Quantitative Trait Loci in Inbred Lines

Jansen, R.C.

Published in:
EPRINTS-BOOK-TITLE

Publication date:
2001

21

Quantitative Trait Loci in Inbred Lines

R.C. Jansen

Wageningen UR Centre for Biometry, Plant Research International, The Netherlands

Quantitative traits result from the influence of multiple genes (quantitative trait loci) and environmental factors. Detecting and mapping the individual genes underlying such 'complex' traits is a difficult task. Fortunately, populations obtained from crosses between inbred lines are relatively ideal for this – at least far more ideal than livestock and human populations – and true multigenic models are now available and have been applied successfully. In this chapter we will introduce the reader to statistical tools for segregation analysis and genetic mapping with the aid of molecular markers.

21.1 INTRODUCTION

21.1.1 Mendelian Factors and Quantitative Traits

The breeding experiments of Mendel in the late nineteenth century led to the foundation of modern genetics. He crossed white- and purple-flowering pea plants, and proposed a simple theory to explain the observed frequencies of white- and purple-flowering plants from generation to generation. With today's knowledge, we can conclude that *one* gene with *qualitative* effect encoded the flower colour studied by Mendel. If many genes and/or genes with quantitative effects encode a trait of interest, then the effects of the individual genes can hardly be distinguished. Therefore dissection of truly quantitative variation into the underlying 'Mendelian factors' is difficult on the basis of phenotypic information only. In this chapter we will see what modern statistical and molecular tools we have in our toolbox for studying quantitative or complex traits in inbred line crosses. These tools are of prime importance for plant breeding, livestock improvement and medical research using model organisms, because application of advanced tools offers great opportunities for understanding and manipulation of complex biological processes.

Handbook of Statistical Genetics. Edited by D.J. Balding et al.

Genes underlying quantitative or complex traits are commonly called *quantitative trait loci* (hereafter QTLs).

We start with a very limited description of inbred line genetics and breeding – all you need to know about inbred lines, seen through a statistician's spectacles – and the reader can generalize if he or she wishes. The next sections deal with what we think are the essentials for proper statistical modelling and analysis, first for so-called segregation analysis and second for genetic mapping of QTLs with the aid of genetic markers. The reader will see that similar statistical models and algorithms will be used for segregation analysis as for QTL mapping with the aid of molecular markers. We will introduce much of the relevant statistical machinery in the section on the simpler 'marker-free' case of segregation analysis.

There is a large amount of literature, and within the scope of this chapter it is hardly possible to refer to all statistical methods and applications published. The author has made his own selection, but hopes that this text, the bibliography and references give the reader sufficient entries to find his own favourite way in the world of QTLs in inbred lines. We sometimes get the feeling that the statistical theory of genetics is presented with a semblance of mystery, as if it were very special and complex. We believe there is no reason for this in the case of QTLs in inbred lines – almost all the theory is closely related to standard analysis of variance and regression analysis. Nevertheless, we will make numerous cautionary remarks – standard methods can still be greatly abused.

21.1.2 The Genetics of Inbred Lines

In order to understand the genetics of inbred lines, we first need to define what one commonly understands by the terms 'inbred' and 'line', and then we can discuss what 'sort of genetics' is involved. Here, we consider only inbred line crosses of diploid organisms. The reader may generalize to similar theoretical results for polyploid organisms and biparental crosses between outbreeding lines. In all cases we start with two parental lines, say P_1 and P_2, and intercross them in order to generate new genetic combinations.

In the genetics literature a *line* is a set of genetically related individuals, which are maintained under one and the same breed identification and/or commercial name. Chromosomes occur in homologous pairs, one originating from the mother, the other from the father. A line P_1 is *homozygous* if at any given gene the maternal and paternal states or *alleles* are identical, say a_1a_1. Thus, any offspring from crossing or mating within the P_1-pool will have the same P_1-genotype. The genotype of P_2 is denoted by a_2a_2. How do you get a line to become homozygous? For instance, in plants by following a single-seed-descent strategy over multiple generations: each generation you randomly take, grow and self a single individual in order to generate the seeds of the next generation of the line. By expectation, the number of heterozygous loci, a_1a_2, will be halved each generation. After eight to ten generations one can stop as the breeding process has led to an (almost) homozygous 'inbred line'. Of course, this scheme is typical of plants and cannot be used in animals. For model animals (mice, rats, hamsters) and livestock animals (chickens) simple designs of brother–sister mating can be used. Various plant species (apple, strawberry, pine) and most animal species (cattle, horses) can hardly be manipulated this way, because it would take too much time and money, because the organism would suffer from severe inbreeding depression, and, last but not least, because it would be immoral (humans). Although we focus here on biparental crosses between divergent inbred lines, most of the theory also applies to the case of a biparental cross

between divergent outbreeding lines. In the latter case, there can be up to four different alleles per locus. We present theory for the simpler two-allele situation and thus leave the extension to four alleles to the reader.

There follows a short description of a number of mating designs. All start with a cross between two divergent inbred lines P_1 and P_2, which generates heterozygous filial F_1-offspring ($a_1a_1 \times a_2a_2 \rightarrow a_1a_2$). In the backcross (BC) design, the F_1 is (back)crossed to one of the parents, say female F_1 to male P_1 ($a_1a_2 \times a_1a_1 \rightarrow a_1a_2$ and a_1a_1 for each locus). In the doubled haploids (DH) design, either male or female gametes of the F_1 are artificially doubled by some kind of treatment, giving rise to homozygous recombinant offspring in one step ($a_1a_2 \rightarrow a_1a_1$ and a_2a_2 per locus). In the BC and DH design one can only trace segregation and recombination events in either the male or the female F_1 gamete. In the filial F_2 design, the F_1 is selfed or two F_1 individuals are crossed, in order to generate offspring, resulting from recombination events in both male and female gamete production ($a_1a_2 \times a_1a_2 \rightarrow a_1a_1, a_1a_2$ and a_2a_2 in the ratio $1:2:1$ per locus). Finally, in the recombinant inbred line (RIL) design one first generates the F_2 progeny and then each F_2 enters individually a single-seed-descent inbreeding programme ($a_1a_2 \times a_1a_2 \rightarrow a_1a_1$ or a_2a_2 per locus). From here on we us the notation A, H and B for a_1a_1, a_1a_2, and a_2a_2, respectively.

21.1.3 Phenotype, Genotype and Environment

The basic model of quantitative genetics is

$$P = G + E,$$

or, in words, the phenotype (trait) is the sum of genetic and environmental factors. By taking two divergent lines, we have a high chance that the two lines have different states (alleles) at all the genes underlying the traits of interest. Unfortunately, the trait difference between the two parents reflects the total effect of the their genes and not their individual gene effects. In statistical terms, the effects of the genetic factors are *confounded*, the factors are *aliased* or *collinear*. The genetic reproduction mechanism will yield us offspring with new allele combinations, generated by independent segregation of different chromosomes and by recombination within chromosomes. Therefore in the offspring, genes on different chromosomes are independent stochastic variables (*unlinked* genes). In contrast, the genes on the same chromosome will show statistical dependence or *linkage association*, although this may be negligible if they are far apart, and we speak of *linked* genes. Unlinked genes will become orthogonal factors (in infinite populations) and closely linked genes become factors with a high degree of collinearity. Any recombination between two flanking genes increases our chance of dissecting their effects. The distance between two loci is usually reported in units of *morgans* (M) or *centimorgans* (cM) using Haldane's mapping function. According to Haldane, a recombination frequency of r between two loci corresponds to a map distance of m morgans, with

$$m = -\tfrac{1}{2}\ln(1-2r);$$

conversely, a distance of m morgans corresponds to a recombination frequency r of

$$r = \tfrac{1}{2}(1-\exp(-2m)).$$

For small distances the relation is more or less proportional ($m = 0.01$ M $\sim r = 0.01$, in which case one in 100 individuals of a DH population will show up as recombinant).

21.2 SEGREGATION ANALYSIS

21.2.1 Visualization of Quantitative Variation in a Histogram

Our first step, once we have collected the phenotypic data, will most probably be to rank the data and draw a histogram to visualize the results for the trait under study. This is 'just' descriptive statistics, but it is an important step, as conclusions drawn from descriptive statistics often have more practical impact than those from inferential statistics.

Figure 21.1 shows typical histograms. What can we learn from these pictures? And, are there any pitfalls? What we know from lessons in genetics is that homologous chromosomes segregate randomly, creating opportunity for 2^n different egg or sperm cells per meiosis and an equal amount of possible homozygous genotypes for an organism with n chromosomes pairs. Nature is even cleverer than that, and devised the cross-over mechanism as the result of which the total number of possible genotypes is much bigger even than 2^n so that usually no two individuals in the population have the same genotype. What we see in Figure 21.1 is a 'mixture' of as many different genotypes as there are individuals.

Figure 21.1(a) clearly shows a bimodal distribution for a segregating DH population. If this is the case for our trait, we will probably be excited! We recognize that the mixture of genotypes falls into two different groups, which can be labelled '*A*' and '*B*', respectively. Of course, it is likely that the two labels correspond to the two *genotypes* *A* and *B* at a yet unmapped major gene. One might say that the action of this gene is semi-qualitative or semi-quantitative, that is, somewhere between pure qualitative, as with Mendel's pea genes, and pure quantitative, as with many complex traits such as body weight or crop yield. The action of our major gene may or may not be modified by other genes of minor effect, but the information in the histogram is too limited for us to come to conclusions on this aspect. In some cases trait values of a number of individuals for each of the parents are available. If a parent line is homozygous, all individuals of that line are genetically identical and any variation observed among them must therefore be

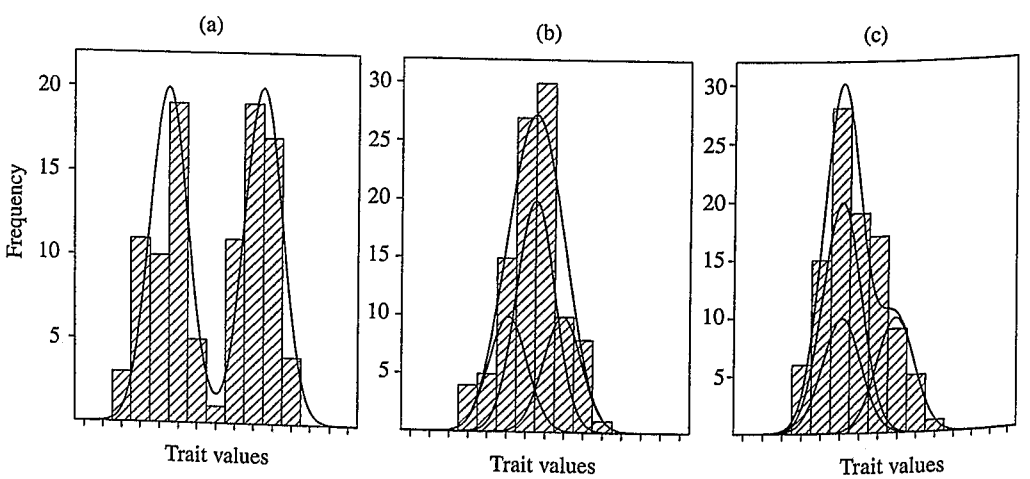

Figure 21.1 Mixture distributions plotted on top of the histograms (a) for a DH population, (b)–(c), for F_2 populations. The component distributions for each of the genotypes *A*, *H* and B are also plotted in (b) and (c).

environmental in origin. This would make it possible to compare within-parents variation, which is supposed to be environmental and not genetic, to within-*A* and within-*B* variation, which is the sum of environmental and possibly residual genetic variation.

