Lecture 9
QTL Mapping

ECOL 519A, Jan 2007. University of Arizona

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.

MAPPING USING INBRED LINE CROSSES

We start by considering crosses between inbred lines. The analysis of such crosses illustrates many of the fundamental features of QTL mapping without the additional complications that arise with outbred populations. Although inbred line crosses are uncommon in animal breeding (outside of rats and mice), crosses between widely-differing lines are often treated as an inbred line cross, as we assume that marker and QTL allele frequencies are very different between the lines.

Experimental Designs

Starting with two completely inbred parental lines, P_1 and P_2, a number of line-cross populations derived from the F_1 can be used for QTL mapping. The **F_2 design** examines marker-trait associations in the progeny from a cross of F_1s, while the **backcross design** examines marker-trait associations in the progeny formed by backcrossing the F_1 to one of the parental lines. While these are the most widely used designs, other line-cross populations can offer further advantages (and disadvantages). Designs using an F_t population (**advanced intercross lines**, formed by randomly mating F_1s for $t-1$ generations) allow for higher resolution of QTL map positions than do F_2s, albeit at the expense of decreased power of QTL detection.

Experimental designs are also classified by the unit of marker analysis chosen by the investigator. Marker-trait associations can be assessed using one-, two-, or multiple-locus marker genotypes. Under a **single-marker analysis**, the distribution of trait values is examined separately for each marker locus. Each marker-trait association test is performed independent of information from all other markers, so that a chromosome with n markers offers n separate single-marker tests. A single-marker analysis is generally a good choice when the goal is simple *detection* of a QTL linked to a marker, rather than *estimation* of its position and effects. Under **interval mapping** (or **flanking-marker analysis**), a separate analysis is performed for each *pair* of adjacent marker loci. The use of such two-locus marker genotypes results in $n-1$ separate tests of marker-trait associations for a chromosome with n markers (one for each marker interval). Interval mapping offers increased of power of detection (albeit usually slight) and more precise estimates of QTL effects and position. Both single-marker and interval mapping approaches are biased when multiple QTLs are linked to the marker/interval being considered. Methods simultaneously using three or more marker loci attempt to reduce or remove such bias. **Composite interval mapping** considers a marker interval plus a few other well-chosen single markers in each analysis, so that (as above) $n-1$ tests for interval-trait associations are performed on a chromosome with n markers. **Multipoint mapping** considers all of the linked markers on a chromosome simultaneously, resulting in a single analysis for each chromosome

Conditional Probabilities of QTL Genotypes

The basic element upon which the formal theory of QTL mapping is built is the conditional probability that the QTL genotype is Q_k, given the observed marker genotype is M_j. From the definition of a conditional probability (Lecture 1),

$$\Pr(Q_k \mid M_j) = \frac{\Pr(Q_k M_j)}{\Pr(M_j)} \tag{9.1}$$

The joint $\Pr(Q_k M_j)$ and marginal $\Pr(M_j)$ probabilities are functions of the experimental design and the linkage map (the position of the putative QTLs with respect to the marker loci). Computing these probabilities is a relatively simple matter of bookkeeping, but can get rather tedious as the number of markers and/or QTLs under consideration increases.

When computing joint probabilities involving more than two loci, one must also account for recombinational interference between loci (Lecture 3). Consider a single QTL flanked by two markers, M_1 and M_2. The gamete frequencies depend on three parameters: the recombination frequency c_{12} between markers, the recombination frequency c_1 between marker M_1 and the QTL, and the recombination frequency c_2 between the QTL and marker M_2. Under the assumption of no interference, $c_{12} = c_1 + c_2 - 2c_1 c_2$, while $c_{12} = c_1 + c_2$ under complete interference. When c_{12} is small, gamete frequencies are essentially identical under either interference assumption. Typically, c_{12} is assumed known, leaving two unknown recombination parameters (c_1 and c_2) under general assumptions about interference. In either case, there is only one parameter to estimate, as assuming complete interference $c_2 = c_{12} - c_1$, or with no interference $c_2 = (c_{12} - c_1)/(1 - 2c_1)$. Hence, for flanking-marker analysis, we restrict attention to the single recombination parameter c_1, the distance from marker locus M_1 to the QTL. When considering analysis of single-marker loci, for notational ease we drop the subscript, using c in place of c_1.

Example: Conditional Probabilities for an F_2

Consider a single-marker analysis using the F_2 formed by crossing two inbred lines, $MMQQ \times mmqq$. If the recombination frequency between the marker locus and the QTL is c, the expected F_1 gamete frequencies are

$$\Pr(MQ) = \Pr(mq) = (1-c)/2, \qquad \Pr(Mq) = \Pr(mQ) = c/2$$

The probability that an F_2 individual is $MMQQ$ is $\Pr(MQ)\Pr(MQ) = [(1-c)/2]^2$. Likewise, $2\Pr(MQ)\Pr(mQ) = 2(c/2)[(1-c)/2]$ is the probability of an $MmQQ$ individual, and so on. Since the probabilities of the marker genotypes MM, Mm, and mm are $1/4$, $1/2$, and $1/4$, Equation 9.1 gives the F_2 conditional probabilities as

$$\Pr(QQ \mid MM) = (1-c)^2, \quad \Pr(Qq \mid MM) = 2c(1-c), \quad \Pr(qq \mid MM) = c^2$$

$$\Pr(QQ \mid Mm) = c(1-c), \; \Pr(Qq \mid Mm) = (1-c)^2 + c^2, \; \Pr(qq \mid Mm) = c(1-c) \qquad (9.2)$$

$$\Pr(QQ \mid mm) = c^2, \quad \Pr(Qq \mid mm) = 2c(1-c), \quad \Pr(qq \mid mm) = (1-c)^2$$

This same logic extends to multiple marker loci. Suppose the QTL is flanked by two scored markers, and consider the F_2 in a cross of lines fixed for $M_1 Q M_2$ and $m_1 q m_2$. What are the conditional probabilities of the three QTL genotypes when the marker genotype is $M_1 M_1 M_2 M_2$? Since all F_1s are $M_1 Q M_2 / m_1 q m_2$, under the assumptions of no interference, the frequency of F_1 gametes involving $M_1 M_2$ are

$$\Pr(M_1 Q M_2) = (1-c_1)(1-c_2)/2, \qquad \Pr(M_1 q M_2) = c_1 c_2/2$$

giving expected frequencies in the F_2 of $M_1 M_1 M_2 M_2$ offspring as

$$\Pr(M_1 Q M_2 / M_1 Q M_2) = [(1-c_1)(1-c_2)/2]^2$$
$$\Pr(M_1 Q M_2 / M_1 q M_2) = 2[(1-c_1)(1-c_2)/2][c_1 c_2/2]$$
$$\Pr(M_1 q M_2 / M_1 q M_2) = (c_1 c_2/2)^2$$

