Mapping Quantitative Trait Loci for Complex Binary Diseases Using Line Crosses

Shizhong Xu* and William R. Atchley†

**Department of Botany and Plant Sciences, University of California, Riverside, California 92521-0124 and †Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695-7614*

Manuscript received November 1, 1995
Accepted for publication April 3, 1996

ABSTRACT

A composite interval gene mapping procedure for complex binary disease traits is proposed in this paper. The binary trait of interest is assumed to be controlled by an underlying liability that is normally distributed. The liability is treated as a typical quantitative character and thus described by the usual quantitative genetics model. Translation from the liability into a binary (disease) phenotype is through the physiological threshold model. Logistic regression analysis is employed to estimate the effects and locations of putative quantitative trait loci (our terminology for a single quantitative trait locus is QTL while multiple loci are referred to as QTLs). Simulation studies show that properties of this mapping procedure mimic those of the composite interval mapping for normally distributed data. Potential utilization of the QTL mapping procedure for resolving alternative genetic models (*e.g.*, single- or two-trait-locus model) is discussed.

COMPLEX disease refers to any disease with unknown mode of inheritance, especially polygenic models. The genetic mechanisms underlying such complex diseases are usually analyzed using quantitative genetics techniques whose classical model partitions a complex trait into genetic and environmental components. The genetic component is thought to be controlled by a number of loci each with a small effect (BULMER 1971; FALCONER 1981).

Many disease-resistant traits in plants and animals are described as quantitative characters. For instance, resistance to *Gibberella zeae* infection in maize was measured as the ratio of the infected area to the total area in the inoculated internode (PE *et al.* 1993), while resistance to blast fungus infection in rice was measured by lesion number and size (WANG *et al.* 1994). Mapping genes for such quantitative disease traits can be accomplished by traditional interval mapping procedures (LANDER and BOTSTEIN 1989; HALEY and KNOTT 1992) and methods of composite interval mapping (JANSEN 1993, 1994; ZENG 1993, 1994).

Some disease-susceptible traits, however, are not quantitative characters, but rather are qualitative traits and usually binary response variables. The vast majority of qualitative disease traits have a polygenic basis, such as the fusiform rust disease resistance in loblolly pine where the trait is described as presence or absence of the formation of galls (WILCOX 1995). The genetic mechanism underlying rust-disease resistance in loblolly pine is still unknown. It is heritable but not inherited in a simple Mendelian fashion. It is not a single gene trait and environment also plays a role. Binary disease traits with a polygenic basis are also categorized as complex diseases.

WRIGHT (1934) proposed a "physiological threshold" theory to explain the link between a continuous latent variable and an observable binary phenotype. The threshold theory states that underlying the dichotomy (phenotype), there is a "scale of factor combinations" to which each factor (locus) makes a fairly constant contribution. More recently, this scale of factor combination (plus a random environmental deviation) was referred to as "liability" (*e.g.*, FALCONER 1981). When liability is below the threshold an individual has the "normal" phenotypic expression, when it is above the threshold the individual has the "affected" phenotypic expression. Therefore, quantitative genetic analysis of a complex disease refers to the genetic study of the liability and the threshold.

Mapping genes for such binary traits is more complicated than that for continuous traits. Current methods are limited to analyses of the association between a marker and a quantitative trait loci (QTL) using a simple 2×2 chi-square test (*e.g.*, WILCOX 1995). The chi-square test uses one marker at a time that does not provide estimates of the effect and position of the QTL. In addition, if multiple QTLs occur in the same linkage group, the chi-square test tends to be biased. HALEY and KNOTT (1992) suggest using generalized linear model approach to analyze such threshold traits. JANSEN (1992) and JANSEN and STAM (1994) investigated a general mixture model and claimed that the general mixture model for mapping QTLs can be used for non-normally distributed data, such as counts or percent-

Corresponding author: Shizhong Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521-0124.
E-mail: xu@genetics.ucr.edu

ages. However, systematic investigation of gene mapping for binary traits under the physiological threshold model has been lacking. Herein, we modify the (composite) interval mapping procedures applied to continuous traits (LANDER and BOTSTEIN 1989; JANSEN 1993, 1994; ZENG 1994) to interval mapping for binary data.

