An improved procedure of mapping a quantitative trait locus via the EM algorithm using posterior probabilities

SAURABH GHOSH and PARTHA P. MAJUMDER*

Anthropology and Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India

Abstract

Mapping a locus controlling a quantitative genetic trait (e.g. blood pressure) to a specific genomic region is of considerable contemporary interest. Data on the quantitative trait under consideration and several codominant genetic markers with known genomic locations are collected from members of families and statistically analysed to estimate the recombination fraction, θ, between the putative quantitative trait locus and a genetic marker. One of the major complications in estimating θ for a quantitative trait in humans is the lack of haplotype information on members of families. We have devised a computationally simple two-stage method of estimation of θ in the absence of haplotypic information using the expectation-maximization (EM) algorithm. In the first stage, parameters of the quantitative trait locus (QTL) are estimated on the basis of data of a sample of unrelated individuals and a Bayes's rule is used to classify each parent into a QTL genotypic class. In the second stage, we have proposed an EM algorithm for obtaining the maximum-likelihood estimate of θ based on data of informative families (which are identified upon inferring parental QTL genotypes performed in the first stage). The purpose of this paper is to investigate whether, instead of using genotypically 'classified' data of parents, the use of posterior probabilities of QT genotypes of parents at the second stage yields better estimators. We show, using simulated data, that the proposed procedure using posterior probabilities is statistically more efficient than our earlier classification procedure, although it is computationally heavier.

[Ghosh S. and Majumder P. P. 2000 An improved procedure of mapping a quantitative trait locus via the EM algorithm using posterior probabilities. *J. Genet.* **79**, 47–53]

Introduction

The recent identification of highly polymorphic DNA markers has resulted in a resurgence of interest in developing statistical techniques for quantitative trait locus (QTL) mapping (Haseman and Elston 1972; Amos and Elston 1989; Lander and Botstein 1989; Goldgar 1990; Haley and Knott 1992; Kruglyak and Lander 1995; Olson 1995; Almasy and Blangero 1998). Many common human disorders (e.g. hypertension, diabetes) are inherently quantitative in nature. Therefore, QTL mapping is of considerable interest in human genetics. Many currently used QTL mapping methods, especially those that have been developed in the context of plant genetics or genetics of inbred animals, assume knowledge of linkage phase in individuals, which imposes a severe restriction on the applicability of these methods in human genetics. One of the major problems in QTL mapping is to accurately infer the genotype of an individual at the major locus controlling variation of the quantitative trait. Ghosh and Majumder (2000) have proposed a method to estimate, via the expectation-maximization (EM) algorithm, the recombination fraction between a marker locus and an autosomal major locus controlling a quantitative trait from data on nuclear families without any assumptions on linkage phase and haplotypes. The proposed method is a two-stage strategy. In the first stage, individuals are probabilistically classified into the major locus genotypes, and in the second stage, the recombination fraction is estimated using the inferences made in the first stage. Monte-Carlo simulation studies showed that this method works well only when the percentage of correct trait locus classification is high and that the performance of the method is quite poor in presence of high degree of dominance in the QT. In this paper, we modify the estimation procedure proposed by Ghosh and Majumder

*For correspondence. E-mail: ppm@isical.ac.in.

Keywords. recombination fraction; linkage; maximum likelihood estimation.

(2000). Instead of classifying each parent into a specific trait locus genotype, we use the posterior probabilities corresponding to each parental genotype in the second stage of our algorithm. We show, using simulated data, that this procedure performs better than the classification procedure.

Model

Consider an autosomal biallelic locus with alleles (A_1, a_1) determining a quantitative trait Y. Suppose the distribution of Y conditioned on the genotype is

$$Y|A_1A_1 \sim N(\alpha, \sigma^2)$$
$$Y|A_1a_1 \sim N(\beta, \sigma^2)$$
$$Y|a_1a_1 \sim N(-\alpha, \sigma^2),$$

where $\beta \leq \alpha$ and σ^2 includes the environmental variance.

Suppose the allele frequency of A_1 is p. Then, assuming Hardy–Weinberg equilibrium proportions at the QTL, Y has a mixture distribution given by

$$p^2N(\alpha, \sigma^2) + 2p(1-p)N(\beta, \sigma^2) + (1-p)^2N(-\alpha, \sigma^2).$$

Consider an autosomal biallelic codominant marker locus with alleles (M_1, m_1) possibly linked to the QTL. The aim is to estimate the recombination fraction, θ, between the two loci, which are assumed to be in linkage equilibrium.