Figure 21.1(b) shows a histogram with a unimodal distribution for a segregating F_2 population. It is not unlikely that our first reaction will be 'a real pity, my trait is complex, probably encoded by many genes of small action'. If many genes encode our trait, then we would expect a 'normal' distribution, as shown in Figure 21.1(b). Again, comparing with parental data, we can get an idea of the amount of genetic variation relative to environmental variation. However, our questions 'how many genes of what sizes of effects', will not be answered unless we invoke the aid of molecular techniques – this will be the topic of the next sections. Are there any potential pitfalls? The reader may feel misled, but the histogram in Figure 21.1(b) actually represents the case of a single additive major gene explaining half of the total variation! The segregating population consists of a mixture of genotypes *A*, *H* (heterozygote) and *B* in the ratio 1 : 2 : 1. Thus even if we observe a unimodal distribution, we may still hope that the underlying genetics is not too complex.

Figure 21.1(c) shows another histogram with a unimodal distribution for a segregating F_2 population, but this time the distribution is skewed. This may give us a hint of dominant major gene action! Suppose individuals with genotypes *A* and *H* group together (i.e. have the same mean value) and underlie the main body of the distribution. The other individuals, with genotype *B*, underlie the 'shoulder' of the distribution. There is little by way of a proof in the picture, but we may think there is a lot of evidence for our hypothesis of a single dominant gene. Again, there is a chance of a serious pitfall. We may have made the implicit assumption that the error distribution or 'component distribution' is symmetric or even 'normal'. Many traits have a natural lower bound (often zero), and variances often increase with the mean. In reality our histogram was generated under a (polygenic) model with a skewed residual error distribution. Thus, seeing a skewed distribution may lead us to believe erroneously that there is a dominant major gene. Some researchers are aware of this pitfall and transform their data, for instance they take the natural logarithm of all trait values, thereby changing the scale of the *x*-axis to a logarithmic scale on which the distribution looks more symmetrical. Is this a good procedure? Not necessarily, because there is the danger of throwing out the baby with the bathwater: we may loose power for detecting dominant gene action. However, there may sometimes be good biological reasons to log-transform. To check the type of distribution one can best look at parental (non-mixture) data and use standard statistical techniques to help find an optimal data transformation – see Atkinson (1985), Jansen and Den Nijs (1993) and Jansen et al. (1993) for illustrative examples in the case of mixture models. Choice of mixing distribution becomes more critical when the model becomes more complex (e.g. if models include gene interactions). It is well known, for instance, that significant interactions can arise as statistical artefacts: the interactions try to compensate for wrong model assumptions. Still, distribution checks are only occasionally reported in genetic analysis.

Quantitative data are not always continuous data. Typically, quantification of disease scores may lead to count data (e.g. number of spots on a leaf), categorical or ordinal data (e.g. a disease score 1–5) or proportions (e.g. number of affected individuals per DH line being tested on multiple seedlings). In many cases the distribution is close to normal after appropriate data transformation (e.g. log, square root, or probit). The type of transformation should preferably be chosen on the basis of non-mixture data. See McCullagh and Nelder (1980) for generalized linear models for many types of distribution.

Finally we refer to Allard (1999) for several illustrative graphs (his Figures 8-1 and 8-2), showing qualitative and quantitative variation resulting from segregation for one, two, or more genes with or without dominance and/or in the presence or absence of environmental noise.

21.2.2 Plotting Mixture Distributions on Top of the Histogram

Histograms may show bimodality or skewness and as such they can indicate segregation of major genes. Still we can not unambiguously assign individuals to genotypes. In statistical terms, a QTL is a latent unobserved categorical variable. Let us look again at Figure 21.1. How can we model, fit and plot the mixture distributions on top of the histogram?

Let us start with a relatively simple model for Figure 21.1(b): a model with a single QTL with *additive* allele effect and normal error in an F_2. Let y_i denote the trait value of the ith individual. We do not know its QTL genotype. There are three possible genotypes, each with its own probability of occurrence and component probability density function (PDF):

			P	PDF
1	y_i	A	0.25	$\phi(y_i; \mu_A, \sigma^2)$
2	y_i	H	0.50	$\phi(y_i; \mu_H, \sigma^2)$
3	y_i	B	0.25	$\phi(y_i; \mu_B, \sigma^2)$

Here

$$\phi(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y-\mu)^2/2\sigma^2)$$

is the PDF of the normal (Gaussian) distribution with mean μ and variance σ^2, and

$$\mu_H = \tfrac{1}{2}(\mu_A + \mu_B),$$

because we assumed additivity of allele effects. Taking the weighted sum over the three components, we get the mixture PDF

$$f_{\text{mix}}(y_i) = \frac{1}{4}\phi(y_i; \mu_A, \sigma^2) + \frac{1}{2}\phi\left(y_i; \frac{\mu_A + \mu_B}{2}, \sigma^2\right) + \frac{1}{4}\phi(y_i; \mu_B, \sigma^2).$$

The simultaneous mixture likelihood of observations $y_1, y_2, \ldots, y_n$ is the product of the individual likelihood contributions:

$$L = \prod_{i=1}^{n} f_{\text{mix}}(y_i).$$

In order to shorten the notation we will often use $\phi_A(\cdot)$, $\phi_H(\cdot)$ and $\phi_B(\cdot)$ to denote the distributions for genotypes A, H and B, respectively.

Suppose – for the moment – that we know the parameter values. How do we show the fit of the model to the data? That is fairly easy: calculate $f_{\text{mix}}(\cdot)$ for values ranging from the minimum to the maximum trait value, multiply by the number of individuals n, and by the width w of the groups of the histogram to account for scaling, and finally plot this on top of the histogram. It will become a little more complicated if we want to show the histogram on the original scale of measurement, while we think that it is more appropriate to fit the mixture of normal distributions on a different scale. Suppose that we fit the normal mixture model to the transformed trait values $T(y_i)$, that is, on a

different scale than the original scale of measurement. The protocol is now as follows: calculate $f_{\text{mix}}(T(\cdot))$ for values ranging from the minimum to the maximum trait value on the original scale of measurement, multiply by the Jacobian $T'(\cdot)$ to calculate the likelihood on the original scale of measurement, next multiply by n and by w, and finally plot this on top of the histogram. The Jacobian $T'(\cdot)$ is the first-order derivative of the transformation function – for example, if $T(y) = \ln(y)$, then $T'(y) = 1/y$.

21.2.3 Fitting Mixture Distributions

What remains is the question of how we estimate the parameters of the finite (normal) mixture distribution. The standard maximum likelihood approach consists of finding the parameter values which maximize the simultaneous (log-)likelihood. To achieve this we can take the first-order derivatives of the log-likelihood l and set these to zero:

$$0 = \frac{\partial}{\partial\theta} l = \frac{\partial}{\partial\theta} \log\left(\prod_{i=1}^{n} f_{\text{mix}}(y_i)\right) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_{\text{mix}}(y_i).$$

We continue with our example of a single QTL with additive effect in an F_2. Now note that

$$\begin{aligned}
\frac{\partial}{\partial\theta} \log f_{\text{mix}}(y_i) &= \frac{1}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} [0.25\phi_A(y_i) + 0.50\phi_H(y_i) + 0.25\phi_B(y_i)] \\
&= \frac{0.25}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \phi_A(y_i) + \frac{0.50}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \phi_H(y_i) + \frac{0.25}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \phi_B(y_i) \\
&= \frac{0.25\phi_A(y_i)}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \log \phi_A(y_i) + \frac{0.50\phi_H(y_i)}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \log \phi_H(y_i) \\
&\quad + \frac{0.25\phi_B(y_i)}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \log \phi_B(y_i) \\
&= P(A|y_i) \frac{\partial}{\partial\theta} \log \phi_A(y_i) + P(H|y_i) \frac{\partial}{\partial\theta} \log \phi_H(y_i) + P(B|y_i) \frac{\partial}{\partial\theta} \log \phi_B(y_i),
\end{aligned}$$

which you can recognize as a sum of weighted 'normal' likelihood contributions, where the weights are conditional probabilities of the genotype given the observed phenotype $\text{P}(A|y)$, $\text{P}(H|y)$ and $\text{P}(B|y)$, which depend on the unknown θ (to shorten the notation we write $\text{P}(\cdot)$ instead of $P_\theta(\cdot)$). Unfortunately the likelihood equations cannot be solved analytically. But there is a simple – thus popular – algorithm, called the expectation-maximization algorithm – for short the EM algorithm. Dempster et al. (1977) considered the mixture problem as one of many examples in which data are incomplete. They interpreted mixture data as incomplete data by regarding an observation on the mixture as missing its component of origin. See their Section 4.3 on finite mixtures. In our context the information on QTL genotype is missing (with three components A, H, and B). The basic idea of the iterative EM algorithm is to replace the incomplete observation y_i by its three complete observations (y_i, A), (y_i, H) and (y_i, B), weighting the three complete observations by specified or updated (conditional) probabilities. Iteration consists of two steps:

(E-step) Specify or update weights $\text{P}(A|y_i)$, $\text{P}(H|y_i)$, and $\text{P}(B|y_i)$.

(M-step) Update estimates of μ_A, μ_B, and σ^2.

In the E-step, conditional probabilities are calculated by using the current parameter estimates. The M-step consists of weighted regression analysis on the triplicate data set, which can be done with most statistical packages, and which requires no more than a routine for weighted least squares. The explicit solution of the M-step can be written as

$$\hat{\mu}_A = \frac{\sum_{i=1}^{n}\left[\mathrm{P}(A|y_i)y_i + \mathrm{P}(H|y_i)\frac{1}{2}y_i\right]}{\sum_{i=1}^{n}\left[\mathrm{P}(A|y_i) + \mathrm{P}(H|y_i)\frac{1}{2}\right]}, \quad \hat{\mu}_B = \frac{\sum_{i=1}^{n}\left[\mathrm{P}(B|y_i)y_i + \mathrm{P}(H|y_i)\frac{1}{2}y_i\right]}{\sum_{i=1}^{n}\left[\mathrm{P}(B|y_i) + \mathrm{P}(H|y_i)\frac{1}{2}\right]},$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\left[\mathrm{P}(A|y_i)(y_i - \hat{\mu}_A)^2 + \mathrm{P}(H|y_i)\left(y_i - \frac{1}{2}(\hat{\mu}_A + \hat{\mu}_B)\right)^2 + \mathrm{P}(B|y_i)(y_i - \hat{\mu}_B)^2\right].$$

Setting parameters equal to (well-chosen) initial values conveniently starts the algorithm.

