromosome. The standard deviations of the estimated QTL positions (obtained from 100 replicates) are given in parentheses.

mapping. The design of the simulation was similar to that described in the first paragraph of the section of simulation studies. We now set the QTL effect at $a = 0.3333$ so that $h^2 = 0.10$. We simulated three levels for the number of categories, 2, 5 and 8, corresponding to 1, 4 and 7 different threshold values (see Table 2 for the simulated threshold values). The frequency ratios of the categories in the three sets of simulations were 1:1, 1:2:4:2:1, and 1:2:3:4:4:3:2:1, respectively, for the three sets of threshold values. The sample size for the BC population was fixed at $n = 200$. The simulations was replicated 100 times for each setting. The results are given in Table 2, which shows that the number of phenotypic categories does not have a dramatic effect on the estimate of the QTL effect and position, but it does affect the statistical power. Increasing the number of categories tends to increase the statistical power. This result may be explained by the fact that increasing the number of categories has increased the information of predicting the liability from the observed categorical phenotype. If the number of categories had been increased to infinity, we would have observed the liability, and thus the power would reach that of QTL mapping for continuous traits. In reality, however, it is impossible to handle a large number of categories because we may encounter a problem of overparameterization due to the large number of thresholds to be estimated. For a large number of categories, the phenotype should be treated as a continuous trait and analyzed using a classical QTL mapping procedure.

Effect of the size of QTL on the result of QTL mapping

This simulation experiment intends to evaluate the effect of QTL size on the result of QTL mapping under a sample size of $n = 200$, which is typically used in QTL mapping experiments. The parameters simulated in this experiment are identical to those reported in the paragraph under the title of 'comparison with logistical regression,' except $n = 200$. The results are given in Table 3. Again, a general trend of higher statistical power for higher heritability was observed. In addition, the QTL position is more precisely estimated for higher heritability than for lower heritability.

Effect of phenotypic distribution on QTL mapping

In this simulation experiment, we investigated the effect of the shape of phenotypic distribution on the result of QTL mapping under a fixed sample size ($n = 200$), a given number of categories ($C = 5$) and a given size of QTL ($a = 0.3333$, that is, $h^2 = 0.10$). We choose the set of threshold values by trial and error so that the phenotypic frequency ratios of the five categories were 1:1:1:1:1 for the first set (uniform distribution), 1:2:4:2:1 for the second set (symmetrical and bell-shaped distribution) and 6:4:3:1:1 for the third set (highly skewed distribution). The simulated threshold values as well as the estimated parameters from 100 replicated simulations are given in Table 4. We found that skewed distribution has decreased the statistical power. The optimal power occurred in the situation where the phenotypic distribution is bell-shaped.

Effect of sample size on QTL mapping

Finally, we investigate the effect of sample size on the result of QTL mapping when the QTL size was fixed at $a = 0.3333$ ($h^2 = 0.10$), the number of categories was $C = 5$ and the shape of the phenotypic distribution was 1:2:4:2:1. We evaluated four levels of sample sizes: 100, 200, 300 and 500. Results of 100 replicated simulations

Table 2 Mean and standard deviation (STD) of the estimated threshold values and QTL effect for various number of phenotypic categories (*C*)

C	*Parameter*			*Estimates*			
	Name		*True value*	*Mean*	*STD*	*Power (%)*	*Position (cM)*
2	Threshold	t_1	0.0000	0.0083	0.0986	77	24.96 (9.92)
	QTL effect		0.3333	0.3547	0.0855		
5	Threshold	t_1	−1.3524	−1.3603	0.1361	86	25.59 (6.96)
		t_2	−0.5542	−0.5578	0.0974		
		t_3	0.5542	0.5578	0.1039		
		t_4	1.3524	1.3689	0.1132		
	QTL effect		0.3333	0.3385	0.0712		
8	Threshold	t_1	−1.7340	−1.7606	0.1593	91	24.71 (10.84)
		t_2	−1.0944	−1.1151	0.1176		
		t_3	−0.5542	−0.5701	0.1055		
		t_4	0.0000	0.0054	0.0935		
		t_5	0.5542	0.5690	0.0994		
		t_6	1.0944	1.0995	0.1075		
		t_7	1.7340	1.7533	0.1778		
	QTL effect		0.3333	0.3450	0.0734		

See the legends in Table 1.

