Mapping a Quantitative Trait Locus Via the EM Algorithm and Bayesian Classification

Saurabh Ghosh and Partha P. Majumder*

Anthropology and Human Genetics Unit, Indian Statistical Institute, Calcutta, India

Mapping a locus controlling a quantitative genetic trait (e.g., blood pressure) to a specific genomic region is of considerable 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 analyzed to draw inferences on the genomic position of the trait locus. The vector of parameters of interest comprises the pairwise recombination fractions, θ, between the putative quantitative trait locus and the marker loci. One of the major complications in estimating θ for a quantitative trait in humans is the lack of haplotype information on members of families. The purpose of this study was to devise a computationally simple and efficient method of estimation of θ in the absence of haplotype information. We have proposed a two-stage estimation procedure using the expectation-maximization (EM) algorithm. In the first stage, parameters of the QTL are estimated based on data of a sample of unrelated individuals. From estimates thus obtained, we have used a Bayes' rule to infer QTL genotypes of parents in families. Finally, in the second stage of the procedure, 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). We have shown, using simulated data, that the proposed procedure is cost-effective, computationally simple, and statistically efficient. As expected, analysis of data on multiple markers jointly is more efficient than the analysis based on single markers. Genet. Epidemiol. 19:97–126, 2000. © 2000 Wiley-Liss, Inc.

Key words: recombination fraction; linkage; maximum likelihood estimation; two-point mapping; multipoint mapping

INTRODUCTION

Developing statistical techniques for the detection and estimation of linkage among marker loci and loci determining a quantitative trait is an active area of research [Jayakar, 1970; Haseman and Elston, 1972; Hill, 1975; Weller, 1986; Amos and Elston, 1989; Lander and Botstein, 1989; Goldgar, 1990; Haley and Knott, 1992; Zeng, 1994; Whittaker et al., 1995; Kruglyak and Lander, 1995; Schork et al., 1996]. Although the idea of mapping quantitative traits (QTL mapping) can be traced back to Sax [1923], the recent identification of highly polymorphic DNA markers in plants and animals and the development of dense maps of such markers have resulted in a resurgence of interest in developing simple and efficient statistical methods for QTL mapping. 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 that 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. The purpose of this paper is to propose a method to estimate, via the expectation maximization (EM) algorithm, the recombination fractions between marker loci 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. In the second stage, the recombination fractions are estimated using the inferences made in the first stage. The proposed procedure also provides estimates of parameters of the QTL. We have examined the efficiency of the estimation procedure using Monte-Carlo simulations and have shown that the proposed procedure works very well.

MODEL

Consider an autosomal biallelic locus, with alleles (A_1, A_2), determining a quantitative trait X. Suppose the distribution of X conditioned on the genotype is:

$$\begin{aligned} X|A_1A_1 &\sim N(\alpha,\sigma^2) \\ X|A_1A_2 &\sim N(\beta,\sigma^2) \\ X|A_2A_2 &\sim N(-\alpha,\sigma^2) \end{aligned} \qquad (1)$$

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, X has a mixture distribution given by:

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

Consider an autosomal biallelic codominant marker locus with alleles (M_1,M_2) possibly linked to the quantitative trait locus (QTL). [Extensions of the proposed method to multiple and multiallelic markers are discussed in later sections.] Our 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_2 or both parents are M_1M_2. (Obviously, if neither parent is heterozygous at the marker locus, the family is not informative for linkage.) For the i^{th} 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_2 or M_2M_2) of each offspring is known. Thus, the data comprise trait values and marker genotypes of parents and offspring in nuclear families.

ESTIMATION PROCEDURE

Although our primary aim is to estimate θ, since the trait parameters α, β, σ^2, and π are unknown, we shall 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, we 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. Our estimation procedure is based on this two-stage strategy.

Let, $f_1(x)$, probability density function (p.d.f.) of $N(\alpha,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$,
π_1, prior probability of $f_1 = p^2$,
$f_2(x)$, p.d.f. of $N(\beta,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\beta)^2}{2\sigma^2}}$,
π_2, prior probability of $f_2 = 2p(1-p)$,
$f_3(x)$, p.d.f. of $N(-\alpha,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x+\alpha)^2}{2\sigma^2}}$ and π_3, prior probability of $f_3 = (1-p)^2$
Thus the p.d.f. of y_{ij} ($i = 1,2,\ldots,K; j = 1,2$) is given by:

$$f(y_{ij}) = \sum_{n=1}^{3}\pi_n f_n(y_{ij}) \qquad (3)$$

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

The likelihood of the parental data is:

$$L(\alpha,\beta,\sigma^2,p|y_{ij}) = \prod_{i=1}^{K}\prod_{j=1}^{2}\sum_{n=1}^{3}\pi_n f_n(y_{ij}) \qquad (4)$$

However, a direct analytical maximization of the above function will not yield closed form estimators and iterative numerical maximization procedures, e.g., scoring method [Rao, 1973], will involve complicated expressions.

A computationally simpler and more elegant procedure is based on the EM algorithm corresponding to a mixture of normal populations [Dempster et al., 1977; 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 as to which of the three component distributions any particular observation belongs. 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:

$$\begin{aligned} z_{ijn} &= 1, \text{ if } y_{ij} \text{ is an observation from p.d.f. } f_n, \\ &= 0, \text{ otherwise;} \end{aligned}$$

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

The introduction of z_{ijn}s thus constitutes the "complete" setup. However, as z_{ijn}s are unknown, we have to estimate them conditioned on the observations y_{ij}. This is the E-step of the EM algorithm.

$$\begin{aligned} \hat{z}_{ijn} &= E(z_{ijn} | y_{ij}) \\ &= \frac{\pi_n f_n(y_{ij})}{\Sigma_{n=1}^{3} \pi_n f_n(y_{ij})}; \end{aligned} \tag{5}$$

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

Having obtained the $\hat{z}_{ijn}$s, we can easily obtain the closed form expressions for the m.l.e. of p, α, and σ^2 in the M-step of the algorithm.

$$L(p,\alpha,\beta,\sigma^2 | y_{ij}, \hat{z}_{ijn}) = \prod_{i=1}^{K} \prod_{j=1}^{2} \prod_{n=1}^{3} \{\pi_n f_n(y_{ij})\}^{\hat{z}_{ijn}}. \tag{6}$$

The m.l.e.s of the parameters are given by:

$$\hat{p} = \frac{\Sigma_{i=1}^{K} \Sigma_{j=1}^{2} (\hat{z}_{ij1} + \frac{1}{2} \hat{z}_{ij2})}{2K}, \tag{7}$$

$$\hat{\alpha} = \frac{\Sigma_{i=1}^{K} \Sigma_{j=1}^{2} (\hat{z}_{ij1} - \hat{z}_{ij3}) y_{ij}}{\Sigma_{i=1}^{K} \Sigma_{j=1}^{2} (\hat{z}_{ij1} + \hat{z}_{ij3})}, \tag{8}$$

$$\hat{\beta} = \frac{\Sigma_{i=1}^{K} \Sigma_{j=1}^{2} \hat{z}_{ij2} y_{ij}}{\Sigma_{i=1}^{K} \Sigma_{j=1}^{2} \hat{z}_{ij2}}, \tag{9}$$

$$\hat{\sigma}^2 = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}\{\hat{z}_{ij1}(y_{ij} - \hat{\alpha})^2 + \hat{z}_{ij2}(y_{ij} - \hat{\beta})^2 + \hat{z}_{ij3}(y_{ij} + \hat{\alpha})^2\}. \tag{10}$$

Thus, the l^{th} step of the EM algorithm is:
E-step:

$$\widehat{z_{ijn}}^{(l)} = \frac{\widehat{\pi_n}^{(l-1)}\,\widehat{f_n(y_{ij})}^{(l-1)}}{\Sigma_{n=1}^{3}\widehat{\pi_n}^{(l-1)}\,\widehat{f_n(y_{ij})}^{(l-1)}}; \tag{11}$$

$i = 1,2,\ldots, K; j = 1,2; n = 1,2,3.$
M-step:

$$\hat{p}^{(l)} = \frac{\Sigma_{i=1}^{K}\Sigma_{j=1}^{2}(\widehat{z_{ij1}}^{(l)} + \tfrac{1}{2}\widehat{z_{ij2}}^{(l)})}{2K}, \tag{12}$$

$$\hat{\alpha}^{(l)} = \frac{\Sigma_{i=1}^{K}\Sigma_{j=1}^{2}(\widehat{z_{ij1}}^{(l)} - \widehat{z_{ij3}}^{(l)})y_{ij}}{\Sigma_{i=1}^{K}\Sigma_{j=1}^{2}(\widehat{z_{ij1}}^{(l)} + \widehat{z_{ij3}}^{(l)})}, \tag{13}$$

$$\hat{\beta}^{(l)} = \frac{\Sigma_{i=1}^{K}\Sigma_{j=1}^{2}\widehat{z_{ij2}}^{(l)}\,y_{ij}}{\Sigma_{i=1}^{K}\Sigma_{j=1}^{2}\widehat{z_{ij2}}^{(l)}}, \tag{14}$$

$$\widehat{\sigma^2}^{(l)} = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}\{\widehat{z_{ij1}}^{(l)}(y_{ij} - \hat{\alpha}^{(l)})^2 + \widehat{z_{ij2}}^{(l)}(y_{ij} - \hat{\beta}^{(l)})^2 + \widehat{z_{ij3}}^{(l)}(y_{ij} + \hat{\alpha}^{(l)})^2\}. \tag{15}$$

We require initial estimates of p, α, β, and σ^2 ($\hat{p}^{(0)}$, $\hat{\alpha}^{(0)}$, $\hat{\beta}^{(0)}$, $\widehat{\sigma^2}^{(0)}$) to implement this iterative algorithm. The method of moments estimators serves as a simple initial choice [see Everitt and Hand, 1981].

As an initial approximation of β, we assume that there is no dominance effect, i.e., $\hat{\beta}^{(0)} = 0$.

Assuming $\beta = 0$, the method of moments yields the following equations:

$$\overline{Y} = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}y_{ij} = \alpha(2p-1), \tag{16}$$

$$s^2 = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{2}(y_{ij} - \overline{Y})^2 = \sigma^2 + 2p(1-p)\alpha^2. \tag{17}$$

As $0 \leq p \leq 1$, we can fix $\hat{p}^{(0)}$, $= p_0$ within this interval. Thus.

$$\hat{\alpha}^{(0)} = \overline{Y}/(2p_0 - 1); \tag{18}$$

$$\widehat{\sigma^2}^{(0)} = s^2 - \frac{2p_o(1-p_0)\bar{Y}^2}{(2p_0-1)^2}. \tag{19}$$

Clearly p_0 cannot be chosen to be 0.5.

