Theor Appl Genet (1990) 79:583–592

THEORETICAL AND APPLIED GENETICS

# Mapping quantitative trait loci using molecular marker linkage maps

**S. J. Knapp** [1], *, **W. C. Bridges, Jr.** [3] **and D. Birkes** [2]

[1] Department of Crop Science, Oregon State University, Corvallis, OR 97331, USA
[2] Department of Statistics, Oregon State University, Corvallis, OR 97331, USA
[3] Department of Experimental Statistics, Clemson University, Clemson, SC 29643, USA

Received August 23, 1989; Accepted November 30, 1989
Communicated by A. R. Hallauer

**Summary.** High-density restriction fragment length polymorphism (RFLP) and allozyme linkage maps have been developed in several plant species. These maps make it technically feasible to map quantitative trait loci (QTL) using methods based on flanking marker genetic models. In this paper, we describe flanking marker models for doubled haploid (DH), recombinant inbred (RI), backcross (BC), $F_1$ testcross ($F_1$ TC), DH testcross (DHTC), recombinant inbred testcross (RITC), $F_2$, and $F_3$ progeny. These models are functions of the means of quantitative trait locus genotypes and recombination frequencies between marker and quantitative trait loci. In addition to the genetic models, we describe maximum likelihood methods for estimating these parameters using linear, nonlinear, and univariate or multivariate normal distribution mixture models. We defined recombination frequency estimators for backcross and $F_2$ progeny group genetic models using the parameters of linear models. In addition, we found a genetically unbiased estimator of the QTL heterozygote mean using a linear function of marker means. In nonlinear models, recombination frequencies are estimated less efficiently than the means of quantitative trait locus genotypes. Recombination frequency estimation efficiency decreases as the distance between markers decreases, because the number of progeny in recombinant marker classes decreases. Mean estimation efficiency is nearly equal for these methods.

**Key words:** Restriction fragment length polymorphisms – Allozymes – Maximum lilkelihood estimators – Normal distribution mixture models

## Introduction

High-density restriction fragment length polymorphism (RFLP) and allozyme linkage maps are the technological tools needed to efficiently map quantitative trait loci (Lander and Botstein 1989). Marker loci are typically dispersed at 1–30 centimorgan (cM) intervals in high-density maps. These maps have several significant features. They provide fairly complete genome coverage and increase the probability of finding QTL. In addition, they enable the use of flanking marker genetic models. In the flanking marker model, it is hypothesized that a quantitative trait locus lies between linked codominant marker loci (Weller 1987; Lander and Botstein 1989). These models are more efficient than individual marker models for estimating the effects of quantitative trait loci (Lander and Botstein 1989). The individual marker locus model describes the cosegregation of a marker locus linked to a quantitative trait locus.

Weller (1987) described the expected values of the means of nonrecombinant $F_2$ marker classes using flanking markers. These values and expected values of individual marker genotype means were used to define an estimator of the recombination frequency between the marker and quantitative trait locus ($r_1$). The expected values of the means of marker genotypic classes have not been described for backcross and $F_2$ progeny using flanking markers, except for those for nonrecombinant $F_2$ marker classes (Weller 1987).

Individual marker models have been described for various progeny types (Soller and Brody 1976; Soller et al. 1979; Weller 1986; Cowen 1988). The expected values of marker genotype means in these models are nonlinear functions of means of QTL genotypes and $r_1$. Because of this, linear models based on individual marker models cannot be used to estimate quantitative trait lo-

* To whom correspondence should be addressed

cus parameters (Weller 1986), unless quantitative trait locus genotypes are used as independent variables and a method is used to simultaneously estimate missing quantitative trait locus genotypic values. This is done in the linear model algorithm proposed by Lander and Botstein (1989). In individual marker models, QTL genotypic values are completely missing.

Linear model analyses of individual marker locus genotypic classes have been widely used in crop species (Edwards et al. 1987; Osborne et al. 1987; Stuber et al. 1987; Tanksley and Hewitt 1988; Young et al. 1988). In these studies, contrasts among marker means have been used to estimate QTL effects. These contrasts underestimate the effects of quantitative trait loci, as is well known, because they are confounded with the recombination frequency between the marker and quantitative trait locus (Lander and Botstein 1989). The expected values of contrasts among homozygote marker means (additive effect contrasts) in $F_2$ and doubled haploid progeny, e.g., are

$$\begin{aligned}\mu^*_{11}-\mu^*_{22}&=(1-2r_1+r_1^2)\,\mu_{11}+r_1^2\,\mu_{22}-r_1^2\,\mu_{11}\\&\quad-(1-2r_1+r_1^2)\mu_{22}=(1-2r_1)\,(\mu_{11}-\mu_{22})\end{aligned}$$

and

$$\begin{aligned}\mu^*_{11}-\mu^*_{22}&=(1-r_1)\,\mu_{11}+r_1\,\mu_{22}-r_1\,\mu_{11}-(1-r_1)\,\mu_{11}\\&=(1-2r_1)\,(\mu_{11}-\mu_{22}),\end{aligned}$$

respectively, where $\mu^*_{11}$ and $\mu^*_{22}$ are means of marker locus genotoypes 11 and 22, respectively, and $\mu_{11}$ and $\mu_{22}$ are means of QTL genotypes 11 and 22, respectively. Thus, $F_2$ and doubled haploid progeny additive effect biases are equal. These biases affect the power of tests for estimating the effects of QTL. In the individual marker model, power is a function of the size of an effect and $r_1$, holding other factors constant. Power decreases as $r_1$ increases for an effect of fixed size (Soller and Brody 1976; Soller et al. 1979).

Weller (1986) described expected values of the means and variances of marker genotypic classes and approximate maximum likelihood methods for estimating the parameters of mixed QTL genotypic distributions for $F_2$ progeny. Simulations were done using $F_2$ populations of 500 or 2,000 individuals. $r_1$ was not well determined using these sample sizes, i.e., variances of $r_1$ estimates were large and the likelihood surface was flat in the $r_1$ dimension. Estimates of the means and variances of quantitative trait locus genotypes were fairly well determined (Weller 1986).

In this paper, we describe flanking marker genetic models for doubled haploid (DH), recombinant inbred (RI), backcross (BC), $F_1$ testcross ($F_1$TC), DH testcross (DHTC), RI testcross (RITC), $F_2$,and $F_3$ progeny. In addition, we describe constrained linear, nonlinear, and normal distribution mixture models for estimating the means of quantitative trait locus genotypes and recombination frequencies between marker and quantitative trait loci.

## Genetic and statistical models

### *Backcross progeny group genetic models*

DH, RI, BC, $F_1$TC, DHTC, and RITC progeny are in the backcross group. Genetic models within this group are algebraically equivalent; thus, there is no need to define separate models for different progeny types within the group. There are notation differences in the group, and differences in the means which can be estimated. But in some cases there are no differences, e.g., DH and RI progeny have identical notation and can be used to estimate homozygote QTL means. Their recombination frequency estimators are different, but the RI estimator is a linear function of the DH estimator.

Let $Q$ denote a quantitative trait locus lying between linked codominant molecular marker loci $A$ and $B$. In addition, let $r_1$, $r_2$, and $r$ denote recombination frequencies between $A$ and $Q$, $B$ and $Q$, and $A$ and $B$, respectively. Suppose $A$, $Q$, and $B$ locus genotypes in inbred lines $P_1$, $P_2$, and $P_3$ are $A_1A_1Q_1Q_1B_1B_1$, $A_2A_2Q_2Q_2B_2B_2$, and $A_3A_3Q_3Q_3B_3B_3$, respectively. The models described in this paper are based on progeny derived from the $F_1$ hybrid between $P_1$ and $P_2$.