We will now use the above rather simple example to introduce the general concept of data completion (augmentation) and parameter estimation via iterative reweighted least squares. Hopefully this will help the reader to understand the more complicated cases that will appear in later sections. Let $\mathbf{y}^{(c)}$ denote the $1 \times 3n$ vector of augmented trait data $(y_1, y_1, y_1, y_2, y_2, y_2, \ldots, y_n, y_n, y_n)'$, where the superscript (c) is used for 'complete'. Furthermore, let $\boldsymbol{\beta}$ denote the $1 \times p$ vector of regression parameters, $\mathbf{X}^{(c)}$ the corresponding design matrix of size $3n \times p$, $\mathbf{e}^{(c)}$ the vector or residuals, and finally $\mathbf{W}^{(c)}$ is the diagonal matrix of conditional probabilities $(\mathrm{P}(A|y_1), \mathrm{P}(H|y_1), \mathrm{P}(B|y_1), \mathrm{P}(A|y_2), \mathrm{P}(H|y_2), \mathrm{P}(B|y_2) \ldots, \mathrm{P}(A|y_n), \mathrm{P}(H|y_n), \mathrm{P}(B|y_n))'$. Note that for the ith individual the model is

$$\begin{pmatrix} y_i \\ y_i \\ y_i \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix}, \quad \text{weighted by} \quad \begin{pmatrix} \mathrm{P}(A|y_i) \\ \mathrm{P}(H|y_i) \\ \mathrm{P}(B|y_i) \end{pmatrix},$$

or in matrix notation $\mathbf{y}^{(c)} = \mathbf{X}^{(c)}\boldsymbol{\beta} + \mathbf{e}^{(c)}$ with weight matrix $\mathbf{W}^{(c)}$. The M-step is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{X}^{(c)})^{-1}\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{y}^{(c)}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'\mathbf{W}^{(c)}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

21.2.4 Wanted: QTLs!

We want to discover all about the QTLs underlying quantitative variation for our trait(s) of interest. How many genes are involved? Where are they located on the chromosomes? What type of (inter)action do they show? It is clear from the Figure 21.1(b), that for truly quantitative variation the effects of the individual genes can hardly be distinguished. But there are clever ways to tackle this problem: use molecular markers, and this is the topic of the following sections! Thus, we now proceed to apply the mixture model tools in a new context where we have partial information, provided by molecular markers, on the underlying genotypes.

21.3 DISSECTING QUANTITATIVE VARIATION WITH THE AID OF MOLECULAR MARKERS

21.3.1 Molecular Markers

Since the early 1980s new ways to unravel complex traits have emerged (Botstein et al., 1980; Beckmann and Soller, 1983). The magic phrase is 'molecular marker'. Remember that a QTL is – in a statistical sense – a categorical variable whose values remain unobserved. A molecular marker is also a locus on the genome, but the genotype can be observed with molecular tools. From the statistical point of view it hardly matters how this molecular technique works; the only important point is that a marker is a categorical variable with observable state. Nowadays hundreds, not to say thousands, of molecular markers are available or will become available over the next few years, all with more or less *known* positions on the genome. In this chapter the focus will be on dense or ultra-dense marker maps. Which means that there is almost always a marker almost on top of any QTL. By 'on top' we mean that in the segregating population no recombination events between the given marker and QTL have appeared.

Markers are often just non-functional or selectively neutral sites. So why should it help to collect information about such loci? It will not help if our population is similar to a random mating population in *linkage equilibrium*, which means that genotypes at different loci are statistically independent. But in this chapter we deal with inbred line crosses for which linkage disequilibrium is at its maximum, i.e. there is maximum statistical correlation between the genotypes at linked loci. For example, between an observable marker that does not affect the phenotype and an unobservable QTL that does affect the trait. Say the homozygous parents P_1 and P_2 have marker-QTL genotype $m_1a_1//m_1a_1$ and $m_2a_2//m_2a_2$, respectively (chromosomes occur in pairs, and the symbol '//' separates the alleles of the first and second chromosome of the pair; the marker has alleles m_1 and m_2; the QTL has alleles a_1 and a_2). The F_1 genotype is $m_1a_1//m_2a_2$. It will produce a mixture of non-recombinant gametes (m_1a_1 or m_2a_2) with probability $1-r$, and recombinant gametes (m_1a_2 or m_2a_1) with probability r. Figure 21.2 is based on the same data as shown in Figure 21.1(a), and shows the effect of tight linkage between marker and QTL on the mixture distribution ($r = 0.01$). There would be weak or no association if the marker and QTL were far apart. Thus, if the marker is close to or on top of a QTL, then the two marker genotype groups (say, A for m_1m_1 and B for m_2m_2) will clearly show different means for the trait. The basic and simple idea is to reverse this statement – if the two marker classes clearly show different means, then there is evidence for 'QTL activity' in the neighbourhood. Note that we have made the reverse statement a little less strong, that is, we cannot claim that there is only one *isolated* QTL involved – there is much more to say about this, but we will postpone this to one of the later sections of this chapter.

Whether two marker classes indeed show different means (or not) can be tested with standard analysis of variance (ANOVA) and regression tools. Suppose that we fit a single QTL at a fully informative marker in an F_2 population. The one-way classification model reads $y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ for the jth individual of the ith genotype ($i = 1, 2, 3$; A, H and B at the marker). This model can easily be extended to two or more marker factors (Cowen 1989; Stam 1991). Assuming that all scores of marker factors are known – that is, their data are complete – the general model for regression on multiple markers is

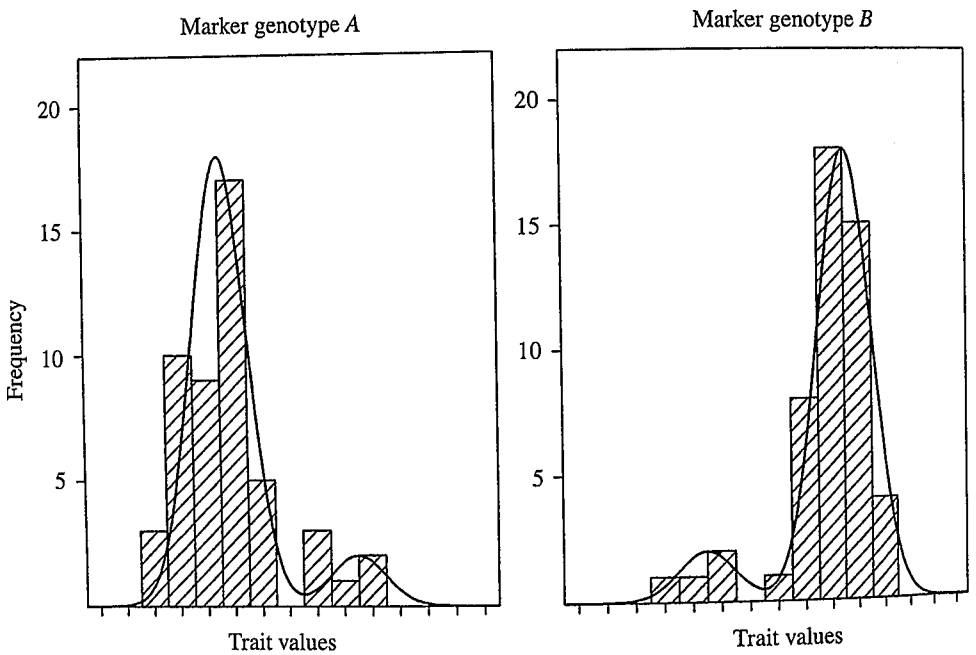

Figure 21.2 Mixture distributions plotted on top of the histograms per marker genotype for the same data as shown in Figure 21.1(a). The marker is close to a QTL, but a few recombinations have occurred. Therefore, the distribution given the marker genotype is a mixture over recombinant (probability r) and non-recombinant individuals (probability $1 - r$).

$\mathbf{y}^{(c)} = \mathbf{X}^{(c)}\boldsymbol{\beta} + \mathbf{e}^{(c)}$. A characteristic of ANOVA and regression models is that they often are over-parameterized, containing more parameters than needed to represent the effects. Usually, setting the sum of allele effects to zero or, equivalently, working in terms of main effect and allele substitution effects compensates for this. The parameters are commonly estimated by the least-squares method, that is,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{(c)\prime}\mathbf{X}^{(c)})^{-1}\mathbf{X}^{(c)\prime}\mathbf{y}^{(c)}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

Note that in the latter formula one divides by $n - p$, the degrees of freedom (df) of the residual error, and not by n; n is the population size, p is the number of regression parameters in $\boldsymbol{\beta}$.

21.3.2 Mixture Models

What happens if some marker scores are missing? ANOVA and regression models do not allow for missing values in the explanatory variables. We could eliminate any individual with one or more missing marker scores and then perform an analysis on the remainder of the data set. But in real experiments most individuals are missing at least some marker scores, so this approach does not seem to be very attractive. Of course, we can think of filling in the gaps *prior to ANOVA analysis* by using the map information. How does

this work? Look at the markers flanking the marker with a missing score. If the markers are close together, we can be pretty sure how to fill in the score and then we can apply ANOVA and regression! This is an *ad hoc* procedure, which works well for ultra-dense maps, but there are more sophisticated methods available for sparse maps. As we have seen in Section 21.2, data completion can be integrated with analysis of quantitative variance – and the type of models we need are mixture models.

We should not be surprised if up to 5% of the marker scores are missing in a real experiment. Often some markers or some individuals are difficult to score. This can be caused by various technical failures such as a given individual's bad DNA sample or problems with the fingerprint image. The statistician assumes that the scores are missing at random, or at least that there is no tendency that one score, say A, is relatively more frequently missing than B or H. Another type of incomplete data occurs when one or more markers are *dominant* and scored in, for instance, an F_2 population. Such a dominant marker still has three different states A, H and B, like a codominant marker. However, for technical reasons either A and H cannot be discriminated, which is often labelled as 'not-B' or just D, or H and B cannot be discriminated, labelled as 'not-A' or just C. A dominant marker is, in statistical terms, an observable categorical variable with two observable states. We can therefore use such a variable straightforwardly in ANOVA. On the other hand, we may want to complete the marker data by again using information from neighbouring markers. For instance, a 'not-A' observation means that the genotype is B or H, and the first one is more likely if we know that an adjacent marker has score B. How successfully can we complete data? This can be quantified by what one calls the *multilocus information content*; this will be defined later. We finish by remarking that the distinction between dominant and codominant markers is not always so absolute. Some types of marker may yield A, H, B, C and D, that is there are five categories (e.g. AFLP© markers).

Currently, marker maps are still not ultra-dense, and this brings us to the third type of incomplete data. Suppose that we have complete marker data, but that our marker map is sparse, that is the markers are spread coarsely over the genome. Our data will show many recombinants between any pair of flanking markers and, if there is a QTL in between, then it is likely that none of the markers is 'on top' of the QTL. In other words, a flanking marker can have the same genotype as the QTL, but not for all individuals at the same time. Therefore, ANOVA on a flanking marker is no longer identical to ANOVA on the QTL. ANOVA may still work relatively well, but the results are biased or less powerful, because we have built in errors by ignoring one or more recombinations between the marker and the QTL. Does statistics offer a solution to this problem? Any locus with incomplete data can be added to the model, whether it is a marker (as discussed above) or a QTL (which is a locus with all its scores missing). The statistical trick of data completion will work in any case, and it will help us to exploit the full multilocus information content. Basically, there is no obstacle to using multiple-QTL models. Over the past decade most theoretical papers on QTL mapping have been devoted to the sparse map situation. The older papers dealt with single-QTL mixture models only (see Weller, 1986; Jensen, 1989; Lander and Botstein, 1989; Simpson, 1989), methods that became known as 'interval mapping'. Jansen (1992) developed a general approach for fitting multiple-QTL models, an approach that became later known as 'MQM', and was elaborated in a number of papers (Jansen, 1993b; 1994b; Jansen and Stam, 1994). Zeng (1994) published a similar strategy and called it 'composite interval mapping' (CIM).