Our next stage is to classify the parents (i.e. $\{(y_{i1},y_{i2})$: $i = 1,2, \ldots, K\}$) into one of the three component distributions. We shall use the usual classification rule given by:

Classify y_{ij} into f_n if and only if

$$\hat{z}_{ijn} = max_{t=1,2,3}\hat{z}_{ijt} \ ;$$

$i = 1,2, \ldots, K$; $j = 1,2$; $n = 1,2,3,$; the $\hat{z}_{ijn}$s being the final (converged) values in the above EM algorithm. This is, in fact, the Bayes' classification rule corresponding to the 0 - 1 loss function and thus minimises 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, we are now in a position to implement another maximum likelihood procedure to estimate θ. Before describing the actual procedure, let us note a few salient points. Information on θ can be obtained from only those offspring who have at least one of doubly heterozygous (i.e., $A_1A_2M_1M_2$) parent. We shall use 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 I and II.

TABLE I. Trait Locus Mating Types Among *MM* × *Mm* Parents, Mating Probabilities, and Probabilities of Trait Locus Genotypes Among Offspring With Marker Genotype *MM**

				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	$A_1A_1 \times A_1a_1$	$p_1^3p_2$	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$	
3	$A_1A_1 \times a_1A_1$	$p_1^3p_2$	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$	0
4	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2p_1^2p_2^2$	0	$\frac{1}{2}$	0
5	$A_1a_1 \times A_1A_1$ $a_1A_1 \times A_1A_1$	$2p_1^3p_2$	$\frac{1}{4}$	$\frac{1}{4}$	0
6	$A_1a_1 \times A_1a_1$ $a_1A_1 \times A_1a_1$	$2p_1^2p_2^2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}$	$\frac{1}{4}\theta$
7	$A_1a_1 \times a_1A_1$ $a_1A_1 \times a_1A_1$	$2p_1^2p_2^2$	$\frac{1}{4}\theta$	$\frac{1}{4}$	$\frac{1}{4}(1-\theta)$
8	$A_1a_1 \times a_1a_1$ $a_1A_1 \times a_1a_1$	$2p_1p_2^3$	0	$\frac{1}{4}$	$\frac{1}{4}$
9	$a_1a_1 \times A_1a_1$	$p_1p_2^3$	0	$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$
10	$a_1a_1 \times a_1A_1$	$p_1p_2^3$	0	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$
11	$a_1a_1 \times a_1a_1$	p_2^4	0	0	$\frac{1}{2}$

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

TABLE II. Trait Locus Mating Types Among $Mm \times Mm$ Parents, Mating Probabilities, and Probabilities of Trait Locus Genotypes Among Offspring With Marker Genotype MM and Mm*

g	Mating type	Probability	$\pi_g(MM)$			$\pi_g(Mm)$		
			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	$A_1A_1 \times A_1a_1$	$2p_1^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
	$A_1a_1 \times A_1A_1$							
3	$A_1A_1 \times a_1A_1$	$2p_1^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}\theta(1-\theta)$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
	$a_1A_1 \times A_1A_1$							
4	$A_1A_1 \times a_1a_1$	$2p_1^2p_2^2$	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$	0
	$a_1a_1 \times A_1A_1$							
5	$A_1a_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[1-2\theta(1-\theta)]$	$\frac{1}{2}\theta(1-\theta)$
6	$A_1a_1 \times a_1A_1$	$2p_1^2p_2^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$
	$a_1A_1 \times A_1a_1$							
7	$a_1a_1 \times A_1a_1$	$2p_1p_2^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}$	$\frac{1}{4}$
	$A_1a_1 \times a_1a_1$							
8	$a_1A_1 \times a_1A_1$	$p_1^2p_2^2$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{2}[1-2\theta(1-\theta)]$	$\frac{1}{2}\theta(1-\theta)$
9	$a_1a_1 \times a_1A_1$	$p_1p_2^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}$	$\frac{1}{4}$
	$a_1A_1 \times a_1a_1$							
10	$a_1a_1 \times a_1a_1$	p_2^4	0	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$

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

Let:

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

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

H_{ij} = trait genotype of j^{th} individual [i.e. $(j-2)^{th}$ offspring] in i^{th} 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 $\gamma_1 = A_1A_1, \gamma_2 = A_1A_2, \gamma_3 = A_2A_2$; $i = 1,2, \ldots ,K; j = 3,4, \ldots ,n_i + 2$;
$n = 1,2,3$.

P_{ijn}s are obviously functions of θ. However, for the same genotype, P_{ijn} may be different for different haplotypes. For example, if $G_{i1} = A_1A_1$, $G_{i2} = A_1A_2$, $M_{i1} = M_1M_1$, $M_{i2} = M_1M_2$, $M_{i3} = M_1M_1$, then $P_{i31} = \theta$ if the haplotype is A_1M_2/A_2M_1. Thus, in estimating θ, we have to consider the different possible haplotypes separately for given trait and marker loci genotypes of each parent. We next classify the offspring into their trait genotypes.

Define:

$$Q_{ijn} = P(H_{ij} = \gamma_n | G_{i1}, G_{i2}, M_{i1}, M_{i2}, M_{ij}, y_{ij}) = \frac{P_{ijn} f_n(y_{ij})}{\Sigma_{n=1}^{3} P_{ijn} f_n(y_{ij})}, \tag{20}$$

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

In the computation of Q_{ijn}, we use $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$ 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:

$$L(\theta) = \prod_{i=1}^{K} L_i(\theta) \tag{21}$$

where $L_i(\theta)$ is the likelihood of the i^{th} 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. For clarity of presentation, let us consider the following example of a nuclear family i with three ($n_i = 3$) offspring. Suppose the parental classified QTL genotypes are A_1A_1 (= G_{i1}) and A_1A_2 (= G_{i2}). Suppose the marker genotypes of these parents are, respectively, M_1M_1 (= M_{i1}) and M_1M_2 (= M_{i2}). Then, the possible haplotypes of the doubly heterozygous parent are: $A_1M_1|A_2M_2$ and $A_1M_2|A_2M_1$. The classification probabilities at the QTL for offspring depend on both marker genotypes of the offspring as also on parental haplotypes. Suppose the marker genotypes of the three offspring are: M_1M_1 (= M_{i3}), M_1M_2 (= M_{i4}), and M_1M_1 (= M_{i5}). Suppose, the classified QTL genotypes of these offspring are, respectively, A_1A_1 (= H_{i3}), A_1A_2 (= H_{i4}) and A_1A_2 (= H_{i5}) when the haplotypic configuration of the doubly heterozygous parent is $A_1M_1|A_2M_2$, and A_1A_2, A_1A_1 and A_1A_2 when the parental haplotypic configuration is $A_1M_2|A_2M_1$. Then,

$$L_i(\theta) = \frac{1}{2}\{\theta(1-\theta)^2 + (1-\theta)^3\} \tag{22}$$

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, we implement an EM procedure. For example, the complete likelihood corresponding to (22) would be:

$$L_i^*(\theta) = \frac{1}{2}\{\theta(1-\theta)^2\}^m\{(1-\theta)^3\}^{1-m} \tag{23}$$

where $m = \dfrac{\theta(1-\theta)^2}{\theta(1-\theta)^2 + (1-\theta)^3} = \theta.$

Thus, $L_i^*(\theta)$ 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,

$$L^*(\theta) = \left\{\prod_{i=1}^{K} c_i\right\}\theta^{\sum_{i=1}^{K} u_i}(1-\theta)^{\sum_{i=1}^{K} v_i} \tag{24}$$

which is easy to maximise giving

$$\hat{\theta} = \frac{\Sigma_{i=1}^{K} u_i}{\Sigma_{i=1}^{K}(u_i + v_i)} \tag{25}$$

Since u_i's and v_i's depend on θ, we need 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 $\hat{\theta}$ exceeds 0.5, we take $\hat{\theta} = 0.5$.

We finally note that in the first stage of this two-stage procedure, the estimated parameters are α, β, p, and σ^2. All these parameters are estimable from a sample of randly drawn individuals from the population. If indeed a random sample of individuals is available, then the above parameters can be estimated with trivial changes in the likelihood function derived above. The E and M steps also require trivial changes. Having estimated these parameters, one can sample families and initially classify only the parents into major QTL genotypes using the proposed classification rule (which requires the value of the quantitative trait of the individual to be classified and estimates of the parameters α, β, p, and σ^2). 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 interest. For this, we have examined the empirical frequency distributions of $\hat{\theta}$ based on multiple replicates of simulated data. 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). In the second step, using the data so generated, the trait parameters (α, β, σ^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' 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 I and II, 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 5 offspring per family, *NFAM* = 100 and *NREP* = 1,000. We have, in a later section entitled "Sample Size Effect", evaluated the effect of sample size.

Classification of Parents With Respect to QTL Genotypes

As mentioned earlier, in the first stage of the present 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 critically depends on the performance at the first stage. It is, therefore, important to evaluate how well parents are classified to their true genotypic classes by the present method. Results pertaining to classification of parents to their true genotypes using the proposed algorithm are provided in Figure 1a–c with *NOBS* = 1,000, *NOBS* = 250, and *NOBS* = 100, respectively. We have observed that though the classification performance was extremely good for *NOBS* = 1,000, the results were sufficiently satisfactory for *NOBS* = 250. We found that when there is no dominance (i.e., $\beta = 0$), between 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. This is expected because increase in the deviation of p from 0.5 increasingly polarises the

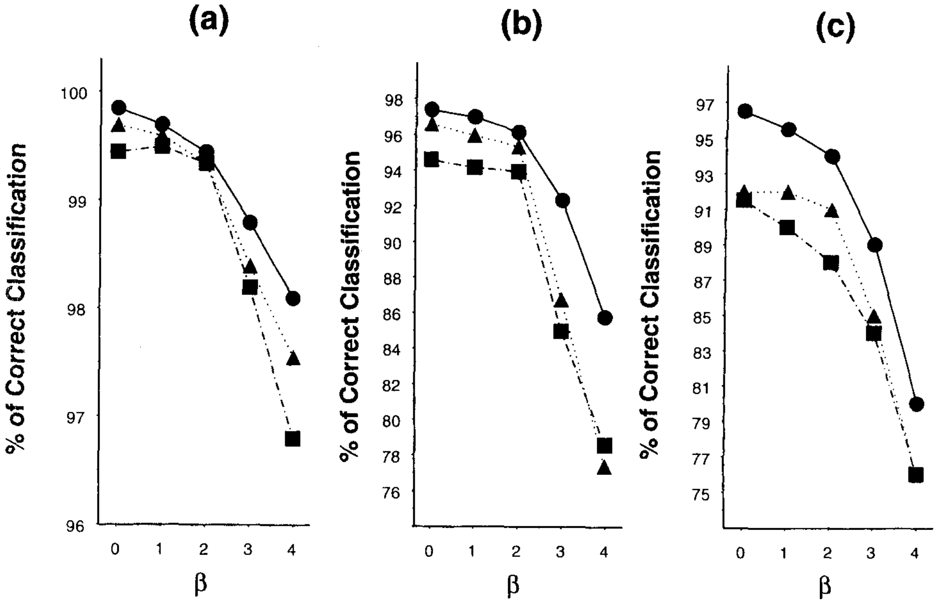

Fig. 1. Percentage of correct classification of parents for simulation parameter values $\alpha = 5$, $\beta = 0, 1, 2, 3, 4$, $\sigma^2 = 1$ and **(a)** *NOBS* = 1,000, **(b)** *NOBS* = 250, and **(c)** *NOBS* = 100. Circles correspond to p = 0.9, triangles to p = 0.7, and squares to p = 0.5.

distributions corresponding to the genotypes. The percentage of correct classification decreased as the extent of dominance (β) increases. The worst classification arose for $\alpha = 5$ and $\beta = 4$. In this case, the overlap between distributions of the A_1A_1 and A_1A_2 genotype classes was the largest. Therefore, a non-informative parent (i.e., with true genotype A_1A_1) has a high probability of being classified as informative (i.e., with true genotype A_1A_2) and vice versa. However, even in this case, the probability of correct classification was about 80%. We also noted that these results are independent of θ. Thus, it is seen that the first stage of the proposed method works extremely well, indicating that evaluation of the next stage, in which an estimate of the recombination fraction is obtained, is worthwhile.

