Genet. Res., Camb. (2001), **78**, *pp.* 79–92. *With 5 figures.* © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005092 Printed in the United Kingdom

Detection and localization of a single binary trait locus in experimental populations

LAUREN M. McINTYRE[1, 2, 3]*, CYNTHIA J. COFFMAN[3, 4] AND R. W. DOERGE[2, 5]
[1] *Computational Genomics, Purdue University, West Lafayette, IN 47907, USA*
[2] *Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA*
[3] *Duke University Medical Center, Division of Biometry, Durham, NC 27710, USA*
[4] *Institute for Clinical and Epidemiological Research Biostatistics Unit, Durham VA Medical Center (152), Durham, NC 27705, USA*
[5] *Department of Statistics, 1399 Mathematical Science Building, Purdue University, West Lafayette, IN 47907, USA*

(*Received 10 April 2000 and in revised form 19 December 2000*)

Summary

The advancements made in molecular technology coupled with statistical methodology have led to the successful detection and location of genomic regions (quantitative trait loci; QTL) associated with quantitative traits. Binary traits (e.g. susceptibility/resistance), while not quantitative in nature, are equally important for the purpose of detecting and locating significant associations with genomic regions. Existing interval regression methods used in binary trait analysis are adapted from quantitative trait analysis and the tests for regression coefficients are tests of effect, not detection. Additionally, estimates of recombination that fail to take into account varying penetrance perform poorly when penetrance is incomplete. In this work a complete probability model for binary trait data is developed allowing for unbiased estimation of both penetrance and recombination between a genetic marker locus and a binary trait locus for backcross and F_2 experimental designs. The regression model is reparameterized allowing for tests of detection. Extensive simulations were conducted to assess the performance of estimation and testing in the proposed parameterization. The proposed parameterization was compared with interval regression via simulation. The results indicate that our parameterization shows equivalent estimation capabilities, requires less computational effort and works well with only a single marker.

1. Introduction

Statistical methods for mapping continuous traits have advanced from methods that consider a single marker and single trait (Wright, 1952; Mérat, 1968; Hammond & James, 1970; O'Donald, 1971; Fain, 1978; Weller, 1986; Beckman & Soller, 1988; Luo & Kearsey, 1989; Luo & Woolliams, 1993) to methods which map multiple quantitative trait loci (QTL) using information from many markers (Jansen, 1992, 1993; Zeng, 1993, 1994). These likelihood interval mapping methods have been extended to handle epistatic interactions between genes (Kao, 1995). Churchill & Doerge (1994) and Doerge & Churchill (1996) have introduced permutation testing as a mechanism for dealing with violations of assumptions. Haley & Knott (1992) and Martinez & Curnow (1992) proposed regression-based methods that are easier to implement and require less computational effort than the likelihood interval mapping methods and have been shown to be comparable to the likelihood methods (Xu, 1996). For comprehensive reviews of these statistical methods, as well as an overview of the tissues involved in searching for genes, see Doerge *et al.* (1997), Elston (1998) and Olson *et al.* (1999).

Binary traits (e.g. susceptibility/resistance), while not quantitative in nature, are equally important for the purpose of detecting and locating significant associations with genomic regions. Much of the complication in binary traits comes from their seemingly simple descriptions (e.g., presence or absence), when in fact their underlying biological model may be more complicated. One approach to the analysis of binary trait data is to consider the trait as a genetic marker (Paterson, 1998), and then map the trait using genetic mapping methodology. Similarly,

* Corresponding author. Department of Agronomy, 1150 Lilly Hall of Science, Purdue University, West Lafayette, IN 47907, USA. Tel: +1 (765) 494 4773. Fax: +1 (765) 496 2926. e-mail: lmcintyre@purdue.edu

simple χ^2 tests (e.g. Wilcox, 1995) using 2×2 associations have been used to test for associations between a single marker and a binary trait. While both these approaches provide information on the association of a binary trait locus (BTL) near previously mapped genetic markers, they do not estimate the genetic distance between the marker and BTL (recombination) and thus are unable to estimate the proportion of individuals for an underlying genotype that express the trait (penetrance).

Recombination for BTL has been estimated by adapting QTL analysis. Hacket & Weller (1995) and Xu & Atchley (1996) used a logistic regression approach applied to ordinal or binary trait data. Visscher *et al*. (1996) used a generalized linear model, and Kruglyak & Lander (1995) proposed a non-parametric approach based on a generalization of the Wilcoxon rank test. Rebaï (1997) compared the methods proposed by Hackett & Weller (1995) and Kruglyak & Lander (1995) with the standard linear regression interval mapping methods (Haley & Knott, 1992) for analysing BTL and reported that the linear regression approach was robust against non-normality and that the loss of power was not significant. Visscher *et al*. (1996) reported similar results in their comparison of generalized linear models and linear regression in their analysis of BTL. The linear and logistic approaches gave similar results in terms of location (recombination) and power for effect, demonstrating that the estimation of recombination between binary trait loci and markers is possible using the QTL (continuous) framework. Xu (1996) compared the performance of regression and maximum likelihood approaches and found that regression-based approaches combined with permutation testing work well. In addition to this work, much has been done in human genetics, where looking at the association of multiple factors with a single binary outcome (disease status) is commonplace. Particularly relevant to the model we propose is an approach proposed by Thompson (1998) where the use of segregation indicators is described.

The existing models for binary traits in experimental populations define the hypothesis tests in terms of the effect. In addition, much of the prior work on BTL focuses on modelling an underlying threshold distribution (Xu & Atchley, 1996; Xu, 1996). The threshold model is an important quantitative genetic model; however, the underlying threshold distribution is unobserved. What is observed is the cumulative probability of the distribution from the threshold point to the limit of the distribution function, or the observed proportion of individuals with the trait. Many different types of threshold models can give the same value for the threshold, and in many cases, the appropriate underlying threshold model is unknown. Ideally, a complete methodological framework for binary traits that allows detection of BTL and estimation of both the recombination and the penetrance between the BTL and the marker locus is needed. This methodology should be expandable, easily interpretable, easily implemented and accuracy of estimates should not require estimation of the threshold model.

In this work, we develop a probability model for a binary trait locus that is based on classical genetic theory. Implementation of the model using segregation indicator variables (Thompson, 1998) is straightforward, and combined with regression techniques can be used to detect associations between the marker and the BTL even when penetrance is incomplete. The standard tests of the regression coefficients are easily interpreted using the probability model regardless of the type of regression (linear or logistic) performed. The power of the tests in both the linear and logistic settings is examined using simulations and reported for the backcross and F_2 experimental designs. The addition of a second flanking marker is considered and estimates of recombination and penetrance are developed in both single and flanking marker cases. Simulations were performed for the purpose of evaluating the performance of these estimators and comparing them to those from interval regression.

2. Methods

(i) *A probability model*

Using notation established by Doerge *et al*. (1997), genetic markers will be represented by **M** and **N** and binary trait loci (BTLs) denoted by **Q**. In the backcross and F_2 designs, for diploid individuals (Fig. 1), there are only two possible alleles for each marker and/or BTL, and they will be denoted by M_1, M_2, and Q_1, Q_2. Recombination, r_{MQ}, is the probability that an observable exchange of genetic material occurred between the BTL (Q_1 or Q_2) and the marker (M_1 or M_2). The amount of recombination is a measure of association between the marker and the BTL. When $r_{MQ} = 0{\cdot}50$, there is no association between the marker and the BTL.