where $c_2 = (c_{12} - c_1)/(1 - 2c_1)$. The overall frequency of $M_1 M_1 M_2 M_2$ individuals is the sum of the three above terms, or $(1 - c_{12})^2/4$. Substituting into Equation 9.1 gives

$$\Pr(QQ \mid M_1 M_1 M_2 M_2) = \frac{(1 - c_1)^2 (1 - c_2)^2}{(1 - c_{12})^2}$$

$$\Pr(Qq \mid M_1 M_1 M_2 M_2) = \frac{2c_1 c_2 (1 - c_1)(1 - c_2)}{(1 - c_{12})^2} \tag{9.3}$$

$$\Pr(qq \mid M_1 M_1 M_2 M_2) = \frac{c_1^2 c_2^2}{(1 - c_{12})^2}$$

Conditional probabilities for other marker genotypes are computed in a similar fashion. Since $c_1 c_2$ is usually very small if c_{12} is moderate to small, essentially all $M_1 M_1 M_2 M_2$ individuals are QQ. For example, assuming $c_1 = c_2 = c_{12}/2$ (the worst case), the conditional probabilities of an $M_1 M_1 M_2 M_2$ individual being QQ are 0.96, 0.98, and 0.99 for $c_1 = c_2 = 0.25, 0.2$, and 0.1.

Expected Marker Means

With these conditional probabilities in hand, the expected trait values for the various marker genotypes follow immediately. Suppose there are N QTL genotypes, $Q_1, \cdots, Q_N$, where the mean of the kth QTL genotype is μ_{Q_k}. The mean value for marker genotype M_j is just

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(Q_k \mid M_j) \tag{9.4}$$

The QTL effects enter through the μ_{Q_k}, while the QTL positions enter through the conditional probabilities $\Pr(Q_k \mid M_j)$. For example, if a QTL with effects $2a : a(1+k) : 0$ is linked (distance c) to a marker, applying Equation 9.2, the F_2 marker means become

$$(\mu_{MM} - \mu_{mm})/2 = a(1 - 2c) \tag{9.5a}$$

$$\frac{\mu_{Mm} - (\mu_{MM} + \mu_{mm})/2}{(\mu_{MM} - \mu_{mm})/2} = k(1 - 2c) \tag{9.5b}$$

Thus using only single marker means we cannot uncouple estimates of QTL effects (a and k) from the distance c from the marker. A small marker difference could be due to a small QTL effect tightly linked to the marker or a QTL of large effect loosely linked to the marker. With markers equallly spaced throughout the genome, say on every c centimorgans, a QTL is no more than $c/2$ from any marker, and this provides a lower bound for the QTL effect.

By considering two-locus (rather than single-locus) marker means, separate estimates of QTL effect and position can be obtained. Taking the genotype at two adjacent marker loci (M_1/m_1 and M_2/m_2) as the unit of analysis, consider the difference between the contrasting double homozygotes in an F_2. If the markers flank a QTL, then under the assumption of no interference, Equation 9.3 (and its analog for $m_1 m_1 m_2 m_2$ probabilities) implies

$$\frac{\mu_{M_1 M_1 M_2 M_2} - \mu_{m_1 m_1 m_2 m_2}}{2} = a \left(\frac{1 - c_1 - c_2}{1 - c_1 - c_2 + 2c_1 c_2} \right)$$

$$\simeq a(1 - 2c_1 c_2) \tag{9.6a}$$

where c_1 is the M_1-QTL recombination frequency. Equation 9.6a is essentially equal to a when the distance between flanking markers $c_{12} \leq 0.20$, as here $(1 - 2c_1 c_2) \geq 0.98$. Thus, recalling from

Equation 9.5a that $\mu_{M_1M_1} - \mu_{m_1m_1} = 2a(1 - 2c_1)$, we can obtain estimates of the recombination frequencies by substituting Equation 9.6a for a and rearranging to give

$$c_1 = \frac{1}{2}\left(1 - \frac{\mu_{M_1M_1} - \mu_{m_1m_1}}{2a}\right)$$

$$\simeq \frac{1}{2}\left(1 - \frac{\mu_{M_1M_1} - \mu_{m_1m_1}}{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}\right) \tag{9.6b}$$

Linear Models for QTL Detection

The simplest linear model considers the phenotypic value z_{ik} of the kth individual of marker genotype i as a mean value μ plus a marker effect b_i and a residual error e_{ik},

$$z_{ik} = \mu + b_i + e_{ik} \tag{9.7a}$$

This is a one-way ANOVA model (Lecture 5), with the presence of a linked QTL being indicated by a significant between-marker variance. Equivalently, we can express this model as a multiple regression, with the phenotypic value for individual j given by

$$z_j = \mu + \sum_{i=1}^{n} b_i\, x_{ij} + e_j \tag{9.7b}$$

where the x_{ij} are n indicator variables (one for each marker genotype),

$$x_{ij} = \begin{cases} 1 & \text{if individual } j \text{ has marker genotype } i, \\ 0 & \text{otherwise.} \end{cases}$$

The number of marker genotypes (n) in Equations 9.7a,b depend on both the number of marker loci and the type of design being used. With a single marker, $n = 2$ for a backcross design, while $n = 3$ for an F_2 design (using codominant markers). When two or more marker loci are simultaneously considered, b_i corresponds to the effect of a *multilocus* marker genotype, and n is the number of such genotypes considered in the analysis. In the regression framework, evidence of a linked QTL is provided by a significant r^2, which is the fraction of character variance accounted for by the marker genotypes.

Estimation of dominance requires information on all three genotypes at a marker locus, i.e., an F_2, F_t, or other design (such as *both* backcross populations). In these cases, dominance can be estimated using an appropriate function of the marker means (e.g., Equation 9.5b). Epistasis between QTLs can be modeled by including interaction terms. Here, an individual with genotype i at one marker locus and genotype k at a second is modeled as $z = \mu + a_i + b_k + d_{ik} + e$, where a and b denote the single-locus marker effects, and d is the interaction term due to epistasis between QTLs linked to those marker loci. In linear regression form this model becomes

$$z_j = \mu + \sum_{i}^{n_1} a_i\, x_{ij} + \sum_{k}^{n_2} b_k\, y_{kj} + \sum_{i}^{n_1}\sum_{k}^{n_2} d_{ik}\, x_{ij}\, y_{kj} + e_j \tag{9.7c}$$

where x_{ij} and y_{kj} are indicator variables for two different marker genotypes (with n_1 and n_2 genotypes, respectively). Significant a_i and/or b_k terms indicate significant effects at the individual marker loci, while significant d_{ik} terms indicate epistasis between the effects of the two markers.

