# Identifying quantitative trait loci in experimental crosses

by

Karl William Broman

B.S. (University of Wisconsin, Milwaukee) 1991
M.A. (University of California, Berkeley) 1994

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA, BERKELEY

Committee in charge:

Professor Terence P. Speed, Chair
Professor John Rice
Professor Jasper D. Rine

Spring 1997

# Identifying quantitative trait loci in experimental crosses



by
Karl William Broman

## Abstract

Identifying quantitative trait loci
in experimental crosses

by

Karl William Broman
Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Terence P. Speed, Chair

Identifying the genetic loci responsible for variation in traits which are quantitative in nature (such as the yield from an agricultural crop or the number of abdominal bristles on a fruit fly) is a problem of great importance to biologists. The number and effects of such loci help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci may aid in designing selection experiments to improve the traits.

We focus on data from a large experimental cross. The usual methods for analyzing such data use multiple tests of hypotheses. We feel the problem is best viewed as one of model selection. After a brief review of the major methods in this area, we discuss the use of model selection to identify quantitative trait loci. Forward selection using a BIC-type criterion is found to perform quite well. Simulation studies are used to compare the performance of the major approaches. In addition, we present the analysis of data from a real experiment.

To my family

# Contents

## Acknowledgements

My years at Berkeley have been very rewarding. The faculty, staff and students here have created a warm and exciting environment. Terry Speed has been a constant source of energy and inspiration for me. I appreciate all of the advice and encouragement he has given me. My fellow students, especially Helge Blaker, Bill Forrest, Chad Heilig, David Levin, Steve Rein, Barathi Sethuraman, Aimee Teo, and the Fighting Sheep, have been of invaluable help during my struggles, have provided many stimulating arguments and discussions, and have filled any excess time with fun. I have had the great pleasure of regular conversations with a very thoughtful geneticist, Mark Neff. Our discussions on statistics and genetics will have a lasting effect on my views of both fields. I wish to thank John Rice and Jasper Rine for carefully reading this dissertation, Bin Yu for her advice on the problem, and Tony Long for providing the data analyzed in Chapter 5 and for answering my questions about it. Without the support of the Statistical Computing Facility, its staff, and the computers bilbo, pooh, and others, this work could not have been done. Finally, I would like to thank my family, who support, and show an interest in, all of my efforts.

# Chapter 1

# Introduction

In this thesis we consider the problem of identifying the genetic loci (called quantitative trait loci or QTLs) that contribute to variation in a quantitative trait. We focus on data from a large experimental cross, and assume that the genes act additively. Most of the current statistical methods for this problem use multiple tests of hypotheses. We feel the problem is best viewed as one of model selection, and so in this thesis, we develop the use of model selection ideas for identifying QTLs, and compare the results of this approach to the methods currently in use. We concentrate almost exclusively on detecting QTLs, considering the estimation of the QTLs' effects and precise locations of secondary importance.

Classical genetics has historically concentrated on binary traits, such as whether or not an individual has a particular disease. Such traits are often the result of a mutation at a single gene. However, most natural traits exhibit quantitative variation. Examples include the yields of agricultural crops, the number of abdominal bristles on fruit flies, and the heights and weights of people. Variation in quantitative traits often results from the action of multiple genes, called polygenes or quantitative trait loci (QTLs). The contribution of each particular gene may be quite small, while environmental (non-heritable) variation may be quite large. As a result, one cannot immediately infer an individual's genotype (its genetic composition) from its phenotype (the trait value), making it a difficult task to identify and characterize the QTLs.

Knowledge of the locations and actions of the QTLs helps us to understand the biochemical basis of these traits, and of their evolution over time, and may aid in designing selection experiments to improve the traits.

The idea that the quantitative variation in a trait could be due to the action of multiple genes was proposed in Gregor Mendel's seminal paper (Mendel 1866), in which he wrote that complex variation in the color of flowers might be due to the independent action of several genetic factors. Nilsson-Ehle (1909) demonstrated that this was indeed true. He showed that differences in the color of the grains of two varieties of wheat were due to segregation at three different loci.

By the 1920's, the chromosome theory and the concept of genetic linkage were well developed, primarily a result of experiments with *Drosophila melanogaster* (the fruit fly) in Thomas Hunt Morgan's lab. Sax (1923) demonstrated an association between seed weight and seed coat color in beans, and proposed that this association was due to linkage between the genes controlling color and one or more genes controlling size.

Thoday (1961) put forth the idea to use multiple genetic markers to systematically map the individual polygenes which control a quantitative trait, and noted that the only barrier to this approach was the small number of available markers. Another problem was that the phenotypic markers in use often displayed a larger effect on the quantitative trait than did the individual polygenes (Tanksley 1993).

Recently, biochemical markers have been developed: first protein polymorphisms and then DNA polymorphisms, such as restriction fragment length polymorphisms (RFLPs) and microsatellites. These markers have a number of useful properties. They are generally phenotypically neutral, they can be highly polymorphic, and, most importantly, they exist in great abundance, spanning entire genomes.

The development of biochemical markers has led to a proliferation of studies aimed at identifying and characterizing the QTLs responsible for variation in quantitative traits. A very large number of traits have been studied in many different organisms, such as pigs (Andersson et al. 1994), maize (Edwards et al. 1987; Beavis et al. 1991; Stuber et al. 1992), mice (Berretini et al. 1994), tomatoes (Paterson et al. 1990, 1991; deVicente and Tanksley 1993) and eucalyptus trees (Grattapaglia et al. 1996).

In the remaining part of this chapter, we describe the typical experiments used to identify QTLs and the statistical models which relate genotype to phenotype. In Chapter 2, we give a critical review of the methods which have been developed to identify QTLs in experimental crosses. In Chapter 3, we discuss the application of model selection ideas to this problem. Chapter 4 contains the results of some large simulation studies to compare the different methods for identifying QTLs, and to evaluate the power to detect QTLs for

different sizes of experiments. Chapter 5 contains an analysis of some data on the number of bristles in *Drosophila*. In Chapter 6, we discuss some further important issues, and summarize our conclusions.

## 1.1 Experiments

Most experiments aimed at identifying quantitative trait loci (QTLs) begin with two pure-breeding lines which differ in the trait of interest. We'll call these the low (L) and high (H) parental lines. The lines are the result of intensive inbreeding, so that each is essentially homozygous at all loci (meaning that, at each locus, they received the same allele from each of their two parents). Crossing these two parental lines gives the first filial (or $F_1$) generation. The $F_1$ individuals receive a copy of each chromosome from each of the two parental lines, and so, wherever the parental lines differ, they are heterozygous. All $F_1$ individuals will be genetically identical, just as the individuals in each of the parental lines were.

In a backcross (see Figure 1.1), the $F_1$ individuals are crossed to one of the two parental lines, for example, the low line. The backcross progeny, which may number from 100 to over 1000, receive one chromosome from the low parental line, and one from the $F_1$. Thus, at each locus, they have genotype either LL or HL. As a result of crossing over during meiosis (the process during which gametes or sex cells are formed), the chromosome received from the $F_1$ parent is a mosaic of the two parental chromosomes. At each locus, there is a half a chance of receiving the allele from the low parental line (L) and a half a chance of receiving the allele from the high parental line (H). The chromosome received will alternate between stretches of L's and H's.

The goal is to look for an association between the phenotype of an individual and whether it received the L or H allele from the $F_1$ parent at various marker loci.

Another common experiment is an intercross (see Figure 1.2). Here, the $F_1$ individuals are either selfed or crossed to each other. The individuals in the resulting $F_2$ generation each receive two chromosomes from the $F_1$ generation, each of which will be a combination of the two parental chromosomes. Thus, at each locus, the $F_2$ individuals will have genotypes LL, HL or HH.

We use the backcross as our chief example, because of its simplicity. At each locus in the genome, the backcross progeny have one of only two possible genotypes. The inter-

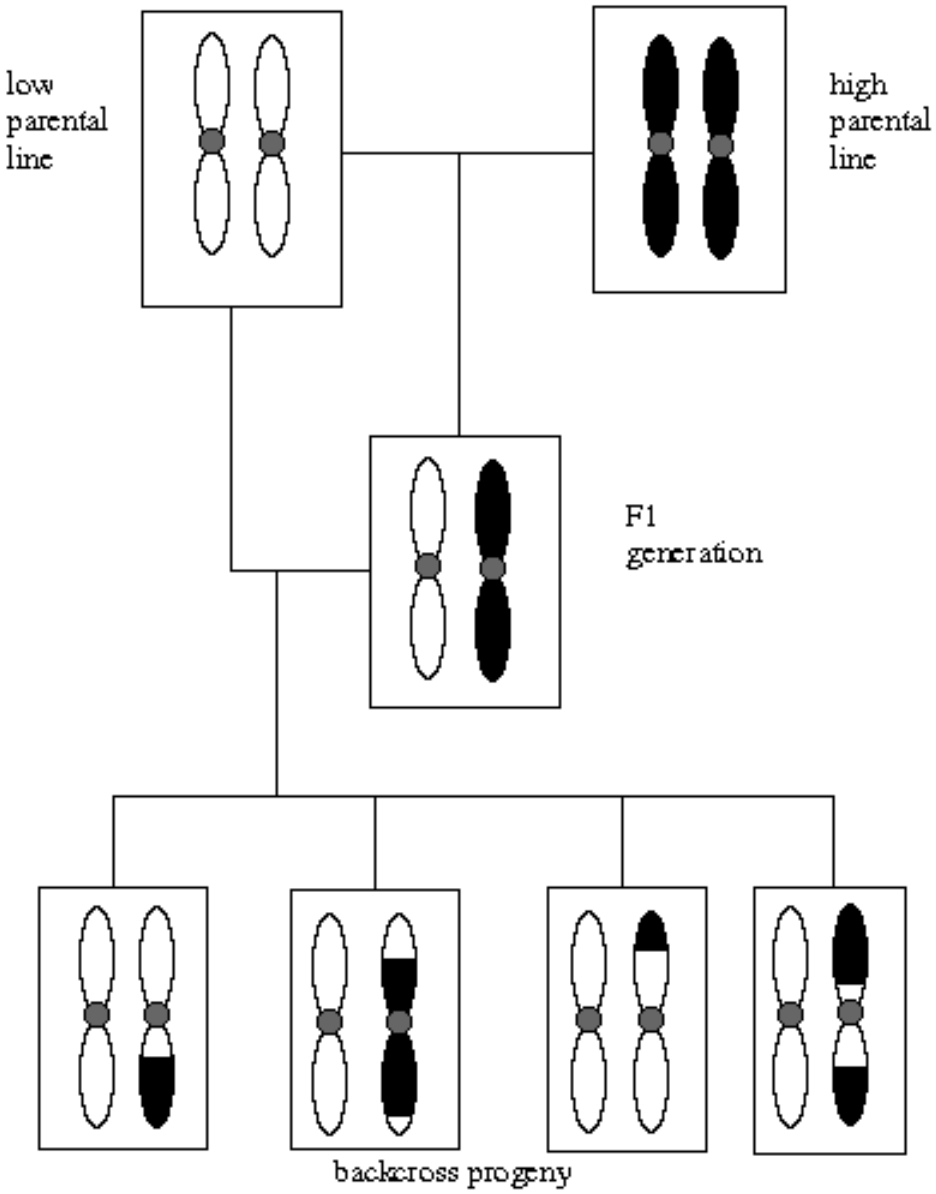


**Figure 1.1**: A backcross experiment, with four progeny. (Typical experiments contain more than 100 progeny.) Only one pair of homologous chromosomes is shown.

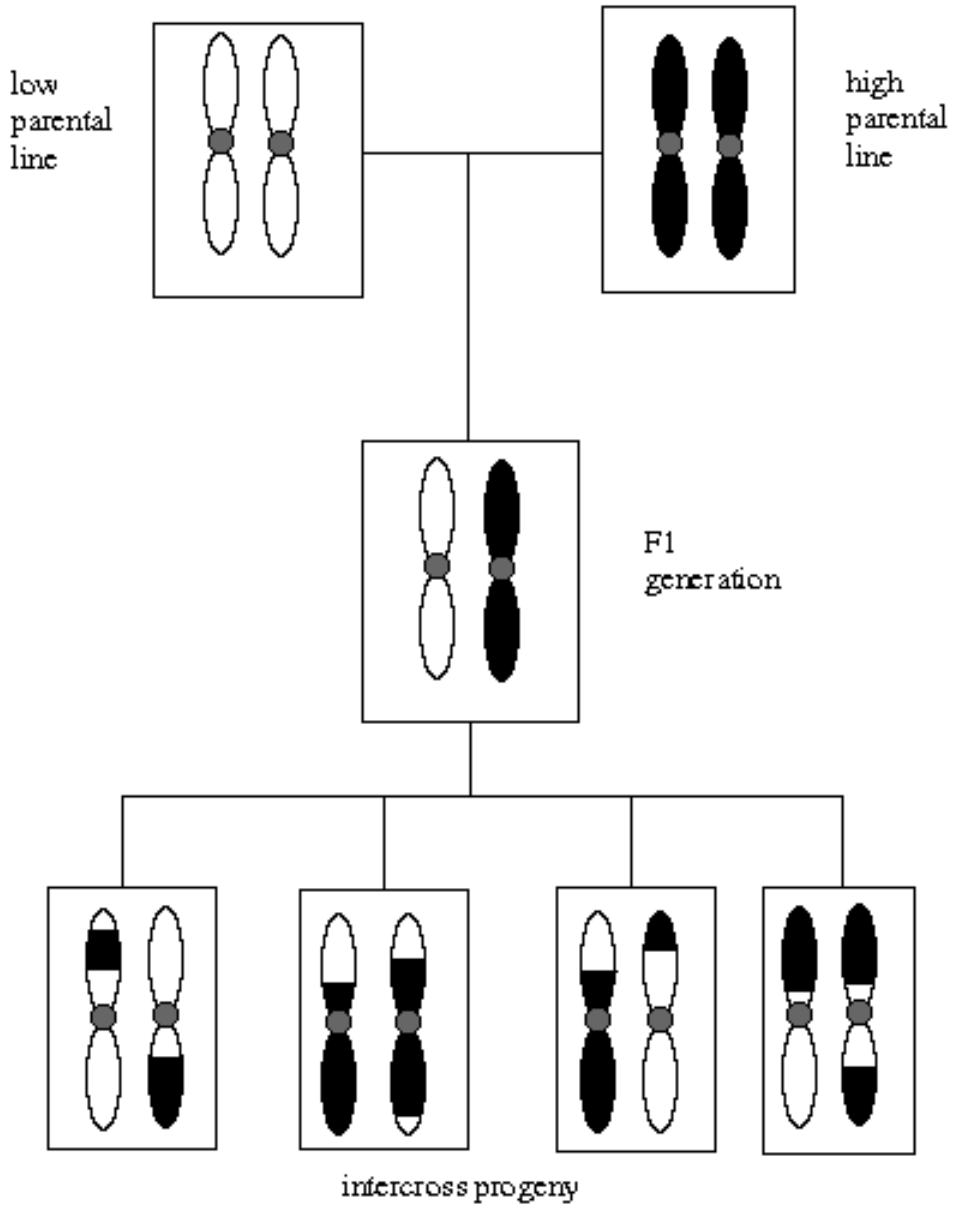


**Figure 1.2**: An intercross experiment, with four progeny. (Typical experiments contain more than 100 progeny.) Only one pair of homologous chromosomes is shown.

cross is more commonly used in practice, but the analysis of the two types of experiments is similar. The strategies developed for analyzing backcross experiments will generally work for intercross experiments as well.

## 1.2 Data

In an experiment like a backcross, each of the progeny is scored for one or more traits. (We'll consider only one trait.) In addition, the progeny are typed at a number of genetic markers: at each marker, it is determined whether the allele an individual received from the $F_1$ parent was that from the low or high parental line. Thus, at each of these marker loci, we determine, for each of the progeny, whether its genotype is LL or HL.

A genetic map, specifying the relative locations of the markers, may be known, or will be estimated using the data from the current experiment. Such a map gives the linear order of the markers on the various chromosomes. The distance between markers in a genetic map is given by genetic distance, in the units centiMorgans (cM). Two markers are separated by $d$ cM, if $d$ is the expected number of crossovers between the markers in 100 meiotic products.

Generally, we'll write $y_i$ for the phenotype (trait value) of individual $i$, and $x_{ij} =$ 1 or 0 according to whether individual $i$ has genotype HL or LL at the $j$th marker.

Typical experiments involve 100 to 1000 progeny, and use between 100 and 300 genetic markers.

## 1.3 Models

### 1.3.1 Model for recombination

A diploid organism has two copies of each chromosome, one from its mother and one from its father. During the formation of gametes (sex cells), in the process of meiosis, the two homologous copies of a chromosome may undergo exchanges, called crossovers. Each of the gametes formed contains one copy of each chromosome, and each of these will be a mosaic of the two original homologs.

The locations of the crossovers along a chromosome are often modelled as a Poisson process (the assumption of "no interference"), with the processes in different individuals

and on different chromosomes in one individual being independent. Moreover, at each locus, there is an equal chance that the allele is either paternally or maternally derived.

Consider a chromosome with $k$ markers, and let $x_{ij} = 1$ or 0 if the $i$th individual has genotype HL or LL, respectively, at the $j$th marker. Then $\mathbf{Pr}(x_{ij} = 1) = \mathbf{Pr}(x_{ij} = 0) = 1/2$, for all $i, j$, and letting $x_j = (x_{ij})$, the $x_j$ form a Markov chain.

Consider markers $j_1$ and $j_2$, separated by a distance of $d$ cM (so that $d$ is the expected number of crossovers between these two markers in 100 meioses). If an odd number of exchanges occur between these markers, then $x_{ij_1} \neq x_{ij_2}$. This event is called a recombination. Let $r = \mathbf{Pr}(x_{ij_1} \neq x_{ij_2})$. Then $r = \frac{1}{2}(1 - e^{-2d/100})$. This is called the Haldane map function (Haldane 1919).

### 1.3.2 Model connecting genotype and phenotype

Let $y$ denote the phenotype for an individual derived from a backcross experiment. Let $g$ be a vector giving its genotype at all loci. Let $\mu_g = \mathbf{E}(y|g)$, the average phenotype for individuals with genotype $g$, and $\sigma_g^2 = \mathbf{var}(y|g)$, the variance of the phenotypes of individuals with genotype $g$. In principle, these could be arbitrary functions of $g$. But imagine that there are a finite number, $p$, of loci which affect the trait. Let $(g_1, \ldots, g_p)$ denote the genotypes of the individual at these loci. Then

$$\begin{aligned} \mathbf{E}(y|g) &= \mu_{g_1 \ldots g_p} \\ \text{and} \quad \mathbf{var}(y|g) &= \sigma^2_{g_1 \ldots g_p}. \end{aligned}$$

Often, we assume that the trait is homoscedastic—that the variance is constant within the genotype groups:

$$\mathbf{var}(y|g) = \sigma^2.$$

There are $2^p$ different possible genotypes at the $p$ QTLs. Each genotype could have a distinct trait mean. But often we assume that the loci act additively. Let $z_j = 1$ or 0, according to whether $g_j =$ HL or LL. We imagine that

$$\mathbf{E}(y|g) = \mu + \sum_{j=1}^{p} \beta_j z_j.$$

Deviation from additivity (i.e. interactions between the QTLs) is called epistasis.

Most current methods use this assumption of additivity. Pairwise interactions are occasionally included, but few studies have found significant effects when using such an

approach (Tanksley 1993), possibly because of the enormous number of pairwise interactions which must be considered. Strong evidence for epistasis has been demonstrated in one of the most studied quantitative traits, the number of abdominal bristles in *Drosophila* (Shrimpton and Robertson 1988; Long et al. 1995). Thus one should not discount the importance of epistasis.

A further often used assumption is that, given the genotypes at the QTLs, the trait $y$ follows a normal distribution. Thus, if we group the backcross progeny according to their genotypes at the $p$ QTLs, the phenotypes within each group will be normally distributed. The phenotypes for the backcross progeny, considered as a whole, will follow a mixture of normal distributions.

In this thesis, we will focus on the case of strict additivity, with the further assumption of normality. This is not because we feel that it is the best approach, but rather because this simple case is still not well solved. We would like to reframe the problem of identifying QTLs as one of model selection rather than hypothesis testing, and this will be done most clearly if we avoid the added difficulties which accompany a search for epistasis.

## 1.4 Goals

Consider a backcross giving $n$ progeny. For individual $i$, we obtain the phenotype, $y_i$, and the genotype at a set of $M$ markers. Let $x_{ij} = 1$ or 0, according to whether individual $i$ has genotype HL or LL at the $j$th marker.

We imagine that there are a set of $p$ QTLs, and write $z_{ij} = 1$ or 0, according to whether individual $i$ has genotype HL or LL at the $j$th QTL. Let

$$y_i = \mu + \sum_{j=1}^{p} \beta_j z_{ij} + \epsilon_i$$

where the $\epsilon_i$ are independent and identically distributed (iid) normal$(0, \sigma^2)$.

The ultimate goal is to estimate the number of QTLs, $p$, the locations of the QTLs, and their effects, $\beta_j$. In estimating the number and locations of the QTLs, we may make two errors: we may miss some of the QTLs, and we may include additional, extraneous loci.

In practice, a scientist may be satisfied with finding a few QTLs with large effect. In QTL experiments aimed at improving an agricultural crop, one seeks only the major QTLs, which may then be introgressed from one line into another. Furthermore, with a

few major QTLs in hand, it may be possible to design experiments which identify the other QTLs segregating in a cross.

How one chooses to balance the two errors, of missing important loci and of including extraneous loci, depends on the goals of the scientists who designed the cross. In some cases, one may wish to find as many of the QTLs as possible and be undeterred by the possibility that several of the identified loci are, in fact, extraneous ones, of no effect. In other situations, one may be satisfied with identifying only a few major QTLs, in order to avoid including extraneous ones.

# Chapter 2

# Major approaches

There are a large number of different methods for identifying the QTLs segregating in an experimental cross (such as a backcross or an $F_2$ intercross, obtained from two inbred lines). In this chapter, we describe most of the proposed methods and briefly discuss their advantages and disadvantages. We focus on the example of a backcross. Two highly inbred lines, differing in the trait of interest, are crossed. The resulting $F_1$ generation is crossed back to one of the two parental lines. The backcross progeny obtained are either heterozygous (with genotype HL, say) or homozygous (with genotype LL) at each locus in the genome.

It's best to distinguish between methods which model a single QTL at a time from those which attempt to model the effects of several QTLs at once. In Section 2.1, we review the single QTL methods, and in Section 2.2, we review the multiple QTL methods. Section 2.3 contains a discussion of the relative merits of the methods.

## 2.1 Single QTL methods

We will consider five basic single QTL methods: analysis of variance at a single marker, maximum likelihood using a single marker, interval mapping (i.e., maximum likelihood using flanking markers), an approximation to interval mapping called "regression mapping," and a further method which gives results approximating interval mapping, called "marker regression." Each of these methods includes a so-called "genome scan." The loci are considered one at a time, and a significance test for the presence of a single QTL is performed at each. Generally, the significance level used for the tests is adjusted to account

for the multiple tests performed. Areas of the genome which give significant results are indicated to contain a QTL.

### 2.1.1 Analysis of variance

Analysis of variance (ANOVA) is the simplest method for identifying QTLs (see Soller et al. 1976). Consider a single marker locus, and group the progeny according to their genotypes at that marker. To test for the presence of a QTL, we look for differences between the mean phenotype for the different groups using ANOVA. If a QTL is tightly linked to the marker, then grouping the progeny according to their marker genotypes will be nearly the same as grouping them according to their (unknown) QTL genotypes, with recombinants being placed in the wrong groups.

Consider a backcross with a single segregating QTL. Suppose that the progeny with QTL genotype HL have mean phenotype $\mu_H$, and that progeny with QTL genotype LL have mean phenotype $\mu_L$, so the QTL has effect $\beta = \mu_H - \mu_L$. Consider a marker locus which is a recombination fraction $r$ away from the QTL. Of the individuals with marker genotype HL, a fraction $(1-r)$ of them have QTL genotype HL, while the remainder have QTL genotype LL, and so these individuals have mean phenotype $(1-r)\mu_H + r\mu_L = \mu_H - \beta r$. The individuals with marker genotype LL have mean phenotype $(1-r)\mu_L + r\mu_H = \mu_L + \beta r$. Thus the mean difference between the two marker genotype groups is $(\mu_H - \beta r) - (\mu_L + \beta r) = \beta(1-2r)$. And so a non-zero mean difference between the marker genotype groups indicates linkage between the marker and a QTL.

There are two drawbacks to this method. First, we do not receive separate estimates of the location of the QTL relative to the marker ($r$) and its effect ($\beta$). QTL location is indicated only by looking at which markers give the greatest differences between genotype group means. Second, when the markers are widely spaced, the QTL may be quite far from all markers, and so the power for detection will decrease, since the difference between marker genotype means decreases linearly as the recombination fraction between the marker and the QTL increases.

### 2.1.2 Maximum likelihood with a single marker

To get around the problems with ANOVA, several authors have proposed to explicitly model the location of the QTL with respect to the marker, and then use maximum

likelihood (ML), or an approximation to ML, to estimate the distance between the marker and the QTL as well as the QTL's effect (Weller 1986, 1987; Simpson 1989). This method makes use of the fact that the marker genotype groups are not quite the same as the QTL genotype groups.

Consider again the backcross discussed in the previous section. Suppose that the individuals who are HL at the QTL have phenotypes which are normal$(\mu_H, \sigma^2)$, and the individuals who are LL at the QTL have phenotypes which are normal$(\mu_L, \sigma^2)$. Then at a marker which is a recombination fraction $r$ away from the QTL, the phenotype distribution for individuals who are HL is a mixture of two normals, with density

$$f_1(y; \mu_H, \mu_L, \sigma, r) = (1-r)\phi\left(\frac{y-\mu_H}{\sigma}\right) + r\phi\left(\frac{y-\mu_L}{\sigma}\right),$$

where $\phi$ is the density of the standard normal distribution. The phenotype distribution for individuals who are LL at the marker has density

$$f_2(y; \mu_H, \mu_L, \sigma, r) = (1-r)\phi\left(\frac{y-\mu_L}{\sigma}\right) + r\phi\left(\frac{y-\mu_H}{\sigma}\right).$$

Let $x_i = 1$ or 0, according to whether individual $i$ has marker genotype HL or LL. Let $y_i$ denote the phenotype for individual $i$. Then the likelihood under this model, letting $\theta$ denote the vector of parameters $(\mu_H, \mu_L, \sigma)$, is

$$L(\theta, r; y, x) = \prod_i [f_1(y_i; \theta, r)]^{x_i} [f_2(y_i; \theta, r)]^{1-x_i}$$

Maximizing this function over $\theta$, using, for example, the EM algorithm (Dempster et al. 1977), gives the maximum likelihood estimates. This is done for a particular value of the recombination fraction $r$. We then maximize the likelihood over $r$ to obtain $\hat{r}$.