Empirical Frequency Distribution of $\hat{\theta}$

If indeed the procedure provides a good estimate of the recombination fraction, θ, then one expects that the probability distribution of $\hat{\theta}$ obtained from multiple replications of simulated data generated using a fixed set of parameter values will be clustered around the true values of θ. Figures 2–10 depict the frequency distributions of $\hat{\theta}$ for simulation parameter values of $\theta = 0$, 0.1, 0.3 and 0.5, separately for p = 0.9, 0.7, 0.5, and $\beta = 0, 2, 4$. The values of the other parameters used in these simulations were: $\alpha = 5$ and $\sigma^2 = 1$. From Figures 2–10 it is seen that in all cases, except when the trait and marker loci are completely unlinked (i.e., $\theta = 0.5$) and the dominance effect (β) is large (Figs. 8d, 9d, and 10d, the distributions were unimodal and leptokurtic. In these extreme cases, there are higher probabilities of misclassification as has been noted in the previous section. For $\theta = 0$, in 80–85% of the replications $\hat{\theta}$

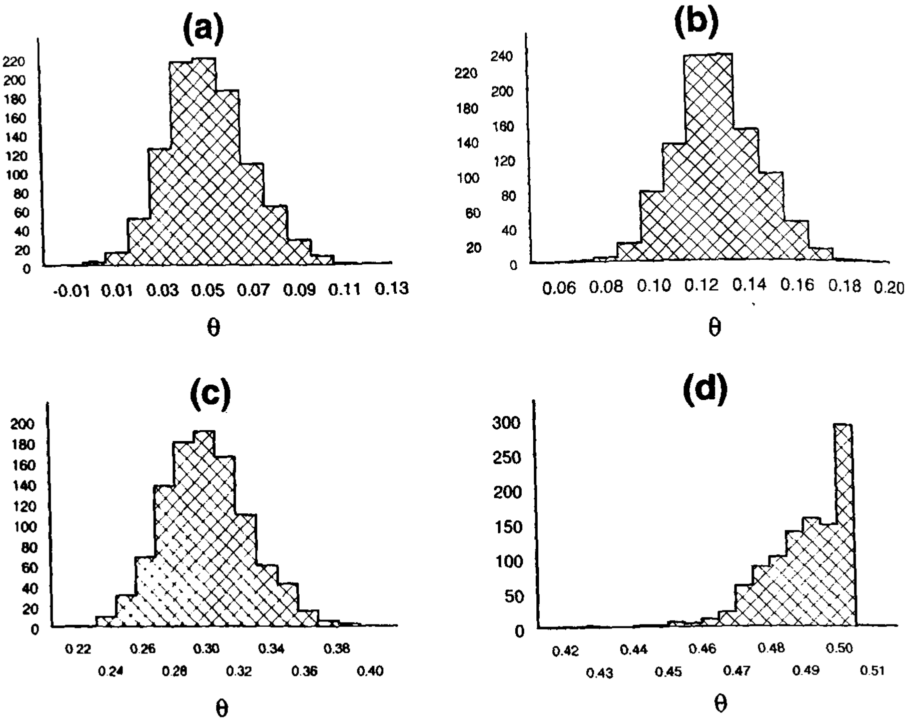

Fig. 2. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .9$, $\alpha = 5$, $\beta = 0$, $\sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$, and (d) $\theta = .5$.

was ≤ 0.08 if $\beta = 0$, while this percentage was between 65–70% if $\beta = 4$. Similarly for $\theta = 0.3$, in 80–90% of the replications $\hat{\theta}$ was in the interval [0.25, 0.35]. However, for $\theta = 0.5$, while 95% of the $\hat{\theta}$ values were between 0.45 and 0.5 for $p = 0.5$, this percentage for $p = 0.9$ was only about 75%. The proportion of $\hat{\theta}$ values lying close to the true value of θ decreased as β increased. Thus, it is seen that the procedure provides good estimates in conformity with expectations, unless the degree of dominance (β) is very high. Therefore, if the estimated values of β is close to that of α, the estimate of θ may be inaccurate.

Mean and Variance of $\hat{\theta}$ and Confidence Interval for θ

To examine the behavior of the estimator with respect to 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 $\theta = 0.0.1, 0.3, 0.5$. We have evaluated the means and variances of $\hat{\theta}$ and have obtained 95% confidence intervals of θ. These results are given in Table II. It is seen from Table III that the true value of θ was always included in the 95% confidence interval of θ. The coefficient of variation of $\hat{\theta}$ was also < 0.5%. These results indicate that the performance of the proposed estimator is extremely good. It is also seen from Table III that when p deviates from 0.5, the mean of $\hat{\theta}$ is closer to the true value of θ and the 95% confidence interval of θ is narrower, unless θ is very close to 0.5. The variance of $\hat{\theta}$ increased when p deviated more from 0.5. Table III also

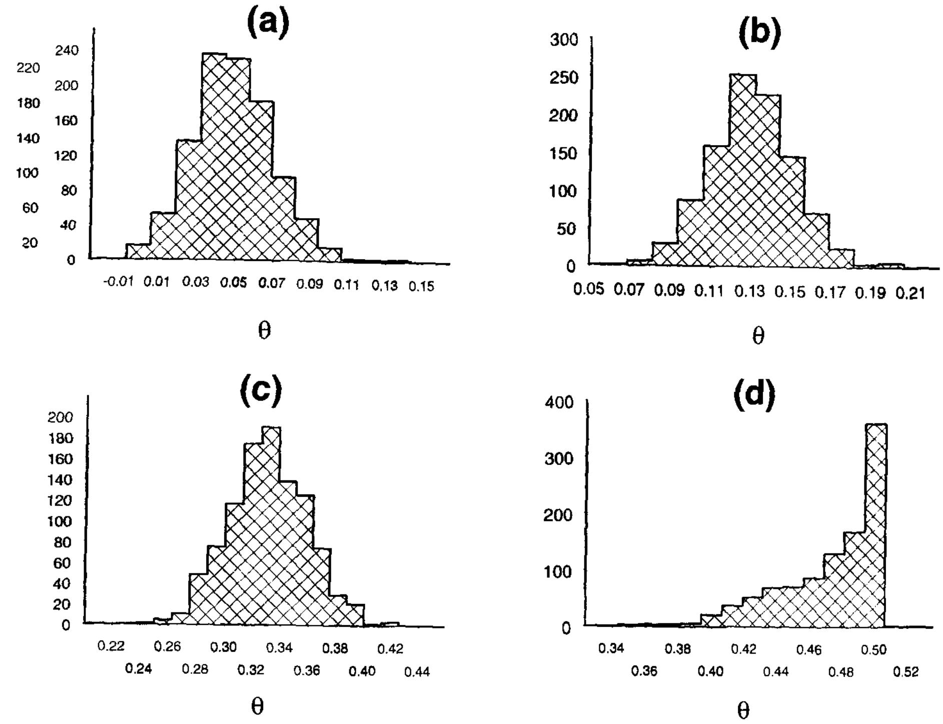

Fig. 3. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .7$, $\alpha = 5$, $\beta = 0$, $\sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$, and (d) $\theta = .5$.

shows that estimates of θ deviated more from the true value of θ, their variances increased, and the 95% confidence intervals became wider as the value of β increased to α. This phenomenon is due to increasing overlap between the two genotypic classes A_1A_1 and A_1A_2 as β becomes closer to α, resulting in errors of parental genotype classification. We also note that for fixed values of α, σ^2, p, and θ, the adverse effect of increase of β on estimation of θ is non-linear in β. For the sets of parameter values investigated, the relative error in estimation of θ never exceeded 20%.

Thus, for a fixed value of β, the efficiency of estimation of θ is dependent on both the true value of θ and the trait allele frequency, p, for any finite sample sizes of families. Since the trait genotype is unknown and needs to be inferred, the inference is better when the value of p is much deviated from 0.5, because in such cases the overlap of trait-value distributions among genotypes is small. However, the efficiency of estimation of θ strongly depends on the "effective" sample size (that is, the number of informative families), especially when the true value of θ is not close to 0. When the value of p deviates from 0.5, the effective sample size decreases, thereby reducing the efficiency of estimation of θ.

Sample Size Effect

We have investigated the performance of the proposed estimator when data on fewer offspring are available. Based on simulated data with 3 offspring per

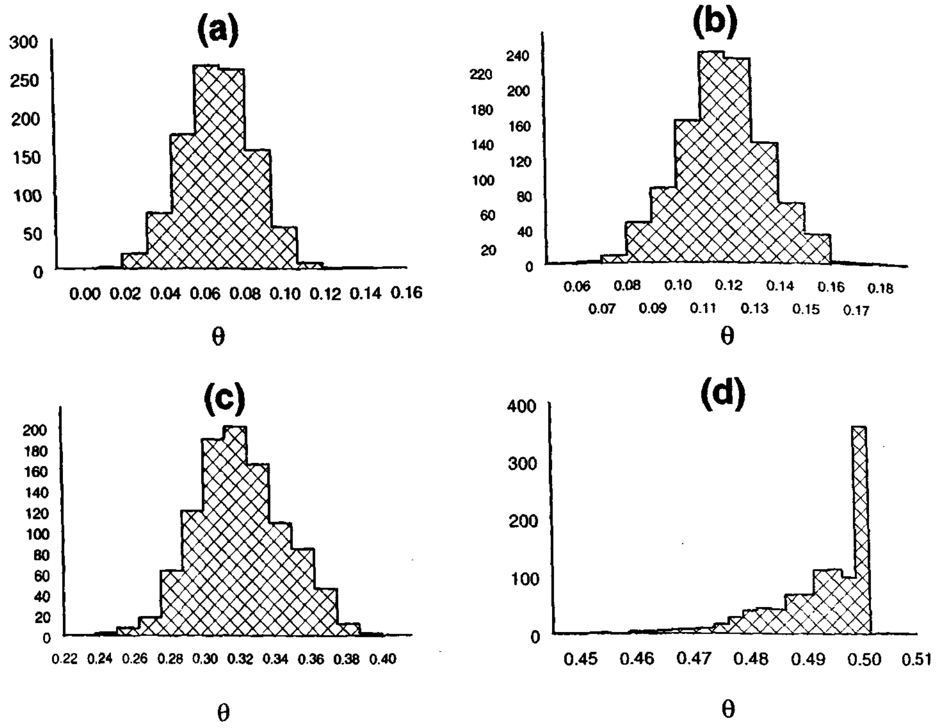

Fig. 4. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .5$, $\alpha = 5$, $\beta = 0$, $\sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$, and (d) $\theta = .5$.