In the backcross design, for each individual there are two possible marker types (MT), M_1/M_1 and M_1/M_2, and two BTL genotypes (GT), Q_1/Q_1 and Q_1/Q_2, giving a total of four possible combinations of marker type and genotype. The number of marker type and genotype combinations will be denoted by c from this point forward. In the F_2, there are three marker types and three genotypes, giving a total of nine possible combinations of marker type and genotype ($c = 9$). The initiating parents are assumed to be homozygous inbred lines differing in the binary trait of interest, meaning that each distribution of the trait for Parent_1 and Parent_2 is a different binomial distribution (see Fig. 1) such that $p_1 \neq p_2$. In addition,

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 1. *Joint and conditional probability distributions for a backcross experimental population*

MT	$P(MT)$	GT	$P(GT\|MT)$	$P(Y\|GT)$	$P(Y,GT\|MT)$	$P(Y,GT,MT)$
M_1/M_1	$\frac{1}{2}$	Q_1/Q_1	$(1-r_{MQ})$	p_1	$(1-r_{MQ})p_1$	$(1-r_{MQ})\frac{p_1}{2}$
		Q_1/Q_2	r_{MQ}	p_3	$r_{MQ}p_3$	$r_{MQ}\frac{p_3}{2}$
M_1/M_2	$\frac{1}{2}$	Q_1/Q_1	r_{MQ}	p_1	$r_{MQ}p_1$	$r_{MQ}\frac{p_1}{2}$
		Q_1/Q_2	$(1-r_{MQ})$	p_3	$(1-r_{MQ})p_3$	$(1-r_{MQ})\frac{p_3}{2}$

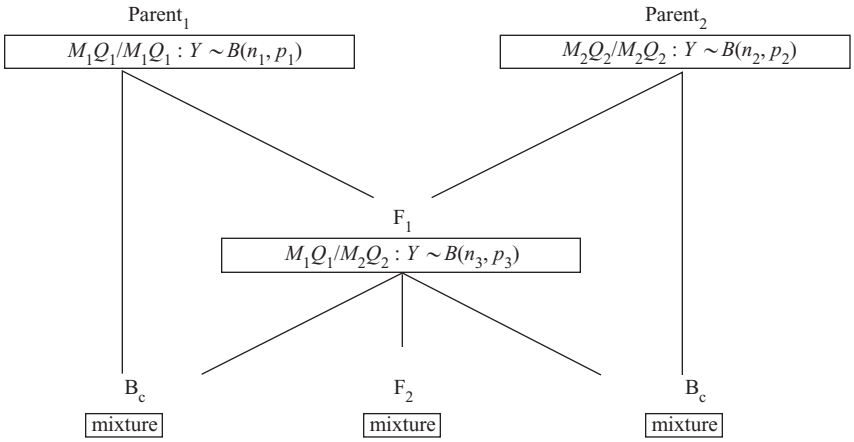

Fig. 1. Trait distributions for backcross and F_2 design.

for clarity of further discussion and without loss of generality, we assume $p_1 < p_2$. The F_1 is assumed to be distributed binomially with p_3.

The binomial probabilities p_1, p_2 and p_3 shown in Fig. 1 represent the probability that a binary trait is present given a specific BTL genotype (GT), or the penetrance of the trait for the specific genotypes Q_1/Q_1, Q_1/Q_2 and Q_2/Q_2, respectively. A dominant model is the special case where $p_2 = p_3$, and the recessive model is the special case where $p_1 = p_3$. Other genetic models can be expressed as combinations of the p_i (i.e. the midparent model is $p_3 = (p_1+p_2)/2$. In this parameterization, the dominant and recessive models are mathematical mirrors of each other.

The joint probability of the genotypes, marker types and the trait for specific experimental designs can be expressed in terms of r_{MQ}, p_1, p_2 and p_3. For the joint and conditional probability distributions of the backcross and the F_2 experimental designs see Tables 1 and 2.

(ii) *Expected values*

The joint probability of Y, GT, MT can be written as follows:

$$\begin{aligned} P(Y,GT,MT) &= P(Y|GT,MT)P(GT,MT) \\ &= P(Y|GT)P(GT,MT) \\ &= P(Y|GT)P(GT|MT)P(MT) \end{aligned}$$

where $P(Y|GT,MT) = P(Y|GT) = p_i$.

Given the joint probability of Y, GT, MT the expected values ($E(Y)$) for individuals in a backcross or F_2 population are as follows.

For the single marker, single BTL backcross model, where the F_1 is crossed with Parent$_1$,

$$\begin{aligned} E(Y) &= Y\sum_c P(Y,GT,MT) \\ &= 1\left[(1-r_{MQ})\frac{p_1}{2}+r_{MQ}\frac{p_3}{2}+r_{MQ}\frac{p_1}{2}\right. \\ &\quad \left.+(1-r_{MQ})\frac{p_3}{2}\right] \\ &\quad +0\left[1-(1-r_{MQ})\frac{p_1}{2}+1-r_{MQ}\frac{p_2}{2}+1-r_{MQ}\frac{p_1}{2}\right. \\ &\quad \left.+1-(1-r_{MQ})\frac{p_3}{2}\right] \\ &= \frac{p_1}{2}+\frac{p_3}{2}, \end{aligned}$$

and similarly for the reciprocal backcross the $E(Y) = \frac{p_2}{2}+\frac{p_3}{2}$.

In the F_2 design,

$$\begin{aligned} E(Y) &= Y\sum_c P(Y,GT,MT) \\ &= \frac{p_1}{4}+\frac{p_3}{2}+\frac{p_2}{4}. \end{aligned}$$

If Parent$_1$ and Parent$_2$ are extreme cases, such that $p_1 = 0$ and $p_2 = 1$, then in the dominant model

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 2. *Joint and conditional probability distributions for an* F_2 *experimental population*

MT	$P(MT)$	GT	$P(GT\|MT)$	$P(Y\|GT)$	$P(Y,GT\|MT)$	$P(Y,GT,MT)$
M_1/M_1	$\frac{1}{4}$	Q_1/Q_1	$(1-r_{MQ})^2$	p_1	$(1-r_{MQ})^2 p_1$	$(1-r_{MQ})^2\ \frac{p_1}{4}$
		Q_1/Q_2	$2(r_{MQ}(1-r_{MQ}))$	p_3	$2(r_{MQ}(1-r_{MQ})p_3)$	$2(r_{MQ}(1-r_{MQ})\ \frac{p_3}{4})$
		Q_2/Q_2	r_{MQ}^2	p_2	$r_{MQ}^2 p_2$	$r_{MQ}^2\ \frac{p_2}{4}$
M_1/M_2	$\frac{1}{2}$	Q_1/Q_1	$r_{MQ}(1-r_{MQ})$	p_1	$r_{MQ}(1-r_{MQ})p_1$	$r_{MQ}(1-r_{MQ})\ \frac{p_1}{2}$
		Q_1/Q_2	$(1-r_{MQ})^2$	p_3	$(1-r_{MQ})^2 p_3$	$(1-r_{MQ})^2\ \frac{p_3}{2}$
		Q_2/Q_1	r_{MQ}^2	p_3	$r_{MQ}^2 p_3$	$r_{MQ}^2 \frac{p_3}{2}$
		Q_2/Q_2	$r_{MQ}(1-r_{MQ})$	p_2	$r_{MQ}(1-r_{MQ})p_2$	$r_{MQ}(1-r_{MQ})\ \frac{p_2}{2}$
M_2/M_2	$\frac{1}{4}$	Q_1/Q_1	r_{MQ}^2	p_1	$r_{MQ}^2 p_1$	$r_{MQ}^2 \frac{p_1}{4}$
		Q_1/Q_2	$2(r_{MQ}(1-r_{MQ}))$	p_3	$2(r_{MQ}(1-r_{MQ}))p_3$	$2(r_{MQ}(1-r_{MQ})\ \frac{p_3}{4})$
		Q_2/Q_2	$(1-r_{MQ})^2$	p_2	$(1-r_{MQ})^2 p_2$	$(1-r_{MQ})^2\ \frac{p_2}{4}$

$(p_2 = p_3), E(Y) = \frac{3}{4}$ and in the recessive model $(p_1 = p_3), E(Y) = \frac{1}{4}$.

(iii) *Regression models*

Now that the probability model for BTL has been described in terms of its parameters (r_{MQ}, p_1, p_2 and p_3), it can be combined with regression techniques to detect associations between a marker and the BTL. In this framework, the test for marker and BTL association is a test of the null hypotheses, $r_{MQ} = 0{\cdot}50$. As previous work has focused on both linear and logistic regression, in this work we examined the interpretation of the tests of the regression coefficients in both a linear and logistic regression model using the probability model developed above (Tables 1, 2).

