Lecture 8

QTL Mapping 1: Overview and Using Inbred Lines

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.
Notes from a short course taught May 2011 at University of Liege

While the machinery of quantitative genetics can blissfully function in complete ignorance of any of the underlying genetic details, we are certainly interested (at least one some level) on the genetic basis of trait variation. At the most practical level, if a gene of major effect is segregating in our population of interest, we would certainly like to not only to be aware of this, but also to fine map it for future exploration/exploitation. Often no single locus contributes more than a fraction of the total genetic variation in a trait. In such cases, we often use the term **polygene** for one of these loci of small effect. The (still evolving) more modern use is to call these genes **quantitative trait loci**, or simply **QTL** or **QTLs** for short.

We start by considering QTL detection using 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 (Figure 8.1). 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 (**AICs, 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. There are two variations based on inbreeding the F_1s. The most dramatic are **DHs, doubled-haploid lines** wherein the haploid gametes from F_1 individuals are duplicated, instantly creating a fully inbred line. Likewise, **RILS, recombinant inbred lines** are obtained by inbreeding the F_1s.

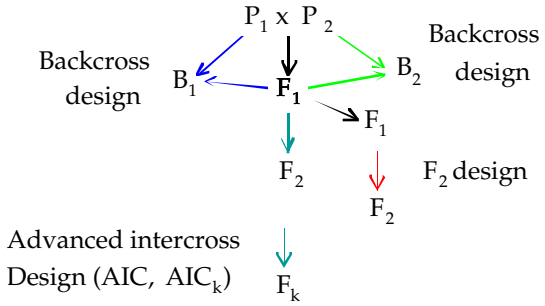

Figure 2.1. Various designs for QTL mapping starting with the F_1 between two lines. Backcross designs cross the F_1 back to either parent, F_2 designs cross the F_1s with each other, advanced intercross lines continue to intermate the F_2 for a number of generations.

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,

$$\Pr(Q_k \mid M_j) = \frac{\Pr(Q_k M_j)}{\Pr(M_j)} \tag{8.1}$$

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 5). 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, $c_{12} = c_1 + c_2 - 2c_1c_2$, 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 8.1 gives the F_2 conditional probabilities as

$$\begin{aligned}
&\Pr(QQ\,|\,MM\,) = (1-c)^2, \quad \Pr(Qq\,|\,MM\,) = 2c(1-c), \quad \Pr(qq\,|\,MM\,) = c^2 \\
&\Pr(QQ\,|\,Mm\,) = c(1-c),\ \Pr(Qq\,|\,Mm\,) = (1-c)^2 + c^2,\ \Pr(qq\,|\,Mm\,) = c(1-c) \qquad (8.2) \\
&\Pr(QQ\,|\,mm\,) = c^2, \quad \Pr(Qq\,|\,mm\,) = 2c(1-c), \quad \Pr(qq\,|\,mm\,) = (1-c)^2
\end{aligned}$$

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_1QM_2 and m_1qm_2. What are the conditional probabilities of the three QTL genotypes when the marker genotype is $M_1M_1M_2M_2$? Since all F_1s are M_1QM_2/m_1qm_2, under the assumptions of no interference, the frequency of F_1 gametes involving M_1M_2 are

$$\Pr(M_1QM_2) = (1-c_1)(1-c_2)/2, \qquad \Pr(M_1qM_2) = c_1\,c_2/2$$

giving expected frequencies in the F_2 of $M_1M_1M_2M_2$ offspring as

$$\begin{aligned}
\Pr(M_1QM_2/M_1QM_2) &= [\,(1-c_1)(1-c_2)/2\,]^2 \\
\Pr(M_1QM_2/M_1qM_2) &= 2\,[\,(1-c_1)(1-c_2)/2\,]\,[\,c_1\,c_2/2\,] \\
\Pr(M_1qM_2/M_1qM_2) &= (\,c_1\,c_2/2\,)^2
\end{aligned}$$

where $c_2 = (c_{12} - c_1)/(1-2c_1)$. The overall frequency of $M_1M_1M_2M_2$ individuals is the sum of the three above terms, or $(1-c_{12})^2/4$. Substituting into Equation 8.1 gives

$$\begin{aligned}
\Pr(QQ\,|\,M_1M_1M_2M_2) &= \frac{(1-c_1)^2(1-c_2)^2}{(1-c_{12})^2} \\
\Pr(Qq\,|\,M_1M_1M_2M_2) &= \frac{2c_1c_2(1-c_1)(1-c_2)}{(1-c_{12})^2} \qquad (8.3) \\
\Pr(qq\,|\,M_1M_1M_2M_2) &= \frac{c_1^2\,c_2^2}{(1-c_{12})^2}
\end{aligned}$$