family, we have obtained the empirical frequency distributions of $\hat{\theta}$ for different values of θ and $p = 0.5, 0.7, 0.9$. The frequency distributions were marginally less well-behaved compared to those based on 5 offspring per family. The estimated $\hat{\theta}$ values were also marginally more deviant from true θ values with smaller sibship sizes. For example, the mean and variance of $\hat{\theta}$ for simulation parameter values $\alpha = 5$, $\beta = 2$, $\sigma^2 = 1$, $p = 0.7$, and $\theta = 0.3$ were 0.342 and 0.00765, respectively, based on 100 informative families with 3 offspring per family, while these figures were 0.317 and 0.000683 with 5 offspring per family. For $p = 0.5$ and the remaining simulation parameter values as above, the mean of $\hat{\theta}$ was 0.377 with 3 offspring per family, while it was 0.321 with 5 offspring per family. The additional deviation of estimated θ from true θ due to decrease in sibship size from 5 to 3 varied between 8 and 20%. Thus, while larger data sets are desirable especially when p is close to 0.5, our method continues to perform rather well even with smaller sibship sizes. We also emphasize that although we have performed our simulation experiments with fixed sibship sizes of 5 and 3, variable sibship sizes pose no problem. Likelihood equations are easily modified and since data on offspring conditional on parental genotypes are independent, our simulation results are based effectively on 5 (or 3) × number of families.

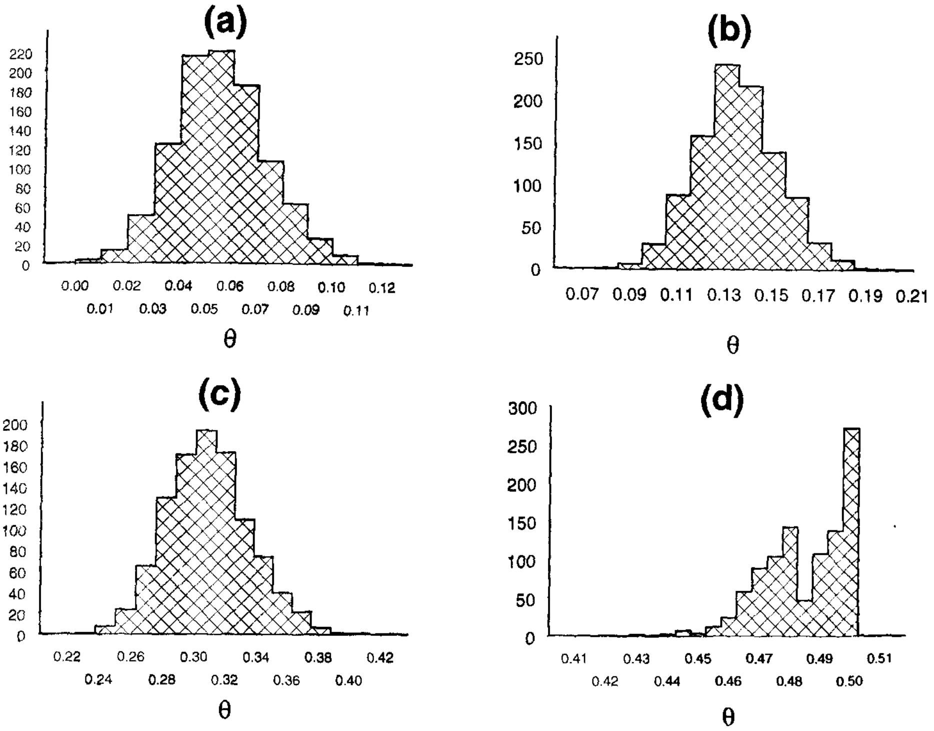

Fig. 5. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .9$, $\alpha = 5$, $\beta = 2$, $\sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3.$, and (d) $\theta = .5$.

Power of Test for Linkage Detection for Varying Degrees of Major Trait Locus Effect

To assess, more clearly, the efficiency of the proposed estimator $\hat{\theta}$ at varying degrees of major locus effect, measured as the proportion of variance of QT explained by the major biallelic locus (Δ), we have computed the empirical power of the test of hypothesis $H_0 : \theta = \theta_0 < 0.5$ vs. $H_1 : \theta = 0.5$. For this assessment, we have generated simulated data for different values of Δ. While such data can be generated for various combinations of parameters, we present results of β and σ^2 kept fixed at 0 and 1, respectively, with α and p varied suitably to attain different values of Δ in the range of 0.2 to 0.9. A fixed value of $\theta_0 = 0.1$ was used throughout. For each set of simulated data, the empirical 5% cut-off points for rejection for the null hypothesis was determined and the power was estimated as the proportion of replications (out of 1,000) with $\theta = 0.5$ in which $\hat{\theta}$ was greater than the empirical 5% cut-off point. The results are graphically presented in Figure 11 from which it is evident that our proposed method performs quite well, at least when the percentage of variance in QT explained by the major locus exceeds 30%. Other combinations of values of parameters β, σ^2, α, p yielding the same value of Δ resulted in approximately the same power; estimates are not provided.

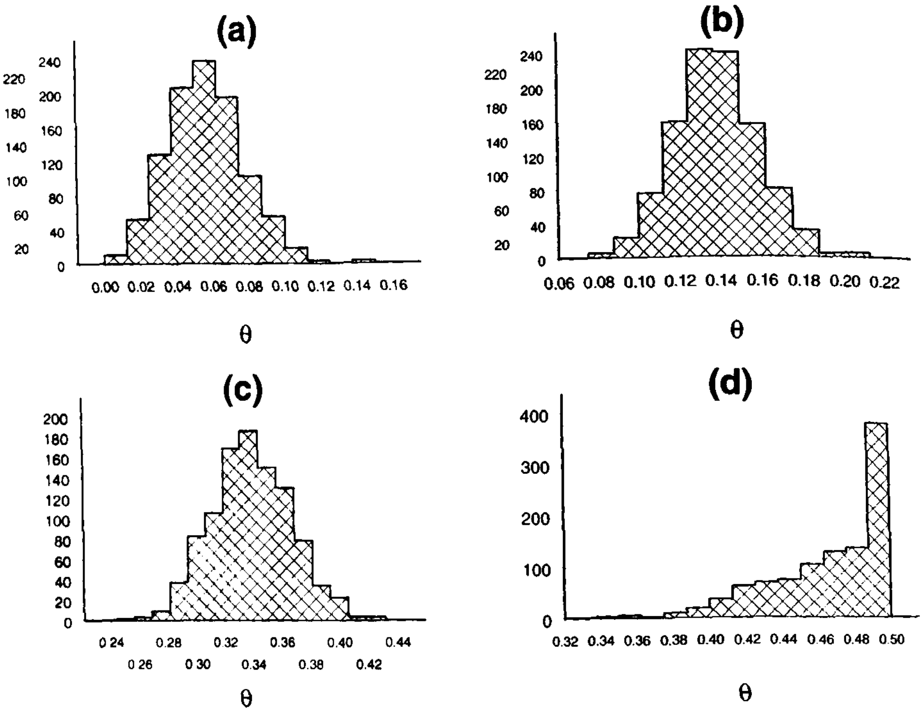

Fig. 6. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .7$, $\alpha = 5$, $\beta = 2$, $\sigma^2 = 1$ and **(a)** $\theta = 0$, **(b)** $\theta = .1$, **(c)** $\theta = .3$, and **(d)** $\theta = .5$.

Effect of Linkage Heterogeneity

Suppose a QT is controlled by a single major biallelic locus that explains Δ % of variance (the remaining variance being environmental), but there is linkage heterogeneity. That is, in a proportion (π) of families, the QT is due to one major locus and in the remaining proportion ($1 - \pi$) of families, the trait is due to an unlinked major biallelic locus. We assume that the values of Δ for both loci are equal, which in a sense is the worst-case scenario. Suppose, a biallelic marker is linked to the first QTL; that is, linkage is present in only π proportion of families. It is pertinent to examine the performance of the proposed procedure in estimating linkage from the pooled set of families, when one is unaware of the existence of the underlying linkage heterogeneity.

We have used simulations to examine this. In generating simulated data, we have used $\theta = 0.1$, two values of $\Delta = 80$ and 60% (in each case, α, β, and σ^2 were kept fixed at 5, 0, and 1, respectively; p was varied suitably to attain appropriate values of Δ, and five values of $\pi = 0.9$, 0.8, 0.7, 0.6, and 0.5 for each value of Δ. Results are presented in Table IV, from which it is seen that the estimated value of θ is reasonably good unless there is considerable linkage heterogeneity (small π) or the proportion of variance explained by the major locus is small (small Δ). When compared to the results presented in the previous section (and Fig. 11), we see that

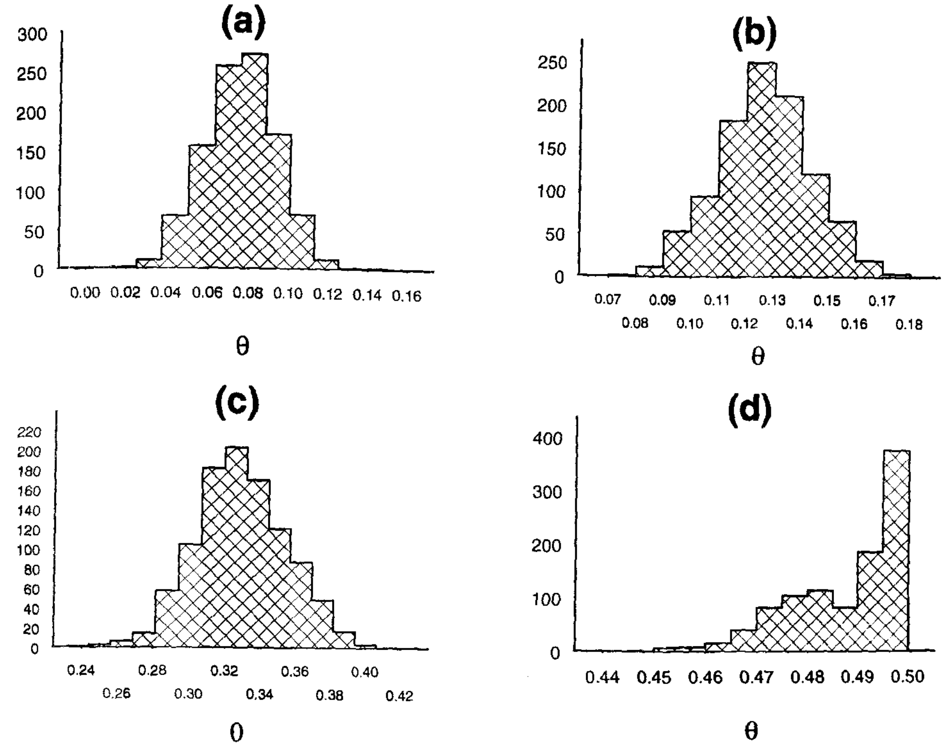

Fig. 7. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .5$, $\alpha = 5$, $\beta = 2$, $\sigma^2 = 1$ and **(a)** $\theta = 0$, **(b)** $\theta = .1$, **(c)** $\theta = .3$, and **(d)** $\theta = .5$.

while in the absence of linkage heterogeneity the proposed method performs quite well even when Δ is as low as 30%, in the presence of linkage heterogeneity the method fails to perform well unless Δ is as high as about 80%.