Linkage between the marker and the QTL is tested by performing a likelihood ratio test, comparing the above model, with a single QTL linked to the marker, to the null hypothesis of no segregating QTLs, where the individuals are assumed to have phenotypes which are normal$(\mu, \sigma^2)$.

The likelihood under the null hypothesis, letting $\theta_0 = (\mu, \sigma)$, is

$$L_0(\theta_0; y) = \prod_i \phi\left(\frac{y_i - \mu}{\sigma}\right).$$

The likelihood ratio test is performed by calculating the likelihood ratio, or, as seems to be preferred by geneticists, the LOD score, which is the log (base 10) likelihood ratio

$$\mathrm{LOD}(r) = \log_{10}\left[\frac{\max_\theta L(\theta, r; y, x)}{\max_{\theta_0} L_0(\theta_0; y)}\right]$$

and comparing it to the distribution of the maximum LOD score under the null hypothesis (that is, under the assumption that no QTLs are segregating).

This method has the advantage of giving separate estimates of the QTL's location with respect to a marker and its effect. One disadvantage is the great increase in computation associated with maximizing the likelihood function to obtain parameter estimates. But a bigger problem involves combining the information for different markers to give a single estimate of the QTL location; it is not at all clear how this can be done.

### 2.1.3 Interval mapping

Lander and Botstein (1989) improved on the previous single marker approaches by considering flanking markers. Their method has been called "interval mapping," and is currently the most popular method for identifying QTLs in experimental crosses.

Again, they assume that there is a single segregating QTL, and that backcross individuals have phenotypes which are normally distributed with mean $\mu_H$ or $\mu_L$, according to whether their QTL genotype is HL or LL, and common variance $\sigma^2$. Further, they use the assumption of no crossover interference, and require a genetic map specifying the locations of the markers.

Consider two markers which are separated by $d$ cM, corresponding to a recombination fraction of $r = \frac{1}{2}(1 - e^{-2d/100})$, and a putative QTL located $d_L$ cM from the left marker, corresponding to a recombination fraction of $r_L = \frac{1}{2}(1 - e^{-2d_L/100})$. The recombination fraction between the QTL and the right marker is thus $r_R = (r - r_L)/(1 - 2r_L)$. There are four possible sets of genotypes at the two marker loci; for each, we can calculate the conditional probability for each of the two QTL genotypes, given the marker genotypes. These are displayed in Table 2.1. Note that, with fully informative markers, the flanking markers provide all of the information about the QTL genotypes.

For each of the four sets of marker genotypes, we can now write down the conditional phenotype density, which has the form of a mixture of two normal distributions, similar to those seen in Section 2.1.2. Thus we can obtain the likelihood for our four parameters, $(\mu_H, \mu_L, \sigma, r_L)$.

Lander and Botstein (1989) proposed to maximize this likelihood, for fixed $r_L$, using the so-called EM algorithm (Dempster et al. 1977). They then calculated the LOD score, which is the log (base 10) likelihood ratio comparing the hypothesis of a single QTL

**Table 2.1**: Conditional probabilities for the QTL genotypes given the two flanking marker genotypes.

| marker genotype | | QTL genotype | |
|---|---|---|---|
| left | right | HL | LL |
| HL | HL | $(1-r_L)(1-r_R)/(1-r)$ | $r_L r_R/(1-r)$ |
| HL | LL | $(1-r_L)r_R/r$ | $r_L(1-r_R)/r$ |
| LL | HL | $r_L(1-r_R)/r$ | $(1-r_L)r_R/r$ |
| LL | LL | $r_L r_R/(1-r)$ | $(1-r_L)(1-r_R)/(1-r)$ |

at the current locus (i.e., the current value of $r_L$) to the null hypothesis of no segregating QTLs (meaning that the individuals' phenotypes follow a normal$(\mu, \sigma^2)$ distribution). The two likelihoods in this ratio must be maximized over their respective parameters.

The procedure outlined above is performed for each locus in the genome. The likelihood under the null hypothesis is calculated just once. The likelihood for the hypothesis of a single QTL must be calculated at each locus in the genome (or, really, just every 1 cM or so), and so the EM algorithm must be performed at each locus.

The LOD score is then plotted against genome location, and is compared to a genome-wide threshold. Whenever the LOD curve exceeds the threshold, we infer the presence of a QTL. The point at which the LOD is maximized is used as the estimate of the QTL location. A one- or two-LOD support interval, the region around the inferred QTL in which the LOD score is within one or two of its maximum, is used as an interval estimate for QTL location.

The genome-wide threshold, used to indicate the significance of a peak in the LOD curve, is obtained by finding the 95th percentile of the maximum LOD score, across the entire genome, under the null hypothesis of no segregating QTLs.

Figure 2.1 gives an example of a LOD curve. We simulated 200 backcross progeny, having a single chromosome of length 100 cM with 11 equally spaced markers, using a model with a single QTL located 35 cM from the left of the chromosome. The effect of the QTL (the difference between the means for HL versus LL individuals) was $0.75\sigma$, giving a heritability, the proportion of the total phenotypic variance due to the QTL, of 0.36. The dots plotted on the curve point out the locations of the marker loci. Using a LOD threshold of 2.5, the observed peak is significant. The inferred QTL is estimated to be at 37 cM, with

a maximum LOD score of 3.4. The one-LOD support interval covers the region from 27 cM to 47 cM, which does indeed include the actual location of the simulated QTL.

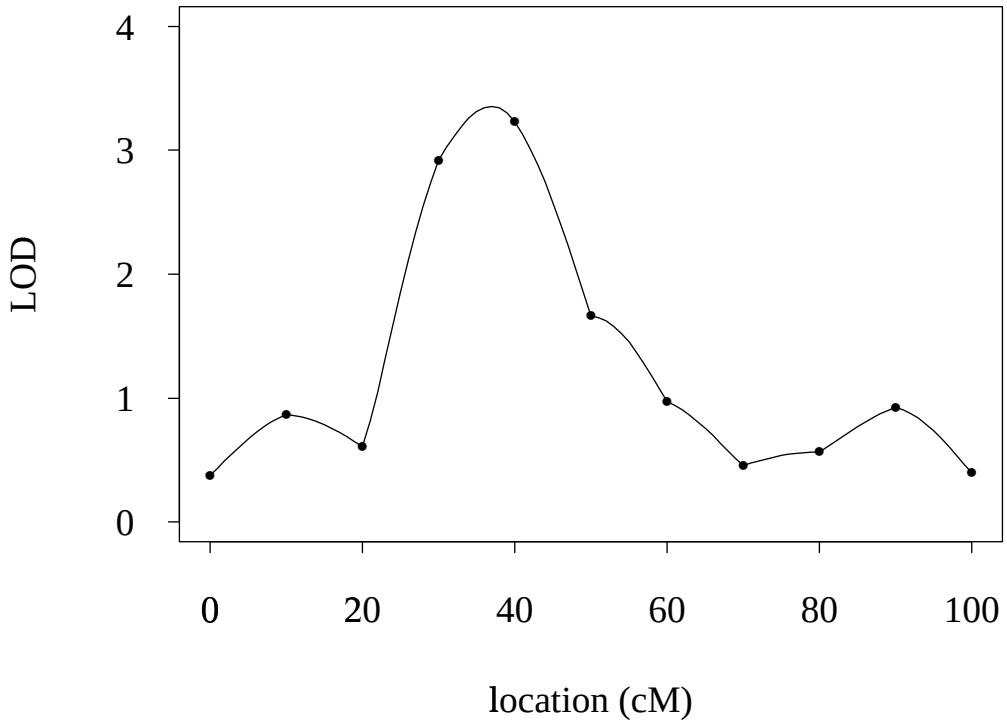


**Figure 2.1**: An example LOD curve for some simulated data.

A great deal of effort has been expended in trying to understand the appropriate LOD threshold to use. Lander and Botstein (1989) performed simulations to estimate the threshold for various different genome sizes and marker densities. They gave analytical calculations for the case of a very dense marker map. These guidelines should suffice for most uses. If one is concerned, additional simulations, conforming to the particular case under study, can be performed quite easily, or one can use a permutation test (Churchill and Doerge 1994), which has the advantage of avoiding the assumption of normally distributed environmental variation.

A number of studies have assessed the performance of interval mapping in comparison to ANOVA (van Ooijen 1992; Knott and Haley 1992; Darvasi et al. 1993; Rebaï et al. 1995; Hyne et al. 1995). The chief benefit of interval mapping is that it gives more precise estimates of the location and effect of a QTL. It does not give an appreciable increase in

the power for detecting QTLs, and it requires a great deal more computational effort than does single marker ANOVA.

Hyne et al. (1995) stated that when a QTL is located very near one end of a linkage group, its estimated location, as given by interval mapping, will be biased, since if one looks for QTLs only within the two extreme markers on the linkage group, its estimated location will never be outside of the last marker. It is possible to extend the LOD curves beyond the most extreme markers, however; outside of these markers, we can use the single marker maximum likelihood method, described in the previous section. Doing this should eliminate the bias problem. (Of course, a slight increase in variance, and a slight decrease in power, will accompany this approach.)

Look again at Figure 2.1. The dots on the LOD curve are at the marker loci. At these points, interval mapping is really just ANOVA, since the genotypes there are known exactly. If we performed only ANOVA, we'd get exactly those points on the LOD curve. Interval mapping links these points together, and indicates that the best estimate for the QTL position is at 37 cM. But the markers at both 30 and 40 cM are within the one-LOD support interval.

### 2.1.4 Regression mapping

Knapp et al. (1990), Haley and Knott (1992), and Martínez and Curnow (1992) independently developed a method which approximates interval mapping quite well, but requires much less computation. The method has come to be called "regression mapping." The presentation in Haley and Knott (1992) is by far the best.

Consider again the model of the previous section, with two markers separated by a recombination fraction $r$, and a putative QTL located between them, at a recombination fraction $r_L$ from the left marker. The conditional expected value of the phenotype for an individual, given its marker genotypes, is

$$\mathbf{E}(y|\text{marker gen.}) = \mu_L + (\mu_H - \mu_L)\mathbf{Pr}(\text{QTL gen. is HL}|\text{marker gen.}),$$

where $\mathbf{Pr}(\text{QTL gen. is HL}|\text{marker gen.})$ is as shown in Table 2.1 (page 14).

In regression mapping, we regress the individuals' phenotypes on their conditional probabilities for having the genotype HL at the putative QTL, given their marker genotypes. The log likelihood is calculated assuming that

$$y|\text{marker gen.} \sim \text{normal}(\tilde{y}, \sigma^2)$$

where $\tilde{y} = \mathbf{E}(y|\text{marker gen.})$. This gives the LOD score

$$\text{LOD} = \frac{n}{2}\log_{10}\left(\frac{\text{RSS}_0}{\text{RSS}}\right)$$

where $n$ is the number of progeny, RSS is the residual sum of squares from the above regression, $\sum_i (y_i - \hat{y}_i)^2$, and $\text{RSS}_0$ is the residual sum of squares under the null hypothesis of no segregating QTLs, $\sum_i (y_i - \bar{y})^2$.

Like interval mapping, the LOD score is calculated at each locus in the genome, but here, we need only calculate a single regression at each locus, rather than perform the EM algorithm at each locus, which requires a number of iterations, each containing a regression. Thus, there is a great savings in computation time. Also, because regression mapping requires only simple regression calculations, it is much easier to include additional effects into the analysis, such as sex or treatment effects. This may translate into large increases in performance.

Figure 2.2 displays the difference between the LOD curves calculated by regression mapping and interval mapping, for the data used in the previous section. The difference between the two curves is very subtle, being less than 0.1 in absolute value. Regression mapping gives results every bit as good as interval mapping, with a great deal less computation.

### 2.1.5 Marker regression

Kearsey and Hyne (1994) and Wu and Li (1994) independently developed a further method, which seems to approximate interval mapping quite well, with less intensive computation. But this method, which Kearsey and Hyne call "marker regression," seems more awkward and less adaptable than Haley and Knott's "regression mapping," and has not been shown to provide any further benefits.

Consider a linkage group with $M$ markers, and fix the location for a putative QTL. Let $r_j$ be the recombination fraction between the QTL and the $j$th marker. Group the individuals according to whether they have genotype HL or LL at marker $j$. Let $\hat{\beta}_j$ be the difference between the phenotype means for these two groups. As shown in Section 2.1.1,

$$\mathbf{E}(\hat{\beta}_j) = \beta(1 - 2r_j),$$

where $\beta = \mu_H - \mu_L$, the effect of the QTL.

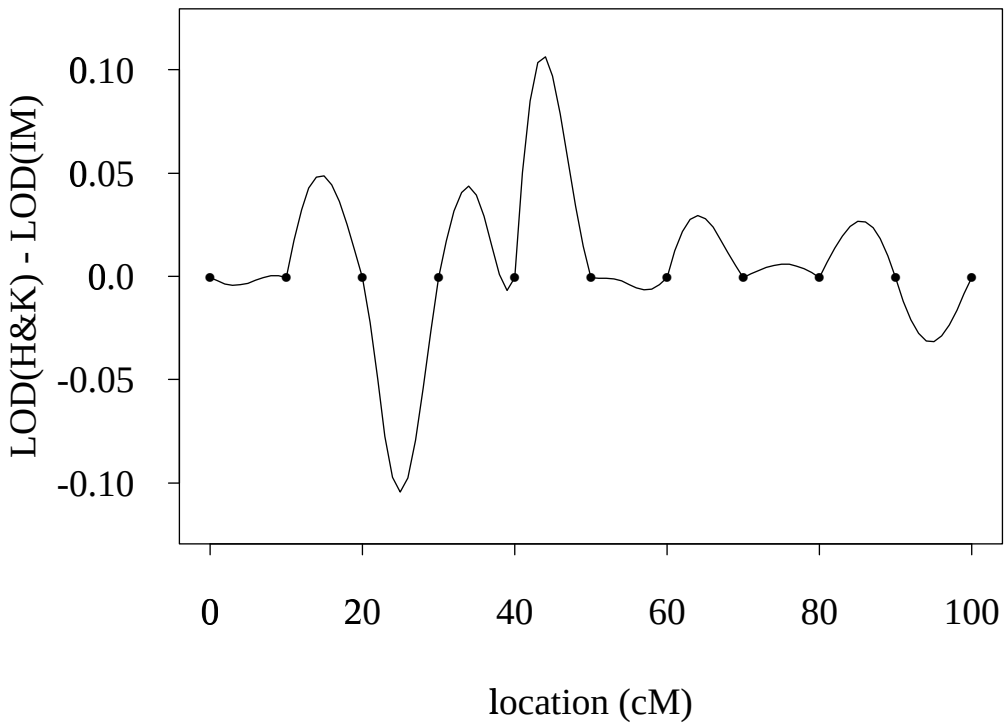


**Figure 2.2**: The difference between the LOD curves calculated using regression mapping and interval mapping for some simulated data.

Kearsey and Hyne (1994) suggest regressing the $\hat{\beta}_j$ for the $M$ markers on the values $(1 - 2r_j)$, without an intercept. This is performed for each locus on the linkage group; we seek the locus giving the minimum residual sum of squares in this regression.

Wu and Li (1994) point out that the $\hat{\beta}_j$ do not have constant variance. The variance of $\hat{\beta}_j$ is approximately $4[\sigma^2 + r_j(1 - r_j)\beta^2]/n$, where $n$ is the number of progeny, and $\sigma^2$ is the environmental variance. They suggest using weighted least squares, using weights inversely proportional to the variances of the $\hat{\beta}_j$. But since $\sigma$ and $\beta$ are not known, it is not clear how to do this, unless one were to use a form of iteratively re-weighted least squares.

Wu and Li (1996) further point out that the $\hat{\beta}_j$ are correlated, and recommend using general least squares using an estimate of the covariance matrix.

We applied the method of Kearsey and Hyne (1994) to the simulated data analyzed in Sections 2.1.3 and 2.1.4. Figure 2.3 displays the residual sum of squares curve. The minimum is realized at 42 cM.

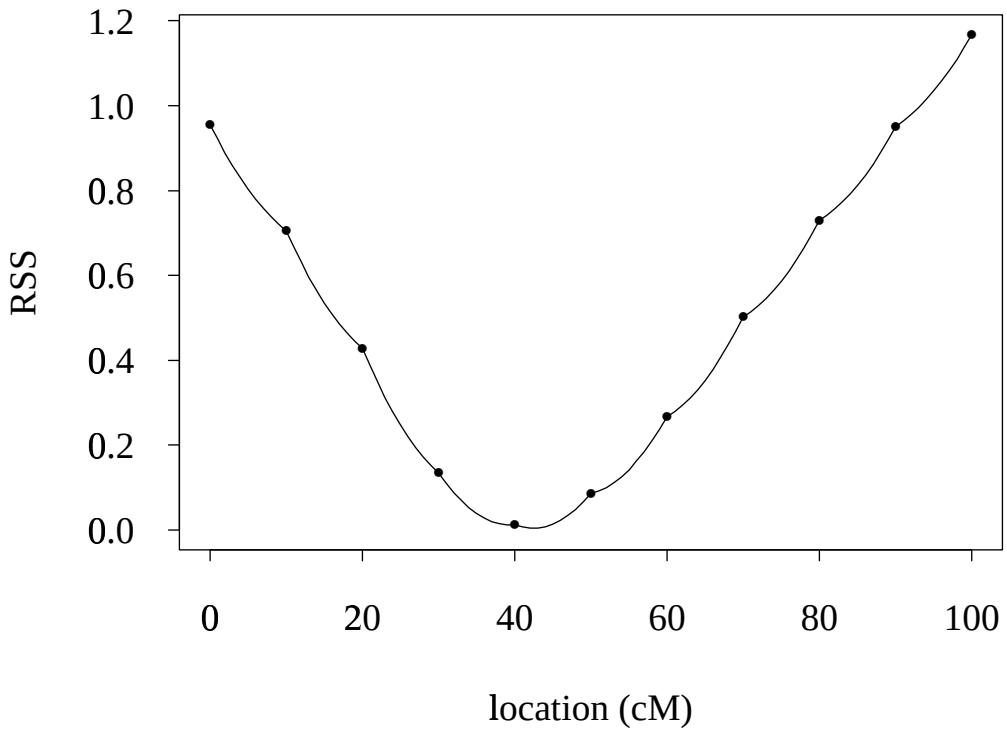


**Figure 2.3**: The residual sum of squares curve calculated using the marker regression method for some simulated data.

Kearsey and Hyne (1994) gave a small amount of simulations which suggested that marker regression performs as well as interval mapping. But they have not made a case for real improvements, aside from ease of computation. The method seems to have no advantages over regression mapping.

## 2.2 Multiple QTL methods

Recent efforts in developing methods to identify QTLs have focused on multiple QTL methods. There are three principal reasons for modelling multiple QTLs: to increase sensitivity, to separate linked QTLs, and to estimate epistatic effects (i.e., interactions between alleles at different QTLs).

When several QTLs are modelled, one can control for much of the genetic variation in a cross, and thus individual QTLs can be more clearly seen. In contrast, when one models a single QTL at a time, the genetic variation due to other segregating QTLs is incorporated

into the "environmental" variation. When two QTLs are linked, single QTL methods, such as interval mapping, often view them as a single QTL. Searches which allow multiple QTLs do a better job of separating the two loci, and identifying them as distinct. The presence of epistasis can only be detected and estimated using models which include multiple QTLs. Incorporating epistatic effects into multiple QTL models will be very difficult, however. If one were to include all possible pairwise interactions, the number of parameters in the model would quickly explode. The methods discussed here all neglect the possibility of epistasis.

In this section, we discuss four important methods which explicitly consider multiple QTLs: multiple regression, interval mapping type methods using either forward selection or multi-dimensional searches, composite interval mapping (also called MQM mapping), and Markov chain Monte Carlo using a full Bayesian model.

### 2.2.1 Multiple regression

The obvious extension of analysis of variance is multiple regression. We attempt to form a model which includes a number of different marker loci, rather than looking at the markers one at a time. Let $M$ be the number of markers, let $x_{ij} = 1$ or 0, according to whether individual $i$ had genotype HL or LL at the $j$th marker, and let $y_i$ be the phenotype for individual $i$. We write

$$\mathbf{E}(y_i|x_i) = \mu + \sum_{j=1}^{M} \beta_j x_{ij}$$

where $x_i = (x_{ij})$. We presume that most of the markers have $\beta_j = 0$. We seek the set of markers, $S$, with non-zero coefficients, $\beta_j$, so that

$$\mathbf{E}(y_i|x_i) = \mu + \sum_{j \in S} \beta_j x_{ij}.$$

The markers in $S$ are indicated to be near QTLs.

There are two problems associated with this method. First, we must find a way to search through the set of possible models, in order to seek good ones. In an experiment with 100 genetic markers, there are $2^{100} \approx 10^{30}$ possible models to consider; it will be impossible to look at each of them. Second, we must form a criterion for choosing from these models. For models that include the same number of markers, one generally picks the one with the smallest residual sum of squares. The difficulty is in choosing between models of different sizes: what change in the residual sum of squares must we see before we'll accept an additional marker into the model?

Cowen (1989) discussed using stepwise selection and backward deletion, and using Mallows' $C_p$ and the adjusted-$R^2$ criteria, when using multiple regression to identify QTLs. More recently, Doerge and Churchill (1996) described using forward selection, with permutation tests to determine the appropriate size of the model. We will discuss these approaches in detail in Chapter 3.

### 2.2.2 Interval mapping revisited

Lander and Botstein (1989) briefly mentioned a method for distinguishing linked loci. If, when performing interval mapping, the LOD curve for a linkage group shows two peaks, or a single very broad peak, Lander and Botstein recommended to fix the position of one QTL at the location of the maximum LOD, and then search for a second QTL on that linkage group. In the model selection literature, this method is generally called forward selection (Miller 1990). Though some authors (Haley and Knott 1992; Satagopan et al. 1996) have interpreted this method as applying interval mapping to the residuals from the best fit of one QTL, it is best to estimate the effects of both QTLs simultaneously, using the original data (cf Dupuis et al. 1995).

We fix the location of the first QTL, and vary the location of the second QTL along the linkage group. At each location for the second QTL, we calculate a LOD score, comparing the maximum likelihood under the hypothesis of two QTLs at these locations, to that with a single QTL, located where the first QTL was placed. Each individual's contribution to the likelihood has the form of a mixture of four normal distributions, the four components corresponding to the four possible QTL genotypes. The EM algorithm can again be used to obtain the maximum likelihood estimates and the corresponding LOD score. (One could also apply the "regression mapping" method.)

Several authors have criticized this method (Haley and Knott 1992; Martínez and Curnow 1992), pointing to the phenomenon of "ghost QTLs." When two or more QTLs are linked in coupling (meaning that their effects have the same sign), interval mapping often gives a maximum LOD score at a location in between the two QTLs.

Consider, for example, a 60 cM segment of a chromosome, with four equally spaced markers (20 cM spacing). Consider a backcross with QTLs located at 15 and 45 cM, acting additively and having equal additive effect $0.5\sigma$. Figure 2.4 gives the expected LOD (ELOD) curve for this situation, when using 200 progeny. (Since there is no closed-form expression

for the ELOD curve, it was estimated by performing 1000 simulations of the above situation and averaging the LOD curves obtained. We also used the fact that the ELOD curve is symmetric about the 30 cM point, and so averaged the pairs of points on the curve which are symmetric about 30 cM.)

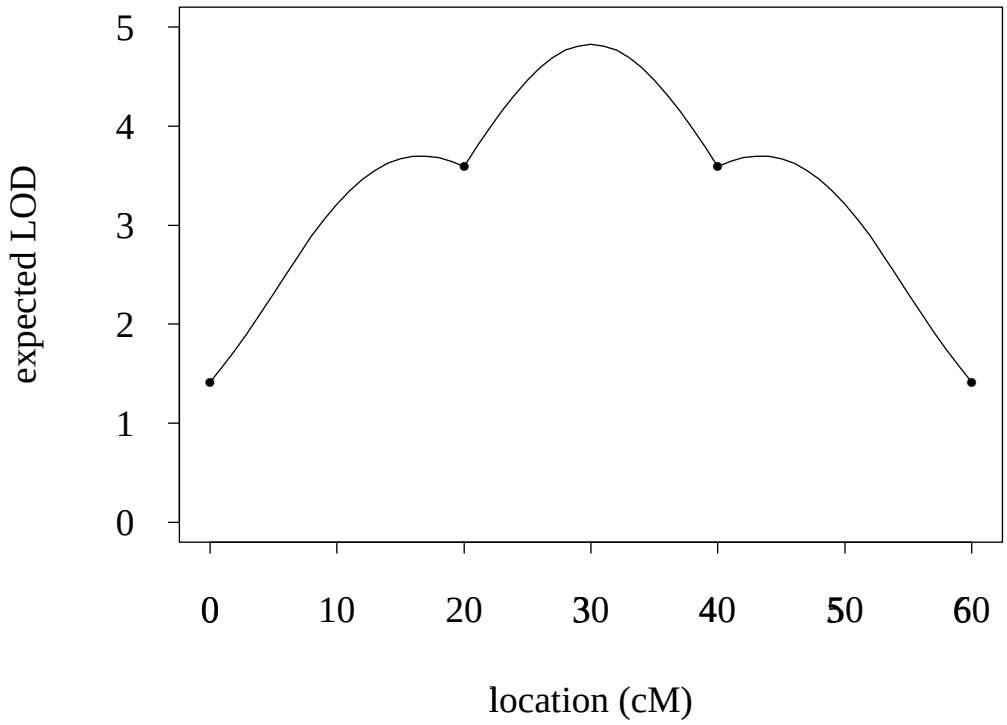


**Figure 2.4**: Expected LOD curve, with two QTLs located at 15 and 45 cM, and with markers at 0, 20, 40 and 60 cM.

Note that the ELOD curve is maximized at 30 cM, even though the simulated QTLs were at 15 and 45 cM. This gives rise to the term "ghost QTL." Forward selection here would give bad results. We would generally pinpoint the first QTL at around 30 cM, and then search for a second QTL, and so would be completely mistaken.

But this "ghost QTL" problem turns out to be an artifact of interval mapping. Figure 2.5 shows the ELOD curves for the above example, using marker spacings of 20, 10 and 5 cM.