Table 3 Mean and standard deviation (STD) of the estimated threshold values and QTL effect under various levels of QTL size

Heritability (h^2)	*Parameter*			*Estimates*			
	Name		*True value*	*Mean*	*STD*	*Power (%)*	*Position (cM)*
0.05	Threshold	t_1	−1.3152	−1.3155	0.1017	45	25.62 (15.10)
		t_2	−0.5383	−0.5393	0.0920		
		t_3	0.5383	0.5379	0.0873		
		t_4	1.3152	1.3349	0.1173		
	QTL effect		0.2294	0.2444	0.0776		
0.10	Threshold	t_1	−1.3524	−1.3730	0.1312	89	25.38 (7.80)
		t_2	−0.5542	−0.5668	0.1089		
		t_3	0.5542	0.5680	0.0963		
		t_4	1.3524	1.3700	0.1290		
	QTL effect		0.3333	0.3530	0.0769		
0.20	Threshold	t_1	−1.4394	−1.4512	0.1126	100	25.20 (3.67)
		t_2	−0.5932	−0.6007	0.0987		
		t_3	0.5932	0.6029	0.1068		
		t_4	1.4394	1.4712	0.1348		
	QTL effect		0.5000	0.5220	0.0744		
0.40	Threshold	t_1	−1.6807	−1.6719	0.1471	100	24.66 (2.17)
		t_2	−0.7212	−0.7268	0.1133		
		t_3	0.7212	0.7357	0.1144		
		t_4	1.6807	1.7190	0.1466		
	QTL effect		0.8165	0.8235	0.0933		

See legends in Table 1.

are summarized in Table 5. We did observe the expected trend of the power increase as the sample size was increased. The accuracy and precision were also increased as the sample size was increased. Note that the statistical power was 90% when the sample size was 200. This situation has been simulated several times in the previous subsections (Tables 2–4). The empirical statistical powers ranged from 86 to 90%, which reflects the stochastical error due to limited number of replicates. The main purpose of the paper was to develop a new method rather than to conduct exhaustive simulations for comparison of statistical power in an exact manner. Therefore, 100 replicates appear to suffice for demonstrating the efficiency of the new method of QTL mapping.

Discussion

We introduced the multicycle ECM algorithm for mapping ordinal traits using a four-way cross model, not because the four-way cross model is more common than the simple line cross model (BC and F_2) but because the former is a general model which covers the simple line crosses as special cases. Note that when we extend the four-way crosses model to BC and F_2 families, the estimated genetic effects need to be rescaled in order to

Table 4 Mean and standard deviation (STD) of the estimated threshold values and QTL effect for various shapes of phenotypic distribution

Phenotypic distribution	*Parameter*			*Estimates*			
	Notation		*True value*	*Mean*	*STD*	*Power (%)*	*Position (cM)*
Uniform (1:1:1:1:1)	Threshold	t_1	−0.8890	−0.8831	0.1080	83	26.45 (8.67)
		t_2	−0.2678	−0.2579	0.0888		
		t_3	0.2678	0.2799	0.0870		
		t_4	0.8890	0.9086	0.1093		
	QTL effect		0.3333	0.3303	0.0791		
Symmetrical distribution (1:2:4:2:1)	Threshold	t_1	−1.3524	−1.3706	0.1266	90	26.26 (7.92)
		t_2	−0.5542	−0.5744	0..0906		
		t_3	0.5542	0.5535	0.0923		
		t_4	1.3524	1.3740	0.1310		
	QTL effect		0.3333	0.3470	0.0813		
Skewed distribution (6:4:3:1:1)	Threshold	t_1	−0.2678	−0.2766	0.0965	80	25.91 (8.54)
		t_2	0.4552	0.4593	0.0927		
		t_3	1.1727	1.2000	0.1283		
		t_4	1.5831	1.6222	0.1500		
	QTL effect		0.3333	0.3386	0.0918		

See legends of Table 1.