Analyzing a Two-Locus QT as a Single-Locus QT: Effect on Estimate of θ

Consider a QT that is controlled by two unlinked, biallelic trait loci. Suppose a biallelic marker is linked to one of the two loci. In the absence of knowledge that the QT is controlled by two major loci, it is reasonable to investigate the effect of analyzing data assuming that the QT is controlled by a single major locus, on the estimate of θ. To examine this issue, we generated simulated data sets for different values of α and p. We denote the values of α as α_1 and α_2 for the two trait loci. The values of β and p were held fixed at 0 and 0.7, respectively, for both loci; θ (recombination fraction between the marker and one QT) was taken as 0.1. Having generated replicate data sets for each combination of the above parameters, we used the proposed method for estimating θ assuming that the QT was controlled by a single major locus. The results are presented in Table V. It is seen from Table V that when the effect on the QT of the trait locus to which the marker is unlinked (the "unlinked QTL") is small relative to the linked trait locus, the method performs quite well even when the data are incorrectly analyzed assuming that the QT is controlled by a single

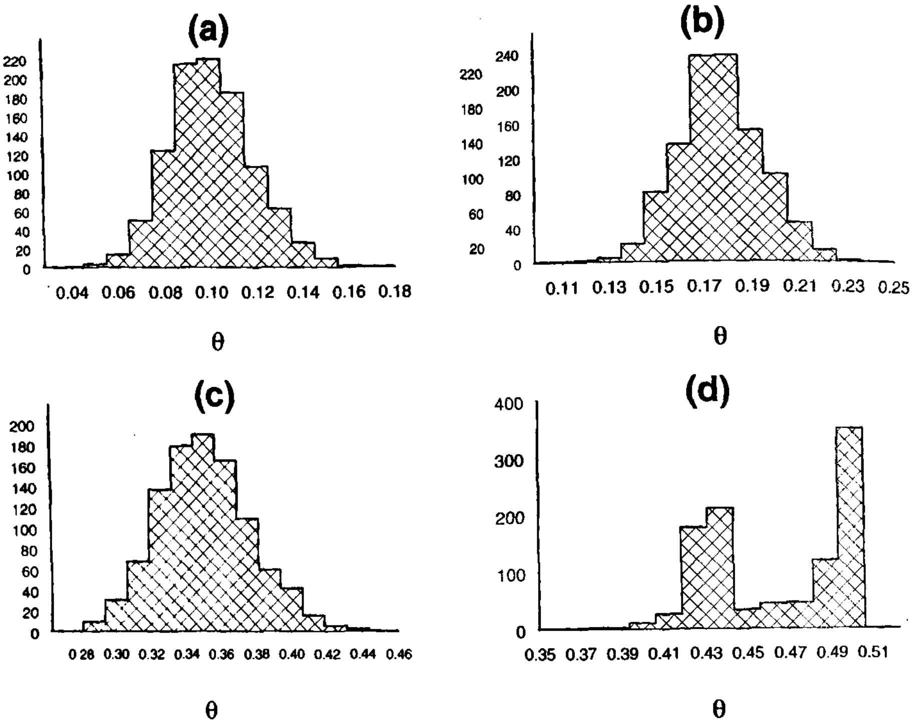

Fig. 8. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .9$, $\alpha = 5$, $\beta = 4$, $\sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$, and (d) $\theta = .5$.

major locus. However, the estimate of θ, which is always upwardly biased, worsens as the relative effect of the linked trait locus decreases. When the relative effects of the two trait loci are equal, analyzing the two-locus QT data as single-locus data leads to hopelessly bad estimates of the recombination fraction.

ESTIMATION OF θ WHEN THE MARKER IS MULTIALLELIC

The above procedure of estimation of θ can be easily shown to hold in the case of a multiallelic locus. Suppose the marker locus has K alleles denoted as $M_1, M_2, \ldots, M_K$. A mating between a homozygote and a heterozygote will be of the form $M_iM_i \times M_jM_k$, while a mating between two heterozygotes will be of the form $M_iM_j \times M_kM_l$.

A $M_iM_i \times M_jM_k$ mating will produce offspring with marker genotypes M_iM_j and M_iM_k with probability 1/2 each. The probabilities of the trait genotypes of the offspring for various parental mating types are identical to those corresponding to marker genotypes MM or Mm given in Table I. In the case of a mating between two heterozygotes, we need to differentiate between matings $M_iM_j \times M_iM_j$ and $M_iM_j \times M_kM_l$ where either $i \neq k$ or $j \neq l$. For $M_iM_j \times M_iM_j$ matings, the distributions of the trait genotypes of the offspring for varioius parental mating types are identical to those corresponding to marker genotypes MM, Mm, or mm given in Table II. $M_iM_j \times M_kM_l$

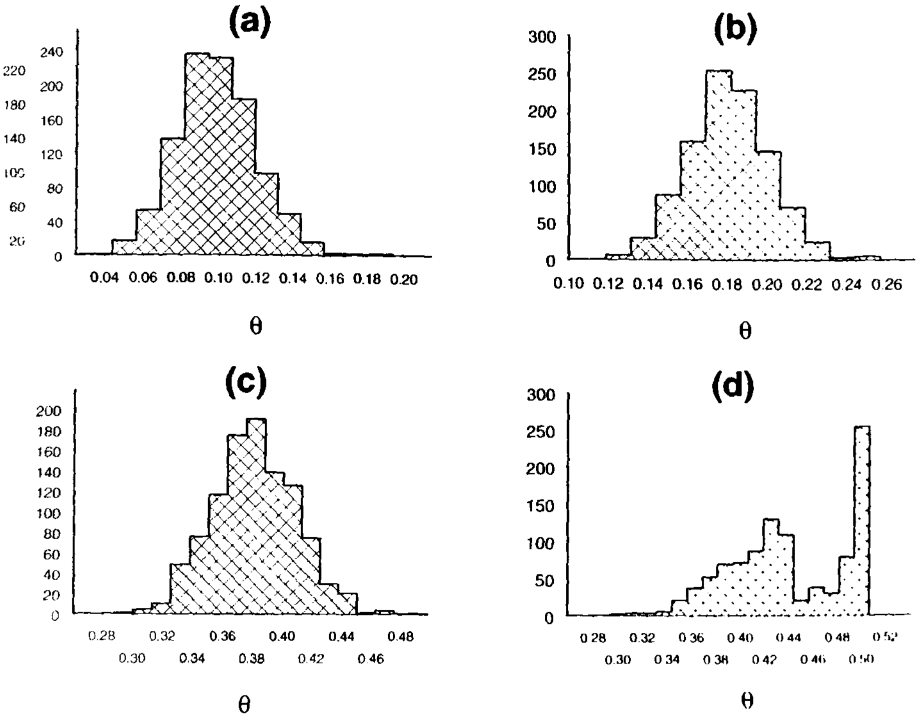

Fig. 9. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .7$, $\alpha = 5$, $\beta = 4$, $\sigma^2 = 1$ and **(a)** $\theta = 0$, **(b)** $\theta = .1$, **(c)** $\theta = .3$, and **(d)** $\theta = .5$.

($i \neq k$ or $j \neq l$) matings can produce offspring with marker genotypes M_iM_k, M_jM_k, M_iM_l, and M_jM_l with probability 1/4 each. The probabilities of the trait genotypes of the offspring for various parental mating types are given in Table VI.

Note that the estimation of the trait parameters α, σ^2, and p does not depend on the marker. Thus, the procedure of estimating these parameters in the case of a multiallelic marker is identical to that in the case of a biallelic marker described earlier. While estimating θ, we should consider the appropriate conditional distribution of the trait genotypes of the offspring given in Tables I, II, and VI. As all the probabilities in these tables are some multiples of θ, θ^2, $(1 - \theta)$, $(1 - \theta)^2$, $[\theta^2 + (1 - \theta)^2]$, we can use the EM procedure described earlier to obtain the m.l.e. of θ.

EM APPROACH IN MULTIPOINT MAPPING

The proposed EM procedure for mapping a trait locus using two-point linkage can be easily extended to the case of multipoint mapping. For ease of exposition, we consider a three-point mapping setup. Suppose the trait locus is flanked by two biallelic, codominant marker loci with alleles (M_1,m_1) and (M_2,m_2), respectively, such that the recombination fractions between the trait locus and the marker loci are θ_1 and θ_2, respectively. We assume that chromatid interference is absent and, hence, the

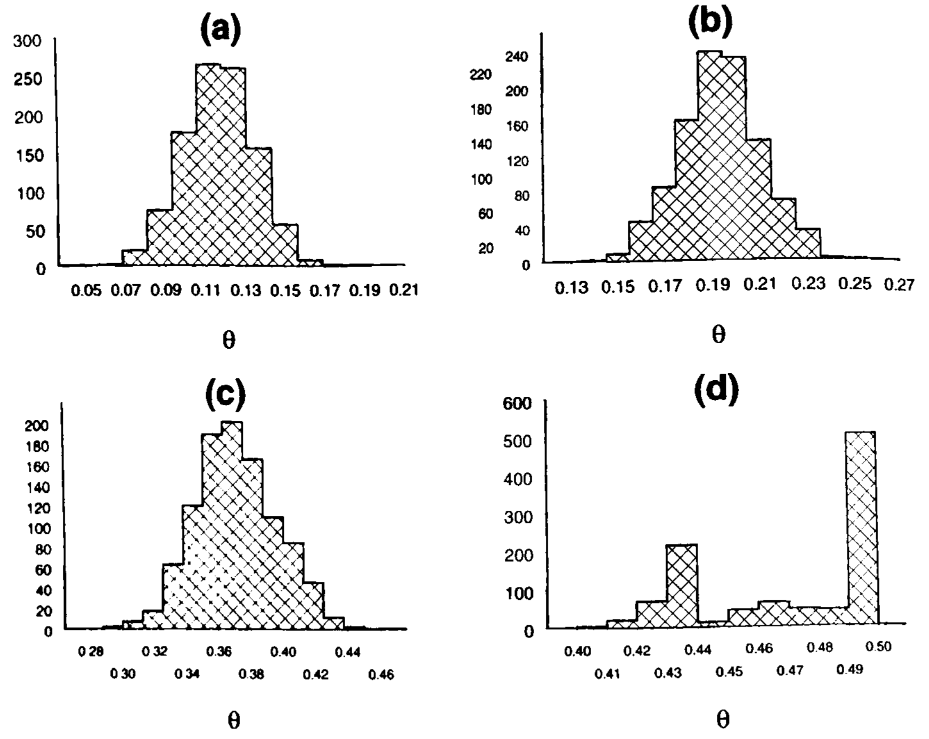

Fig. 10. Empirical frequency distributions of $\hat{\theta}$ for simulation parameter values $p = .5$, $\alpha = 5$, $\beta = 4$, $\sigma^2 = 1$ and (a) $\theta = 0$, (b) $\theta = .1$, (c) $\theta = .3$, and (d) $\theta = .5$.

recombination fraction between the two flanking markers is $\theta = \theta_1 + \theta_2 - 2\theta_1\theta_2$. The conditional probabilities of the trait genotypes of offspring given the parental trait and marker genotypes as well as the offspring marker genotypes are given in Table VII for backcross at both marker loci, in Table VIII for intercross at both marker loci, and in Table IX for backcross at one marker locus and intercross at the other.