MODEL OF LIABILITY

A complex disease trait is assumed to be controlled by a latent variable, referred to as liability, which is considered to be continuous and normally distributed. It can be described by the usual linear model

$$z_i = b_0 + \sum_{j=1}^{m} b_j x_{ij} + e_i, \qquad (1)$$

where z_i is the liability for the ith individual, b_0 is the grand mean (intercept), x_{ij} is the jth explanatory variable, b_j is the regression coefficient and e_i is the residual with a distribution of $N(0, \sigma_e^2)$. Since the liability is unobservable, the mean and residual variance can be set at any arbitrary values. For simplicity, we choose $b_0 = 0$ and $\sigma_e^2 = 1$ throughout the presentation.

Our purpose is to map QTLs controlling disease trait using molecular markers, thus, the explanatory variables are now defined as indicator variables of marker genotypes. In fact, other fixed effects, such as sex, age and location of field, can be incorporated into the model to control the residual variance, but they are ignored here for convenience. Let us consider, for simplicity, a backcross population derived from two inbred parental lines, P_1 and P_2, fixed for alternative alleles at several QTLs and m markers. Let us assume that the backcross population is derived from $F_1 \times P_1$ so that a backcross individual can be either homozygous with P_1 allelic type or heterozygous with one P_1 allele and one P_2 allele at a particular locus. If the ith individual is homozygous at the jth marker, $x_{ij} = 1$, otherwise, $x_{ij} = 0$. The expected values of b_j's are given by ZENG (1993).

A disease susceptible trait (y_i), determined by the underlying liability, is a realized binary variable, defined as

$$y_i = \begin{cases} 1 & \text{if affected} \\ 0 & \text{otherwise.} \end{cases}$$

The device that translates liability into disease phenotype is the physiological threshold model (WRIGHT 1934). It assumes that there is a threshold (θ) in the scale of liability, below which the individual has the normal phenotype, and above which it is affected. The translation can be summarized by

$$y_i = \begin{cases} 1 & \text{if } z_i \geq \theta \\ 0 & \text{if } z_i < \theta \end{cases}$$

Estimation of regression coefficients requires the conditional probability of $y_i = 1$ given x_{ij}'s, the marker genotypes. This conditional probability can be obtained by integrating out the random noise. Let us define $z_i \mid X$ as a conditional variable of z_i given the QTL and marker genotypes. From model 1 we know that the conditional mean and variance are $E(z_i \mid X) = \sum_{j=1}^{m} b_j x_{ij}$ and Var $(z_i \mid X) = 1$, respectively. We also know that the conditional variable is normally distributed because the residual term is assumed to be normal. Therefore, the density function of the conditional variable is

$$f(z_i \mid X) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(z_i \mid X - \sum_{j=1}^{m} b_j x_{ij} \right)^2 \right\}.$$

The conditional probability of $y_i = 1$ given X is obtained by

$$\begin{aligned} \Pr(y_i = 1 \mid X) &= \int_{\theta}^{\infty} f(z_i \mid X)\, d(z_i \mid X) \\ &= 1 - \int_{-\infty}^{\theta} f(z_i \mid X)\, d(z_i \mid X) = 1 - \Phi\left(\theta - \sum_{j=1}^{m} b_j x_{ij} \right) \\ &= \Phi\left(\sum_{j=1}^{m} b_j x_{ij} - \theta \right), \qquad (2) \end{aligned}$$

where $\Phi(\xi)$ stands for the standardized cumulative normal distribution function and ξ is the argument. Analysis involving $\Phi(\xi)$ is referred to as probit analysis. We chose the probit model because the parameters are easy to interpret. However, the probit model is difficult to manipulate because numerical integration is required. So, a logistic model is employed to approximate $\Phi(\xi)$ for estimation purpose. Logistic regressions have been used by human geneticists in segregation analysis (*e.g.*, BONNEY 1986). The logistic model is expressed by

$$\psi(\xi) = \frac{\exp\{\xi\}}{1 + \exp\{\xi\}}.$$

The approximate relationship between a probit model and a logistic model is $\Phi(\xi) \approx \psi(c\xi)$, where $c = \pi/\sqrt{3}$. Therefore,

$$\Pr(y_i = 1 \mid X) \approx \frac{\exp\left\{ c\left[\sum_{j=1}^{m} b_j x_{ij} - \theta \right] \right\}}{1 + \exp\left\{ c\left[\sum_{j=1}^{m} b_j x_{ij} - \theta \right] \right\}}. \qquad (3)$$

This approximation is remarkably close when $0.1 < \Phi(\xi) < 0.9$ (LIAO 1994). Hereafter, the probit model is replaced by its logistic approximation.