Let us now work out an example. For convenience, but without loss of generality, we will zoom in on one F_2 individual with trait value y_i and five loci with scores

$$A\,A\,D\,U\,A,$$

where D means 'not-B' and U means 'genotype unknown'. We suppose that the loci 1, 2, 3 and 5 are markers, and that the fourth locus is a QTL, which is located at a map position within the interval between loci 3 and 5. All genotype scores of the QTL are unknown. The observed data can be completed for the missing locus information, giving rise to six different complete genotypes:

		1	2	3	**4**	5	Simultaneous genotype probability	Component density
1	y_i	A	A	A	**A**	A	$\frac{1}{4}(1-q)^2(1-r)^2(1-s)^2(1-t)^2$	$\phi_A(y_i)$
2	y_i	A	A	A	**H**	A	$\frac{1}{4}(1-q)^2(1-r)^2\,2(1-s)s(1-t)t$	$\phi_H(y_i)$
3	y_i	A	A	A	**B**	A	$\frac{1}{4}(1-q)^2(1-r)^2s^2t^2$	$\phi_B(y_i)$
4	y_i	A	A	H	**A**	A	$\frac{1}{4}(1-q)^2\,2(1-r)r(1-s)s(1-q)^2$	$\phi_A(y_i)$
5	y_i	A	A	H	**H**	A	$\frac{1}{4}(1-q)^2\,2(1-r)r((1-s)^2+s^2)(1-t)t$	$\phi_H(y_i)$
6	y_i	A	A	H	**B**	A	$\frac{1}{4}(1-q)^2\,2(1-r)r(1-s)st^2$	$\phi_B(y_i)$

Suppose you know the frequencies of recombination q, r, s and t between the loci. The multilocus genotype probabilities can be derived straightforwardly: $\frac{1}{4}$ for 'genotype A at the first locus, then $(1-q)^2$ for no recombination between A on the first locus and the A at the second locus, etc. For the moment, our model will include only one explanatory variable – the fourth locus is a QTL – and we can assign component densities $\phi_A(y_i)$, $\phi_H(y_i)$ and $\phi_B(y_i)$. By multiplying each of the six genotype probabilities with the corresponding component densities, and then summing these products over the six genotypes, we can see that the contribution of this individual to the *multilocus* or *multipoint* simultaneous likelihood is

$$\begin{aligned} f_{\text{mix}}(y_i, AADUA) = {} & \text{P}(AAAAA)\phi_A(y_i) + \text{P}(AAAHA)\phi_H(y_i) + P(AAABA)\phi_B(y_i) \\ & + \text{P}(AAHAA)\phi_A(y_i) + P(AAHHA)\phi_H(y_i) + P(AAHBA)\phi_B(y_i). \end{aligned}$$

The crux is that we complete the explanatory variable(s), here the fourth locus, so that we can perform an ANOVA-like analysis. There is no need to complete the data for the other non-explanatory loci, and we can rewrite the above expression as

$$\begin{aligned} f_{\text{mix}}(y_i, AADUA) = {} & [\text{P}(AAAAA) + \text{P}(AAHAA)]\phi_A(y_i) \\ & + [\text{P}(AAAHA) + \text{P}(AAHHA)]\phi_H(y_i) \\ & + [\text{P}(AAABA) + \text{P}(AAHBA)]\phi_B(y_i). \end{aligned}$$

The genotype probabilities can easily be calculated recursively. The regression model for the ith individual is

$$\begin{pmatrix} y_i \\ y_i \\ y_i \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix}, \text{ weighted by } \begin{pmatrix} \text{P}(A|y_i) \\ \text{P}(H|y_i) \\ \text{P}(B|y_i) \end{pmatrix},$$

in which

$$P(A|y_i) = \frac{[P(AAAAA) + P(AAHAA)]\phi_A(y_i)}{f_{mix}(y_i)},$$

and $P(H|y_i)$ and $P(B|y_i)$ are calculated analogously. It is easy to add extra QTLs to the model. For instance, we can insert a second QTL (a locus with only U scores) in the first marker interval. Doing so, the PDF of the ith individual is

$$f_{mix}(y_i, AUADUA) = \sum_{\substack{u=A,H,B \\ v=A,H,B}} [P(AuAAAvA) + P(AuAAHvA)]\phi_{uv}(y_i).$$

The general regression model, for all individuals together and for any number of markers and QTLs, reads $\mathbf{y}^{(c)} = \mathbf{X}^{(c)}\boldsymbol{\beta} + \mathbf{e}^{(c)}$, with weight matrix $\mathbf{W}^{(c)}$. The parameters can be conveniently estimated via the EM algorithm by iterating in two steps, similar to what we have seen for the marker-free case of segregation analysis:

(E-step) Specify or update weights P(multilocus genotype $|y_i$).

(M-step) Update estimates of $\boldsymbol{\beta}$ and σ^2.

In the E-step conditional multipoint genotype probabilities are calculated by using the current parameter estimates. The M-step is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{X}^{(c)})^{-1}\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{y}^{(c)}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}((\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}))'\mathbf{W}^{(c)}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

In our small example presented above, data augmentation for the given individual gave rise to six different complete genotypes. However, different individuals may have different amounts of missing genetic (QTL and marker) information. In this augmentation approach there is no restriction on multiplying any other individual with any other value (3, 6, 10, 100, etc.) within the same analysis. For instance, 'triplicate' the jth individual with observed data (y_j, $AAAUA$) to complete data {(y_j, $AAAAA$), (y_j, $AAAHA$), (y_j, $AAABA$)}.

Use of the EM algorithm (generally) yields maximum likelihood (ML) estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Jansen (1994b) proposed to use restricted ML, i.e. adjust the ML estimate $\hat{\sigma}^2$ for the df used. Divide by the df of residual error instead of by n, just as in ANOVA and regression analysis:

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'\mathbf{W}^{(c)}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

As we will see in a later section, this can lead to several appealing advantages during selection and inference, at least if multipoint linkage information is high.

We would like to emphasize that we model the *simultaneous* likelihood of the trait and of multiple markers and QTLs. Genotypes are *multilocus* and genotype probabilities, are *unconditional* (Jansen, 1992; 1993b; 1994b). This is in contrast to most literature on inbred line analysis, in which one usually calculates the likelihood of the trait *conditional*

on the two markers flanking the interval study (e.g. Lander and Botstein, 1989). If the flanking marker scores are missing or incomplete for a given individual, then the nearest informative marker in the same direction is taken. This 'conditional marker-QTL-marker approach' works well for simple single-QTL models, but is definitely less 'transparent' for multiple-QTL models. Only unconditional multilocus likelihoods can form the basis for direct comparison of the fit of several competing models.

How do we define multilocus information content at a given map location? In the ideal case, one of the conditional probabilities $P(A|y_i)$, $P(H|y_i)$, or $P(B|y_i)$ is 1 and the two others are 0. This happens only if the QTL is on top of a complete-data marker. In the worst case of no marker data, the three probabilities are 0.25, 0.50, and 0.25. One simple way to visualize the quality of the data works as follows. Take for each individual the most likely genotype and average the corresponding probabilities over all individuals. If the average is close to 1, then our data are excellent at the given map position. If it is close to 0.5, then the multimarker information is really poor. This is just our own definition of information content, and others exist. See, for instance, the definition, based on variances of conditional probabilities, in Spelman et al. (1996). Is it a major problem if the information content is substantially lower than 1? This has not been studied in much detail in the QTL literature, but it may not be much of a problem: Redner and Walker (1984) pointed out that small to moderate proportions of completely informative individuals are sufficient for good ML estimation in mixture problems.

21.3.3 Alternative Regression Mapping

Let us look again at our simple example in which the fourth locus was the QTL and

$$\begin{aligned} f_{\text{mix}}(y_i, AADUA) = {} & [\text{P}(AAAAA) + \text{P}(AAHAA)]\phi_A(y_i) \\ & + [\text{P}(AAAHA) + \text{P}(AAHHA)]\phi_H(y_i) \\ & + [\text{P}(AAABA) + \text{P}(AAHBA)]\phi_B(y_i). \end{aligned}$$

Note that $f_{\text{mix}}(y_i, AADUA) = f_{\text{mix}}(y_i|AADUA)\text{P}(AADUA)$. Conditional on observed genotype, the expected trait value of the ith individual, μ_i, is given by

$$\begin{aligned} \text{P}(AADUA)\mu_i = {} & [\text{P}(AAAAA) + \text{P}(AAHAA)]\mu_A \\ & + [\text{P}(AAAHA) + \text{P}(AAHHA)]\tfrac{1}{2}(\mu_A + \mu_B) \\ & + [\text{P}(AAABA) + \text{P}(AAHBA)]\mu_B \\ = {} & [\text{P}(AAAAA) + \text{P}(AAHAA) + \tfrac{1}{2}\text{P}(AAAHA) + \tfrac{1}{2}\text{P}(AAHHA)]\mu_A \\ & + [\text{P}(AAABA) + \text{P}(AAHBA) + \tfrac{1}{2}\text{P}(AAAHA) + \tfrac{1}{2}\text{P}(AAHHA)]\mu_B, \end{aligned}$$

which we can write as

$$\mu_i = p_1\mu_A + p_2\mu_B.$$

Haley and Knott (1992) and Martinez and Curnow (1992) proposed to use the regression model for the ith individual

$$y_i = [p_1\, p_2]\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \varepsilon_i,$$

and parameters can be estimated easily by the least-squares approach, if p_1 and p_2 are assumed to be known. Here we refer to this method as 'regression mapping'. In an ML framework, their approach would be equivalent to approximating the mixture density $f_{\text{mix}}(y_i)$ by a normal distribution $\phi(y_i; \mu_i, \tau^2)$. Figure 21.2 visualizes the worst case of a major QTL for which the approximation can only be poor. In many cases, however, this approach works reasonably well. Where the EM algorithm often requires five to seven iterations of iterative weighted least squares, the regression approach obviously only needs one step.

21.3.4 Highly Incomplete Marker Data

Let us now briefly discuss problems with *highly* incomplete genetic data (e.g. if many marker scores are missing or if we postulate multiple QTLs). The PDF is then a sum over many different candidate genotypes, in which case computation may become time consuming. In the above example, the complete genotypes *AAABA* (no. 3) and *AAHBA* (no. 6) have negligibly small probabilities compared to the most likely genotype *AAAAA*. We can set some threshold for simply ignoring relatively unlikely genotypes, to save us a lot of computer time in genuine multilocus problems. An alternative approach is to use so-called Markov Chain Monte Carlo (MCMC) sampling techniques. Such techniques have been developed for the more complex situation of outbred line crosses (e.g. Jansen, 1996). The basic MCMC idea is simple: if you cannot enumerate all possible genotypes, then just sample a representative set. We refer to our Bibliographic Notes (Section 21.5) and Hoeschele (Chapter 22, this volume) and Stephens (Chapter 8, this volume) for more details.

21.3.5 ANOVA and Regression Tests

Now suppose that we fit a single QTL at a fully informative marker in an F_2 population. As before, the one-way classification model reads $y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ for the jth individual of the ith genotype ($i = 1, 2, 3$; A, H and B). The ANOVA table is shown in Table 21.1. We refer to Soller et al. (1976) and Soller and Genizi (1978) for some of the early references.