Maximum Likelihood Methods for QTL Mapping and Detection

Maximum likelihood (ML) methods are especially popular in the QTL mapping literature. While linear models use only marker means, ML uses the full information from the marker-trait distribution and, as such, is expected to be more powerful. The tradeoff is that ML is computationally intensive, requiring rather special programs to solve the likelihood equations, while linear model analysis can be performed with almost any standard statistical package. Further, while modifying the basic model (such as adding extra factors) is rather trivial in the linear model framework, with ML new likelihood functions need to be constructed and solved for each variant of the original model.

Assuming that the distribution of phenotypes for an individual with QTL genotype Q_k is normal with mean μ_{Q_k} and variance σ^2, the likelihood for an individual with phenotypic value z and marker genotype M_j becomes

$$\ell(z \mid M_j) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2) \Pr(Q_k \mid M_j) \tag{9.8}$$

where $\varphi(z, \mu_{Q_k}, \sigma^2)$ denotes the density function for a normal distribution with mean μ_{Q_k} and variance σ^2, and a total of N QTL genotypes is assumed. This likelihood is a mixture-model distribution (Lecture 15). The mixing proportions, $\Pr(Q_k \mid M_j)$, are functions of the genetic map (the position(s) of the QTL(s) with respect to the observed markers) and the experimental design, while the QTL effects enter only though the means μ_{Q_k} and variance σ^2 of the underlying distributions.

The likelihood equations can be modified to account for dichotomous (binary) and polychotomous (ordinal) characters through the use of logistic regressions and probit scales. Alternatively, one can simply ignore the discrete structure of the data, treating them as if they were continuous (e.g., coding alternative binary characters as 0/1) and applying ML. When flanking markers are used, this approach gives essentially the same power and precision as methods specifically designed for polychotomous traits, but when single markers are used, this approach can give estimates for QTL position that are rather seriously biased.

Likelihood Maps

In the likelihood framework, tests of whether a QTL is linked to the marker(s) under consideration are based on the likelihood-ratio statistic,

$$\mathrm{LR} = -2 \ln \left[\frac{\max \ell_r(\mathbf{z})}{\max \ell(\mathbf{z})} \right]$$

where $\max \ell_r(\mathbf{z})$ is the maximum of the likelihood function under the null hypothesis of no segregating QTL (i.e., under the assumption that the phenotypic distribution is a single normal). This test statistic is approximately χ^2-distributed, with the degrees of freedom given by the extra number of fitted parameters in the full model. For a model assuming a single QTL, most designs have five parameters in the full model (the three QTL means, the variance, and the QTL position), and two in the reduced model (the mean and variance), giving three degrees of freedom. Certain designs (such as a backcross) involve situations where only two QTL means enter (e.g., Qq and QQ or qq for a backcross), and here the likelihood ratio has two degrees of freedom.

The amount of support for a QTL at a particular map position is often displayed graphically through the use of **likelihood maps** (Figure 9.1), which plot the likelihood-ratio statistic (or a closely related quantity) as a function of map position of the putative QTL. For example, the value of the likelihood map at $c = 0.05$ gives the likelihood-ratio statistic that a QTL is at recombination fraction 0.05 from the marker vs. a model assuming no QTL. This approach for displaying the support for a QTL was introduced by Lander and Botstein (1989), who plotted the LOD (**likelihood of odds**)

scores (Morton 1955). The LOD score for a particular value of c is related to the likelihood-ratio test statistic (LR) by

$$\text{LOD}(c) = -\log_{10}\left[\frac{\max \ell_r(\mathbf{z})}{\max \ell(\mathbf{z}, c)}\right] = \frac{\text{LR}(c)}{2 \ln 10} \simeq \frac{\text{LR}(c)}{4.61} \tag{9.9}$$

showing that the LOD score is simply a constant times the likelihood-ratio statistic. Here $\max \ell(\mathbf{z}, c)$ denotes the maximum of the likelihood function given a QTL at recombination frequency c from the marker. Another variant is simply to plot $\max \ell(\mathbf{z}, c)$ instead of the likelihood-ratio statistic, as the restricted likelihood, $\max \ell_r(\mathbf{z})$, is the same for each value of c.
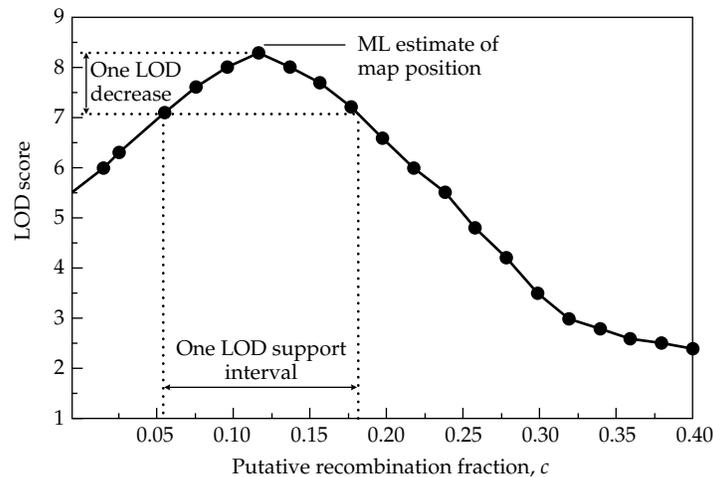

Figure 9.1 Hypothetical likelihood map for the marker-QTL recombination frequency c in a single-marker analysis. Points connected by straight lines are used to remind the reader that likelihood maps are computed by plotting the maximum of the likelihood function for each c value, usually done by considering steps of 0.01 to 0.05. A QTL is indicated if any part of the likelihood map exceeds a critical value. In such cases, the ML estimate for map position is the value of c giving the highest likelihood. Approximate confidence intervals for QTL position (one-LOD support intervals) are often constructed by including the set of all c values giving likelihoods within one LOD score of the maximum value.

The likelihood map projects the multidimensional likelihood surface (which is a function of the QTL means, variance, and map position) on to a single dimension, that of the map position, c. The ML estimate of c is that which yields the maximum value on the likelihood map, and the values for the QTL means and variance that maximize the likelihood given this value of c are the ML estimates for the QTL effects. Thus, in the likelihood framework, *detection* of a linked QTL and *estimation* of its position are coupled — if the likelihood ratio exceeds the critical threshold for that chromosome, it provides evidence for a linked QTL, whose position is estimated by the peak of the likelihood map. If the peak does not exceed this threshold, there is no evidence for a linked QTL.

