# Mapping a human quantitative trait

**Saurabh Ghosh and Partha P. Majumder***

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

**Quantitative traits in humans are generally determined jointly by interacting genetic and environmental factors. For many traits that are highly heritable, environmental factors are generally found to be relatively unimportant. Genes underlying such traits may vary from one to a few, which explains most of the variance in trait values. Such genes are known as 'major genes'. It is of interest and importance to determine the locations of such major genes on chromosomes. This is done by statistical analysis of data on members of families or on sib-pairs. The data on each individual belonging to such sets of relatives comprise measurement of the value of the quantitative trait and genotypes at one or more marker loci whose chromosomal locations are known *a priori*. Various statistical approaches have been developed for this purpose. This paper provides a brief overview of some of these approaches and presents relevant extensions to more complex, but more realistic scenarios. The performance of the proposed methods has been examined using computer simulations and it has been shown that the proposed approaches perform efficiently under a wide variety of scenarios.**

MANY human quantitative traits such as blood pressure, are known to be determined primarily, though not exclusively, by inherited genetic factors. Early biometrical geneticists considered that such traits were determined by 'blending inheritance'. However Fisher[1] showed that the inheritance of such characters can be modelled under the Mendelian paradigm of discrete genes. Thoday[2] first clearly laid the conceptual foundation of locating on chromosomes, the gene(s) that may be involved in the determination of quantitative traits. Considerable developments have taken place in mapping quantitative trait loci (QTL mapping). All methods rely on identifying cosegregation of alleles or associations between allelic differences at a genetic marker locus (whose chromosomal location is known *a priori*) and of alleles at a putative trait locus or differences among individuals in phenotype. The methods vary in statistical approach, for example, likelihood vs variance components, or in data requirement, for example, observations on pairs of relatives or on all members of nuclear or extended families. A key parameter used in linkage analysis[3] – the class of statistical techniques used for localization of genes on chromosomes – is the recombination fraction between a putative QTL and the marker locus. The recombination fraction is a function of the physical distance (measured in number of base pairs on DNA) separating the putative QTL and the marker locus. A reliable estimate of the recombination fraction provides the basic evidence needed to map a QTL. Although statistical methodologies for mapping genes determining dichotomous qualitative characters in humans are well-developed, the development of such methodologies, especially those that are statistically and computationally efficient, for human quantitative traits is an active area of current research in human genetics. It has been emphasized that many traits that have traditionally been treated as qualitative are inherently quantitative in nature. Often for ease of analysis, quantitative data are dichotomized using a threshold. This results in loss of information and is statistically undesirable. An example is hypertension/normotension which is a dichotomized trait on underlying quantitative traits (systolic and diastolic blood pressures).

Experimental studies on quantitative characters in plants[4], dairy cattle[5–7], etc. have revealed that quantitative traits may often be determined by multiple loci. There is also increasing evidence[8–11] that alleles at a locus determining a quantitative trait may exhibit dominance over other alleles within that locus and may interact epistatically with alleles at other loci controlling that trait. Although, a quantitative trait may be determined by multiple loci, often the effects of the loci on the trait are highly variable, so that it may suffice to consider only those loci which have large effects, which are generally few in number, and are, therefore, more easily mapped than those loci with small effects. Such factors (dominance, epistasis, etc.) are often ignored[12] to simplify statistical analysis. However, in view of the experimental observations cited above, it is necessary to consider multiple loci and the possibility of epistatic interactions among the loci. In this paper, we propose two computationally simple statistical techniques, by extending some traditional techniques, for mapping QTLs when the trait is actually determined by a set of unlinked, autosomal, epistatically interacting loci. The two methodologies pertain to two different types of data; nuclear families and sib-pairs. We also elaborately examine the performance of these modified and extended methodologies from various statistical considerations which results in insights on the performances of these methodologies under various scenarios.

*For correspondence. (e-mail: ppm@isical.ac.in)

We first consider parental and offspring data separately on families in which only one parent is heterozygous at the marker locus and those in which both parents are heterozygous and suitably modify the estimator proposed by Jayakar[13] based on variance components. We show, based primarily on the widths of confidence intervals, that for a wide range of parameter values the proposed estimator is quite efficient. Additionally, we suggest a non-parametric procedure for testing null hypotheses regarding $\theta$ and show that the power function of the test has desirable statistical properties. We also show that analyses of data ignoring epistatic interactions, when in fact these are present, may lead to grossly inaccurate inferences about linkage. However, the variance of the proposed estimator is found to be larger than that of the maximum likelihood estimator (m.l.e.). Our results provide statistical insights on the major reasons why Jayakar's[13] estimators may not perform well in practice.

Our second data type includes quantitative trait values of sib-pairs and their estimated identity-by-descent (i.b.d.) scores at the marker locus. A pair of related individuals shares an allele i.b.d. if that allele has a common ancestral source. For sib-pairs, the common ancestors are their parents. One of the popular statistical techniques to analyse such data is based on the regression of squared difference in trait values of sib-pairs on their estimated marker i.b.d. scores. Under a very general set-up, even in the presence of dominance and epistatic effects, Tiwari and Elston[14] have extended the classical regression method for QTL mapping when the trait is controlled by two unlinked, autosomal, biallelic loci. Since this general model involves too many parameters, insights into effects of variation of individual parameters on the performance of the method were difficult to obtain. We, therefore, examine the performance of the method under the specific digenic interaction model[15]. We also extend the method to the case of a quantitative trait that is controlled by multiple unlinked loci. The competing strategies of analysing the data by simultaneous, as opposed to sequential, consideration of the markers are quantitatively assessed using simulation studies. As is intuitively expected, the simultaneous strategy is found to be more optimal and cost-effective.

## Nuclear family data

Based on observations on members of nuclear families, i.e. observations on parents and their offspring, Jayakar[13] derived an estimator of $\theta$, the recombination fraction between the putative trait locus (a single locus is assumed to determine the quantitative trait) and a marker locus, as a function of the variances of the quantitative trait in the population and among offspring of specified marker genotypes within and across various parental mating types.

Jayakar assumed that a quantitative character $Y$ is controlled by an autosomal biallelic locus with alleles $A_1$ and $a_1$. The allele frequencies are $p_1$ and $q_1 = 1 - p_1$, respectively and the population is in Hardy–Weinberg equilibrium. Suppose the probability density function of $Y$ for the trait genotypes $A_1A_1$, $A_1a_1$ and $a_1a_1$ are $f_1$, $f_2$ and $f_3$, respectively, where $f_i$s are independent of $\theta$. Let the expectation of $Y$ given the trait genotypes be $\alpha$, $\beta$ and $-\alpha$, respectively and the variance of $Y$ given any trait genotype be $\sigma^2$, which includes the environmental variance. It is assumed that the trait locus is in linkage equilibrium with an autosomal, biallelic, codominant marker locus with alleles $M_1$ and $m_1$.

To ensure informativeness for linkage, it is necessary to only consider matings for which at least one parent is heterozygous at the marker locus. Jayakar distinguished, at the marker locus, the two types of families, backcross ($M_1M_1 \times M_1m_1$) and intercross ($M_1m_1 \times M_1m_1$). It is obvious that $m_1m_1 \times M_1m_1$ families can be handled in the same manner as $M_1M_1 \times M_1m_1$ families by relabelling alleles. Families in which neither parent is heterozygous at the marker locus are excluded from analyses.

The probability distribution of offspring genotypes at the $(A_1, a_1)$ locus for various backcross parental genotypic matings is provided in Table 1.

For any particular parental genotypic mating $g$ (= 1, 2, . . ., 11), let $\pi_{gi}$ and $Y_{gi}$ denote the probability and value of the quantitative trait, respectively, for an offspring of type $i$; $i = 1 = A_1A_1M_1M_1$, $i = 2 = A_1a_1M_1M_1$, $i = 3 =$

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

| | | | $\pi_g$ | | |
|---|---|---|---|---|---|
| $g$ | 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 | $A_1A_1 \times a_1a_1$<br>$a_1a_1 \times A_1A_1$ | $2p_1^2p_2^2$ | 0 | $\frac{1}{2}$ | 0 |
| 5 | $A_1a_1 \times A_1A_1$<br>$a_1A_1 \times A_1A_1$ | $2p_1^3p_2$ | $\frac{1}{4}$ | $\frac{1}{4}$ | 0 |
| 6 | $A_1a_1 \times A_1a_1$<br>$a_1A_1 \times A_1a_1$ | $2p_1^2p_2^2$ | $\frac{1}{4}(1-\theta)$ | $\frac{1}{4}$ | $\frac{1}{4}\theta$ |
| 7 | $A_1a_1 \times a_1A_1$<br>$a_1A_1 \times a_1A_1$ | $2p_1^2p_2^2$ | $\frac{1}{4}\theta$ | $\frac{1}{4}$ | $\frac{1}{4}(1-\theta)$ |
| 8 | $A_1a_1 \times a_1a_1$<br>$a_1A_1 \times a_1a_1$ | $2p_1p_2^3$ | 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 $M_1m_1$ can be obtained by replacing $\theta$ by $(1-\theta)$ in this table.

$a_1a_1M_1M_1$, $i = 4 = A_1A_1M_1m_1$, $i = 5 = A_1a_1M_1m_1$ and $i = 6 = a_1a_1M_1m_1$.

Let,

$T_g$ = Var $\{Y_{g1}, Y_{g2}, \ldots, Y_{g6}\}$ = Variance of trait values among all offspring;

$V_{g1}$ = Var $\{Y_{g1}, Y_{g2}, Y_{g3}\}$ = Variance of trait values among offspring of marker genotype $M_1M_1$;

$V_{g2}$ = Var $\{Y_{g4}, Y_{g5}, Y_{g6}\}$ = Variance of trait value among offspring of marker genotype $M_1m_1$;

$V_g = V_{g1} + V_{g2}$; and

$V_p$ = Variance of the trait value $Y$ in the whole population.

If $T = E_g(T_g)$ and $V = E_g(V_g)$, then Jayakar showed that:

$$\theta = \frac{1}{2}\left[1 - \sqrt{\frac{4T - 2V}{V_p - \sigma^2}}\right],$$

the estimator of θ being obtained by plugging in observed values of $T$, $V$, $V_p$ and $\sigma^2$.

When the family is an intercross ($M_1m_1 \times M_1m_1$), the probabilities of different offspring types for various parental genotypic matings are given in Table 2. As in the case of backcross, let, for any particular genotypic mating $g$ (= 1, 2, . . ., 10), $\pi_{gi}$ and $Y_{gi}$ denote the probability and quantitative trait value, respectively, for an offspring of type $i$; $i = 1 = A_1A_1M_1M_1$, $i = 2 = A_1a_1M_1M_1$, . . ., $i = 9 = a_1a_1m_1m_1$.

Let,

$V_{g1}$ = Var $\{Y_{g1}, Y_{g2}, Y_{g3}\}$

$V_{g2}$ = Var $\{Y_{g4}, Y_{g5}, Y_{g6}\}$

$V_{g3}$ = Var $\{Y_{g7}, Y_{g8}, Y_{g9}\}$

If $V_1 = E_g(V_{g1})$, $V_2 = E_g(V_{g2})$ and $V_3 = E_g(V_{g3})$, then Jayakar showed that:

$$\theta = \frac{1}{2}\left[1 - \sqrt{\frac{2V_2 - (V_1 + V_3)}{(V_p - \sigma^2)}}\right].$$

In this section, we modify Jayakar's estimator of θ to the case of a quantitative trait being determined by multiple, unlinked, epistatically interacting loci. We also examine the statistical properties of the modified estimator.