Data description

We consider data on nuclear families. Suppose $\{(y_{i1}, y_{i2}) : i = 1, 2, \ldots, K\}$ are the observed values of the quantitative trait of K pairs of parents such that, in each pair, either one parent is M_1M_1 and the other M_1m_1 or both parents are M_1m_1. (Obviously, if neither parent is heterozygous at the marker locus, the family is not informative for linkage.) For the ith pair of parents with n_i offspring, the known trait values will be denoted as $(y_{i3}, y_{i4}, \ldots, y_{in_i+2})$, $i = 1, 2, \ldots, K$. We further assume that the marker genotype (M_1M_1, M_1m_1, or m_1m_1) of each offspring is known. Thus, the data comprise trait values and marker genotypes of parents and offspring in nuclear families.

An outline of the classification procedure

Estimation algorithm

Although the primary aim is to estimate θ, since the trait parameters α, β, σ^2 and p are unknown, one can estimate these also to facilitate estimation of θ. Knowledge of α, β, σ^2 and p facilitates estimation of θ because using the estimated values of α, β, σ^2 and p, and the observed values of the quantitative trait, one can classify each parent, albeit probabilistically, to a specific trait locus genotype. When trait locus genotypes are known for the parents in a nuclear family, then obtaining an estimate of θ from the remaining data (marker genotypes of parents and offspring, and values of the quantitative trait of the offspring) becomes much simpler. The estimation procedure is based on this two-stage strategy.

Let $f_1(x)$, probability density function (pdf) of $N(\alpha, \sigma^2), = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right)$,

π_1, prior probability of $f_1, = p^2$,

$f_2(x)$, pdf of $N(\beta, \sigma^2), = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\beta)^2}{2\sigma^2}\right)$,

π_2, prior probability of f_2, $= 2p(1-p)$,

$f_3(x)$, pdf of $N(-\alpha, \sigma^2)$, $= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x+\alpha)^2}{2\sigma^2}\right)$, and

π_3, prior probability of f_3, $= (1-p)^2$.

Thus the pdf of y_{ij} $(i = 1, 2, \ldots, K; j = 1, 2)$ is given by $f(y_{ij}) = \sum_{n=1}^{3} \pi_n f_n(y_{ij})$.

The parameters to be estimated in this mixture model are α, σ^2 and p. One can estimate these parameters by the maximum-likelihood method.

The likelihood of the parental data is $L(\alpha, \beta, \sigma^2, p|y_{ij}) = \prod_{i=1}^{K}\prod_{j=1}^{2}\sum_{n=1}^{3}\pi_n f_n(y_{ij})$.

A computationally simple and elegant procedure of estimating the parameters is based on the EM algorithm (Dempster *et al.* 1977) corresponding to a mixture of normal populations (see McLachlan and Krishnan 1997). A sketch of the algorithm is presented below.

The mixture distribution can be viewed as an 'incomplete' setup in the sense that we have no a priori knowledge of which of the three component distributions any particular observation belongs to. The first step (E-step) in this algorithm is therefore to estimate the probabilities with which an observation may belong to any of the three component distributions. The second step (M-step) uses these estimates to build up the 'complete' likelihood function, which is easily maximized to yield relevant parameter estimates.

Define:

$$z_{ijn} = 1, \quad \text{if } y_{ij} \text{ is an observation from pdf } f_n,$$
$$= 0, \text{otherwise},$$

where $i = 1, 2, \ldots, K; j = 1, 2; n = 1, 2, 3$.

The E-step of the EM algorithm is

$$\widehat{z}_{ijn} = E(z_{ijn}|y_{ij})$$
$$= \frac{\pi_n f_n(y_{ij})}{\sum_{n=1}^{3}\pi_n f_n(y_{ij})},$$

where $i = 1, 2, \ldots, K;\ j = 1, 2; n = 1, 2, 3$. We note that these estimators are Bayes's.