The test statistic

$$F_{\text{QTL}} = \frac{\text{MS between}}{\text{MS within}}$$

has an F-distribution with 2 and $n-3$ df under the null hypothesis of no segregating QTL (i.e. a ratio of two independent χ^2 variables, $(\chi_2^2/2)/(\chi_{n-3}^2/(n-3))$). The ANOVA

Table 21.1 Analysis of variance table for a single QTL at a fully informative marker in an F_2 population.

Source	df	SS	MS	E(MS)
Between QTL genotypes	2	$\sum_{i=1}^{3} n_i(\bar{y}_i - \bar{y})$	s^2_{QTL}	$\sigma_e^2 + \frac{1}{2}\sum_{i=1}^{3} n_i\beta_i^2$
Within QTL genotypes	$\sum_{i=1}^{3}(n_i - 1) = n - 3$	$\sum_{i=1}^{3} n_i(y_{ij} - \bar{y}_i)^2$	s_e^2	σ_e^2
Total	$\left(\sum_{i=1}^{3} n_i\right) - 1 = n - 1$	$\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$		

model can easily be extended to two or more factors, but the interpretation of the table is now complicated by the possible collinearity of the putative QTLs. Each sum of squares represents the variation accounted for by the given QTL, having eliminated all the effects of the QTLs above it in the table, but ignoring any terms of QTLs below it (e.g. McCullagh and Nelder, 1989). Thus, in the case of strong collinearity the order of fitting terms in the model may affect the sum of squares value of a given QTL. Alternatively, we can tabulate sum of squares values representing the variation accounted for by a given QTL having eliminated all the effects of the other QTLs postulated in the model (no matter whether they are tabulated above or below the given QTL).

21.3.6 Maximum Likelihood Tests

Let us now look at the same one-way QTL classification, but use ML procedures for estimation of the parameters of the normal distributions involved. We still assume that we fit a single QTL at a fully informative marker. The *full* model (fitting the QTL) reads $y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ with residual variance σ^2_{full}, whereas the *reduced* model (omitting the QTL) simply reads $y_{ij} = \mu + \varepsilon_{ij}$ with residual variance $\sigma^2_{\text{reduced}}$. Suppose that the marker data are complete. Maximum likelihood estimates can be obtained via least-squares analysis as usual. Note that

$$\hat{\sigma}^2_{\text{full}} = \frac{\text{SS within}}{n}$$

and

$$\hat{\sigma}^2_{\text{reduced}} = \frac{\text{SS total}}{n}.$$

Then, the log-likelihood of the full model, l_{full}, is

$$l_{\text{full}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\text{full}}}} - \frac{(y_{ij} - x_i'\hat{\beta})^2}{2\hat{\sigma}^2_{\text{full}}} \right) = -\frac{1}{2} n \log(2\pi\hat{\sigma}^2_{\text{full}}) - \frac{1}{2} n;$$

and the log-likelihood of the reduced model, l_{reduced}, is

$$l_{\text{reduced}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\text{reduced}}}} - \frac{(y_{ij} - \hat{\mu})^2}{2\hat{\sigma}^2_{\text{reduced}}} \right) = -\frac{1}{2} n \log(2\pi\hat{\sigma}^2_{\text{reduced}}) - \frac{1}{2} n.$$

The likelihood ratio (LR) test statistic is

$$LR = 2 \log \frac{L_{\text{full}}}{L_{\text{reduced}}} = 2(l_{\text{full}} - l_{\text{reduced}}) = n \log \frac{\hat{\sigma}^2_{\text{reduced}}}{\hat{\sigma}^2_{\text{full}}} = n \log \frac{\text{SS total}}{\text{SS within}}$$

It is asymptotically χ^2-distributed with 2 df (Wilks, 1938). In general, in a combined test for p parameters, the LR test is χ^2-distributed with p df. The models can easily be extended to two or more factors (QTLs). In all cases one tests on the basis of LR between a full model and a reduced model.

Some readers may not be familiar with the calculation of LR test statistic values on a natural logarithm scale. In applied genetics literature one frequently uses the logarithm

with 10 as base, in which case one uses the notation *LOD* instead of *LR*. It is defined as

$$LOD = \log_{10} \frac{L_{\text{full}}}{L_{\text{reduced}}}$$

We can easily convert a score on one scale to a score on the other scale via

$$LR = 2 \times \log_e(10) \times LOD \approx 4.6 \times LOD.$$

For large n the F-test is approximately distributed as $\chi^2(2)/2$ under the null hypothesis of no QTL (Lynch and Walsh, 1998), in which case $2LR \sim F$. Thus, in general QTL detection via ANOVA/regression is not identical to QTL detection via ML – slightly different outcomes may be expected, even for complete data!

21.3.7 Analysis-of-deviance Tests

We will now describe a procedure which can be used for complete data and also for *incomplete* data, that is, all cases in which we do not fit a QTL at a fully informative marker (Jansen, 1994b). We refer to the textbook by McCullagh and Nelder (1989) on generalized linear models (GLMs) for more information.

In comparing a sequence of models, an unbiased estimate of the residual variance can be obtained from the full 'most complex' model. This estimate is used for all models in the sequence to make the comparison fair, between the *full* model and any *reduced* model and amongst reduced models. We can then generate an ANOVA-like table, which in the GLM literature is commonly called an *analysis-of-deviance table*, where the term *deviance* is used for the difference $2(l_{\text{full}} - l_{\text{reduced}})$. The procedure has three steps (Jansen, 1994b):

1. Calculate ML estimates $\hat{\beta}_{\text{full}}$ and $\hat{\sigma}^2_{\text{full}}$. Adjust $\hat{\sigma}^2_{\text{full}}$ for bias, i.e. divide by the residual df and not by n.
2. Calculate ML estimates $\hat{\beta}_{\text{reduced}}$ given 'known' residual error as estimated in the full model ($\hat{\sigma}^2_{\text{full}}$ in step 1).
3. Calculate the deviance (*LR* test statistic).

Let us look again at the simple one-way ANOVA for complete data. The log-likelihood of the *full* model (QTL model) reads, as before,

$$l_{\text{full}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\text{full}}}} - \frac{(y_{ij} - x_i'\hat{\beta})^2}{2\hat{\sigma}^2_{\text{full}}} \right).$$

The real difference comes for the log-likelihood of the *reduced* model (the no-QTL model), which now reads

$$l_{\text{reduced}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\text{full}}}} - \frac{(y_{ij} - \hat{\mu})^2}{2\hat{\sigma}^2_{\text{full}}} \right).$$

Next we calculate the deviance

$$deviance = 2(l_{\text{full}} - l_{\text{reduced}}) = \frac{\text{SS total} - \text{SS within}}{\hat{\sigma}^2_{\text{full}}} = 2 \times \frac{\text{MS between}}{\text{MS within}},$$

where the likelihood ratio test has an F-distribution with 2 and $n-3$ df. In general,

$$deviance = p \times \frac{\text{MS between}}{\text{MS within}} = pF,$$

where p is the difference between the parameter numbers in the two models being compared. The maximum achievable likelihood is

$$l_{\max} = -0.5n \log(2\pi\hat{\sigma}^2_{\text{full}}),$$

which we obtain if we use as many parameters as we have individuals. If our model contains all the important parameters, the deviance $2(l_{\max} - l_{\text{model}})$ will be close to the difference between the number of individuals, n, and the number of parameters in the model, p. One can use this as a measure for goodness of fit (McCullagh and Nelder, 1989).

We have described the analysis-of-deviance approach for a simple one-factor model, but this can easily be extended to two or more factors. The analysis-of-deviance approach is a unified approach, which can be applied to many types of data (McCullagh and Nelder, 1989). It can also be applied to incomplete data, provided that at least a proportion of the individuals are completely informative at the region under study (Jansen, 1994b). Of course, with rather incomplete information the fitting of many parameters becomes dangerous and therefore undesirable; one can then use the ML approach. We close this section with a remark on the use of non-segregating data. If we have data from parental or F_1 populations, we can estimate the environmental variance, $\hat{\sigma}^2_e$, in the non-segregating populations. It could be an option to use $\hat{\sigma}^2_e$ instead of $\hat{\sigma}^2_{\text{full}}$ during the selection procedure, or at least to define the maximum achievable likelihood as

$$l_{\max} = -0.5n \log(2\pi\hat{\sigma}^2_e).$$

21.3.8 How Many Parameters Can we Fit Safely?

To obtain a good estimate of the residual error variance in ANOVA and regression analysis, it is necessary to have at least 10 residual error degrees of freedom, and many statisticians would take 12–20 df as their preferred lower limit. Increasing the residual error df makes little difference to the significance threshold in F-tests. Say we have a DH population of 100 individuals and suppose our organism has 10 chromosomes of length 200 cM each (e.g. maize). We can then spend as many as 80 df on explanatory variables, for instance by postulating eight QTLs per chromosome equally spaced every ~20 cM.

How many parameters can we fit in the ML approach? In the case of large numbers of parameters, residual error variance can be severely underestimated and asymptotic relations such as the χ^2 approximations do not necessarily hold. Therefore the number of parameters should not be too large, preferably less than $2\sqrt{}$(number of observations) (Jansen and Stam, 1994). In a DH population of 100 individuals, we have df for fitting approximately $2\sqrt{100} = 20$ QTLs, considerably less than in the ANOVA and regression framework. A critic could well state that we are wasting our resources by not asking enough questions of our data. Fortunately, this serious disadvantage of the ML approach can be overcome by the general analysis-of-deviance approach described above, by which we can fit up to 80 QTLs (Jansen and Stam, 1994).

21.4 QTL DETECTION STRATEGIES

21.4.1 Model Selection and Genome Scan

Having read the previous sections you have almost all the statistical tools in hand for building the QTL model and for estimating its parameters. You may think the major hurdles have been overcome, and that all that remains – mapping of all QTLs – must be more or less straightforward from the statistical point of view. Here are some cautionary remarks. There are several different statistical techniques for mapping and they do not necessarily all lead to the same answer. Moreover, by no means should the results from statistical analysis be considered as a proof of how nature shaped our trait – any model is wrong because it is a simplification and the possibility of ending up with erroneous conclusions must be kept in mind. The more complex the model, the greater the dangers.

Suppose that we have formulated and fitted a number of different models; which one should we adhere to? Probably the model with a good fit to the data and the lowest number of parameters. In statistics this is called the principle of 'parsimony'. Unfortunately there is no general theory for testing one model against another, and as such no claim such as 'this is the significantly best model at a confidence level of 95%' is possible.

In the ideal case all genetic variance of the trait is explained by detected QTLs only. In practice, a number of QTLs may be missed (error type 2) and at the same time a number of false positives may occur, indicating QTLs at map positions (or regions) where actually no QTLs are present (error type 1). The actual balance between the cost of false positives and the benefit of detected QTLs depends on the aim of the experiment (e.g. map-based cloning or marker-assisted breeding). Nevertheless, one often strives to keep the probability of a type 1 error below 5%. At the same time one should minimize the probability of an error of type 2. Approaches for QTL mapping generally comprise one or both of the following two steps:

- *Model selection*: Compare different QTL models and select the best one(s).
- *Genome scan*: Plot QTL likelihood along the genome using the selected model.

In our opinion the first step is the more critical one, whereas the second step is merely a good way of visualizing QTL results along the genetic map.