When the markers are more tightly spaced, the ghost QTL disappears. The ELOD curves are not maximized exactly at the true QTL locations, but things do get better as marker density increases. Note that if one considered only the marker loci, one would not

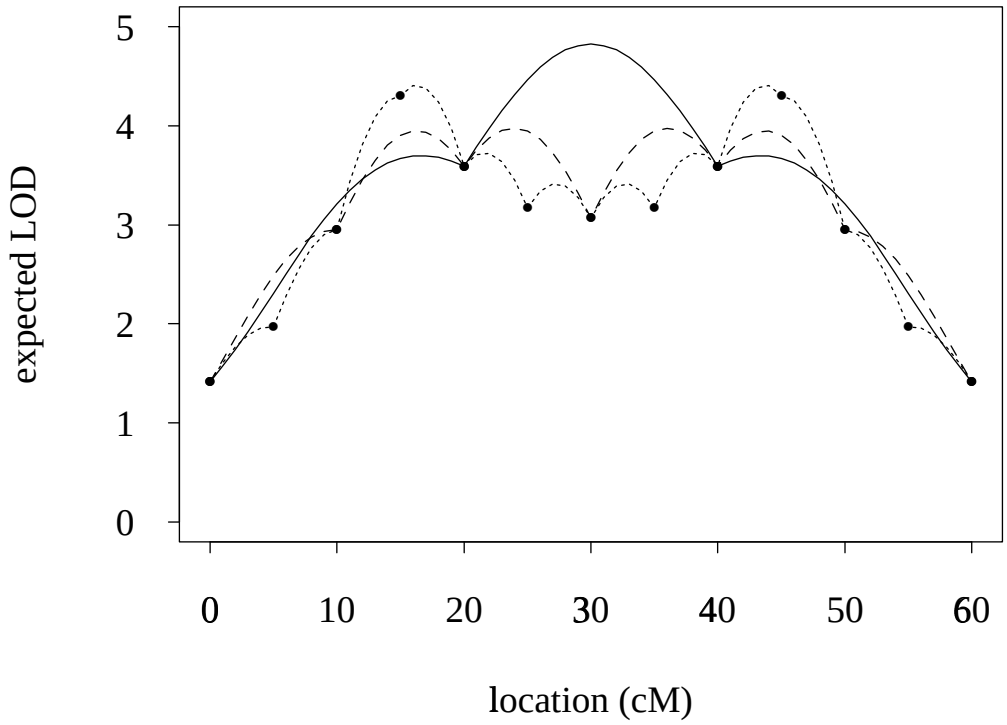


**Figure 2.5**: Expected LOD curves, with two QTLs located at 15 and 45 cM. The solid line, dashed line, and dotted line correspond to using equally-spaced markers at spacings of 20, 10 and 5 cM, respectively.

be so misled. The marker loci at which the LOD is maximized are those closest to the true QTLs. We will investigate this problem further in Chapter 3.

As an alternative to forward selection, several authors have recommended performing a full two-dimensional search for QTLs (Haley and Knott 1992; Hyne and Kearsey 1995; Wu and Li 1994, 1996). Instead of fixing the location of one QTL and then searching for an additional one, the locations of both QTLs are allowed to vary simultaneously. A great deal more computation must be performed. Extending this method to more than two QTLs, as recommended by Wu and Li (1996), is possible in principle, but the computation requirements would very quickly become prohibitive.

One problem that these authors have not discussed carefully is the question of when to add an additional QTL: how much of an increase in LOD should we require before allowing an additional QTL? Such guidelines are necessary, if one is to use these methods in practice.

### 2.2.3 Composite interval mapping and MQM mapping

Jansen and Zeng independently developed a method which attempts to reduce the multi-dimensional search for identifying multiple QTLs to a one-dimensional search (Jansen 1993; Jansen and Stam 1994; Zeng 1993, 1994). This is done using a hybrid between interval mapping and multiple regression on marker genotypes. By including other markers (on the same chromosome and on different chromosomes) as regressors while doing interval mapping, one hopes to control for the effects of QTLs in other intervals, so that there will be greater power in detecting a QTL, and so that the effects of the QTLs will be estimated more precisely. Jansen called the method MQM mapping (short for "marker-QTL-marker" or "multiple QTL models"); Zeng called it composite interval mapping.

The method is performed as follows. We choose a set of markers, $S$, to control for background genetic variation. Then, we perform a genome scan, like in interval mapping. At each locus in the genome, we hypothesize the presence of a QTL, and we write

$$y \sim \text{normal}(\mu + \beta z + \sum_{j \in S^\star} \beta_j x_j, \sigma^2),$$

where $y$ is the phenotype, $z = 1$ or 0, according to the whether the genotype at the putative QTL is HL or LL, $x_j = 1$ or 0, according to whether the genotype at the $j$th marker is HL or LL, and $S^\star$ is a subset of our set of markers, $S$, where we exclude any markers that are within, say, 10 cM of the putative QTL. Under this model, the contribution of each individual to the likelihood has the form of a mixture of two normal distributions with means $\mu + \sum_{j \in S^\star} \beta_j x_j$ and $\mu + \beta + \sum_{j \in S^\star} \beta_j x_j$, with mixing proportions equal to the conditional probabilities of the individual having QTL genotype HL and LL, given its marker genotypes. The EM algorithm, or a variant called the ECM algorithm (Meng and Rubin 1993), can be used to maximize the likelihood function.

As in interval mapping, at each locus, a likelihood ratio or LOD score is calculated, comparing the likelihood assuming that there is a QTL at that locus, to the likelihood assuming that there is not a QTL there, in which case we imagine that all progeny have phenotypes which are normally distributed with mean $\mu + \sum_{j \in S^\star} \beta_j x_j$ and variance $\sigma^2$. The LOD score is plotted as a function of genome position, and is compared to a genome-wide threshold. As in interval mapping, areas of the genome for which the LOD curve exceeds the threshold are said to contain a QTL.

The genome-wide threshold is obtained by considering the distribution of the max-

imum LOD score under the hypothesis of no segregating QTLs anywhere in the genome. This distribution should take into account the selection of the set of marker regressors, $S$. The distribution can be estimated by simulating a set of data under the hypothesis of no segregating QTLs, performing the entire procedure, and calculating the maximum LOD curve obtained, and then repeating the process a number of times. The 95th percentile of these maximum LOD scores is used as the threshold.

The key problem in this method is the choice of which markers to use as regressors: using too many markers will increase the variance of the LOD score, and thus will decrease the power for detecting QTLs. Jansen (1993) and Jansen and Stam (1994) used backward deletion, with Akaike's Information Criterion (AIC) or a slight variant, to pick the subset of markers. Zeng (1994) recommended using either all markers, dropping those within 10 cM of the putative QTL, or using all markers that are not linked to the putative QTL. Basten et al. (1996), in a manual for the program QTL Cartographer, recommend using forward selection up to a fixed number of markers, say five, and then dropping any markers that are within 10 cM of the putative QTL.

We have found that the methods that Zeng (1994) originally recommended, using all markers or all markers not linked to the putative QTL, work very badly. Including so many markers increases the corresponding LOD threshold to such a large value that power is reduced to almost zero. Only QTLs with extremely large effect will be found by this method.

The performance of the other methods for choosing the set of marker regressors depends on how many markers are chosen. And once we have found a way to choose this set, the task of identifying QTLs is essentially done: the best set of markers to use is exactly the set of markers which are closest to the underlying QTLs. In Chapter 4, we present some simulation studies which assess the performance of these methods.

### 2.2.4 Markov chain Monte Carlo

Satagopan et al. (1996) have applied the Markov chain Monte Carlo (MCMC) method to the problem of identifying QTLs. MCMC is a very popular approach to solving very complex statistical problems, especially those which include a large amount of missing information. Gelman et al. (1995) gives a very good introduction to the subject.

Consider again a backcross. Satagopan et al. (1996) consider a single linkage group.

(The method can be extended to several linkage groups in a straightforward way.) Consider $n$ progeny. Let $y_i$ be the phenotype for individual $i$. Suppose there are $M$ markers, at locations $D = (D_1, D_2, ..., D_M)$, in cM, from the left end of the linkage group. Let $x_{ij} = 1$ or 0, according to whether individual $i$ has genotype HL or LL at the $j$th marker.

Let $S$ be the number of segregating QTLs, and let $\lambda = (\lambda_1, \ldots, \lambda_S)$ be their locations, in cM, from the left end of the linkage group. Let $z_{ij} = 1$ or 0, according to whether individual $i$ has genotype HL or LL at the $j$th QTL. Let $\beta_j$ be the effect of the $j$th QTL, and assume that the environmental variation is normally distributed, with variance $\sigma^2$. Let $\mu$ be the mean of individuals for whom $z_{ij} = 0$ for all $j$.

As shorthand, we'll write $y = (y_1, \ldots, y_n)$, $x_i = (x_{i1}, \ldots, x_{iM})$, $x = (x_1, \ldots, x_n)$, and similarly for $z_i$, $z$ and $\beta$. Also, let $\theta = (\mu, \beta, \sigma)$.

We have

$$y_i | z_i, \theta \sim \text{normal}(\mu + \sum_{j=1}^{S} \beta_j z_{ij}, \sigma^2).$$

This gives the likelihood

$$L(\lambda, \theta | y, x, D) = \prod_{i=1}^{n} \sum_{q} f(y_i | z_i = q, \theta) \mathbf{Pr}(z_i = q | \lambda, x_i, D)$$

where the sum over $q$ is over the $2^S$ possible QTL genotypes for individual $i$ and where $f$ is the conditional (normal) density for $y$.

Satagopan et al. (1996) use a full Bayesian framework, meaning that they assign a prior probability distribution to the unknown parameters $(\lambda, \theta)$, say $p(\lambda, \theta)$, and then look at their posterior distribution, given the data, $p(\lambda, \theta | y, x, D)$.

The goal of the MCMC method is thus to estimate the posterior distribution of the unknown parameters. This is done by creating a Markov chain whose stationary distribution is the desired posterior distribution.

Simulating from this chain gives a sequence $(\lambda_0, \theta_0), (\lambda_1, \theta_1), \ldots (\lambda_N, \theta_N)$. Estimates of the desired parameters, such as the QTL effects, $\beta_j$, are obtained by averaging over these samples. Interval estimates for the QTL locations can be obtained by looking at the smallest intervals which contain, say, 95% of the samples.

In order to determine the number of QTLs, $S$, Satagopan et al. (1996) run separate chains for different values of $S$, and use Bayes factors. In brief, for each value of $S$, they use their samples to estimate the probability of the data given the model, $p(y, x | S)$. They estimate the number of QTLs to be the value of $S$ for which this estimated probability

is large. If one were willing to give a prior on the number of QTLs, say $\mathbf{Pr}(S = s)$, the posterior distribution for $S$ could be calculated

$$\mathbf{Pr}(S = s|y, x) = \frac{p(y, x|S = s)\mathbf{Pr}(S = s)}{\sum_s p(y, x|S = s)\mathbf{Pr}(S = s)}$$

The estimated number of QTLs would then simply be the value of $S$ with the largest posterior probability.

A later report (Satagopan and Yandell 1996), using an idea developed by Green (1995), describes how to allow the unknown number of QTLs, $S$, to be included as an unknown parameter, so that a single Markov chain can be used to estimate $S$ along with the other parameters. Doing this requires placing a prior distribution on the number of QTLs.

We have skipped all of the details of the MCMC method. The difficulties in applying this approach are entirely in those details. First, you need to create a Markov chain which has your posterior distribution as its stationary distribution. There are a number of standard ways to do this, such as the Gibbs sampler (Geman and Geman 1984) and the Metropolis-Hastings algorithm (Hastings 1970). The most important characteristic in the chain is that it mixes well: that it moves around the parameter space rather easily, and that it very quickly reaches its stationary distribution. Forming good Markov chains, and monitoring their behavior, is a delicate and sophisticated art.

The other important problem is in the determination of the number of QTLs. Whether we assign a prior to the unknown number of QTLs or use Bayes factors, we must make choices which balance the problem of missing real QTLs with that of including extraneous loci.

## 2.3 Discussion

Interval mapping and its approximations have been shown to provide little improvement in power over simple ANOVA at the marker loci. The advantage to interval mapping is that it gives improved estimates for QTL locations and effects. But all of these single QTL methods have difficulty in separating linked QTLs. Moreover, when searching for multiple QTLs all at once, one can control for some of the genetic variance, which may increase the power for detecting additional QTLs.

Interval mapping with forward selection has been harshly criticized, but large

multi-way searches are infeasible. Composite interval mapping, and the very high-powered Markov chain Monte Carlo, may provide ways around this, but there can be no universal solution. The problems of searching through the space of possible models and of determining the number of QTLs have not been solved. With composite interval mapping, one must determine how many markers to use as regressors; but having done that, one has practically determined the number of QTLs already. With Markov chain Monte Carlo, one must place a prior on the number of QTLs, or at least form a rule for determining the number of QTLs, given a set of Bayes factors.

Our feeling is that model selection methods using multiple regression methods were discarded too early, and should be considered further. MCMC is quite a hefty bit of machinery, and will no doubt perform quite well when in the hands of a highly skilled, experienced user. But it may be like using a chain saw to cut a loaf of bread; tearing a bit off with your hands does just as well, without the mess (or the gasoline). No one has shown that MCMC will perform better than multiple regression. In fact, for the data in Satagopan et al. (1996), interval mapping seems to give nearly identical results.

# Chapter 3

# Model Selection

Identifying quantitative trait loci (QTLs) is a model selection problem. We imagine that there are a finite number of QTLs segregating in the cross under study (and here we assume that they act additively), and we wish to estimate their number, their locations, and their effects. With this aim, we obtain data on the phenotypes of the progeny from an experimental cross, as well as the genotypes of these progeny at a set of marker loci, for which we have a genetic map.

In Chapter 2, we pointed out that, for typical QTL experiments, one is unable to resolve the locations of QTLs to positions within an interval between markers. Thus, there is very little lost in assuming that QTLs are located exactly at marker loci. This is the approach that we recommend: we dispense of interval mapping (inferring between marker loci), and attempt to choose a set of markers, which we identify as being at or near QTLs.

Consider again the example of a backcross. For each of $n$ progeny, we obtain phenotypes $(y_i)$ and genotypes $(x_{ij}) = 1$ or 0 for a set of $M$ markers, indexed by $j$. We write

$$y_i = \mu + \sum_{j=1}^{M} \beta_j x_{ij} + \epsilon_i$$

where the $\epsilon_i$ are iid normal$(0, \sigma^2)$. We seek the set of markers for which $\beta_j \neq 0$.

In identifying this set of markers, we can make two errors: we can miss some markers that are important, and we can include some extraneous ones. How we choose to balance these two errors should depend on the goal of the QTL experiment.

Our hope, in viewing the problem in this way, was that we could be guided by previous work on selecting subsets of regressor variables, a problem which has been studied

extensively over many years. Unfortunately, most research in this area has focused on choosing models for prediction. In that scenario, most of the coefficients are believed to be non-zero, but, because of the large number of regressors, they are estimated with large variance. By dropping some of the regressors, a small bias is introduced, but the variances of the estimated coefficients may be reduced radically, and so the overall prediction error may be made smaller. (For a review of this subject, see Miller (1990).) Still, this work on prediction has much to say on our problem of finding the underlying model.

The problem of identifying the underlying model has two components. First, one must form a criterion for comparing models. Models with the same number of regressors are generally compared by the residual sum of squares (RSS) obtained after estimating the coefficients by least squares. The model with the smaller RSS is preferred. When comparing models with different numbers of regressors, one cannot simply choose the model with the smallest RSS, since when adding an additional regressor to a model, the RSS never increases. Thus, one must make a decision about what decrease in RSS is required before accepting an additional regressor.

The second part of the problem is that of searching through the space of models. With $M$ markers, there are $2^M$ possible models that must be considered. When $M = 40$, so that $2^M \approx 10^{12}$, it may be feasible to fit each possible model. But when $M = 200$, $2^M \approx 10^{60}$, and it will be impossible to fit each of the models. Thus, one must find a way to search through this large space, fitting a subset of the models which hopefully contains the ones that would be chosen if one were able to fit all models.

In the next section, we discuss various criteria for choosing between models. In Section 3.2, we discuss approaches to searching through model space. At the end of the chapter, we summarize our recommended approach for identifying QTLs.

## 3.1 Comparing models

Imagine that we were able to fit all possible models. Let $\Gamma$ denote the set of models, with $\gamma \in \Gamma$ written as an $M$-vector whose $i$th element $\gamma_i = 1$ or 0 according to whether or not the $i$th marker is included in that model. Let $q_\gamma$ denote the number of markers in the model $\gamma$, and let $\mathrm{RSS}(\gamma)$ denote the residual sum of squares after fitting $\gamma$ by least squares.

As mentioned above, in most approaches to selecting a subset of regressors, when choosing among models with the same number of regressors, the chosen model is that with

the smallest RSS. We'll write $\gamma_k$ to be the model with the smallest RSS, among models with $k$ regressors, so that

$$\gamma_k = \arg \max_{\gamma: q_\gamma = k} \mathrm{RSS}(\gamma)$$

Thus, $\gamma_M$ is the full model, with all markers included, and $\gamma_0$ is the model including no markers.

Two of the most well known methods for choosing subsets of regressor variables are Mallows' $C_p$ and adjusted-$R^2$ (Miller 1990). Both are included in most standard statistical packages. Mallows' $C_p$ has the form

$$C_p(\gamma) = \frac{\mathrm{RSS}(\gamma)}{\hat{\sigma}^2} - (n - 2q_\gamma)$$

where $\hat{\sigma}^2$ is some estimate of $\sigma^2$, for example $\mathrm{RSS}(\gamma_M)/(n - M)$. The chosen model is that which minimizes this criterion.

Adjusted-$R^2$ has the form

$$R^2_a(\gamma) = 1 - \frac{\mathrm{RSS}(\gamma)}{\mathrm{RSS}(\gamma_0)} \cdot \frac{(n-1)}{(n - q_\gamma)}$$

It is easy to see that maximizing this criterion is equivalent to minimizing

$$\hat{\sigma}^2(\gamma) = \mathrm{RSS}(\gamma)/(n - q_\gamma)$$

Both Mallows' $C_p$ and adjusted-$R^2$ tend to include a large number of extraneous variables, and so are unsatisfactory for our purposes.

Two more modern approaches for choosing subsets of regressor variables are cross-validation and the bootstrap. In both of these approaches, an estimate of the mean squared error of prediction (MSEP) is obtained. The chosen model has the smallest estimated MSEP.

In the simplest form of cross-validation, one of the observations is dropped, an estimated regression equation is formed using least squares with the other $(n-1)$ observations, and the value for the dropped observation is predicted using this regression equation, giving, say, $\hat{y}_{(i)}(\gamma)$. The process is repeated, dropping each of the $n$ observations one at a time, and the MSEP is estimated using

$$\sum_{i=1}^{n} [y_i - \hat{y}_{(i)}(\gamma)]^2$$

This is often called the PRESS statistic (Miller 1990). This can be written in the form

$$\sum_{i=1}^{n}[y_i - \hat{y}_i(\gamma)]^2/(1-h_i)^2$$

where $\hat{y}_i(\gamma)$ is the predicted value for the $i$th observation, using the regression equation estimated with all $n$ observations, and where $h_i$ is the $i$th diagonal element of the so-called "hat" matrix, $X_\gamma(X_\gamma'X_\gamma)^{-1}X_\gamma'$. As a result, one need not actually perform $n$ different regressions, with the observations dropped one at a time; the statistic can be calculated with information obtained from the single regression using the model $\gamma$.

More generally, one may consider dropping several observations at once. Dropping $k$ observations at a time leads to a $k$-fold cross-validation estimate of the MSEP. The trick used above will not work for $k$-fold cross-validation, and so a great deal of computation will have to be expended in estimating the MSEP in this way. The bootstrap, described in Shao (1996), gives similar estimates of the MSEP, and will not be discussed further here.

Another approach is to minimize a criterion of the form

$$\Psi(\gamma) = \log \mathrm{RSS}(\gamma) + q_\gamma D(n)/n$$

where $D(n)$ is some function of $n$. Taking $D(n) = 2$ gives Akaike's information criterion (Akaike 1969). Taking $D(n) = \log n$, one obtains Schwartz's BIC (Schwartz 1978), and with $D(n) = \log\log n$, one obtains the criterion of Hannan and Quinn (1979).

Consider our sequence of models $\gamma_0, \gamma_1, \ldots, \gamma_M$, and suppose that $\gamma_k$ minimizes the above criterion $\Psi(\gamma)$. Then

$$\begin{aligned} & \Psi(\gamma_{k+1}) \geq \Psi(\gamma_k) \\ \Longrightarrow \quad & \log \mathrm{RSS}(\gamma_{k+1}) + (k+1)D(n)/n \geq \log \mathrm{RSS}(\gamma_k) + kD(n)/n \\ \Longrightarrow \quad & \log[\mathrm{RSS}(\gamma_{k+1})/\mathrm{RSS}(\gamma_k)] \geq -D(n)/n \\ \Longrightarrow \quad & \mathrm{RSS}(\gamma_{k+1})/\mathrm{RSS}(\gamma_k) \geq \exp[-D(n)/n] \end{aligned}$$

Similarly,

$$\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1}) \leq \exp[-D(n)/n]$$

The ratio $\mathrm{RSS}(\gamma_{k+1})/\mathrm{RSS}(\gamma_k)$ approaches 1 as $k$ becomes large, and is often strictly increasing in $k$. In that case, minimizing the above criterion $\Psi(\gamma)$ is equivalent to choosing the model $\gamma_k$ where $\mathrm{RSS}(\gamma_{k+1})/\mathrm{RSS}(\gamma_k)$ is above and $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ is below the value $\exp[-D(n)/n]$ (a function of the sample size $n$ and the choice of the function $D$). Figure 3.1

contains a plot of $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ versus $k$, for a set of simulated data, with $n = 200$. The dashed line corresponds to using $D(n) = 2\log n$. In this case, we would choose a model with three regressors.

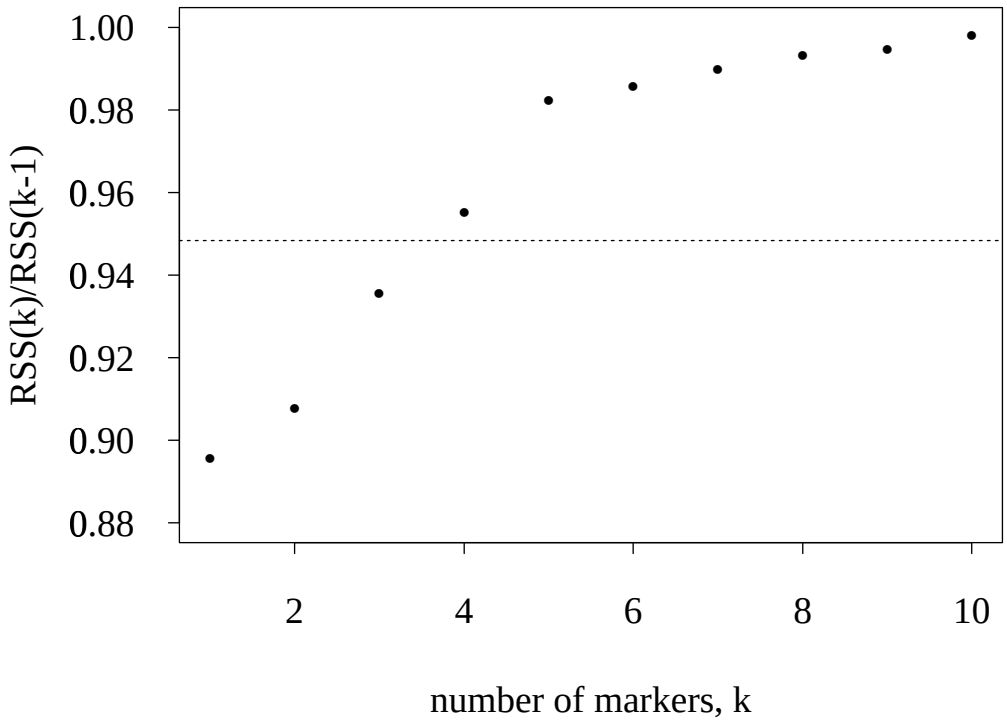


**Figure 3.1**: A plot of $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ against $k$ for some simulated data, where $\gamma_k$ minimizes $\mathrm{RSS}(\gamma)$ among models of size $k$.

Thus, with criteria of this form, the statistic $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ is of chief interest. When considering whether to add an additional regressor, it is the increase in RSS when adding the regressor, as a proportion of the RSS of the best model with one fewer regressors, that is deemed important. The amount of increase required, once the form of the function $D(n)$ has been chosen, depends only on the sample size, $n$.

After viewing it in this way, this type of criterion seems a quite reasonable one. Further support for these criteria lies in the consistency of the resulting procedures. With a fixed number of possible regressors, and provided that $D(n)/n \to 0$ and $D(n)/\log\log n \to \infty$, the criterion $\Psi(\gamma)$ gives a consistent estimate of the underlying model, meaning that, as the sample size increases, the probability that the correct model is chosen converges to 1 (Rao and Wu 1989).

We have concentrated on $D(n) = \delta \log n$, so that we have the criterion

$$\text{BIC}(\gamma)\text{-}\delta = \log \text{RSS}(\gamma) + q_\gamma \delta \log n/n$$

Letting $\delta = 1$, this gives Schwartz's BIC criterion. We've found that using $\delta = 1$ works rather poorly, including far too many extraneous regressor variables. Using instead $\delta = 2$ or 3 gives much better results. A smaller $\delta$ will include more regressors, thus giving a better chance to find the correct ones, but also including extraneous ones with greater frequency.

A further approach to this model selection problem is to place prior probabilities on each of the possible models, as well as on the parameters (the coefficients $\beta$ and the residual variance $\sigma^2$), and to use Bayes' theorem to calculate the posterior distribution of the models given the data. If the goal were to pick out just one model, one could choose that which gives the largest posterior probability.

As an example, consider one of the priors discussed in Smith (1996). We let $y|\gamma, \beta_\gamma, \sigma^2 = X_\gamma \beta_\gamma + \epsilon$, where $\epsilon \sim \text{normal}(0, \sigma^2)$, and use the prior

$$\beta_\gamma \sim \text{normal}(0, c\sigma^2 (X_\gamma' X_\gamma)^{-1})$$

$$p(\sigma^2|\gamma) \propto 1/\sigma^2$$

$$p(\gamma) \propto (c/d)^{q_\gamma/2}$$

Let $c \to \infty$, resulting in a diffuse improper prior. Then, integrating out $\beta_\gamma$ and $\sigma^2$,

$$p(\gamma|y) = d^{-q_\gamma/2} \text{RSS}(\gamma)^{-n/2}$$

and so

$$-\frac{2}{n} \log p(\gamma|y) = \log \text{RSS}(\gamma) + q_\gamma \log d/n$$

Thus, taking $\log d = D(n)$, we see that, with this prior, maximizing the posterior probability for $\gamma$ is equivalent to minimizing the criterion discussed above. One might consider this as further support for the use of the criterion $\Psi(\gamma)$. The only real justification for a procedure, however, is in its performance. We will study the performance of this criterion in the next chapter.

There are a few additional methods which require a nested sequence of models, as obtained from methods such as forward selection and backward deletion (which are discussed in the next section). With such a sequence, one may work from the null model $\gamma_0$ to the full model $\gamma_M$, performing a hypothesis test at each step, testing whether the coefficient for

the added regressor is 0. The first time that the null hypothesis is not rejected, one stops, and picks the model with all regressors whose coefficients had been found to be non-zero. The test performed may be a simple t-test, or may be a permutation test, as discussed in Doerge and Churchill (1996).

To summarize, there are a large number of procedures for choosing between subsets of regressor variables. Many of them, such as Mallows' $C_p$, cross-validation and the bootstrap, are aimed at finding a model with minimum mean squared prediction error. Criteria such as those we denote BIC-$\delta$ have a reasonable interpretation, have been shown to give a consistent procedure in the case of a fixed number of possible regressors, and can be interpreted as the negative log posterior probability under a certain type of prior.