Suppose DH lines are produced by self-fertilizing doubled haploids derived from the $F_1$. Expected values of marker genotype means in DH progeny are functions of $r_1$, $r_2$, $\mu_{11}$, and $\mu_{22}$ are means of QTL genotypes $Q_1Q_1$ and $Q_2Q_2$, respectively. Let $\theta_1$, $\theta_2$, $\theta_3$, and $\theta_4$ denote expected values of the means of marker genotypes $A_1A_1B_1B_1$, $A_1A_1B_2B_2$, $A_2A_2B_1B_1$, and $A_2A_2B_2B_2$, respectively. Relative frequencies of QTL genotypes within marker classes were used to derive expected values. Expected values were derived using no-double-crossover gamete frequencies (Table 1 and Table 2) and double-crossover gamete frequencies (Table 1 and Table 3). For example, relative frequencies of QTL genotypes $Q_1Q_1$ and $Q_2Q_2$ in the $A_1A_1B_1B_1$ marker genotypic class are

$$1/2(1-r_1-r_2)/[1/2(1-r_1-r_2)]=1.0$$

and 0.0, respectively, when there are no double-crossovers. Likewise, relative frequencies of QTL genotypes $Q_1Q_1$ and $Q_2Q_2$ in the $A_1A_1B_2B_2$ marker genotypic class are $r_2/(r_1+r_2)$ and $r_1/(r_1+r_2)$, respectively. Except for notation differences, expected values of other progeny types in the backcross group are identical to those for DH progeny.

In addition to the no-double-crossover model, we derived a genetic model using arbitrary double-crossover frequencies (coefficient of coincidence, Table 3). This model is theoretically preferable but, as we elaborate throughout the paper, there are strong arguments to support the use of the no-double-crossover model (Table 2). There are, for example, linear model estimators of BC and $F_2$ progeny group QTL means and recombination frequencies based on the no-double-crossover model.

**Table 1.** Genotypes and genotype frequencies for doubled haploid lines derived from an $F_1$ ($A_1A_2Q_1Q_2B_1B_2$) based on no-double-crossovers and double-crossovers

| Genotype | No-double-crossover frequency | Double-crossover frequency[a] |
|---|---|---|
| $A_1A_1Q_1Q_1B_1B_1$ | $1/2(1-r_1-r_2)$ | $1/2(1-r_1-r_2+\delta r_1 r_2)$ |
| $A_2A_2Q_2Q_2B_2B_2$ | $1/2(1-r_1-r_2)$ | $1/2(1-r_1-r_2+\delta r_1 r_2)$ |
| $A_1A_1Q_2Q_2B_2B_2$ | $1/2r_1$ | $1/2(r_1-\delta r_1 r_2)$ |
| $A_2A_2Q_1Q_1B_1B_1$ | $1/2r_1$ | $1/2(r_1-\delta r_1 r_2)$ |
| $A_1A_1Q_1Q_1B_2B_2$ | $1/2r_2$ | $1/2(r_1-\delta r_1 r_2)$ |
| $A_2A_2Q_2Q_2B_1B_1$ | $1/2r_2$ | $1/2(r_1-\delta r_1 r_2)$ |
| $A_1A_1Q_2Q_2B_1B_1$ | 0 | $1/2\delta r_1 r_2$ |
| $A_2A_2Q_1Q_1B_2B_2$ | 0 | $1/2\delta r_1 r_2$ |

[a] $r_1$ and $r_2$ are recombination frequencies between $A$ and $Q$ and $B$ and $Q$, respectively, and $\delta$ is the coefficient of coincidence

**Table 2.** Marker locus genotypes, QTL genotype mixtures, and expected values of marker genotype means for doubled haploid progeny based on the no-double-crossover flanking marker model

| Marker locus genotype | QTL genotype mixture | Expected value of marker genotype mean[a] |
|---|---|---|
| $A_1A_1B_1B_1$ | $Q_1Q_1$ | $\theta_1=\mu_{11}$ |
| $A_1A_1B_2B_2$ | $Q_1Q_1+Q_2Q_2$ | $\theta_2=(r_2\,\mu_{11}+r_1\,\mu_{22})/(r_1+r_2)$<br>$=(1-\varrho)\,\mu_{11}+\varrho\,\mu_{22}$ |
| $A_2A_2B_1B_1$ | $Q_1Q_1+Q_2Q_2$ | $\theta_3=(r_1\,\mu_{11}+r_2\,\mu_{22})/(r_1+r_2)$<br>$=\varrho\,\mu_{11}+(1-\varrho)\,\mu_{22}$ |
| $A_2A_2B_2B_2$ | $Q_2Q_2$ | $\theta_4=\mu_{22}$ |

[a] $\varrho=r_1/r$ where $r=r_1+r_2$ and $r_1$, $r_2$, and $r$ are recombination frequencies between $A$ and $Q$, $B$ and $Q$, and $A$ and $B$, respectively. $\mu_{11}$ and $\mu_{22}$ are means of $Q_1Q_1$ and $Q_2Q_2$ genotypes, respectively

**Table 3.** Marker locus genotypes, QTL genotype mixtures, and expected values of marker genotype means for doubled haploid progeny based on the double crossover flanking marker model

| Marker locus genotype | QTL genotype mixture | Expected value of marker genotype mean[a] |
|---|---|---|
| $A_1A_1B_1B_1$ | $Q_1Q_1+Q_2Q_2$ | $[\mu_{11}(1-r_1-r_2+\delta r_1 r_2)$<br>$+\mu_{22}\,\delta r_1\,\delta r_2]/\gamma_1$ |
| $A_1A_1B_2B_2$ | $Q_1Q_1+Q_2Q_2$ | $[\mu_{11}(r_2-\delta r_1 r_2)$<br>$+\mu_{22}(r_1-\delta r_1 r_2)]/\gamma_2$ |
| $A_2A_2B_1B_1$ | $Q_1Q_1+Q_2Q_2$ | $[\mu_{11}(r_1-\delta r_1 r_2)$<br>$+\mu_{22}(r_2-\delta r_1 r_2)]/\gamma_2$ |
| $A_2A_2B_2B_2$ | $Q_1Q_1+Q_2Q_2$ | $[\mu_{11}\,\delta r_1 r_2+\mu_{22}$<br>$\cdot(1-r_1-r_2+\delta r_1 r_2)]/\gamma_1$ |

[a] $\gamma_1=1-r_1-r_2+2\delta r_1 r_2$ and $\gamma_2=r_1+r_2-2\delta r_1 r_2$ where $r_1$ and $r_2$ are recombination frequencies between $A$ and $Q$ and $B$ and $Q$, respectively, and $\delta$ is the coefficient of coincidence. $\mu_{11}$ and $\mu_{22}$ are means of $Q_1Q_1$ and $Q_2Q_2$ genotypes, respectively

Recombinant inbred or single-seed descent lines are produced by self-fertilizing one individual per line per generation for five or more generations beyond the $F_2$. RI progeny genotypes are the same as those for double haploid progeny genotypes; however, genotype frequencies differ because there are several additional meioses in RI progeny (Haldane and Waddington 1931; Cowen 1988). Expected values of means of RI line marker genotypes are defined by substituting $R_1=2r_1/(1+2r_1)$ and $R_2=2r_2/(1+2r_2)$ for $r_1$ and $r_2$ (Haldane and Waddington 1931) in DH expected values (Table 2). $R_1$ and $R_2$ are estimated instead of $r_1$ and $r_2$ and solved for $r_1$ and $r_2$, respectively. The respective estimators of $r_1$ and $r_2$ are $\hat{r}_1=\hat{R}_1/(2-2\hat{R}_1)$ and $\hat{r}_2=\hat{R}_2/(2-2\hat{R}_2)$.