METHODS OF ESTIMATION

Marker-QTL association: The method of maximum likelihood is used to estimate the regression coefficients. A significant value of a regression coefficient

indicates linkage of the marker with a QTL. Let p_i denote $\Pr(y_i = 1 | X)$, then the log likelihood is

$$L = \sum_{i=1}^{n} y_i \log(p_i) + \sum_{i=1}^{n} (1 - y_i) \log(1 - p_i). \quad (4)$$

The unknown parameters are θ and b_j's, but θ is a nuisance parameter in linkage analysis and only b_j's are of interest. The maximum likelihood estimators are found by setting partial derivatives of L with respect to the parameters equal to zero. The first and second partial derivatives can be found in COX (1970). A statistical test for $H_0: b_j = 0$ is carried out by the likelihood ratio (LR) approximation. The likelihood ratio test involves calculation of L under the full model, denoted by L_1, and under the restricted model ($b_j = 0$), denoted by L_0. The likelihood ratio is $-2(L_0 - L_1)$, which asymptotically follows a chi-square distribution with one degree of freedom under the null hypothesis.

Standard computer programs for logistic regression analysis can be found in some commercial statistical packages such as PROC LOGISTIC in SAS (SAS Institute 1988).

Interval mapping: The multiple logistic regression analysis provides a test for marker QTL association, but it does not give estimates of the size and location of a tested QTL. Assuming one QTL on a chromosome, LANDER and BOTSTEIN (1989) developed the interval mapping procedure, which can separate the QTL effect and the linkage parameter. Let us set the effect of the heterozygous genotype to zero and that of the homozygous genotype to a for a particular QTL. Note that arbitrarily setting the effect of the heterozygous genotype to zero does not change the estimation and test of the QTL effect, but it will affect the estimation of the threshold (θ). We describe the liability using the model of ZENG (1994) where an indicator variable representing the QTL genotype is included in the model. His model is different from that of LANDER and BOTSTEIN (1989) in that other important markers are incorporated as covariates to control the genetic background of other chromosomal regions. ZENG's model is described by

$$z_i = b^* x_i + \sum_{j \in \Omega} b_j x_{ij} + e_i, \quad (5)$$

where b^* is the effect of a putative QTL ($b^* = a$) in the tested interval, x_i^* is an indicator variable representing the genotype of the putative QTL, Ω indexes the markers excluding the two flanking ones. Note that x_i^* is no longer known for sure and it takes a value of 1 or 0 with a probability depending on the genotypes of the two flanking markers and the QTL position. The conditional probability of $x_i^* = 1$ is expressed by $f_i = \Pr(x_i^* = 1 | x_{i1}\, x_{i2})$, where x_{i1} and x_{i2} are the indicators of the two flanking marker genotypes. Let r_1 and r_2 be the recombination fractions of the QTL with the left and the right markers, respectively, and denote the recombination fraction between the two markers by r. Without interference, the conditional probability of $x_i^* = 1$ is (DOERGE *et al.* 1994)

$$f_i = \begin{cases} (1 - r_1)(1 - r_2)/(1 - r) & \text{if } x_{i1} = x_{i2} = 1 \\ (1 - r_1) r_2 / r & \text{if } x_{i1} = 1 \text{ and } x_{i2} = 0 \\ r_1 (1 - r_2)/r & \text{if } x_{i1} = 0 \text{ and } x_{i2} = 1 \\ r_1 r_2/(1 - r) & \text{if } x_{i1} = x_{i2} = 0. \end{cases}$$

Since the putative QTL is assumed to be inside the interval, r_2 is a function of r_1 and r, as shown by

$$r_2 = \frac{r - r_1}{1 - 2r_1}.$$

It is taken that r is known, so that there is only one independent unknown recombination fraction.

The conditional probability of $y_i = 1$ given the marker genotypes is partitioned into two parts, p_{i1} and p_{i0}. If $x_i^* = 1$, this probability is p_{i1}, where

$$\text{Logit}(p_{i1}) = \text{Log}[p_{i1}/(1 - p_{i1})] = c\left[b^* + \sum_{j \in \Omega} b_j x_{ij} - \theta\right]. \quad (6)$$

If $x_i^* = 0$, the conditional probability becomes p_{i0}, where

$$\text{Logit}(p_{i0}) = \text{Log}[p_{i0}/(1 - p_{i0})] = c\left[\sum_{j \in \Omega} b_j x_{ij} - \theta\right]. \quad (7)$$

The likelihood function becomes

$$L = \prod_{i=1}^{n} [f_i p_{i1}^{y_i} (1 - p_{i1})^{1 - y_i} + (1 - f_i) p_{i0}^{y_i} \times (1 - p_{i0})^{1 - y_i}], \quad (8)$$

whose solutions can be solved via the expectation-maximization (EM) algorithm. In this particular case, the EM algorithm requires the first and second partial derivatives of L with respect to the unknown parameters. These partial derivatives and the EM steps are given in the APPENDIX.