Having obtained the $\widehat{z}_{ijn}$s, we can easily obtain the closed form expressions for the maximum likelihood estimate (mle) of p, α and σ^2 in the M-step of the algorithm:

$$L(p, \alpha, \beta, \sigma^2|y_{ij}, \widehat{z}_{ijn}) = \prod_{i=1}^{K}\prod_{j=1}^{2}\prod_{n=1}^{3}\{\pi_n f_n(y_{ij})\}^{\widehat{z}_{ijn}}.$$

The mle's of the parameters are given by

$$\widehat{p} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}(\widehat{z}_{ij1} + \frac{1}{2}\widehat{z}_{ij2})}{2K},$$

$$\widehat{\alpha} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}(\widehat{z}_{ij1} - \widehat{z}_{ij3})y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{2}(\widehat{z}_{ij1} + \widehat{z}_{ij3})},$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{2}\widehat{z}_{ij2}y_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{2}\widehat{z}_{ij2}},$$

$$\widehat{\sigma}^2 = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}\{\widehat{z}_{ij1}(y_{ij} - \widehat{\alpha})^2 + \widehat{z}_{ij2}(y_{ij} - \widehat{\beta})^2 + \widehat{z}_{ij3}(y_{ij} + \widehat{\alpha})^2\}.$$

As this algorithm is an iterative procedure, one requires initial estimates of p, α, β and σ^2 $(\widehat{p}^{(0)}, \widehat{\alpha}^{(0)}, \widehat{\beta}^{(0)}, \widehat{\sigma^2}^{(0)})$ to implement this iterative algorithm. As an initial approximation of β, one can assume that there is no dominance effect, i.e. $\widehat{\beta}^{(0)} = 0$. As $0 \le p \le 1$, one can fix $\widehat{p}^{(0)} = p_0$ within this interval. One can obtain the initial estimates of α and σ^2 using the method of moments.

In the next stage the parents are classified (i.e. $\{(y_{i1}, y_{i2}) : i = 1, 2, \ldots, K\}$) into one of the three component distributions. One uses the usual classification rule given by:

Classify y_{ij} into f_n if and only if

$$\widehat{z}_{ijn} = \max_{t=1,2,3}\widehat{z}_{ijt}, \qquad (1)$$

where $i = 1, 2, \ldots, K$; $j = 1, 2$; $n = 1, 2, 3$, the $\widehat{z}_{ijn}$s being the final (converged) values in the above EM algorithm. This is, in fact, the Bayes's classification rule corresponding to the 0–1 loss function and thus minimizes the error in classification under such loss functions (Fergusson 1967).

Having estimated α, β, σ^2, p and having classified the parents into the trait genotypes, one is now in a position to implement another maximum-likelihood procedure to estimate θ. One uses the conditional trait distribution of the offspring given the trait genotypes of the parents and the marker genotypes of both parents and the offspring in order to estimate θ. We provide these distributions in tables 1 and 2.

Let

M_{ij} = marker genotype of jth individual in ith family, $i = 1, 2, \ldots, K$; $j = 1, 2, \ldots, n_i + 2$

G_{i1}, G_{i2} = classified trait genotypes of the parents in ith family, $i = 1, 2, \ldots, K$; $j = 1, 2$

H_{ij} = trait genotype of jth individual [i.e. $(j-2)$th offspring] in ith family, $i = 1, 2, \ldots, K$; $j = 3, 4, \ldots, n_i + 2$

$P_{ijn} = P\{H_{ij} = \gamma_n | G_{i1}, G_{i2}, M_{i1}, M_{i2}, M_{ij}\}$, where $\gamma_1 = A_1A_1, \gamma_2 = A_1A_2, \gamma_3 = A_2A_2$, $i = 1, 2, \ldots, K$; $j = 3, 4, \ldots, n_i + 2$; $n = 1, 2, 3$.

The P_{ijn}s are obviously functions of θ. However, for the same genotype, P_{ijn} may be different for different haplotypes. Thus, in estimating θ, one has to consider the different possible haplotypes separately for given trait and marker loci genotypes of each parent. Next, one classifies the offspring into their trait genotypes.

Define:

$$Q_{ijn} = P(H_{ij} = \gamma_n | G_{i1}, G_{i2}, M_{i1}, M_{i2}, M_{ij}, y_{ij}) = \frac{P_{ijn} f_n(y_{ij})}{\sum_{n=1}^{3} P_{ijn} f_n(y_{ij})},$$

$i = 1, 2, \ldots, K$; $j = 3, 4, \ldots, n_i + 2$; $n = 1, 2, 3$.