Precision of ML Estimates of QTL Position

Since ML estimates are approximately normally distributed for large sample sizes, confidence intervals for QTL effects and position can be constructed using the sampling variances for the ML estimates (Lecture 1). Approximate confidence intervals are often constructed using the **one-LOD rule** (Figure 9.1), with the confidence interval being defined by all those values falling within one LOD score of the maximum value. The motivation for such **one-LOD support intervals** follows from the fact that the large-sample distribution of the LR statistic follows a χ^2 distribution. If only one parameter in the likelihood function is allowed to vary, as when testing whether c equals a particular value (say the observed ML estimate), the LR statistic has one degree of freedom. Because a

one-LOD change corresponds to an LR change of 4.61 (Equation 9.9), which for a χ^2 with one degree of freedom corresponds to a significance value of 0.04 (e.g., $\Pr(\chi_1^2 \geq 4.61) = 0.04$), it follows that one-LOD support intervals approximate 95% confidence intervals under the appropriate settings. However, the one-LOD rule often gives confidence intervals that are too narrow.

The length of the confidence interval is influenced by the number of individuals sampled, the effect of the QTL in question, and the marker density. Precision is not significantly increased by increasing marker density beyond a certain point (around one marker every 5 to 10 cM). Given such a dense map, ML mapping using flanking markers with reasonable sample sizes (200–300 F_2 or backcross individuals) allows a QTL accounting for 5% of the total variance to be mapped to a 40 cM interval, while one accounting for 10% can be mapped to a 20 cM interval.

Interval Mapping with Marker Cofactors

When multiple linked QTLs are present, single marker and interval methods often place QTLs in the wrong location, for example generating a ghost QTL in the position between the two real QTLs. One approach for dealing with multiple QTLs is to modify standard interval mapping to include additional markers as cofactors in the analysis. Using the appropriate unlinked markers can partly account for the segregation variance generated by unlinked QTLs, while the effects of linked QTLs can be reduced by including markers linked to the interval of interest. This general approach of adding marker cofactors to an otherwise standard interval analysis, often referred to as **composite interval mapping** or **CIM**, results in substantial increases in power to detect a QTL and in the precision of estimates of QTL position.

Suppose the interval of interest is flanked by markers i and $i + 1$. One way to incorporate information from additional markers is to consider the sum over some collection of markers outside the interval of interest,

$$\sum_{k \neq i, i+1} b_k \cdot x_{kj} \tag{9.10a}$$

where k denotes a marker locus and j the individual being considered. Letting M_k and m_k denote alternative alleles at the kth marker, the values of the indicator variable x_{kj} depend on the marker genotype of j, with

$$x_{kj} = \begin{cases} 1 & \text{if individual } j \text{ has marker genotype } M_k M_k \\ 0 & \text{if individual } j \text{ has marker genotype } M_k m_k \\ -1 & \text{if individual } j \text{ has marker genotype } m_k m_k \end{cases} \tag{9.10b}$$

This is simply a convenient recoding of a regression of trait value on the number of M_k alleles. Hence, b_k is an estimate of the additive marker effect for locus k. For a backcross design, each marker has only two genotypes and the indicator variable takes on values 1 and -1. More generally, if there is considerable dominance, the effects of the kth marker locus can be more fully accounted for by considering a more complex regression with a term for each genotype, e.g., $b_{k1} x_{k1j} + b_{k2} x_{k2j} + b_{k3} x_{k3j}$, where the indicator variable x_{k1j} is one if j has marker genotype $M_k M_k$, else it is zero. The other two indicator variables for this marker locus are defined accordingly. Composite interval mapping proceeds by adding this regression term to the particular model being considered.

Just which markers should be added? While there is no single solution, the two markers directly flanking the interval being analyzed should always be included. Suppose the interval of interest is delimited by markers i and $i + 1$ (Figure 9.2). Adding markers $i - 1$ and $i + 2$ as cofactors accounts for all linked QTLs to the left of marker $i - 1$ and to the right of marker $i + 2$. Thus, while these cofactors do not account for the effects of linked QTLs in the intervals immediately adjacent to the one of interest (i.e., the intervals $(i - 1, i)$ and $(i + 1, i + 2)$ in Figure 9.2), they do account for all other linked QTLs.
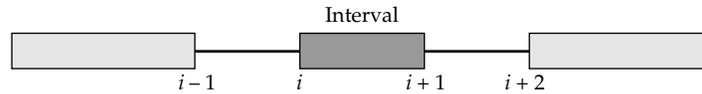
Figure 9.2 Suppose the interval being examined by CIM is between markers i and $i + 1$. Addition of the adjacent markers $i - 1$ and $i + 2$ as cofactors absorbs the effects of any linked QTLs to the left of marker $i - 1$ and to the right of marker $i + 2$. Their inclusion, however, does not remove the effects of QTLs present in the two intervals, $(i - 1, i)$ and $(i + 1, i + 2)$, flanking the interval of interest.

The number of *unlinked* markers that should be used as cofactors is unclear, as inclusion of too many factors greatly reduces power. The number of cofactors should not exceed $2\sqrt{n}$, where n is the number of individuals in the analysis. A first approach would be to include all unlinked markers showing significant marker-trait associations (detected, for example, by standard single-marker regression). If several linked markers from a single chromosome all show significant effects, one might just use the marker having the largest effect. A related strategy is to first perform a multiple regression using all markers unlinked to the region of interest and then eliminate those that are not significant.

Power and Repeatability: The Beavis Effect

Even under designs where power is low, if the number of QTLs is large, it is likely that at least a few will be detected. In such cases of low power, the contributions of detected QTLs can be significantly (often *very* significantly) overestimated. Such a scenario, wherein we detect a small number of QTLs that appear to account for a significant fraction of the total character variation, can lead to the false conclusion that character variation is largely determined by a few QTLs of major effect. Such overestimation is often called a **Beavis effect**, after it was discovered in simulation studies by Beavis (1994).
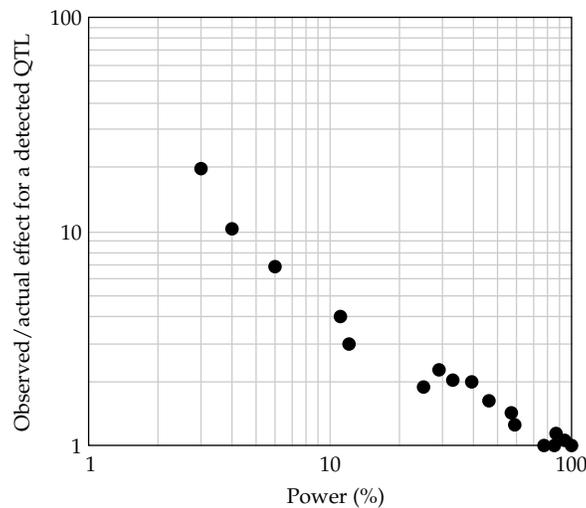

Figure 9.3 Relationship between the probability (power) of detecting a QTL and the amount by which the estimated effect of a *detected* QTL overestimates it actual value. (Based on Beavis 1994.)