We assume that a quantitative trait $Y$ is controlled by $L$ autosomal biallelic loci. Let $A_l$ and $a_l$ denote the alleles at the $l$th locus, $l = 1, 2, \ldots, L$. We assume that the loci are mutually unlinked and that the population is in Hardy–Weinberg equilibrium in respect of each of these loci. Let the allele frequencies at the $l$th locus be denoted as $p_l$ and $q_l = 1 - p_l$. Let the expectation, $E(Y)$, of the quantitative character, $Y$, given the genotypes of the $l$th locus be $\alpha_l$, 0

**Table 2.** Trait locus mating types among $M_1m_1 \times M_1m_1$ (intercross) parents, mating probabilities and probabilities of trait locus genotypes among offspring with marker genotypes $M_1M_1$ 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 | $A_1A_1 \times A_1a_1$<br>$A_1a_1 \times A_1A_1$ | $2p_1^3p_2$ | $\frac{1}{4}(1-\theta)$ | $\frac{1}{4}\theta$ | 0 | $\frac{1}{4}$ | $\frac{1}{4}$ | 0 |
| 3 | $A_1A_1 \times a_1A_1$<br>$a_1A_1 \times A_1A_1$ | $2p_1^3p_2$ | $\frac{1}{4}\theta$ | $\frac{1}{4}(1-\theta)$ | 0 | $\frac{1}{4}$ | $\frac{1}{4}$ | 0 |
| 4 | $A_1A_1 \times a_1a_1$<br>$a_1a_1 \times A_1A_1$ | $2p_1^2p_2^2$ | 0 | $\frac{1}{4}$ | 0 | 0 | $\frac{1}{2}$ | 0 |
| 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 | $A_1a_1 \times a_1A_1$<br>$a_1A_1 \times A_1a_1$ | $2p_1^2p_2^2$ | $\frac{1}{4}\theta(1-\theta)$ | $\frac{1}{4}[1-2\theta(1-\theta)]$ | $\frac{1}{4}\theta(1-\theta)$ | $\frac{1}{4}[1-2\theta(1-\theta)]$ | $\theta(1-\theta)$ | $\frac{1}{4}[1-2\theta(1-\theta)]$ |
| 7 | $a_1a_1 \times A_1a_1$<br>$A_1a_1 \times a_1a_1$ | $2p_1p_2^3$ | 0 | $\frac{1}{4}(1-\theta)$ | $\frac{1}{4}\theta$ | 0 | $\frac{1}{4}$ | $\frac{1}{4}$ |
| 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 | $a_1a_1 \times a_1A_1$<br>$a_1A_1 \times a_1a_1$ | $2p_1p_2^3$ | 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.

and $-\alpha_l$ for $A_lA_l$, $A_la_l$ and $a_la_l$, respectively, i.e. there is no dominance at any of the QTLs. We assume that the variance of $Y$ within each single-locus genotype is the same. For the $l$th locus, let this variance be denoted as $\sigma_l^2$. In the absence of epistatic interactions, effects of the loci on $E(X)$ are assumed to be additive. Thus for example, $E(Y \mid A_1A_1A_2A_2 \ldots A_LA_L) = \Sigma_{l=1}^{L}\alpha_i$; $E(Y|A_1A_1A_2A_2 \ldots A_{L-1}A_{L-1}a_La_L) = \Sigma_{i=1}^{L-1}\alpha_i - \alpha_L$, etc. when epistatic interactions are absent. Since the trait loci are assumed to be unlinked, the variance of the trait among individuals of any multilocus genotype is the same, $\sigma^2 = \sigma_G^2 + \sigma_E^2$, where $\sigma_G^2 = \Sigma_{l=1}^{L} = \sigma_l^2$ and $\sigma_E^2$ = environmental variance. We use a simple model of additive epistatic interaction at the different trait loci. This model is prompted by experimental observations on some plants and animals[16] and has been termed as the digenic interaction model[15] when $L = 2$. We assume that there are epistatic interactions only among homozygotes between pairs of loci. Between loci $i$ and $j$ ($i \neq j = 1, 2, \ldots, L$), epistatic interaction effects are assumed to be: for $A_iA_iA_jA_j$ and $a_ia_ia_ja_j$ the effect is $\Delta_{ij}$, for $A_iA_ia_ja_j$ and $a_ia_iA_jA_j$ the effect is $-\Delta_{ij}$; the effect for all other two-locus genotypes is 0. Thus, for example, under additive epistatic effects,

$$E(Y \mid A_1A_1A_2A_2 \ldots A_LA_L) = \sum_{i=1}^{L}\alpha_i + \sum_{i=1}^{L}\sum_{j>1}^{L}\Delta_{ij},$$

$$E(Y \mid A_1A_1A_2A_2 \ldots A_{L-1}A_{L-1}a_La_L) =$$

$$\sum_{i=1}^{L-1}\alpha_i - \alpha_L + \sum_{i=1}^{L-1}\sum_{j>1}^{L-1}\Delta_{ij} - \sum_{i=1}^{L-1}\Delta_{iL}, \text{ etc.}$$

For clarity, and to fix ideas, we provide, in Table 3, the genotypes ($G$), their population frequencies and expectation of the quantitative trait given genotype [$E(Y \mid G)$], for $L = 2$.

**Table 3.** Genotypes ($G$) of individuals at two autosomal, unlinked, epistatically interacting biallelic loci, relative frequencies of these genotypes in the population and expected values of quantitative trait $Y$ corresponding to these genotypes under the digenic interaction model

| $G$ | Relative frequency | $E(Y\|G)$ |
|---|---|---|
| $A_1A_1A_2A_2$ | $p_1^2p_2^2$ | $\alpha_1 + \alpha_2 + \Delta_{12}$ |
| $A_1A_1A_2a_2$ | $2p_1^2p_2q_2$ | $\alpha_1$ |
| $A_1A_1a_2a_2$ | $p_1^2q_2^2$ | $\alpha_1 - \alpha_2 - \Delta_{12}$ |
| $A_1a_1A_2A_2$ | $2p_1p_2^2q_1$ | $\alpha_2$ |
| $A_1a_1A_2a_2$ | $4p_1p_2q_1q_2$ | 0 |
| $A_1a_1a_2a_2$ | $2p_1q_1q_2^2$ | $-\alpha_2$ |
| $a_1a_1A_2A_2$ | $p_2^2q_1^2$ | $-\alpha_1 + \alpha_2 - \Delta_{12}$ |
| $a_1a_1A_2a_2$ | $2p_2q_1^2q_2$ | $-\alpha_1$ |
| $a_1a_1a_2a_2$ | $q_1^2q_2^2$ | $-\alpha_1 - \alpha_2 + \Delta_{12}$ |

We assume, without loss of generality, that the trait locus ($A_1$, $a_1$) is linked to an autosomal biallelic codominant marker locus with alleles $M_1$ and $m_1$. These two loci are assumed to be in linkage equilibrium. Let the recombination fraction between the loci be denoted as $\theta$. Our purpose is to estimate $\theta$ from observations on the quantitative trait and the genotypes at the marker locus on members of families.

Since only the ($A_1$, $a_1$) locus is linked to ($M_1$, $m_1$), information on $\theta$ is contained only in two locus gametotypes obtainable upon joint consideration of these two loci only.

For a backcross family, we define $T$, $V$ and $V_p$ as in Jayakar's derivation. For the model considered, we can show that:

$$(1 - 2\theta)^2 = \frac{2T - V}{V_p - T}$$

$$\Rightarrow \theta = \frac{1}{2}\left[1 - \sqrt{\frac{2T - V}{V_p - T}}\right]. \quad (1)$$

The estimate of $\theta$ is obtained from eq. (1) by plugging in observed values of $T$, $V$ and $V_p$. We note that although eq. (1) is independent of the parameters underlying the model governing the trait and marker loci (i.e. $\alpha$s, $\Delta$s, $\sigma^2$ and allele frequencies), the sampling distribution of the proposed estimator of $\theta$ is a function of these parameters.

Similarly for an intercross family, we define $V_1$, $V_2$, $V_3$ and $V_p$ as before.

We can show that:

$$(1 - 2\theta)^2 = \frac{2V_2 - (V_1 + V_3)}{2(V_p - V_2)}$$

$$\Rightarrow \theta = \frac{1}{2}\left[1 - \sqrt{\frac{2V_2 - (V_1 + V_3)}{2(V_p - V_2)}}\right]. \quad (2)$$

Before proceeding further, we wish to note that these computationally simple estimators are analogous, but not identical, to the estimators obtained by Jayakar[13]. The equations corresponding to eqs (1) and (2) derived by Jayakar[13] for a single quantitative trait locus, has in the denominator the term $V_p - \sigma_1^2$. These equations fail to hold when $\sigma_1^2$ is replaced by $\Sigma_{l=1}^{L}\sigma_l^2$, if there are $L$ trait loci, even in the absence of any interactions.

### *Test procedure and evaluation of power*

Having estimated $\theta$, one is obviously interested in testing the null hypothesis $\theta = 0.5$. We suggest a non-parametric test procedure that is analogous to the permutation test.

For the observed values of the marginal totals of offspring, one can generate simulated data under the null hypothesis $\theta = 0.5$. Based on the simulated data on a number of offspring (NOFF), one obtains an estimate of $\theta$ by eqs (1) or (2). When the simulation is replicated a large number of times (NREP), an empirical probability distribution of $\theta$ can be obtained and empirical cut-off point(s), for a predetermined level of significance, determined. An inspection of whether the observed value of $\theta$ is outside the interval determined by the empirical cut-off point(s) provides the decision on rejection of the null hypothesis.

For obtaining the power at $\theta = \theta_1$, the simulation is carried out with $\theta = \theta_1$ and at each replication, a check is made whether the estimated value of $\theta$ lies outside of the interval defined by the empirical cut-off points determined earlier (using $\theta = \theta_0 = 0.5$, say). If the number of replications is $n_1$, then the power at $\theta_1$ is $a/n_1$, where $a$ is the number of replications for which $\theta = \theta_0$ is rejected. For every set of parameter values, these evaluations are performed with NOFF = 1000 and NREP = 10000. We emphasize that NOFF is the total number of offspring in the pooled data set of a particular mating type (backcross or intercross). If each family comprises 4 offspring, we are in effect dealing with 250 families.

### *Mean and variance of* $\hat{\theta}$

To examine the behaviour $\hat{\theta}$ of the estimators in respect of variations of values of the underlying parameters ($p_1$, $p_2$, $\alpha_1$, $\alpha_2$, $\Delta_{12}$, $\sigma^2$), we perform simulations for different sets of values of the parameters and evaluate the means and variances of $\hat{\theta}$. These results are given in Table 4. It is seen from this table that the mean values of $\hat{\theta}$ for both backcross and intercross matings deviate more from the true value of $\theta$ and the variances of $\hat{\theta}$ increase with increase in the value of the interaction parameter $\Delta_{12}$. Similar deviations in Mean($\hat{\theta}$) and similar increases in Var($\hat{\theta}$) are observed when (a) the variance $\sigma^2$ of the quantitative trait increases; (b) the expected value of the quantitative trait given the genotype at the second trait locus, $\alpha_2$, increases; and (c) when $p_1$ deviates from 0.5. Variation in $p_2$ has virtually no effect on Mean($\hat{\theta}$) or Var($\hat{\theta}$). Although for brevity, results for only a selected number of sets of parameter values are provided in Table 4, we have verified the above facts for a large number of sets of parameter values.