(a) *Linear model*

The linear regression model parameterized for a backcross, single BTL, single marker model is written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \quad i = 1, \ldots, n,$$

where

$$Y_i = \begin{cases} 1, & \text{Trait} = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$X_i = \begin{cases} 1, & \text{if MT} = M_1/M_2 \\ 0, & \text{if MT} = M_1/M_1. \end{cases}$$

This representation using the X_i as indicator functions allows the parameters β_0 and β_1 in the regression model to be expressed as

$$\beta_0 = \mu_{M_1/M_1}$$
$$\beta_1 = \mu_{M_1/M_2} - \mu_{M_1/M_1}.$$

However, as with any linear model the error terms are assumed to be independent and normally distributed with mean zero and constant variance σ^2. While this does not affect the estimation of the regression coefficients (Searle, 1997), it can affect the distribution of the test statistic for the test $\beta_1 = 0$. Therefore, the distribution of the test statistic and the power of the test for the regression parameters must be carefully examined. In this paper, all P values were determined using permutation theory (Churchill & Doerge, 1994; Doerge & Churchill, 1996).

Using the conditional probabilities given in Table 1, the difference between marker class means is expressed as a function of r_{MQ}, p_1 and p_3, where

$$\mu_{M_1/M_2} - \mu_{M_1/M_1} = (1-2r_{MQ})(p_3 - p_1).$$

In this case, the test of the regression coefficient $\beta_1 = 0$ is a test of $r_{MQ} = 0{\cdot}50$ and $p_1 = p_3$. Assuming that $p_1 \neq p_3$, this is a direct test of $r_{MQ} = 0{\cdot}50$.

Similarly, in an F_2 population the linear regression model is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i; \quad i = 1, \ldots, n \tag{1}$$

where

$$Y_i = \begin{cases} 1, & \text{Trait} = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$X_{1i} = \begin{cases} 1, & \text{if MT} = M_1/M_2 \\ 0, & \text{otherwise} \end{cases}$$

$$X_{2i} = \begin{cases} 1, & \text{if MT} = M_2/M_2 \\ 0, & \text{otherwise.} \end{cases}$$

Using the X_i as specific indicator functions allows the regression the parameters β_0, β_1 and β_2 to be expressed as

$$\beta_0 = \mu_{M_1/M_1},$$
$$\beta_1 = \mu_{M_1/M_2} - \mu_{M_1/M_1},$$
$$\beta_2 = \mu_{M_2/M_2} - \mu_{M_1/M_1}.$$

As with the backcross, the conditional probabilities given in Table 2 for the F_2 are used to express the

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 3. *Estimates of r_{MQ} and p_3 for single marker backcross and F_2 populations*

	$\hat{r}_{MQ}$	$\hat{p}_3$
Backcross	$\frac{p_1-\mu_{M_1/M_1}}{2p_1-\mu_{M_1/M_2}-\mu_{M_1/M_1}}$	$\mu_{M_1/M_2}+\mu_{M_1/M_1}-p_1$
F_2	$\frac{1}{2}\frac{p_1+\mu_{M_2/M_2}-\mu_{M_1/M_1}-p_2}{p_1-p_2}$	$\frac{1}{2}\mu_{M_2/M_2}+\mu_{M_1/M_2}+\frac{1}{2}\mu_{M_1/M_1}-\frac{1}{2}p_1-\frac{1}{2}p_2$

Table 4. *Expected trait distributions for binary traits in a backcross with two markers for linkage map MQN*

MT	$P(MT)$	GT	$P(GT\|MT)$	$P(Y\|GT)$	$P(Y,GT\|MT)$	$P(Y,GT,MT)$
M_1N_1/M_1N_1	$\frac{1}{2}(1-r_{MN})$	Q_1/Q_1	$\frac{(1-r_{MQ})(1-r_{NQ})}{1-r_{MN}}$	p_1	$\frac{(1-r_{MQ})(1-r_{NQ})}{1-r_{MN}}p_1$	$\frac{1}{2}(1-r_{MQ})(1-r_{NQ})p_1$
		Q_1/Q_2	$\frac{r_{MQ}r_{NQ}}{1-r_{MN}}$	p_3	$\frac{r_{MQ}r_{NQ}}{1-r_{MN}}p_3$	$\frac{1}{2}r_{MQ}r_{NQ}p_3$
M_1N_1/M_1N_2	$\frac{1}{2}r_{MN}$	Q_1/Q_1	$\frac{(1-r_{MQ})(r_{NQ})}{r_{MN}}$	p_1	$\frac{(1-r_{MQ})(r_{NQ})}{r_{MN}}p_1$	$\frac{1}{2}(1-r_{MQ})(r_{NQ})p_1$
		Q_1/Q_2	$\frac{r_{MQ}(1-r_{NQ})}{r_{MN}}$	p_3	$\frac{r_{MQ}(1-r_{NQ})}{r_{MN}}p_3$	$\frac{1}{2}r_{MQ}(1-r_{NQ})p_3$
M_1N_1/M_2N_1	$\frac{1}{2}r_{MN}$	Q_1/Q_1	$\frac{(r_{MQ})(1-r_{NQ})}{r_{MN}}$	p_1	$\frac{(r_{MQ})(1-r_{NQ})}{r_{MN}}p_1$	$\frac{1}{2}(r_{MQ})(1-r_{NQ})p_1$
		Q_1/Q_2	$\frac{(r_{NQ})(1-r_{MQ})}{r_{MN}}$	p_3	$\frac{(r_{NQ})(1-r_{MQ})}{r_{MN}}p_3$	$\frac{1}{2}(r_{NQ})(1-r_{MQ})p_3$
M_1N_1/M_2N_2	$\frac{1}{2}(1-r_{MN})$	Q_1/Q_1	$\frac{(r_{MQ})(r_{NQ})}{1-r_{MN}}$	p_1	$\frac{(r_{MQ})(r_{NQ})}{1-r_{MN}}p_1$	$\frac{1}{2}(r_{MQ})(r_{NQ})p_1$
		Q_1/Q_2	$\frac{(1-r_{MQ})(1-r_{NQ})}{1-r_{MN}}$	p_3	$\frac{(1-r_{MQ})(1-r_{NQ})}{1-r_{MN}}p_3$	$\frac{1}{2}(1-r_{MQ})(1-r_{NQ})p_3$

standard tests of the regression parameters $\beta_1 = 0$ and $\beta_2 = 0$ in terms of r_{MQ}, p_1, p_2 and p_3, where

$$\mu_{M_1/M_2}-\mu_{M_1/M_1} = (1-2r_{MQ})[r_{MQ}(p_1-2p_3+p_2) + (p_3-p_1)]$$

$$\mu_{M_2/M_2}-\mu_{M_1/M_1} = (1-2r_{MQ})(p_2-p_1).$$

In this model, the test of the regression coefficient $\beta_1 = 0$ is a test of $r_{MQ} = 0{\cdot}50$ and $p_1 = p_2 = p_3$, and $\beta_2 = 0$ is a test of $r_{MQ} = 0{\cdot}50$ and $p_1 = p_2$. Assuming that $p_1 \neq p_2$, both tests are a direct test of $r_{MQ} = 0{\cdot}50$. The test of β_2 in the F_2 is conceptually equivalent to the test of β_1 in the backcross.

(b) *Logistic model*

The logistic model has been suggested in binary trait analysis. The logistic regression model is written as:

$$Y_i = \pi(X_i)+\epsilon_i; \quad i = 1, \ldots, n$$

where for a backcross, single BTL, single marker model

$$\pi(X_i) = \frac{\exp(\beta_0+\beta_1X_i)}{1+\exp(\beta_0+\beta_1X_i)},$$

and X_i is the marker class as defined in (1).

For each trait and marker status classification we derived $\pi(X_i = 0)$ and $\pi(X_i = 1)$, giving rise to the log odds ratio

$$\beta_1 = \ln\left[\frac{\pi(X_i=1)}{1-\pi(X_i=1)}\right] \Big/ \ln\left[\frac{\pi(X_i=0)}{1-\pi(X_i=0)}\right]$$

(Hosmer & Lemeshow, 1989; Agresti, 1990). This expression lends interpretability to the model , β_1 now represents a comparative assessment of the odds among individuals with the trait ($Y = 1$) and marker genotype $X = 1$, compared with the odds among individuals with the trait and marker genotype $X = 0$.