Table 1. Trait locus mating types among $M_1M_1 \times M_1m_1$ parents, mating probabilities, and probabilities of trait locus genotypes among offspring with marker genotype M_1M_1*.

			π_g		
g	Mating type	Probability	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	$\frac{1}{2}$	0	0
2	$A_1A_1 \times A_1a_1$	${p_1}^3p_2$	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$	0
3	$A_1A_1 \times a_1A_1$	${p_1}^3p_2$	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$	0
4	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2{p_1}^2{p_2}^2$	0	$\frac{1}{2}$	0
5	$A_1a_1 \times A_1A_1$ $a_1A_1 \times A_1A_1$	$2{p_1}^3p_2$	$\frac{1}{4}$	$\frac{1}{4}$	0
6	$A_1a_1 \times A_1a_1$ $a_1A_1 \times A_1a_1$	$2{p_1}^2{p_2}^2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}$	$\frac{1}{4}\theta$
7	$A_1a_1 \times a_1A_1$ $a_1A_1 \times a_1A_1$	$2{p_1}^2{p_2}^2$	$\frac{1}{4}\theta$	$\frac{1}{4}$	$\frac{1}{4}(1-\theta)$
8	$A_1a_1 \times a_1a_1$ $a_1A_1 \times a_1a_1$	$2p_1{p_2}^3$	0	$\frac{1}{4}$	$\frac{1}{4}$
9	$a_1a_1 \times A_1a_1$	$p_1{p_2}^3$	0	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$
10	$a_1a_1 \times a_1A_1$	$p_1{p_2}^3$	0	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$
11	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	$\frac{1}{2}$

*Probabilites of trait locus genotypes among offspring with marker genotype M_1m_1 can be obtained by replacing θ by $(1-\theta)$ in this table.

Table 2. Trait locus mating types among $M_1m_1 \times M_1m_1$ parents, mating probabilities, and probabilities of trait locus genotypes among offspring with marker genotype M_1M_1 and M_1m_1*.

			$\pi_g(M_1M_1)$			$\pi_g(M_1m_1)$		
g	Mating type	Probability	A_1A_1	A_1a_1	a_1a_1	A_1A_1	A_1a_1	a_1a_1
1	$A_1A_1 \times A_1A_1$	${p_1}^4$	$\frac{1}{4}$	0	0	$\frac{1}{2}$	0	0
2	$A_1A_1 \times A_1a_1$ $A_1a_1 \times A_1A_1$	$2{p_1}^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
3	$A_1A_1 \times a_1A_1$ $a_1A_1 \times A_1A_1$	$2{p_1}^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
4	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2{p_1}^2{p_2}^2$	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$	0
5	$A_1a_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[1-2\theta(1-\theta)]$	$\frac{1}{2}\theta(1-\theta)$
6	$A_1a_1 \times a_1A_1$ $a_1A_1 \times A_1a_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$
7	$a_1a_1 \times A_1a_1$ $A_1a_1 \times a_1a_1$	$2p_1{p_2}^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}$	$\frac{1}{4}$
8	$a_1A_1 \times a_1A_1$	${p_1}^2{p_2}^2$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[1-2\theta(1-\theta)]$	$\frac{1}{2}\theta(1-\theta)$
9	$a_1a_1 \times a_1A_1$ $a_1A_1 \times a_1a_1$	$2p_1{p_2}^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}$	$\frac{1}{4}$
10	$a_1a_1 \times a_1a_1$	${p_2}^4$	0	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$

*Probabilites of trait locus genotypes among offspring with marker genotype m_1m_1 can be obtained by replacing θ by $(1-\theta)$ in the block corresponding to the genotype M_1M_1 in this table.

In the computation of Q_{ijn}, one uses $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\sigma}^2$ obtained using the EM algorithm described previously.