Suppose BC progeny are produced using $P_1$ ($A_1A_1Q_1Q_1B_1B_1$) as the recurrent parent, i.e., $(P_1 \times P_2) \times P_1$. Given this cross, expected values of means of marker genotypes are functions of $r_1$, $r_2$, $\mu_{11}$, and $\mu_{12}$ where $\mu_{12}$ is the mean for $Q_1Q_2$ genotypic class; thus, BC equations are obtained by substituting $\mu_{12}$ and $\mu_{22}$ in the DH equations (Table 2). In addition, in backcross progeny, $\theta_1$, $\theta_2$, $\theta_3$, and $\theta_4$ are expected values of the means of marker genotypes $A_1A_1B_1B_1$, $A_1A_1B_1B_2$, $A_1A_2B_1B_1$, and $A_1A_2B_1B_2$, respectively.

Let $F_1$TC, DHTC, and RITC progeny be produced by crossing an $F_1$ or DH or RI line, respectively, to $P_3$. Cowen (1988) described individual marker models for these progeny. In the flanking marker model, $\theta_1$, $\theta_2$, $\theta_3$, and $\theta_4$ are expected values of the means of marker genotypes $A_1A_3B_1B_3$, $A_1A_3B_2B_3$, $A_2A_3B_1B_3$, $A_2A_3B_2B_3$, respectively, and $\mu_{13}$ and $\mu_{23}$ are substituted for $\mu_{11}$ and $\mu_{22}$, respectively, in DH equations (Table 2). In RITC equations, $R_1$ and $R_2$ are substituted for $r_1$ and $r_2$, respectively.

## *Backcross progeny group linear and nonlinear models*

The flanking marker doubled haploid model (Table 2) led to several useful statistical models and estimators of recombination frequencies between marker and quantitative trait loci and means of QTL genotypes. The model arising directly from these equations is

$$y=\mu_{11}\,x_1+[(r_2\,\mu_{11}+r_1\,\mu_{22})/(r_1+r_2)]\,x_2 + [(r_1\,\mu_{11}+r_2\,\mu_{22})/(r_1+r_2)]\,x_3+\mu_{22}\,x_4+e \qquad (1)$$

where $y$ is a dependent variable, $x_1$, $x_2$, $x_3$, and $x_4$ are independent variables indexing marker genotypes, and $e$ is a random experimental error. If the marker genotype is $A_1A_1B_1B_1$, then $x_1=1$, otherwise $x_1=0$. If the marker genotpye is $A_1A_1B_2B_2$, then $x_2=1$, otherwise $x_2=0$. If the marker genotype is $A_2A_2B_1B_1$, then $x_3=1$, otherwise $x_3=0$. If the marker genotype is $A_2A_2B_2B_2$, then $x_4=1$, otherwise $x_4=0$.

Model (1) is an overparameterized nonlinear model because $r_1$ and $r_2$ are not identifiable. Putting $\varrho = r_1/(r_1+r_2) = r_1/\mathrm{r}$ in Eq. (1) leads to the equivalent model

$$y = \mu_{11}\, x_1 + [(1-\varrho)\, \mu_{11} + \varrho\, \mu_{22}]\, x_2 + [\varrho\, \mu_{11} + (1-\varrho)\, \mu_{22}]\, x_3 + \mu_{22}\, x_4 + e \qquad (2)$$

with the constraint $0 \leq \varrho \leq 1$. Depending on the nature of the dependent variable, other constraints such as $\mu_{11} > 0$ and $\mu_{22} > 0$ may be appropriate. The parameters of Eq. (2) ($\mu_{11}$, $\mu_{22}$, and $\varrho$) are identifiable and estimable. Convergence of the estimation procedure should be improved by rewriting the model

$$y = \mu_{11}(x_1 + x_3) + [(1-\varrho)\, \mu_{11} + \varrho\, \mu_{22}]\, (x_2 - x_3) + \mu_{22}(x_3 + x_4) + e \qquad (3)$$

and using the same constraints as those used in model (2).

The problem of estimating $r_1$ and $r_2$ is circumvented in practice by estimating $r = r_1 + r_2$ from marker phenotypes (Allard 1956) and using this estimate ($\hat{r}$) and Eq. (3) to estimate $\varrho$ and the means. $r_1$ and $r_2$ can be estimated by putting $\hat{r}_1 = \hat{\varrho}\, \hat{r}$ and $\hat{r}_2 = \hat{r} - \hat{r}_1$. Equivalently, we can define $\varrho = r_1/\hat{r}$, rather than $r_1/r$, and impose the constraint $\hat{r} = r_1 + r_2$. This gives Eq. (3), but the definition of $\varrho$ is slightly different. Estimates of $\mu_{11}$, $\mu_{22}$, $r_1$, and $r_2$ based on either definition of $\varrho$ and (3) are equivalent.

There are several useful linear models for the backcross progeny group. Putting $\theta_1 = \mu_{11}$, $\theta_2 = (1-\varrho)\, \mu_{11} + \varrho\, \mu_{22}$, $\theta_3 = \varrho\, \mu_{11} + (1-\varrho)\, \mu_{22}$, and $\theta_4 = \mu_{22}$ in Eq. (2) leads to

$$y = \theta_1\, x_1 + \theta_2\, x_2 + \theta_3\, x_3 + \theta_4\, x_4 + e, \qquad (4)$$

where the $\theta_i$ are marker means. Suppose the $\theta_i$ are functionally independent parameters and no constraints are imposed, in particular, the equality $\theta_1 + \theta_4 = \theta_2 + \theta_3$ is ignored, then estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ from Eq. (4) can be used to estimate the means of QTL genotypes. Suppose, however, this equality is used to redefine Eq. (4) by substituting $\theta_3 = \theta_1 - \theta_2 + \theta_4$, then we get the linear model

$$y = \theta_1(x_1 + x_3) + \theta_2(x_2 - x_3) + \theta_4(x_3 + x_4) + e. \qquad (5)$$

Note that the estimators $\theta_1$, $\theta_2$, and $\theta_4$ of Eq. (5) do not coincide with the estimators $\theta_1$, $\theta_2$, and $\theta_4$ of Eq. (4). An estimator of $\varrho$ based on Eq. (5) is

$$\hat{\varrho} = (\hat{\theta}_2 - \hat{\theta}_1)/(\hat{\theta}_4 - \hat{\theta}_1). \qquad (6)$$

Since Eq. (5) has no constraints, this $\hat{\varrho}$ may not satisfy $0 \leq \hat{\varrho} \leq 1$. As before, $r_1$ and $r_2$ can be estimated using $\hat{r}_1 = \hat{\varrho}\, \hat{r}$ and $\hat{r}_2 = \hat{r} - \hat{r}_1$. Suppose $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_4$ are the least-squares estimators of Eq. (5). Set $\hat{\mu}_{11} = \hat{\theta}_1$ and $\hat{\mu}_{22} = \hat{\theta}_4$. If $\hat{\mu}_{11}$, $\hat{\mu}_{22}$, and $\hat{\varrho}$ of Eq. (5) satisfy the constraints of Eq. (3), then they coincide with the least-squares estimators of Eq. (3).

The analysis of linear models (4) and (5) and nonlinear model (3) is straightforward. Partial derivatives of model (3) are needed to implement certain nonlinear model estimation algorithms, e.g., the Gauss-Newton algorithm (Jennrich and Ralston 1978; Gallant 1987). These derivatives are given in the Appendix. An estimation example based on Model (3) and the Gauss-Newton algorithm is described in the next section.

### *Backcross progeny group normal distribution mixture models*

Lander and Botstein (1989) used the EM algorithm (Little and Rubin 1987) to develop mapping methods where missing genotypic values and quantitative trait locus means are simultaneously estimated. They used a linear model based on the no-double-crossover model where quantitative trait locus genotypes are used as independent variable values (Lander and Botstein 1989). In this section, we describe mapping methods using normal distribution mixture models where QTL genotypes are used as independent variable values. We use maximum likelihood methods based on the EM algorithm (Dempster et al. 1977; Little and Rubin 1987; McLachlan and Basford 1988) to simultaneously estimate missing QTL genotypic values and means and variances of QTL genotypes. In the proposed analysis, mixing weights are equal to the segregation ratios for QTL genotypes.