As shown in Figure 9.3, the lower the power, the more the effects of a *detected* QTL are overestimated. For example, a QTL accounting for 0.75% of the total F_2 variation has only a 3% chance of being detected with 100 F_2 progeny with markers spaced at 20 cM. However, for cases in which such a QTL is detected, the average estimated total variance it accounts for is 8.4%, a 19-fold overestimate

of the correct value. With 1,000 F_2 progeny, the probability of detecting such a QTL increases to 25%, and each detected QTL on average accounts for approximately 1.5% of the total variance, only a twofold overestimate. Further, these are the *average* values for the estimates. As shown in Figure 9.4, the distribution of observed effects is skewed, with a few loci having large estimated effects, and the rest small to modest effects. Such distributions of effects, commonplace in QTL mapping studies, have usually been taken as being representative of the true distribution of effects. Beavis's simulation studies show that they can be spuriously generated by a set of loci with equal effects.
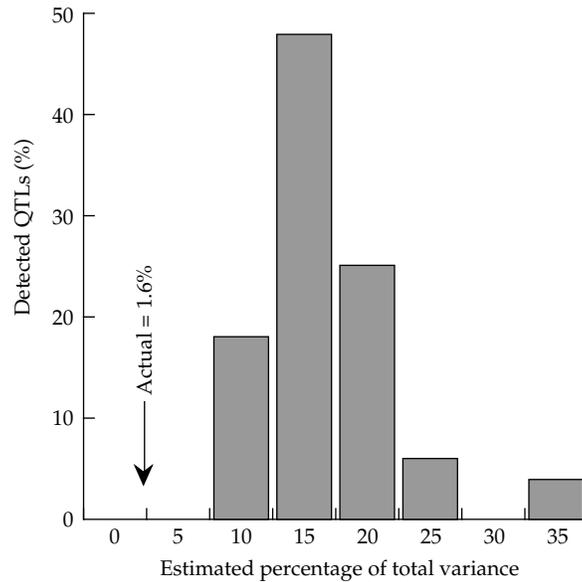

Figure 9.4 Distribution of the estimated effects of detected QTLs. Here 40 QTLs, each accounting for 1.58% of the variance, are assumed. Using 100 F_2 individuals, only 4% of such loci were detected. The average estimated fraction of total variation fraction accounted for by each detected QTLs was 16.3%, with the distribution of estimates skewed towards larger values. (From Beavis 1994.)

MAPPING IN OUTBRED POPULATIONS

The major difference between QTL analysis using inbred-line crosses vs. outbred populations is that while the parents in the former are genetically uniform, parents in the latter are genetically variable. This distinction has several consequences. First, only a fraction of the parents from an outbred population are **informative**. For a parent to provide linkage information, it must be heterozygous at both a marker *and* a linked QTL, as only in this situation can a marker-trait association be generated in the progeny. Only a fraction of random parents from an outbred population are such double heterozygotes. With inbred lines, F_1's are heterozygous at all loci that differ between the crossed lines, so that all parents are fully informative. Second, there are only two alleles segregating at any locus in an inbred-line cross design, while outbred populations can be segregating any number of alleles. Finally, in an outbred population, individuals can differ in marker-QTL linkage phase, so that an *M*-bearing gamete might be associated with QTL allele Q in one parent, and with q in another. Thus, with outbred populations, marker-trait associations must be examined *separately* for each parent. With inbred-line crosses, all F_1 parents have identical genotypes (including linkage phase), so one can simply average marker-trait associations over all offspring, regardless of their parents.

Before considering the variety of QTL mapping methods for outbred populations, some comments on the probability that an outbred family is informative are in order. A parent is **marker-**

informative if it is a marker heterozygote, **QTL-informative** if it is a QTL heterozygote, and simply **informative** if it is both. Unless *both* the marker and QTL are highly polymorphic, most parents will not be informative. Given the need to maximize the fraction of marker-informative parents, classes of marker loci successfully used with inbred lines may not be optimal for outbred populations. For example, historically RFLPs (and more recently SNPs) are widely used in inbred lines, but these markers are typically diallelic and hence have modest polymorphism (at best). Microsatellite (STR) marker loci, on the other hand, are highly polymorphic and hence much more likely to yield marker-informative individuals.

Table 9.1 Types of marker-informative matings.

Fully informative: $\quad M_i M_j \times M_k M_\ell$

Parents are different marker heterozygotes.

All offspring are informative in distinguishing alternative alleles from both parents.

Backcross: $\quad M_i M_j \times M_k M_k$

One parent is a marker heterozygote, the other a marker homozygote.

All offspring informative in distinguishing heterozygous parent's alternative alleles.

Intercross: $\quad M_i M_j \times M_i M_j$

Both parents are the same marker heterozygote.

Only homozygous offspring informative in distinguishing alternative parental alleles.

Note: Here i, j, k, and ℓ index different marker alleles.

As shown in Table 9.1, there are three kinds of marker-informative crosses. With a highly polymorphic marker, it may be possible to examine marker-trait associations for both parents. With a **fully** (marker) **informative family** ($M_i M_j \times M_k M_\ell$) all parental alleles can be distinguished, and both parents can be examined by comparing the trait values in M_i– vs. M_j– offspring and M_k– vs. M_ℓ– offspring. With a **backcross family** ($M_i M_j \times M_k M_k$), only the heterozygous parent can be examined for marker-trait associations. Finally, with an **intercross family** ($M_i M_j \times M_i M_j$), homozygous offspring ($M_i M_i$, $M_j M_j$) are unambiguous as to the origin of parental alleles, while heterozygotes are ambiguous, because allele M_i (M_j) could have come from either parent.

In designing experiments, it is useful to estimate the fraction of families expected to be marker-informative. One measure of this is the **polymorphism information content**, or **PIC**, of the marker locus,

$$\text{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\, p_i^2\, p_j^2 \leq \frac{(n-1)^2 (n+1)}{n^3} \tag{9.11}$$

which is the probability that one parent is a marker heterozygote and its mate has a *different* genotype (i.e., a backcross or fully informative family, but excluding intercross families). In this case, we can distinguish between the alternative marker alleles of the first parent in all offspring from this cross. The upper bound (given by the right hand side of Equation 9.11) occurs when all marker alleles are equally frequent, $p_i = 1/n$.

QTL Mapping Using Sib Families

One can use family data to search for QTLs by comparing offspring carrying alternative marker alleles from the same parent. Consider half-sib, where the basic linear model is a nested ANOVA

(Lecture 5), with marker effects nested within each sibship,

$$z_{ijk} = \mu + s_i + m_{ij} + e_{ijk} \tag{9.12}$$

where z_{ijk} denotes the phenotype of the kth individual of marker genotype j from sibship i, s_i is the effect of sire i, m_{ij} is the effect of marker genotype j in sibship i (typically, $j = 1, 2$ for the alternative sire marker alleles), and e_{ijk} is the within-marker, within-sibship residual. It is assumed that s, m, and e have means equal to zero, are uncorrelated, and are normally distributed with variances σ_s^2 (the between-sire variance), σ_m^2 (the between-marker, within-sibship variance), and σ_e^2 (the residual or within-marker, within-sibship variance). A significant marker variance indicates linkage to a segregating QTL, and is tested by using the statistic

$$F = \frac{\mathrm{MS}_m}{\mathrm{MS}_e} \tag{9.13}$$

where the mean squares are similiar to those in Table 5.3. Assuming normality, Equation 9.13 follows an F distribution under the null hypothesis that $\sigma_m^2 = 0$. Assuming a balanced design with N sires, each with $n/2$ half-sibs in each marker class, Equation 9.13 has N and $N(n-2)$ degrees of freedom.