The unknown parameters are b^*, b_j's and θ, but only $b^* = 0$ is tested. The test can be accomplished by the likelihood ratio approximation. The permutation test of CHURCHILL and DOERGE (1994) offers a very robust alternative to the likelihood ratio test. The recombination fraction between the QTL and a flanking marker is also an unknown parameter, but it is usually treated as a known constant and then the whole interval is searched from one end to another with an increment of 1 or 2 cM.

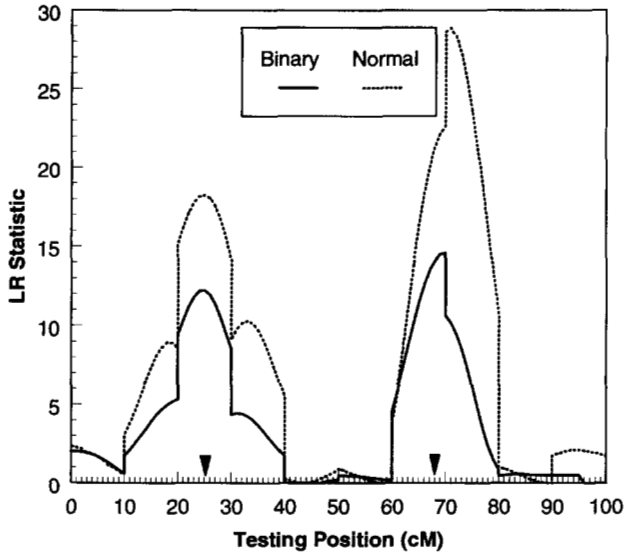

FIGURE 1.—Likelihood ratio profiles of interval mapping from one replicate of simulation in a backcross population of size 500. There are two QTLs located at 25 and 68 cM positions of the chromosomal segment. The solid curve represents QTL mapping for binary data using the logistic regression presented in this paper. The dotted curve stands for QTL mapping for normally distributed data of the liability (as if it were the observed phenotype) using ZENG's composite interval mapping.

SIMULATION

An example: To illustrate the properties of the method, a simulation study was performed. One chromosome with 11 markers separated in 10 10-cM intervals was simulated for a backcross population. The underlying liability is affected by two QTLs located in the positions of 25 and 68 cM (depicted in Figure 1) with gene effects of $a_1 = 0.931$ and $a_2 = -0.931$ units, respectively. Dominance and epistasis were assumed to be absent. Using HALDANE's mapping function, the recombination frequency between the two QTLs is $r = 0.2884$. The additive genetic variance is $\sigma_A^2 = [a_1^2 + a_2^2 + 2(1 - 2r)\, a_1 a_2]/4 = 0.25$. The liability of each individual was generated by adding a random normal deviate, $e \sim N(0, 1)$, to the additive genetic value. The mean and variance of the liability are $\bar{z} = (a_1 + a_2)/2 = 0.0$ and $\sigma_z^2 = \sigma_A^2 + \sigma_e^2 = 0.25 + 1.0 = 1.25$. Each QTL alone accounts for 17.34% of the total variation and the two QTLs jointly account for 20% of the total variation (due to a negative covariance between the two loci). To convert the continuous liability into a binary responsible variable, we set the threshold at $\theta = 0.0$, which leads to 50% of the individuals being affected. Sample size of this simulation was 500, which is sufficiently large to demonstrate the general properties of the method.

The data set generated under the genetic model described above was used for the composite interval mapping analysis. For comparison, the liability was treated as if it were the observed phenotypic variable and the continuous data of the liability were analyzed using ZENG's (1994) composite interval mapping procedure. Figure 1 shows the likelihood profiles from one replicate of the simulation. Both methods had successfully detected the two QTLs in the right locations. However, analysis of the binary data shows a lower profile than that of the normal data because some information has been lost when converting normal into binary data. A lower profile implies a lower statistical power. Similar results have been observed from analyses of more replicates (not shown).