Maximum likelihood methods are often used to estimate mixture model parameters. The likelihood function used depends on whether or not there are observations of known group origin and, if there are, whether or not the observations arise in the proportions expected for the population under study (Titterington et al. 1985). Let $L_1$ and $L_2$ be the likelihood functions appropriate for situations where there are categorized observations. $L_1$ is used when the categorized observations arise in proportions different from expected segregation ratios. $L_2$ is used when the categorized observations arise in proportions equal to expected segregation ratios. Let $L_0$ be the likelihood function appropriate for situations where there are no categorized observations. Parameter estimation efficiency using these likelihoods usually increases as the information increases ($L_0 < L_1 < L_2$) (Titterington et al. 1985; McLachlan and Basford 1988).

Mixture model mapping methods can be applied to the no-double-crossover (Table 2) or double-crossover (Table 3) model. In the no-double-crossover model (Table 2), there are categorized observations for each group or QTL genotypic class. The means of these classes are biased by double-crossovers, but this bias is often negligible. In the double-crossover model (Table 3), each marker class is comprised of mixtures of the two genotypes; thus, there are no categorized observations (known QTL genotypes).

These differences have important ramifications. If the no-double-crossover model is used, then $L_2$ should be used. If the double-cross-over model is used, then $L_0$ must be used, although estimates based on the no-double-crossover model can be used as starting values. In the double crossover model, standard Mendelian methods can be used to estimate recombination frequencies and map distances from estimated QTL genotypic values. This is an attractive feature of this model. In the no-double-crossover model, the estimation of recombination frequencies between marker and quantitative trait loci is constrained by the assumption of no double crossovers. Thus, recombination frequencies can be estimated directly from the hypothesized (categorized) and estimated (uncategorized) genotypic values using the constraint $r = r_1 + r_2$.

The use of the no-double-crossover model may be justified because estimation based on $L_0$ is usually less efficient than estimation based on $L_2$ (Titterington et al. 1985), but numerical studies have not been done to investigate the statistical properties of recombination frequency estimators based on these likelihood functions. The difference in efficiency between $L_0$ and $L_2$ may overshadow the bias caused by double-crossovers in the no-double-crossover model. We have implemented mapping algorithms based on both models, but their performance has not been investigated.

In the mixture model, a phenotypic distribution is hypothesized to be a mixture of $g$ quantitative trait locus genotypic distributions. A mixture model for either doubled haploid genetic model is

$$f(y|\psi) = \pi_1\, \phi(y|\mu_{11}, \sigma_{11}^2) + (1-\pi_1)\, \phi(y|\mu_{22}, \sigma_{22}^2) \qquad (7)$$

where $\phi(y|\mu, \sigma^2)$ denotes a univariate normal density (Titterington et al. 1985). The objective is to estimate

$$\psi = [\mu_{11}\ \mu_{22}\ \sigma_{11}^2\ \sigma_{22}^2\ \pi_1]'$$

where $\sigma_{11}^2$ and $\sigma_{22}^2$ are phenotypic variances of $Q_1Q_1$ and $Q_2Q_2$ genotypic distributions, respectively, and $\pi_1$ is the mixing weight for the $Q_1Q_1$ genotypic distribution. The mixing weight for the $Q_2Q_2$ genotypic distribution is $\pi_2 = (1-\pi_1)$.

To estimate mixing weights, the posterior probabilities of group membership and genotypic values are estimated for observations in $A_1A_1B_2B_2$ and $A_2A_2B_1B_1$ marker genotypic classes – the unclassified observations (McLachlan and Basford 1988). Two types of errors arise when assigning genotypic values in the no-double-crossover model – those due to double-crossovers in noncombinant classes ($A_1A_1B_1B_1$ and $A_2A_2B_2B_2$) and those due to misclassification in recombinant classes ($A_1A_1B_2B_2$ and $A_2A_2B_1B_1$). Misclassification errors, as was implied in the discussion above, are greater than errors caused by double-crossovers. This is where the efficiency difference arises between $L_0$ and $L_2$. In the double-crossover model, where $L_0$ is used, every observation must be classified.

The mapping algorithm we developed uses maximum likelihood methods based on the EM algorithm to estimate normal distribution mixture model parameters (McLachlan and Basford 1988). Because expected mixing weights or proportions are equivalent to expected segregation ratios for QTL genotypes, classified observations supply information about $\psi$. In our algorithm, observations in the $A_1A_1B_1B_1$ and $A_2A_2B_2B_2$ marker genotypic classes are used to estimate means, variances, and mixing weights of the $Q_1Q_1$ and $Q_2Q_2$ genotypic classes. These estimates are used as starting values in the EM algorithm. Quantitative trait locus means and variances, mixing weights, posterior probabilities of group membership, and genotypic values are subsequently estimated using an implementation of the EM algorithm (McLachlan and Basford 1988). The genotypic values are directly used to estimate recombination frequencies between marker and quantitative trait loci. We are developing software (GENEMAP) using these methods. This software was used for the estimation example described in the next section.

### *$F_2$ and $F_3$ progeny genetic models*

$F_2$ progeny are produced by self-fertilizing the $F_1$. $F_3$ lines are produced by self-fertilizing $F_2$ individuals and maintaining individual $F_3$ lines. $F_2$ individuals and $F_3$ lines have identical expected values when dependent variables are observed from bulks of individuals within lines. Expected values of means of $F_2$ and $F_3$ progeny marker genotype means are functions of $r_1$, $r_2$, $\mu_{11}$, $\mu_{12}$ and $\mu_{22}$. Let $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$, $\theta_5$, $\theta_6$, $\theta_7$, $\theta_8$, and $\theta_9$ denote expected values of the means of $A_1A_1B_1B_1$, $A_1A_1B_1B_2$, $A_1A_1B_2B_2$, $A_1A_2B_1B_1$, $A_1A_2B_1B_2$, $A_1A_2B_2B_2$, $A_2A_2B_1B_1$, $A_2A_2B_1B_2$, and $A_2A_2B_2B_2$ marker genotypes, respectively.

No double-crossover gamete frequencies (Table 4) were used to derive the expected values of $F_2$ progeny marker genotype means (Table 5). Their derivation is described in the Appendix.

### *$F_2$ and $F_3$ progeny linear and nonlinear models*

Just as the expected values in Table 2 led to nonlinear model (1) with parameters $\mu_{11}$, $\mu_{22}$, $r_1$, and $r_2$, the expected values in Table 5 led to a nonlinear model with parameters $\mu_{11}$, $\mu_{12}$, $\mu_{22}$, $r_1$, and $r_2$. This model, putting $r = r_1 + r_2$ and $\varrho = r_1/r$, is

$$\begin{aligned} y = {} & \mu_{11}\, x_1 + [(1-\varrho)\, \mu_{11} + \varrho\, \mu_{12}]\, x_2 \\ & + [(1-\varrho)^2\, \mu_{11} + 2(1-\varrho)\, \varrho\, \mu_{12} + \varrho^2\, \mu_{22}]\, x_3 \\ & + [\varrho\, \mu_{11} + (1-\varrho)\, \mu_{12}]\, x_4 + c^{-1}\{(1-\varrho)\, \varrho(\mu_{11} + \mu_{22}) \\ & + [c - 2(1-\varrho)\, \varrho]\, \mu_{12}\}\, x_5 + [(1-\varrho)\, \mu_{12} + \varrho\, \mu_{22}]\, x_6 \\ & + [\varrho^2\, \mu_{11} + 2(1-\varrho)\, \varrho\, \mu_{12} + (1-\varrho)^2\, \mu_{22}]\, x_7 \\ & + [\varrho\, \mu_{12} + (1-\varrho)\, \mu_{22}]\, x_8 + \mu_{22}\, x_9 + e \qquad (8) \end{aligned}$$