### *Power function*

The power functions for different values of $\theta$, separately for backcross and intercross cases, are depicted in Figures 1 and 2, respectively, for a single set of parameter values $p_1 = 0.5$, $p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$ and $\sigma^2 = 1$. It is seen from Figures 1 and 2 that the power functions are very well-behaved for values of $\theta$ in the range $0 < \theta < 0.5$. For $\theta = 0$ or 0.5, the powers are rather high even for values of $\theta$ quite close to that specified under the null hypothesis. However, from a practical viewpoint, this undesirable fact for the two extreme values of $\theta$ may not

**Table 4.** Means and variances of estimated values of recombination fraction, $\theta$, each based on 10,000 replications of data simulated at given sets of values of underlying parameters for backcross and intercross families

| | | | | | | | Backcross | | Intercross | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $p_1$ | $p_2$ | $\alpha_1$ | $\alpha_2$ | $\Delta_{12}$ | $\sigma^2$ | Mean($\hat{\theta}$) | Var($\hat{\theta}$) | Mean($\hat{\theta}$) | Var($\hat{\theta}$) |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0121 | 0.00027 | 0.0180 | 0.00049 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 2.00 | 1.00 | 0.0122 | 0.00028 | 0.0186 | 0.00058 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 3.00 | 1.00 | 0.0128 | 0.00032 | 0.0214 | 0.00080 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 4.00 | 1.00 | 0.0132 | 0.00037 | 0.0251 | 0.00114 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 5.00 | 1.00 | 0.0136 | 0.00040 | 0.0301 | 0.00169 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0121 | 0.00027 | 0.0180 | 0.00049 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 5.00 | 0.0139 | 0.00040 | 0.0236 | 0.00097 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 10.00 | 0.0158 | 0.00055 | 0.0326 | 0.00196 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0121 | 0.00027 | 0.0180 | 0.00049 |
| 0.00 | 0.50 | 0.50 | 5.00 | 5.00 | 1.00 | 1.00 | 0.0165 | 0.00062 | 0.0332 | 0.00213 |
| 0.00 | 0.50 | 0.50 | 5.00 | 10.00 | 1.00 | 1.00 | 0.0214 | 0.00131 | 0.0811 | 0.01463 |
| 0.00 | 0.50 | 0.10 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0121 | 0.00027 | 0.0175 | 0.00050 |
| 0.00 | 0.50 | 0.20 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0123 | 0.00029 | 0.0181 | 0.00051 |
| 0.00 | 0.50 | 0.30 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0122 | 0.00028 | 0.0183 | 0.00053 |
| 0.00 | 0.50 | 0.40 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0117 | 0.00026 | 0.0179 | 0.00050 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0121 | 0.00027 | 0.0180 | 0.00049 |
| 0.00 | 0.10 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0178 | 0.00053 | 0.0300 | 0.00119 |
| 0.00 | 0.20 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0135 | 0.00033 | 0.0215 | 0.00067 |
| 0.00 | 0.30 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0124 | 0.00029 | 0.0184 | 0.00051 |
| 0.00 | 0.40 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0120 | 0.00026 | 0.0182 | 0.00049 |
| 0.00 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0121 | 0.00027 | 0.0180 | 0.00049 |
| 0.10 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.0983 | 0.00081 | 0.1028 | 0.00166 |
| 0.30 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.2888 | 0.00084 | 0.2982 | 0.00582 |
| 0.50 | 0.50 | 0.50 | 5.00 | 1.00 | 1.00 | 1.00 | 0.4228 | 0.00072 | 0.4135 | 0.00750 |

imply a serious limitation of the test procedure. As we have already noted in an earlier subsection, for both backcross and intercross matings, at these two extreme values of $\theta$, in the vast majority of replications the estimated $\hat{\theta}$ is quite close to the true $\theta$. We have evaluated the power functions for many other sets of parameter values; the results are not provided for brevity, since the general feature described above is true for the other sets of values also.

### *Effect of ignoring epistatic interactions*

We investigate the effect of ignoring epistatic interactions when in fact these are present, using the following strategy. For a set of parameter values $\theta$, $p_1$, $p_2$, $\alpha_1$, $\alpha_2$, $\sigma^2$ and $\Delta_{12} = 0$, we first obtain, based on 10,000 simulation replications, the 95% confidence interval of $\theta$ using the procedure outlined earlier. Then, for the same fixed values of $\theta$, $p_1$, $p_2$, $\alpha_1$, $\alpha_2$ and $\sigma^2$, but with $\Delta_{12} = \Delta_0 \neq 0$, we generate 1000 simulated data sets. For each such simulated data set, we estimate $\theta$ using eq. (1) or eq. (2), as appropriate and check whether $\hat{\theta}$ is included in the confidence interval obtained earlier (with $\Delta_{12} = 0$). Inclusion of $\hat{\theta}$ in the confidence interval implies that the estimate of $\theta$ is not significantly adversely affected in spite of ignoring the effect of epistatic interaction, when in fact it is present.

For several sets of parameter values, we find, using this procedure, that for most sets of parameter values, the percentage of inclusion of $\hat{\theta}$ in the appropriate confidence interval varies from about 40% to about 60%. For example, for backcross families with $p_1 = 0.5$, $p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$ and $\sigma^2 = 1$, this value is 47.3%. Thus, there is a strong adverse effect of ignoring epistatic interactions for estimating $\theta$, when in fact such interactions are present.

### *Comparison with the maximum likelihood estimator*

Since the character $Y$ is controlled by $L$ trait loci, the number of possible trait genotypes of an individual is $3^L$. Let the probability density function of $Y$ given these trait genotypes be $f_1, f_2, \ldots, f_{3^L}$, respectively. Then the likelihood of the offspring data given the parental data is:

$$L(\theta) = \prod_{j=1}^{n} f(Y_j)\pi_j,$$

where $n$ is the number of observations, $f$ takes values $f_1, f_2, \ldots f_{3^L}$ and $\pi_j$ is the probability of the quantitative trait.

We note that, since the only trait locus linked to the marker is $(A_1, a_1)$, $\pi_j$ can be interpreted as the probability of the quantitative trait with respect to this locus only. As mentioned earlier, $f(Y_j)$ is independent of $\theta$ and thus the m.l.e. of $\theta$ turns out to be a simple function of the number of observations in those genotypic classes for which $\pi_{gi}$ is not independent of $\theta$.

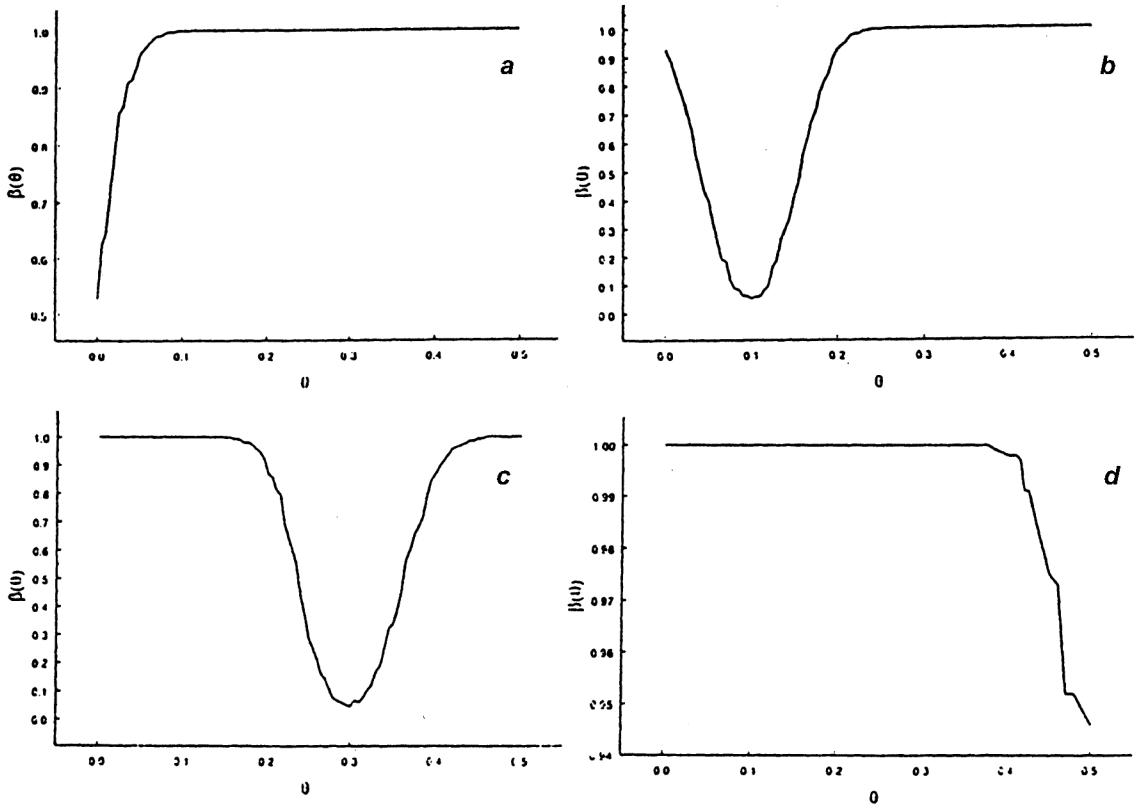


**Figure 1.** Power functions of the proposed test procedure for backcross families (with NOFF = 1000) at simulation parameter values $p_1 = p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$, $\sigma^2 = 1$ and (*a*) $\theta = 0$, (*b*) $\theta = 0.1$, (*c*) $\theta = 0.3$ and (*d*) $\theta = 0.5$.

We compare the efficiency of the proposed variance method in a two trait loci set-up with that of the maximum likelihood approach through simulation studies. We performed simulations for parameter values of $\alpha_1 = 5$; $\alpha_2 = 1$; $\sigma^2 = 1$; $\Delta_{12} = 1$; $p_1 = 0.5, 0.3, 0.1$; $p_2 = 0.5$ and $\theta = 0, 0.1, 0.3, 0.5$ separately for backcross and intercross matings. The results are given in Table 5. It is seen that the m.l.e. of $\theta$ has more precision than the modified-Jayakar estimator in terms of variance of $\hat{\theta}$. At the boundary values of $\theta$, i.e. $\theta = 0$ and $\theta = 0.5$, the mean of the m.l.e. is also closer to the true value of $\theta$ than the modified-Jayakar estimator. We note that the m.l.e. of $\theta$ is independent of the trait values of individuals, while Jayakar's estimator is not. However, in spite of using the additional information on trait values, the relative efficiency of the modified-Jayakar estimator is much lower than the m.l.e.

### *Sib-pair data*

Haseman and Elston[17] proposed a regression-based QTL mapping method using sib-pair data. Their mapping procedure was based on squared difference in quantitative trait values of sib-pairs ($Y$) and their estimated marker i.b.d. scores ($\hat{\pi}_m$). They obtained the regression equation:

$$E(Y \mid \hat{\pi}_m) = \alpha + \beta \hat{\pi}_m,$$

where there is no dominance in the trait and $\beta$ is a one-to-one function of the recombination fraction $\theta$, between the QTL and the marker locus. The test for no linkage (i.e. $\theta = 0.5$) is equivalent to testing $\beta = 0$.