Power studies: To compare the statistical powers and estimation errors of interval mapping for binary data with those for normal data, more simulations were conducted. One QTL located in the middle of a single interval of 20 cM was simulated. We considered the following factors that may have great influence on the performance of the mapping procedures: (1) sample size, (2) size of the QTL, and (3) threshold. Two levels were investigated for each factor and simulation was repeated 100 times in each parameter combination. The effect of the QTL was set at $a = 0.459$ and 0.667, corresponding to $h^2 = 0.05$ and 0.10, respectively. The threshold determines the proportion of disease infection (disease incidence) and it was set at values such that the disease incidences were at 25 and 50%. A critical value of 3.84 (the critical value at $\alpha = 0.05$ of the χ^2 distribution with one degree of freedom) in the test statistic was chosen to determine the statistical powers. The actual critical value may be slightly higher than 3.84 (HALEY and KNOTT 1992). Results are given in Table 1 and Table 2 for situations where disease incidences are 50 and 25%, respectively. In general, the gene mapping procedure for binary data performs very well. Compared with the analyses when the liability is treated as if it were observed normal phenotype, the binary method has lower power and larger estimation error, especially when the disease incidence deviated from 50%. Note that the logistic regression approximation requires that the disease cannot be too rare, else the approximation does not hold.

DISCUSSION

A QTL mapping procedure for binary disease traits is proposed in this paper. A binary trait is assumed to be controlled by an underlying liability with normal distribution. The liability is treated as a usual quantitative character and thus described by the usual linear model (ZENG 1994). Translation from the liability into the binary phenotype is through the physiological threshold model. The conditional probability of disease infection given QTL and marker genotypes is described by the probit model but approximated by a logistic regression for convenience. The genotype of a putative QTL is uncertain, but it is inferred from the genotypes of two flanking markers. Therefore, a mixture of likeli-

TABLE 1

Statistical powers and estimation errors when the disease incidence (proportion of affected individuals) is about 50%

Heritability[a]	Sample size	Data type	cM_A[b]	$\hat{a}$[c]	Test statistic	Power (%)[d]
0.05	200	Binary	9.79 (7.616)	0.42 (0.19)	7.33 (5.74)	64
		Normal	9.34 (7.13)	0.46 (0.16)	10.36 (6.59)	87
	500	Binary	10.25 (6.29)	0.39 (0.11)	13.69 (7.05)	98
		Normal	9.95 (4.97)	0.46 (0.09)	22.02 (8.42)	100
0.10	200	Binary	9.86 (5.52)	0.63 (0.17)	13.11 (6.16)	96
		Normal	10.00 (4.64)	0.71 (0.15)	20.94 (8.27)	100
	500	Binary	9.92 (4.37)	0.60 (0.11)	28.17 (10.08)	100
		Normal	10.20 (3.53)	0.67 (0.10)	44.47 (12.29)	100

The table shows the average estimates and standard deviations (in parentheses) from 100 replicates of simulation.

[a] Proportion of the total variation in liability explained by the QTL.

[b] Estimated position of the QTL. The parametric value is 10 cM.

[c] Estimated effect of the QTL.

[d] Statistical power at an error rate of 0.05.

hood is used for the logistic regression analysis. Properties of this mapping technique resembles those of the composite interval mapping of ZENG (1994) for normal data.

The interval mapping procedure for binary traits is usually less powerful than the well developed mapping methods for continuous traits. This is because some information will be lost during the translation from the underlying liability into the observed binary phenotype. The threshold (θ) is the leading parameter that determines the amount of information loss. The threshold determines the disease incidence (proportion of infected individuals) in the population of interest. The efficiency of the interval mapping largely depends on the disease incidence. The maximum efficiency occurs when the disease incidence is 50%. Consider that statistical power is a monotonically increasing function of the heritability of the putative. This heritability has a maximum value when the disease incidence is 50% (EDWARDS 1969). As the disease incidence deviates from 50%, more information will be lost. As a consequence, the method presented here is not applicable to gene mapping for rare diseases. Although θ cannot be controlled in natural populations, it can be manipulated in designed experiments. In QTL mapping experiments, plants are usually artificially inoculated by spraying a certain amount of pathogen spore suspension. Under some circumstances, one may adjust the amount of spore suspension to make the disease incidence close to 50% as much as possible.