The usual classification rule is given by:

Classify y_{ij} into f_n if and only if

$$Q_{ijn} = \max_{t=1,2,3} Q_{ijt},$$

$i = 1, 2, \ldots, K;\ j = 3, 4, \ldots, n_i + 2;\ n = 1, 2, 3.$

The likelihood of θ is given by

$$L(\theta) = \prod_{i=1}^{K} L_i(\theta), \tag{2}$$

where $L_i(\theta)$ is the likelihood of the ith family based on the classified genotypes of the n_i offspring of that family. Note that as haplotypic information is usually unavailable from nuclear family data, $L_i(\theta)$ would be a mixture of the different conditional trait distributions of the offspring corresponding to the different possible haplotypes. In fact $L_i(\theta)$ is a mixture with components of the form $c_{i0}\theta^{i1}(1-\theta)^{i2}$ or $c_{i0}\theta^{i1}(1-\theta)^{i2}\{\theta^2 + 1 - \theta)^2\}^{i3}$, where c_{i0} is some constant. Since a direct analytical maximization procedure is complicated, one implements an EM procedure. $L_i^*(\theta)$ would be of the form $c_i\theta^{u_i}(1-\theta)^{v_i}$, where c_i is some constant while u_i and v_i are functions of θ. Thus,

$$L^*(\theta) = \left\{\prod_{i=1}^{K} c_i\right\}\theta^{\sum_{i=1}^{K} u_i}(1-\theta)^{\sum_{i=1}^{K} v_i},$$

which is easy to maximize, giving

$$\widehat{\theta} = \frac{\sum_{i=1}^{K} u_i}{\sum_{i=1}^{K}(u_i + v_i)}.$$

Since u_is and v_is depend on θ, one needs an initial approximation for implementing the EM algorithm. As $0 \le \theta \le 0.5$, $\theta = 0.25$ may be used as an initial approximation. If the final (converged) value of $\widehat{\theta}$ exceeds 0.5, one takes $\widehat{\theta} = 0.5$.

We finally note that families in which neither parent is classified as a heterozygote at the major QTL can be discarded even before marker-typing because these families will not provide any information for estimating θ. This strategy will be cost-effective.

Efficiency of the estimation procedure

Assessment of the efficiency of the estimation procedure is of obvious importance. Before providing the results, we describe the simulation procedure for fixed values of p, α, β, σ^2 and θ. In the first step, we randomly generated the trait values of a fixed number (*NOBS*) of pairs of unrelated parents from appropriate (selected randomly using a trinomial random number generator with cell probabilities p^2, $2pq$ and q^2) normal distributions (see Model section above). In the second step, using the data so generated, the trait parameters $(\alpha, \beta, \sigma^2, p)$ were estimated using the EM algorithm. (We emphasize that, for the purpose of estimating the trait parameters, it is not essential to obtain data on pairs of parents; only data on randomly sampled unrelated individuals suffice.) In the third step, the QTL genotypes of the parents are inferred using the Bayes's rule. For further computations, only those pairs of parents with at least one inferred QTL heterozygote are retained. In the fourth step, for each parent in the retained pairs, marker genotype was determined using a trinomial random number generator. For subsequent computations, only those parental pairs with at least one double heterozygote were retained. In the fifth step, we randomly generated the marker genotype of an offspring by sampling either from a binomial distribution with success probability $1/2$ for a parental mating in which one parent is M_1M_1 or M_2M_2 and the other parent is M_1M_2 at the marker locus, or from a trinomial distribution with cell probabilities $(1/4, 1/2, 1/4)$ for a parental mating in which both parents are M_1M_2. In the sixth step, based on the conditional probabilities of offspring genotypes given parental mating type as provided in tables 1 and 2, we generated, using a trinomial random number generator, the genotype of the offspring with respect to the trait locus. These steps were repeated until the required number of informative families *(NFAM)* were obtained. Using the data so generated, we again used the EM algorithm to estimate θ. Replication of this procedure a large number of times *(NREP)* yielded the empirical frequency distribution. For every set of parameter values, we have evaluated the performance of the estimator with five offspring per family, $NFAM = 100$ and $NREP = 1000$.

Classification of parents with respect to QTL genotypes

As mentioned earlier, in the first stage of the procedure parents are classified into genotype classes on the basis of their observed trait values. Success of estimating the recombination fraction accurately by the present procedure depends critically on the classification performance at the first stage. We find that when there is no dominance (i.e. $\beta = 0$) more than 95% and 99.5% of the parents were correctly classified into their true genotypic classes. The percentage of correct classification increased as p deviated more from 0.5. The percentage of correct classification decreased as the extent of dominance (β) increased. The worst classification arose when the overlap between distributions of the A_1A_1 and A_1a_1 genotype classes was the largest.