Conditional probabilities for other marker genotypes are computed in a similar fashion. Since c_1c_2 is usually very small if c_{12} is moderate to small, essentially all $M_1M_1M_2M_2$ individuals are QQ. For example, assuming $c_1 = c_2 = c_{12}/2$ (the worst case), the conditional probabilities of an $M_1M_1M_2M_2$ 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

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(\,Q_k\,|\,M_j\,) \qquad (8.4)$$

The QTL effects enter through the μ_{Q_k}, while the QTL positions enter through the conditional probabilities $\Pr(\, Q_k \,|\, M_j \,)$. For example, if a QTL with effects $2a : a(1+k) : 0$ is linked (distance c) to a marker, applying Equation 8.2, the F_2 marker means become

$$(\mu_{MM} - \mu_{mm})/2 = a(1-2c) \tag{8.5a}$$

$$\frac{\mu_{Mm} - (\mu_{MM} + \mu_{mm})/2}{(\mu_{MM} - \mu_{mm})/2} = k(1-2c) \tag{8.5b}$$

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 equally 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 8.3 (and its analog for $m_1m_1m_2m_2$ probabilities) implies

$$\begin{aligned} \frac{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}{2} &= a\left(\frac{1-c_1-c_2}{1-c_1-c_2+2c_1\,c_2}\right) \\ &\simeq a\,(1-2c_1\,c_2) \end{aligned} \tag{8.6a}$$

where c_1 is the M_1-QTL recombination frequency. Equation 8.6a 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 8.5a that $\mu_{M_1M_1} - \mu_{m_1m_1} = 2a(1-2c_1)$, we can obtain estimates of the recombination frequencies by substituting Equation 8.6a for a and rearranging to give

$$\begin{aligned} c_1 &= \frac{1}{2}\left(1 - \frac{\mu_{M_1M_1} - \mu_{m_1m_1}}{2a}\right) \\ &\simeq \frac{1}{2}\left(1 - \frac{\mu_{M_1M_1} - \mu_{m_1m_1}}{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}\right) \end{aligned} \tag{8.6b}$$

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},

$$z_{ik} = \mu + b_i + e_{ik} \tag{8.7a}$$

This is a one-way ANOVA model, 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

$$z_j = \mu + \sum_{i=1}^{n} b_i\, x_{ij} + e_j \tag{8.7b}$$

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 8.7a,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 8.5b). 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

$$z_j = \mu + \sum_{i}^{n_1} a_i \, x_{ij} + \sum_{k}^{n_2} b_k \, y_{kj} + \sum_{i}^{n_1} \sum_{k}^{n_2} d_{ik} \, x_{ij} \, y_{kj} + e_j \tag{8.7c}$$

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

$$\ell(z \mid M_j) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2) \, \Pr(Q_k \mid M_j) \tag{8.8}$$

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. 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,

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

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.

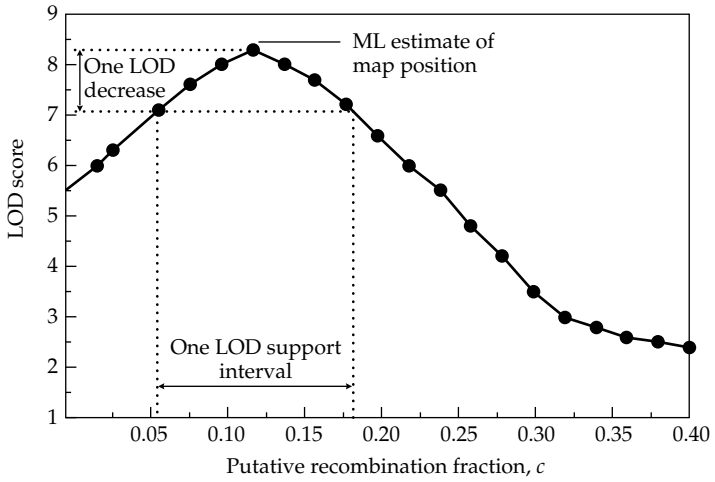

Figure 8.21 Hypothetical likelihood map for the marker-QTL recombination frequency c in a single-marker analysis. Points connected by straight lines are used to remind the reader that likelihood maps are computed by plotting the maximum of the likelihood function for each c value, usually done by considering 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 amount of support for a QTL at a particular map position is often displayed graphically through the use of **likelihood maps** (Figure 8.2), 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 1955b). The LOD score for a particular value of c is related to the likelihood-ratio test statistic (LR) by

$$\text{LOD}(c) = -\log_{10} \left[\frac{\max\ \ell_r(\mathbf{z})}{\max\ \ell(\mathbf{z}, c)} \right] = \frac{\text{LR}(c)}{2 \ln 10} \simeq \frac{\text{LR}(c)}{4.61} \tag{8.9}$$

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.