Applying the probability model derived above (Table 1), the expected values of the log odds are

$$e^{\beta_0} = \left[\frac{p_1-r_{MQ}p_1+r_{MQ}p_3}{1-(p_1-r_{MQ}p_1+r_{MQ}p_3)}\right]$$

and

$$e^{\beta_1} = \left[\frac{r_{MQ}p_1+p_3-r_{MQ}p_3}{1-(r_{MQ}p_1+p_3-r_{MQ}p_3)}\right] \Big/ e^{\beta_0}.$$

The logistic regression model for an F_2, single BTL, single marker model is written as:

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0+\beta_1X_{1i}+\beta_2X_{2i})}{1+\exp(\beta_0+\beta_1X_{1i}+\beta_2X_{2i})}$$

where X_{1i} and X_{2i} are indicator variables as defined in (1).

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

As in the backcross, the expected log odds for M_1M_1 (e^{β_0}), M_1M_2 (e^{β_1}) and M_2M_2 (e^{β_2}) are derived from the conditional probabilities in Table 2 (see Appendix).

The tests of the regression parameters in the logistic model are tests of $\beta_i = 1$ as they are a ratio of means, compared with the linear regression which is a difference between means. The tests in the logistic regression have the same interpretation as the tests in the linear regression.

(iv) *Power*

The power of linear and logistic regression coefficient tests to detect BTL were directly compared. The tests between the regression models were evaluated in terms of their power to detect BTL under a variety of genetic models in the backcross and the F_2 experimental designs. We computed *P* values via permutation (Churchill & Doerge, 1994; Doerge & Churchill, 1996) to ensure that the type I error was valid and was comparable between the two regression strategies. As mentioned previously, with our parameterization, the null hypothesis tests of the β_i are the same in the linear and logistic models. Therefore, we considered the regression model with the higher power of the tests to be the better method of implementation.

(v) *Estimation of recombination and penetrance*

Once a BTL is detected, the distance from the marker to the BTL can be estimated. Currently, recombination can be estimated directly from the observed trait and marker data. The unadjusted estimate is $\hat{r}_{MQ} = n_r/n$, where n_r is the number of individuals with the trait and marker type that are discordant according to the genetic model, and n is the total number of individuals (Lynch & Walsh, 1998). This estimate does not allow for reduced penetrance and implicitly assumes that $p_1 = 0$ and $p_2 = 1$.

We derive moment estimators for recombination that are adjusted for penetrance in both the backcross and F_2 experimental designs assuming p_1 and/or p_2 are known. In the backcross, the moment estimators are constructed using the equations for the marker means in the backcross, where

$$\mu_{M_1/M_1} = (1-r_{MQ})p_1 + r_{MQ}p_3$$
$$\mu_{M_1/M_2} = r_{MQ}p_1 + (1-r_{MQ})p_3.$$

The first equation was solved for r_{MQ} giving the estimate

$$r_{MQ} = \frac{p_1 - \mu_{M_1/M_1}}{p_1 - p_3}.$$

We then substituted this estimate of r_{MQ} into the second equation and solved for p_3 giving the estimate $p_3 = \mu_{M_1/M_2} + \mu_{M_1/M_1} - p_1$. We then substituted this estimate back into the equation for r_{MQ}, resulting in the estimate

$$r_{MQ} = \frac{p_1 - \mu_{M_1/M_1}}{2p_1 - \mu_{M_1/M_2} - \mu_{M_1/M_1}}.$$

Similarly, moments estimates for the F_2 experimental design were constructed (Table 4).

When $p_1 = 0$ and $p_3 = 1$ in the backcross, the adjusted estimate of r_{MQ} reduces to the unadjusted estimate. Estimates of p_3 are also important as they give insight into the genetic model (Table 4). The moment estimator for p_3 was derived assuming that p_1 and p_2 were known. The estimates of r_{MQ} and p_3 depend upon p_1 and p_2, and if p_1 and p_2 are known exactly, then the moment estimators of r_{MQ} and p_3 are unbiased. However, in experimental situations the precise values of p_1 and p_2 may not be known and, therefore, we evaluated the sensitivity of the estimates of r_{MQ} and p_3 to misspecification of p_1 and p_2.

(vi) *Two markers*

We now extend our parameterization to include a second marker. For a two marker backcross design, with a map *MQN*, the joint and conditional probability distributions can be described in terms r_{MQ}, r_{NQ}, r_{MN}, p_1 and p_3. There are four observable marker genotypes, and two unobservable BTL genotypes within each marker type. The full joint and conditional probability distributions are given in Table 3.

In order to derive moments estimators we assumed r_{MN} and p_1 were known. We then used the method of moments, as described for the single marker case, to derive moment estimators for r_{MQ}, r_{NQ} and p_3. These estimators are as follows:

$$\hat{p}_3 = (\mu_{M_2N_1} + \mu_{M_1N_2} - \mu_{M_2N_2} - \mu_{M_1N_1}) \times r_{MN} + \mu_{M_2N_2} + \mu_{M_1N_1} - p_1)$$

$$\hat{r}_{MQ} = \frac{p_1 - (\mu_{M_1N_2} - \mu_{M_1N_1})r_{MN} - \mu_{M_1N_1}}{p_3 - p_1}$$

$$\hat{r}_{NQ} = \frac{p_1 - (\mu_{M_2N_1} - \mu_{M_1N_1})r_{MN} - \mu_{M_1N_1}}{p_3 - p_1}.$$

(vii) *Comparison with interval regression*

Interval regression has been described in detail by Martinez & Curow (1992) for the backcross and by Haley & Knott (1992) for the F_2. Briefly, this approach uses flanking markers to define the coefficients of the regression as mean, additive or dominance effects. For *s* steps along the interval between markers *M* and *N* values of *X* are calculated according to the conditional probability of a QTL in that location. In the backcross design, at each position in the interval, the estimated regression parameter $\hat{\beta}_1$ provides an estimate of the

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 5. *Simulation conditions for single marker backcross and* F_2 *populations*

Model	n	r_{MQ}	p_1	p_2	p_3	No. of combinations
Backcross	50, 100, 200, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·0	N/A	0·10, 0·40, 0·60, 0·80, 1·00	120
	100, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·10, 0·20	N/A	0·10, 0·40, 0·60, 0·80, 1·00	80
F_2	50, 100, 200, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·0	p_3	0·10, 0·40, 0·60, 0·80, 1·00	120
	100, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·10, 0·20	p_3	0·10, 0·40, 0·60, 0·80, 1·00	80
F_2	100, 200, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·0	0·10, 0·40, 0·60, 0·80, 1·00	p_1	120
	100, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·10, 0·20	0·10, 0·40, 0·60, 0·80, 1·00	p_1	80
F_2	100, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·10, 0·20	0·10, 0·40, 0·60, 0·80, 1·00	$\frac{p_1+p_2}{2}$	80
Total						680

additive genetic effect a for a specific r_{MQ}. Thus, the test of the regression parameter $\beta_1 = 0$ in interval regression is the test $a = 0$ for a specific r_{MQ}. This is a statistical test for an additive effect. Similarly, in the F_2, the tests of the regression coefficients in interval regression are tests of additive and dominance effect. The test $\beta_1 = 0$ is the test $a = 0$ and the test $\beta_2 = 0$ is the test $d = 0$. The test of the entire model is a test of $a+d = 0$. Using our notation, the test $a+d = 0$ corresponds to a test of $p_3 - p_1 = 0$.

The interval parameterization thus provides a mechanism to test for effect using tests of the regression parameters. In our parameterization, the regression coefficients are tests for detection. Thus, the two parameterizations have different null hypotheses for the tests of the regression coefficients and are not comparable in terms of power. However, the estimates of r_{MQ} and p_3 produced by both parameterizations are comparable.