We have used simulated data to assess the relative efficiency of multipoint linkage analysis over two-point linkage analysis. The simulation parameter values used were $\alpha = 1$; $\sigma^2 = 1$; $\beta = 0, 2, 4$; $p = 0.9, 0.7, 0.5$ and different values of θ_1 and θ_2. We first note that since the estimation of the trait parameters α, β, σ^2, and p do not involve marker information, the classification of parents into their true genotypes is identical to the case of two-point linkage. The EM algorithm invoked in the second stage of our proposed procedure to estimate θ_1 and θ_2 is also similar to the previous case, except that we need to consider the conditional trait genotypic distribution of the offspring given information at both the marker loci. We assume that the value of θ, i.e., the recombination fraction between the two marker loci is known a priori. Based on our simulated data, we have found that the histograms of $\hat{\theta}_1$ and $\hat{\theta}_2$ (i.e., the estimated values of θ_1 and θ_2) are much more well behaved and concentrated than in the case of two-point linkage, especially when a marker locus is unlinked to the trait locus (i.e., when θ_1 or θ_2 is 0.5). Two representative histograms are presented in Figures 12 and 13. We also find that using three-

TABLE III. Mean and Variance of $\hat{\theta}$ and 95% Confidence Interval of θ for $\alpha = 5$, $\sigma^2 = 1$, $p = .9$, 0.7, 0.5; $\beta = 0, 2, 4$; $\theta = 0, 0.1, 0.3, 0.5$

p	True θ	β	Mean ($\hat{\theta}$)	Var ($\hat{\theta}$)	95% C.I. of θ
.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)
.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)
.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)

point linkage analysis, the mean of $\hat{\theta}_1$ (or $\hat{\theta}_2$) is closer to the true value of θ_1 (or θ_2) than in the case of two-point linkage. The variances of the estimates in the case of multipoint linkage are also lower than those in the case of two-point linkage. Relevant statistics are provided in Table X. The relative efficiency of the three-point linkage analysis over the two-point linkage analysis (defined as the ratio of the variance of the estimate in the case of two-point linkage to that in the case of three-point linkage) was found to be about 1.3.

DISCUSSION

The proposed method of linkage detection exploits the fact that knowledge of parental genotypes at the QTL greatly eases statistical estimation of θ. Since for a

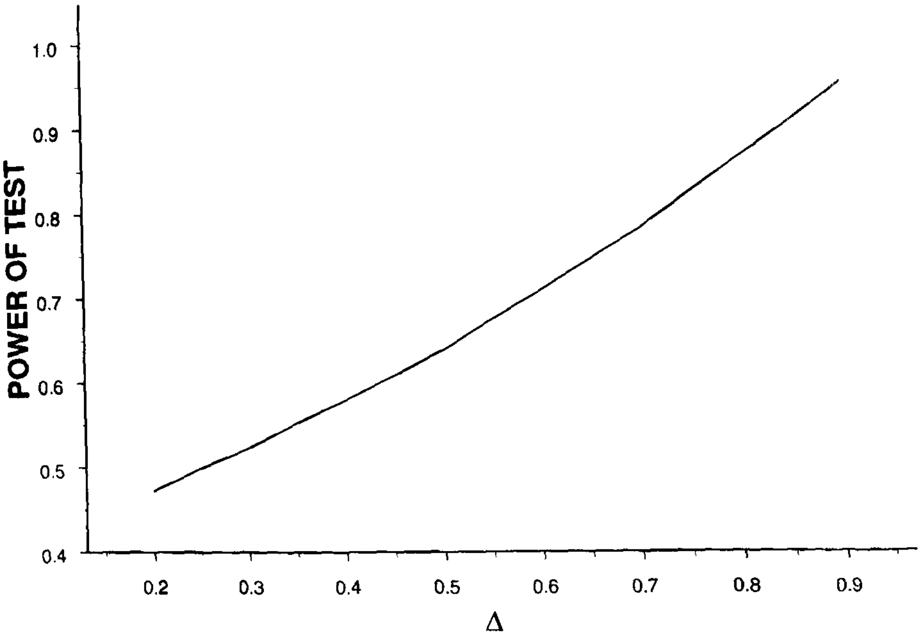

Fig. 11. Empirical power of the test procedure for detecting linkage for different values of the proportion of variance in QT explained by the major QTL.

quantitative character, the QTL genotype of an individual cannot be inferred with certainty because of intrinsic variability within genotype classes, we have used the EM algorithm coupled with a Bayes' classification procedure to classify parents into QTL genotype classes. A similar EM approach to estimate the trait parameters was used by Kao and Zeng [1997] in mapping a quantitative trait locus in an interval flanked by two markers. However, their procedure was based on inherent knowledge of haplotype information that is not readily available in human genetic studies. Moreover, the effect of marker genotype on trait value was assumed to be linear. The procedure proposed by us does not use these assumptions. In our procedure, estimates of trait parameters and recombination fraction are obtained. The estimates of

TABLE IV. Mean and Variance of Recombination Fraction in the Presence of Linkage Heterogeneity (π = Proportion of Linked Families) Estimated From Simulated Data Sets With Differing Values of Δ (Percentage of Variance Explained by the Major QTL) and Recombination Fraction = 0.1

	Δ = 80%		Δ = 60%	
π	Mean ($\hat{\theta}$)	Var ($\hat{\theta}$)	Mean ($\hat{\theta}$)	Var ($\hat{\theta}$)
0.9	0.126	0.0041	0.1466	0.0075
0.8	0.1415	0.0062	0.1683	0.0112
0.7	0.167	0.0091	0.1975	0.0172
0.6	0.194	0.0143	0.223	0.0237
0.5	0.2268	0.0197	0.2549	0.0294

TABLE V. Effect of Analyzing Two-Locus QT Data as a Single-Locus Data, on Estimated θ; When Its True Value is 0.1 (α_1 and α_2 Denote the Effects of the Two QTLs), Δ Is the Proportion of Variance Explained Jointly by the Two QTLs and Δ_1 Is the Proportion Explained by the Linked QTL

α_1	α_2	Δ	Δ_1	Mean ($\hat{\theta}$)	Var ($\hat{\theta}$)
5	1	0.91	0.88	0.1088	0.0007
5	2	0.92	0.79	0.1291	0.0016
5	3	0.93	0.68	0.1516	0.0031
5	4	0.94	0.57	0.1774	0.0058
5	5	0.95	0.47	0.2056	0.0094

trait parameters are used in inferring the parental QTL genotypes. 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. The first stage of our procedure does not use marker genotype information. Thus, even if families are sampled, it will be prudent to initially obtain only measurements of the quantitative trait on parents. The EM algorithm implemented in the first stage will provide estimates of trait parameters. Having obtained these estimates, we can classify parents in families into major QTL genotypes, using the proposed classification rule. This enables identification of potentially informative families (i.e., at least one parent heterozygous at the major QTL). Then, the investigator can obtain genotype information at marker loci on both parents and measurements of quantitative trait on offspring in those families in which at least one parent is doubly heterozygous. Thus, the proposed two-stage procedure provides cost effectiveness in terms of data collection. In the second stage, based on data on only informative families, the proposed EM algorithm provides the maximum likelihood estimate of the recombination fraction between a marker locus and the major QTL. We note that in the context of human pedigree analysis, the EM algorithm was first applied by Ott [1977].

We have shown that our proposed method results in virtually error-free classification of parental QTL genotypes, unless the dominance effect is very large. We have also shown, using simulations, that for a wide range of parameter values that corresponds to widely different values of the proportion of variance in QT explained by the major locus, the estimates of recombination fractions and the power to detect linkage are quite good for reasonable sample sizes. We have shown that our method performs more efficiently when data on multiple markers flanking the trait locus are used. One major advantage of the proposed method is that the estimation of recombination fractions is not as strongly tied to estimates of QTL parameters as in lod-score analysis. Through the use of EM algorithm, the present procedure extracts appropriate information from the quantitative data and then uses a Bayesian classification rule to transform the QT data to qualitative genotypes before estimating θ from the transformed data by a likelihood-based method. This reduces the impact of error in estimating trait parameters on the estimate of θ. Because of the weak dependence of $\hat{\theta}$ on estimates of trait parameters, and because no separate segregation and linkage analyses need to be performed in the present approach, the earlier observations that model misspecification can seriously affect estimates of trait parameters

TABLE VI. Trait Locus Mating Types Among $M_iM_j \times M_kM_l$ Parents, Mating Probabilities, and Probabilities of Trait Locus Genotypes Among Offspring With Marker Genotype M_iM_k and M_jM_k*

			$\pi_g(M_iM_k)$			$\pi_g(M_jM_k)$		
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	$\frac{1}{4}$	0	0
2	$A_1A_1 \times A_1a_1$	$p_1^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0
3	$A_1a_1 \times A_1A_1$	$p_1^3p_2$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0
4	$A_1A_1 \times a_1A_1$	$p_1^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0
5	$a_1A_1 \times A_1A_1$	$p_1^3p_2$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0
6	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2p_1^2p_2^2$	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$	0
7	$A_1a_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$
8	$A_1a_1 \times a_1A_1$	$p_1^2p_2^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$
9	$a_1A_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}\theta(1-\theta)^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}\theta^2$
10	$a_1a_1 \times A_1a_1$	$p_1p_2^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$
11	$A_1a_1 \times a_1a_1$	$p_1p_2^3$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$
12	$a_1A_1 \times a_1A_1$	$p_1^2p_2^2$	$\frac{1}{4}\theta^2$	$\frac{1}{2}\theta(1-\theta)$	$\frac{1}{4}(1-\theta)^2$	$\frac{1}{4}\theta(1-\theta)$	$\frac{1}{4}[1-2\theta(1-\theta)]$	$\frac{1}{4}\theta(1-\theta)$
13	$a_1a_1 \times a_1A_1$	$p_1p_2^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$
14	$a_1A_1 \times a_1a_1$	$p_1p_2^3$	0	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	0	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$
15	$a_1a_1 \times a_1a_1$	p_2^4	0	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$

*Probabilities of trait locus genotypes among offspring with markes genotype M_jM_l and M_iM_l can be obtained by replacing θ by $(1-\theta)$ in the blocks corresponding to the genotype M_iM_k and M_jM_k, respectively, in this table.