**Table 4.** Genotypes and gamete frequencies for $F_2$ progeny derived from an $F_1$ ($A_1A_2Q_1Q_2B_1B_2$) based on no double crossovers

| Genotype | Frequency[a] |
|---|---|
| $A_1A_1Q_1Q_1B_1B_1$ | $[1/2(1-r_1-r_2)]^2$ |
| $A_1A_1Q_1Q_2B_1B_1$ | 0 |
| $A_1A_1Q_2Q_2B_1B_1$ | 0 |
| $A_1A_1Q_1Q_1B_1B_2$ | $2[1/2(1-r_1-r_2)1/2r_2]$ |
| $A_1A_1Q_1Q_2B_1B_2$ | $2[1/2(1-r_1-r_2)1/2r_1]$ |
| $A_1A_1Q_2Q_2B_1B_2$ | 0 |
| $A_1A_1Q_1Q_1B_2B_2$ | $(1/2r_1)^2$ |
| $A_1A_1Q_1Q_2B_2B_2$ | $2(1/2r_1\,1/2r_2)$ |
| $A_1A_1Q_2Q_2B_2B_2$ | $(1/2r_1)^2$ |
| $A_1A_2Q_1Q_1B_1B_1$ | $2[1/2(1-r_1-r_2)1/2r_1]$ |
| $A_1A_2Q_1Q_2B_1B_1$ | $2[1/2(1-r_1-r_2)1/2r_2]$ |
| $A_1A_2Q_2Q_2B_1B_1$ | 0 |
| $A_1A_2Q_1Q_1B_1B_2$ | $2[1/2r_1\,1/2r_2]$ |
| $A_1A_2Q_1Q_2B_1B_2$ | $2[1/2(1-r_1-r_2)]^2+2(1/2r_1)^2+2(1/2r_2)^2$ |
| $A_1A_2Q_2Q_2B_1B_2$ | $2(1/2r_1\,1/2r_2)$ |
| $A_1A_2Q_1Q_1B_2B_2$ | 0 |
| $A_1A_2Q_1Q_2B_2B_2$ | $2[1/2(1-r_1-r_2)1/2r_2]$ |
| $A_1A_2Q_2Q_2B_2B_2$ | $2[1/2(1-r_1-r_2)1/2r_1]$ |
| $A_2A_2Q_1Q_1B_1B_1$ | $[1/2r_1]^2$ |
| $A_2A_2Q_1Q_2B_1B_1$ | $2(1/2r_1\,1/2r_2)$ |
| $A_2A_2Q_2Q_2B_1B_1$ | $(1/2r_2)^2$ |
| $A_2A_2Q_1Q_1B_1B_2$ | 0 |
| $A_2A_2Q_1Q_2B_1B_2$ | $2[1/2(1-r_1-r_2)1/2r_1]$ |
| $A_2A_2Q_2Q_2B_1B_2$ | $2[1/2(1-r_1-r_2)1/2r_2]$ |
| $A_2A_2Q_1Q_1B_2B_2$ | 0 |
| $A_2A_2Q_1Q_2B_2B_2$ | 0 |
| $A_2A_2Q_2Q_2B_2B_2$ | $[1/2(1-r_1-r_2)]^2$ |

[a] $r_1$ and $r_2$ are recombination frequencies between $A$ and $Q$ and $B$ and $Q$, respectively

where $c=(r^{-1}-1)^2+1$, $y$ is a dependent variable, $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$, and $x_9$ are dummy variables indexing marker genotypic classes, and $e$ is a random error. The parameters $\mu_{11}$, $\mu_{12}$, $\mu_{22}$, $\varrho$, and $c$ of Eq. (8) satisfy the constraints $0\leq\varrho\leq 1$ and $1\leq c\leq\infty$. These parameters are identifiable and estimable. $r$ may be estimated and substituted in Eq. (8), thus reducing the parameters to $\mu_{11}$, $\mu_{12}$, $\mu_{22}$, and $\varrho$.

The linear model

$$y=\theta_1 x_1+\theta_2 x_2+\theta_3 x_3+\theta_4 x_4+\theta_5 x_5+\theta_6 x_6+\theta_7 x_7 + \theta_8 x_8+\theta_9 x_9+e \tag{9}$$

is useful because functions of the $\hat\theta_i$ can be used to estimate the means of QTL genotypes, even the mean of the $Q_1Q_2$ genotypic class, and $r_1$ and $r_2$. $\hat\theta_1$ and $\hat\theta_9$ can be used to estimate $\mu_{11}$ and $\mu_{22}$, respectively, but, other than $\hat\theta_5$, there is no obvious estimator of $\mu_{12}$. $\hat\theta_5$, however, is a genetically biased estimator of $\mu_{12}$ (Table 5). Because the number of $Q_1Q_1$ and $Q_2Q_2$ progeny in the $A_1A_2B_1B_2$ marker genotypic class decreases as $r$ decreases, the size of the bias, which is often negligible, decreases as $r$ decreases.

We found a genetically unbiased estimator of $\mu_{12}$ using linear functions of the $\theta_i$ (expected values of marker classes). If the coefficients of Eq. (9) are equated with the coefficients of Eq. (8), then

$$\theta_2+\theta_4+\theta_6+\theta_8=\mu_{11}+2\mu_{12}+\mu_{22};$$

hence, a genetically unbiased estimator of $\mu_{12}$ is

$$\hat\mu_{12}=(\hat\theta_2+\hat\theta_4+\hat\theta_6+\hat\theta_8-\hat\theta_1-\hat\theta_9)/2. \tag{10}$$

Thus, $2\hat\theta_1-\hat\theta_2-\hat\theta_4-\hat\theta_6-\hat\theta_8+2\hat\theta_9$ and $\hat\theta_1-2\hat\theta_5+\hat\theta_9$ may be used to estimate the contrast $\mu_{11}+\mu_{22}-2\mu_{12}$.

We found an estimator of $\varrho$ based on linear functions of the $\theta_i$ of Eq. (9):

$$\hat\varrho=\frac{-\hat\theta_1+\hat\theta_2+\hat\theta_8-\hat\theta_9}{-2\hat\theta_1+\hat\theta_2+\hat\theta_4+\hat\theta_6+\hat\theta_8-2\hat\theta_9};$$

thus, $r_1$ and $r_2$ can be estimated by putting $\hat r_1=\hat\varrho\,\hat r$ and $\hat r_2=\hat r-\hat r_1$.

The extension of the mixture model to the $F_2$ model is straightforward. There are, nevertheless, differences between the $F_2$ and BC progeny group models. In the $F_2$ model, for example, there are categorized observations for the $Q_1Q_1$ and $Q_2Q_2$ genotypic classes. We propose using the mean and variance of the $A_1A_2B_1B_2$ marker class as initial estimates of the mean and variance of the $Q_1Q_2$ genotypic class, even though a fraction of the progeny are $Q_1Q_1$ and $Q_2Q_2$. This allows for the use of $L_2$.

## Discussion

We used simulated doubled haploid data to illustrate the main features of the nonlinear model analysis. The parameter values used to simulate the data were $r_1=0.1$, $r_2=0.2$, $\mu_{11}=50.0$, $\mu_{22}=51.5$, $\sigma^2=4$, and $n=100$. The program used for this example and programs for other nonlinear models have been compiled (Knapp 1989; Knapp and Bridges 1990). Wald-statistics ($W$) were used for hypothesis tests (Gallant 1987). Reciprocals of marker genotypic class variances were used to define the weight matrix $\Sigma^-$, and $\hat\mu_{11/11}$, $\hat\mu_{22/22}$, and $\hat r/2$ were used as starting values.