## 3.2 Searching model space

The number of possible models is very large. If there are more than around 40 markers, it will be impossible to fit each of them. For example, with 200 markers, there are $2^{200} \approx 10^{60}$ different models. Thus, one must find a way to search through this large space of models, hopefully in a way which allows one to pick out the good ones—those that would have been chosen if one could fit all possible models.

In the case that the number of markers is only marginally large, one may use a branch-and-bound procedure to pick out the best subsets of each size, without actually fitting all of the possible subsets (Miller 1990). Suppose, for example, that one has fit a subset with four markers, including the marker $x_1$, and has found that the RSS for this model is smaller than that for the model containing all markers *except* $x_1$. Then one can conclude that the best model with four markers *must* include $x_1$. This technique finds the best subsets of each size, with considerable computational savings over an exhaustive search.

With many markers, a branch-and-bound type of search cannot be used any more than an exhaustive one can. And so one is led to techniques such as forward selection and backward deletion. In forward selection, one first looks at all models with one regressor, and chooses the one which gives the smallest RSS, say $x_{j_1}$. Next, one considers models which include $x_{j_1}$ and one other regressor, and adds the marker which gives the greatest decrease in RSS, say $x_{j_2}$. Then, one looks at models with $x_{j_1}$, $x_{j_2}$ and one other regressor. This process is continued until all regressors have been added. The result is a sequence of $M$

nested models, where $M$ is the number of possible regressors. In obtaining this sequence, $M + (M-1) + (M-2) + \ldots + 1 = M(M+1)/2$ models were fitted. Of course, one need not fit this entire sequence of models. After fitting 10 or so regressors, it may become clear that no further regressors will improve the model, and so the process can be stopped early. If $M = 200$, there are $2^{200} \approx 10^{60}$ possible models. Performing forward selection all the way to the full model, $200 \times 201/2 = 20,100$ models would be fit. If one stopped at 10 regressors, only 1,955 models would be fit.

In backward deletion, one first fits models with all but one of the possible regressors. The model with the *largest* RSS is chosen, so that the regressor resulting in the smallest increase in RSS (apparently the least important regressor) is dropped. One then considers this model with one regressor dropped, and again drops the regressor which results in the smallest increase in RSS. The process is repeated until all regressors have been discarded. Again, the result is a sequence of nested models. The sequence may be quite different from that obtained using forward selection. Backward deletion cannot be performed if one has more regressors than observations (i.e., more markers than progeny), but hopefully that does not often occur, since, as will be seen in Chapter 4, large numbers of progeny may be more important than large numbers of markers, provided that one has enough markers to at least cover the genome.

Forward selection and backward deletion are quite easy to implement using the Sweep algorithm (see Thisted (1988)). There is a great savings in computation when using these methods, since a small fraction of the possible models are fit. This savings is also a cost, however: we see only a fraction of the possible models, and we might not see the good ones. With forward selection, once a regressor is included, it will be retained in all further models. With backward deletion, once a regressor is dropped, it will be excluded in all further models.

Forward selection has a particularly bad reputation. One can find quite simple situations in which forward selection will miss the correct model, even when the sample size is very large. Consider, for example, three regressors, $x_1$, $x_2$ and $x_3$, which are multivariate normal, each with mean 0 and variance 1, and with the following correlation matrix

$$\begin{pmatrix} 1.0 & 0.0 & 0.7 \\ 0.0 & 1.0 & 0.7 \\ 0.7 & 0.7 & 1.0 \end{pmatrix}$$

Thus, $x_1$ and $x_2$ are independent, but they are each highly correlated with $x_3$. Now suppose that $y = x_1 + x_2 + \epsilon$, where $\epsilon \sim \text{normal}(0, \sigma^2)$. The partial regression coefficients of $y$ on each of $x_1$, $x_2$ and $x_3$ are 1.0, 1.0, and 1.4, respectively. Thus, with a large amount of data, forward selection would first choose $x_3$, even though it does not belong in the model.

Backward elimination does not suffer from this problem, at least with a large sample. An and Gu (1985) showed that, when using BIC, and in the case of a fixed number of regressors, the backward elimination procedure is consistent, meaning that, as the sample size increases, the probability of choosing the correct model converges to 1. The proof is easily seen to apply to our more general criteria, denoted BIC-$\delta$.

The problem with forward selection, as seen above, arises when one of the extraneous regressors ($x_3$ in our example) mimics a set of regressors which belong in the model ($x_1$ and $x_2$ in our example). But in the situation discussed in this thesis, the regressors are genetic markers which, under the assumption of no interference, follow a Markov chain. Thus, given any one marker, the markers to its left are conditionally independent of the markers to its right. This raises the possibility that the forward selection procedure, using a BIC-type criterion, is consistent, and indeed we've shown this to be true. The proof is given in the Appendix, beginning on page 93.

One may also combine the forward selection and backward deletion methods with a branch-and-bound approach: one may perform forward selection or backward deletion to obtain a model with, say, 30 regressors, and then use branch-and-bound to pick out the best subsets which contain only regressors from that set of 30. This method is computationally feasible, and improves on forward selection and backward deletion, in that it sees a great deal more models.

A different approach to searching the space of models is to use a randomized algorithm, such as Markov chain Monte Carlo (MCMC). For this method, one places a prior probability on each model and on the other unknown parameters, and then forms the posterior distribution of the models given the data. One then forms a Markov chain whose stationary distribution is this posterior distribution. Simulating the Markov chain gives a sequence of models, a sort of walk through the space of models, which will, eventually, spend more time at models which have a high posterior. Whereas this method is usually used to obtain an approximation of the posterior distribution, and especially to find the region with highest posterior, here we consider it simply as a method of searching the space of models.

There are a number of standard ways to form a Markov chain with the desired stationary distribution. Using the prior discussed in the previous section, Smith (1996) used a Gibbs sampler to obtain a Markov chain whose stationary distribution is the posterior distribution for the models given the data,

$$p(\gamma|y) = d^{-q_\gamma/2}\mathrm{RSS}(\gamma)^{-n/2}$$

The method is as follows. First, pick an initial model, $\gamma^{(0)}$, for example, the null model, $\gamma_0$, or the model obtained after performing a method such as forward selection. Then, at step $t$, we cycle through the $M$ different markers. For each $j = 1, 2, \ldots, M$, we draw $\gamma_j^{(t)}$ from the distribution

$$p(\gamma_j|\gamma_{-j}^{(t-1)}, y)$$

where $\gamma_{-j}^{(t-1)}$ is composed of all of the elements of $\gamma$, except for $\gamma_j$, at their current values. So for $i < j$, it contains the $\gamma_i$ for the current step, $t$, and for $i > j$, it contains $\gamma_i$ for the previous step, $t-1$. Thus we have

$$\gamma_{-j}^{(t-1)} = (\gamma_1^{(t)}, \gamma_2^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_{j+1}^{(t-1)}, \ldots, \gamma_M^{(t-1)})$$

It is easy to see that, for the posterior written above,

$$\mathbf{Pr}(\gamma_j = 1|\gamma_{-j}, y) = \frac{\mathrm{RSS}(\gamma_1, \ldots, \gamma_{j-1}, 1, \gamma_j, \ldots, \gamma_M)^{-n/2}}{\mathrm{RSS}(\ldots, 1, \ldots)^{-n/2} + \sqrt{d} \cdot \mathrm{RSS}(\ldots, 0, \ldots)^{-n/2}}$$

The most important characteristic for the Markov chian is that it mixes well—that it travels through the space of models with relative ease, not getting stuck in certain places. We have implemented the above Markov chain, and have found that it works very well. In 1000 steps of the chain, it will visit around 300–500 different models, and it will almost always visit the "best" of those models (i.e., the one with the largest posterior probability) within the first 100 steps.

In this chapter, we've discussed several different methods for searching through the space of models: branch-and-bound (which is appropriate only when the number of markers is small), forward selection, backward deletion, forward selection or backward deletion followed by branch-and-bound, and MCMC. These methods differ greatly in the amount of computation (and programming effort) required. The obvious question is: does this additional computation buy anything? To answer this question, we have simulated data from a backcross, applied each of the methods, and compared the results.

We simulated a backcross of 250 progeny, obtained from inbred lines, with nine chromosomes, each of length 100 cM and having 11 equally spaced markers per chromosome (thus at a 10 cM spacing). The recombination process was assumed to exhibit no interference. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$.

We modelled five QTLs, with equal additive effect 0.5. One QTL was located at the center of chromosome 1, two QTLs were located on chromosome 2 at 30 and 70 cM, and two QTLs were located on chromosome 3 at 30 and 70 cM. The linked QTLs were either in coupling (effects of equal sign) or repulsion (effects of opposite sign). The QTLs were assumed to act additively. The heritability was 0.30 and 0.17 when the linked QTLs were in coupling and repulsion, respectively. Note that all QTLs were located exactly at marker loci.

For each QTL model and for each sample size, we performed 1000 simulations, and applied five methods: forward selection, backward deletion, forward selection to 30 markers, followed by branch-and-bound, backward deletion to 30 markers, followed by branch-and-bound, and the MCMC method described above (started at the null model, $\gamma_0$, and taking 200 steps). The BIC-2.5 criterion was used, so that the chosen model, by each method, was that which minimizied $\log \mathrm{RSS}(\gamma) + 2.5 q_\gamma \log n/n$, among the models seen.

The result of the application of each method was a set of marker loci indicated to be at or near QTLs. In assessing the results, we defined a chosen marker to be correctly identifying a QTL if it was within 20 cM of a QTL; otherwise it was deemed incorrect. If more than one chosen marker were within 20 cM of the same QTL, one was called correct and the others were called incorrect.

Table 3.1 displays the distribution of the number of correctly chosen markers and the number of simulations with at least one incorrectly chosen marker, in 1000 simulations of the model in which the linked QTLs were in coupling. Note that, for each of the methods, no more than 3 in 1000 simulations resulted in more than one incorrectly identified marker; thus we summarize these data only with the number of simulations giving at least one incorrect marker.

The most striking feature of this table is that the different methods perform substantially the same. The additional effort expended in the branch-and-bound approach and in MCMC buys nothing; forward selection performs as well as the other methods. The methods generally find around three QTLs, and seldom find all five or only one.

**Table 3.1**: Distribution of the number of correctly chosen markers, and number of simulations with at least one incorrectly chosen marker, in 1000 simulations of a model containing five QTLs with two pairs linked in coupling, and using 250 progeny.

| | | | branch-and-bound | | |
|---|---|---|---|---|---|
| correct | forw | back | forw | back | MCMC |
| 5 | 13 | 9 | 13 | 13 | 12 |
| 4 | 120 | 129 | 105 | 126 | 125 |
| 3 | 502 | 478 | 488 | 469 | 499 |
| 2 | 356 | 369 | 386 | 371 | 355 |
| 1 | 9 | 14 | 8 | 20 | 9 |
| 0 | 0 | 1 | 0 | 1 | 0 |
| # incor. | 59 | 56 | 52 | 66 | 59 |

**Table 3.2**: Distribution of the number of correctly chosen markers, and number of simulations with at least one incorrectly chosen marker, in 1000 simulations of a model containing five QTLs with two pairs linked in repulsion.

| | | | branch-and-bound | | |
|---|---|---|---|---|---|
| correct | forw | back | forw | back | MCMC |
| 5 | 10 | 8 | 11 | 8 | 11 |
| 4 | 17 | 9 | 15 | 10 | 17 |
| 3 | 91 | 92 | 84 | 86 | 93 |
| 2 | 164 | 142 | 141 | 135 | 163 |
| 1 | 402 | 392 | 414 | 376 | 401 |
| 0 | 316 | 357 | 335 | 385 | 315 |
| # incor. | 43 | 45 | 48 | 48 | 42 |

Table 3.2 displays the distribution of the number of correctly chosen markers and the number of simulations with at least one incorrectly chosen marker, in 1000 simulations of the model in which the linked QTLs were in repulsion.

Again, all five methods perform substantially the same, though here, they do a much worse job of identifying the QTLs. More than 30% of the time, they do not find even one QTL. And they find more than one QTL less than 30% of the time.

In summary, we've shown that, when using a BIC-type criterion, forward selection, a method which generally has a very bad reputation, is consistent. More importantly, though, our simulation studies suggest that it performs as well as more complicated approaches for searching the space of models. We do not want to discourage the use of methods such as MCMC, but it should be stressed that the very simple forward selection

approach, which is very much faster, and which is implemented in most standard statistical packages, may perform just as well. The difficulty here may not be the search through model space so much as the choice of criterion for comparing models, as well as the very weak information contained in the data.

## 3.3 Recommended approach

A brief summary of our recommended approach for identifying QTLs, assumed to act additively, is in order. First, we dispense of interval mapping, and consider only the marker loci, since, with the typical experiments performed, one is generally unable to localize a QTL to a particular interval, and so little is lost by doing this.

Next, we compare models using a BIC-type criterion, of the form

$$\text{BIC-}\delta = \log \text{RSS}(\gamma) + \delta q_\gamma \log n/n$$

This type of criterion is reasonable, in that it focuses on the change in RSS when a marker is added, as a proportion of the RSS of the best model with one fewer markers. Moreover, BIC-type criteria have been shown to give a consistent estimate of the model. The value $\delta$ should be chosen to give an appropriate balance between the error of missing important QTLs and the error of including extraneous loci. Values between 2 and 3 seem reasonable; smaller values of $\delta$ give a greater chance of detecting QTLs, but also include extraneous loci with greater frequency. The appropriate $\delta$ to use will be seen to depend on the number of markers used and the number and effects of the QTLs segregating in the cross.

The search through the space of models can be done with simple forward selection. A more extensive search, using, for example, Markov chain Monte Carlo, appears to provide little improvement over forward selection.

The great benefit of this approach, using forward selection at marker loci with a BIC-type criterion, is that it can be performed quite easily with the standard statistical software. The more flashy methods described in the previous chapter require specially designed computer programs, and may give no gains in performance.

The uncertainty in the estimated location of a QTL is a very important issue, and the above approach does not immediately address it. In interval mapping, this problem is dealt with by comparing the maximum LOD score to LOD scores at loci near the maximum.

(Recall that the LOD score is a $\log_{10}$ likelihood ratio.) A one- or two-LOD support interval, the interval in which the LOD score is within one or two of its maximum, is used to indicate the most likely location for the QTL. A similar approach can be used with our method. One can compare the chosen model with models in which a putative QTL is relaced by its neighbors. Let $\gamma$ denote the chosen model, and let $\gamma'$ denote the model with a putative QTL replaced by a neighboring marker. Under the normal model, the change in LOD score is simply

$$\frac{n}{2} \log_{10} \left[ \frac{\mathrm{RSS}(\gamma')}{\mathrm{RSS}(\gamma)} \right]$$

Finding the neighboring markers for which this value is less than one or two, one can obtain a rough idea of the uncertainty in the estimated location of the QTL.

# Chapter 4

# Simulations

In this chapter, we present the results of a number of simulation studies whose aim was to compare the performance of the more prominent methods for identifying QTLs and to explore our ability to identify QTLs using different numbers of progeny and different marker densities, as a function of the number of segregating QTLs and the sizes of the effects of the QTLs. Simulations are necessary, because the methods for identifying QTLs are too complex to be assessed by analytical means, at least in the situations in which they would be used in practice.

Most authors have used simulations to demonstrate their methods for finding QTLs. Many have presented the results of applying their method to a single data set (Jansen 1993; Knapp 1991; Lander and Botstein 1989; Zeng 1994), a practice which precludes a true assessment of the method's performance. Others consider only very simple situations, such as simulating only one or two chromosomes with one or two segregating QTLs (Haley and Knott 1992; Kearsey and Hyne 1994). In practice, most QTL studies involve a search over ten or more chromosomes, and very often there is evidence for at least a moderate number of segregating QTLs (from three or four to as many as a dozen). A method's ability to detect QTLs in simulation studies which use very limited searches and in which only a small number of QTLs are allowed will say little about its performance in the more complex situations where the method is anticipated to be used.

Also missing from the literature is a careful comparison of the performance of the many methods available for identifying QTLs. It is surprising that such comparisons are not a routine part of the presentation of a new method. Before dropping a simple approach in favor of a more complex one, we should have evidence that the complexities of the new

approach will be accompanied by a real improvement in performance.

To design QTL mapping experiments, scientists require information about the power for detection that different methods provide. This will allow them to determine the number of progeny required to have a reasonable chance of achieving their goals. Lander and Botstein (1989) provided calculations of the approximate power for interval mapping in the case of a single segregating QTL, for backcrosses and intercrosses with different numbers of progeny and with different sizes of effect for the QTL. Others have followed this up, mainly with simulation studies, and generally considering a single QTL and a single chromosome (Carbonell et al. 1993; Darvasi et al. 1993; Knott and Haley 1992, Rebaï et al. 1995; van Ooijen 1992). The performance of interval mapping and other methods in the presence of multiple QTLs when searching over multiple chromosomes has not been well studied.

Our focus in this chapter will be on our ability to identify the QTLs segregating in a cross. We view the problems of estimating effects as well as the precise location of QTLs as secondary issues. In Section 4.1, we present the results of a study to compare the most promising methods. Section 4.2 contains a reproduction of the simulations found in Doerge and Churchill (1994). In the final section, we present a study of the ability of multiple regression, using forward selection, to identify QTLs using different sizes of experiments, varying the number of QTLs and their effects.

## 4.1 A comparison of methods

In this section, we discuss the results of a study aimed at comparing several different methods. Our focus is on identifying QTLs, and so we look only at whether the methods detect the simulated QTLs, and not at the estimated effects and the precision with which the location is estimated. We simulated a backcross with a moderate number of QTLs of small effect.

### 4.1.1 Methods

We compared four different methods for identifying QTLs: analysis of variance (ANOVA) at the marker loci, the method of Zeng (1994), forward selection using a BIC-type criterion, and forward selection using a permutation test at each stage (Doerge and Churchill 1994). These methods are described in Chapters 2 and 3.

Interval mapping (IM) was ignored, because it provides no improvement over simple ANOVA when using a relatively dense marker map (10 cM spacing or less) and a small or moderate number of progeny (500 or less), at least when it comes to identifying QTLs. This can easily be seen when inspecting the one- or two-LOD support intervals which accompany any application of IM: they invariably span several markers. The benefit of IM is in providing more precise estimates of QTL location and effects.

For Zeng's method, we used forward selection up to either 3, 5, 7 or 9 markers to obtain the set of regressors, and limited the search for QTLs to marker loci. With ANOVA and Zeng's method, we obtained genome-wide thresholds by performing 1000 simulations under the hypothesis of no segregating QTLs: the estimated threshold was the 95th percentile of the maximum LOD score across all markers. In addition, for these two methods, we required that the LOD dropped by at least 2.2 in base 10 (corresponding to 5 in base $e$) between "peaks" before we declared that two QTLs were identified. This value was obtained empirically (in other words, by trial and error).

The BIC-type criterion used is $\log \mathrm{RSS} + \delta q \log n/n$, where RSS is the residual sum of squares, $n$ is the number of progeny, $q$ is the number of markers in the model, and $\delta$ is either 2, 2.5 or 3. We use BIC-2, BIC-2.5 and BIC-3 to identify these criteria. For the permutation method, at each stage we used the 95th percentile of 500 permutations to determine whether to add another marker.

In the study described in this section, we simulated a backcross obtained from inbred lines, with nine chromosomes, each of length 100 cM and having 11 equally spaced markers per chromosome (thus at a 10 cM spacing). The recombination process was assumed to exhibit no interference. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$. We simulated experiments with 100, 250 and 1000 progeny.

We modelled either three or five QTLs, with equal additive effect 0.5. One QTL was located at the center of chromosome 1, and two QTLs were located on chromosome 2 at 30 and 70 cM. In the case with five QTLs, an additional two QTLs were located on chromosome 3 at 30 and 70 cM. The linked QTLs were either in coupling (effects of equal sign) or repulsion (effects of opposite sign). The QTLs were assumed to act additively. In the models containing three QTLs, the heritability (defined as the ratio of the genetic variance to the total phenotypic variance) was 0.20 and 0.12 when the linked QTLs were in coupling and repulsion, respectively. In the models containing five QTLs, the heritability was 0.30 and 0.17 when the linked QTLs were in coupling and repulsion, respectively. Note

**Table 4.1**: Estimated genome-wide LOD thresholds for a backcross with nine 100 cM chromosomes each containing 11 equally-spaced markers.

| sample | | Zeng | | | |
|---|---|---|---|---|---|
| size | ANOVA | 3 | 5 | 7 | 9 |
| 100 | 2.5 | 3.5 | 4.1 | 4.7 | 5.1 |
| 250 | 2.5 | 3.3 | 3.6 | 3.8 | 4.0 |
| 1000 | 2.5 | 3.2 | 3.3 | 3.4 | 3.4 |

that all QTLs were located exactly at marker loci.

For each QTL model and for each sample size, we performed 1000 simulations. The result of the application of each method was a set of marker loci indicated to be at or near QTLs. In assessing the results, we defined a chosen marker to be correctly identifying a QTL if it was within 20 cM of a QTL; otherwise it was deemed incorrect. If more than one chosen marker were within 20 cM of the same QTL, one was called correct and the others were called incorrect.

### 4.1.2 Results

The estimated genome-wide LOD (base 10) thresholds for ANOVA and Zeng's method (using forward selection up to 3, 5, 7 and 9 markers) are displayed in Table 4.1. The estimated standard errors for the thresholds, obtained using a bootstrap (Venables and Ripley 1994), are approximately 0.1.

For ANOVA, the threshold was constant across sample sizes and corresponded closely to the threshold in Figure 4 of Lander and Botstein (1989). For Zeng's method, the threshold increased with the number of regressors used and decreased with sample size. With 1000 progeny, the threshold for Zeng's method was nearly constant for the different numbers of regressors used.

In Table 4.2, we display the joint distribution, across the 1000 simulations, of the numbers of correctly and incorrectly chosen markers for the case of three QTLs with two QTLs linked in coupling, and using 250 progeny. The four columns labelled "Zeng" correspond to Zeng's method using forward selection up to either 3, 5, 7 or 9 markers. The three columns labelled "BIC" correspond to forward selection using the BIC-2, BIC-2.5 and BIC-3 criteria. The column "permu" gives the results for using forward selection with a permutation test at each stage. The second-to-last row in the table includes all simulations

**Table 4.2**: Distribution of the numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in coupling, and using 250 progeny.

| # | # | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| corr | incorr | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| 3 | 0 | 69 | 31 | 25 | 19 | 13 | 180 | 65 | 19 | 133 |
| 2 | 0 | 526 | 412 | 315 | 240 | 199 | 509 | 496 | 395 | 539 |
| 1 | 0 | 332 | 429 | 443 | 430 | 421 | 199 | 395 | 554 | 246 |
| 0 | 0 | 1 | 97 | 184 | 281 | 334 | 0 | 2 | 9 | 0 |
| 3 | 1 | 4 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 6 |
| 2 | 1 | 26 | 6 | 7 | 5 | 6 | 59 | 13 | 5 | 37 |
| 1 | 1 | 40 | 18 | 12 | 18 | 13 | 35 | 28 | 17 | 34 |
| 0 | 1 | 0 | 6 | 11 | 5 | 12 | 0 | 0 | 0 | 0 |
| other | | 2 | 1 | 3 | 2 | 2 | 11 | 1 | 1 | 5 |
| $\geq$ 1 wrong | | 72 | 31 | 33 | 30 | 33 | 112 | 42 | 23 | 82 |

with two or more incorrectly chosen markers. The last row in the table gives the number of simulations in which at least one incorrect marker was chosen.

ANOVA nearly always found at least one QTL, and often found two, but it had difficulty in separating the two linked QTLs. ANOVA added incorrect markers about 7% of the time. Zeng's method did worse than ANOVA in this situation. It suffered from low power for detection, and the power decreased sharply as the number of markers used as regressors increased; using three markers as regressors worked best in this case. Forward selection using BIC-2 did a better job of detecting the QTLs, but included incorrect markers 11% of the time—much more often than the other methods. The use of a larger multiplier helped to avoid this problem, but at the expense of a lower power for detection. Forward selection using a permutation test did well: it detected more QTLs than ANOVA and Zeng's method, while including incorrect markers only 8% of the time.

Table 4.3 shows which of the QTLs were correctly identified by the different methods. The first three columns, labelled "model," correspond to the three QTLs: first the QTL on chromosome 1, and then the two linked QTLs on chromosome 2. A one in these columns indicates that the QTL was correctly identified; a zero indicates that it was not found. Note that in this table, we ignore the markers which were incorrectly identified. For example, in the column labelled "ANOVA," the model "1 1 1" was identified 73 times out of 1000 simulations; this includes 69 times in which no extraneous markers were included,

**Table 4.3**: Models identified in 1000 simulations of the model containing three QTLs with two QTLs (represented in the second and third columns) linked in coupling, and using 250 progeny.

| | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|
| model | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| 1 1 1 | 73 | 31 | 25 | 19 | 13 | 187 | 65 | 19 | 139 |
| 1 1 0 | 254 | 189 | 140 | 102 | 83 | 258 | 239 | 189 | 260 |
| 1 0 1 | 257 | 191 | 144 | 112 | 92 | 264 | 241 | 192 | 274 |
| 0 1 1 | 41 | 39 | 40 | 33 | 31 | 52 | 30 | 20 | 45 |
| 1 0 0 | 3 | 115 | 173 | 182 | 180 | 1 | 3 | 8 | 1 |
| 0 1 0 | 200 | 161 | 136 | 131 | 120 | 131 | 218 | 293 | 152 |
| 0 0 1 | 171 | 171 | 147 | 135 | 135 | 107 | 202 | 270 | 129 |
| 0 0 0 | 1 | 103 | 195 | 286 | 346 | 0 | 2 | 9 | 0 |

**Table 4.4**: Average numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in coupling.

| | sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| Corr. | 100 | 0.91 | 0.55 | 0.40 | 0.29 | 0.26 | 1.18 | 0.94 | 0.74 | 0.90 |
| | 250 | 1.70 | 1.38 | 1.18 | 1.00 | 0.89 | 1.95 | 1.64 | 1.43 | 1.86 |
| | 1000 | 2.68 | 3.00 | 2.87 | 2.78 | 2.70 | 3.00 | 2.99 | 2.97 | 3.00 |
| Incorr. | 100 | 0.08 | 0.02 | 0.04 | 0.03 | 0.05 | 0.28 | 0.12 | 0.05 | 0.10 |
| | 250 | 0.07 | 0.03 | 0.04 | 0.03 | 0.04 | 0.12 | 0.04 | 0.02 | 0.09 |
| | 1000 | 0.07 | 0.03 | 0.02 | 0.04 | 0.06 | 0.06 | 0.02 | 0.01 | 0.10 |

and 4 times in which one extraneous marker was included (see Table 4.2).