We start with the simple digenic-interaction model described earlier and in Table 3, where the QT is determined by two unlinked, autosomal, epistatically interacting biallelic loci, and then extend it further to multiple QT loci. We note that for two loci, our model is a special case of the more general model considered by Tiwari and Elston[14]. They assumed that the QT is controlled by two autosomal, unlinked, epistatically interacting loci with dominance present as each locus. In their model, the epistatic interaction between loci which are both homozygous is identical to that in the digenic-interaction model. Moreover, they assumed epistatic interaction to be present between other pairs of loci as well. They showed that:

$$E(Y \mid \hat{\pi}_{m_1}, \hat{\pi}_{m_2}, f_1, f_2) = \alpha + \beta_1 \hat{\pi}_{m_1} + \beta_2 \hat{\pi}_{m_2}$$
$$+ \text{terms involving } f_1 \text{ and } f_2 + \text{cross-product terms of}$$
$$\hat{\pi}_{m_1}, \hat{\pi}_{m_2}, f_1, f_2,$$

where $\hat{\pi}_{m_1}$ and $\hat{\pi}_{m_2}$ are the estimated i.b.d. scores at two marker loci which are assumed to be in linkage equilibrium with the two QTLs, respectively, and $f_j = P(\hat{\pi}_{m_j} = \frac{1}{2} \mid \hat{\pi}_{m_j})$; $j = 1, 2$. However, as mentioned earlier, our model of epistasis is prompted by experimental

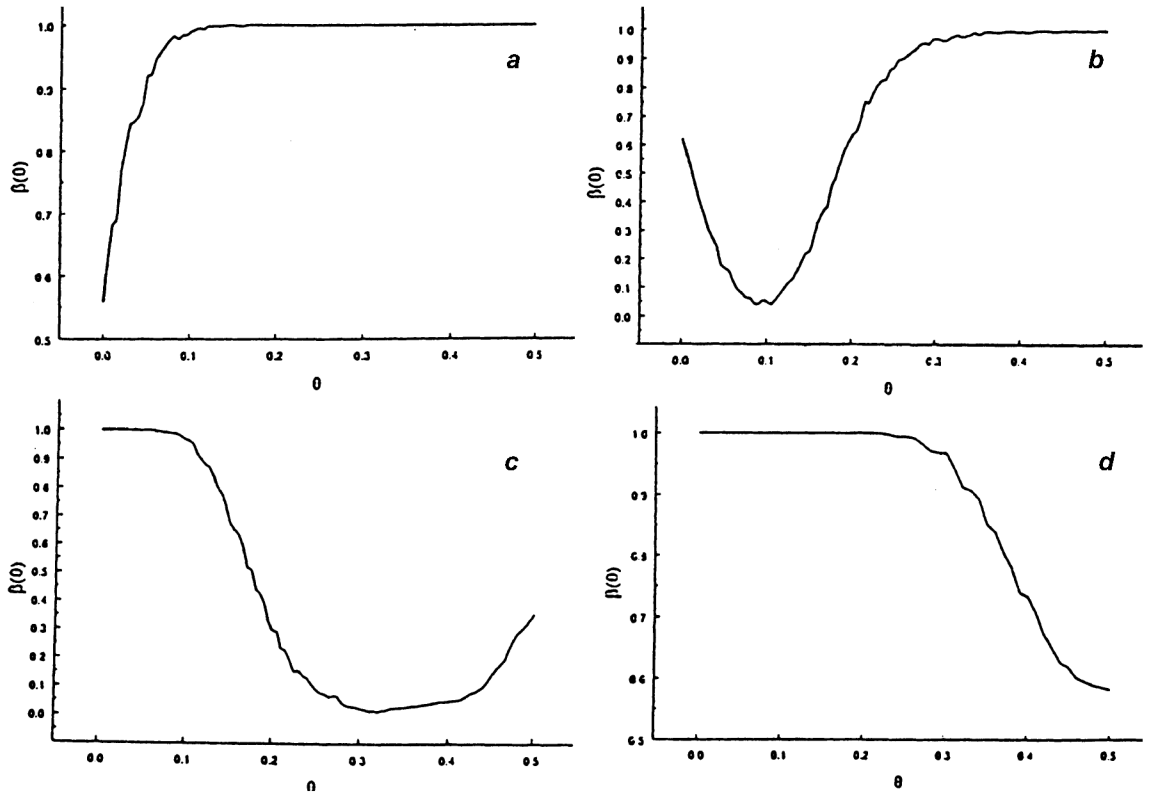


**Figure 2.** Power functions of the proposed test procedure for intercross families (with NOFF = 1000) at simulation parameter values $p_1 = p_2 = 0.5$, $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta_{12} = 1$, $\sigma^2 = 1$ and (*a*) $\theta = 0$, (*b*) $\theta = 0.1$, (*c*) $\theta = 0.3$ and (*d*) $\theta = 0.5$.

observations. The small number of parameters in our model enables clearer evaluation of the marginal effects of different trait and linkage parameters on the sample size requirement to detect linkage.

In our derivations, we have suppressed the suffix $_{12}$ and have denoted $\Delta_{12}$ as $\Delta$. We assume that the trait locus $(A_l, a_l)$ is linked to an autosomal, biallelic codominant marker locus with alleles $M_l$ and $m_l$; $l = 1, 2$. The loci $(A_l, a_l)$ and $(M_l, m_l)$ are assumed to be in linkage equilibrium. Our aim is to make inferences on $\theta_l$, the recombination fraction between $(A_l, a_l)$ and $(M_l, m_l)$; $l = 1, 2$, based on data on the quantitative trait values of sib-pairs. Suppose $(y_{j1}, y_{j2})$: $j = 1, 2, \ldots, n$ are the observed values of the quantitative trait of $n$ independent sib-pairs. We assume that $(y_{j1}, y_{j2})$s are distributed with an identical covariance structure given by

$$\sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Let $\pi_{j1}$ and $\pi_{j2}$ be the proportions of alleles shared i.b.d. at the loci $(A_1, a_1)$ and $(A_2, a_2)$, respectively, for the $j$th sib-pair. These proportions can assume values $0, \frac{1}{2}, 1$. The conditional probabilities of genotypes of sib-pairs with respect to the locus $(A_l, a_l)$ given $\pi_{jl}$ are provided in table 1 in Haseman and Elston[17]. As the loci $(A_1, a_1)$ and $(A_2, a_2)$ are unlinked, the joint conditional probability of the trait locus genotypes given the trait i.b.d. scores is the product of the marginal conditional probability of each trait locus genotype given the corresponding trait i.b.d. score. For example, $P(\text{Sib 1} = A_1A_1A_2A_2 \text{ and Sib 2} = A_1A_1A_2A_2 \mid \pi_{j1} = \frac{1}{2}, \pi_{j2} = 1) = p_1^3 p_2^2$.

Define $Y_j = (Y_{j1} - Y_{j2})^2$, $j = 1, 2, \ldots, n$; i.e. $Y_j$ denotes the squared pair difference in the trait values for the $j$th sib-pair. Note that $V(Y_{j1} - Y_{j2}) = 2\sigma^2(1 - \rho) = \phi^2$, $\forall j = 1, 2, \ldots, n$.

### *The regression equation*

Let $\pi_{jm_1}$ and $\pi_{jm_2}$ denote the proportions of alleles shared i.b.d. at the marker loci $(M_1, m_1)$ and $(M_2, m_2)$ respectively for the $j$th sib-pair. Let $f_{ji}^{(l)}$ denote the probability that the $j$th sib-pair has $i$ alleles shared i.b.d. at the marker locus $(M_l, m_l)$, $i = 0, 1, 2$; $l = 1, 2$. Then the estimator of $\pi_{jm_l}$ is given by $\hat{\pi}_{jm_l} = f_{j2}^{(l)} + \frac{1}{2} f_{j1}^{(l)}$; $l = 1, 2$. Haseman and Elston[17] have explicitly calculated $f_{ji}^{(l)}$ for different mating types and in the case of missing parental information, they have suggested an algorithm considering phenosets[18].

Suppose now we are interested in evaluating $E(Y_j \mid \hat{\pi}_{jm_1}, \hat{\pi}_{jm_2})$. Combining the different values of $\hat{\pi}_{jm_1}$ and $\hat{\pi}_{jm_2}$, we obtain the relation:

$$E(Y_j \mid \hat{\pi}_{jm_1}, \hat{\pi}_{jm_2}) = \beta_0 + \beta_1 \hat{\pi}_{jm_1} + \beta_2 \hat{\pi}_{jm_2},$$

where:

$$\begin{aligned}
\beta_0 &= \phi^2 + 4p_1q_1\{\alpha_1^2 + \Delta^2(p_2^2 + q_2^2) + 2\alpha_1\Delta(p_2 - q_2)\} \times \\
&\quad (1 - 2\theta_1 + 2\theta_1^2) + 4p_2q_2\{\alpha_2^2 + \Delta^2(p_1^2 + q_1^2) \\
&\quad + 2\alpha_2\Delta(p_1 - q_1)\}(1 - 2\theta_2 + 2\theta_2^2); \\
\beta_1 &= -4p_1q_1\{\alpha_1^2 + \Delta^2(p_2^2 + q_2^2) + 2\alpha_1\Delta(p_2 - q_2)\} \times \\
&\quad (1 - 2\theta_1)^2; \\
\beta_2 &= -4p_2q_2\{\alpha_2^2 + \Delta^2(p_1^2 + q_1^2) + 2\alpha_2\Delta(p_1 - q_1)\} \times \\
&\quad (1 - 2\theta_1)^2. \qquad (3)
\end{aligned}$$

This provides the motivation to set up the linear model:

$$Y_j = \beta_0 + \beta_1 \hat{\pi}_{jm_1} + \beta_2 \hat{\pi}_{jm_2} + e_j, \quad j = 1, 2, \ldots, n$$

where $e_j$s are i.i.d. $N(0, \tau^2)$.

**Table 5.** Comparison between means and variances of estimated values of recombination fraction, $\theta$, each based on 10,000 replications of data simulated at parameter values $\alpha_1 = 5$, $\sigma^2 = 1$ for backcross and intercross families in the two trait loci set-up using Jayakar's approach and maximum likelihood approach

| | | Backcross | | | | Intercross | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $p_1$ | $M(\hat{\theta}_J)$ | $V(\hat{\theta}_J)$ | $M(\hat{\theta}_M)$ | $V(\hat{\theta}_M)$ | $M(\hat{\theta}_J)$ | $V(\hat{\theta}_J)$ | $M(\hat{\theta}_M)$ | $V(\hat{\theta}_M)$ |
| 0 | 0.50 | 0.0121 | 0.00027 | 0.00003 | 0.00001 | 0.0180 | 0.00049 | 0.00003 | 0.00001 |
| | 0.30 | 0.0122 | 0.00028 | 0.0001 | 0.00002 | 0.0186 | 0.00058 | 0.0001 | 0.00002 |
| | 0.10 | 0.0128 | 0.00032 | 0.0001 | 0.00004 | 0.0214 | 0.00080 | 0.0001 | 0.00004 |
| 0.10 | 0.50 | 0.0983 | 0.00081 | 0.0997 | 0.00005 | 0.1028 | 0.00166 | 0.1248 | 0.00004 |
| | 0.30 | 0.1029 | 0.00105 | 0.0994 | 0.00012 | 0.1041 | 0.00210 | 0.1165 | 0.00007 |
| | 0.10 | 0.1045 | 0.00117 | 0.1013 | 0.00016 | 0.1056 | 0.00229 | 0.1187 | 0.00009 |
| 0.30 | 0.50 | 0.2888 | 0.00084 | 0.3001 | 0.00012 | 0.2984 | 0.00582 | 0.3312 | 0.00008 |
| | 0.30 | 0.3101 | 0.00118 | 0.3013 | 0.00023 | 0.2970 | 0.00639 | 0.3282 | 0.00011 |
| | 0.10 | 0.3138 | 0.00142 | 0.3067 | 0.00049 | 0.3026 | 0.00725 | 0.3304 | 0.00015 |
| 0.50 | 0.50 | 0.4228 | 0.00072 | 0.04953 | 0.00005 | 0.4135 | 0.00750 | 0.4964 | 0.00003 |
| | 0.30 | 0.4183 | 0.00154 | 0.04929 | 0.00006 | 0.4102 | 0.00848 | 0.4951 | 0.00005 |
| | 0.10 | 0.4117 | 0.00202 | 0.04908 | 0.00008 | 0.4036 | 0.00971 | 0.4926 | 0.00006 |

$\theta_J$ refers to Jayakar's estimator and $\hat{\theta}_M$ refers to maximum likelihood estimator.