Again referring to Table 5.3, we see that for a balanced design the mean squares have expected values of

$$E(\mathrm{MS}_m) = \sigma_e^2 + (n/2)\sigma_m^2 \qquad \text{and} \qquad E(\mathrm{MS}_e) = \sigma_e^2 \tag{9.14}$$

where

$$\sigma_m^2 = \frac{E(\mathrm{MS}_m) - E(\mathrm{MS}_e)}{n/2} = (1 - 2c)^2 \frac{\sigma_A^2}{2} \tag{9.15}$$

Thus, an estimate of the QTL effect (measured by its additive variance σ_A^2, scaled by the distance c between QTL and marker), can be obtained from the observed mean squares.

One immediate drawback of measuring a QTL effect by its variance in an outbred population is that even a completely linked QTL with a large effect can nonetheless have a small σ_m^2. Consider a strictly additive diallelic QTL with allele frequency p, where $\sigma_A^2 = 2a^2 p(1 - p)$. Even if a is large, the additive genetic variance can still be quite small if the QTL allele frequencies are near zero or one. An alternative way of visualizing this relationship is to note that the probability of a QTL-informative sire is $2p(1 - p)$. If this is small, even if a is large, σ_A^2 will be small, as most families will not be informative. In those rare informative families, however, the between-marker effect is large. Contrast this to the situation with inbred-line crosses, where the QTL effect estimates $2a$, since here all families are informative, rather that the fraction $2p(1 - p)$ seen an outbred population.

One approach for increasing power is the **granddaughter design** (Weller et al. 1990), under which each sire produces a number of sons that are genotyped for sire marker alleles (Figure 9.5), and the trait values for each son are taken to be the mean value of the traits in offspring from the son (rather than the direct measures of the son itself). This design was developed for milk-production characters in dairy cows, where the offspring are granddaughters of the original sires. The linear model for this design is

$$z_{ijk\ell} = \mu + g_i + m_{ij} + s_{ijk} + e_{ijk\ell} \tag{9.16}$$

where g_i is the effect of grandsire i, m_{ij} is the effect of marker allele j ($= 1, 2$) from the ith sire, s_{ijk} is the effect of son k carrying marker allele j from sire i, and $e_{ijk\ell}$ is the residual for the ℓth offspring of this son. Sire marker-allele effects are halved by considering granddaughters (as opposed to daughters), as there is only a 50% chance that the grandsire allele is passed from its son onto its granddaughter. However, this reduction in the expected marker contrast is usually more than countered by the smaller standard error associated with each contrast due to the large number of offspring used to estimate trait value.
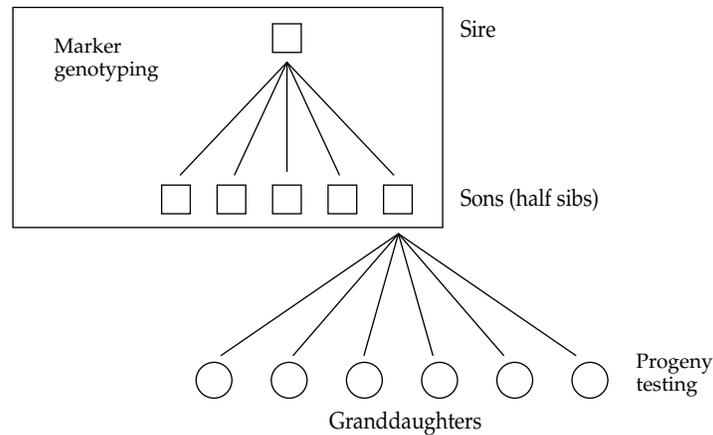
Figure 9.5 The granddaughter design of Weller et al. (1990). Here, each sire produces a number of half-sib sons that are scored for the marker genotypes. The character value for each son is determined by progeny testing, with the trait value being scored in a large number of daughters (again half-sibs) from each son.

General Pedigree Methods

Likelihood models can also easily be developed for QTL mapping using sib families. These explicitly model the transmission of QTL genotypes from parent to offspring, requiring estimation of QTL allele frequencies and genotype means (as well as assumptions about the number of segregating alleles). While this approach can be extended to multi-generational pedigrees, the number of possible combinations of genotypes for individuals in the entire pedigree increases exponentially with the number of pedigree members, and solving the resulting likelihood functions becomes increasingly more difficult. An alternative is to construct likelihood functions using the **variance components** associated with a QTL (or linked group of QTLs) in a genetic region of interest, rather than explicitly modeling all of the underlying genetic details. This approach allows for very general and complex pedigrees. The basic idea is to use marker information to compute the fraction of a genetic region of interest that is identical by descent between two individuals. Recall that two alleles are identical by descent, or **ibd**, if we can trace them back to a single copy in a common ancestor (Lecture 4). This is a *random-effects* approach, estimating a single quantity (the variance associated with the marker) versus a fixed- effects approach were we would have to estimate the frequency of each QTL genotype and their means (and variances).

Consider the simplest case, in which the genetic variance is additive for the QTLs in the region of interest as well as for background QTLs unlinked to this region. Under this model, an individual's phenotypic value is decomposed as

$$z_i = \mu + A_i + A_i^* + e_i \tag{9.17}$$

where μ is the population mean, A is the contribution from the chromosomal interval being examined, A^* is the contribution from QTLs outside this interval, and e is the residual. The random effects A, A^*, and e are assumed to be normally distributed with mean zero and variances σ_A^2, $\sigma_{A^*}^2$, and σ_e^2. Here σ_A^2 and $\sigma_{A^*}^2$ correspond to the additive variances associated with the chromosomal region of interest and background QTLs in the remaining genome, respectively. We assume that none of these background QTLs are linked to the chromosome region of interest so that A and A^* are uncorrelated, and we further assume that the residual e is uncorrelated with A and A^*. Under these assumptions, the phenotypic variance is $\sigma_A^2 + \sigma_{A^*}^2 + \sigma_e^2$.