(viii) *Simulations*

Data were simulated for the single marker backcross, two marker backcross, and single marker F_2 experimental frameworks for the cases given in Table 5. There were a total of 680 combinations of parameters simulated. For each combination of parameters, 1000 replicates of the simulation were performed. For each replicate, the null hypothesis was rejected when the empirical P value for that replicate was less than the nominal alpha, 0·05. For each simulation (set of 1000 replicates), the power of the linear model was compared with that of the logistic model using McNemar's test (Agresti, 1990). The power for each test of a regression parameter was estimated as the number of times the empirical P value for that replicate was less than 0·05 divided by the number of replicates.

For each replicate, recombination was estimated using the unadjusted $\hat{r}_{MQ} = n_r/n$ and adjusted estimates, and p_3 was estimated (Table 4). The adjusted estimates assumed p_1 and p_2 were known and the sensitivity of the adjusted r_{MQ} and p_3 estimates to misspecification of p_1 and p_2 was examined by setting p_1 and p_2 to incorrect values in the estimation of r_{MQ} and p_3. Incorrect values ranged across all possible values of p_1 and p_2 consistent with the initial assumption $p_1 < p_2$.

Two marker backcross populations were simulated for the cases given in Table 6. For each replicate, r_{MQ} and p_3 were estimated using the moment estimators derived from our single marker model, from our two marker model, and using the interval regression method. For the interval regression method, we stepped through the interval from $r_{MQ} = 0{\cdot}00$ to $r_{MQ} = \hat{r}_{MN}$ using increments of 0·005. We selected the value of r_{MQ} in the interval from 0·00 to r_{MN}

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 6. *Simulation conditions for two marker backcross populations*

Population	n	r_{MQ}	r_{NQ}	p_1	p_2	p_3	No. of combinations
Backcross	100, 500	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·00, 0·10, 0·20, 0·30, 0·40, 0·50	0·00	N/A	0·10, 0·40, 0·60, 0·80, 1·00	360

that produced the lowest value for the approximate likelihood ratio test statistic

$$\left(n * \log \frac{SSE_R}{SSE_F} \right)$$

where SSE_R is the sum of squared errors for the reduced model, and SSE_F is the sum of squared errors for the full model.

3. Results

(i) *Power*

In the backcross, power for the test of β_1 was similar in the linear and logistic regression models except when the marker was close to the BTL ($r_{MQ} < 0{\cdot}10$), the difference between the parental lines was small ($(p_3 - p_1) < 0{\cdot}4$) and the sample size was small. In these cases, the linear model had significantly better power than the logistic model ($p \leqslant 0{\cdot}05$, McNemar's test). Similarly, in the F_2 the power for the logistic model test of the β_2 was significantly lower than that of the linear model ($P \leqslant 0{\cdot}05$, McNemar's test) under similar conditions. Additionally, examining the power curves for β_2 under different values of recombination for these conditions revealed that power was not monotonic for the logistic regression (Fig. 2).

Power also depends on the absolute difference in the penetrance of the parental lines ($p_2 - p_1$). As mentioned in Section 2, we assumed that the penetrance of the parental lines was different, which made the tests of the β_i simple tests of the recombination between marker and trait. We found that the power of the tests of the $\beta_i = 0$ depended upon maximizing the difference in the parental lines (Fig. 3). The larger the difference between the parental lines, the higher the power.

We can derive the relationship between the power of the test and the difference between parental lines by examining the expected values of β_2 in the linear model. In the F_2 design, the $E(\beta_2) = (1 - 2r_{MQ})(p_2 - p_1)$. When the parental lines are equal ($p_2 - p_1 = 0$), the $E(\beta_2) = 0$ and the power of the test is the nominal alpha level specified. As the difference between the parents increases, the expected value moves further from zero for a fixed value of r_{MQ}. Correspondingly, the permuted power for the test of β_2 from our simulations increased as the difference in the parental lines increased (Fig. 3). Similarly, power for the test of β_1 in the backcross decreased monotonically as the difference ($p_3 - p_1$) decreased. Our simulations showed that as the difference in the means between parental and F_1 lines increased, power increased.

Power for tests on β_1 in the backcross and β_2 in the F_2 in the linear regression framework increased as linkage between marker and BTL increased, or as

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

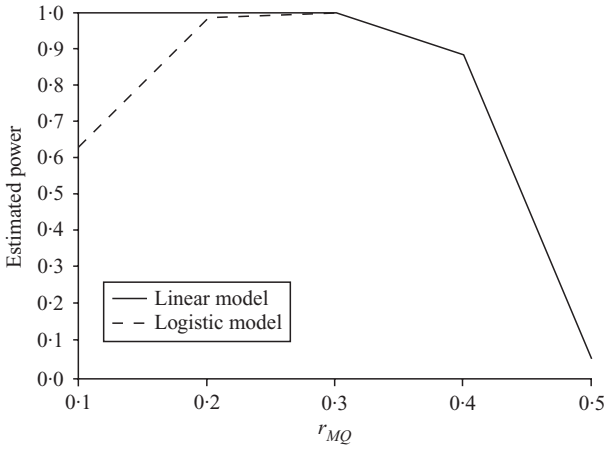

Fig. 2. Power of the test of β_1 in linear and logistic models as a function of recombination.

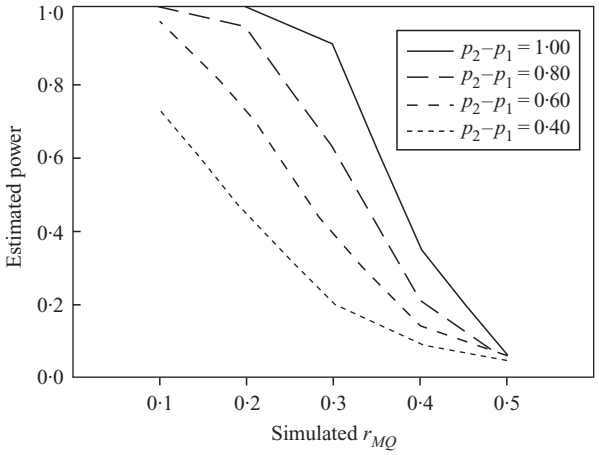

Fig. 3. Power in the linear model as a function of recombination and difference between parental lines.

recombination between marker and BTL decreased (Fig. 3). Thus, the closer a marker to the true BTL location the higher the power for detecting the BTL. Power also increased as sample size increased.

The power of the test of β_1 in the F_2 depends on the relationship between the F_1 and the parental lines which is described by the genetic model. However, if we define our indicator variables differently than what is shown in (1), the test of β_1 and the power of the test of β_1 will be different. For example, if $X_{2i} = 1$ when $MT = M_1/M_1$ instead of when $MT = M_2/M_2$, the $E(\beta_1) = \mu_{M_1/M_2} - \mu_{M_2/M_2} = (1-2r_{MQ})[r_{MQ}(p_1 - 2p_3 + p_2) + (p_3 - p_2)]$. This demonstrates the importance of careful specification of the indicator variables and the impact of the specification on power for tests of β_1 in the F_2.

(ii) *Estimates of r_{MQ} and p_3*

In the backcross, estimates of r_{MQ} unadjusted for penetrance (n_r/n) are unbiased only when $Parent_1$ had no individuals expressing the trait ($p_1 = 0$) and the F_1 had all individuals expressing the trait ($p_3 = 1$). As penetrance parameters moved away from these extreme values, the estimate of recombination using this formulation became increasingly biased (Fig. 4). In contrast, estimates of r_{MQ} adjusted for penetrance (Table 4) were unbiased.

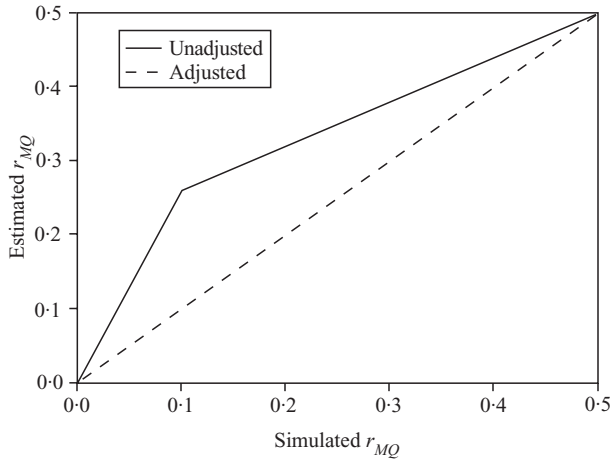

Fig. 4. Estimation of recombination unadjusted for incomplete penetrance compared with estimation of recombination adjusted for incomplete penetrance.