We now note that for $l = 1, 2$; $\beta_l = 0 \Leftrightarrow \theta_l = 0.5$ and $\beta_l < 0 < \theta_l < 0.5$ as $\beta_l$ is an increasing 1-1 function of $\theta_l$. Thus, a test for linkage at the $l$th locus (i.e. $H_0$: $\theta_l = 0.5$ vs $H_1$: $\theta_l < 0.5$), is equivalent to testing for $H_0$: $\beta_l = 0$ vs $H_1$:$\beta_l < 0$ in the above linear model. The test statistic is given by

$$T_l = \frac{\hat{\beta}_l}{\widehat{s.e.}(\hat{\beta}_l)},$$

where $\hat{\beta}_l$ is the least squares estimator of $\beta_l$. In order to compute the standard error of $\hat{\beta}_l$, consider the design matrix $X$ given by:

$$\begin{pmatrix} \hat{\pi}_{1m_1} & \hat{\pi}_{1m_2} & 1 \\ . & . & . \\ . & . & . \\ . & . & . \\ \hat{\pi}_{nm_1} & \hat{\pi}_{nm_2} & 1 \end{pmatrix}.$$

Let $S = (X'X)$. Then

$$\widehat{s.e.}(\hat{\beta}_l) = \sqrt{\frac{s^{ll} R_0^2}{n-3}},$$

where $s^{ll} = (S^{-1})_{ll}$ and $R_0^2$ = residual sum of squares = $\Sigma_{j=1}^{n}(Y_j - \hat{\beta}_0 - \hat{\beta}_1\hat{\pi}_{jm_1} - \hat{\beta}_2\hat{\pi}_{jm_2})^2$. Under $H_0$, $T_l$ follows a $t$-distribution with $(n-3)$ degrees of freedom. Thus the critical region for a level $\alpha$ test is given by: $T_l < t_{n-3,1-\alpha}$.

If $n$ is sufficiently large, by the central limit theorem (CLT), we can approximate the critical region by: $T_l < z_{1-\alpha}$ where $z_p$ is the $(1-p)$th quantile of a standard normal variate. Using CLT the power function can be expressed as:

$$P(\beta_l) = \Phi\left\{ z_{1-\alpha} - \frac{\beta_l}{\widehat{s.e.}(\hat{\beta}_l)} \right\},$$

where $\Phi$ is the c.d.f. of $N(0, 1)$.

## *Determination of sample size required to detect linkage*

Having derived the power function of the proposed test, one is obviously interested in determining the minimum sample size required to detect linkage at the $l$th locus; $l = 1, 2$. In order for the test to have a power $\beta$ at $\beta_l$ (which is a 1 – 1 increasing function of $\theta_l$), we require the condition:

$$\Phi\left\{ z_{1-\alpha} - \frac{\beta_l}{\sqrt{\frac{s^{ll} R_0^2}{n_l^{-3}}}} \right\} = \beta$$

$$\Rightarrow z_{1-\alpha} - \frac{\beta_l}{\sqrt{\frac{s^{ll} R_0^2}{n_l^{-3}}}} = z_{1-\beta}$$

$$\Rightarrow n_l = \frac{(z_{1-\alpha} - z_{1-\beta})^2 s^{ll} R_0^2}{\beta_l^2} + 3. \quad (4)$$

Thus the required sample size to detect linkage at both the loci is given by $n = \max(n_1, n_2)$. In order to examine the effects of different trait parameters and linkage parameters on sample size requirement, we prove the following proposition.

*Proposition*: The sample size ($n_l$) required to detect linkage at the $l$th locus is: (i) an increasing function of $\theta_l$; (ii) an increasing function of $p_l$; (iii) a decreasing function of $p_i$, $i \neq l$; (iv) a decreasing function of $\alpha_l$; (v) a decreasing function of $\Delta$; (vi) independent of $\alpha_i$, $i \neq l$; (vii) independent of $\sigma^2$ and $\rho$.

*Proof*: Equation (3) implies that $\beta_l$ is an increasing function of $\theta_l$ and $p_l$ and a decreasing function of $\alpha_l$, $\Delta$ and $p_i$ ($i \neq l$). Now, $\beta_l < 0 < \beta_l^2$ is a decreasing function of $\beta_l$. Considering eq. (4), points (i)–(v) follow immediately.

Again, eqs (3) and (4) are both independent of $\alpha_i$ ($i \neq l$), $\sigma^2$ and $\rho$. Thus, points (vi) and (vii) are obviously true.

Hence, as intuitively expected, if the strength of linkage between a trait locus and a marker locus is higher, a smaller sample size suffices to detect linkage. Moreover, if a locus is controlled by several loci with comparable effects, then the sample size required for mapping the QTL with the highest level of heterozygosity is the smallest. Further, if among several QTLs, the marginal effect of one QTL increases, then smaller sample sizes are required to map that locus. Thus, among several QTLs, the QTLs with major effects are easiest to map. Moreover, if two QTLs have equal effects, then smaller sample sizes are required to map them if they epistatically interact, than if they do not. A similar result holds if there are multiple loci even with unequal effects.

## *Simultaneous detection vs. sequential detection as strategies to reduce sample size*

One interesting question that may arise in the determination of sample size to detect linkage at both the loci is whether it is more optimal to analyse the data by considering both the markers simultaneously (as illustrated in the previous subsection) or by considering them sequentially, one by one. In order to resolve this problem, let us first obtain expressions for $E(Y_j | \hat{\pi}_{jm_l})$, $l = 1, 2$. Using similar arguments as in the previous subsection, we can easily show that:

$$E(Y_j \mid \hat{\pi}_{jm_l}) = \beta_0 + \beta_l \hat{\pi}_{jm_l}, \quad l = 1, 2.$$

We find that $\beta_l$ is a $1-1$ increasing function of $\theta_l$ and $\beta_l = 0 \Leftrightarrow \theta_l = 0.5$, while $\beta_l < 0 < \theta_l < 0.5$. Thus, as discussed in the previous section, in order to detect linkage at the $l$th locus, we use a test statistic which follows a $t$-distribution with $(n-2)$ degrees of freedom under the null hypothesis of no linkage at the $l$th locus. The minimum sample size to detect linkage at the $l$th locus (i.e. to attain a power of $\beta$ at $\beta_l$) is given by:

$$N_l = \frac{(z_{1-\alpha} - z_{1-\beta})^2}{\beta_l^2} \frac{\sum_{j=1}^{n} (y_j - \hat{\beta}_0 - \hat{\beta}_l \hat{\pi}_{jm_l})^2}{(n-2)\sum_{j=1}^{n} (\hat{\pi}_{jm_l} - \overline{\hat{\pi}_{m_l}})^2} + 2.$$

Thus the minimum sample size to detect linkage at both the loci is given by $N = \max(N_1, N_2)$. Though $N$ (as given above) and $n$ (as determined in the previous section) cannot be compared analytically, we show through simulation studies that $n$ is, in general, smaller than $N$, i.e., we require a smaller sample size to detect linkage at both the loci if we analyse the data by considering both the markers simultaneously as opposed to considering them sequentially, one by one.

We note here that since the two QTLs are unidentifiable, in the sequential strategy, evidence of linkage would imply that the chosen marker locus is linked to one of the two trait loci. As the chosen markers are themselves unlinked, evidence of linkage for two different markers would indicate that we are able to map both the QTLs. Moreover the sequential analysis of markers is equivalent to analysing the data under the misspecified model of a single QTL. We thus observe that under this misspecified model, linkage of the chosen marker locus with the QTL can be correctly detected, but it requires a larger sample to map that QTL.

### *Extension of the regression procedure when the quantitative trait is controlled by more than two loci*

The regression procedure described above can be easily extended when the quantitative trait $Y$ is controlled by $k$ autosomal, biallelic, unlinked loci $(A_1, a_1)$, $(A_2, a_2)$, . . ., $(A_k, a_k)$. The generalized epistatic interaction model considered in the case of multiple loci has been described earlier.

Suppose the trait locus $(A_l, a_l)$ is in linkage equilibrium with an autosomal, biallelic, codominant marker locus $(M_l, m_l)$, $l = 1, 2, \ldots, k$, and the recombination fraction between these two loci is $\theta_l$. Our aim is to make inferences on $\theta_l$ based on observations on the quantitative trait of $n$ sib-pairs given by $\{(y_{j1}, y_{j2}): j = 1, 2, \ldots, n\}$.

Suppose $\pi_{jm_l}: \; l = 1, 2, \ldots, k$ denote the estimated proportions of alleles shared i.b.d. at the $l$th marker locus $\{(M_l, m_l): \; l = 1, 2, \ldots, k\}$. Defining $Y_j = (Y_{ji} - Y_{j2})^2$, we can show that:

$$E(Y_j \mid \hat{\pi}_{jm_l}, \ldots, \hat{\pi}_{jm_k}) = \beta_0 + \sum_{l=1}^{k} \beta_l \hat{\pi}_{jm_l},$$

where $\beta_l$ is a $1-1$ increasing function of $\theta_l$ and $\theta_l = 0.5 \Leftrightarrow \beta_l = 0$ and $\theta_l < 0.5 \Leftrightarrow \beta_l < 0$.

Thus, as in the two-loci set-up, we can test for linkage based on the linear model:

$$Y_j = \beta_0 + \sum_{l=1}^{k} \beta_l \hat{\pi}_{jm_l} + e_j, \quad j = 1, 2, \ldots, n,$$

where $e_j$s are i.i.d. $N(0, \tau^2)$.

### *Simulation results*

In order to assess the performance of our proposed regression strategy, we generate data on trait values of sib-pairs and estimated marker i.b.d. scores for different parameter values. Having generated the required data on 100 sib-pairs, we regress the squared difference in trait values on the different estimated marker i.b.d. scores. Based on the regression coefficients obtained, we evaluate the sample size requirements for detecting linkage for different values of recombination fractions. We perform the regression analysis both by considering the two markers simultaneously as well as sequentially, one by one. In each case we determine the sample size requirement (i.e. $n$ and $N$) and compare them by an 'efficiency' ratio $E = N/n$.

In our simulation examples, we assume the quantitative trait to be controlled by two autosomal loci and thus consider two marker loci which are in linkage equilibrium with the trait loci. Table 6 provides the results of the regression of squared difference in trait values on the two estimated marker i.b.d. scores and Tables 7 and 8 provide the sample sizes necessary to detect linkage at the two trait loci for simulation parameter values of $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta = 1$, $\sigma^2 = 1$ and different parameter values of $p_1$, $p_2$, $\rho$, $\theta_1$ and $\theta_2$. We perform the tests of linkage at 5% level of significance and determine the sample size requirements to attain a power of 0.9 for each test.