In the model selection step we will try to relate the trait variable to one or more explanatory loci (possibly at marker positions) or 'factors'. Different sets of factors can be considered as competing statistical models and many statistical criteria and selection approaches can be used. It is important to note that in genetic mapping one has a major advantage over many other model selection applications. In most applications there is no theory explaining the fact that two or more factors are correlated. But markers and QTLs are known to be located on the genome map, and thus genetic linkage theory tells us that we can replace one factor in the model by another one from the same region without much changing the fit of the model. When two or more explanatory factors are strongly correlated to one other, it is difficult to disentangle their individual effects – this is known as multicollinearity – but this is less of a problem in QTL analysis and in some sense simplifies the difficult task of model selection.

In this next section we will review several major techniques in use for model selection and inference in QTL mapping: single-marker analysis; interval mapping (IM); composite

interval mapping; and multiple-QTL mapping (MQM). With our cautionary remarks in mind, one can safely apply the procedures as given, without being too concerned with the actual significance levels.

21.4.2 Single-marker Analysis and Interval Mapping

We have depicted QTL mapping as a problem of selection among many possible models. Despite the potential benefits of genuine multifactor approaches, its statistical complications lead some users to adopt single-factor ANOVA analyses. How does it work? Suppose we have our known map with markers covering part or all of the genome. We then calculate the QTL likelihood (F- or LR-score) at each marker position and plot it along the map. In statistical terms, we produce the likelihood profile along the map. Next we look at our plot and we think there is evidence for QTL activity in regions where the QTL likelihood peaks and exceeds a significance threshold. If the markers are sparse, we can also calculate QTL likelihood at any position within a marker interval to get a smooth QTL likelihood curve, by using IM (see our Figure 21.3 Lander and Botstein, 1989) or the regression mapping approach (Haley and Knott, 1992; Martinez and Curnow, 1992). In this section we will discuss several features of this single-factor type of analysis.

What can we do in order to avoid reporting too many false positives (i.e. non-existing QTLs)? The significance of an effect plays a dominant role in the genetics literature. The common rule is that reviewers will accept a QTL which is significant at a 95% experimentwise confidence level ($\alpha = 0.05$). We do multiple tests, and what we obviously need is therefore a chromosome-wide or genome-wide significance threshold. Test scores at linked markers are strongly correlated. No unifying analytical solutions for the genome-wide distribution of the test statistic in a general inbred line cross are known. Several authors developed (complex) formulae for specific cases (Lander and Botstein, 1989;

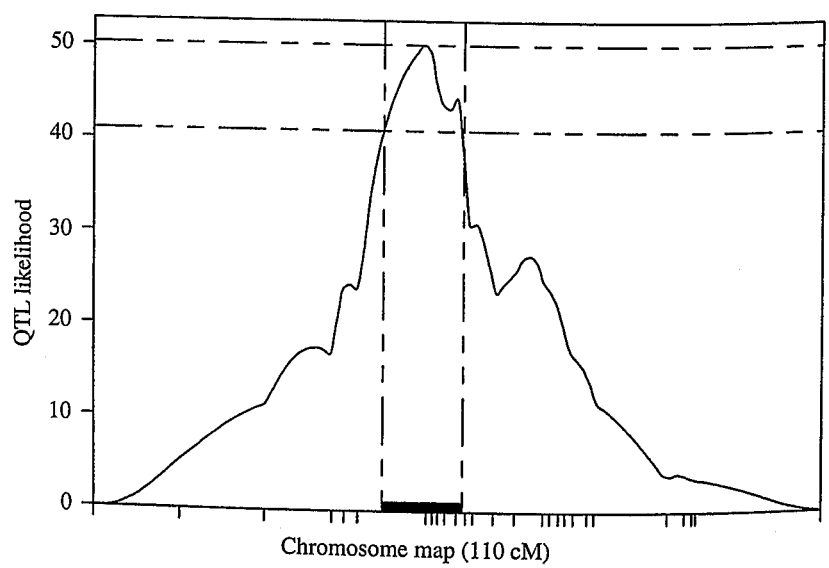

Figure 21.3 QTL likelihood profile obtained with interval mapping for the case of an isolated QTL. Marks below the x-axis indicate positions of markers along the chromosome under study. The bar indicates the 2 *LOD* support interval for the QTL. The genome-wide significance threshold is ~14.

Rebaï et al., 1994; Doerge and Rebaï, 1996), while others published tables resulting from extensive simulations (Van Ooijen, 1999). We can also use simple permutation strategies (Churchill and Doerge, 1994) or bootstrap strategies (Jansen, 1993b; 1994b; Visscher et al., 1996b). Suppose that we use one of these methods. Let us briefly look at permutation and the bootstrap. In both approaches new sets of data will be generated and analysed over many runs. In each run the 'old' marker data will be completed with 'new' trait data. The artificial data set is then analysed for QTLs and the maximum test score over the genome is calculated and stored. The entire procedure is repeated many (say, 1000–10 000) times in order to generate an empirical cumulative distribution for the test statistic. In the permutation approach, observed trait data are reshuffled over the individuals under the null hypothesis of no QTL, thereby breaking down any existing marker – QTL associations. Under the null hypothesis any of the trait data for any of the individuals could have come equally well from either of the marker or QTL classes. In the parametric bootstrap approach new trait data are generated from a standard normal distribution postulating no QTL.

Suppose that we have detected a 'genuine' QTL. If we replicated the experiment, the gene's location would of course remain identical, but the size of its effect can easily change significantly due to QTL by environment interaction. Therefore we believe that the gene's location and the sign of its effect are probably the most relevant parameters. Thus, we question the value of putting standard errors on effects, but we would certainly like to report a 95% confidence region for the QTL location. The inverse of Fisher's information matrix is the standard tool in a statistician's ML toolbox for calculation of standard errors. Computation of this matrix is a difficult task in the case of mixture models, but at relatively low cost the matrix can be approximated by using first – order partial derivatives, which are directly available from the EM algorithm (Redner and Walker, 1984; Jansen and Stam, 1994). But there is a more direct way to construct a confidence region for the QTL location. We usually calculate and plot the QTL likelihood profile at each map position along the map. A clear peak in this profile is taken as the most likely QTL position. Suppose we have a clear peak in the profile. According to standard ML theory, a 95% confidence region for QTL location is then bounded by the map positions where the profile is $\chi^2_{0.05}(1) = 3.84$ less than at the peak (equivalent to 0.84 *LOD*). Let us call this region a *support region* (Figure 21.3). This region is a 95% confidence region if and only if the confidence interval falls within one marker interval. What is the critical problem? If the confidence region spans more than one interval, then we actually compare several different statistical models. But ML properties only hold within one model and not across models. In other words, we have no general theory to guarantee that the support interval is a 95% confidence region. Various simulation studies have shown that a 2 *LOD* support interval is a safe choice in most cases yielding at least 95% confidence regions (e.g. Van Ooijen, 1992).

Here are some more warnings. Most of them are pretty obvious from the statistical point of view, knowing that what can go wrong in ANOVA and regression analysis is often due to multicollinearity. Suppose that you have used the appropriate genome-wide significance threshold and detected a significant QTL at a certain map position where the test score peaks highest. You may yell 'Hurray, I have found a QTL!'. But you have only found a statistical association and not a gene, and there are at least four traps that may lead you to erroneous inference. The first trap is that there are actually two or more linked QTLs with effects of equal sign (QTLs are in coupling phase), in which case it is not unlikely that the analysis will reveal a single QTL in the middle of the two true

QTLs. This is known as the detection of a ghost QTL (Martinez and Curnow, 1992), an error of type 1. The second trap is that an unlinked major QTL has inflated the test score. Incidental association can arise due to deviations from expected segregation ratios for any pair of loci on the genome. This is especially probable in the case of severe segregation distortion, or in small populations where larger deviations from expected segregation ratios may arise as the result of natural sampling variation. Of course, this is another example of a ghost QTL, but the general user less often anticipates it. The third trap is even nastier than the other two, although it can be considered as the multi-QTL extension. What we test with interval mapping is nothing more or less than an *average* effect of all QTLs in the region under study. With interval mapping there is no way to dissect their effects; any effect detected can be the sum of many possibly small QTL effects instead of the effect of an isolated QTL. If we repeat our experiment, the effects of the QTLs can be modulated in different ways and therefore the average effect may give rise to a peak of the test score at another map location. Furthermore, if two or more QTLs have effects of different sign (QTLs are in repulsion phase), then the joint effect can be close to zero. In other words, in such cases (major) QTLs will remain undetected. Finally, the fourth trap has to do with variable information content. Suppose that the information content is relatively low in a region containing a QTL. What may happen is that the peak is shifted towards more informative regions (Knott and Haley, 1992).

21.4.3 Composite Interval Mapping

Is there a way to avoid the traps described above? In fact we have already seen almost all the essentials for such an approach. In Section 21.3 we mentioned the solution: tabulate the variation accounted for by a given QTL, having eliminated all the effects of the other QTLs postulated in the model. (Recall the DH example, in which we postulated up to eight QTLs per chromosome, equally spaced every ~20 cM.) As in Section 21.4.2, we calculate the QTL likelihood at each marker position, plot it along the map and look for peaks. If the marker map is sparse, we can calculate QTL likelihood at multiple positions within a marker interval to get a smooth QTL likelihood profile. In principle we can postulate QTLs at any position we like – on top of markers or inside marker intervals (Jansen, 1992; Kao et al., 1999). In the practice of working with dense maps we often place them on top of markers. We perform a genome scan by moving a QTL along the chromosomes while using a pre-identified set of markers as cofactors (Jansen, 1992; 1993b; Zeng, 1994). Or, in other words, for the sparse map case, we combine interval mapping with multiple regression on markers. Stam (1991), and later Rodolphe and Lefort (1993) and Zeng (1993), demonstrated that in regression the effect of a QTL is absorbed only by its flanking cofactors, at least if progeny size is large. So, suppose we test for the presence of a QTL at a certain map position with a cofactor at some distance on the left and one on the right-hand side of the QTL. The test is now asymptotically unaffected by any QTLs located to the left of the left cofactor and to the right of the right cofactor (Stam, 1991).

Let us first pay some attention to what is known about setting the genome-wide significance threshold. Zeng (1994) studied the sparse map case and proposed to use all markers as cofactors, except the ones flanking the interval under study. Thus, the model for the ith individual in a DH population is

$$y_i = \mu + \sum_{j=1}^{M} x_{ij}\beta_j + x_{i0}\beta_0 + \varepsilon_i$$

where summation is over M marker cofactors not flanking the interval under study, x_{ij} are 0–1 indicator variables and β_j are substitution effects for the jth marker cofactor ($j = 1, \ldots, M$), or for the QTL ($j = 0$). Zeng (1994) used maximum likelihood and called the method 'composite interval mapping' or CIM. Simulation work with backcross populations demonstrated that $\chi^2_{\alpha/M}$ (2) can be used as an upper bound for the $100\alpha\%$ genome-wide threshold on a genome with M marker intervals, unless the number of parameters is too large (Zeng, 1994; see his Figure 21.1). The $\chi^2_{\alpha/M}$ (2) relation does not hold if the number of parameters exceeds $2\sqrt{}$(number of observations) (Jansen and Stam, 1994). With the analysis-of-deviance approach (Jansen and Stam, 1994), however, there is no limitation on the number of parameters and $2F_{a/M}$ (2, df) can be used as an upper bound, where df are the degrees of freedom for estimating the residual error variance.