Mean and variance of $\widehat{\theta}$ and confidence interval for θ

To examine the behaviour of the estimator in respect of variation in values of p and β, we have performed simulations for fixed parameter values $\alpha = 5$, $\sigma^2 = 1$, and for values of $p = 0.9, 0.7, 0.5$; $\beta = 0, 2, 4$; and $\theta = 0.1, 0.3, 0.5$. We have evaluated the means and variances of $\widehat{\theta}$ and have obtained 95% confidence intervals of θ. The results are given in table 3. These results indicate that the performance of the estimator is poor when p is close to 0.5 and the degree of dominance (β) is high. When p is close to 0.5, the mean of $\widehat{\theta}$ is more deviant from the true value of θ and the 95%

Table 3. Mean and variance of $\widehat{\theta}$ and 95% confidence interval of θ using classification procedure for $\alpha = 5$; $\sigma^2 = 1$; $p = 0.9, 0.7, 0.5$; $\beta = 0, 2, 4$; $\theta = 0, 0.1, 0.3, 0.5$.

p	True θ	β	Mean($\widehat{\theta}$)	Var($\widehat{\theta}$)	95% C.I. of θ
0.9	0	0	0.015	0.000174	(0.009, 0.026)
		2	0.044	0.000432	(0.017, 0.048)
		4	0.075	0.000695	(0.051, 0.097)
	0.1	0	0.103	0.000084	(0.099, 0.114)
		2	0.117	0.000277	(0.095, 0.126)
		4	0.172	0.001008	(0.131, 0.195)
	0.3	0	0.303	0.000452	(0.291, 0.311)
		2	0.313	0.000747	(0.286, 0.328)
		4	0.368	0.001739	(0.345, 0.401)
	0.5	0	0.478	0.000397	(0.438, 0.500)
		2	0.471	0.000902	(0.415, 0.500)
		4	0.409	0.001335	(0.395, 0.487)
0.7	0	0	0.021	0.000154	(0.019, 0.041)
		2	0.053	0.000312	(0.023, 0.057)
		4	0.081	0.000865	(0.063, 0.101)
	0.1	0	0.107	0.000087	(0.095, 0.122)
		2	0.122	0.000290	(0.097, 0.128)
		4	0.182	0.001064	(0.143, 0.204)
	0.3	0	0.308	0.000497	(0.293, 0.317)
		2	0.317	0.000683	(0.284, 0.321)
		4	0.373	0.001867	(0.357, 0.408)
	0.5	0	0.491	0.000083	(0.477, 0.500)
		2	0.487	0.000118	(0.472, 0.500)
		4	0.413	0.001146	(0.401, 0.494)
0.5	0	0	0.038	0.000186	(0.022, 0.058)
		2	0.067	0.000299	(0.035, 0.073)
		4	0.105	0.001018	(0.071, 0.112)
	0.1	0	0.113	0.000129	(0.097, 0.123)
		2	0.115	0.000283	(0.089, 0.124)
		4	0.196	0.001153	(0.162, 0.208)
	0.3	0	0.314	0.000512	(0.291, 0.325)
		2	0.321	0.000630	(0.287, 0.329)
		4	0.381	0.001794	(0.358, 0.416)
	0.5	0	0.497	0.000056	(0.486, 0.500)
		2	0.491	0.000068	(0.478, 0.500)
		4	0.421	0.001062	(0.411, 0.498)

confidence interval of θ is wider, particularly when θ is very close to 0.5. We also note that for fixed values of α, σ^2, p and θ the estimator is adversely affected in a nonlinear fashion by increase in β.

Effect of using posterior probabilities at the second stage

As described in the previous section, Ghosh and Majumder (2000) classified each parent into a most likely trait genotype using Bayes's 0–1 classification rule. As we note from our simulation results in the previous section, the performance of the estimator is strongly dependent on the percentage of correct genotypic classification of the parents. The estimator does not perform well for high degrees of dominance in the trait.

In this section, we investigate whether the performance of the estimator can be improved by using posterior probabilities of the three possible parental trait genotypes given the trait values of the parents in the second stage of the proposed procedure instead of classifying each parent into one specific trait genotype [which is equivalent to using one of posterior probability distributions (1,0,0), (0,1,0) or (0,0,1)].