In both cases, we find that the regression procedure detects linkage quite efficiently and the sample size requirements are in accordance with our proposition stated in an earlier subsection, [i.e. $n_l$ increases with $p_l$ and decreases with $\alpha_l$, $\Delta$ and $p_i$ $(i \neq l)$]. The significance of the regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ (i.e. the extent of linkage at the two loci) depends not only on $\theta_1$ and $\theta_2$, but also on $\alpha_1$ and $\alpha_2$ (i.e. on the effect of each trait locus on the quantitative trait). When this effect is small, the corresponding regression coefficient tends to be less significant even if the trait locus is actually tightly linked to the marker locus. Similarly when the effect is large, the corresponding regression coefficient tends to be less insignificant even if the trait locus is actually unlinked to the marker locus.

**Table 6.** Regression and multiple correlation coefficients in the two types of regression analyses (simultaneous and sequential) for different sets of trait parameter values

| | | $\beta_0$ | | | $\beta_1$ | | | $\beta_2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Type of analysis | $R^2$ | Est | S.E. | $t$-val | Est | S.E. | $t$-val | Est | S.E. | $t$-val |
| Parameter values: $p_1 = 0.9$, $p_2 = 0.9$, $\rho = 0.8$, $\theta_1 = 0.1$ and $\theta_2 = 0.1$ | | | | | | | | | | |
| Simult. | 0.94 | 10.19 | 4.05 | 2.51 | – 14.67 | 7.63 | – 1.92* | – 4.04 | 2.28 | – 1.70* |
| Seq. using: | | | | | | | | | | |
| Marker 1 | 0.74 | 12.18 | 6.69 | 1.82 | – 16.73 | 8.06 | – 2.07* | | | |
| Marker 2 | 0.63 | 11.74 | 5.06 | 2.32 | | | | – 7.83 | 4.36 | – 1.80* |
| Parameter values: $p_1 = 0.5$, $p_2 = 0.5$, $\rho = 0.1$, $\theta_1 = 0.5$ and $\theta_2 = 0.5$ | | | | | | | | | | |
| Simult. | 0.24 | 11.07 | 6.24 | 1.77 | – 2.70 | 8.55 | – 0.32 | – 1.96 | 8.69 | – 0.23 |
| Seq. using: | | | | | | | | | | |
| Marker 1 | 0.12 | 13.43 | 7.00 | 1.92 | – 3.61 | 5.83 | – 0.63 | | | |
| Marker 2 | 0.08 | 14.08 | 8.72 | 1.62 | | | | – 2.36 | 6.95 | – 0.34 |

*Significant at 5% level.

**Table 7.** Efficiency of the simultaneous strategy over the sequential strategy for simulation parameter values $p_1 = 0.9$, $p_2 = 0.9$, $\theta_1 = 0.1$, $\theta_2 = 0.1$

| $\theta_1$ | $\theta_2$ | $n_1$ | $n_2$ | $N_1$ | $N_2$ | $E$ |
|---|---|---|---|---|---|---|
| 0 | 0.0 | 63 | 68 | 72 | 76 | 1.12 |
| | 0.1 | 65 | 123 | 72 | 143 | 1.16 |
| | 0.2 | 64 | 176 | 72 | 205 | 1.16 |
| | 0.3 | 68 | 254 | 72 | 293 | 1.15 |
| | 0.4 | 68 | 298 | 72 | 351 | 1.18 |
| 0.1 | 0.0 | 112 | 70 | 124 | 76 | 1.11 |
| | 0.1 | 114 | 126 | 124 | 143 | 1.14 |
| | 0.2 | 117 | 173 | 124 | 205 | 1.18 |
| | 0.3 | 115 | 258 | 124 | 293 | 1.14 |
| | 0.4 | 120 | 303 | 124 | 351 | 1.16 |
| 0.2 | 0.0 | 176 | 70 | 187 | 76 | 1.06 |
| | 0.1 | 178 | 128 | 187 | 143 | 1.05 |
| | 0.2 | 176 | 181 | 187 | 205 | 1.13 |
| | 0.3 | 180 | 260 | 187 | 293 | 1.13 |
| | 0.4 | 179 | 305 | 187 | 351 | 1.15 |
| 0.3 | 0.0 | 251 | 72 | 262 | 76 | 1.04 |
| | 0.1 | 254 | 131 | 262 | 143 | 1.03 |
| | 0.2 | 253 | 185 | 262 | 205 | 1.04 |
| | 0.3 | 256 | 264 | 262 | 293 | 1.11 |
| | 0.4 | 259 | 312 | 262 | 351 | 1.13 |
| 0.4 | 0.0 | 310 | 73 | 326 | 76 | 1.05 |
| | 0.1 | 312 | 135 | 326 | 143 | 1.04 |
| | 0.2 | 308 | 188 | 326 | 205 | 1.06 |
| | 0.3 | 311 | 267 | 326 | 293 | 1.05 |
| | 0.4 | 313 | 316 | 326 | 351 | 1.11 |

**Table 8.** Efficiency of the simultaneous strategy over the sequential strategy for simulation parameter values $p_1 = 0.5$, $p_2 = 0.5$, $\theta_1 = 0.5$, $\theta_2 = 0.5$

| $\theta_1$ | $\theta_2$ | $n_1$ | $n_2$ | $N_1$ | $N_2$ | $E$ |
|---|---|---|---|---|---|---|
| 0 | 0.0 | 27 | 30 | 36 | 38 | 1.27 |
| | 0.1 | 28 | 75 | 36 | 84 | 1.12 |
| | 0.2 | 30 | 107 | 36 | 120 | 1.12 |
| | 0.3 | 30 | 158 | 36 | 185 | 1.17 |
| | 0.4 | 32 | 199 | 36 | 232 | 1.17 |
| 0.1 | 0.0 | 72 | 32 | 78 | 38 | 1.09 |
| | 0.1 | 74 | 77 | 78 | 84 | 1.09 |
| | 0.2 | 74 | 109 | 78 | 120 | 1.10 |
| | 0.3 | 73 | 159 | 78 | 185 | 1.16 |
| | 0.4 | 75 | 203 | 78 | 232 | 1.14 |
| 0.2 | 0.0 | 104 | 34 | 112 | 38 | 1.08 |
| | 0.1 | 103 | 77 | 112 | 84 | 1.09 |
| | 0.2 | 105 | 111 | 112 | 120 | 1.08 |
| | 0.3 | 107 | 162 | 112 | 185 | 1.14 |
| | 0.4 | 109 | 208 | 112 | 232 | 1.12 |
| 0.3 | 0.0 | 159 | 35 | 168 | 38 | 1.06 |
| | 0.1 | 161 | 79 | 168 | 84 | 1.04 |
| | 0.2 | 162 | 114 | 168 | 120 | 1.04 |
| | 0.3 | 164 | 167 | 168 | 185 | 1.11 |
| | 0.4 | 165 | 210 | 168 | 232 | 1.10 |
| 0.5 | 0.0 | 200 | 36 | 211 | 38 | 1.06 |
| | 0.1 | 198 | 81 | 211 | 84 | 1.07 |
| | 0.2 | 202 | 117 | 211 | 120 | 1.04 |
| | 0.3 | 203 | 170 | 211 | 185 | 1.04 |
| | 0.4 | 205 | 214 | 211 | 232 | 1.08 |

We assess the nature of sample size requirements under various scenarios assuming that the QT is controlled by two loci. First, in the absence of epistatic interaction ($\Delta = 0$), if both loci have equal effects ($\alpha_1 = \alpha_2 = \alpha$), then the sample size required to map the first (or the second) trait locus decreases as heterozygosity at that locus increases (Figure 3). The rate of decrease, however, is greater when the locus has a smaller effect on the QT. Second, in the absence of epistatic interaction ($\Delta = 0$), the sample size required to map the locus which has a greater effect on the QT decreases as its relative effect increases (Figure 4 *a*). However although the rate of decrease in sample size depends largely on the heterozygosity of the locus with the greater effect, it also varies with the hetero zygosity of the second locus. For example, we find that while the sample size requirement to map the first locus in the case $p_1 = p_2 = 0.75$, is more than in the case $p_1 = p_2 = 0.5$ when the marginal effects of the two loci do not differ significantly, it requires a smaller sample to do so in the case $p_1 = p_2 = 0.75$, if the marginal effect of the first locus is quite large compared to that of the second locus. Moreover, while it requires a smaller sample to detect linkage at the first trait locus for $p_1 = p_2 = 0.5$, $p_1 = p_2 = 0.75$ and $p_1 = 0.5$, $p_2 = 0.75$ whenever the marginal effect

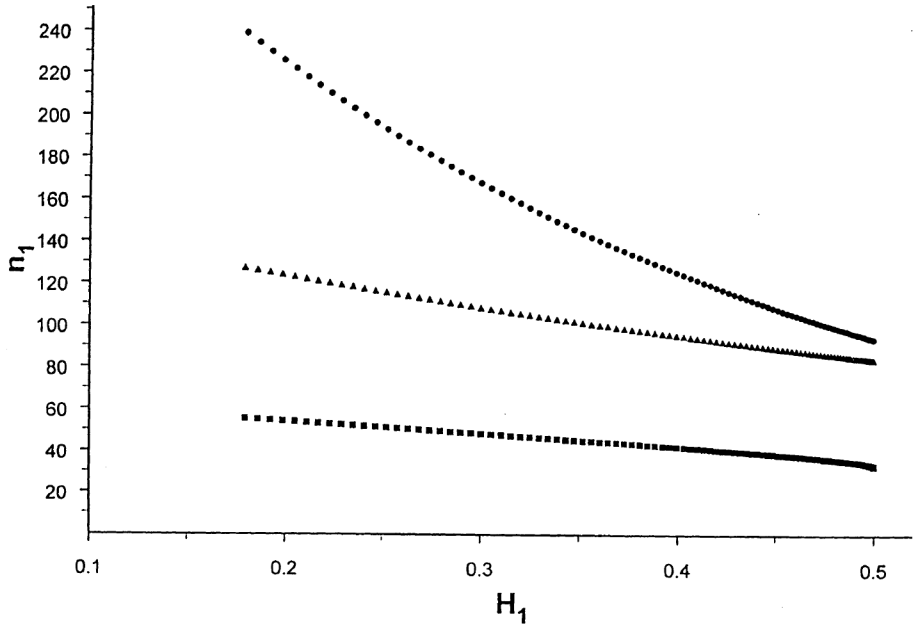


**Figure 3.** Sample size requirement to map the first trait locus for simulation parameter values $p_2 = 0.5$ and $\Delta = 0$. Circles correspond to $\alpha_1 = \alpha_2 = 2$, triangles to $\alpha_1 = \alpha_2 = 5$ and squares to $\alpha_1 = \alpha_2 = 10$.