Although attractive in properly disentangling the effects of multiple (linked) QTLs, the precision mapping approach has some serious drawbacks. Suppose there is a QTL located in the interval currently under study. What can happen? First, any linked cofactor in the (wide) neighbourhood of our testing position will absorb at least part of the effect of the QTL, because we use conditional tests. Second, in models with many cofactors it is even likely that a linear combination of unlinked cofactors explains a considerable part of the variation of our QTL (Jansen, 1994b). Third, the genome-wide threshold is a function of the number of intervals M, which means that the significance threshold increases with the number of cofactors. Putting the pieces together, the conclusion is that with too many cofactors we will simply end up with high precision (we will make no type 1 errors), but no power (we will miss most QTLs). In a later paper Kao et al. (1999) proposed an alternative version of CIM, which they called multiple-interval mapping (MIM), and which included a stepwise selection phase. Such strategies will be discussed in more detail in the next subsection.

21.4.4 Multiple-QTL Mapping

Parsimony of parameters is one of the basic paradigms in statistics. A parsimonious model does not include parameters which are unnecessary or which (even worse) decrease the chance (power) of reaching our goals (Draper and Smith, 1981; McCullagh and Nelder, 1989). In our context, we do not want the model to include cofactors in regions where there are no QTLs. There is no unique best statistical method for finding the 'important' cofactors, but there are several good and general criteria, penalizing the use of extra parameters (e.g. AIC criterion). But selection of a useful set of cofactors from a large set of possible cofactors to form a parsimonious model is a non-trivial task with both statistical and computational problems. Below we will discuss forward selection, backward elimination and stepwise regression. We think that selection is the most important step in QTL analysis. Subsequent to cofactor selection, we perform a genome scan by moving a QTL along the chromosomes while using a preselected set of markers as cofactors (Jansen, 1993b; 1994b; Jansen and Stam, 1994). If the moving QTL gets close to a cofactor, say within a preset window of 10 or 20 cM, then we drop the cofactor (Figure 21.4). This approach is called multiple-QTL mapping. Zeng (1994) also suggested the use of a thinned set of cofactors (e.g. via stepwise regression), and elaborated on it in detail in Kao et al. (1999).

We mentioned three selection strategies: forward, backward and stepwise. In the forward selection approach, at each stage the best new cofactor satisfying the selection criterion is added until no further candidates remain. This approach is often used in QTL analysis, on top of the IM approach. Once we have found one or more QTLs with IM, we add

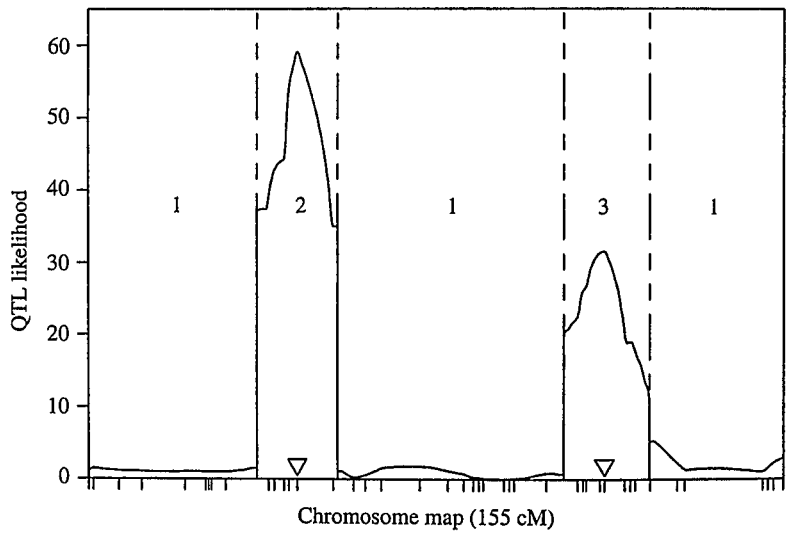

Figure 21.4 QTL likelihood (deviance) profile obtained with MQM mapping for the case of two linked QTLs. Marks below the x-axis indicate positions of markers along the chromosome under study. Triangles indicate positions of selected cofactors. Different tests for the presence of a QTL are performed in regions indicated by '1', '2', and '3', respectively. In region '1' the test is conditional on the cofactors at 36 and 76 cM. In region '2' it is conditional on the cofactor at 76 cM, and in region '3' it is conditional on the cofactor at 36 cM. Regions '2' and '3' are chosen (i) symmetrically around their respective cofactor, and (ii) wide enough to span their respective 2*LOD* support interval for the QTL (support intervals not shown). The genome-wide significance threshold is ~14.

them to our model and rerun the genome scan. Although a natural procedure, we do not recommend it for two reasons. First, because we do not avoid the traps described above for the single-QTL analysis. Second, the approach does not exploit power to the full: each test is based on the ratio between variance explained by the factor under study and the unexplained variance. In a forward selection approach unexplained variance contains environmental plus as yet unexplained genetic variance, which decreases power relative to a backward elimination procedure.

The backward elimination procedure starts with a multiple regression model, using a full set of cofactors (putative QTLs/markers) evenly spread over the genome. The unimportant or least important cofactors are dropped one by one until all remaining cofactors are essential given the selection criterion. This is a satisfactory approach, especially if we wish to see all the variables in the model in order 'not to miss any QTL'. The full model gives an unbiased estimate of the maximum amount of variance explainable by (non-interacting) cofactors (QTLs). In order to exclude redundant cofactors, the selection criterion should be stringent, but not so stringent that important cofactors (those flanking the QTLs) are thereby excluded. We can use one of the two approaches that we described above: the maximum likelihood approach (Jansen, 1993b; Zeng, 1994; Kao et al., 1999) or the analysis-of-deviance approach (Jansen and Stam, 1994; Jansen, 1994b).

For the ML approach, Jansen (1993b) proposed to maximize the log-likelihood (l) minus the number of free parameters (k) in the model; this is equivalent to minimizing Akaike's information criterion (AIC), given by $-2(l-k)$. In general, a penalty in the range of k to $3k$ may provide plausible initial models (McCullagh and Nelder, 1989).

In 'ordinary' regression with adequate degrees of freedom to estimate σ^2, a penalty of k is equivalent to the use of (about) the 16% point of the F-test for the comparison of two nested models, which differ only by the inclusion of one free parameter; a penalty of $3k$ is equivalent to the use of (about) the 2% point (McCullagh and Nelder, 1989). Note that we can also compare *non-nested* models by using the AIC. And we can even compare several models fitted on different scales, having multiplied $f_{\text{mix}}(T(\cdot))$ by its Jacobian $T'(\cdot)$. Finally, we recall here that there are two disadvantageous features of the ML approach. First, the number of parameters should remain relatively small – less than $2\sqrt{}$(number of observations). Second, there is the danger of over-fitting and thereby underestimating the error variance.

In the analysis-of-deviance approach, one can use partial F-tests conditional on the other cofactors in the current model. Note that the same approach is valid for regression mapping (Haley and Knott, 1992; Martinez and Curnow, 1992). The partial F-test values are calculated for each cofactor. The cofactor with the lowest partial F-test value is removed if its effect is less significant than a preset significance level. This process is repeated until all remaining cofactors have a partial F-test value exceeding the threshold. Jansen (1994b) used a 2% threshold in simulations. It was demonstrated that this penalty is stringent, since no or only a few cofactors are generally selected in the case of no QTLs segregating. Moreover, it was shown that this penalty is still not too stringent, since cofactors are selected for those QTLs that considerably affect the test statistic in their nearby region; the effects of such QTLs are satisfactorily eliminated by selected cofactors. Because of this feature, the IM thresholds, which were obtained for the case that no QTLs are segregating, are also suitable for MQM mapping. These thresholds can be used when no QTLs are segregating, since in that case no or only a few cofactors will be selected; moreover, these thresholds can still be used when there are QTLs segregating, the effects of which are eliminated by cofactors. Detection and unraveling of the separate QTL effects in the case of linked loci is much easier in MQM mapping than in IM (Jansen, 1994b).

One of the disadvantages of the backward elimination procedure is that for a cofactor 'once out' means 'always out'. Backward elimination followed by a stepwise procedure, including new cofactors and dropping old ones, may help to overcome this – at the cost of more computation. Alternatively, replacing important cofactors by neighbours, not present in the initial set of cofactors, can also help in fine-tuning the model.

We emphasize one important difference between MQM, on the one hand, and IM and CIM, on the other hand, that arises because of MQM's use of a fixed residual error for both full and reduced models. Suppose we compare two nested models, in the simplest case a model with a QTL versus a model with no QTL. In regression mapping and MQM the *LR*-test or the equivalent F-test is based on the fit of the full model versus that of a reduced model with the residual error fixed at the value estimated from the full model. If the QTL has a major effect, then the no-QTL model will fit badly and the test score will be high. If the (putative) QTL has a minor or no effect, then the residual variance estimate of the full model is fully satisfactory for the no-QTL model. What happens in (composite) interval mapping? The residual error variance in the no-QTL model is not fixed, but it is estimated and it will absorb the QTL variation. As a result the test score will not be as high as in the MQM and regression mapping approaches. A minor detail? Not really, because the effect on the test score can be considerable. For instance, in Figure 21.3 the test score is 50 for IM and 66 for MQM (the latter is not shown). Only

one QTL was simulated on the whole genome. Thus in this case the difference between interval mapping and MQM is solely due to the way the residual variance is treated.

21.4.5 Uncritical use of Model Selection Procedures

We would like to emphasize the following warning often posted in the statistical literature: *the uncritical use of the results of selection procedures can be very dangerous.* Here is one example of what can go wrong. Suppose we fit two cofactors, which are not far apart on the map; say that there is only one recombinant individual in a DH population. What can happen? The model for the ith individual is

$$y_i = \mu + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i,$$

where x_{ki} and β_{ki} are the 0–1 indicator variables and effect variables for the two closely linked cofactors, respectively. For all individuals but one, $x_{1i} = x_{2ij}$, that is either

$$y_i = \mu + \varepsilon_i$$

or

$$y_i = \mu + \beta_1 + \beta_2 + \varepsilon_i$$

holds. In the latter case only the sum $\beta = \beta_1 + \beta_2$ is relevant. Suppose the single recombinant individual has $x_{1i} = 0$ and $x_{2i} = 1$, in which case

$$y_i = \mu + \beta_2 + \varepsilon_i.$$

For a given value of μ and β, we can obtain a perfect fit for the ith individual by setting

$$\beta_2 = y_i - \mu,$$

$$\beta_1 = \beta - \beta_2.$$

If there are no real QTLs in this region, then one would expect that $\beta = \beta_1 + \beta_2 = 0$. But we can still get large values $\beta_1 = -\beta_2$, which we might take as evidence for two linked QTLs with opposite effects. In some cases this may be true, but more frequently we probably just misinterpret what is actually no more than a statistical artefact. Other artificial patterns of near-collinearity between sets of cofactors can arise. Moreover, any real data set will contain erroneous trait and marker scores (genotypic mis-scores, phenotypic outliers and so forth), for which our models may want to compensate for by an increase in the number of parameters. Several things can be done to avoid serious problems – on the biological and on the statistical side of the problem. First, make sure that the trait and marker data are very well checked! We prefer that labels *U*, *C* or *D* replace suspicious marker scores. Second, avoid simultaneous fitting of closely linked cofactors. For instance, in a DH population of 100 individuals we should make sure that simultaneously fitted cofactors stay at least 10–20 cM apart. If we do so, then there are 10–20 recombinants expected between any pair of cofactors. Third, we can set the genome-wide significance threshold by an empirical (permutation or bootstrap) strategy, in which we mimic the complete selection protocol at each run.