When forward selection and ANOVA identified just one QTL, it was almost always one of the two linked QTLs, but Zeng's method often picked only the QTL on chromosome 1. When two QTLs were identified, all of the methods tended to pick the QTL on chromosome 1 and one of the two linked QTLs. Note that the two linked QTLs on chromosome 2 were chosen at approximately equal frequencies, by all of the methods: the models "1 1 0" and "1 0 1" were chosen nearly the same number of times, as were the models "0 1 0" and "0 0 1."

The average numbers of correctly and incorrectly chosen markers provides nearly as much information as the full distribution shown in Table 4.2. Table 4.4 displays these averages for all three sample sizes, for the simulations of the three-QTL model with two QTLs linked in coupling. The estimated standard errors for these averages vary from around 0.01 to 0.03.

**Table 4.5**: Average numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in repulsion.

| | sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| Corr. | 100 | 0.23 | 0.22 | 0.19 | 0.15 | 0.15 | 0.53 | 0.28 | 0.15 | 0.23 |
| | 250 | 0.87 | 1.18 | 1.15 | 1.02 | 0.90 | 1.37 | 0.90 | 0.58 | 1.03 |
| | 1000 | 2.54 | 2.99 | 2.86 | 2.78 | 2.69 | 3.00 | 2.96 | 2.92 | 2.98 |
| Incorr. | 100 | 0.08 | 0.05 | 0.04 | 0.04 | 0.04 | 0.27 | 0.10 | 0.04 | 0.08 |
| | 250 | 0.07 | 0.04 | 0.04 | 0.04 | 0.03 | 0.12 | 0.04 | 0.02 | 0.08 |
| | 1000 | 0.06 | 0.02 | 0.02 | 0.03 | 0.04 | 0.03 | 0.01 | 0.00 | 0.08 |

The behavior observed in Table 4.2, with 250 progeny, held true for the other two sample sizes as well, except that at $n = 1000$, ANOVA no longer performed well at all, while Zeng's method performed much better, at least when using only three markers as regressors. The frequency with which incorrect markers were added was stable across sample sizes for ANOVA, Zeng's method, and forward selection using a permutation test. Forward selection using the BIC-type criteria included many more incorrect markers when using 100 progeny than when using 1000 progeny. ANOVA included an average of about 0.07 incorrect markers, Zeng's method included an average of about 0.04, and the permutation test method included an average of about 0.1. For BIC-2, the average number of incorrect markers dropped from 0.3 to 0.06 as the sample size went from 100 to 1000.

Table 4.5 gives the average numbers of correctly and incorrectly identified markers for the three-QTL model where the linked QTLs are in repulsion. At the smaller sample sizes, the methods did not perform as well when the linked QTLs were in repulsion; ANOVA and forward selection suffered much more than Zeng's method. The number of incorrectly chosen markers showed little change from the case of coupling, for all of the methods. But the number of correctly identified QTLs, in comparison to coupling, was halved for ANOVA and forward selection, when using 250 progeny. Zeng's method, on the other hand, showed very little change in its ability to identify QTLs, with the result that here his method worked better than ANOVA. With 1000 progeny, the methods all performed much the same as in the coupling case.

Table 4.6 gives the average numbers of correctly and incorrectly identified markers for the five-QTL model with two pairs of QTLs linked in coupling. In terms of identifying QTLs, the behavior of the different methods was similar to that of the three-QTL case with linkage in coupling: Zeng's method didn't find as many QTLs as ANOVA when the sample

**Table 4.6**: Average numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing five QTLs with two pairs of QTLs linked in coupling.

| | sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| Corr. | 100 | 1.43 | 1.03 | 0.77 | 0.55 | 0.46 | 2.05 | 1.62 | 1.26 | 1.54 |
| | 250 | 2.67 | 2.50 | 1.95 | 1.70 | 1.49 | 3.13 | 2.70 | 2.40 | 3.04 |
| | 1000 | 4.22 | 4.98 | 4.99 | 4.74 | 4.60 | 5.00 | 4.98 | 4.94 | 5.00 |
| Incorr. | 100 | 0.08 | 0.03 | 0.03 | 0.03 | 0.03 | 0.29 | 0.12 | 0.05 | 0.09 |
| | 250 | 0.06 | 0.03 | 0.02 | 0.01 | 0.02 | 0.12 | 0.04 | 0.02 | 0.08 |
| | 1000 | 0.08 | 0.30 | 0.04 | 0.02 | 0.04 | 0.12 | 0.06 | 0.03 | 0.20 |

**Table 4.7**: Average numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing five QTLs with two pairs of QTLs linked in repulsion.

| | sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| Corr. | 100 | 0.27 | 0.24 | 0.26 | 0.23 | 0.21 | 0.67 | 0.33 | 0.17 | 0.27 |
| | 250 | 1.00 | 1.42 | 1.61 | 1.53 | 1.40 | 1.89 | 1.11 | 0.63 | 1.25 |
| | 1000 | 3.93 | 4.27 | 4.99 | 4.72 | 4.58 | 4.99 | 4.95 | 4.83 | 4.96 |
| Incorr. | 100 | 0.07 | 0.04 | 0.04 | 0.04 | 0.04 | 0.24 | 0.08 | 0.02 | 0.07 |
| | 250 | 0.08 | 0.05 | 0.04 | 0.03 | 0.03 | 0.14 | 0.05 | 0.02 | 0.08 |
| | 1000 | 0.07 | 0.02 | 0.02 | 0.02 | 0.03 | 0.06 | 0.02 | 0.01 | 0.12 |

size was 100 or 250, while forward selection found more. With 1000 progeny, ANOVA did worse than Zeng's method. Regarding the inclusion of incorrect markers, ANOVA continued to add an average of around 0.08 incorrect markers. Zeng's method continued to control the inclusion of incorrect markers, except in the case of 1000 progeny, when using three markers as regressors; there, the average number of incorrectly identified markers was 0.30. Forward selection using BIC-2 continued to add a high number of incorrect markers, and the marked decrease at 1000 progeny seen previously, no longer seemed to hold. Forward selection using a permutation test gave a great deal more incorrect markers with 1000 progeny than with 100 or 250 progeny.

Table 4.7 gives the average numbers of correctly and incorrectly identified markers for the five-QTL model with two pairs of QTLs linked in repulsion. As with the three-QTL model, Zeng's method performed substantially better than ANOVA with the QTLs in repulsion, and here Zeng's method performed better when using five markers as regressors, rather than three. The decrease in power for Zeng's method accompanied by using more regressors was still observed. The behavior of forward selection was similar to the three-QTL case.

### 4.1.3 Discussion

In this simulation study, we compared the performance of ANOVA, Zeng's method (using varying numbers of regressors) and forward selection (using BIC-type criteria and permutation tests to infer the number of QTLs). Forward selection and Zeng's method showed good benefits over ANOVA, at least when the sample size was large.

With smaller samples, the performance of the methods depended on whether linked QTLs were in coupling or repulsion. In the case of coupling, Zeng's method performed quite poorly, even in comparison to ANOVA. But in the case of repulsion, Zeng's method performed as well as or better than the others. The reason that Zeng's method is more successful in teasing out a pair of QTLs linked in repulsion, may be that such QTLs look more important when both are included in the model. Zeng's method forces the fit of the larger model, whereas forward selection considers the markers one at a time. This difference is best illustrated in Table 4.3. When identifying just one QTL, ANOVA and forward selection generally pick one of the two linked QTLs, whereas Zeng's method picks from the three QTLs at nearly equally proportions.

But the great difficulty with Zeng's method is in choosing how many markers to use as regressors. When too many markers are used, the method suffers from a great loss of power to detect QTLs. The principal reason for this drop in power is the great increase in threshold which comes with the larger number of regressors: using more regressors adds a great deal of noise. When too few markers are used, the method seems to give lower power when linked QTLs are in repulsion, and a high rate of incorrect markers when linked QTLs are in coupling. The correct number of regressors to use seems to be related not to the number of QTLs segregating in the cross, but to the number of QTLs the given experiment is able to identify. That number is, of course, not known.

The method of forward selection improves on ANOVA, but doesn't perform quite as well as Zeng's method when linked QTLs are in repulsion and the sample size is moderate. Using a permutation test to infer the size of the model (i.e., the number of QTLs to add) gave a relatively consistent rate of inclusion of incorrect markers, but one that was much higher than 5%. When using the BIC-type criteria to infer the size of the model, the rate at which incorrect markers were added decreased greatly with sample size, suggesting that the $\log(n)/n$ penalty was not quite right for these sample sizes.

It is important to point out that forward selection with BIC-2 consistently identi-

fied the most QTLs—in all scenarios at all sample sizes. At the same time, however, it also included the most extraneous QTLs. One may feel that a 10–15% rate of extraneous QTLs is perfectly acceptable for many purposes, and so this method may seem best in such situations. But note that, if the threshold used in the other methods (ANOVA, Zeng's method, permutation test) were lowered, one may be able to match the level of performance achieved by BIC-2. So, when comparing forward selection with BIC-type penalties with, e.g., Zeng's approach, it is best to use a criterion and threshold which lead to similar rates of extraneous QTLs. Ideally, we would choose the multiplier $\delta$ in the BIC-$\delta$ and the threshold for Zeng's method (with a given number of markers used as regressors) to give a prescribed rate of inclusion of extraneous QTLs, and would then compare the power of the methods, using these values. Unfortunately, however, the appropriate $\delta$ and threshold depend on both the sample size and the underlying QTL model. As a result, it is not feasible to carry out this approach in practice. One must be satisfied with the approach used above, and use one's judgement about which of the methods performs best.

The results of these simulations recommend the use of forward selection for identifying QTLs. Using permutation tests to determine how many markers to include worked well, but one must keep in mind that the use of 5% level tests does not imply that incorrect markers are included only 5% of the time.

The use of the BIC-type criteria to infer model size has the great benefit of ease of computation, but the lack of control over the rate of including incorrect markers may be a concern. Still, BIC-3 is consistently conservative, at least with the genome size considered here, and so a reasonable approach would be to place more confidence in the markers which enter the model when using BIC-3, and to consider markers which enter with BIC-2 but not BIC-3 as possibly but not definitely in the model.

## 4.2 The study of Doerge and Churchill (1994)

Doerge and Churchill (1994) described a simulation of a backcross with four chromosomes and one or two QTLs. We have attempted to reproduce their results, and have used their models to further compare the major methods for identifying QTLs.

### 4.2.1 Methods

Doerge and Churchill (1994) simulated a backcross with four chromosomes, each 100 cM in length. The first and third chromosomes contained 50 randomly placed markers, and the second and fourth chromosomes contained 10 randomly placed markers. Since we didn't know the exact placement of their markers, we created our own map by throwing down markers at random onto the chromosomes according to their specifications. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$. They simulated 100 and 200 progeny.

They considered cases with 0, 1 and 2 QTLs. In the case with one QTL, a QTL with additive effect 1.0 was placed on chromosome 2 at 61.6 cM, giving a heritability of 0.2. In the case with two QTLs, a second QTL, with additive effect 0.75, was placed on chromosome 1 at 44.4 cM. The heritability in this case was 0.28.

Doerge and Churchill (1994) applied forward selection using a permutation test at each stage, but allowed the inclusion of no more than one marker per chromosome. The 95th percentile of 1000 permutations was used to determine whether or not to add a marker.

We also applied the four methods used in the previous section: ANOVA, Zeng's method (using forward selection up to 3, 5, 7 or 9 markers to choose regressors), forward selection using a BIC-type penalty (with a multiplier of 2, 2.5 or 3), and forward selection using a permutation test, but allowing the inclusion of more than one marker per chromosome. As in Section 4.1, we'll say that a chosen marker is correct if it is within 20 cM of a QTL. We performed 1000 simulations, but in comparing our results to Doerge and Churchill (1994), we considered only the first 500, since they performed just 500 simulations.

### 4.2.2 Results

Table 4.8 contains our reproduction of the results of Doerge and Churchill (1994), as well as those found in Table 3 of their paper. The first four columns of the table, labelled "model," correspond to the four chromosomes: a one in this column indicates that at least one of the markers on that chromosome was chosen, and a zero indicates that none of the markers on that chromosome were chosen. The numbers in the table give the numbers of simulations for which the different models were identified. Periods in the table represent zeros. The columns labelled "new" give the results for our simulation; the columns labelled "D&C" give the results found in Table 3 of Doerge and Churchill (1994).

**Table 4.8**: Reproduction of a simulation in Doerge and Churchill (1994).

| | sample size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 100 | | | | | | 200 | | | | | |
| | number of QTLs | | | | | | number of QTLs | | | | | |
| | 0 | | 1 | | 2 | | 0 | | 1 | | 2 | |
| model | new | D&C | new | D&C | new | D&C | new | D&C | new | D&C | new | D&C |
| 0 0 0 0 | 476 | 500 | 50 | 33 | 21 | 6 | 478 | 480 | . | 1 | . | . |
| 1 0 0 0 | 8 | . | 3 | . | 28 | 64 | 9 | . | . | . | 1 | . |
| 0 1 0 0 | 2 | . | 416 | 452 | 106 | 161 | 2 | 1 | 483 | 498 | 9 | 15 |
| 0 0 1 0 | 4 | . | 2 | . | . | . | 8 | 1 | . | . | . | . |
| 0 0 0 1 | 9 | . | . | . | . | . | 2 | 18 | . | . | . | . |
| 1 1 0 0 | 1 | . | 9 | . | 324 | 266 | . | . | 4 | 1 | 469 | 447 |
| 1 0 1 0 | . | . | . | . | 2 | . | . | . | . | . | . | . |
| 1 0 0 1 | . | . | . | . | 1 | . | . | . | . | . | . | . |
| 0 1 1 0 | . | . | 15 | 15 | 3 | . | 1 | . | 11 | . | . | . |
| 0 1 0 1 | . | . | 4 | . | . | . | . | . | 2 | . | . | . |
| 0 0 1 1 | . | . | . | . | . | . | . | . | . | . | . | . |
| 0 1 1 1 | . | . | . | . | . | . | . | . | . | . | . | . |
| 1 0 1 1 | . | . | . | . | . | . | . | . | . | . | . | . |
| 1 1 0 1 | . | . | . | . | 2 | 2 | . | . | . | . | 8 | 37 |
| 1 1 1 0 | . | . | 1 | . | 13 | 1 | . | . | . | . | 12 | 1 |
| 1 1 1 1 | . | . | . | . | . | . | . | . | . | . | 1 | . |

**Table 4.9**: Estimated genome-wide LOD thresholds for the simulated backcross described in Doerge and Churchill (1994).

| sample | | Zeng | | | |
|---|---|---|---|---|---|
| size | ANOVA | 3 | 5 | 7 | 9 |
| 100 | 2.2 | 3.5 | 4.0 | 4.3 | 4.6 |
| 200 | 2.2 | 3.5 | 3.9 | 3.9 | 4.1 |

**Table 4.10**: Average number incorrectly chosen markers in 1000 simulations of a model with no QTLs

| sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|
| size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| 100 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.10 | 0.01 | 0.00 | 0.05 |
| 200 | 0.04 | 0.04 | 0.03 | 0.04 | 0.05 | 0.05 | 0.00 | 0.00 | 0.05 |

Our simulations gave results quite different than the published ones. In the case of 100 progeny and no QTLs, our simulations gave 24 cases in which at least one marker was identified (incorrectly) as a QTL. In the published results, all 500 simulations identified no QTLs. For the models with one QTL, at both sample sizes, our simulations included fewer cases in which the correct model was identified. For the models with two QTLs, at both sample sizes, our simulations included more cases in which the correct model was identified.

The estimated genome-wide LOD (base 10) thresholds for ANOVA and Zeng's method (using forward selection up to 3, 5, 7 and 9 markers) are displayed in Table 4.9. The estimated standard errors for the thresholds, obtained using a bootstrap (Venables and Ripley 1994), are approximately 0.1. The thresholds exhibit similar behavior as those found in Section 4.1.

Table 4.10 gives the average number of incorrectly chosen markers for the models with no QTLs. The estimated standard errors are around 0.1. ANOVA, Zeng's method, and forward selection using permutation tests all displayed the appropriate 5% rate of incorrectly included markers. BIC-2 gave a rate of 10% and 5% when the sample size was 100 and 200, respectively. BIC-2.5 and BIC-3 identified QTLs very seldom in this case.

Table 4.11 gives the average numbers of correctly and incorrectly identified markers for the model with one QTL. ANOVA performed best here, though the performance of forward selection was not very different. Zeng's method performed quite badly, and, as was seen previously, its power decreased sharply with the number of regressors used. The

**Table 4.11**: Average numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing one QTL

| | sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| Corr. | 100 | 0.87 | 0.53 | 0.39 | 0.32 | 0.25 | 0.90 | 0.75 | 0.51 | 0.87 |
| | 200 | 0.94 | 0.58 | 0.41 | 0.36 | 0.30 | 0.95 | 0.87 | 0.72 | 0.94 |
| Incorr. | 100 | 0.09 | 0.03 | 0.03 | 0.04 | 0.04 | 0.15 | 0.05 | 0.02 | 0.10 |
| | 200 | 0.08 | 0.02 | 0.02 | 0.03 | 0.04 | 0.10 | 0.03 | 0.01 | 0.08 |

**Table 4.12**: Average numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing two QTLs

| | sample | | Zeng | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | size | ANOVA | 3 | 5 | 7 | 9 | 2 | 2·5 | 3 | permu |
| Corr. | 100 | 1.42 | 0.99 | 0.73 | 0.56 | 0.43 | 1.60 | 1.18 | 0.70 | 1.48 |
| | 200 | 1.67 | 1.26 | 0.90 | 0.70 | 0.56 | 1.77 | 1.50 | 1.14 | 1.72 |
| Incorr. | 100 | 0.14 | 0.05 | 0.03 | 0.03 | 0.04 | 0.22 | 0.08 | 0.03 | 0.15 |
| | 200 | 0.12 | 0.04 | 0.02 | 0.03 | 0.03 | 0.17 | 0.06 | 0.02 | 0.13 |

numbers of incorrect markers added for the different methods were similar to what was seen in Section 4.1.

Table 4.12 gives the average numbers of correctly and incorrectly identified markers for the model with two QTLs. Zeng's method again performs quite badly, and forward selection shows little improvement over ANOVA.

### 4.2.3 Discussion

The large differences between the results in Doerge and Churchill (1994) and our reproduction of those results are disconcerting; one might be led to the conclusion that we've done something wrong. But the results of Doerge and Churchill (1994) displayed a number of anomalies. With data in which no QTLs are segregating, we expect the permutation test method to still identify QTLs about 5% of the time. But in the case of no QTLs and 100 progeny, all 500 simulations in Doerge and Churchill (1994) identified the null model. Moreover, when incorrect markers are identified, we expect them to be distributed approximately equally across the chromosomes, though maybe a few more will be found on chromosomes with more densely spaced markers. But consider the case of two QTLs and 200 progeny; the results in Doerge and Churchill (1994) deviate quite markedly from what we would expect: of the 38 times in which a marker on chromosome 3 or 4 was

(incorrectly) identified, 37 times that marker was on chromosome 4, whereas only once was it on chromosome 3. We would have expected the incorrect marker to be on chromosomes 3 and 4 with approximately equal frequency, or maybe more often on chromosome 3, since it had 50 markers, as compared to only 10 on chromosome 4. Thus, we place more confidence in our own results than in those of Doerge and Churchill (1994).

The most striking result in Tables 4.11 and 4.12 is the superiority of ANOVA in these situations. But this should be no surprise, since the basic assumption underlying ANOVA is that there is just one QTL. The advantage of the other methods comes only when there are several QTLs segregating. Still, the models simulated here are really much simpler than is generally seen in experimental crosses. For example, a study of grain yield in maize (Stuber et al. 1992) revealed as many as eight QTLs segregating in a single experiment.

There are two important points to be emphasized in light of this study. First, when presenting a new method, one should not study it in isolation, but rather should compare its performance to that of other methods. Second, the models that one uses to assess the performance of a method should be reasonable representations of the sorts of situations in which it will be used in practice.

## 4.3 Power to detect QTLs

Before beginning a QTL experiment, it is important to think carefully about the number of progeny required to have a reasonable chance of success. Many authors have studied the power of ANOVA and interval mapping to identify QTLs, with different sizes of experiments and with QTLs of varying effects (Soller et al. 1976; Lander and Botstein 1989; Knott and Haley 1992; van Ooijen 1992; Carbonell et al. 1993; Darvasi et al. 1993; Jansen 1994; Rebaï et al. 1995). Unfortunately, these studies have considered very simple cases, generally simulating just one or two chromosomes and only a single QTL.

In this section, we consider the problem further. Our goals are, first, to study the power for detecting QTLs in somewhat more complex situations (where there are several segregating QTLs, and a number of chromosomes), and, second, to obtain results which may guide researchers who are planning to carry out QTL experiments. Typically, a researcher knows, or has some idea of, the heritability of the trait of interest for their particular cross (that is, the proportion of the total phenotypic variance due to the QTLs segregating in the cross). The question then is, how many progeny should be generated in order to have

a particular chance of identifying QTLs of large effect? One might also ask, what is the expected proportion of QTLs that will be identified, and what is the chance of identifying at least one QTL? Also, will we be able to distinguish linked QTLs? Finally, a researcher will ask, how dense of a genetic map should be used?

We attempt to answer these questions in the context of an $F_2$ intercross, since it is most commonly used in practice. We will apply solely the method of forward selection using the BIC-2 criterion.

### 4.3.1 Proportion of QTLs identified

First, we look at the proportion of segregating QTLs that are identified in a cross. We consider an $F_2$ intercross, with 100, 200, 400, 600, 800 or 1000 progeny, and with 9 chromosomes of length 100 cM each, and having 11 equally spaced markers on each chromosome (at a 10 cM spacing). We model either 2, 4, 6 or 8 QTLs, acting additively, having equal effects, and all located on separate chromosomes. The locations of the QTLs were chosen randomly, and were fixed for all simulations. The eight QTLs were on chromosomes 1–8 at positions 7.0, 57.8, 60.2, 89.0, 52.3, 37.5, 81.4, and 21.0 cM, respectively. (When fewer than eight QTLs were used, the last few QTLs were dropped. For instance, in the model with two QTLs, we used QTLs on chromosomes 1 and 2 at positions 7.0 and 57.8 cM, respectively.)

The environmental variation was normally distributed, with standard deviation $\sigma = 1$. The QTLs had dominance deviation, $d = 0$, and additive effect, $a$, chosen to give heritability $h^2 = 0.2$, 0.4, 0.6 or 0.8. In other words, the three QTL genotypes, LL, HL and HH, had effects $-a$, 0 and $+a$, where $a$ was chosen to give the prescribed heritability. Let $S$ denote the number of QTLs. Then

$$h^2 = \frac{a^2 S/2}{a^2 S/2 + 1},$$

and so

$$a = \sqrt{\frac{2}{S} \cdot \frac{h^2}{1 - h^2}}$$

The values of $a$ that we used are shown in Table 4.13.

Figure 4.1 (on page 63) displays the average, across 200 simulations, of the proportion of QTLs which were correctly identified, where we say that a chosen marker is correctly identifying a QTL if it is within 20 cM of a QTL. The proportion of QTLs detected increases

**Table 4.13**: Additive effect ($a$) of each QTL in an intercross with environmental variance $\sigma^2 = 1$, for a given number ($S$) of unlinked QTLs of equal effect, to give a prescribed heritability ($h^2$).

| | $h^2$ | | | |
|---|---|---|---|---|
| $S$ | 0.2 | 0.4 | 0.6 | 0.8 |
| 2 | 0.50 | 0.82 | 1.22 | 2.00 |
| 4 | 0.35 | 0.58 | 0.87 | 1.41 |
| 6 | 0.29 | 0.47 | 0.71 | 1.15 |
| 8 | 0.25 | 0.41 | 0.61 | 1.00 |

with sample size and with heritability, and decreases with the number of segregating QTLs responsible for the given heritability. With a sample size of only 100 or 200, only a very small proportion of the QTLs will be detected, unless there are a small number of QTLs of large effect.

This figure gives a rather pessimistic view of the problem of identifying QTLs. However, one does not usually expect that a single experiment will result in the detection of a majority of the QTLs which are segregating in a cross; the goal of most experiments is rather to identify a least a few QTLs of large effect. Still, it is important to keep in mind that, as this figure shows, the QTLs detected in an analysis will generally be only a subset of the segregating QTLs.

### 4.3.2 Chance of finding at least one QTL

In this section, we estimate the chance of finding at least one of the segregating QTLs in a cross. Using exactly the same simulations as the previous subsection, Figure 4.2 (on page 64) displays the proportion of the 200 simulations in which at least one QTL was correctly identified, meaning that there was a chosen marker within 20 cM of a QTL.

This figure *is* rather depressing. We see that, when the sample size is 100 or 200, the chance of detecting even one QTL is quite small, unless the heritability is high, and the number of QTLs contributing to that heritability is rather small. However, when the sample size is large, one will be assured of detecting at least one QTL.

### 4.3.3 Chance of finding a particular QTL

In this section, we consider the chance of identifying a particular QTL of large effect. We again use an $F_2$ intercross, with 100, 200, 400, 600, 800 or 1000 progeny, and

**Table 4.14**: Additive effect ($a$) of a QTL responsible for a proportion $p$ of the genetic variance, in an intercross with environmental variance $\sigma^2 = 1$, to give a prescribed heritability ($h^2$).

| | $h^2$ | | |
|---|---|---|---|
| $p$ | 0.2 | 0.4 | 0.6 |
| 0.15 | 0.27 | 0.45 | 0.67 |
| 0.30 | 0.39 | 0.63 | 0.95 |
| 0.45 | 0.47 | 0.77 | 1.16 |

with 9 chromosomes of length 100 cM each, and having 11 equally spaced markers on each chromosome (at a 10 cM spacing), and we model either 2, 4, 6 or 8 QTLs, acting additively, all located on separate chromosomes. (The QTL locations are the same as for the simulations in the previous two subsections.) Here, we let the first QTL (on chromosome 1) be responsible for either 15, 30 or 45% of the genetic variance. The other QTLs have equal effects. Note that the first QTL does not always have the largest effect.

The environmental variation was again normally distributed, with standard deviation $\sigma = 1$, and the QTLs again had dominance deviation $d = 0$, and additive effects chosen to give heritability $h^2 = 0.2$, 0.4 or 0.6. Let $a$ be the additive effect of the first QTL, and let $a'$ be the additive effect of the remaining $S - 1$ QTLs. Letting $p$ denote the proportion of the genotypic variance ascribed to the first QTL, we have

$$p = \frac{\frac{1}{2}a^2}{\frac{1}{2}a^2 + \frac{S-1}{2}(a')^2}$$

so that

$$a = a'\sqrt{\frac{p}{1-p}(S-1)}.$$

Also

$$h^2 = \frac{\frac{1}{2}a^2 + \frac{S-1}{2}(a')^2}{\frac{1}{2}a^2 + \frac{S-1}{2}(a')^2 + 1},$$

and so

$$a' = \sqrt{\frac{2}{S-1}(1-p)\frac{h^2}{1-h^2}}$$

and

$$a = \sqrt{2p\frac{h^2}{1-h^2}}.$$

Note that this effect, $a$, is independent of the number of other QTLs segregating in the cross. Table 4.14 gives the values of $a$ that we used.