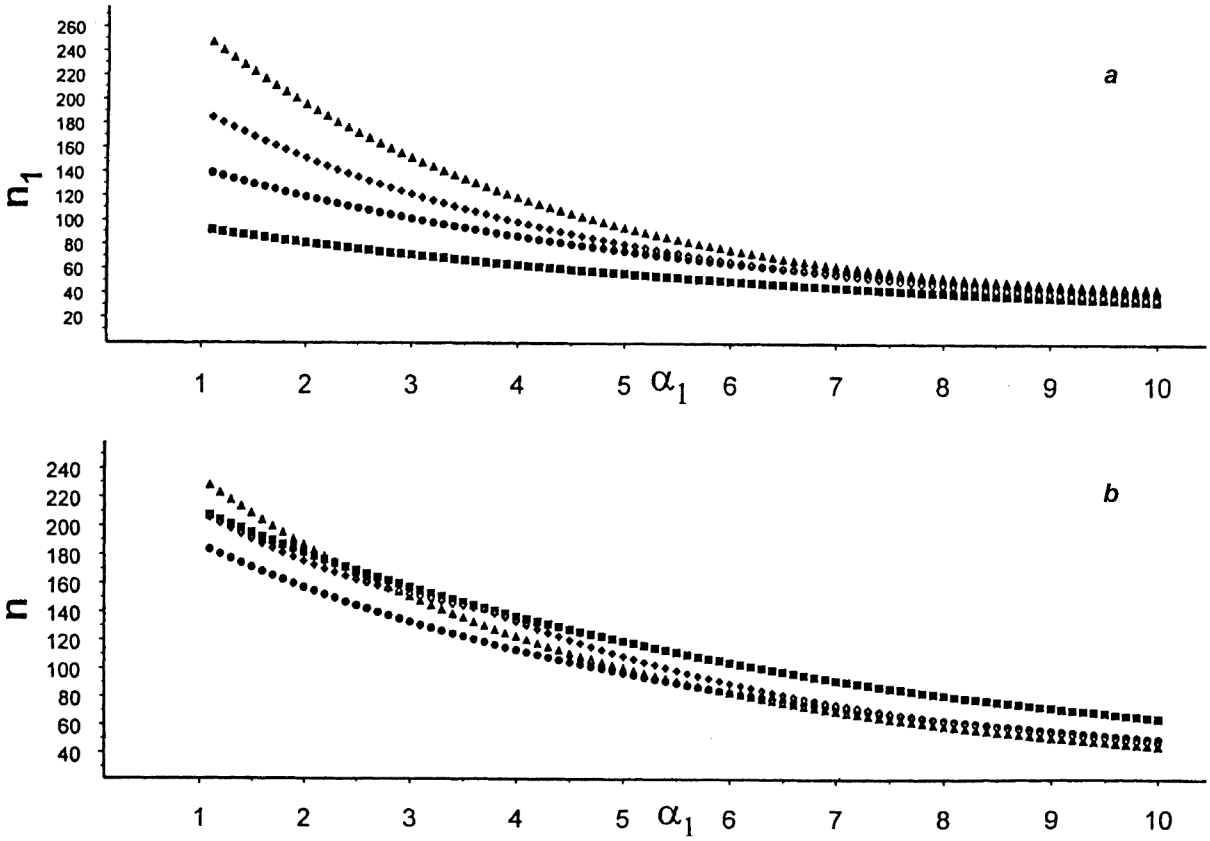


**Figure 4.** Sample size requirement to map (*a*) the first trait locus and (*b*) both the trait loci for simulation parameter values $\alpha_2 = 1$ and $\Delta = 0$. Triangles correspond to $p_1 = 0.75$, $p_2 = 0.5$, diamonds to $p_1 = 0.75$, $p_2 = 0.75$, circles to $p_1 = 0.5$, $p_2 = 0.5$ and squares to $p_1 = 0.5$, $p_2 = 0.75$.

of the first locus is higher, the sample size requirement in the case $p_1 = 0.75$, $p_2 = 0.5$ is more for the first locus, if the marginal effects of the two loci do not differ significantly, but is less if the marginal effect of the first locus is very large compared to that of the second locus (Figure 4 *b*). Next, we find that the sample size required to map either of the two loci decreases as the degree of epistatic interaction ($\Delta$) between the two loci increases (Figures 5–7). Moreover in the presence of epistatic interaction, it requires a smaller sample to map the locus with a greater marginal effect on the QT (Figures 5 *b*, 6 *b*, 7 *b*).

Comparing the simultaneous strategy with the sequential strategy (the results are presented in Tables 7 and 8), we find that the sample size requirement is, in general, less when we analyse the data by considering the two markers simultaneously. The 'efficiency' ratio $E$ (defined earlier in this subsection) is found to be greater than 1 in all our simulation studies.

### *Comparison with the Tiwari–Elston method*

As we noted earlier, the digenic interaction model is a special case of the more general epistasis model assumed by Tiwari and Elston[14], in which the epistasis parameters can vary arbitrarily. However, as their model involves a larger number of regressors, the tests for linkage (i.e. $H_0$: $\theta_i = 0.5$ vs $H_1$: $\theta_i < 0.5$) are much more conservative. We compare the powers of the two procedures under our proposed model using simulated data. We generate data sets with simulation parameter values of $\alpha_1 = 5$, $\alpha_2 = 1$, $\Delta = 1$, $\sigma^2 = 1$, $\theta_1 = 0.5$ and different values of $p_1$, $p_2$, $\theta_2$ for varying sample sizes. We perform 100 replications of regression using each set of parameter values and evaluate the power of the test $H_0$: $\theta_1 = 0.5$ vs $H_1$: $\theta_1 < 0.5$ at $\theta_1 = 0.1$. The average power of the 100 replications for each set of parameter values is presented in Table 9. If our model is indeed true, we find that the tests for linkage under their regression set-up are less powerful, especially if the sample size is small. Moreover, in that case, their regression equation is a gross overfit to our model.

### *Some recently developed sib-pair methods*

Multipoint mapping techniques (i.e. in which information on more than one marker is incorporated simultaneously to detect linkage with the putative trait locus) have been proved to be much more powerful than two-point linkage analyses (i.e. in which information on a single marker is used at a time; multipoint extension of Haseman–Elston[19], non-parametric technique[20], QLOD score[21], maximum likelihood binomial method[22]). We have recently developed a robust, two-stage, semi-parametric method of

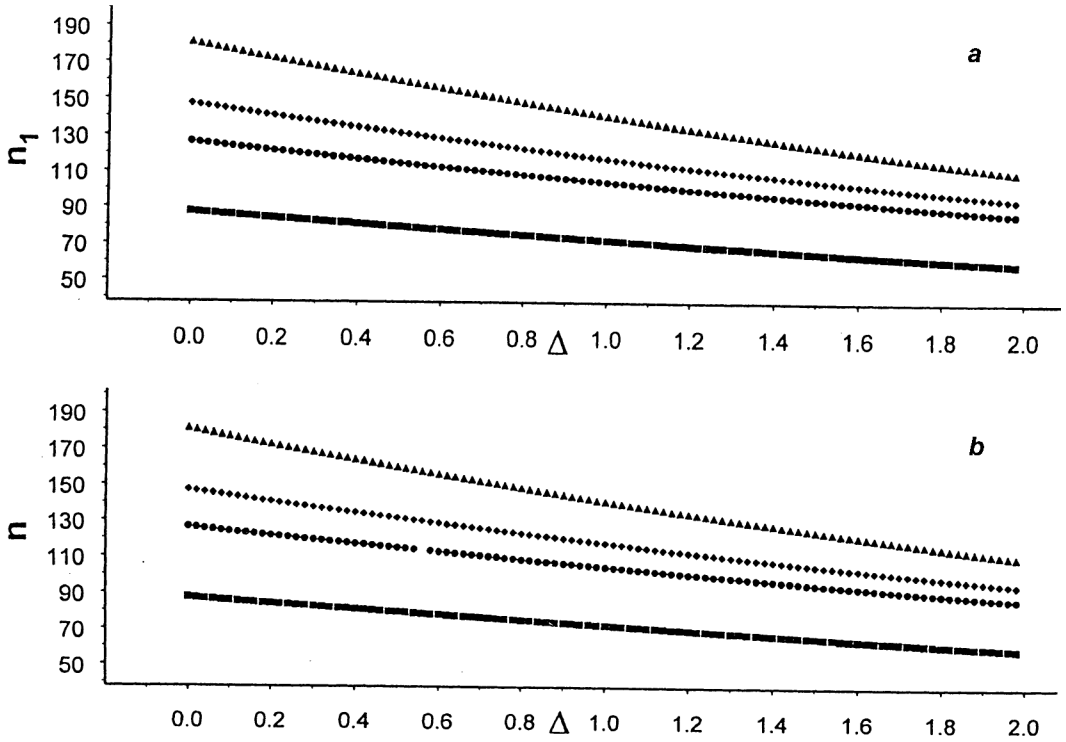


**Figure 5.** Sample size requirement to map (***a***) the first trait locus and (***b***) both the trait loci for simulation parameter values $\alpha_1 = 2$ and $\alpha_2 = 1$. Triangles correspond to $p_1 = 0.75$, $p_2 = 0.5$, diamonds to $p_1 = 0.75$, $p_2 = 0.75$, circles to $p_1 = 0.5$, $p_2 = 0.5$ and squares to $p_1 = 0.5$, $p_2 = 0.75$.

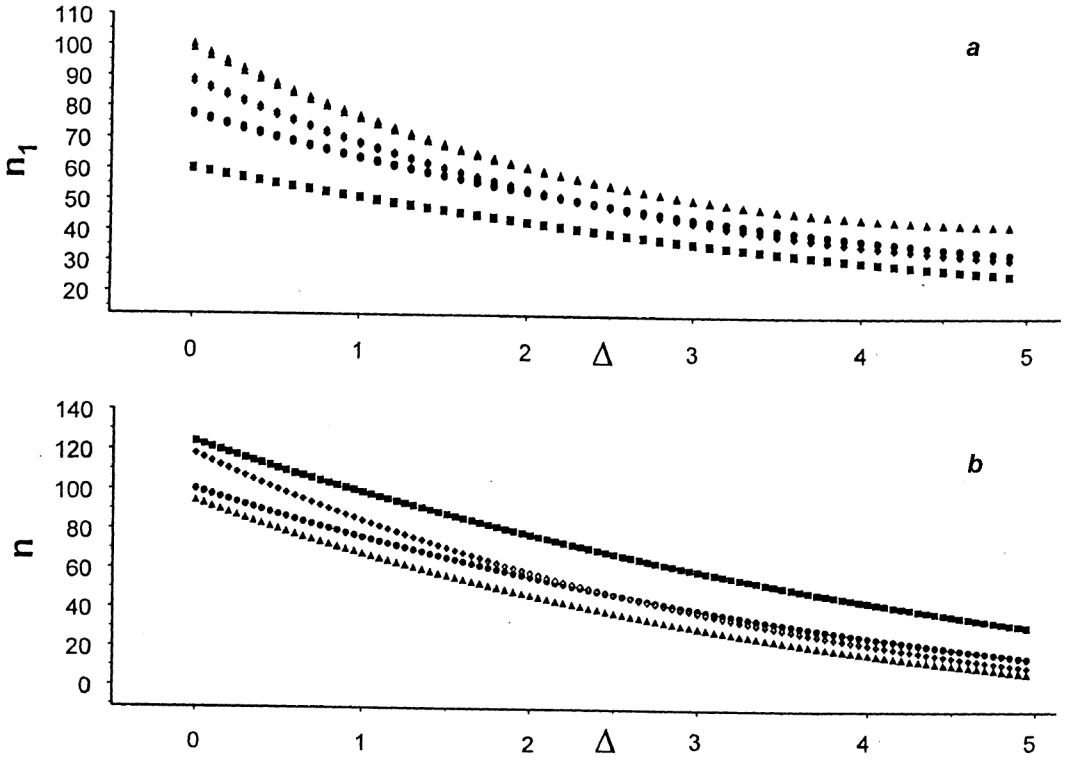


**Figure 6.** Sample size requirement to map (***a***) the first trait locus and (***b***) both the trait loci for simulation parameter values $\alpha_1 = 5$ and $\alpha_2 = 1$. Triangles correspond to $p_1 = 0.75$, $p_2 = 0.5$, diamonds to $p_1 = 0.75$, $p_2 = 0.75$, circles to $p_1 = 0.5$, $p_2 = 0.5$ and squares to $p_1 = 0.5$, $p_2 = 0.75$.