TABLE VII. Trait Locus Mating Types Among *MMNN* × *MnNn* and *MMNn* × *MmNN* Parents, Mating Probabilities, and Probabilities of Trait Locus Genotypes Among Offspring With Marker Genotype *MMNN**

			$\pi_g(MMNN \times MmNn)$			$\pi_g(MMNn \times MmNN)$		
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	1	0	0	1	0	0
2	$A_1A_1 \times A_1a_1$	$p_1^3p_2$	$\frac{(1-\theta_1)(1-\theta_1)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$(1-\theta_1)$	θ_1	0
3	$A_1A_1 \times a_1A_1$	$p_1^3p_2$	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	θ_1	$(1-\theta_1)$	0
4	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2p_1^2p_2^2$	0	1	0	0	1	0
5	$A_1a_1 \times A_1A_1$ $a_1A_1 \times A_1A_1$	$2p_1^3p_2$	$\frac{1}{2}$	$\frac{1}{2}$	0	$1-\theta_2$	θ_2	0
6	$A_1a_1 \times A_1a_1$ $a_1A_1 \times A_1a_1$	$2p_1^2p_2^2$	$\frac{(1-\theta_1)(1-\theta_2)}{2(1-\theta)}$	$\frac{1}{2}$	$\frac{\theta_1\theta_2}{2(1-\theta)}$	$(1-\theta_1)(1-\theta_2)$	$1-\theta$	$\theta_1\theta_2$
7	$A_1a_1 \times a_1A_1$ $a_1A_1 \times a_1a_1$	$2p_1^2p_2^2$	$\frac{\theta_1\theta_2}{2(1-\theta)}$	$\frac{1}{2}$	$\frac{(1-\theta_1)(1-\theta_2)}{2(1-\theta)}$	$\theta_1\theta_2$	$1-\theta$	$(1-\theta_1)(1-\theta_2)$
8	$A_1a_1 \times a_1a_1$ $a_1A_1 \times a_1a_1$	$2p_1p_2^3$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$1-\theta_2$	θ_2
9	$a_1a_1 \times A_1a_1$	$p_1p_2^3$	0	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$(1-\theta_1)$	θ_1
10	$a_1a_1 \times a_1A_1$	$p_1p_2^3$	0	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	θ_1	$(1-\theta_1)$
11	$a_1a_1 \times a_1a_1$	p_2^4	0	0	1	0	0	1

*Probabilities of trait locus genotypes among offspring with marker genotypes *MmNn*, *MMNn* and *MmNN* can be obtained by replacing θ_1 by $(1-\theta_1)$ and θ_2 by $(1-\theta_2)$; θ_2 by $(1-\theta_2)$ and θ by $(1-\theta)$; and θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$, respectively, in this table.

TABLE VIII. Trait Locus Mating Types Among *MmNn* × *MmNn* Parents, Mating Probabilities, and Probabilities of Trait Locus Genotypes Among Offspring With Marker Genotypes *MMNN*, *MMNn*, and *MmNn**

			$\pi_g(MMNN)$			$\pi_g(MMNn)$			$\pi_g(MmNn)$		
g	Mating type	Probability	A_1A_1	A_1a_1	a_1a_1	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	1	0	0	1	0	0	1	0	0
2	$A_1A_1 \times A_1a_1$ $A_1a_1 \times A_1A_1$	$2p_1^3p_2$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$\frac{1}{2}(1-\theta_1)\left\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\right\}$	$\frac{1}{2}\theta_1\left\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\right\}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0
3	$A_1A_1 \times a_1A_1$ $a_1A_1 \times A_1A_1$	$2p_1^3p_2$	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	$\frac{1}{2}\theta_1\left\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\right\}$	$\frac{1}{2}(1-\theta_1)\left\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\right\}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0
4	$A_1a_1 \times A_1A_1$ $a_1a_1 \times A_1A_1$	$2p_1^2p_2^2$	0	1	0	0	1	0	0	1	0
5	$A_1a_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{(1-\theta_1)^2(1-\theta_2)^2}{(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)^2(1-\theta_2)^2}{(1-\theta)^2}$	$\frac{\theta_1^2\theta_2^2}{(1-\theta)^2}$	$\frac{(1-\theta_1)^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$\frac{\theta_1^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{2\theta_1(1-\theta_1)\theta_2(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$
6	$a_1A_1 \times A_1a_1$ $A_1a_1 \times a_1A_1$	$2p_1^2p_2^2$	$\frac{\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$1-\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$\frac{\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$\frac{\theta_2(1-\theta_2)\{1-2\theta_1(1-\theta_1)\}}{\theta(1-\theta)}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$
7	$a_1A_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{\theta_1^2\theta_2^2}{(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{(1-\theta)^2}$	$\frac{(1-\theta_1)^2(1-\theta_2)^2}{(1-\theta)^2}$	$\frac{\theta_1^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{\theta_1(1-\theta_1)\{1-2\theta_2(1-\theta_2)\}}{\theta(1-\theta)}$	$\frac{(1-\theta_1)^2\theta_2(1-\theta_2)}{\theta(1-\theta)}$	$\frac{2\theta_1(1-\theta_1)\theta_2(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$1-\frac{4\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$	$\frac{2\theta_1\theta_2(1-\theta_1)(1-\theta_2)}{\theta^2+(1-\theta)^2}$
8	$A_1a_1 \times a_1a_1$ $a_1a_1 \times A_1a_1$	$2p_1p_2^3$	0	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	$\frac{\theta_1\theta_2}{1-\theta}$	0	$\frac{1}{2}(1-\theta_1)\left\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\right\}$	$\frac{1}{2}\theta_1\left\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\right\}$	0	$\frac{1}{2}$	$\frac{1}{2}$
9	$a_1a_1 \times a_1A_1$ $a_1A_1 \times a_1a_1$	$2p_1p_2^3$	0	$\frac{\theta_1\theta_2}{1-\theta}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\theta}$	0	$\frac{1}{2}\theta_1\left\{\frac{1-\theta_2}{\theta}+\frac{\theta_2}{1-\theta}\right\}$	$\frac{1}{2}(1-\theta_1)\left\{\frac{\theta_2}{\theta}+\frac{1-\theta_2}{1-\theta}\right\}$	0	$\frac{1}{2}$	$\frac{1}{2}$
10	$a_1a_1 \times a_1a_1$	p_2^4	0	0	1	0	0	1	0	0	1

*Probabilities of trait locus genotypes among offspring with marker genotype *MMnn*, *mmNN* and *mmmm* can be obtained by replacing θ_2 by $(1-\theta_2)$ and θ by $(1-\theta)$; θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$; and θ_1 by $(1-\theta_1)$ and θ_2 by $(1-\theta_2)$ in the block corresponding to genotype *MMNN* and those of marker genotype *mmNn*, *MmNN*, and *Mmnn* can be obtained by replacing θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$; θ_1 by θ_2 by θ_1; and θ_1 by $(1-\theta_2)$, θ_2 by $(1-\theta_1)$, and θ by $(1-\theta)$, respectively, in the block corresponding to genotype *MMNn* in this table.

TABLE IX. Trait Locus Mating Types Among *MMNn* × *MmNn* Parents, Mating Probabilities, and Probabilities of Trait Locus Genotypes Among Offspring With Marker Genotypes *MMNN* and *MMNn**

			π_g (*MMNN*)			π_g (*MMNn*)		
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	1	0	0	1	0	0
2	$A_1A_1 \times A_1a_1$	$p_1^3p_2$	$\frac{(1-\theta_1)(1-\theta_2)}{1-6}$	$\frac{\theta_1\theta_2}{1-6}$	0	$1-\theta_1$	θ_1	0
3	$A_1A_1 \times a_1A_1$	$p_1^3p_2$	$\frac{\theta_1\theta_2}{1-6}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-6}$	0	θ_1	$1-\theta_1$	0
4	$A_1A_1 \times a_1a_1$ $a_1a_1 \times A_1A_1$	$2p_1^2p_2^2$	0	1	0	0	1	0
5	$A_1a_1 \times A_1A_1$	$p_1^3p_2$	$1-\theta_2$	θ_2	0	$\theta(1-\theta_2)+\theta_2(1-\theta)$	$\theta\theta_2+(1-\theta)(1-\theta_2)$	0
6	$a_1A_1 \times A_1A_1$	$p_1^3p_2$	θ_2	$1-\theta_2$	0	$\theta\theta_2+(1-\theta)(1-\theta_2)$	$\theta(1-\theta_2)+th_2(1-\theta)$	0
7	$A_1a_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{(1-\theta_1)(1-\theta_2)^2}{1-6}$	$\frac{\theta_2(1-\theta_2)}{1-6}$	$\frac{\theta_1\theta_2}{1-6}$	$2(1-\theta_1)\theta_2(1-\theta_2)$	$1-2\theta_2(1-\theta_2)$	$2\theta_1\theta_2(1-\theta_2)$
8	$a_1A_1 \times A_1a_1$	$p_1^2p_2^2$	$\frac{(1-\theta_1)\theta_2(1-\theta_2)}{1-6}$	$1-\frac{\theta_2(1-\theta_2)}{1-6}$	$\frac{\theta_1\theta_2(1-\theta_2)}{1-6}$	$\frac{1}{2}\theta_1\{1-\theta_2(1-\theta_2)\}$	$2\theta_2(1-\theta_2)$	$\frac{1}{2}(1-\theta_2)\{1-\theta_2(1-\theta_2)\}$
9	$A_1a_1 \times a_1A_1$	$p_1^2p_2^2$	$\frac{\theta_1\theta_2(1-\theta_2)}{1-6}$	$1-\frac{\theta_2(1-\theta_2)}{1-6}$	$\frac{(1-\theta_1)\theta_2(1-\theta_2)}{1-6}$	$\frac{1}{2}\theta_1\{1-\theta_2(1-\theta_2)\}$	$2\theta_2(1-\theta_2)$	$\frac{1}{2}(1-\theta_2)\{1-\theta_2(1-\theta_2)\}$
10	$a_1A_1 \times a_1A_1$	$p_1^2p_2^2$	$\frac{\theta_1\theta_2^2}{1-6}$	$\frac{\theta_2(1-\theta_2)}{1-6}$	$\frac{(1-\theta_1)(1-\theta_2)^2}{1-6}$	$2\theta_1\theta_2(1-\theta_2)$	$1-2\theta_2(1-\theta_2)$	$2(1-\theta_1)\theta_2(1-\theta_2)$
11	$A_1a_1 \times a_1a_1$	$p_1p_2^3$	0	$1-\theta_2$	θ_2	0	$\theta(1-\theta_2)+\theta_2(1-\theta)$	$\theta\theta_2+(1-\theta)(1-\theta_2)$
12	$a_1A_1 \times a_1a_1$	$p_1p_2^3$	0	θ	$1-\theta_2$	0	$\theta\theta_2+(1-\theta)(1-\theta_2)$	$\theta(1-\theta_2)+th_2(1-\theta)$
13	$a_1a_1 \times A_1a_1$	$p_1p_2^3$	0	$\frac{(1-\theta_1)(1-\theta_2)}{1-6}$	$\frac{\theta_1\theta_2}{1-6}$	0	$1-\theta_1$	θ_1
14	$a_1a_1 \times a_1A_1$	$p_1p_2^3$	0	$\frac{\theta_1\theta_2}{1-6}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-6}$	0	θ_1	$1-\theta_1$
15	$a_1a_1 \times a_1a_1$	p_2^4	0	0	1	0	0	1