Table 5 Mean and standard deviation (STD) of the estimated threshold values and QTL effect for various sample sizes (*n*)

Sample size	*Parameter*			*Estimates*			
	Name		*True value*	*Mean*	*STD*	*Power (%)*	*Position (cM)*
100	Threshold	t_1	−1.3524	−1.4029	0.1972	50	29.38 (20.48)
		t_2	−0.5542	−0.5846	0.1433		
		t_3	0.5542	0.5564	0.1325		
		t_4	1.3524	1.4033	0.1764		
	QTL effect		0.3333	0.3630	0.1139		
200	Threshold	t_1	−1.3524	−1.3732	0.1376	90	24.31 (7.32)
		t_2	−0.5542	−0.5636	0.0953		
		t_3	0.5542	0.5654	0.0923		
		t_4	1.3524	1.3794	0.1311		
	QTL effect		0.3333	0.3524	0.0819		
300	Threshold	t_1	−1.3524	−1.3533	0.0941	98	25.67 (5.71)
		t_2	−0.5542	−0.5439	0.0784		
		t_3	0.5542	0.5661	0.0844		
		t_4	1.3524	1.3721	0.1097		
	QTL effect		0.3333	0.3443	0.0586		
500	Threshold	t_1	−1.3524	−1.3712	0.0730	100	25.51 (3.65)
		t_2	−0.5542	−0.5525	0.0605		
		t_3	0.5542	0.5622	0.0667		
		t_4	1.3524	1.3695	0.0831		
	QTL effect		0.3333	0.3221	0.0480		

See legends of Table 1.

be comparable with the results using the traditional BC and F_2 models. Recall that the design matrix for the linear model in the four-way cross is denoted by $\mathbf{Z}_j = [Z_{1j}\ Z_{2j}\ Z_{3j}]$ for the *j*th individual. The coefficient of each genetic effect takes one of two possible values, 1 and −1, with an equal probability. Therefore, they all have a zero expectation and a unity variance, and are orthogonal to each other. When extended to the BC family, Z_{2j} and Z_{3j} have vanished from the model. The only coefficient left in the model is Z_{1j}, which takes value 1 for a heterozygote and −1 for a homozygote. In the traditional BC model, however, the coefficient is defined as 1 for a heterozygote and 0 for a homozygote, which leads to an expectation of 1/2 and a variance of 1/4. Therefore, when the traditional BC model is compared with our extended BC model, we should take into consideration the scale difference. The estimated effect of the extended BC model would be half the effect of the traditional BC model. When extended to the F_2 family, Z_{1j} and Z_{2j} have been combined because $a^m = a^f = a$. Therefore, the coefficient of the additive effect is $Z_{1j} + Z_{2j}$, with a zero expectation and a variance of 2. This means that the coefficient of the additive effect is defined as −2 for one homozygote, 0 for the heterozygote and 2 for the other homozygote. In the traditional F_2 model, however, the coefficient of the additive effect is defined as 0 for one homozygote, 1 for the heterozygote and 2 for the other homozygote. In such a scale, the expectation of the additive coefficent is 1 and the variance is 1/2. Therefore, when the traditional F_2 model is compared with our extended F_2 model, we should take into consideration the scale difference. The estimated additive effect of the traditional F_2 model would be twice the effect of the extended F_2 model. The coefficient of the dominance effect in the extended F_2 model is defined as 1 for the homozygote and -1 for the heterozygote, whereas, in the traditional F_2 model, this coefficient is defined as 1 for the heterozygote and 0 for the homozygote. Therefore, the estimated dominance effect in the extended F_2 model should be half the effect of the traditional F_2 model with an opposite sign.