Figure 4.3 (on page 65) displays the fraction, of 200 simulations, in which this first QTL was chosen. The four lines in each plot correspond to having a total of 2, 4, 6 or 8 segregating QTLs. We are better able to detect the first QTL when there are fewer QTLs segregating in the cross. After detecting some of the other QTLs, one has a greater chance of detecting this particular one, since the residual variance has been reduced. This effect is strongest when the heritability is large, and the proportion of the genetic variance due to the first QTL is small.

Note, again, that if the sample size is only 100 or 200, one has a very small chance of detecting this QTL, unless its effect becomes rather large. Even when the QTL is responsible for 30% of the genetic variance, and the heritability is 0.4 (so that the QTL has effect $a/\sigma = 0.63$ ), with only 100 progeny, the chance of detecting the QTL is less than 30%.

### 4.3.4 Separating linked QTLs

In this section, we study our ability to separate linked QTLs. We consider a setup similar to the previous sections, though here there are 5 chromosomes with 21 markers per chromosome (at a 5 cM spacing). We simulate two QTLs on chromosome 1, separated by 15, 25, 35 or 45 cM, and centered around the 50 cM position. The QTLs are either in coupling or repulsion, and have dominance deviation $d = 0$ and an additive effect $a$, chosen to give a heritability $h^2 = 0.2$ or 0.4. The environmental variation was, again, normally distributed with standard deviation $\sigma = 1$.

In this case,

$$h^2 = \begin{cases} \frac{2a^2(1-r)}{2a^2(1-r)+1} & \text{in coupling} \\ \frac{2a^2 r}{2a^2 r+1} & \text{in repulsion} \end{cases}$$

where $r = \frac{1}{2}(1 - e^{-2d/100})$ is the recombination fraction between the two QTLs, separated by $d$ cM. And so,

$$a = \begin{cases} \sqrt{\frac{h^2}{1-h^2} \cdot \frac{1}{2(1-r)}} & \text{in coupling} \\ \sqrt{\frac{h^2}{1-h^2} \cdot \frac{1}{2r}} & \text{in repulsion} \end{cases}$$

Table 4.15 displays the values of $a$ used.

Figure 4.4 (on page 66) displays the fraction of 200 simulations in which both of the two QTLs were detected. Here, we required that a chosen marker be within 5 cM of

**Table 4.15**: Absolute additive effect ($a$) for each of two QTLs separated by $d$ cM, linked in either coupling or repulsion, giving a prescribed heritability ($h^2$), when the environmental variance is 1.

| | coupling $h^2$ | | repulsion $h^2$ | |
|---|---|---|---|---|
| $d$ (cM) | 0.2 | 0.4 | 0.2 | 0.4 |
| 15 | 0.38 | 0.62 | 0.98 | 1.60 |
| 25 | 0.39 | 0.64 | 0.80 | 1.30 |
| 35 | 0.41 | 0.67 | 0.70 | 1.15 |
| 45 | 0.42 | 0.69 | 0.65 | 1.06 |

a QTL before we would say it was correctly identifying that QTL. (We used this more rigorous criterion, since in one case, our two QTLs were separated by only 15 cM.)

It is interesting to see that, when heritability is held constant, the separation between two QTLs (of equal effect) linked in repulsion has only a small effect on the ability to identify both of those QTLs. However, note that, as seen in Table 4.15, when the QTLs are linked in repulsion, one must radically change the size of the QTLs' effects if the heritability is to be kept constant. One can infer from these results that, if the QTLs' *effects* were kept constant, it would be much more difficult to separate QTLS that were close together.

In the case of coupling, the effect of each QTL does not change much with separation, when holding heritability constant (see Table 4.15). As a result, with the heritability held constant (and hence the QTL effect held approximately constant), the separation between the two QTLs has a large effect on the ability to distinguish them. With a sample size of 100, we are not able to distinguish the QTLs, even when they are 45 cM apart.

### 4.3.5 Effect of marker density

In order to study the effect of marker density on the ability to detect a QTL, we simulated an intercross with five chromosomes each of length 100 cM, and with 6, 11 or 21 equally spaced markers, at spacings 20, 10 or 5 cM, respectively. We simulated 4 QTLs, at the same positions used in the simulations described in the previous subsections. The QTL on chromosome one was responsible for 25% or 50% of the total genotypic variance; the heritability was $h^2 = 0.25$ or 0.50. Figure 4.5 (on page 67) displays the proportion of 200 simulations in which the QTL on chromosome 1 was detected. Note that map density shows only a small effect on our ability to detect the presence of a QTL.

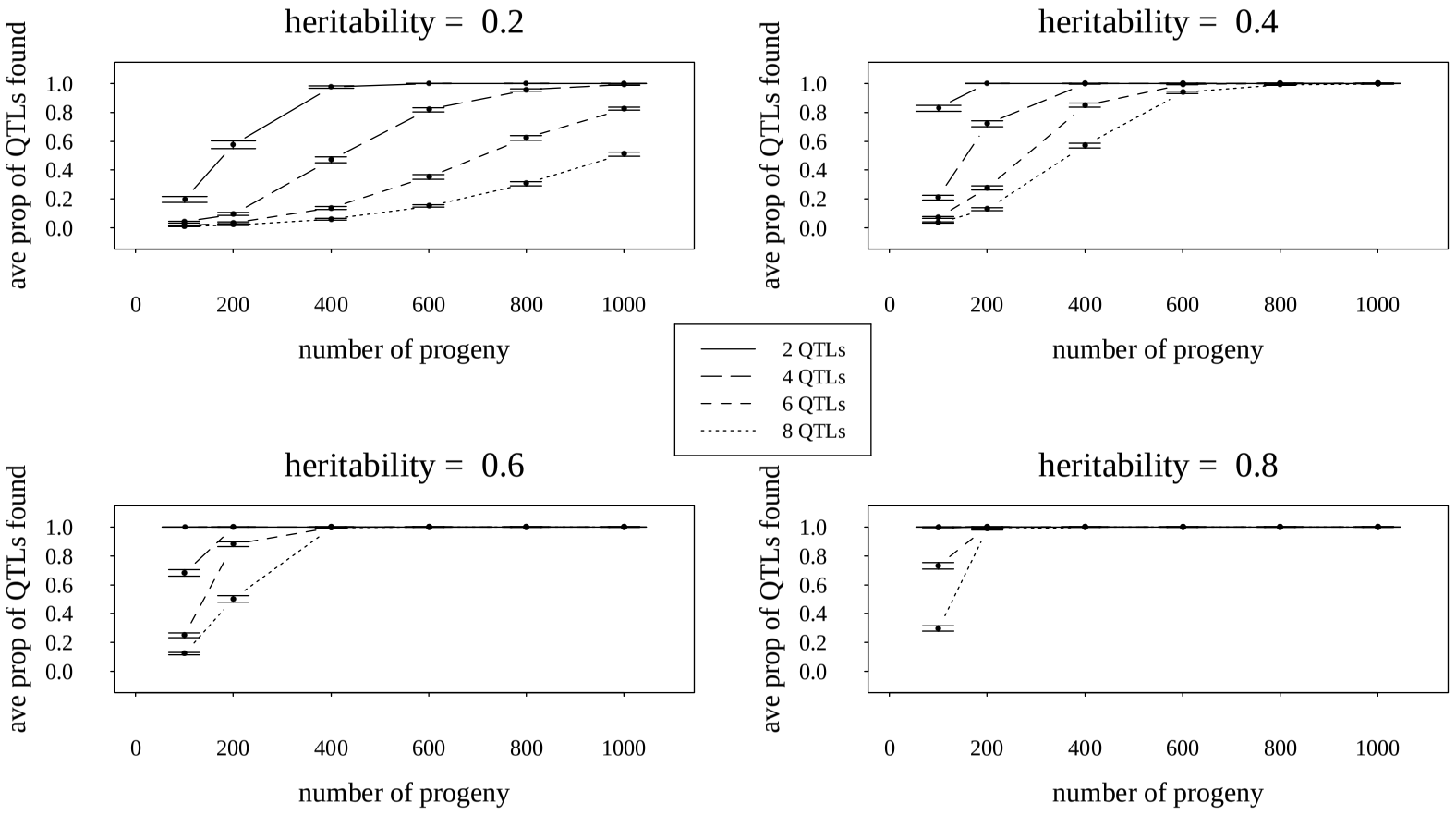


**Figure 4.1**: Average proportion of QTLs identified in 200 simulations, using nine chromosomes of length 100 cM, with 11 equally spaced markers (10 cM apart), and QTLs of equal, additive effects, on separate chromosomes. Forward selection with BIC-2 was used to detect QTLs. Error bars correspond to $\pm 1$ SE.

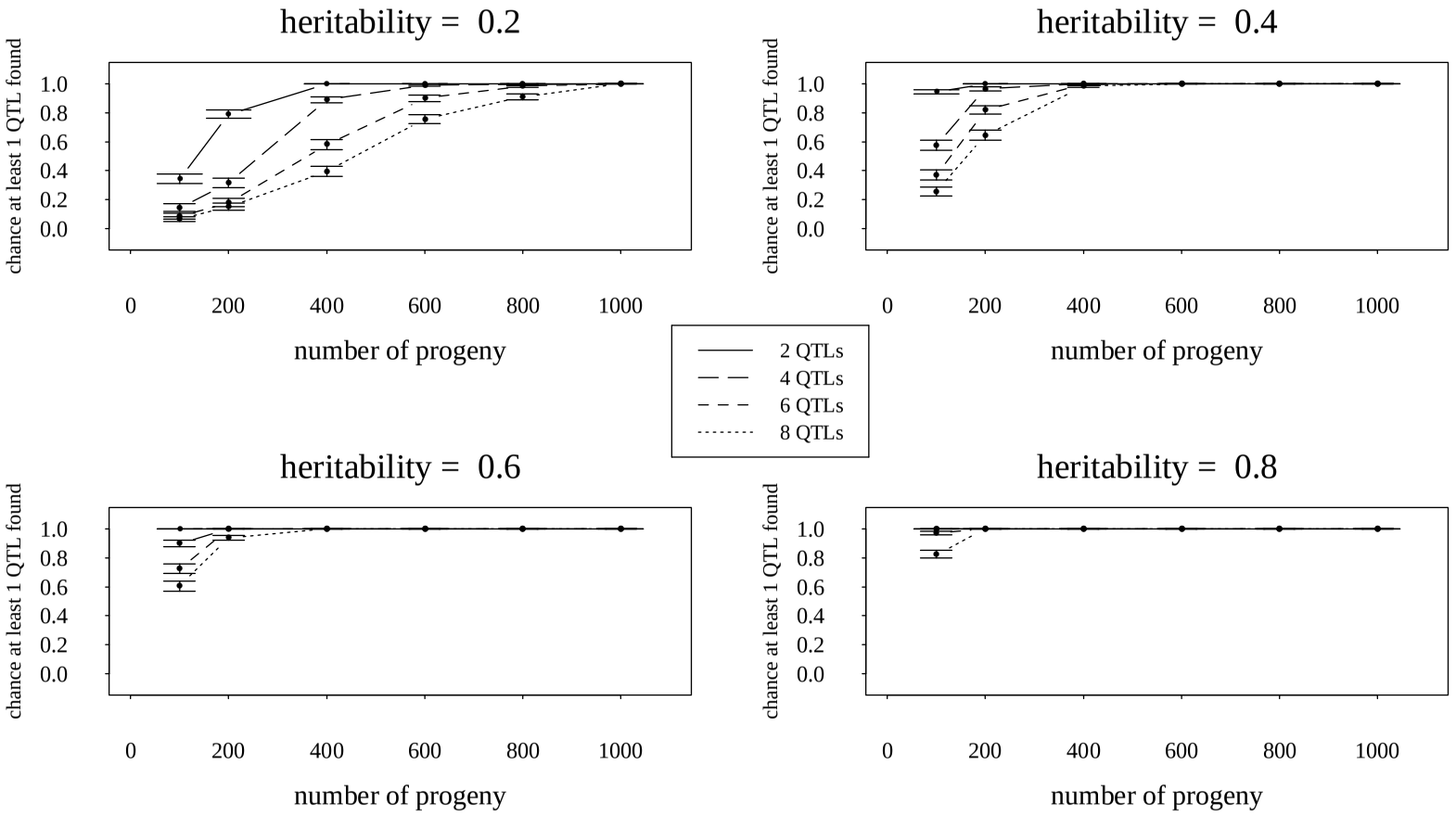


**Figure 4.2**: Chance that at least one QTL is identified, among 200 simulations using nine chromosomes of length 100 cM, with 11 equally spaced markers (10 cM apart), and QTLs of equal, additive effects, on separate chromosomes. Forward selection with BIC-2 was used to detect QTLs. Error bars correspond to $\pm 1$ SE.

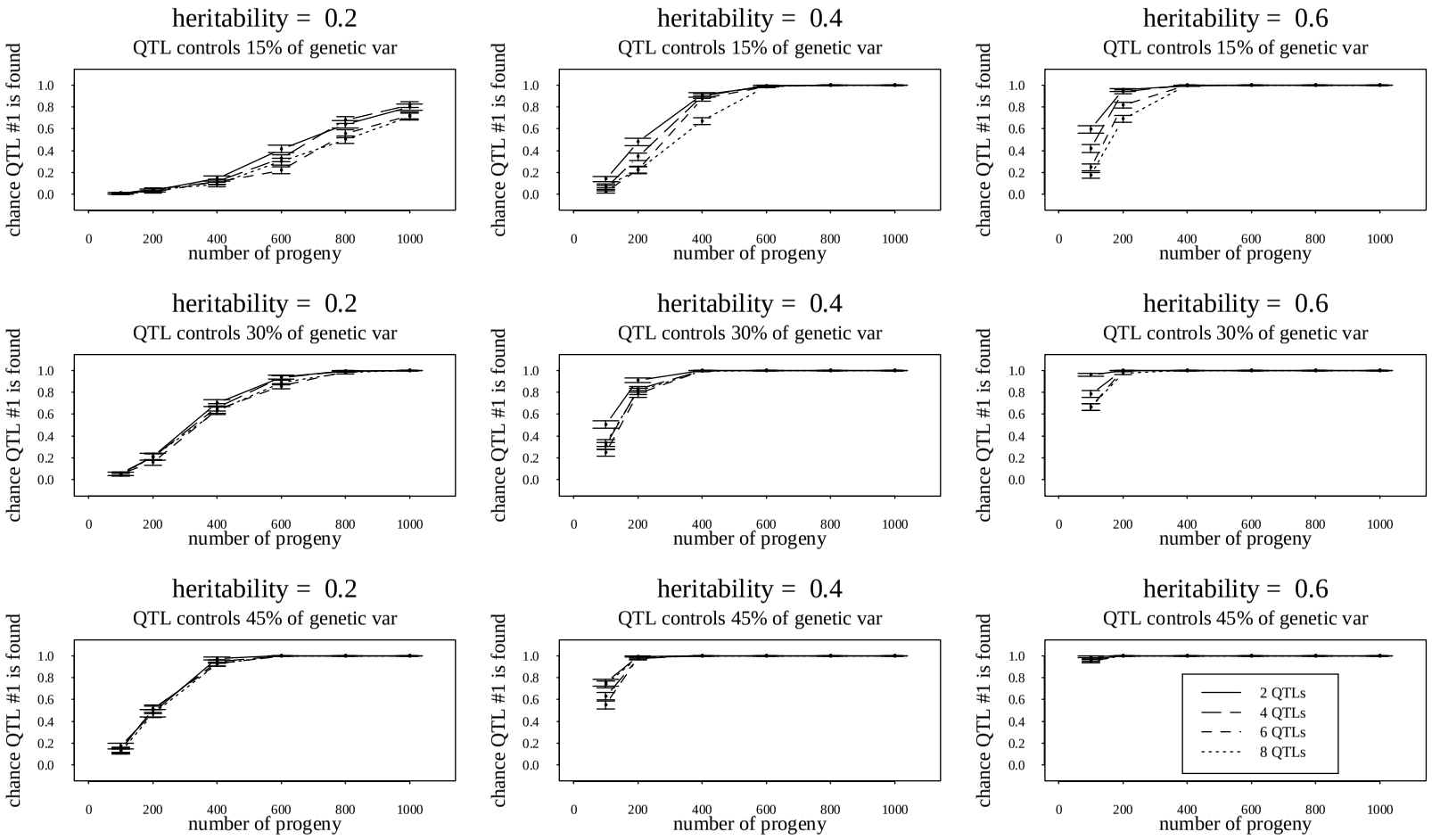


**Figure 4.3**: Chance that a particular QTL is identified, among 200 simulations using nine chromosomes of length 100 cM, with 11 equally spaced markers (10 cM apart). The QTLs had additive effects and were on separate chromosomes. Forward selection with BIC-2 was used to detect QTLs. Error bars correspond to $\pm 1$ SE.

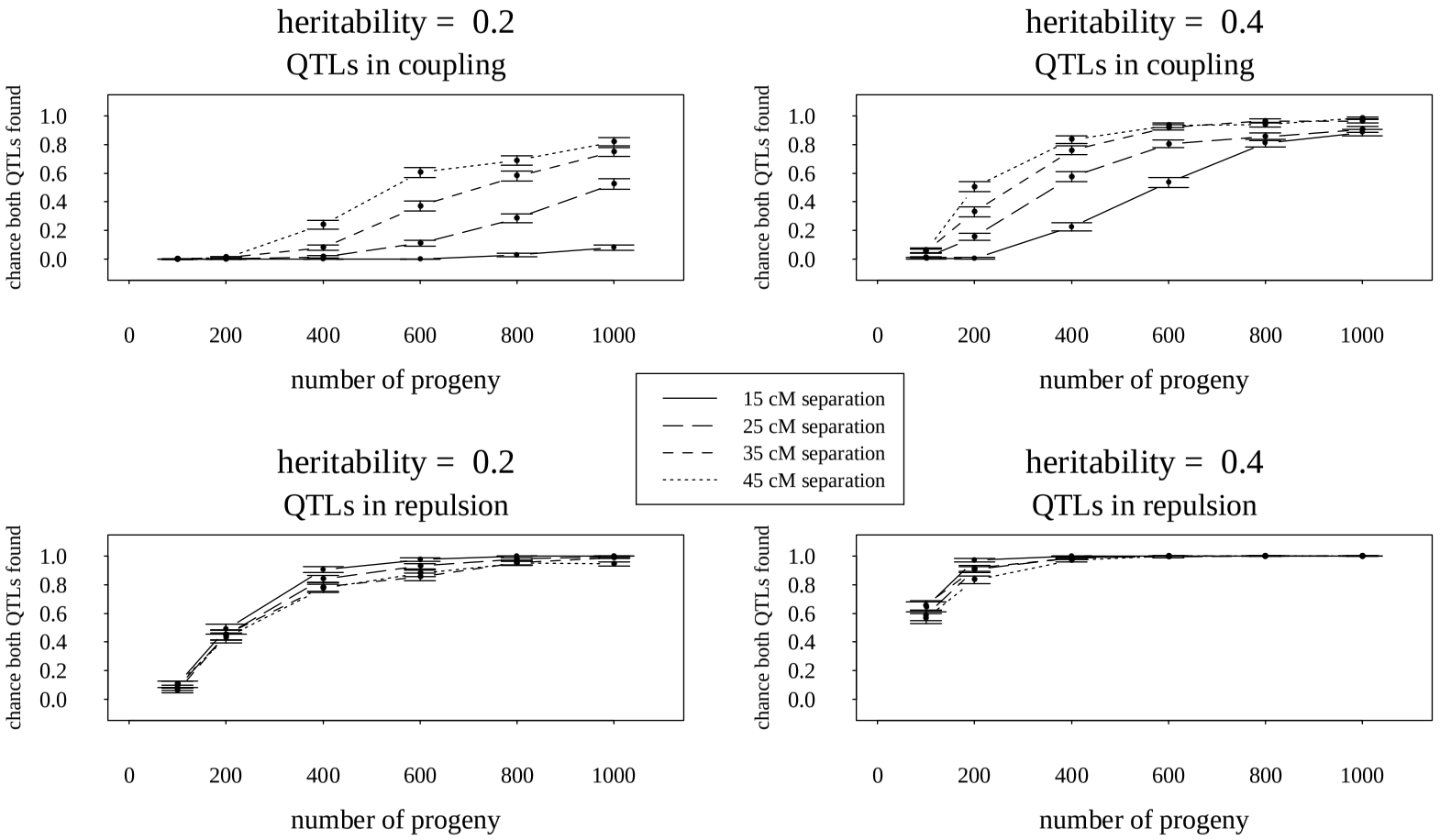


**Figure 4.4**: Chance of identifying both of two linked QTLs, among 200 simulations using 5 chromosomes of length 100 cM, with 21 equally spaced markers (5 cM apart). Forward selection with BIC-2 was used to detect QTLs. Error bars correspond to $\pm 1$ SE.

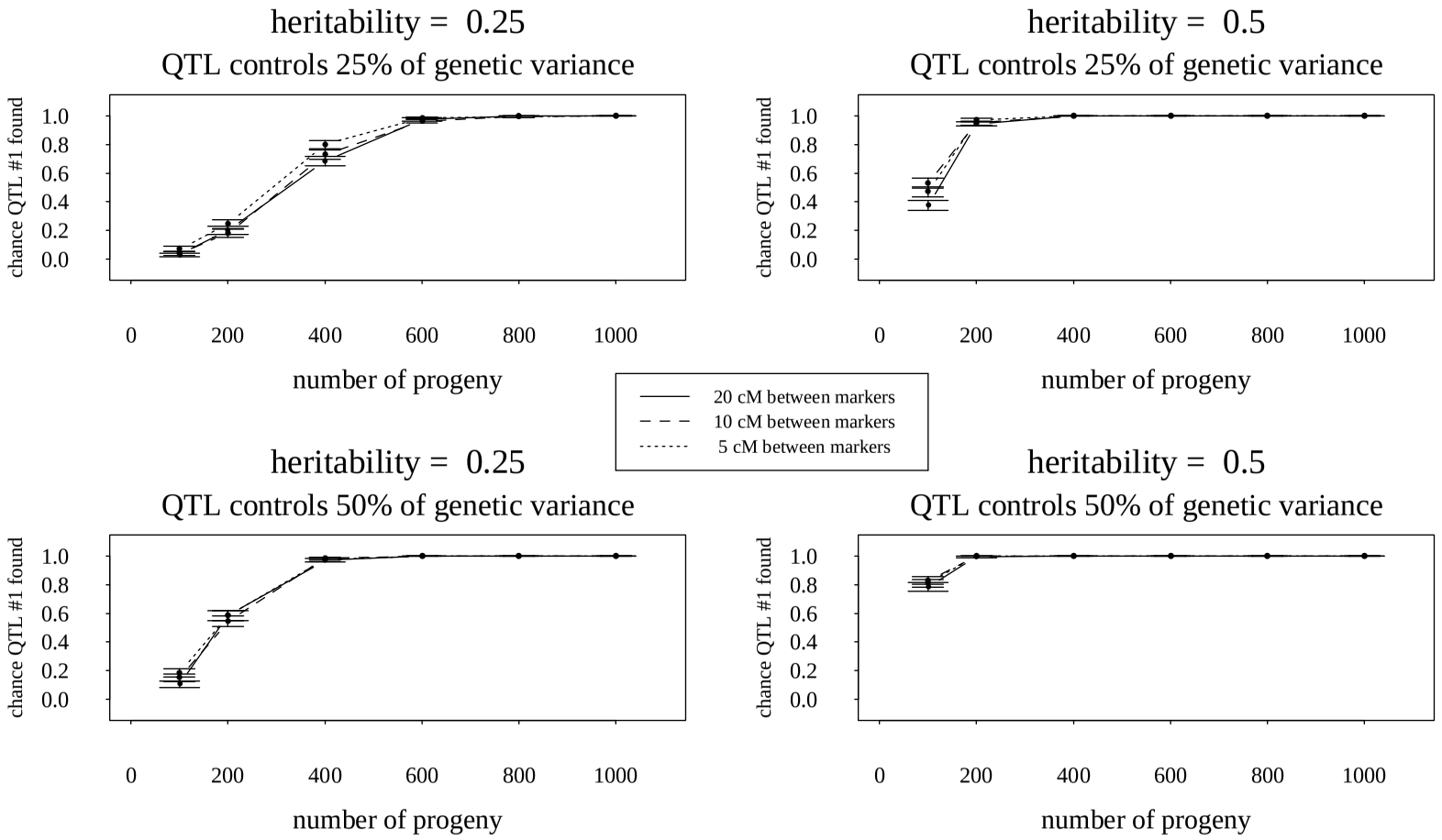


**Figure 4.5**: Chance of identifying a particular QTL for different marker spacings, among 200 simulations using 5 chromosomes of length 100 cM. The QTLs had equal, additive effects and were on separate chromosomes. Forward selection with BIC-2 was used to detect QTLs. Error bars correspond to $\pm 1$ SE.

**Table 4.16**: Power to detect a QTL responsible for 5 or 10% of the phenotypic variance.

| | van Ooijen (1992) | | Lander and Botstein (1989) | | our simulations | |
|---|---|---|---|---|---|---|
| $n$ | 5% | 10% | 5% | 10% | 5% | 10% |
| 100 | 6 | 31 | 23 | 66 | 5 | 13 |
| 200 | 29 | 79 | 67 | 90 | 21 | 47 |
| 400 | 76 | 100 | 90 | 97 | 63 | 93 |

### 4.3.6 Discussion

In this section, we have presented the results of a number of simulations aimed at assessing the power for detecting QTLs in $F_2$ intercross experiments. The models that we used are more complex than those of past work on the problem, but are still much simpler than what one would expect to see in a real experiment.

One can conclude from the results here that an experiment with only 100 or 200 progeny will only be able to detect QTLs of very large effect. For example, with eight QTLs of equal, additive effects, and giving a heritability of 0.4, there is a greater than 30% chance that none of the QTLs will be detected in an experiment of 200 progeny; an experiment of 100 progeny has a greater than 70% chance of finding no QTLs in this situation.

These results are somewhat more pessimistic than previous studies. Van Ooijen (1992) simulated an $F_2$ intercross with a single chromosome of length 120 cM, with equally spaced markers, 5 cM apart, and a single QTL responsible for 5 or 10% of the total phenotypic variance. Table 4.16 displays the apparent power in 1000 simulations, as shown in that paper, as well as the power for this situation, as calculated using the approximation described in Lander and Botstein (1989). We also display the power we found in the simulations of Section 4.3.3, in which there were nine chromosomes of length 100 cM, with markers every 10 cM, and with eight segregating QTLs giving a heritability of 0.2, and where one QTL was responsible for either 30 or 45% of the genetic variance. (These two cases correspond approximately to those of van Ooijen (1992).)