In the F_2, the adjusted r_{MQ} estimates (Table 4) and p_3 estimates were unbiased when p_1 and p_2 are specified correctly as shown in a subset of the simulation runs in Tables 7–10. As sample size decreased, the standard errors of the estimates for r_{MQ} and p_3 increased, and as the difference between parental lines decreased ($p_2 - p_1$), the standard errors for the estimates of r_{MQ} increased.

The estimates of both r_{MQ} and p_3 were biased when the values of p_1 and p_2 were misspecified. For estimates of r_{MQ}, if the penetrance of $Parent_1$ (p_1) was specified too low (further from p_2), then estimates were biased towards 0·50 while if p_1 is specified too high (closer to p_2), estimates were biased towards zero (Fig. 5*a*). If misspecification was large, then estimates of r_{MQ} were sometimes greater than 0·50 or less than 0·00. However, when $r_{MQ} = 0{\cdot}50$ the estimated value was approximately 0·50 regardless of the misspecification of p_1. Results were similar for misspecification of p_2 (Fig. 5*c*). In both cases, misspecification of less than approximately 20% resulted in estimates of r_{MQ} reasonably close to the true value even for samples of size 100. For estimates of p_3, if p_1 was specified too low (further from p_2), then estimates were biased towards p_2, while if p_1 was specified too high (closer to p_2), estimates were biased towards p_1 (Fig. 5*b*). Results were similar for misspecification of p_2 (Fig. 5*d*).

(iii) *Two markers*

The estimates and the standard error of the estimates of r_{MQ} were not improved by expanding our para-

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 7. *Estimates of recombination parameter* ($\hat{r}_{MQ}$) *and penetrance parameter* ($\hat{p}_3$) *for data simulated for an* F_2, *dominant model*, ($p_1 = 0$, *and* $p_2 = p_3$), $n = 100$

p_1	p_2	p_3	r_{MQ}	$\hat{p}_3$(se($\hat{p}_3$))	$\hat{r}_{MQ}$(se($\hat{r}_{MQ}$))
0·00	1·00	1·00	0·10	0·9997 (0·001866)	0·1004 (0·001341)
			0·20	0·9975 (0·002297)	0·1996 (0·001654)
			0·30	0·997 (0·002549)	0·2996 (0·001938)
			0·40	1·003 (0·002736)	0·4022 (0·002014)
			0·50	0·9947 (0·002723)	0·501 (0·002004)
0·00	0·80	0·80	0·10	0·8004 (0·002671)	0·1007 (0·002112)
			0·20	0·8013 (0·002893)	0·2048 (0·002427)
			0·30	0·8012 (0·003007)	0·3045 (0·002732)
			0·40	0·8067 (0·003154)	0·4015 (0·002793)
			0·50	0·7958 (0·003125)	0·5036 (0·002743)
0·00	0·60	0·60	0·10	0·6014 (0·002915)	0·0927 (0·003113)
			0·20	0·5945 (0·003076)	0·2084 (0·003378)
			0·30	0·5935 (0·003112)	0·3008 (0·00358)
			0·40	0·6008 (0·003163)	0·3949 (0·003675)
			0·50	0·6004 (0·00318)	0·4954 (0·003864)
0·00	0·40	0·40	0·10	0·4004 (0·002844)	0·1043 (0·004436)
			0·20	0·4041 (0·002839)	0·2063 (0·00488)
			0·30	0·3965 (0·002938)	0·3057 (0·004895)
			0·40	0·394 (0·002834)	0·4074 (0·005231)
			0·50	0·3961 (0·002762)	0·4935 (0·005119)

Table 8. *Estimates of recombination parameter* ($\hat{r}_{MQ}$) *and penetrance parameter* ($\hat{p}_3$) *for data simulated for an* F_2, *recessive model*, ($p_1 = p_3 = 0$), $n = 100$

p_1	p_2	p_3	r_{MQ}	$\hat{p}_3$(se($\hat{p}_3$))	$\hat{r}_{MQ}$ (se($\hat{r}_{MQ}$))
0·00	1·00	0·00	0·10	0·002871 (0·001821)	0·09887 (0·001331)
			0·20	−0·003263 (0·002406)	0·1994 (0·001618)
			0·30	−0·001751 (0·002611)	0·3028 (0·001842)
			0·40	2·354e-05 (0·002739)	0·4028 (0·001927)
			0·50	−0·00419 (0·002773)	0·5002 (0·001977)
0·00	0·80	0·00	0·10	0·001787 (0·001998)	0·09931 (0·00198)
			0·20	0·002596 (0·002282)	0·1984 (0·00213)
			0·30	0·003764 (0·002443)	0·2992 (0·002276)
			0·40	0·002801 (0·00241)	0·4021 (0·002206)
			0·50	0·001556 (0·002608)	0·5006 (0·002239)
0·00	0·60	0·00	0·10	−0·0007081 (0·001866)	0·1016 (0·002641)
			0·20	−0·000394 (0·002061)	0·2029 (0·002688)
			0·30	−0·001148 (0·002142)	0·3007 (0·002785)
			0·40	−0·0007598 (0·002215)	0·3964 (0·00272)
			0·50	−9·109e-05 (0·002246)	0·4992 (0·002781)
0·00	0·40	0·00	0·10	−0·0003342 (0·001709)	0·09977 (0·003703)
			0·20	0·0002425 (0·001831)	0·1963 (0·003605)
			0·30	−0·0009838 (0·001816)	0·3004 (0·00343)
			0·40	−0·001232 (0·001905)	0·404 (0·003402)
			0·50	0·0001936 (0·001873)	0·4976 (0·003295)

meterization to include the second marker (Table 11). However, it is likely that additional markers will help position the BTL to the left or right of the primary marker due to the gain in available information provided by these markers.

(iv) *Comparison with interval regression*

The estimates of r_{MQ} and the standard errors of r_{MQ} were similar between the interval regression and our estimators for both the single and two marker

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 9. *Estimates of recombination parameter* ($\hat{r}_{MQ}$) *and penetrance parameter* ($\hat{p}_3$) *for data simulated for an* F_2, $p_3 = \frac{p_1+p_2}{2}$, $n = 100$

p_1	p_2	p_3	r_{MQ}	$\hat{p}_3$(se($\hat{p}_3$))	$\hat{r}_{MQ}$ (se($\hat{r}_{MQ}$))
0·10	1·00	0·55	0·10	0·5519 (0·002689)	0·1007 (0·001667)
			0·20	0·5507 (0·002972)	0·2014 (0·002064)
			0·30	0·5449 (0·00315)	0·3021 (0·002391)
			0·40	0·549 (0·003237)	0·3992 (0·002541)
			0·50	0·5492 (0·003271)	0·5016 (0·002478)
0·10	0·80	0·45	0·10	0·4473 (0·002963)	0·1017 (0·002598)
			0·20	0·4537 (0·003002)	0·2063 (0·002893)
			0·30	0·4538 (0·003117)	0·2907 (0·003097)
			0·40	0·4521 (0·003216)	0·3979 (0·003169)
			0·50	0·4444 (0·003172)	0·497 (0·003235)
0·10	0·60	0·35	0·10	0·3496 (0·002865)	0·1006 (0·003822)
			0·20	0·3482 (0·00291)	0·201 (0·004081)
			0·30	0·3459 (0·002985)	0·2994 (0·004262)
			0·40	0·3531 (0·003113)	0·3972 (0·004255)
			0·50	0·35 (0·003052)	0·4958 (0·004339)