The difference between QTL genotype means was approximately one standard deviation (Table 6). The hypothesis of no additive effect of the QTL was rejected ($W=5.5$ and $p=0.021$). In addition to this hypothesis, we tested $H_0{:}r_1=0$. $W$ for this test was 0.3 ($p=0.58$); thus, we failed to reject the null hypothesis. $r_1$ was inefficiently estimated in this example. The $r_1$ interval estimate covered the entire parameter space (Table 6). This example illustrates an important characteristic about the nonlinear model analysis – recombination frequencies are less efficiently estimated than means.

**Table 5.** Marker locus genotypes, QTL genotype mixtures, and expected values of marker genotypes means for $F_2$ progeny based on the no-double-crossover flanking marker model

| Marker locus genotype | QTL genotype mixture | Expected value of marker genotype mean[a] |
|---|---|---|
| $A_1A_1B_1B_1$ | $Q_1Q_1$ | $\theta_1=\mu_{11}$ |
| $A_1A_1B_1B_2$ | $Q_1Q_1+Q_1Q_2$ | $\theta_2=\frac{\mu_{11}r_2+\mu_{12}r_1}{r}$ |
| $A_1A_1B_2B_2$ | $Q_1Q_1+Q_1Q_2+Q_2Q_2$ | $\theta_3=\frac{\mu_{11}r_2^2+\mu_{12}2r_1r_2+\mu_{22}r_1^2}{r^2}$ |
| $A_1A_2B_1B_1$ | $Q_1Q_1+Q_1Q_2$ | $\theta_4=\frac{\mu_{11}r_1+\mu_{12}r_2}{r}$ |
| $A_1A_2B_1B_2$ | $Q_1Q_1+Q_1Q_2+Q_2Q_2$ | $\theta_5=\frac{r_1r_2\mu_{11}+r_1r_2\mu_{22}}{(1-2r+2r^2)}+\frac{[(1-r)^2+r_1^2+(r-r_1)^2]\,\mu_{12}}{(1-2r+2r^2)}$ |
| $A_1A_2B_2B_2$ | $Q_1Q_2+Q_2Q_2$ | $\theta_6=\frac{\mu_{12}r_2+\mu_{22}r_1}{r}$ |
| $A_2A_2B_1B_1$ | $Q_1Q_1+Q_1Q_2+Q_2Q_2$ | $\theta_7=\frac{\mu_{11}r_1^2+\mu_{12}2r_1r_2+\mu_{22}r^2}{r^2}$ |
| $A_2A_2B_1B_2$ | $Q_1Q_2+Q_2Q_2$ | $\theta_8=\frac{\mu_{12}r_1+\mu_{22}r_2}{r}$ |
| $A_2A_2B_2B_2$ | $Q_2Q_2$ | $\theta_9=\mu_{22}$ |

[a] $r=r_1+r_2$ and $r_1$, $r_2$, and $r$ are recombination frequencies between $A$ and $Q$, $B$ and $Q$, and $A$ and $B$, respectively. $\mu_{11}$, $\mu_{12}$ and $\mu_{22}$ are means of $Q_1Q_1$, $Q_1Q_2$, and $Q_2Q_2$ genotypes, respectively

**Table 6.** Maximum likelihood estimates of doubled haploid QTL parameters for a simulated example. The Gauss-Newton algorithm was used to estimate the parameters of model (3). The parameter values used to simulate the data were $r_1=0.1$, $r_2=0.2$, $\mu_{11}=50.0$, $\mu_{22}=51.5$, $\sigma^2=4.0$, and $n=100$ where $n$ is the number of doubled haploid lines. The estimated variance, coefficient of determination, and recombination frequency between marker loci $A$ and $B$ were $\hat{\sigma}^2=3.91$, $R^2=0.113$, and $\hat{r}=0.25$, respectively

| Parameter | Estimate | Standard error | Interval estimate[a] |
|---|---|---|---|
| $\mu_{11}$ | 50.07 | 0.29 | 49.5, 50.6 |
| $\mu_{22}$ | 51.17 | 0.33 | 50.5, 51.8 |
| $r_1$ | 0.057 | 0.10 | −0.14, 0.26 |

[a] $1-\alpha=0.95$

The difference in estimation efficiency is partly caused by the disparity in information for estimating these parameters. The number of observations making up the $A_1A_1B_2B_2$ and $A_2A_2B_1B_1$ classes decreases as $r$ decreases. The $A_1A_1B_1B_1$ and $A_2A_2B_2B_2$ classes do not contribute information to the estimation of $r_1$; thus, information is limited to observations in the $A_1A_1B_2B_2$ and $A_2A_2B_1B_1$ classes. In our example, there were 25 observations in these classes, i.e., 25 observations out of 100 for estimating $r_1$. Thus, there is less information for estimating recombination frequencies than means. The variance of $r_1$ increases as the distance between $A$ and $B$ decreases. In contrast to recombination frequencies, QTL genotype means are efficiently estimated. They are estimated with slightly less power than expected when QTL genotypes are known. The difference in power decreases as the distance between $A$ and $B$ decreases.

We generated an error sum of squares surface by grid searching the mean and recombination frequency parameter space. This surface further illustrates the behavior of Eq. (3). An additional example was simulated for the error surface analysis using $r_1=0.1$, $r_2=0.2$, $\mu_{11}=20.0$, $\mu_{22}=22.0$, and $\sigma^2=16$. Bounds used in the grid search were $r_1=0.0$–$0.30$, $\mu_{11}=16$–$24$, and $\mu_{22}=18$–$26$. The surfaces shown in Figs. 1 and 2 were generated by fitting a second-order response surface described by

$$\begin{aligned}\mathrm{SSE}(\hat{\theta})=&\,19{,}162.94-529.05\,\mu_{11}-439.07\,\mu_{22}-199.63\,r_1\\&+13.00\,\mu_{11}^2+9.94\,\mu_{22}^2+1518.58\,r_1^2\\&+0.38\,\mu_{11}\,\mu_{22}-35.08\,\mu_{11}\,r_1+25.06\,\mu_{22}\,r_1\,.\end{aligned}$$

$r_1$ had no effect on $\mathrm{SSE}(\hat{\theta})$; however, interactions between $r_1$ and QTL means were significant (Table 7). QTL means had significant linear and quadratic effects on $\mathrm{SSE}(\hat{\theta})$ (Table 7). The shape of the surface was convex when $r_1$ was held constant (Fig. 1), but approached a stationary ridge when $\mu_{11}$ or $\mu_{22}$ was held constant (Fig. 2). The contours were nearly circular (indicative of linearity) when $r_1$ was held constant (Fig. 1). The $\mathrm{SSE}(\hat{\theta})$

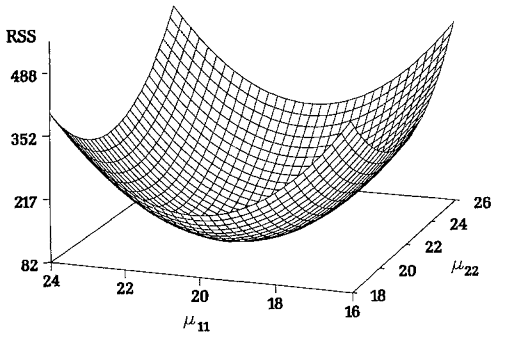

**Fig. 1.** SSE($\hat{\theta}$) surface for a simulated doubled haploid experiment. A second-order response surface ($R^2=0.97$) was estimated using SSE($\hat{\theta}$) (RSS) values from a grid search of the parameter space of $\hat{\theta}$ for a simulated doubled haploid experiment. The SSE($\hat{\theta}$) surface shown was generated by holding $r_1$ and $r_2$ constant at 0.070 and 0.180, respectively, and varying $\mu_{11}$ from 16 to 24 and $\mu_{22}$ from 18 to 26. A population of 100 doubled haploid lines was simulated using parametric values of $r_1=0.1$, $r_2=0.2$, $\mu_{11}=20$, $\mu_{22}=22$, and $\sigma^2=4$