Our simulations gave a somewhat lower power for detection, which can be ascribed to three things. First, we required that the QTL be detected to within 20 cM, whereas van Ooijen (1992) simply looked for a significant LOD score anywhere on the same chromosome as the QTL. Second, we performed a search over nine chromosomes, instead of just one. Third, we used a different criterion for detection: van Ooijen (1992) and Lander and Botstein (1989) use an approximate 5% significance threshold, whereas we used the BIC-2 criterion,

which, in this case, is somewhat more conservative. In the simulations described above, forward selection with BIC-2 included an extraneous marker around 3–5% of the time, for all sample sizes. Note that this is a much smaller rate than was seen in the simulations of backcross experiments of Section 4.1.

It is apparent that the appropriate multiplier $\delta$ to use in the BIC-$\delta$ criterion depends not only on how one wishes to balance the errors of missing QTLs and of including extraneous ones, but also on the type of experiment performed: the behavior of the criterion is quite different for the two types of experiments considered here: the backcross and the $F_2$ intercross. The values that we used in this thesis should not be taken as given. The choice of criterion should be made carefully, and if permitted, the problem should be revisited with each new experiment.

We hope that the simulations described above may help to guide researchers in future work to assess the performance of methods for detecting QTLs. The questions we have considered (What proportion of QTLs are detected? What is the chance that at least one QTL is detected? What is the chance that a particular QTL is detected? What is the chance of identifying both of two linked QTLs?) are important to biologists doing QTL experiments, and should be studied further.

# Chapter 5

# Application

In this chapter, we apply the methods developed in this thesis to data on the number of bristles on *Drosophila melanogaster* (the fruit fly). This trait has been studied for over fifty years, because bristles are easy to count, and because the trait is highly heritable, with mostly additive genetic variation (Falconer 1989). Our analysis is intended to be an illustration of our approach, and should not be considered definitive.

Long et al. (1995) gathered a sample of fruit flies and performed 25 generations of selection, to obtain two lines which differed greatly in the number abdominal bristles. They then synthesized a set of recombinant inbred lines, with a recombinant third chromosome in an isogenic low background. The lines were genotyped at a number of genetic markers, and 40 individuals in each line (20 males and 20 females) were scored for the number of abdominal bristles and the number of sternopleural bristles. The objective of this experiment was to identify QTLs segregating in a natural population, and to map them with sufficient precision that candidate genes might be identified.

Long et al. (1995) applied a method similar to that of Zeng (1994); a half dozen possible QTLs, of quite strong effect, were identified. Looking at the pairwise interactions between the identified loci revealed evidence of strong epistatic effects.

## 5.1 Methods

### 5.1.1 Experimental methods

Long et al. (1995) took a sample of 62 flies from the Raleigh Farmer's Market, and performed 25 generations of selection for abdominal bristles to create high and low isogenic lines. (At each generation, they chose 25 extreme individuals of each sex among 100 individuals scored for each sex.) The high third chromosome was placed in an isogenic low background, and 66 recombinant inbred chromosome 3 lines were synthesized. (See, for example, Frankel (1995).) The individuals in a line are genetically identical. All lines have two copies of chromosomes 1 and 2 from the low parental line. One copy of chromosome 3 is from the low parental line; the other is a product of recombination between the third chromosomes from the low and high parental lines.

The genetic markers used in this experiment involved the *in situ* hybridization of *roo* transposable elements to the polytene chromosomes. Polytene chromosomes, found in the cells of the salivary glands in *Drosophila*, consist of a large bundle of chromatids (typically 1024), the result of repeated chromosome duplications without cell replication, and thus are considerably thicker than chromosomes at meiosis. (See, for example, Russell (1992).) When stained, they exhibit a distinct banding pattern which can be easily seen under a microscope. *Roo* transposable elements appear a large number of times in the *Drosophila* genome. Biotin-labelled *roo* elements are hybridized with the polytene chromosomes, so that the positions of the elements can be observed relative to the bands on the chromosomes. When a *roo* element is found at a particular position in only one of the two parental chromosomes, it can be used as a genetic marker.

Twenty-nine such markers were obtained. At each marker, we know whether a line received the high (H) or low (L) allele. The locations of the markers are known from their position relative to the cytological bands on the polytene chromosomes, for which map distances have been precisely estimated from numerous prior experiments. Only nineteen of the markers gave distinct H/L patterns for the 66 lines, and so we discarded those markers which showed no recombination with their left adjacent marker. The markers span 108 cM, with the largest distance between markers being 13.5 cM.

For each line, 10 males and 10 females, from two replicate vials, were scored for abdominal and sternopleural bristles. We have data on the averages and the variances of the scores for each sex, for each of the two traits. To simplify the analysis, which is intended

to be illustrative rather than definitive, we ignored the variances.

Thus, the data consist of the genotypes for the 66 lines, at each of 19 markers, the markers' map locations, and the average phenotypic score for males and females in each line, for each of two traits.

### 5.1.2 Statistical methods

We considered the two traits, abdominal bristles and sternopleural bristles, separately. Let $y_{ij}$, for $i = 1, \ldots, 66, j = 1, 2$, denote the average phenotype for line $i$, with $j = 1$ corresponding to males and $j = 2$ corresponding to females. Let $s_1 = -1$ and $s_2 = +1$, corresponding to sex, and let $x_{ik} = -1$ or 1, according to whether line $i$ had genotype LL or HL at the $k$th marker, for $k = 1, \ldots, 19$.

Initially, we considered models of the form

$$y_{ij} = \mu + \alpha_j s_j + \sum_k \beta_k x_{ik} + \sum_k \gamma_k x_{ik} s_j + \epsilon_{ij}$$

There were 39 regressors: sex, the 19 markers, and 19 sex $\times$ marker interactions.

Parameter estimates were obtained by least squares. Models were compared using BIC-type penalties, of the form

$$\text{BIC-}\delta = \log \text{RSS} + \delta q \log n / n$$

where $\delta = 2$, 2.5 or 3 and $q$ is the number of regressor variables used. The chosen models minimize the above score.

We searched through the space of models using forward selection, backward elimination and a global search over the set of $2^{39} \approx 5 \times 10^{11}$ possible models.

We also applied ANOVA and the method of Zeng (1994). In the ANOVA method, we considered the markers one at a time, forming, for marker $k$, the model

$$y_{ij} = \mu + \alpha_j s_j + \beta x_{ik} + \gamma x_{ik} s_j + \epsilon_{ij}$$

and calculating the LOD score (the $\log_{10}$ likelihood ratio) comparing the hypotheses $H_a$: $\beta \neq 0$ or $\gamma \neq 0$ and $H_0$: $\beta = \gamma = 0$. The LOD scores were compared to an overall 5% empirical threshold (Churchill and Doerge 1994): we permuted the $(y_{ij})$ and calculated the LOD score at each marker, using this new dataset. The process was repeated 1000 times, and the 95th percentile of the maximum LOD score was used as the threshold.

For Zeng's method, we determined a set of markers, $S$, either as all markers, or by performing forward selection to 3, 5 or 7 markers (during forward selection, the marker effect and the sex $\times$ marker interaction were required to enter together). Then, the markers were considered one at a time. For the $k$th marker, we looked at the following model.

$$y_{ij} = \mu + \alpha_j s_j + \beta x_{ik} + \gamma x_{ik} s_j + \sum_{l \in S^\star} \beta_l x_{il} + \sum_{l \in S^\star} \gamma_l x_{il} s_j + \epsilon_{ij}$$

where $S^\star$ is the set $S$, with markers within 5 cM of the $k$th marker (that under consideration) removed. We then calculated the LOD score comparing $H_a$: $\beta \neq 0$ or $\gamma \neq 0$ and $H_0$: $\beta = \gamma = 0$. The LOD scores were compared to a 5% empirical threshold, calculated similarly to that for the ANOVA method: the $(y_{ij})$ were permuted, and then the entire procedure to obtain the LOD scores was performed. This was repeated 1000 times, and the 95th percentile of the maximum LOD score was used as the threshold.

For the data on abdominal bristles, we also considered the inclusion of pairwise interactions, though we restricted attention to interactions between the loci which were identified in our previous analyses. Let $A$ be the chosen set of markers. We considered models of the form

$$y_{ij} = \mu + \alpha_j s_j + \sum_{k \in A} \beta_k x_{ik} + \sum_{k \in A} \gamma_k x_{ik} s_j + \sum_{k,l \in A} \beta_{kl} x_{ik} x_{il} + \sum_{k,l \in A} \gamma_{kl} x_{ik} x_{il} s_j + \epsilon_{ij}$$

The entire space of models was searched. Parameters were estimated by least squares, and models were compared using BIC-type criteria.

## 5.2 Results

Table 5.1 displays the means and SDs of the numbers of abdominal and sternopleural bristles, by sex, in the low and high parental lines. Forty individuals, for each sex and in each line, were scored. The lines show a greater difference in the number of abdominal bristles, since they were obtained by selecting for differences in this trait. The males and females in the low parental line show a striking difference in the average number of abdominal bristles.

Table 5.2 gives the genetic map for the 29 cytogenetic markers. The ten markers which are indicated by a star did not recombine with their left adjacent marker, and were dropped from the analysis. In the results below, we'll denote the markers by their locations

**Table 5.1**: Means and SDs of the numbers of abdominal and sternopleural bristles, by sex, for the two parental lines.

| | | Abdominal bristles | | Sternopleural bristles | |
|---|---|---|---|---|---|
| | | mean | SD | mean | SD |
| Low | male | 10.0 | 2.5 | 15.8 | 1.4 |
| | female | 5.5 | 2.8 | 16.9 | 1.8 |
| High | male | 22.2 | 2.3 | 20.4 | 1.6 |
| | female | 20.7 | 2.6 | 21.5 | 1.5 |

**Table 5.2**: Genetic map for the 29 cytogenetic markers. The markers indicated with a star did not recombine with the left adjacent marker in the 66 recombinant inbred lines.

| | | | | | |
|---|---|---|---|---|---|
| 61A1 | 0.0 | | 87A1 | 51.0 | ⋆ |
| 64C1 | 13.5 | | 87B1 | 51.0 | |
| 64D1 | 19.5 | ⋆ | 88B1-4 dis | 54.0 | |
| 66A1 | 23.0 | | 88E1 | 56.0 | |
| 67C4 | 23.9 | | 89D1 | 58.8 | ⋆ |
| 67F1 | 34.0 | ⋆ | 92E1 | 69.5 | |
| 68A1 | 34.5 | ⋆ | 93F1 | 73.0 | |
| 68C1 | 35.0 | | 94B1 | 76.0 | ⋆ |
| 68E1 | 37.0 | | 95A1 | 79.7 | |
| 69A1 | 38.0 | | 96B5 | 85.0 | |
| 70A1 | 40.0 | | 96F5 | 90.0 | |
| 75C1 | 46.0 | | 99F1 | 102.0 | |
| 85E1 | 49.0 | ⋆ | 100C2 | 105.0 | |
| 85F1 dis | 49.0 | ⋆ | 100F1 | 108.0 | ⋆ |
| 85F4 | 49.0 | ⋆ | | | |

as indicated on this map. For instance, the marker 64C1, at map position 13.5, will be denoted M13.5.

Table 5.3 displays the quantiles for the average numbers of abdominal and sternopleural bristles, by sex, for the 66 recombinant inbred lines. The maxima and minima are not far from the averages of the high and low parental lines, respectively (shown in Table 5.1).

Table 5.4 displays the estimated correlations between the two traits in males and females, across the 66 recombinant inbred lines. The average number of abdominal bristles for males and females in a line are highly correlated, as is the average number of sternopleural bristles between sexes. The correlation between the average number of abdominal bristles and the average number of sternopleural bristles, within males or within females,

**Table 5.3**: Quantiles for the average numbers of abdominal and sternopleural bristles, by sex, for the recombinant inbred chromosome 3 lines.

| | Abdominal bristles | | Sternopleural bristles | |
|---|---|---|---|---|
| | male | female | male | female |
| min | 7.9 | 2.8 | 14.6 | 15.3 |
| 25th %ile | 10.9 | 5.7 | 16.1 | 17.4 |
| median | 13.4 | 10.5 | 16.7 | 17.7 |
| 75th %ile | 17.7 | 16.6 | 18.7 | 20.1 |
| max | 20.2 | 22.7 | 20.6 | 22.2 |

**Table 5.4**: Estimated correlations between traits across the recombinant inbred chromosome 3 lines. (AB denotes abdominal bristles, and SB denotes sternopleural bristles.)

| | female AB | male SB | female SB |
|---|---|---|---|
| male AB | 0.93 | 0.49 | 0.48 |
| female AB | | 0.44 | 0.41 |
| male SB | | | 0.95 |

are much less strongly correlated.

### 5.2.1 Abdominal bristles

Forward selection using each of the three criteria, BIC-2, BIC-2.5 and BIC-3, indicated a model for abdominal bristles which contained seven variables: sex, M13.5, M35, M46, M69.5, M90 and sex $\times$ M35. Backward elimination gave similar results, though the effect sex $\times$ M46 was given in place of sex $\times$ M35. A search over all $2^{39}$ models (by branch-and-bound) showed that the model indicated by forward selection gave the global minimum for each of the three BIC-type criteria.

Table 5.5 contains the estimated coefficients and their estimated standard errors (SEs) for this model. (Note that the estimated regression coefficients are expected to exhibit selection bias. See the discussion in Section 6.1, beginning on page 83.) All of the QTLs have positive effects, indicating that the high allele gives an increase in the number of bristles over the low allele. The QTLs identified had effects of 1.1–1.8 bristles, corresponding to differences of 2.2–3.6 bristles between the high and low lines. The effect for sex was negative, indicating that females had fewer abdominal bristles, on average, than males. The sex $\times$ M35 interaction indicates that, for the lines with genotype HL at M35, the two sexes had nearly the same average number of bristles. The estimated residual SD was $\hat{\sigma} = 1.65$.

**Table 5.5**: Estimated coefficients and estimated standard errors (in parentheses) for the chosen model for abdominal bristles.

| | | |
|---|---|---|
| intercept | 12.9 | (0.2) |
| sex | –1.4 | (0.1) |
| M13.5 | 1.2 | (0.2) |
| M35 | 1.1 | (0.3) |
| M46 | 1.8 | (0.3) |
| M69.5 | 1.1 | (0.2) |
| M90 | 1.7 | (0.2) |
| sex × M35 | 0.9 | (0.1) |

**Table 5.6**: Estimated 5% LOD thresholds for ANOVA and Zeng's method, obtained by 1000 permutations of the data on abdominal bristles.

| | Zeng, forward selection | | | |
|---|---|---|---|---|
| ANOVA | 3 | 5 | 7 | Zeng, all |
| 2.3 | 3.1 | 3.4 | 3.5 | 3.3 |

Replacing a marker by its neighbor resulted in a change in $\log_{10}$ likelihood of more than 1. Thus the locations of the QTLs are quite well resolved.

One of the observations, for the females of line 55, deviated quite markedly from the expected. The average number of abdominal bristles for these flies was $y = 12.7$, but the fitted value for the above model was $\hat{y} = 4.7$. The standardized, studentized residual (see, for example, McCullogh and Nelder 1989) was 4.9. It is interesting to note that this line had genotype LL at all markers. Thus, we would expect its average to be similar to that of the low parental line. (Nine lines were LL at all markers; eight were HL at all markers. All of these lines, except for the one noted above, correspond well to the average counts of the corresponding parental line, shown in Table 5.1.) This line, indicated as an outlier, may have been contaminated with another line after genotyping but prior to phenotyping (T. Long, personal communication). In any case, removal of this point had little effect on the estimated coefficients and SEs, and had no effect on the choice of models.

Table 5.6 contains the 5% LOD thresholds for ANOVA and for the method of Zeng (1994), using all markers, and using forward selection up to 3, 5 and 7 markers. The thresholds were estimated by repeated permutations of the phenotype data, as described in Churchill and Doerge (1994). The estimated standard errors for these thresholds are all around 0.1.

**Table 5.7**: The main effects included in the chosen models for abdominal bristles, for each of the methods used.

| BIC | | | | Zeng | | | |
|---|---|---|---|---|---|---|---|
| 2 | 2.5 | 3 | ANOVA | 3 | 5 | 7 | all |
| M13.5 | M13.5 | M13.5 | | M13.5 | M13.5 | | |
| M35 | M35 | M35 | | | M35 | M35 | |
| M46 | M46 | M46 | M46 | M46 | M46 | M46 | (none) |
| M69.5 | M69.5 | M69.5 | | M69.5 | M69.5 | M69.5 | |
| M90 | M90 | M90 | | M90 | M90 | M90 | |

Application of ANOVA gives a single broad peak: nearly all markers give significant LOD scores, and so we cannot separate the individual QTLs at all.

Zeng's approach, using all markers which are at least 5 cM away from the marker being tested, gave no significant markers. Zeng's approach using forward selection to 5 markers indicated the same model as BIC, except that when using this method, we always included the sex $\times$ marker interaction, when a marker was identified. Using forward selection to 3 markers, M35 was not significant. Using forward selection to 7 markers, M13.5 was not significant.

For ease of comparison, Table 5.7 displays the markers whose main effects were included in the chosen models for abdominal bristles obtained by each of the different methods. For the models obtained using the BIC-type criteria, we display those indicated by a complete search over the model space.

### 5.2.2 Sternopleural bristles

Forward selection using BIC-2, BIC-2.5 and BIC-3 indicated a model for sternopleural bristles with just sex and M0. Backward elimination, and a complete search over all $2^{39}$ models (by branch-and-bound), using BIC-2.5 and -3, also indicated this model. Backward elimination and a complete search, using BIC-2, indicated a model containing sex, M0, M51, M54, M90 and M102.

Table 5.8 contains the estimated coefficients and estimated SEs for the larger model. (These estimates are expected to exhibit selection bias.) The estimated residual SD for this model was $\hat{\sigma} = 0.74$. The estimated residual SD for the model containing only sex and M0 was 0.86. The putative QTLs which were not found by forward selection, and which show up only with the BIC-2 criterion, appear as two pairs of QTLs tightly linked in

**Table 5.8**: Estimated coefficients and estimated standard errors (in parentheses) for the chosen model for sternopleural bristles.

| | | |
|---|---|---|
| intercept | 18.2 | (0.1) |
| sex | 0.6 | (0.1) |
| M0 | 1.5 | (0.1) |
| M51 | –1.2 | (0.3) |
| M54 | 1.3 | (0.3) |
| M90 | 0.4 | (0.1) |
| M102 | –0.4 | (0.1) |

repulsion.

Replacing the markers, one at a time, by their neighbors shows that marker M56 works as well as M54, and M105 works as well as M102.

A number of moderately outlying observations were seen, but dropping these had little effect on the results.

The estimated LOD thresholds for ANOVA and Zeng's methods, obtained by permuting the sternopleural bristle phenotype data, were within one SE of those obtained for the abdominal bristle data, shown in Table 5.6.

With ANOVA, marker M0 gave the largest LOD score, but markers M13.5–M35 also gave LOD scores above the 5% empirical threshold. The other markers had LOD scores below the threshold. Zeng's approach using forward selection to 3 markers indicated only marker M0. Using forward selection to 5 markers indicated markers M0 and M102. Using forward selection to 7 markers indicated markers M0, M51 and M105. Using all markers indicated M0 and M56.

For ease of comparison, Table 5.9 displays the markers whose main effects were included in the chosen models for sternopleural bristles obtained by each of the different methods. For the models obtained using the BIC-type criteria, we display those indicated by a complete search over the model space.

### 5.2.3 Epistasis

We considered including pairwise interactions, for the model for the average number of abdominal bristles. We restricted attention to the markers which were indicated in the previous analyses: M13.5, M35, M46, M69.5 and M90. Thus, we had 31 regressors: sex, 5 markers, 5 sex $\times$ marker interactions, 10 marker $\times$ marker interactions, and 10 sex

**Table 5.9**: The main effects included in the chosen models for sternopleural bristles, for each of the methods used.

| BIC | | | | Zeng | | | |
|---|---|---|---|---|---|---|---|
| 2 | 2.5 | 3 | ANOVA | 3 | 5 | 7 | all |
| M0 | M0 | M0 | M0 | M0 | M0 | M0 | M0 |
| M51 | | | | | | M51 | |
| M54 | | | | | | | |
| | | | | | | | M56 |
| M90 | | | | | | | |
| M102 | | | | | M102 | | |
| | | | | | | M105 | |

**Table 5.10**: Estimated coefficients and estimated standard errors for the chosen model for abdominal bristles, when pairwise interactions were allowed.

| | | |
|---|---|---|
| intercept | 12.5 | (0.2) |
| sex | –1.4 | (0.1) |
| M13.5 | 1.2 | (0.2) |
| M35 | 0.8 | (0.3) |
| M46 | 1.8 | (0.3) |
| M69.5 | 1.5 | (0.2) |
| M90 | 0.6 | (0.2) |
| M35 × M69.5 | 0.7 | (0.2) |
| sex × M35 | 0.9 | (0.1) |

× marker × marker interactions. A full search of the space of models gave, for BIC-2, the model previously found, with the addition of a single marker × marker interaction: M35 × M69.5. With BIC-2.5 and BIC-3, the markers M46 and M90 were dropped. Table 5.10 contains the estimated coefficients and estimated SEs for the larger model. (Again, these estimates are expected to exhibit selection bias.) Comparing these coefficients to those in Table 5.5 (page 76), one observes that the main effects for the markers have not changed dramatically. The coefficient for M35 went from 1.1 to 0.8, and the coefficient for M69.5 went from 1.1 to 1.5.

## 5.3 Discussion

Since the "truth" is not known for the data analyzed in this chapter, we cannot know how well the methods have performed in uncovering it, but this application is still

useful in comparing the different methods in the face of a real problem.

For the abdominal bristles trait, forward selection performed very well, giving the model which minimized the BIC criteria globally. The poor performance of ANOVA-type methods (including interval mapping), in the face of linked QTLs, is clearly seen. At least five QTLs for abdominal bristles were segregating in this experiment, and ANOVA was unable to distinguish them. With an appropriate choice of markers to use as regressors, Zeng's method gave the same model as chosen by BIC. However, different methods of choosing the set of regressors gave quite different results; when all markers were used, no markers had LOD scores above the 5% empirical threshold. This is the clear drawback to Zeng's approach: how to choose this set of regressors appropriately is not clear, and the results depend greatly on this decision.

The sternopleural bristle data may indicate the presence of QTLs tightly linked in repulsion. These appear only when using BIC-2, however, and so they may be extraneous. Forward selection did not pick them out, whereas backward elimination did. (This behavior corresponds to our experience with simulations, discussed in Chapter 4. When feasible, one should apply not just forward selection, but also backward elimination, and maybe a branch-and-bound approach, to insure that the best models are seen.) Zeng's method indicated only single QTLs in the regions of the pairs of QTLs identified with BIC-2. The very tight linkage of these QTLs would make it difficult for Zeng's method to uncover them, since when the LOD scores are calculated, tightly linked loci are never considered together.

Our work to search for pairwise interactions should be considered only as preliminary. The big problem in this situation is the search through possible models, since the inclusion of interactions leads to enormous increases in the number of regressors. Our approach, of considering interactions only between loci with clear main effects, was one of convenience, and should not be recommended generally. Loci with negligible main effects may be found to be important when epistasis is considered. One marker $\times$ marker interaction, comparable in size to the main effects of many loci, was clearly present. Thus, the presence of epistasis in quantitative traits should not be disregarded.

It is prudent that we compare the results obtained here to the original analysis presented in Long et al. (1995). They used a method similar to the approach of Zeng (1994), using all markers, though they considered *intervals* rather than markers, calling an interval "high" if both of the flanking markers were high, and low if both flanking markers were low, and dropping data for lines in which an interval showed a recombination event. The

intervals that they identified were adjacent to the markers we chose in our analysis, except for two cases, where they were next-to-adjacent. The biggest difference seen was in the results of the analysis of epistasis. Long et al. (1995) considered the pairwise interactions one at a time, and found a large number of them giving significant effects, whereas our analysis identified only one pairwise interaction.

# Chapter 6

# Conclusions and discussion

In this thesis, we have considered the problem of identifying quantitative trait loci (QTLs) in large experimental crosses. We have assumed that the QTLs act additively, and that crossovers in meiosis occur with no interference. The standard approaches to this problem involve multiple tests of hypotheses, with a genome-wide significance threshold which controls for the multiple tests. Our approach has been to view the problem as one of model selection, and to use standard methods for selecting subsets of variables in regression. This approach gives quite good results, and helps to focus on the important issue in the problem: the balance between the problems of missing important loci and including extraneous ones.

In light of the observation that the variation in most quantitative traits appears to be the result of the action of multiple loci, methods which model a single QTL at a time, such as analysis of variance and interval mapping, should be expected to perform poorly in comparison to the methods which model multiple QTLs, such as multiple regression and composite interval mapping. In addition, since, with most experiments, interval mapping is unable to resolve the location of QTLs to within a single marker interval, it is clear that, in the identification of QTLs, very little is lost by considering only the marker loci themselves. We are thus lead to the approach studied in this thesis: choosing a subset of markers by applying well-known methods for subset selection in regression.

There are a number of very important issues which have been neglected in this thesis. We conclude our discussion with brief statements on several of these issues.

## 6.1 Selection bias

The estimated effects of identified QTLs, and the estimated genetic variance ascribed to those QTLs, will be greatly biased. If a locus has a small estimated coefficient, it will not in the identified as a QTL. Thus, the expected values of the estimated coefficients for chosen loci, given that they have been chosen, will be too large. This is true for all of the methods described in this thesis, including interval mapping, composite interval mapping, and our own model selection approach.

Severe bias in the least squares estimates of the coefficients in a regression problem, when the variables in the regression equation were obtained by subset selection, is a well known problem (Miller 1990). This bias appears to not be so well known among scientists studying QTLs. We have seen no mention of the issue accompanying the analyses of QTL experiments.

To illustrate the possible size of the bias, we performed a small simulation. We used a backcross of 250 progeny, with nine chromosomes, of length 100 cM each, with 11 equally spaced markers (10 cM spacing). We used four QTLs, at the center of chromosomes 1–4, with effects 1.0, 0.75, 0.5 and 0.25. The environmental variation was normally distributed with standard deviation $\sigma = 1$. We used forward selection with the BIC-2.5 criterion, and performed 10,000 simulations.

If the correct model is fit, the estimated coefficients are all unbiased, and have standard error 0.13. In Table 6.1, we display the results of the simulations. The second column gives the percent of the simulations in which each of the QTLs were chosen, using forward selection with BIC-2.5. The third column gives the estimated selection bias in the estimated coefficients, as a percent of the true effects, $\beta$. For each QTL, we take the average of the estimated coefficients, among the simulations in which that QTL was chosen. The fourth column gives the root mean square (RMS) of the nominal standard errors for the coefficient estimates. Again, for each QTL, we use only those simulations for which it was detected. The nominal standard errors for the regression coefficients are the square root of the diagonal elements of $\hat{\sigma}^2(X'X)^{-1}$, where $\hat{\sigma}^2$ is the estimated residual variance. The selection bias in the estimated QTL effects is negligible for the QTLs which could be detected with high power. For the QTL with effect $0.5\sigma$, the bias is moderately large, and for the QTL with very small effect, which was identified only 3% of the time, the bias was very large. The estimated standard errors show no selection bias.