**Table 9.** Comparison of the powers of the digenic interaction model and the Tiwari–Elston model at $\theta_1 = 0.1$ for simulated parameter value of $\theta_1 = 0.5$ and different values of $p_1$, $p_2$, $\theta_2$ and number of sib-pairs $n = 50$, 100, 200

| | | | | $n = 50$ | | $n = 100$ | | $n = 200$ | |
|---|---|---|---|---|---|---|---|---|---|
| $p_1$ | $p_2$ | $\theta_1$ | $\theta_2$ | $P_{DI}$ | $P_{TE}$ | $P_{DI}$ | $P_{TE}$ | $P_{DI}$ | $P_{TE}$ |
| 0.7 | 0.5 | 0.5 | 0.0 | 0.75 | 0.67 | 0.83 | 0.80 | 0.91 | 0.90 |
| 0.7 | 0.9 | 0.5 | 0.1 | 0.81 | 0.74 | 0.88 | 0.85 | 0.95 | 0.94 |
| 0.9 | 0.9 | 0.5 | 0.3 | 0.72 | 0.64 | 0.81 | 0.79 | 0.88 | 0.88 |
| 0.5 | 0.5 | 0.5 | 0.5 | 0.78 | 0.72 | 0.85 | 0.82 | 0.90 | 0.89 |

$P_{DI}$ and $P_{TE}$ denote the powers of the digenic interaction model and the Tiwari–Elston model at $\theta_1 = 0.1$.

mapping QTLs based on genome-wide scan data (i.e. where a chromosome is saturated with a large number of markers and the interval location(s) of the QTL(s) are inferred) on sib-pairs[23]. In the first-stage, we use a statistic based on Spearman's rank correlation[24] between squared difference ($Y_j$) in trait values of each sib-pair and its estimated i.b.d. score at each marker locus ($\hat{\pi}_{m_j}$s), while in the second stage, we use a non-parametric regression of $Y_j$ on $\hat{\pi}_{m_j}$s with kernel smoothing[25].

One of the major current challenges in genetic epidemiology is to unravel genetic architectures of complex traits. Quantitative variables, possibly correlated, generally underlie complex traits. A heritable multivariate quantitative phenotype comprises several correlated component phenotypes that are usually pleiotropically controlled by multiple loci and environmental factors. One approach to decipher the genetic architecture of a multivariate phenotype, in particular, to map the underlying loci, is to reduce the dimensionality of the data by a data-reduction technique, such as principal component analysis. The extracted principal components can then be analysed in conjunction with marker data to map the underlying loci. We have used our proposed two-stage semi-parametric method on the extracted principal components of the multivariate phenotype of the sib-pairs and their identity-by-descent scores on several marker loci[26] to map the QTLs controlling the phenotype. We have found that if principal components are extracted without consideration of the underlying correlation structure of the multivariate phenotype, mapping the loci controlling the phenotype cannot be done efficiently and may require huge sample sizes.

## Discussion

Since there is increasing evidence of epistatic interactions among the loci determining a quantitative trait[8–11,16], we have attempted to devise an efficient estimator of the recombination fraction, $\theta$, between a quantitative trait locus and a marker locus in the presence of such inter-

actions. We have used a simple model of interaction among homozygotes at the different trait loci. This model is one of the basic models used in the study of epistatic interactions[14] and is helpful for capturing some essential features and complexities that underline QTL mapping in presence of epistatic interactions. Using an approach originally proposed by Jayakar[13], we have proposed separate, computationally simple, estimators for families in which only one parent is heterozygous at the marker locus (backcross type families) and those in which both parents are heterozygous (intercross type families). We have studied the efficiencies of these estimators when there are two trait loci and have shown that for a wide range of parameter values the estimators are quite efficient. We have proposed a non-parametric procedure for testing null hypotheses regarding $\theta$ and have shown that the power function of the test has desirable properties. We have also shown that analyses of data ignoring epistatic interactions, when in fact these are present may lead to grossly inaccurate inferences about linkage. Although most of our results pertain to the case of the marker locus being biallelic, we have theoretically shown that extension to a multiallelic marker locus is straightforward. However, we have found that the estimators obtained by this approach, although simple to use in practice, are not as efficient as the maximum likelihood estimator. We also note that our procedure does not provide simultaneous estimates of recombination and other parameters (i.e. quantitative trait locus effects, epistatic interaction effects, etc.). Independent estimates of these other parameters have to be obtained to estimate the recombination fraction and test hypotheses concerning this parameter. The upshot is that although the modified-Jayakar estimator proposed by us is computationally simple and enjoys some desirable statistical properties, in practice it is preferable to use the maximum likelihood estimator in view of its superior performance over a wider range of scenarios and parameter values.

For sib-pair data, we have extended the regression procedure proposed by Haseman and Elston[17] to map a single QTL, to the case of mapping two unlinked QTLs in the presence of epistatic interactions. Our proposed procedure provides a test for detecting linkage between a trait locus and a marker locus but fails to provide, as in the Haseman–Elston approach, an estimate of the recombination fraction between the two loci. We have derived expressions for the sample size requirement to map the two QTLs. The marginal effects of the different trait and linkage parameters on the sample size requirement have been analysed theoretically. In particular, we have shown that the sample size requirement for mapping a QTL is smaller if its marginal effect and heterozygosity are larger. Moreover, the presence of epistatic interactions reduces the sample size requirement compared to the situation in which the marginal effects are same, but epistatic interactions are absent. We have also assessed the nature of dependence of sample size requirements on different trait parameters considered simultaneously. We have shown through simulation studies that the simultaneous analysis of markers reduces the sample size requirements and thus

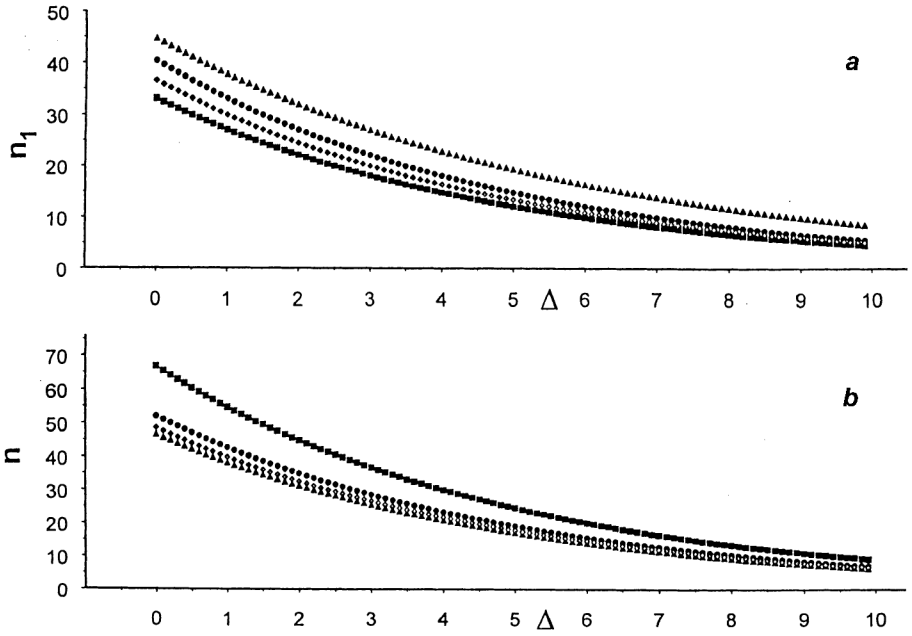


**Figure 7.** Sample size requirement to map (*a*) the first trait locus and (*b*) both the trait loci for simulation parameter values $\alpha_1 = 10$ and $\alpha_2 = 1$. Triangles correspond to $p_1 = 0.75$, $p_2 = 0.5$, diamonds to $p_1 = 0.75$, $p_2 = 0.75$, circles to $p_1 = 0.5$, $p_2 = 0.5$ and squares to $p_1 = 0.5$, $p_2 = 0.75$.

is more cost-effective compared to the sequential analysis. This equivalently implies that under a misspecified model of a single QTL, we would require a larger sample to map the QTL. The proposed regression approach has been extended to the case of multiple QTLs with a typical epistatic interaction structure. The results are similar to the case of two QTLs with digenic epistatic interaction.


1. Fisher, R. A., *Trans. R. Soc. Edinburgh*, 1918, **52**, 399–433.
2. Thoday, J. M., *Nature*, 1961, **191**, 368–370.
3. Ott, J., *Analysis of Human Genetic Linkage*, Johns Hopkins University Press, Baltimore, 1999, 3rd edn.
4. Tanksley, S. D., *Annu. Rev. Genet.*, 1993, **27**, 207–233.
5. Georges, M. *et al.*, *Genetics*, 1995, **139**, 907–920.
6. Berrethini, W. H., Ferraro, T. N., Alexander, R. C., Buchberg, A. M. and Vogel, W. H., *Nat. Genet.*, 1994, **7**, 54–58.
7. Schork, N. J. *et al.*, *Genome Res.*, 1995, **5**, 164–172.
8. Lark, K. G., Chase, Adler, F., Mansur, L. M. and Orf, J. H., *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 4656–4660.
9. Coupland, G., *Trends Genet.*, 1995, **11**, 393–397.
10. Fijneman, R. J. A., de Vries, S. S., Jensen, R. C. and Demant, P., *Nat. Genet.*, 1996, **9**, 465–467.
11. van Wezel, T., Stassen, A. P. M., Moen, C. J. A., Hart, A. A. M., van der Valk, M. A. and Demant, P., *Nat. Genet.*, 1996, **9**, 468–470.
12. Frankel, W. N. and Schork, N. J., *Nat. Genet.*, 1996, **9**, 371–373
13. Jayakar, S. D., *Biometrics*, 1970, **26**, 451–464.
14. Tiwari, H. K. and Elston, R. C., *Ann. Hum. Genet.*, 1997, **61**, 253–261.
15. Kearsey, M. J. and Pooni, H. S., *The Genetical Analysis of Quantitative Traits*, Chapman and Hall, London, 1996.
16. Chang, B. *et al.*, *Nat. Genet.*, 1999, **21**, 405–409.
17. Haseman, J. K. and Elston, R. C., *Behav. Genet.*, 1972, **2**, 3–19.
18. Cotterman, C. W., in *Applications in Genetics*, University of Hawaii Press, Honolulu, 1969, pp. 1–19.
19. Olson, J. M., *Genet. Epidemiol.*, 1995, **12**, 177–193.
20. Kruglyak, L. and Lander, E. S., *Genetics*, 1995, **139**, 1421–1428.
21. Page, G. P., Amos, C. I. and Boerwinkle, E., *Am. J. Hum. Genet.*, 1998, **62**, 962–968.
22. Alcais, A. and Abel, L., *Genet. Epidemiol.*, 1999, **17**, 102–117.
23. Ghosh, S and Majumder, P. P., *Am. J. Hum. Genet.*, 2000, **66**, 1046–1061.
24. Randles, R. H. and Wolfe, D. A., *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons, New York, 1979.
25. Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
26. Ghosh, S. and Majumder, P. P., *Adv. Genet.*, 2001, **42**, 323–347.