Usually, the mechanisms underlying complex binary disease traits are not known. The physiological threshold model is only a hypothesis that is hardly tested. However, a plausible biological interpretation of the threshold model can be drawn from the threshold characteristic of enzyme activity. MOZHAEV *et al.* (1989) found that the dependence of the catalytic activities of α-chymotrypsin and lactase on the concentration of organic cosolvents in mixed aqueous media has a pronounced threshold character: the activity does not change up to a critical concentration of the nonaqueous cosolvents added, yet further increase of the latter

TABLE 2

Statistical powers and estimation errors when the disease incidence (proportion of affected) is 25%

Heritability	Sample size	Data type	cM_A	$\hat{a}$	Test statistic	Power (%)
0.05	200	Binary	10.93 (7.97)	0.46 (0.21)	5.17 (3.67)	57
		Normal	11.24 (7.14)	0.44 (0.14)	9.43 (5.81)	83
	500	Binary	9.85 (6.83)	0.44 (0.15)	11.28 (6.62)	86
		Normal	9.88 (5.09)	0.46 (0.10)	22.90 (9.99)	100
0.10	200	Binary	9.76 (7.31)	0.65 (0.29)	8.42 (5.33)	78
		Normal	10.02 (5.58)	0.65 (0.16)	18.05 (8.41)	99
	500	Binary	9.24 (5.37)	0.72 (0.17)	22.05 (8.67)	99
		Normal	9.32 (3.68)	0.68 (0.11)	46.32 (13.43)	100

The table shows the average estimates and standard deviations (in parentheses) from 100 replicates of simulation. See Table 1 for the legends.

(by only a small amount) leads to an abrupt decrease in enzyme activity. Consider that disease resistance is determined by the activity of a particular enzyme. The locus coding for this enzyme may be called the major gene. In a population where the major gene has been fixed, the disease phenotype may still show polymorphism. This may occur when the enzyme activity is determined by the level of gene products of several QTL. The collective effect of the gene products of the QTLs is analogous to the liability. When the level of the gene products reaches a certain threshold, the enzyme becomes inactive, leading to disease infection.

A well-known alternative to the threshold model is that the expression of a disease trait is determined by the expression of one or two loci (SCHORK 1993). If the disease in question is controlled by one locus and environment does not play a role, the disease expression in a backcross population will show simple Mendelian segregation. In this case, there is no need to invoke the threshold model of gene mapping. However, the threshold model is appropriate in dealing with situations where the disease is controlled by a single locus but environmental effect also plays a role in the expression of the disease trait. In genetic mapping of diseases under one- or two-locus model, the conditional probability of trait expression of a given genotype is usually referred to as penetrance (SCHORK 1993). For example, the penetrances of homozygotes ($x^* = 1$) and heterozygotes ($x^* = 0$) in a backcross population are denoted by

$$p_1 = \Pr(y = 1 \mid x^* = 1) \quad \text{and} \quad p_0 = \Pr(y = 1 \mid x^* = 0),$$

respectively. It is then natural to use the difference between p_1 and p_0 to detect the effect of the putative locus on the disease. Recall that x^* is unobservable but its conditional distribution given marker genotypes is known (denoted by f_i). Therefore, the log likelihood function can be constructed as

$$L_1 = \sum_{i=1}^{n} \log[f_i p_1^{y_i}(1 - p_1)^{1-y_i} + (1 - f_i) p_0^{y_i}(1 - p_0)^{1-y_i}].$$

The maximum likelihood extimates of p_1 and p_0 are solved iteratively by

$$\hat{p}_1 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad \text{and} \quad \hat{p}_0 = \frac{\sum_{i=1}^{n} (1 - w_i) y_i}{\sum_{i=1}^{n} (1 - w_i)},$$

where

$$w_i = \frac{f_i \hat{p}_1^{y_i}(1 - \hat{p}_1)^{(1-y_i)}}{f_i \hat{p}_1^{y_i}(1 - \hat{p}_1)^{(1-y_i)} + (1 - f_i)\hat{p}_0^{y_i}(1 - \hat{p}_0)^{(1-y_i)}}.$$

Under the null hypothesis H_0: $p_1 = p_0 = p$, the log likelihood function becomes

$$L_0 = \sum_{i=1}^{n} [y_i \log(p) + (1 - y_i) \log(1 - p)],$$

whose maximum likelihood solution is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

As usual, the likelihood ratio statistic is used to test the null hypothesis.