*Probabilities of trait locus genotypes among offspring with marker genotype *MMnn*, *MmNN*, and *Mmnn* can be obtained by replacing θ_2 by $(1-\theta_2)$ and θ by $(1-\theta)$; θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$; and θ_1 by $(1-\theta_1)$ and θ_2 by $(1-\theta_2)$, respectively, in the block corresponding to the genotype *MMNN* and that of marker genotype *MmNn* can be obtained by replacing θ_1 by $(1-\theta_1)$ and θ by $(1-\theta)$ in the block corresponding to the genotype *MMNn* in this table.

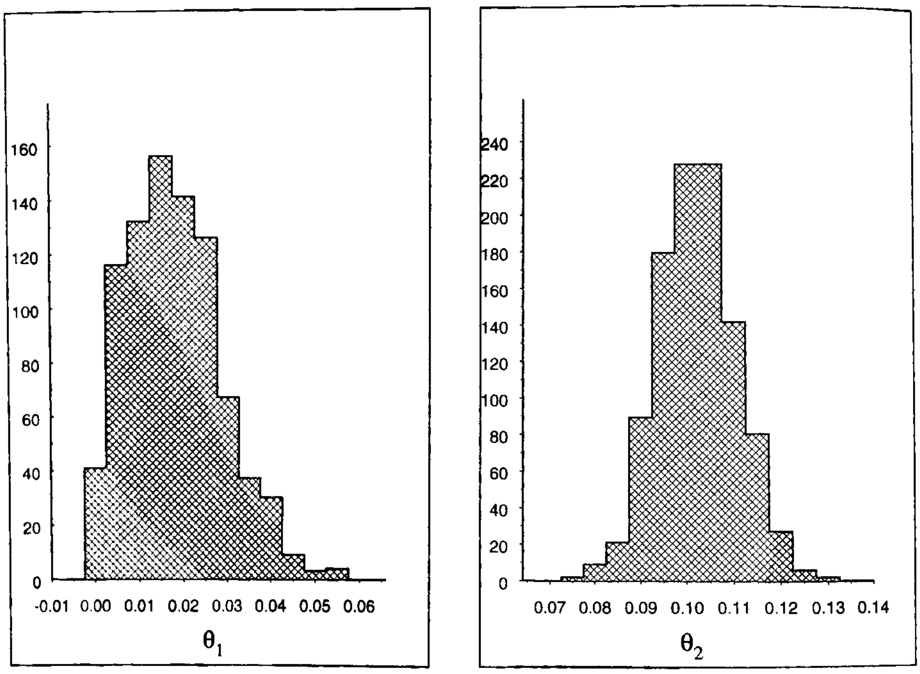

Fig. 12. Empirical frequency distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ for simulation parameter values $p = .9$, $\alpha = 5$, $\beta = 0$, $\sigma^2 = 1$, $\theta_1 = 0$, $\theta_2 = 0.1$.

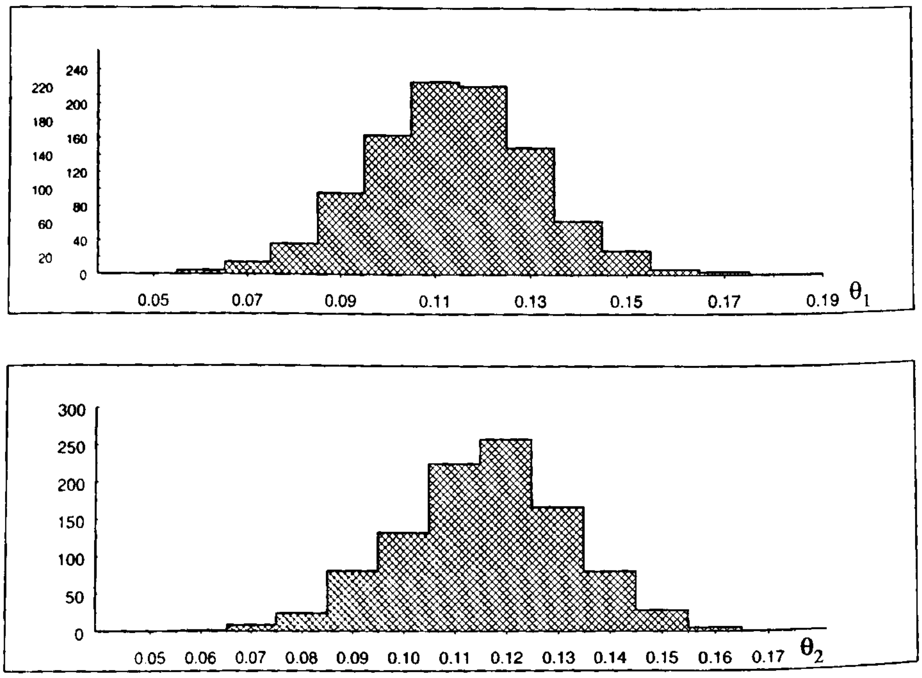

Fig. 13. Empirical frequency distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ for simulation parameter values $p = .7$, $\alpha = 5$, $\beta = 2$, $\sigma^2 = 1$, $\theta_1 = 0.1$, $\theta_2 = 0.1$.

TABLE X. Mean and Variance of $(\hat{\theta}_1)$ and $(\hat{\theta}_2)$ and 95% C.I. of θ_1 and θ_2 for $\alpha = 5$ and $\sigma^2 = 1$

p	β	θ_1	θ_2	Mean($\hat{\theta}_1$)	Var($\hat{\theta}_2$)	95% C.I. of θ_1	Mean($\hat{\theta}_1$)	Var($\hat{\theta}_2$)	95% C.I. of θ_2
.9	0	0	.1	.013	.000156	(.007, .023)	.102	.000067	(.097, .115)
.7	2	.1	.1	.113	.000238	(.098, .124)	.115	.000242	(.096, .126)
.9	4	.1	.3	.170	.000945	(.132, .194)	.362	.001583	(.341, .395)
.5	2	.3	.3	.316	.000586	(.287, .328)	.314	.000588	(.288, .325)
.5	4	.3	.5	.372	.001658	(.359, .416)	.444	.000971	(.420, .498)
.7	0	.5	.1	.492	.000065	(.481, .500)	.101	.000068	(.096, .119)

[Dizier et al., 1993; Atwood et al., 1995] and that prior segregation analysis can reduce the power to detect linkage [Atwood and Slifer, 1997] become less relevant. Further, this approach leads to a considerable reduction in computational load, which in usual parametric segregation and linkage analyses of a QT can be sufficiently heavy to require the use of a supercomputer [Atwood and Slifer, 1997]. Joint segregation and linkage are computationally even more expensive, although there are some indications that it may be more powerful [Gauderman et al., 1997] than separate segregation and linkage analyses. We have shown that for reasonable levels of linkage heterogeneity, the proposed method performs quite well. Model misspecification, treating a two-locus QT as a single locus QT, even though yields biased estimates of θ, leads to gross errors in inference only when both trait loci have nearly equal effects. Thus, the present method appears to be quite useful and robust for QTL mapping. Compared to numerical maximization of the likelihood 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.

ACKNOWLEDGMENT

We are grateful to an anonymous reviewer for comments that led to substantial improvement of this paper.

REFERENCES

Amos CI, Elston RC. 1989. Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 6:349–60.

Atwood LD, Slifer SH. 1997. Prior segregation analysis and the power to detect linkage. Genet Epidemiol 14:755–60.

Atwood LD, Mitchell BD, Stowell NC. 1995. Segregation and linkage analysis of the complex trait Q1. Genet Epidemiol 12:713–8.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. R Stat Soc, Ser B 39:1–38.

Dizier M, Bonaiti-Pellie C, Clerget-Darpoux F. 1993. Conclusions of segregation analysis for family data generated under two locus models. Am J Hum Genet 53:1338–46.

Everitt BS, Hand DJ. 1981. Finite mixture distributions. London: Chapman and Hall.

Fergusson TS. 1967. Mathematical Statistics: A Decision-theoretic Approach. New York: Academic Press.

Gauderman WJ, Faucett CL, Morrison JL, Carpenter CL. 1997. Joint segregation and linkage analysis of a quantitative trait compared to separate analyses. Genet Epidemiol 14:993–98.

Goldgar DE. 1990. Multipoint analysis of human quantitative genetic variation. Am J Hum Genet 47:957–67.

Haley CS, Knott SA. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–24.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19.

Hill A. 1975. Quantitative linkage: a statistical procedure for its detection and estimation. Ann Hum Genet 38:439–49.

Jayakar SD. 1970. On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. Biometrics 26:451–64.

Kao CH, Zeng ZB. 1997. General formulas for observing the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics 53:653–65.

Kruglyak L, Lander ES. 1995. A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421–28.

Lander ES, Botstein D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–99.

McLachlan GJ, Krishnan T. 1997. The EM algorithm and extensions. New York: John Wiley and Sons.

Ott J. 1977. Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. Ann Hum Genet 40:443–54.

Rao CR. 1973. Linear Statistical Inference and its applicatons, 2nd ed. New York: Wiley.

Sax K. 1923. The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics 8:552–60.

Schork NJ, Nath SK, Lindpainter K, Jacob HJ. 1996. Extensions to quantitative trait locus mapping in experimental organisms. Hypertension 28:1104–11.

Weller JI. 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–40.

Whittaker JC, Thompson R, Visscher PM. 1995. On the mapping of QTL by regression of phenotype on markertype. Heredity 77:23–32.

Zeng ZB. 1994. Precision mapping of quantitative trait loci. Genetics 136:1457–68.