Table 10. *Estimates of recombination parameter* ($\hat{r}_{MQ}$) *and penetrance parameter* ($\hat{p}_3$) *for data simulated for an* F_2, $p_1 = p_3$, $n = 100$

p_1	p_2	p_3	r_{MQ}	$\hat{p}_3$(se($\hat{p}_3$))	$\hat{r}_{MQ}$ (se($\hat{r}_{MQ}$))
0·10	1·00	0·10	0·10	0·1009 (0·002401)	0·1022 (0·001785)
			0·20	0·1005 (0·00265)	0·2008 (0·002067)
			0·30	0·101 (0·002862)	0·2999 (0·002258)
			0·40	0·09833 (0·002953)	0·4012 (0·002267)
			0·50	0·09954 (0·00297)	0·4983 (0·002344)
0·10	0·80	0·10	0·10	0·1006 (0·00244)	0·09807 (0·00257)
			0·20	0·1011 (0·002732)	0·2005 (0·002673)
			0·30	0·0992 (0·00275)	0·2998 (0·00282)
			0·40	0·09831 (0·002709)	0·395 (0·002878)
			0·50	0·1007 (0·002945)	0·499 (0·003015)
0·10	0·60	0·10	0·10	0·105 (0·00241)	0·09658 (0·003651)
			0·20	0·09976 (0·002629)	0·1952 (0·003741)
			0·30	0·1021 (0·002674)	0·2951 (0·003782)
			0·40	0·09402 (0·002611)	0·4038 (0·003753)
			0·50	0·09876 (0·002642)	0·4991 (0·00376)

simulations (Table 11). Estimates of p_3 were also comparable across all three approaches. When the flanking markers were unlinked $r_{MN} = 0{\cdot}50$, the interval regression tended to produced biased results.

4. Discussion and conclusions

We have described a simple parameterization that can detect and localize BTL in plant and animal populations for backcross and F_2 experiments. The method relies on simple linear regression combined with a permutation algorithm that can be easily applied and implemented in commercially available software such as SAS or Splus with little effort.

In our analysis, using the logistic regression models, the power of the logistic regression tests of the β_2 parameter in the F_2 and the β_1 parameter in the backcross was not monotonic in all cases. In fact, the power was low precisely when evidence for linkage was the strongest, namely when r_{MQ} was small and the difference $(p_2 - p_1)$ was large. This failure occurred due to a quasi-separation of points in the logit function. When r_{MQ} was small and the difference $(p_2 - p_1)$ was large, there are very few individuals with discordant marker trait pairs, resulting in difficulty fitting the logit function. The test statistic in the logistic regression is based upon a ratio of means, while the test statistic in the linear model is based upon a difference of means. Thus, when there are few

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

Table 11. *Comparison of estimates of* r_{MQ} *and* p_3 *between the method of McIntyre* et al. *and interval regression* ($n = 100$, $r_{NQ} = 0{\cdot}10$)

			McIntyre *et al.*'s method				Interval regression	
			Single marker		Two marker		Two marker	
p_1	p_3	r_{MQ}	$\hat{r}_{MQ}$ (se($\hat{r}_{MQ}$))	$\hat{p}_3$(se($\hat{p}_3$))	$\hat{r}_{MQ}$ (se($\hat{r}_{MQ}$))	$\hat{p}_3$(se($\hat{p}_3$))	$\hat{r}_{MQ}$ (se($\hat{r}_{MQ}$))	$\hat{p}_3$(se($\hat{p}_3$))
0·00	1·00	0·20	0·197 (0·00148)	1·002 (0·0026)	0·197 (0·001374)	0·999 (0·0018)	0·198 (0·0013)	1·002 (0·0015)
		0·30	0·298 (0·00162)	0·999 (0·0029)	0·299 (0·001536)	0·9987 (0·0019)	0·301 (0·0016)	1·010 (0·0035)
		0·40	0·402 (0·00162)	0·996 (0·0031)	0·403 (0·001585)	0·9993 (0·0019)	0·402 (0·00158)	1·051 (0·0077)
		0·50	0·501 (0·00162)	0·997 (0·0032)	0·501 (0·001581)	0·9989 (0·0020)	0·470 (0·00085)	1·035 (0·0100)
0·20	0·80	0·20	0·194 (0·00306)	0·800 (0·0030)	0·195 (0·003005)	0·8000 (0·0029)	0·202 (0·00162)	0·801 (0·0022)
		0·30	0·291 (0·00299)	0·799 (0·0030)	0·293 (0·002871)	0·7994 (0·0027)	0·297 (0·00179)	0·807 (0·0026)
		0·40	0·401 (0·00271)	0·806 (0·0032)	0·401 (0·002687)	0·8035 (0·0027)	0·396 (0·00171)	0·843 (0·0050)
		0·50	0·499 (0·00271)	0·803 (0·0032)	0·499 (0·0027)	0·8012 (0·0027)	0·464 (0·001)	0·908 (0·0050)

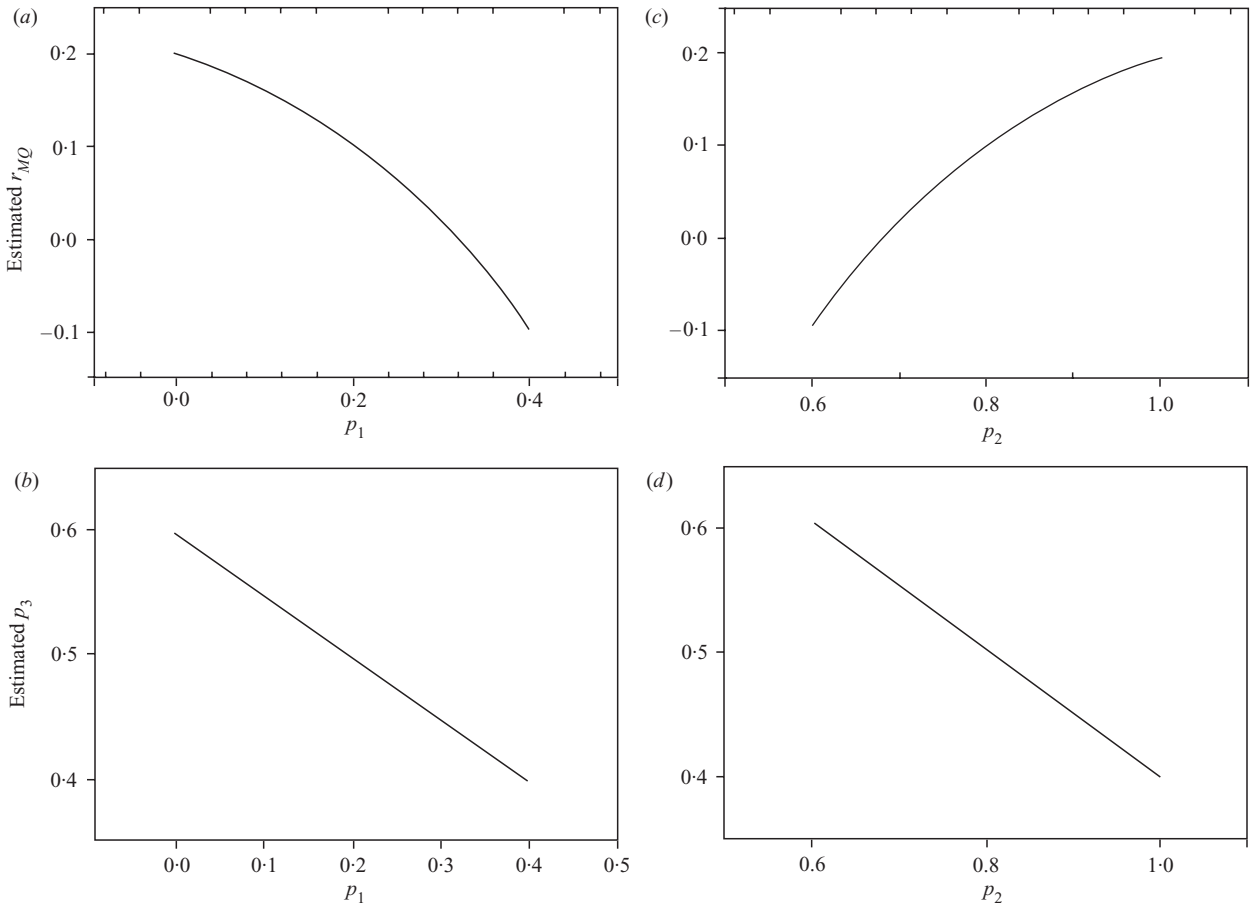

Fig. 5. (*a*) The effect of misspecification of p_1 on the estimation of recombination. (*b*) The effect of misspecification of p_1 on the estimation of p_3. (*c*) The effect of misspecification of p_2 on the estimation of recombination. (*d*) The effect of misspecification of p_2 on the estimation of p_3.

discordant marker trait pairs in the linear model the difference between the marker means is maximized and the power of the test in the linear regression is correspondingly high. In our simulations, we found no cases where the power of the logistic model was better than that of the linear model. In fact, the power of the tests using the logistic model was no better and may be worse than the tests in the linear model. On

https://doi.org/10.1017/S0016672301005092 Published online by Cambridge University Press

the basis of these results, our recommendation is to use the linear model for detection of BTL with an appropriate corresponding permutation algorithm to determine the P value empirically.