As mentioned in the previous paragraph, we do not use the classification rule given by equation 1. We note that the posterior probability of the jth parent of the ith family belonging to the tth trait genotype is given by $\widehat{z_{ijt}}, i = 1, 2, \ldots, K$; $j = 1, 2$; $t = 1, 2, 3$, which will be used in the second stage of our estimation procedure.

In the present setup, we need to redefine G_{i1}, G_{i2} and P_{ijn} as:

G_{i1}, G_{i2} = trait genotypes of parents in the ith family.

$$P_{ijn}^{l,m} = P(H_{ij} = \gamma_n | G_{i1} = \gamma_l, G_{i2} = \gamma_m, M_{i1}, M_{i2}, M_{ij}),$$

where $\gamma_1 = A_1A_1, \gamma_2 = A_1a_1, \gamma_3 = a_1a_1$.

Similarly, Q_{ijn} has to be redefined as:

$$Q_{ijn}^{l,m} = P(H_{ij} = \gamma_n | G_{i1} = \gamma_l, G_{i2} = \gamma_m, M_{i1}, M_{i2}, M_{ij}, y_{ij}) = \frac{P_{ijn}{}^{l,m} f_n(y_{ij})}{\sum_{n=1}^{3} P_{ijn}{}^{l,m} f_n(y_{ij})}.$$

Thus, at the trait genotype classification stage of each offspring, we need to classify the offspring for every possible trait genotype combination of the parents (i.e. for each combination of (l, m), $l, m = 1, 2, 3$). The likelihood function $L(\theta)$ is identical to equation 2 except that each $L_i(\theta)$ comprises more complex mixture components than in the classification procedure, with the mixture proportions being functions of the product $(\widehat{z_{i1l}} \times \widehat{z_{i2m}})$, for each combination of (l, m), i.e. the posterior trait genotype probabilities of the parents in the ith family.

We use simulated data with the same sets of trait and linkage parameters as in the previous section to compare the performances of the estimators under the two strategies. The results based on the present strategy are given in table 4. Comparing this table with table 3, we find that means of the estimates of θ are, in general, more close to the true values of θ and have less variance compared to the earlier procedure based on parental classification. Moreover, the confidence intervals of θ are less wide under this strategy. The two procedures perform similarly when the proportion of homozygotes is high and dominance at the trait locus is low. However, as the proportion of heterozygotes or the degree of dominance at the trait locus increases, the performance of this procedure becomes increasingly better. This is due to the fact that unlike our proposed procedure, this procedure does not depend on the performance of parental trait genotype classification. Thus, the performance of this procedure is not affected by parameters that increase the misclassification probabilities like trait locus heterozygosity and dominance. The estimator under this strategy has more desirable statistical properties than the earlier estimator (Ghosh and Majumder 2000), though data analysis using this strategy is computationally more complex.

Table 4. Mean and variance of $\widehat{\theta}$ and 95% confidence interval of θ using posterior probabilities for $\alpha = 5; \sigma^2 = 1; p = 0.9, 0.7, 0.5; \beta = 0, 2, 4; \theta = 0, 0.1, 0.3, 0.5$.