Assuming no shared environmental effects, the phenotypic covariance between two individuals

is

$$\sigma(z_i, z_j) = R_{ij}\,\sigma_A^2 + 2\Theta_{ij}\,\sigma_{A^*}^2 \qquad (9.18)$$

where R_{ij} is the fraction of the chromosomal region shared ibd between individuals i and j, and $2\Theta_{ij}$ is twice Wright's coefficient of coancestry (i.e., $2\Theta_{ij} = 1/2$ for full sibs, see Lecture 4). For a vector $\mathbf{z}$ of observations on n individuals, the associated covariance matrix $\mathbf{V}$ can be expressed as contributions from the region of interest, from background QTLs, and from residual effects,

$$\mathbf{V} = \mathbf{R}\,\sigma_A^2 + \mathbf{A}\,\sigma_{A^*}^2 + \mathbf{I}\,\sigma_e^2 \qquad (9.19a)$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{R}$ and $\mathbf{A}$ are matrices of known constants,

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases} \qquad (9.19b)$$

The elements of $\mathbf{R}$ contain the estimates of ibd status for the region of interest based on marker information, while the elements of $\mathbf{A}$ are given by the pedigree structure.

The resulting likelihood is a multivariate normal with mean vector $\boldsymbol{\mu}$ (all of whose elements are μ) and variance-covariance matrix $\mathbf{V}$,

$$\ell(\,\mathbf{z}\,|\,\mu, \sigma_A^2, \sigma_{A^*}^2, \sigma_e^2\,) = \frac{1}{\sqrt{(2\pi)^n\,|\mathbf{V}|}}\,\exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T\,\mathbf{V}^{-1}\,(\mathbf{z} - \boldsymbol{\mu})\right] \qquad (9.20)$$

This likelihood has four unknown parameters (μ, σ_A^2, $\sigma_{A^*}^2$, and σ_e^2). A significant σ_A^2 indicates the presence of at least one QTL in the interval being considered, while a significant $\sigma_{A^*}^2$ implies background genetic variance contributed from QTLs outside the focal interval. Both of these hypotheses can be tested by likelihood-ratio tests (using $\sigma_A^2 = 0$ and $\sigma_{A^*}^2 = 0$, respectively).

Haseman-Elston Regressions

Starting with Haseman and Elston (1972), human geneticists have developed a number of methods for detecting QTLs using pairs of relatives as the unit of analysis. The idea is to consider the number of alleles identical by descent (ibd) between individuals for a given marker. If a QTL is linked to the marker, pairs sharing ibd marker alleles should also tend to share ibd QTL alleles and hence are expected to be more similar than pairs not sharing ibd marker alleles. This fairly simple idea is the basis for a large number of relative-pair methods (often referred to as **allele sharing** methods).

Haseman and Elston regress (for each marker) the squared difference $Y_i = (z_{i1} - z_{i2})^2$ in trait value in two relatives on the proportion π_{im} of alleles ibd at the marker of interest,

$$Y_i = a + \beta\pi_{im} + e \qquad (9.21)$$

Here the slope β and intercept a depend on the type of relatives and the recombination fraction c. For full sibs,

$$\beta = -2\,(1 - 2c)^2\,\sigma_A^2 \qquad (9.22)$$

A significant negative slope provides evidence of a QTL linked to the marker, with the power of this test scaling with $(1 - 2c)^2$ and σ_A^2. The expected slopes for other pairs of relatives are

$$\beta = \begin{cases} -2\,(1 - 2c)\,\sigma_A^2 & \text{grandparent–grandchild;} \\ -2\,(1 - 2c)^2\,\sigma_A^2 & \text{half-sibs;} \\ -2\,(1 - 2c)^2\,(1 - c)\,\sigma_A^2 & \text{avuncular (aunt/uncle–nephew/niece).} \end{cases} \qquad (9.23)$$

The Haseman-Elston test is quite simple: for n pairs of the same type of relatives, one regresses the squared difference of each pair on the fraction of alleles ibd at the marker locus. A significant negative slope for the resulting regression indicates linkage to a QTL. This is a one-sided test, as the null hypothesis (no linkage) is $\beta = 0$ versus the alternative $\beta < 0$.

There are several caveats with this approach. First, different types of relatives cannot be mixed in the standard H-E test, requiring separate regressions for each type of relative pair. This procedure can be avoided by modifying the test by using an appropriately weighted multiple regression. Second, parents and their offspring share *exactly* one allele ibd and hence cannot be used to estimate this regression, as there is no variability in the predictor variable π. Finally, QTL position (c) and effect (σ_A^2) are confounded and cannot be separately estimated from the regression slope β. Thus, in its simplest form, the H-E method is a *detection* test rather than an *estimation* procedure. This conclusion is not surprising, given that the H-E method is closely related to the single-marker linear model. Estimation of both position (c) and and effect (σ_A^2) is possible by extending the H-E regression by using ibd status of two (or more) linked marker loci to estimate π_{jt}.

Affected Sib Pair Methods

When dealing with a dichotomous (i.e., presence/absence) character, pairs of relatives can be classified into three groups: pairs where both are normal, **singly affected** pairs with one affected and one normal member, and **doubly affected** pairs. The first and last pairs are also called **concordant**, while pairs that differ are called **discordant**. The motivation behind relative-pair tests is that if a marker is linked to a QTL influencing the trait, concordant and discordant pairs should have different distributions for the number of ibd marker alleles.

In addition to being much more robust than ML methods for dichotomous characters, relative-pair tests also have the advantage of selective genotyping in that pairs are usually chosen so that at least one member is affected. The pairs of relatives considered are usually full sibs, and a number of variants of these **affected sib-pair**, or ASP, methods have been proposed. Most of these are detection tests, rather than estimation procedures, as they cannot provide seperate estimates of QTL effect and position. While our attention focuses on full-sib pairs, this basic approach can easily be applied to any pair of relatives, *provided* there is variability in the number of ibd alleles. (This excludes parent-offspring pairs, as these share exactly one allele ibd.) Most affected sib-pair tests have the basic structure of comparing the observed ibd frequencies (or some statistic based on them) of doubly affected pairs with either their expected values under no linkage or with the corresponding values in singly affected pairs. There are many possible tests based on this idea and most, it seems, have made their way into the literature. We consider three here.

Among those n_i pairs with i affected members ($i = 0, 1, 2$), let p_{ij} denote the frequency of such pairs with j ibd marker alleles ($j = 0, 1, 2$). From the binomial distribution, the estimator $\widehat{p}_{ij}$ has mean p_{ij} and variance $p_{ij}(1-p_{ij})/n_i$. One ASP test is based on $\widehat{p}_{22}$, the observed frequency of doubly affected pairs that have two marker alleles ibd. Under the assumption of no linkage, $\widehat{p}_{22}$ has mean $1/4$ (as full sibs have a 25% chance of sharing both alleles ibd) and variance $(1/4)(1-1/4)/n_2 = 3/(16\,n_2)$, suggesting the test

$$T_2 = \frac{\widehat{p}_{22} - 1/4}{\sqrt{\dfrac{3}{16\,n_2}}} \tag{9.24a}$$

For a large number of doubly affected pairs, T_2 is approximately distributed as a unit normal under the null hypothesis of no linkage. This test is one-sided, as $p_{22} > 1/4$ under linkage.