The ECM algorithm developed in this study depends on the probit model rather than the logistic model (Hackett and Weller, 1995; Rao and Xu, 1998). The probit model uses a normal link function, which is more natural than the logistic link function because the residual error is assumed to be normally distributed in the probit model. With the normal link function, QTL effect is estimated in the original scale rather than in a logit scale and then converted into the probit scale using a constant $\sqrt{3}/\pi$, which is an approximate factor. The probit model serves as an alternative but slightly better model than the logistic analysis because of the normal distribution of the residual error. The two models take different approaches, in that the probit model simply tries to take advantage of existing QTL mapping theory for regular quantitative traits whereas the logistic regression tries to take advantage of the simple form of the link function. The logistic link function can be easily calculated without using numerical integration, whereas the probit link function may require numerical integration because there is no closed form of the normal distribution function. This disadvantage is less relevant as most modern computer programs, such as SAS (SAS Institute, 1999), can make use of a function to call the normal distribution function and its reverse function. Binary data QTL mapping is a special case of the ordinal data QTL mapping where there are only two categories in the phenotype. In binary trait QTL mapping, Rebai (1997) and Kadarnideen *et al* (2000) compared the threshold model with a simple regression analysis where the binary phenotype, coded as 0 or 1, was simply analyzed as if it were continuous. They showed that the power loss in the simple regression analysis was almost negligible compared with the threshold model. We present the threshold model to give the users an alternative but statistically more rigorous method for ordinal data analysis. Users may choose either method for their data analysis. If users prefer rapid results, then simple regression is the choice; otherwise, the threshold model implemented via the EM algorithm should be choice, because the EM method is at least as efficient as the regression method.

We have written a computer program implementing the above data analyses. The program is written in SAS 8.2, called QTL-By-SAS, which runs on both the Windows and Unix platform. The program codes and a user manual can be downloaded from our website at www.statgen.ucr.edu.

Acknowledgements

We are grateful to two anonymous reviewers for their constructive suggestions on revising the manuscript. This research was supported by the National Institutes of Health Grants R01-GM55321 and the USDA National Research Initiative Competitive Grants Program 00-35300-9245 to SX.

References

Caranta C, Plieger S, Lefebvre V, Daubeze AM, Thabuis A, Palloix A (2002). QTLs involved in the restriction of cucumber mosaic virus (CMV) long-distance movement in pepper. *Theor Appl Genet* **104**: 586–591.

Galecki A, Ten Have AT, Molenberghs G (2001). A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Comp Statist Data Analysis* **35**: 265–281.

Hackett CA, Weller JI (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**: 1252–1263.

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

Jansen RC (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.

Jiang C, Zeng ZB (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.

Kadarnideen HN, Janss LLG, Dekkers JCM (2000). Power of quantitative trait locus mapping for polygenic binary traits using generalized and regression interval mapping in multi-family half-sib designs. *Genet Res* **76**: 305–317.

Kao CH, Zeng ZB, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.

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

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Genetics*. Sinauer Associates: MA.

Manzanares-Dauleux MJ, Delourme R, Baron F, Thomas G (2000). Mapping of one major gene and of QTLs involved in resistance to clubroot in *Brassica napus*. *Theor Appl Genet* **101**: 885–891.

Meng XL, Rubin DB (1993). Maximum likelihood estimation via the ECM algorithm. *Biometrika* **80**: 267–278.

Nelder JA, Mead A (1965). A simplex method for function minimization. *Comput J* **7**: 308–313.

Piepho H-P (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**: 425–432.

Rao S, Xu S (1998). Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* **81**: 214–224.

Rao S, Li X (2001). Strategies for genetic mapping of categorical traits. *Genetica* **109**: 183–197.

Rebai A (1997). Comparision of methods for regression interval mapping in QTL analysis with non-normal traits. *Genet Res* **69**: 69–74.

Rush MC, Hoff BJ, Mcllrath WO (1976). A uniform disease rating system for rice disease in the United States. Proceedings of the 16th Rice Tech Working Group, Lake Charles, LA, USA, p 64.

SAS Institute Inc (1999). *SAS Language User's Guide*, Version 8. Cary, NC.

Visscher PM, Haley CS, Knott SA (1996). Mapping QTLs for binary traits in backcross and F_2 populations. *Genet Res* **68**: 55–63.

Xu S (1996). Mapping quantitative trait loci using four-way crosses. *Genet Res* **68**: 175–181.

Xu S (1998). Iteratively reweighted least squares mapping of quantitative trait loci. Behavior. *Genetics* **28**: 341–355.

Xu S, Atchley WR (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* **143**: 1417–1424.

Xu S, Yi N, Burke D, Galecki A, Miller RA (2003). An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. *Genet Res* **82**: 127–138.

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

Zou J, Pan XB, Chen ZX, Xu JY, Lu JF, Zhai WX *et al* (2000). Mapping quantitative trait loci controlling sheath blight resistance in two rice cultivars. *Theor Appl Genet* **101**: 569–573.