21.4.6 Final Comments

What if the trait is encoded by many QTLs, say 40, each of small effect? It will definitely be hard to decompose the phenotypic variation into the underlying QTL components. The

number of recombinants is the limiting factor, and the only way out is to generate more of them by increasing the population size or by choosing another type of cross with more informative meioses. See Beavis (1996) and Visscher et al. (2000) for some illustrative examples of the hard job of dissecting polygenic variation.

Last but not least, we would like to state that the proof of the QTL pudding is not in the eating of the results from statistical QTL analysis. It is in the eating of results from cross-validation and/or new biological experiments.

21.5 BIBLIOGRAPHIC NOTES

The textbook *Genetics and Analysis of Quantitative Traits* by Lynch and Walsh (1998) provides a wealth of information for the interested reader. The book brings together basic biology and many methods of analysis.

In this chapter QTL analysis is accomplished by embedding it in a more or less standard statistical framework of (generalized) linear models. For general discussion of linear models we refer to the textbook *Applied Regression Analysis* by Draper and Smith (1981), and the textbook *Generalized Linear Models* by McCullagh and Nelder (1989). GLMs provide a unified analysis of diverse types of trait data (e.g. normal, binary, ordinal) via a simple approach of iterative reweighted least squares.

Jansen (1993a; 1994a) embedded mixture models in the GLM framework, which is of prime importance to segregation analysis and QTL analysis, where one deals with mixtures of different genotypes. Here we have focused on QTL analysis in inbred lines for normally distributed traits; see, for instance, Hackett and Weller (1995) for QTL analysis in inbred lines with ordinal trait data and Visscher et al. (1996a) for QTL analysis with binary data. For extensive rigorous statistical treatment of mixture models we refer to the review by Redner and Walker (1984), the textbook by Titterington et al. (1985), and Dempster et al.'s (1977) paper on the EM algorithm. In recent years Bayesian modelling of mixtures with an unknown number of underlying components has made enormous progress (Richardson and Green, 1997) and is beginning to have an impact on various applications, including genetics (see below and the chapters by Hoeschele and by Gianola in this book).

During the past decade a number of important papers have been published on QTL analysis with sparse marker maps. We list some of the early and most influential papers: Lander and Botstein (1989), who coined the name 'interval mapping'; Haley and Knott (1992) and Martinez and Curnow (1992), with the 'regression mapping' approach; Jansen (1992; 1993b; 1994b) and Jansen and Stam (1994), who developed the general GLM framework for QTL analysis and the MQM approach; Zeng (1993; 1994) and Kao et al. (1999), with composite interval mapping and multiple interval mapping; and finally, Churchill and Doerge (1994) and Doerge and Churchill (1996), with the permutation test for setting the genome-wide significance threshold. There are various software packages available, among them Map Manager, MapMaker/QTL, MapQTL, Multi-QTL, PlabQTL, QTL Cartographer (further information can easily be found on the world wide web).

From a statistical viewpoint there is not much special to QTL modelling. If we look in our statistical toolbox, we will see, for instance, tools for interaction models, for mixed models with random and fixed effects, for multivariate instead of univariate analysis, and for Bayesian analysis. For each we list some key references.

QTL by environment interaction is a very important issue in plant breeding. We refer to Jansen (1992), Jansen et al. (1995), Tinker and Mather (1995), and Jiang and Zeng (1995). A second important type of interaction is epistasis, that is QTL by QTL interaction. See, for instance, Chase et al. (1997), Fijneman et al. (1996), and Kao et al. (1999). One of the major obstacles with epistatic interactions is the increase of the number of parameters, which is generally not compensated by a similar increase of the population size. In a recent paper Wang et al. (1999) combine QTL by environment and additive QTL by QTL interaction and propose the use of fixed effect terms for the QTL, and random effect terms for cofactors and interactions.

In most experiments two or more traits are scored on the same individual. Jiang and Zeng (1995) and Korol et al. (1995) are good entries to multitrait QTL analysis. For correlated traits a higher resolution can be obtained by multitrait analysis than by separate univariate analyses.

In recent years Bayesian statistics has received much interest in genetics. Undoubtedly, that popularity is largely due to the (MCMC) algorithms that enable the estimation of complex genetic models. Bayesian methods are particularly good for missing data augmentation (which we have emphasized here for the EM and MCEM algorithm). With ultra-dense maps we can obtain nearly complete data, and the use of Bayesian methods in the relatively simple case of inbred line genetics is perhaps overkill. However, we have seen that model selection – how many QTLs and on which map positions – is still a hard problem, and Bayesian methods are becoming more prominent for this. We refer to Green (1995), certainly one of the pioneering papers on Bayesian model selection. Satagopan et al. (1996) described the first Bayesian method for inbred line crosses. In their approach models with different (fixed) numbers of QTLs were compared via Bayes factors (similar to the *LR*-test). Sillanpää and Arjas (1998) allowed for a variable number of linked QTLs on the chromosome under study. The effects of unlinked QTLs were taken into account by fixed preselected cofactors (as in CIM and MQM). Stephens and Fisch (1998) allowed for a variable number of linked and unlinked QTLs. There is much more to say about the potential of Bayesian methodology, and we refer the interested reader to the chapter by Gianola and by Hoeschele in this volume for many more details.

In a single cross, inferences about QTLs and their estimated effects are limited to the particular cross. A breeder, however, usually produces many crosses instead of one. A multipopulation analysis would allow us to extend inferences across populations. Rebaï and Goffinet (1993) dealt with QTL analysis of diallel crosses between inbred lines, Xu (1998) with multiple unrelated families. Jansen and Beavis (2000) developed an approach based on haplotyped putative QTL alleles. Finally, we refer to Hoeschele (this volume) for more general theory about QTL analysis in complex populations.

Acknowledgments

The author would like to thank Ina Hoeschele, Jean-Luc Jannink, Piet Stam and Mathisca de Gunst for many useful comments on earlier versions of this chapter.

REFERENCES

Allard, R.W. (1999). *Principles of Plant Breeding*. Wiley, New York.
Atkinson, A.C. (1985). *Plots, Transformations and Regression*. Clarendon Press, Oxford.

Beavis, W.D. (1996). QTL analyses: power, precision and accuracy. In *Molecular Analysis of Complex Traits*, A.H. Paterson ed., CRC Press, Boca Raton, FL.

Beckmann, J.S. and Soller, M. (1983). Restriction fragment length polymorphisms in genetic improvement methodologies, mapping and costs. *Theoretical and Applied Genetics* **67**, 35–43.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980). Construction of a genetic map in man using fragement restriction length polymorphisms. *American Journal of Human Genetics* **32**, 314–331.

Chase, K., Adler, F.R. and Lark, K.G. (1997). Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci. *Theoretical and Applied Genetics* **94**, 724–730.

Churchill, G.A. and Doerge, R.W. (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. In *Development and Application of Molecular Markers to Problems in Plant Genetics*, T. Helentjaris and B Burr, eds. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 113–116.

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

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

Doerge, R.W. and Rebaï, A. (1996). Significance thresholds for QTL interval mapping tests. *Heredity* **76**, 459–464.

Draper, N.R. and Smith, H. (1981). *Applied regression analysis*. Wiley, New Work.

Fijneman, R.J.A., de Vries, S.S., Jansen, R.C. and Demant, P. (1996). Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility of lung cancer in the mouse. *Nature Genetics* **14**, 465–467.

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

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

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

Jansen, R.C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.

Jansen, R.C. (1993a). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* **49**, 227–231.

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

Jansen, R.C. (1994a). Maximum likelihood in a finite mixture model by exploiting the GLM facilities of Genstat. *Genstat Newsletter* **30**, 25–27.

Jansen, R.C. (1994b). Controlling the type 1 and type 2 errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Jansen, R.C. (1996). A general Monte Carlo method for mapping quantitative trait loci. *Genetics* **142**, 305–311.

Jansen, R.C., Reinink, K. and van der Heijden, G.W.A.M. (1993). Analysis of grey level histograms by using statistical methods for mixtures of distributions. *Pattern Recognition Letters* **14**, 585–590.

Jansen, R.C. and Den Nijs, A.P.M. (1993). A statistical mixture model for estimating the proportion of unreduced pollen grains in perennial ryegrass (lolium perenne) via the size of pollen grains. *Euphytica* **70**, 205–215.

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

Jansen, R.C., van Ooijen, J.W., Stam, P., Lister, C. and Dean, C. (1995). Genotype by environment interaction in genetic mapping of multi quantitative trait loci. *Theoretical and Applied Genetics* **91**, 33–37.

Jansen, R.C. and Beavis, W.D. (2000). MQM mapping using haplotyped putative QTL-alleles: a simple approach for mapping QTLs in plant breeding populations. Filed Patent Application 04–010110US.

Jensen, J. (1989). Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theoretical and Applied Genetics* **78**, 613–618.

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

Kao, C.-H., Zeng, Z.-B. and Teasdale, R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Knott, S.A. and Haley, C.S. (1992). Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**, 1211–1222.

Korol, A.B., Ronin, Y.I. and Kirzhner, V.M. (1995). Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* **140**, 1137–1147.

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

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

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

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

Rebaï, A. and Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies desired from a di allel cross. *Theoretical and Applied Genetics*, **86**, 1014–1022.

Rebaï, A., Goffinet, B. and Mangin, B. (1994). Approximate thresholds for interval mapping tests for QTL detection. *Genetics*, **138**, 235–240.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B* **59**, 731–792.

Rodolphe, F. and Lefort, M. (1993). A multi-marker model for detecting chromosomal segments displaying QTL activity. *Genetics* **134**, 1277–1288.

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

Sillanpää, M.J. and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373–1388.

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.

Soller, M. and Genizi, A. (1978). The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* **34**, 47–55.

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

Spelman, R.J., Coppieters, W. Karim, L. Van Arendonk, J.A.M. and Bovenhuis, H. (1996). Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics* **144**, 1799–1808.

Stam, P. (1991). Some aspects of QTL analysis. In: *Proceedings of the Eighth Meeting of the Eucarpia Section 'Biometrics in Plant Breeding'*. Brno.

Stephens, D.A. and Fisch, R.D. (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**, 1334–1347.

Tinker, N.A. and Mather, D.E. (1995). Methods for QTL analysis with progeny replicated in multiple environment. *Journal of Agricultural Genomics* **1**(1). http://www.ncgr.org/research/jag/.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

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

Van Ooijen, J.W. (1999). LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity* **83**, 613–624.

Visscher, P.M., Haley, C.S. and Knott, S.A. (1996a). Mapping QTLs for binary traits in backcross and F_2 populations. *Genetical Research* **68**, 55–63.

Visscher, P.M., Thompson, R. and Haley, C.S. (1996b). Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020.

Visscher, P.M., Whittaker, J.C. and Jansen, R.C. (2000). Mapping multiple QTL of different effects: comparison of a simple sequential testing strategy and MQM. *Molecular Breeding* **6**, 11–24.

Wang, D.L., Zhu, J., Li, Z.K. and Paterson, A.H. (1999). Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches. *Theoretical and Applied Genetics* **99**, 1255–1264.

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.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 60–62.

Xu, S.-Z. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics* **148**, 517–524.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of National Academy of Sciences* **90**, 10972–10976.

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