p	True θ	β	Mean $(\widehat{\theta})$	Var$(\widehat{\theta})$	95% C.I. of θ
0.9	0	0	0.018	0.000182	(0.011, 0.028)
		2	0.040	0.000234	(0.021, 0.044)
		4	0.053	0.000316	(0.038, 0.067)
	0.1	0	0.104	0.000091	(0.098, 0.116
		2	0.116	0.000274	(0.096, 0.124)
		4	0.131	0.000457	(0.115, 0.143)
	0.3	0	0.305	0.000471	(0.294, 0.315)
		2	0.310	0.000619	(0.292, 0.323)
		4	0.331	0.000715	(0.316, 0.347)
	0.5	0	0.484	0.000384	(0.458, 0.500)
		2	0.477	0.000353	(0.443, 0.500)
		4	0.465	0.000505	(0.432, 0.500)
0.7	0	0	0.014	0.000106	(0.008, 0.024)
		2	0.025	0.000165	(0.017, 0.032)
		4	0.036	0.000255	(0.021, 0.052)
	0.1	0	0.102	0.000082	(0.098, 0.110)
		2	0.111	0.000227	(0.097, 0.120)
		4	0.119	0.000336	(0.109, 0.130)
	0.3	0	0.302	0.000404	(0.295, 0.311)
		2	0.311	0.000508	(0.294, 0.322)
		4	0.320	0.000609	(0.310, 0.335)
	0.5	0	0.495	0.000084	(0.485, 0.500)
		2	0.490	0.000281	(0.474, 0.500)
		4	0.479	0.000362	(0.455, 0.500)
0.5	0	0	0.010	0.000082	(0.005, 0.018)
		2	0.017	0.000112	(0.010, 0.025)
		4	0.023	0.000194	(0.014, 0.038)
	0.1	0	0.102	0.000075	(0.098, 0.107)
		2	0.105	0.000186	(0.098, 0.115)
		4	0.111	0.000265	(0.102, 0.120)
	0.3	0	0.300	0.000257	(0.297, 0.308)
		2	0.305	0.000338	(0.296, 0.315)
		4	0.313	0.000426	(0.301, 0.326)
	0.5	0	0.498	0.000064	(0.491, 0.500)
		2	0.494	0.000167	(0.485, 0.500)
		4	0.491	0.000245	(0.477, 0.500)

Discussion

The classification procedure for linkage detection proposed by Ghosh and Majumder (2000) exploits the fact that knowledge of parental genotypes at the QTL greatly eases statistical estimation of θ. Since for a quantitative character the QTL genotype of an individual cannot be inferred with certainty because of intrinsic variability within genotype classes, Ghosh and Majumder (2000) had used the EM algorithm coupled with a Bayes's classification procedure to classify parents into QTL genotype classes. Here we have modified this procedure by introducing posterior probabilities of each parental trait genotype in the second stage of our algorithm instead of classifying each parent into a specific trait locus genotype. In this procedure, estimates of trait parameters and recombination fraction are obtained. The estimates of trait parameters are used in obtaining the posterior probabilities of the parental QTL genotypes, which are then used in obtaining an estimate of the recombination fraction. The estimation of trait parameters, in the first stage of the proposed two-stage procedure, can be based either on data of a random sample of individuals or on data of parents (assumed to be unrelated) in families.

We have shown using simulations that our proposed method provides very good estimates of θ for a wide range of parameter values and reasonable sample sizes. Moreover, unlike the earlier procedure proposed by Ghosh and Majumder (2000), which is strongly dependent on the quality of classification of parental QT genotypes, the present procedure does not involve any parental trait locus classification and performs well even when heterozygosity is less and dominance is high in the QT. Compared to numerical maximization of the likelihood (Lincoln *et al.* 1993) of parental and offspring data, on all families jointly with respect to all parameters (recombination fraction, trait parameters and allele frequencies), the proposed stagewise procedure using the EM algorithm is computationally much more efficient and provides reduction of data collection costs.

References

Almasy L. and Blangero J. 1998 Multipoint quantitative trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**, 1198–1211.

Amos C. I. and Elston R. C. 1989 Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet. Epidemiol.* **6**, 349–360.

Dempster A. P., Laird N. M. and Rubin D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **B39**, 1–38.

Fergusson T. S. 1967 *Mathematical statistics: a decision-theoretic approach.* Academic Press, New York.

Ghosh S. and Majumder P. P. 2000 Mapping quantitative trait loci via the EM algorithm and Bayesian classification. *Genet. Epidemiol.* **19**, 97–126.

Goldgar D. E. 1990 Multipoint analysis of human quantitative genetic variation. *Am. J. Hum. Genet.* **47**, 957–967.

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

Haseman J. K. and Elston R. C. 1972 The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19.

Kruglyak L. and Lander E. S. 1995 A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

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

Lincoln S. E., Daly M. J. and Lander E. S. 1993 MAPMAKER/QTL version 1.1. http://www.genome.wi.mit.edu.

McLachlan G. J. and Krishnan T. 1997 *The EM algorithm and extensions.* Wiley, New York.

Olson J. M. 1995 Robust multipoint linkage analysis: an extension of the Haseman–Elston method. *Genet. Epidemiol.* **12**, 177–193.

Received 7 August 2000