An alternative approach is to consider statistics that employ the mean number of ibd marker alleles, $p_{i1} + 2\,p_{i2}$. Under the hypothesis of no linkage, this has expected value $1 \cdot (1/2) + 2 \cdot (1/4) = 1$ and variance $[\,1^2 \cdot (1/2) + 2^2 \cdot (1/4)\,] - 1^2 = 1/2$. For doubly affected pairs, the test statistic becomes

$$T_m = \sqrt{2\,n_2}\,(\,\widehat{p}_{21} + 2\,\widehat{p}_{22} - 1\,) \tag{9.24b}$$

which again for large samples is approximately distributed as a unit normal and is a one-sided test, as $p_{21} + 2\,p_{22} > 1$ under linkage.

Finally, maximum likelihood-based goodness-of-fit tests can be used (Risch 1990a,b). In keeping with the tradition of human geneticists, ML-based tests usually report LOD (likelihood of odds) scores in place of the closely related likelihood ratio (LR). (Recall from Equation 9.9 that 1 LR = 4.61 LOD.) Here the data are n_{20}, n_{21}, and n_{22}, the number of doubly affected sibs sharing zero, one, or two marker alleles ibd, with the unknown parameters to estimate being the population frequencies of these classes (p_{20}, p_{21}, p_{22}). The MLEs for these population frequencies are given by $\widehat{p}_{2i} = n_{2i}/n_2$. The LOD score for the test of no linkage becomes

$$ MLS = \log_{10}\left[\prod_{i=0}^{2} \left(\frac{\widehat{p}_{2i}}{\pi_{2i}} \right)^{n_{2i}} \right] = \sum_{i=0}^{2} n_{2i} \, \log_{10}\left(\frac{\widehat{p}_{2i}}{\pi_{2i}} \right) \tag{9.25} $$

where π_{2i} is the probability that the pair of doubly affected sibs shares i alleles ibd in the absence of linkage to a QTL. (For full sibs, $\pi_{20} = \pi_{22} = 1/4$, $\pi_{21} = 1/2$.) The test statistic given by Equation 9.25 is referred to as the **maximum LOD score**, or **MLS**, with a score exceeding three being taken as significant evidence for linkage (Risch 1990a, Morton 1955).

An alternative formulation for the MLS test is to consider each informative parent separately, simply scoring whether or not a doubly affected sib pair shares a marker allele from this parent. This approach generates 0 (match, both affected sibs share the allele) or 1 (no match) ibd data. Under the null hypothesis of no linkage, each state (0 or 1) has probability 1/2, and the MLS test statistic becomes

$$ MLS = (1 - n_1) \log_{10}\left(\frac{1 - \widehat{p}_1}{1/2} \right) + n_1 \log_{10}\left(\frac{\widehat{p}_1}{1/2} \right) \tag{9.26} $$

where n_1 and p_1 are, respectively, the number and frequency of sibs sharing the parental allele. This method has the advantage that sibs informative for only one parental marker can still be used. Using this approach, Davies et al. (1994) did a genome-wide search (also commonly called a **genomic scan**) for markers linked to DS genes influencing human type 1 diabetes. Among doubly affected sibs, one marker on chromosome 6, *D6S273*, had 92 pairs sharing parental alleles and 31 pairs not sharing parental alleles. A second marker on the opposite end of this chromosome, *D6S415*, had 74 pairs sharing parental alleles and 60 not sharing alleles. The MLS scores for these two markers are

$$ MLS(D6S273) = 31 \cdot \log_{10}\left(\frac{2 \cdot 31}{123} \right) + 92 \cdot \log_{10}\left(\frac{2 \cdot 92}{123} \right) = 6.87 $$

$$ MLS(D6S415) = 60 \cdot \log_{10}\left(\frac{2 \cdot 60}{134} \right) + 74 \cdot \log_{10}\left(\frac{2 \cdot 74}{134} \right) = 0.32 $$

Thus, the first marker shows significant evidence of linkage, while the second does not. Translating these LOD scores into LR values (the latter being distributed as a χ^2 with one degree of freedom) gives LR $= 4.61 \cdot 6.87 = 31.6$ ($P < 0.001$) for *D6S273* and LR $= 4.61 \cdot 0.32 = 1.47$ ($P = 0.2$) for *D6S415*.

Lecture 17 Problems

1. Suppose you observe the following marker means in an F_2 population:

Marker	Trait Mean
MM	10.5
Mm	12.5
mm	16.2

    a: Suppose your sample size is large enough that the standard error on these means is 0.25. Is there evidence of a QTL linked to this marker?

    b: What can you say about a and k?

    c: Suppose you have markers spaced every 10 centimorgans. What now can you say about a and k?

2.   Consider an outbred population. The allele frequencies at a marker are freq(M) = 0.3 and freq(M) = 0.7, while the allele frequencies at a QTL are freq(Q) = 0.1 and freq(q) = 0.9.

    a: What is the probability that a random invidual is marker-informative?

    b: What is the probability that a random invidual is QTL-informative?

    c: What is the probability that a random invidual is informative (assume, in the population, that the marker and QTL are in linkage equilibriumm)?

    d: How many individuals do you need to sample to have a 90% chance that at least one is informative?

Solutions to Lecture 17 Problems

1. Yes. The marker means are significantly different.

    b:
    $$(\mu_{MM} - \mu_{mm})/2 = (16.2 - 10.5)/2 = 2.85 = a(1 - 2c)$$

Hence, $a \geq 2.854$

$$\frac{\mu_{Mm} - (\mu_{MM} + \mu_{mm})/2}{(\mu_{MM} - \mu_{mm})/2} = \frac{13.35 - 12.5}{2.85} = 0.30 = k(1 - 2c)$$

Thus $k \geq 0.30$

   c: If markers are 10cM apart, them the QTL is no further than 5cM from the marker with greatest effect. hence, $(1 - 2c) = 0.9$ and thus $2.85 \leq a \leq 2.85/.9 = 3.2$. Likewise, $0.3 \leq k \leq 0.33$

2. a. $2 \cdot 0.3 \cdot 0.7 = 0.42$

    b. $2 \cdot 0.1 \cdot 0.9 = 0.18$

    c. $0.42 \cdot 0.18 = 0.0756$

    d. Prob(Not informative) = $1 - 0.0756 = 0.9244$. Prob(n individuals all not informative = 0.9244^n. Prob(at least one informative individual) = $1 - 0.9244^n$. Solve form n in $1 - 0.9244^n = 0.9$, or $0.9244^n = 0.1$, or

    $$n = \log(0.1)/\log(0.9244) = 29.2$$