We now show that the above test is equivalent to that under the threshold model. When the expression of a disease trait is determined by a single locus, it is not necessary to include nonflanking markers in the model of liability. Thus, the penetrance can be expressed as

$$p_1 = \frac{\exp\{c(b^* - \theta)\}}{1 + \exp\{c(b^* - \theta)\}},$$

when $x^* = 1$, and

$$p_0 = \frac{\exp\{-c\theta\}}{1 + \exp\{-c\theta\}},$$

when $x^* = 0$. Instead of maximizing the log likelihood function with respect to p_1 and p_0, we now maximize the likelihood with respect to b^* and θ (c is a constant). According to the invariance property of the maximum likelihood method (DEGROOT 1986), the maximum likelihood estimates of θ and b^* are

$$\hat{\theta} = -\frac{1}{c} \log[\hat{p}_0/(1 - \hat{p}_0)]$$

and

$$\hat{b}^* = \frac{1}{c} \log\left[\frac{\hat{p}_1(1 - \hat{p}_0)}{\hat{p}_0(1 - \hat{p}_1)}\right],$$

respectively. Clearly, the null hypothesis that $b^* = 0$ is equivalent to that $p_1 = p_0$.

When the disease is controlled by two loci and environment plays a role in the expression of disease trait, the threshold model proposed in this paper is still valid as long as there is no epistatic interaction between the two loci. In the presence of epistatic effect, the current threshold model must be modified so that two QTLs are mapped simultaneously using a two-dimensional search strategy (*e.g.*, HALEY and KNOTT 1992), an area that deserves further investigation.

The interval mapping procedure presented in this paper is for binary disease data only. Many disease traits are measured as ordinary data, *i.e.*, among the individuals classified as affected, the degree of infection may be different. Individuals may be classified into several groups depending on the degree of disease infection. QTL mapping for ordinary data can be accomplished by ordinal logistic regression analysis (LIAO 1994). The underlying liability for an ordinary disease trait can be described by the same linear model given in (1), but

there are several thresholds to translate the liability into observed ordinary disease phenotypes (LANGE *et al.* 1976). The binary data analysis will be a special case of the general procedure for ordinary data analysis for which further investigation is required.

This work was supported by National Research Institute Competitive Grants Programs / U.S. Department of Agriculture 95-37205-2313 to S. X. and National Institutes of Health grant GM-45344 and National Science Foundation BSR-910718 to W. R. A.

LITERATURE CITED

BONNEY, G. E., 1986 Regressive logistic models for familial disease and other binary traits. Biometrics **42:** 611–625.

BULMER, M. G., 1971 The effect of selection on genetic variability. Am. Nat. **104:** 201–211.

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

COX, D. R., 1970 *The Analysis of Binary Data.* Methuen & Co. Ltd., London.

DEGROOT, M. H., 1986 *Probability and Statistics,* Ed. 2. Addison-Wesley Publishing Co., Reading, MA.

DOERGE, R. W., Z.-B. ZENG and B. S. WEIR, 1994 Statistical issues in the search for genes affecting quantitative traits in populations, pp. 15–26 in *Analysis of Molecular Marker Data, Proceedings of Joint Plant Breeding Symposia Series,* Corvallis, OR.

EDWARDS, J. H., 1969 Familial predisposition in man. Brit. Med. Bull. **25:** 58–63.

FALCONER, D. S., 1981 *Introduction to Quantitative Genetics,* Ed. 2, Longman, New York.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

JANSEN, R. C., 1992 A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. **85:** 252–260.

JANSEN, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

JANSEN, R. C., 1994 Controlling the Type I and Type II errors in mapping quantitative trait loci. Genetics **138:** 871–881.

JANSEN, R. C., and P. STAM, 1995 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447–1455.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

LANGE, K., J. WESTLAKE and M. A. SPENCE, 1976 Extensions to pedigree analysis: II. Recurrent risk calculation under the polygenic threshold model. Hum. Hered. **26:** 337–348.

LIAO, T. F., 1994 *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-101. Thousand Oaks, CA.

MOZHAEV, V. V., Y. L. KHMELNITSKY, M. V. SERGEEVA, A. B. BELOVA, N. L. KLYACHKO *et al.,* 1989 Catalytic activity and denaturation of enzymes in water / organic cosolvent mixtures. Eur. J. Biochem. **184:** 597–602.

PE, M. E., L. GIANFRANCESCHI, G. TARAMINO, R. TARCHINI, P. ANGELINI *et al.,* 1993 Mapping quantitative trait loci (QTLs) for resistance to *Gibberella zeae* infection in maize. Mol. Gen. Genet. **241:** 11–16.

SAS INSTITUTE INC., 1988 SAS / STAT User's Guide, Release 6.03 Edition. SAS Institute Inc., Cary, NC.