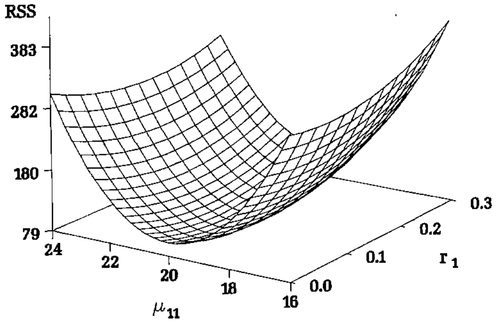

**Fig. 2.** SSE($\hat{\theta}$) surface for a simulated doubled haploid experiment. A second-order response surface ($R^2=0.97$) was estimated using SSE($\hat{\theta}$) (RSS) values from a grid search of the parameter space of $\hat{\theta}$ for a simulated doubled haploid experiment. The SSE($\hat{\theta}$) surface shown was generated by holding $\mu_{22}$ and $r_2$ constant at 21.79 and 0.180, respectively, and varying $\mu_{11}$ from 16 to 24 and $r_1$ from 0.00 to 0.30. A population of 100 doubled haploid lines was simulated using parameteric values of $r_1=0.1$, $r_2=0.2$, $\mu_{11}=20$, $\mu_{22}=22$, and $\sigma^2=4$

surface increases and the variance of $r_1$ decreases as sample size increases.

We used maximum likelihood methods (a prototype of GENEMAP) to estimate QTL means, variances, segregation ratios (mixing weights), and recombination frequencies for a simulated doubled haploid example. The parameter values used to generate data for this example were $r_1=0.1$, $r_2=0.2$, $\mu_{11}=50.0$, $\mu_{22}=55.0$ and $\sigma^2=25$. The estimators of $r_1$ and $\mu_{22}$ we propose for mixture models are the usual Mendelian estimators, e.g., for doubled haploids $\hat{r}_1=(n_1+n_2)/n$ where $n_1$ and $n_2$ are numbers of lines having genotypes $A_1A_1Q_2Q_2$ and $A_2A_2Q_1Q_1$, respectively. The usual variances of recombination frequencies may not be valid because the variances of $r_1$ and $r_2$ are functions of QTL genotypic value misclassification error, in addition to random experimental or sampling error.

**Table 7.** Analysis of a second-order response surface of residual sum of squares estimated by grid searching the parameter space of $r_1$ from 0.00 to 0.30, $\mu_{11}$ from 16 to 24, and $\mu_{22}$ from 18 to 26 for a doubled haploid experiment simulated using $r_1=0.1$, $r_2=0.2$, $\mu_{11}=20$, $\mu_{22}=22$, $\sigma^2=4$, and $n=100$ where $n$ is the number of doubled haploid lines

| Source of variation | Degrees of freedom | Mean square | $P$-value |
|---|---|---|---|
| $\mu_{11}$ Linear ($L$) | 1 | 641,573.0 | <0.0001 |
| $\mu_{22}$ $L$ | 1 | 387,007.6 | <0.0001 |
| $r_1$ $L$ | 1 | 405.8 | 0.35 |
| $\mu_{11}$ Quadratic ($Q$) | 1 | 757,552.4 | <0.0001 |
| $\mu_{22}$ $Q$ | 1 | 442,667.4 | <0.0001 |
| $r_1$ $Q$ | 1 | 23,060.8 | <0.0001 |
| $\mu_{11}$ $L\times\mu_{22}$ $L$ | 1 | 922.4 | 0.16 |
| $\mu_{11}$ $L\times r_1$ $Q$ | 1 | 12,305.6 | <0.0001 |
| $\mu_{22}$ $Q\times r_1$ $L$ | 1 | 6,280.5 | 0.0004 |
| Residual | 90 | 462.4 | |

Likelihoods were estimated using $g=1$ and $g=2$ where $g$ is the number of groups. We tested the hypothesis $H_0\colon g=1$ against $H_1\colon g=2$ using the log likelihood ratio. This statistic is approximately distributed $\chi^2_{df}$ where $df$ is the degrees of freedom for the test (Titterington et al. 1985; McLachlan and Basford 1988). Degrees of freedom is approximately equal to two times the difference in the number of parameters in the two tests excluding mixing weights (Titterington et al. 1985); $df=4$ for our example. The log likelihoods for $H_0$ and $H_1$ were $-287.6$ and $-244.8$, respectively; thus, the likelihood ratio was 85.6. The probability of this value arising by chance is less than 0.0001; thus, we rejected $H_0$ in favor of $H_1$. The evidence supports the existence of a QTL.

Estimated QTL genotypic values were used to estimate $r_1$ and $r_2$. The estimates were $\hat{r}_1=0.09$ and $\hat{r}_2=0.21$. The statistical properties, distribution, and variances of these estimators are not known. Numerical studies are needed to investigate these properties and methods for estimating variances of mixture model recombination frequency estimators, e.g., bootstrapping.

Differences in power for estimating the effects of QTL are predictable. The power of the linear, nonlinear, and mixture model methods described in this paper and the linear model method described by Lander and Botstein (1989) are nearly equivalent, although the power of methods based on estimated QTL genotypic values, i.e., the interval mapping method of Lander and Botstein (1989) and our mixture model method is often slightly less than the power of the linear and nonlinear model methods described in this paper. This happens because QTL genotypic values are estimated with error, and this error is expected to decrease power.

The power of the backcross group linear model (5) is as great or greater than the power of a linear model where missing genotypic values are estimated. Suppose the model

$$y=\mu+\tau_i+e_{ij} \tag{11}$$

is used where $y$ is the dependent variable, $\mu$ is the population mean, $\tau_i$ is the effect of the $i^{\text{th}}$ QTL genotype, $e_{ij}$ is the random error of the $j^{\text{th}}$ line of the $i^{\text{th}}$ genotype, and missing QTL genotypic values are estimated (Lander and Botstein 1989). If there are no double-crossovers, then independent variable values (QTL genotypic values) are missing for $y$ having recombinant marker phenotypes but not for $y$ having nonrecombinant marker phenotypes. The residual degrees of freedom ($df_e$) of Eq. (5) is $n-3$ because $\mu_{11}$, $\mu_{22}$, and $\theta_2$ are estimated, whereas $df_e$ of Eq. (11) is $n-3$ because $\mu$, $\mu_{11}$, and $\mu_{22}$ are estimated. Because $n$ is equal for Eqs. (5) and (11), the power of these methods is nearly equal; however, the power of Eq. (11) is less than the power of Eq. (5) because of the misclassification problem.

This rationale holds for $F_2$ progeny if marker classes other than the parental and double heterozygote classes are pooled and used to estimate a fourth parameter $\theta_p$, where $\theta_p$ is the mean of the pooled class. $\theta_p$ has no particular biological meaning, but including observations from the marker classes used to estimate $\theta_p$ increases $n$ and $df_e$. $df_e$ is $n-4$ because $\mu_{11}$, $\mu_{12}$, $\mu_{22}$, and $\theta$ are estimated. $df_e$ of Eq. (11) is $n-4$ as well because $\mu$, $\mu_{11}$, $\mu_{12}$, and $\mu_{22}$ are estimated. Thus, the difference in power of these methods is minor and probably unimportant. What is not clear is the efficiency of different recombination frequency estimators. Numerical studies are needed to address this.

*Acknowledgements.* We are grateful to D. Solmon and B. Fein in the Department of Mathematics at OSU for letting us use their Sun workstation, and to the OSU College of Agriculture for funding our use of OSU computers. Oregon Agricultural Experiment Station Technical Paper No. 8875.