The estimators we present are moments estimators and are unbiased. However, they share a limitation inherent in all moment estimators, as they are not restricted to the parameter space in that the estimation of the penetrance p_3 is not automatically $0 \leqslant p_3 \leqslant 1$, or $\hat{r}_{MQ}$ is not bounded in the interval $0 \leqslant \hat{r}_{MQ} \leqslant 0{\cdot}50$. We do not present the joint estimation of all the p_i, instead we assume that p_1 and p_2 are known. We have shown that misspecification of p_1 and p_2 does lead to biased results, but that the impact is not large when misspecification is less than 20%.

Our parameterization requires much less computational effort than previously published methods (Martinez & Curnow, 1992; Haley & Knott, 1992) as no interval stepping is required. The estimates of location r_{MQ} are as accurate as the interval estimates and we can use a single marker for estimates of location, with no loss in accuracy (Table 11). Power is not directly comparable between our methods and other published methods (Martinez & Curnow, 1992; Haley & Knott, 1992) since the null hypotheses for the test statistics are different.

Based upon our parameterization, estimation of underlying threshold distributions is not needed, nor is any advance knowledge of the true genetic model. An experiment based upon two inbred lines, known to be different, can be conducted without the need for extensive testing of the F_1 phenotype. The F_1 phenotype or penetrance (p_3) can be estimated from the backcross or F_2 progeny even without precise knowledge of the penetrance of the parental lines. Incorporating the estimation of the penetrance in the F_1 expands the options available to the experimenter, allowing for the dual purpose of detecting BTL and identifying the genetic model in a single set of measurements.

The method described is for a single gene and a binary trait. The model parameterization applies directly to categorical traits if the categorical trait is modelled in pairs of outcomes. To expand the parameterization to multiple genes requires a formulation of the joint distribution of (Y, G, M) that includes multiple genes. Once the joint distribution has been expanded the expected values of the regression coefficients can be determined and the additional parameters estimated. It is important to note that in this approach every additional gene locus modelled requires an additional marker locus. This ensures that the model will be identifiable. That is, estimates for the additional recombination and penetrance parameters will be possible. If the model is expanded in this way, then the results of the current work should generalize to multiple genes.

Appendix

The expected log odds equations for the logistic model for the F_2 experimental design

$$e^{\beta_0} = \left[\frac{p_1 - 2p_1 r_{MQ} + r_{MQ}^2 p_1 + 2r_{MQ}p_3 - 2r_{MQ}^2 p_3 + p_2 r_{MQ}^2}{1 - (p_1 - 2p_1 r_{MQ} + r_{MQ}^2 p_1 + 2r_{MQ}p_3 - 2r_{MQ}^2 p_3 + p_2 r_{MQ}^2} \right]$$

$$e^{\beta_1} = e^{\beta_0} \left[\frac{p_1 r_{MQ} - p_1 r_{MQ}^2 + p_3 - 2r_{MQ}p_3 + 2r_{MQ}^2 p_3 + p_2 r_{MQ} - p_2 r_{MQ}^2}{1 - (p_1 r_{MQ} - p_1 r_{MQ}^2 + p_3 - 2r_{MQ}p_3 + 2r_{MQ}^2 p_3 + p_2 r_{MQ} - p_2 r_{MQ}^2)} \right]$$

$$e^{\beta_2} = e^{\beta_0} \left[\frac{r_{MQ}^2 p_1 + 2r_{MQ}p_3 - 2r_{MQ}^2 p_3 + p_2 - 2p_2 r_{MQ} + p_2 r_{MQ}^2}{1 - (r_{MQ}^2 p_1 + 2r_{MQ}p_3 - 2r_{MQ}^2 p_3 + p_2 - 2p_2 r_{MQ} + p_2 r_{MQ}^2)} \right].$$

This work is supported by NSF grant DBI 98-08026/00-96044 (L.M.M., C.J.C., R.W.D.), NIH grant NIA-AG16996 (L.M.M.), USDA grant 98-35300-6173 (R.W.D.) and a Veterans Affairs Health Services Research Post-doctoral Fellowship (C.J.C.). The authors would like to thank Katy Simonsen, Assistant Professor of Statistics at Purdue University, Marie Davidian, Professor of Statistics North Carolina State University, and James Holland, USDA-ARS Research Geneticist, Department of Crop Science, North Carolina State University.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Beckmann, J. & Soller, M. (1988). Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. *Theoretical and Applied Genetics* **76**, 228–236.

Churchill, G. & Doerge, R. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Doerge, R. & Churchill, G. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Doerge, R. W., Zeng, Z.-B. & Weir, B. S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12**, 195–219.

Elston, R. (1998). Methods of linkage analysis – and assumptions underlying them. *American Journal of Human Genetics* **63**, 931–934.

Fain, P. (1978). Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance. *Annals of Human Genetics* **42**, 109–120.

Hackett, C. & Weller, J. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

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

Hammond, K. & James, J. (1970). Genes of large effect and the shape of the distribution of a quantitative character. *Australian Journal of Biological Sciences* **23**, 867–876.

Hosmer, D. & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.
Jansen, R. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.
Jansen, R. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.
Kao, C. (1995). Statistical methods for locating the positions and analyzing epistasis of multiple quantitative trait genes using molecular marker information. PhD thesis, North Carolina State University.
Kruglyak, L. & Lander, E. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.
Luo, Z. & Kearsey, M. (1989). Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. *Heredity* **63**, 401–408.
Luo, Z. & Woolliams, J. (1993). Estimation of genetic parameters using linkage between a marker gene and a locus underlying a quantitative character in F_2 populations. *Heredity* **70**, 245–253.
Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.
Martinez, O. & Curnow, R. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical Applied Genetics* **85**, 480–488.
Mérat, P. (1968). Distributions de fréquences, interpretation du déterminisme génétique des caractères quantitatifs et recherche de 'genes majeurs'. *Biometrics* **24**, 277–293.
O'Donald, P. (1971). The distribution of genotypes produced by alleles segregating at a number of loci. *Heredity* **26**, 233–241.
Olson, J., Witte, J. & Elston, R. (1999). Tutorial in biostatistics genetic mapping of complex traits. *Statistics in Medicine* **18**, 2961–2981.
Paterson, A. (1998). *Molecular Dissection of Complex Traits*. Boca Raton, FL: CRC Press.
Rebaï, A. (1997). Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genetical Research* **69**, 69–74.
Searle, S. R. (1997). *Linear Models*. New York: Wiley.
Thompson, E. (1998). Inferring gene ancestry: estimating gene descent. *International Statistical Review* **66**, 29–40.
Visscher, P., Haley, C. & Knott, S. (1996). Mapping QTLs for binary traits in backross and F_2 populations. *Genetical Research* **68**, 55–63.
Weller, J. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.
Wilcox, P. (1995). Genetic dissection of fusiform rust resistance in loblolly pine. PhD thesis, North Carolina State University
Wright, S. (1952). The genetics of quantitative variability. In *Bull. Agricultural Research Council: Quantitative Inheritance*, pp. 5–41. London: Her Majesty's Stationery Office.
Xu, S. (1996). Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics* **144**, 1951–1960.
Xu, S. & Atchley, W. (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* **143**, 1417–1424.
Zeng, Z.-B. (1993). Theoretical basis of precision mapping of quantitative trait loci. *Proceedings of the National Academy of Sciences of the USA* **90**, 10972–10976.
Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.