SCHORK, N. J., M. BOEHNKE, J. D. TERWILLIGER, and J. OTT, 1993 Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits. Am. J. Hum. Genet. **53:** 1127–1136.

WANG, G.-L., D. J. MACKILL, M. BONMAN, S. R. MCCOUCH, M. C. CHAMPOUX *et al.,* 1994 RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. Genetics **136:** 1421–1434.

WILCOX, P. L., 1995 *Genetic Dissection of Fusiform Rust Resistance in Loblolly Pine.* Ph.D. Dissertation, Department of Forestry, North Carolina State University, Raleigh.

WRIGHT, S., 1934 The results of crosses between inbred strains of guinea pigs differing in number of digits. Genetics **19:** 537–551.

ZENG, Z.-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

ZENG, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: M. LYNCH

APPENDIX

This appendix gives the partial derivatives of the log likelihood function with respect to the unknown parameters and describes the EM steps.

Let $x_{i0} = -1$ for $i = 1, \ldots, n$ and j indexes the threshold (when $j = 0$, $b_j = \theta$) and all other markers excluding the two flanking ones, then

$$p_{i1} = \frac{\exp\left\{c\left(b^* + \sum_{j\in\Omega} b_j x_{ij}\right)\right\}}{1 + \exp\left\{c\left(b^* + \sum_{j\in\Omega} b_j x_{ij}\right)\right\}},$$

and

$$p_{i0} = \frac{\exp\left\{c\left(\sum_{j\in\Omega} b_j x_{ij}\right)\right\}}{1 + \exp\left\{c\left(\sum_{j\in\Omega} b_j x_{ij}\right)\right\}}.$$

The log likelihood function is

$$L = \sum_{i=1}^{n} \log[f_i p_{i1}^{y_i}(1 - p_{i1})^{1-y_i} + (1 - f_i) p_{i0}^{y_i}(1 - p_{i0})^{1-y_i}].$$

The first partial derivatives are

$$\frac{\partial L}{\partial b^*} = c \sum_{i=1}^{n} w_i (y_i - p_{i1})$$

and

$$\frac{\partial L}{\partial b_j} = c \sum_{i=1}^{n} [w_i (y_i - p_{i1}) + (1 - w_i)(y_i - p_{i0})] x_{ij},$$

for $j = 0, \ldots, \Omega$, where

$$w_i = \frac{f_i p_{i1}^{y_i}(1 - p_{i1})^{(1-y_i)}}{f_i p_{i1}^{y_i}(1 - p_{i1})^{(1-y_i)} + (1 - f_i) p_{i0}^{y_i}(1 - p_{i0})^{(1-y_i)}}$$

is the posterior probability of $x^* = 1$.

The second partial derivatives are messy because w_i is a function of the unknown parameters. However, if these w_i's are treated as constants, then the second partial derivatives have simple forms as shown:

$$\frac{\partial^2 L}{\partial b^{*2}} = -c^2 \sum_{i=1}^{n} w_i p_{i1}(1 - p_{i1}),$$

$$\frac{\partial^2 L}{\partial b^* \partial b_j} = -c^2 \sum_{i=1}^{n} w_i p_{i1}(1 - p_{i1}) x_{ij},$$

and

$$\frac{\partial^2 L}{\partial b_j \partial b_k} = -c^2 \sum_{i=1}^{n} [\, w_i p_{i1}(1 - p_{i1}) + (1 - w_i) p_{i0}(1 - p_{i0}) \,] x_{ij} x_{ik}$$

for $j, k = 0, \ldots, \Omega$

The EM steps are as follows

1. Set up initial values of b^* and b_j for $j = 0, \ldots, \Omega$;
2. Calculate w_i (E-step);
3. Given w_i, solve for b^* and b_j using the NEWTON-RAPHSON iteration (M-step);
4. Update the initial values and go to step 2);
5. Repeat steps 2–4 until convergence.

The maximization step is accomplished via the NEWTON-RAPHSON iteration, which is described as follows. Let $\mathbf{d}$ be a vector of the first partial derivatives and $\mathbf{J}$ be a matrix of the second partial derivatives. If $\beta(t)$ is a vector of solutions at the tth step, the solutions at the $t + 1$ step is $\beta(t + 1) = \beta(t) - \mathbf{J}^{-1}\mathbf{d}$, where $\mathbf{J}$ and $\mathbf{d}$ are evaluated at $\beta(t)$.

With the EM algorithm, convergence is guaranteed because the matrix of the second partial derivatives is always negative definite.