## Appendix

### *Expected values of marker genotype means*

Expected values of the means of marker genotypes are functions of relative frequencies of QTL genotypes within marker genotypic classes. For example, frequencies of $Q_1Q_1$, $Q_1Q_2$, and $Q_2Q_2$ genotypes (Table 2) within the $A_1A_1B_1B_2$ genotypic class, assuming no double-crossovers, are $1/2(1-r_1-r_2)\,r_2$, $1/2(1-r_1-r_2)\,r_1$, and 0, respectively; therefore, the expected value of the mean is

$$\theta_2=\frac{[1/2(1-r_1-r_2)\,r_2]\,\mu_{11}+[1/2(1-r_1-r_2)\,r_1]\,\mu_{12}}{1/2(1-r_1-r_2)\,r_2+1/2(1-r_1-r_2)\,1/2r_1}$$

$$=\frac{(1-r_1-r_2)\,r_2\,\mu_{11}+(1-r_1-r_2)\,r_1\,\mu_{12}}{(1-r_1-r_2)\,r_2+(1-r_1-r_2)\,r_1}.$$

The expected value of $\theta_2$, substituting $r-r_1$ for $r_2$, is

$$\frac{[1-r_1-(r-r_1)]\,(r-r_1)\,\mu_{11}+[1-r_1-(r-r_1)]\,r_1\,\mu_{12}}{[1-r_1-(r-r_1)](r-r_1)+[1-r_1-(r-r_1)]\,r_1}$$

$$=\frac{(1-r)(r-r_1)\,\mu_{11}+(1-r)\,r_1\,\mu_{12}}{(1-r)(r-r_1)+(1-r)\,r_1}=\frac{(r-r_1)\,\mu_{11}+r_1\,\mu_{12}}{r}.$$

Similarly, frequencies of $Q_1Q_1$, $Q_1Q_2$, and $Q_2Q_2$ genotypes within the $A_1A_1B_2B_2$ genotypic class, assuming no double-crossovers, are $(1/2r_2)^2$, $1/2r_1r_2$, and $(1/2\,r_1)^2$, respectively; therefore, the expected value of the mean is

$$\theta_3=\frac{1/4r_2^2\,\mu_{11}+1/2r_1r_2\,\mu_{12}+1/4r_1^2\,\mu_{22}}{1/4r_2^2+1/2r_1\,r_2+1/4r_1^2}$$

$$=\frac{r_2^2\,\mu_{11}+2r_1r_2\,\mu_{12}+r_1^2\,\mu_{22}}{r_2^2+2r_1\,r_2+r_1^2}$$

$$=\frac{(r-r_1)^2\,\mu_{11}+2(r-r_1)\,r_1\,\mu_{12}+r_1^2\,\mu_{22}}{(r-r_1)^2+2(r-r_1)\,r_1+r_1^2}$$

$$=\frac{r_2^2\,\mu_{11}+2r_1r_2\,\mu_{12}+r_1^2\,\mu_{22}}{r^2}.$$

### *DH and $F_2$ nonlinear model derivatives*

As explained above, partial first-order derivatives of Eqs. (3) and (8) are needed to implement Gauss-Newton or Marquardt algorithms (Gallant 1987). For DH model (3), they are

$$\partial/\partial\mu_{11}=x_1+x_3+(r-r_1)\,(x_2-x_3)/r$$

$$\partial/\partial\mu_{22}=x_3+x_4+r_1\,(x_2-x_3)/r$$

$$\partial/\partial r_1\ \ =(\mu_{22}-\mu_{11})\,(x_2-x_3)/r.$$

For $F_2$ model (8), they are

$$\begin{aligned}\partial/\partial\mu_{11}=\,&x_1-x_2(r_1-r)/r+x_3(r^2-2rr_1+r_1^2)/r+x_4r_1/r\\&-x_5(r_1^2-rr_1)/(1-2r+2r^2)+x_7r_1^2/r^2,\end{aligned}$$

$$\begin{aligned}\partial/\partial\mu_{12}=\,&x_2\,r_1/r+x_3(2rr_1-2r_1^2)/r^2+x_4(r-r_1)/r-x_6(r_1-r)/r\\&-x_5(2r-2r^2-2r_1-1+2r_1r)/(1-2r+2\,r^2)\\&+x_7(2rr_1-2r_1^2)/r^2+x_8\,r_1/r,\end{aligned}$$

$$\begin{aligned}\partial/\partial\mu_{22}=\,&x_9-x_8(r_1-r)/r+x_7(r^2-2rr_1+r_1^2)/r^2+x_6r_1/r\\&-x_5(r_1^2-rr_1)/(1-2r+2r^2)+x_3r_1^2/r^2,\end{aligned}$$

$$\begin{aligned}\partial/\partial r_1\ \ =\,&[(x_4-x_2)\,(\mu_{11}-\mu_{12})]/r\\&+x_3(2r_1\,\mu_{11}-2r\,\mu_{11}-4r_1\,\mu_{12}+2r\,\mu_{12}+2r_1\,\mu_{22})/r^2\\&+\frac{x_5(2r_1\,\mu_{11}-r\,\mu_{11}-4r_1\,\mu_{12}+2r\,\mu_{12}+2r_1\,\mu_{22}-r\,\mu_{22})}{1-2r^2+2r^2}\\&+[(x_8-x_6)\,(\mu_{12}-\mu_{22})]/r\\&+x_7(2r_1\,\mu_{11}-2r\,\mu_{12}-4r_1\,\mu_{12}-2r\,\mu_{22}+2r_1\,\mu_{22})/r^2.\end{aligned}$$

## References

Allard RW (1956) Formulas and tables to facilitate the calculation of recombination values in heredity. Hilgardia 24:235–278

Cowen NM (1988) The use of replicated progenies in marker-based mapping of QTLs. Theor Appl Genet 75:857–862

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

Edwards MD, Stuber CW, Wendel JF (1987) Molecular marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, distribution, and types of gene action. Genetics 116:113–125

Gallant RA (1987) Nonlinear statistical models. Wiley, New York

Haldane JBS, Waddington CH (1931) Inbreeding and linkage. Genetics 16:357–374

Jennrich RI, Ralston ML (1978) Fitting nonlinear models to data. Tech Rept 46. BMDP, Los Angeles

Knapp SJ (1989) Quasi-Mendelian analysis of quantitative trait loci using molecular marker linkage maps: an overview of parameter estimation methods. In: Roebellen G (ed) Proc 12th Eucarpia Congr, Goettingen, FRG, pp 51–67

Knapp SJ, Bridges WC (1990) Programs for mapping quantitative trait loci using flanking markers and nonlinear models. J Hered (in press)

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

Little RJA, Rubin DB (1987) Stastical analysis with missing data. Wiley, New York

McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York

Osborne TC, Alexander DC, Fobes JF (1987) Identification of restriction fragment length polymorphisms linked to genes controlling soluble solids content in tomato fruit. Theor Appl Genet 73:350–356

Soller M, Brody T (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–59

Soller M, Brody T, Genizi A (1979) The expected distribution of marker-linked quantitative effects in crosses between inbred lines. Heredity 43:179–190

Stuber CW, Edwards MD, Wendel JF (1987) Molecular marker-facilitated investigations of quantitative trait loci. II. Factors influencing yield and its component traits. Crop Sci 27:639–648

Tanksley SD, Hewitt J (1988) Use of molecular markers in breeding for soluble solids content in tomato – a reexamination. Theor Appl Genet 75:811–823

Titterington DM, Smith AFM, Markow UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640

Weller JI (1987) Mapping and analysis of quantitative trait loci in *Lycopersicon* (tomato) with the aid of genetic markers using approximate maximum likelihood methods. Heredity 59:413–421

Young ND, Zamir D, Ganal MW, Tanksley SD (1988) Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the *Tm-2a* gene in tomato. Genetics 120:579–585