**Table 6.1**: Results of simulations to study selection bias in QTL effects and estimated SEs.

| true effect | % chosen | ave$(\hat{\beta} - \beta)/\beta$ | RMS$(\widehat{\mathrm{SE}})$ |
|---|---|---|---|
| 1.0 | 96 | 0% | 0.13 |
| 0.75 | 88 | 2% | 0.13 |
| 0.5 | 46 | 18% | 0.13 |
| 0.25 | 3 | 110% | 0.13 |

Miller (1990) discusses a number of methods to estimate and adjust for selection bias. We neglect this issue here, because we are chiefly interested in the problem of identifying the QTLs, and are less interested in the estimated effects of those loci.

## 6.2 Missing data

It is not unusual in QTL experiments to find that the genotype data is incomplete: not all progeny were typed for all genetic markers. The complex biochemical reactions which are performed to obtain genotypes will occasionally fail, and the resulting holes in the genotype data may be difficult to fill.

If only a very small proportion of genotypes are missing, it may be possible to simply drop any observations whose genotypes are missing at the markers under consideration. For example, in performing ANOVA at a marker, one could use only those observations which were typed at that marker. Thus, different sets of observations would be used at different markers. There is some loss of efficiency in this approach, since the genotype at markers near the one for which the data is missing provide some information about the likely genotype at that position.

When there is a great deal of missing data, so that there could be a great loss of efficiency when using the above approach, it will be important to use a method which helps to fill in the missing data. The most likely candidate would be a method like interval mapping, using the EM algorithm (Dempster et al. 1977).

A similar problem arises when some of the markers are less than fully informative. In a backcross or intercross using highly inbred lines, all markers are fully informative (meaning that we have complete knowledge about the grandparental origins of the alleles received by an individual). However, more complex experiments are sometimes performed, such as an intercross between outbred lines.

For example, consider an experiment involving four outbred grandparents, two parents, and a large number of progeny. Each of the two parents receives one allele from each of its parents; label these alleles A, B, C and D, so that one parent has genotype AB and the other has genotype CD. Then the progeny will have four possible genotypes: AC, AD, BC or BD.

At some markers, one of the two parents may be homozygous. For example, the alleles C and D may really be the same. Thus, the progeny will have one of only two genotypes at that marker: AC or BC. But in this sort of experiment, we will want to follow the effects of the allele from each grandparent, and so we will need to use nearby markers which are fully informative to try to fill in this missing information. Again, an approach similar to interval mapping may be required.

To summarize, the model selection approach described in this thesis, which uses multiple regression at marker loci, may suffer from a loss of efficiency when faced with a substantial amount of missing data or when a good portion of the markers are not fully informative. In such situations, it may be important to use a method like interval mapping to fill in the missing information.

## 6.3 Epistasis

It cannot be denied that the situation discussed in this thesis, in which QTLs are assumed to act additively, is a great simplification of reality. Experiments on bristles in *Drosophila* (Shrimpton and Robertson 1988; Long et al. 1995) supply strong evidence for epistatic interactions between QTLs, with some epistatic effects being as large as the main effects of many loci.

All of the common statistical methods used to detect QTLs neglect epistasis. Yet it may be that some loci show an effect only in the presence of a particular allele at another locus. (For an example, see Shrimpton and Robertson (1988).) Indeed, the term *epistasis* derives from the situation in which two genes are related in some biochemical pathway, so that a mutation in the gene which is "downstream" in the pathway shows no effect when there is a mutation in the gene which is "upstream." Thus, if epistasis is ignored, one may miss important loci, whose effects are apparent only when considering interactions.

The usual method used to detect epistatic effects is to consider pairwise interactions between loci, generally by performing multiple tests of hypotheses, looking at the pair-

wise interactions one at a time (Tanksley 1993). The biggest difficulty with this approach, and with the epistasis problem generally, is the enormous number of possible interactions to consider. With 100 markers, there are $100 \times 99/2 = 4,950$ different pairwise interactions. As a result, even if some of these show a large effect, one is left with very little power to detect epistasis.

We see no easy solution to this problem. One is faced with "the curse of dimensionality." With 100 markers, there are $2^{100} \approx 10^{30}$ possible models which include only main effects, and there are $2^{5050} \approx 10^{1520}$ possible models which also include pairwise interactions.

A quite different approach may help. One might consider going after the interactions from the start, rather than tacking them on as pairwise interactions at the end. It is natural, in this instance, to consider tree-based models, for which interactions are the rule. (See Breiman et al. 1984). The search through the space of models is still, and it will always be, a problem. But the results may be somewhat improved.

## 6.4 Multiple traits

Because so much effort is expended in generating and then genotyping the progeny in a QTL experiment, scientists are rarely satisfied with measuring only a single quantitative trait. Sometimes as many as 40 different traits are measured on each individual (Edwards et al. 1987). It is hoped that large QTLs will be detected for at least some of these traits. More importantly, there are a number of questions that can be answered only with an experiment that looks at several traits simultaneously. Chief among these regards the phenomenon of pleiotropy, in which action at a single locus leads to variation in a number of traits. Pleiotropy is of special interest to scientists who are performing selection experiments, in which one is trying to simultaneously improve several traits. For example, one might desire a eucalyptus tree that not only grows more quickly, but also has more dense wood, traits that tend to be negatively correlated. If this negative association is primarily due to pleiotropy, it will be very difficult to improve both traits at once.

In the analysis of QTL data, multiple traits are often considered one at a time. It should be emphasized, however, that there can be a great advantage to analyzing the traits simultaneously. In particular, pleiotropy is best tackled by considering models for multiple traits, and perhaps testing the hypothesis that two QTLs, each acting on a different trait, correspond to the same locus. For a discussion of this topic, see Jiang and Zeng (1995).

# References

Akaike, H. (1969) Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics **21**:243–247.

An, H., and L. Gu (1985) On the selection of regression variables. Acta Mathematicæ Applicatæ Sinica **2**:27–36.

Andersson, L., C. S. Haley, H. Ellegren, S. A. Knott, M. Johansson, K. Andersson, L. Andersson-Eklund, I. Edfors-Lilja, M. Fredholm, I. Hansson, J. Håkansson and K. Lundström (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263**:1771–1774.

Basten, C. J., B. S. Weir and Z.-B. Zeng (1996) *QTL Cartographer: A reference manual and tutorial for QTL mapping.* Program in Statistical Genetics, Department of Statistics, North Carolina State University.

Beavis, W. D., D. Grant, M. Albertsen and R. Fincher (1991) Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theoretical and Applied Genetics **83**:141–145.

Berrettini, W. H., T. N. Ferraro, R. C. Alexander, A. M. Buchberg and W. H. Vogel (1994) Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains. Nature Genetics **7**:54–58.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984) *Classification and regression trees.* Wadsworth, Pacific Grove, California.

Carbonell, E. A., M. J. Asins, M. Baselga, E. Balansard and T. M. Gerig (1993) Power studies in the estimation of genetic parameters and the localization of quantitative trait

loci for backcross and doubled haploid populations. Theoretical and Applied Genetics **86**:411–416.

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

Cowen, N. M. (1989) Multiple linear regression analysis of RFLP data sets used in mapping QTLs. Pages 113–116 in *Development and application of molecular markers to problems in plant genetics*, edited by T. Helentjaris and B. Burr. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Darvasi, A., A. Weireb, V. Minke, J. I. Weller and M. Soller (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134**:943–951.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**:1–38.

deVicente, M. C., and S. D. Tanksley (1993) QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics **134**:585–596.

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

Dupuis, J., P. O. Brown and D. Siegmund (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics **140**:843–856.

Edwards, M. D., C. W. Stuber and J. F. Wendel (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116**:113–125.

Falconer, D. S. (1989) *Introduction to quantitative genetics*, third edition. Wiley, New York.

Frankel, W. J. (1995) Taking stock of complex trait genetics in mice. Trends in Genetics **11**:471–477.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (1995) *Bayesian data analysis.* Chapman and Hall, New York.

Geman, S., and D. Geman (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence **6**:721–741.

Grattapaglia, D., F. L. G. Bertolucci, R. Penchel and R. R. Sederoff (1996) Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. Genetics **144**:1205–1214.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**:711–732.

Haldane, J. B. S. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics **8**:299–309.

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

Hannan, E. J., and B. G. Quinn (1979) The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B **41**:190–195.

Hastings, W. F. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**:97–109.

Hyne, V., M. J. Kearsey, D. J. Pike and J. W. Snape (1995) QTL analysis: unreliability and bias in estimation procedures. Molecular Breeding **1**:273–282.

Jansen, R. C. (1993) Interval mapping of multiple quantitative trait loci. Genetics **135**:205–211.

Jansen, R. C. (1994) Controlling the type I and type II errors in mapping quantitative trait loci. Genetics **138**:871–881.

Jansen, R. C., and P. Stam (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136**:1447-1455.

Jiang, C., and Z.-B. Zeng (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140**:1111–1127.

Kearsey, M. J., and V. Hyne (1994) QTL analysis: a simple 'marker-regression' approach. Theoretical and Applied Genetics **89**:698–702.

Knapp, S. J. (1991) Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. Theoretical and Applied Genetics **81**:333–338.

Knapp, S. J., W. C., Bridges, Jr., and D. Birkes (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theoretical and Applied Genetics **79**:583–592.

Knott, S. A., and C. S. Haley (1992) Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genetics Research **60**:139–151.

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

Long, A. D., S. L. Mullaney, L. A. Reid, J. D. Fry, C. H. Langley and T. F. C. Mackay (1995) High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. Genetics **139**:1273–1291.

McCullagh, P., and J. A. Nelder (1989) *Generalized linear models*, second edition. Chapman and Hall, New York.

Martínez, O., and R. N. Curnow (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoretical and Applied Genetics **85**:480–488.

Mendel, G. J. (1866) Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn **4**:3–47.

Meng, X.-L., and D. B. Rubin (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80**:267–278.

Miller, A. J. (1990) *Subset selection in regression*. Chapman and Hall, New York.

Nilsson-Ehle, H. (1909) Kreuzungsuntersuchungen an Hafer und Weizen. Lunds Universitets Årsskrift.

Paterson, A. H., J. W. DeVerna, B. Lanini and S. D. Tanksley (1990) Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes in an interspecies cross of tomato. Genetics **124**:735–742.

Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S. E. Lincoln, E. S. Lander and S. D. Tanksley (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics **127**:181–197.

Rao, C. R., and Y. Wu (1989) A strongly consistent procedure for model selection in a regression problem. Biometrika **76**:369–374.

Rebaï, A., B. Goffinet and B. Mangin (1995) Comparing power of different methods for QTL detection. Biometrics **51**:87–99.

Russell, P. J. (1992) *Genetics*, third edition. HarperCollins, New York.

Satagopan, J. M., and B. S. Yandell (1996) Estimating the number of quantitative trait loci via Bayesian model determination. Special contributed paper session on genetic analysis of quantitative traits and complex diseases, Biometrics section, Joint Statistical Meetings, Chicago, Illinois.

Satagopan, J. M., B. S. Yandell, M. A. Newton and T. C. Osborn (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144**:805–816.

Sax, K. (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Journal of Theoretical Biology **117**:1-10.

Schwarz, G. (1978) Estimating the dimension of a model. The Annals of Statistics **6**:461–464.

Shao, J. (1996) Bootstrap model selection. Journal of the American Statistical Association **91**:655–665.

Shrimpton, A. E., and A. Robertson (1988) The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. I. Allocation of third chromosome sternopleural bristle effects to chromosome sections. Genetics **118**:437–443.

Simpson, S. P. (1989) Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theoretical and Applied Genetics **77**:815–819.

Smith, M. (1996) Nonparametric regression: a Markov chain Monte Carlo approach. Ph.D. dissertation, University of New South Wales.

Soller, M., T. Brody and A. Genizi (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theoretical and Applied Genetics **47**:35–39.

Stuber, C. W., S. E. Lincoln, D. W. Wolff, T. Helentjaris and E. S. Lander (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics **132**:823–839.

Tanksley, S. D. (1993) Mapping polygenes. Annual Review of Genetics **27**:205–233.

Thisted, R. A. (1988) *Elements of statistical computing.* Chapman and Hall, New York.

Thoday, J. M. (1961) Location of polygenes. Nature **191**:368–370.

van Ooijen, J. W. (1992) Accuracy of mapping quantitative trait loci in autogamous species. Theoretical and Applied Genetics **84**:803–811.

Venables, W. N. and Ripley, B. D. (1994) *Modern applied statistics with S-Plus.* Springer-Verlag, New York.

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

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

Wu, W.-R., and W.-M. Li (1994) A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. Theoretical and Applied Genetics **89**:535–539.

Wu, W.-R., and W.-M. Li (1996) Model fitting and model testing in the method of joint mapping of quantitative trait loci. Theoretical and Applied Genetics **92**:477–482.

Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90**:10972–10976.

Zeng, Z.-B. (1994) Precision mapping of quantitative trait loci. Genetics **136**:1457–1468.

# Appendix

This appendix contains a proof of the proposition in Chapter 3, that, in the case of strictly additive QTLs which are located exactly at marker loci, and assuming no interference, forward selection using a BIC-type criterion is consistent.

An and Gu (1985) showed that, in the context of a linear model with a fixed number of independent variables, minimizing the BIC criterion over all possible models gives a consistent estimate of the model. (The argument holds for any criterion of the form $\log \mathrm{RSS} + \delta n / \log n$, where $\delta > 0$.) Moreover, they showed that using backward elimination with BIC still gives a consistent procedure. Forward selection, on the other hand, is overconsistent, meaning that the estimated model will contain all of the correct independent variables, but may contain some extraneous ones as well. In the situation discussed in this thesis, the independent variables have a Markov structure. As a result, forward selection using a BIC-type criterion is consistent.

Consider a backcross. Let $y$ be the vector of trait values. Assume that there are $p$ QTLs, with $z_1, z_2, \ldots, z_p$ the vectors of genotypes, coded as –1 and +1. Assume that

$$y = \mu + \sum_{s=1}^{p} \beta_s z_s + \epsilon$$

Let $X$ be the matrix of marker genotypes, the first column of $X$ containing all 1's (corresponding to the intercept).

Consider two markers, with genotypes $x_i$ and $x_j$, coded as –1 and +1. Note that $\mathbf{E}(x_i) = 0$, $\mathbf{E}(x_i^2) = 1$, and $\mathbf{E}(x_i x_j) = (1 - 2r)$, where $r$ is the recombination fraction between the two markers. Assuming no interference, $r = (1 - e^{-2d/100})/2$, where $d$ is the

distance (in cM) between the two markers. Note also that

$$\begin{aligned}\mathbf{E}(x_i y) &= \mu\mathbf{E}(x_i) + \sum_s \beta_s \mathbf{E}(x_i z_s) + \mathbf{E}(x_i \epsilon) \\ &= \sum_s \beta_s (1 - 2r_{x_i z_s})\end{aligned}$$

The sum above is over only those QTLs which are linked to the marker $x_i$, since when a QTL is not linked to a marker, the recombination fraction $r_{x_i z_s} = 1/2$, and so $1 - 2r_{x_i z_s} = 0$.

In the following, terms like $(1-2r)$ show up regularly. Consider three markers $x_i$, $x_j$ and $x_k$, with $x_j$ located between the other two. Let $r_{ij}$, $r_{jk}$ and $r_{ik}$ denote the recombination fractions between pairs of the markers. Under no interference, $r_{ik} = r_{ij} + r_{jk} - 2r_{ij}r_{jk}$. Let $\rho_{ij} = 1 - 2r_{ij}$, and define $\rho_{ij}$ and $\rho_{ik}$ similarly. Then $\rho_{ik} = \rho_{ij}\rho_{jk}$. This property will be used extensively.

Suppose there are $c$ chromosomes. By the Law of Large Numbers,

$$(X'X)/n \xrightarrow{\text{a.s.}} \begin{pmatrix} 1 & & & & \\ & \Lambda_1 & & & \\ & & \Lambda_2 & & \\ & & & \ddots & \\ & & & & \Lambda_c \end{pmatrix}$$

where each matrix $\Lambda_k$ is symmetric and is of the form $\Lambda = (\lambda_{ij})$ with elements

$$\lambda_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho_i \cdots \rho_{j-1} & \text{if } i < j \end{cases}$$

where $\rho_i = 1 - 2r_i$ and $r_i$ is the recombination fraction between the $i$th and $(i+1)$st markers on the chromosome. (Note that, to simplify the notation, we neglect the dependence on $k$.)

The inverse of such $\Lambda$ has elements

$$\lambda^{ij} = \begin{cases} 1 + \frac{\rho_1^2}{1-\rho_1^2} & \text{if } i = j = 1 \\ 1 + \frac{\rho_{M-1}^2}{1-\rho_{M-1}^2} & \text{if } i = j = M \\ 1 + \frac{\rho_1^2}{1-\rho_1^2} + \frac{\rho_{M-1}^2}{1-\rho_{M-1}^2} & \text{if } i = j \text{ and } i \neq 1 \text{ or } M \\ -\frac{\rho_i}{1-\rho_i^2} & \text{if } j = i+1 \text{ or } i-1 \\ 0 & \text{otherwise} \end{cases}$$

where $M$ is the number of markers on the chromosome.

The matrix $(X'y)/n$ converges almost surely to a matrix whose elements are of the form $\mathbf{E}(xy) = \sum_s \beta_s \rho_{xz_s}$, where $\rho_{xz_s} = 1 - 2r_{xz_s}$, and, as mentioned above, the sum is over only those QTLs which are linked to the marker.

Because of the form of the $X'X$ and $X'y$ matrices, only a single chromosome need be considered. Consider the change in residual sum of squares (RSS) associated with adding a new marker; we wish to show that the marker giving the maximum change in the RSS is always at a QTL. Note that

$$\begin{aligned}
\text{RSS} &= \left(y - X\hat{\beta}\right)' \left(y - X\hat{\beta}\right) \\
&= \left[y - X\left(X'X\right)^{-1} X'y\right]' \left[y - X\left(X'X\right)^{-1} X'y\right] \\
&= y'y - y'X\left(X'X\right)^{-1} X'y
\end{aligned}$$

Thus minimizing the RSS is equivalent to maximizing $y'X(X'X)^{-1}X'y$. Because of the almost sure convergence described above, we deal only with the limiting value of the change in RSS.

Suppose the chromosome has $p$ QTLs, at locations $z_1 < z_2 < \ldots < z_p$. Consider a location $x$, between QTLs $z_u$ and $z_{u+1}$. Suppose that the QTLs $z_i$ and $z_j$ are the closest flanking QTLs which are currently in the model for the trait $y$, with $z_i < z_j$. We want to show that there is a greater change in RSS when adding $z_u$ or $z_{u+1}$ rather than $x$.

Let $\rho_s = 1 - 2r_s$, where $r_s$ is the recombination fraction between the QTL $z_s$ and the location $x$. Let $\rho_{st} = 1 - 2r_{st}$, where $r_{st}$ is the recombination fraction between the QTL $z_s$ and the QTL $z_t$.

The absolute decrease in RSS when adding the locus $x$ is

$$\begin{aligned}
D(x) = &\left(\sum \beta_s \rho_s\right)^2 \left(1 + \frac{\rho_i^2}{1-\rho_i^2} + \frac{\rho_j^2}{1-\rho_j^2}\right) \\
&+ \left(\sum \beta_s \rho_{si}\right)^2 \left(\frac{\rho_i^2}{1-\rho_i^2} - \frac{\rho_{ij}}{1-\rho_{ij}^2}\right) + \left(\sum \beta_s \rho_{sj}\right)^2 \left(\frac{\rho_j^2}{1-\rho_j^2} - \frac{\rho_{ij}}{1-\rho_{ij}^2}\right) \\
&- 2\frac{\rho_i}{1-\rho_i^2}\left(\sum \beta_s \rho_{si}\right)\left(\sum \beta_s \rho_s\right) - 2\frac{\rho_j}{1-\rho_j^2}\left(\sum \beta_s \rho_{sj}\right)\left(\sum \beta_s \rho_s\right) \\
&+ 2\frac{\rho_{ij}}{1-\rho_{ij}^2}\left(\sum \beta_s \rho_{si}\right)\left(\sum \beta_s \rho_{sj}\right) \\
= &\sum \beta_s \beta_t A_{s,t}
\end{aligned}$$

where

$$A_{s,t} = \rho_s\rho_t + \frac{\rho_i}{1-\rho_i^2}\left(\rho_{si}\rho_{ti}\rho_i + \rho_s\rho_t\rho_i - \rho_{si}\rho_t - \rho_{ti}\rho_s\right)$$
$$+ \frac{\rho_j}{1-\rho_j^2}\left(\rho_{sj}\rho_{tj}\rho_j + \rho_s\rho_t\rho_j - \rho_{sj}\rho_t - \rho_{tj}\rho_s\right)$$
$$+ \frac{\rho_{ij}}{1-\rho_{ij}^2}\left(\rho_{si}\rho_{tj} + \rho_{sj}\rho_{ti} - \rho_{si}\rho_{ti}\rho_{ij} - \rho_{sj}\rho_{tj}\rho_{ij}\right)$$

We now simplify the form of the above. First, consider the case where $z_s \le z_i < x < z_j$. Using the fact that $\rho_s = \rho_{si}\rho_i$ and $\rho_{sj} = \rho_{si}\rho_{ij}$, it is not hard to show that $A_{s,t} = 0$. By symmetry, then, $A_{s,t} = 0$ whenever either $z_s$ or $z_t$ is outside of $z_i$ and $z_j$. Thus, the change in RSS, associated with including the marker $x$, depends only on QTLs between $z_i$ and $z_j$.

Now, consider $z_i < z_s < x < z_t < z_j$. After another bit of algebra, we obtain

$$A_{s,t} = \rho_{st}\frac{(1-\rho_{is}^2)(1-\rho_{jt}^2)}{1-\rho_{ij}^2}$$

So when $z_s$ and $z_t$ are on different sides of $x$, the value $A_{s,t}$ doesn't depend on the location of $x$.

Finally, consider $z_i < z_s < z_t < x < z_j$. A bit more algebra gives

$$A_{s,t} = \rho_s\rho_t\frac{(1-\rho_{si}^2)(1-\rho_{ti}^2)}{1-\rho_i^2} - \rho_{sj}\rho_{tj} + \frac{\rho_{ij}\rho_{si}\rho_{sj}}{1-\rho_{ij}^2}\left[1 + \rho_{st}^2\left(1 - \rho_{si}^2 - \rho_{tj}^2\right)\right]$$

And so we can write

$$D(x) = K + \sum_{i<s\neq t\le u}\beta_s\beta_t(1-\rho_{si}^2)(1-\rho_{ti}^2)\frac{\rho_s\rho_t}{1-\rho_i^2}$$
$$+ \sum_{u+1\le s\neq t<j}\beta_s\beta_t(1-\rho_{sj}^2)(1-\rho_{tj}^2)\frac{\rho_s\rho_t}{1-\rho_j^2}$$
$$= K + \frac{1}{1-\rho_i^2}\left(\sum_{s\le u}\beta_s(1-\rho_{si}^2)\rho_s\right)^2 + \frac{1}{1-\rho_j^2}\left(\sum_{u+1\le s}\beta_s(1-\rho_{sj}^2)\rho_s\right)^2$$

where the constant $K$ depends only on the QTLs which are strictly between $z_i$ and $z_j$, and does not depend on the location of the locus $x$.

Setting

$$R = \left(\sum_{i<s\le u}\beta_s(1-\rho_{si}^2)\rho_{su}\right)^2\frac{1}{1-\rho_{iu}^2}$$
$$Q = \left(\sum_{u+1\le s<j}\beta_s(1-\rho_{sj}^2)\rho_{s,u+1}\right)^2\frac{1}{1-\rho_{j,u+1}^2}$$

we obtain

$$D(x) = K + R\rho_u^2 \left( \frac{1-\rho_{i,u}^2}{1-\rho_i^2} \right) + Q\rho_{u+1}^2 \left( \frac{1-\rho_{j,u+1}^2}{1-\rho_j^2} \right)$$

The end is near. Without loss of generality, assume that $D(z_u) \geq D(z_{u+1})$. Then

$$R + Q\rho_{u,u+1}^2 \left( \frac{1-\rho_{j,u+1}^2}{1-\rho_{j,u}^2} \right) \geq R\rho_{u,u+1}^2 \left( \frac{1-\rho_{i,u}^2}{1-\rho_{i,u+1}^2} \right) + Q$$

which gives

$$R \geq Q \left( \frac{1-\rho_{i,u+1}^2}{1-\rho_{j,u}^2} \right)$$

Now suppose that $D(z_u) \leq D(x)$ This gives

$$R \leq Q \left[ \frac{1-\rho_{j,u+1}^2}{1-\rho_{u,j}^2} \right] \left[ \frac{\rho_{u+1}^2(1-\rho_i^2)}{1-\rho_j^2} \right]$$

But this second inequality cannot be true, since

$$\left[ \frac{1-\rho_{j,u+1}^2}{1-\rho_{u,j}^2} \right] \left[ \frac{\rho_{u+1}^2(1-\rho_i^2)}{1-\rho_j^2} \right] < \left( \frac{1-\rho_{i,u+1}^2}{1-\rho_{j,u}^2} \right)$$

And so, we find that if $D(z_u) \geq D(z_{u+1})$, then $D(z_u) > D(x)$.

Further, $D(x)$ is convex in the region between $z_u$ and $z_{u+1}$. The calculations again involve a messy bit of algebra. After writing $D(x)$ as a function of the genetic distance between $z_u$ and $x$, differentiate twice, and see that the second derivative is non-negative.

Note that if there are no QTLs located between $z_{u+1}$ and $z_j$, $Q = 0$, and so $D(x)$ is easily seen to be maximized at $z_u$. If there are no QTLs at all between $z_i$ and $z_j$, $D(x) = 0$ for any location $x$ between $z_i$ and $z_j$.

Also note that, by setting $\rho_j$, $\rho_{j,u}$ and $\rho_{j,u+1} = 0$, the above covers the case in which there is no $z_j$ to the right of $x$ which has already been included in the model for $y$. Setting $\rho_i$, $\rho_{i,u}$ and $\rho_{i,u+1} = 0$ as well gives the case in which no markers on the chromosome have yet been included in the model for $y$.

And so, to wrap up our argument, we see that if there are no QTLs on a chromosome, the change in RSS associated with adding in a marker converges to 0 almost surely. If there are QTLs on the chromosome, then markers located at QTLs will, in the limit, give a greater change in RSS than any other loci. Once all the QTLs on a chromosome have entered the model, the change in RSS corresponding to adding any other marker has a limiting value of 0.

Let $I_n(s)$ denote the set of markers obtained by using forward selection up to $s$ variables, for a sample size of $n$. Consider a situation with a finite set of markers, and with $p$ QTLs all located exactly at marker loci. Then with probability 1, there exists an $N$ such that for all $n > N$, the set $I_n(p)$ is exactly the set of $p$ QTLs. Combined with the result in An and Gu (1985), that minimizing a BIC-type criterion over all possible models gives a consistent estimate of the true model, this shows that using forward selection with a BIC-type criterion is also a consistent procedure.