The likelihood map projects the multidimensional likelihood surface (which is a function of the QTL means, variance, and map position) onto 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.

Permutation Tests: Finding the Significance Threshold

How is the significance threshold obtained in a ML analysis (or other complex analysis)? A very general approach is the **permutation test**. Here, one imagines each individual as having two values: a phenotypic value for the trait of interest and a vector of marker genotypes. The permutation test keeps the vector of marker information for each individual intact (thereby preserving the covariance structure between markers) by randomly shuffling the trait values over these marker vectors. This creates a data set with zero marker-trait associations, and one runs the analysis on such a data set, recording the largest test statistic. Several hundreds to thousands of such shuffling are used, generating an empirical distribution of the test statistic under the null hypotheses. For example, suppose that 95% of all such values are below a value of 10 for our test statistics, and 99% are below a value of 17. The resulting 5% and 1% significance levels are just 10 and 17, respectively.

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. Approximate confidence intervals are often constructed using the **one-LOD rule** (Figure 8.2), 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 8.9), which for a χ^2 with one degree of freedom corresponds to a significance value of 0.04 (e.g., $\Pr(\chi^2_1 \geq 4.61) = 0.04$), it follows that one-LOD support intervals approximate 95% confidence intervals under the appropriate settings. However, while widely used in QTL mapping, the one-LOD rule typically gives confidence intervals that are too narrow, and a 1.5 to 2-LOD support interval is a better choice.

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.

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. Darvasi

and Soller (1997) used simulations to obtain approximate expressions for the expected length of a 95% confidence interval under an F_2 design. If N individuals are scored, and our QTL of interest accounts for a fraction ν of the total phenotypic variance, then the expected approximate length of the confidence interval (in centimorgans) is $530/(N\,\nu)$, so that to map a QTL to a 1cM region (the minimal size to start positional cloning) requires a sample size of $N = 530/\nu$, or 2100, 5300, 10600, and 53,000 for QTL that account for 25%, 10%, 5% and 1% (respectively) of the total variation.

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,

$$\sum_{k \neq i, i+1} b_k \cdot x_{kj} \tag{8.10a}$$

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

$$x_{kj} = \begin{cases} 1 & \text{if individual } j \text{ has marker genotype } M_kM_k \\ 0 & \text{if individual } j \text{ has marker genotype } M_km_k \\ -1 & \text{if individual } j \text{ has marker genotype } m_km_k \end{cases} \tag{8.10b}$$

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_kM_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.

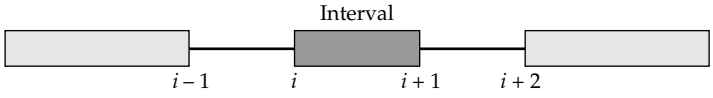

Figure 8.3 Suppose the interval being examined by CIM is between markers i and $i+1$. Addition of the adjacent markers $i-1$ and $i+2$ as cofactors absorbs the effects of any linked QTLs to the left of marker $i-1$ and to the right of marker $i+2$. Their inclusion, however, does not remove the effects of QTLs present in the two intervals, $(i-1, i)$ and $(i+1, i+2)$, flanking the interval of interest.

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 8.3). 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 8.3), they do account for all other linked QTLs.

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 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 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).

As shown in Figure 8.4, 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 8.5, 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.

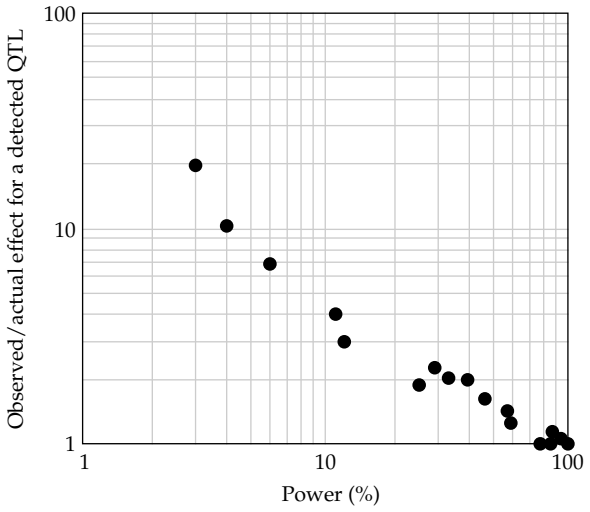

Figure 8.4 Relationship between the probability (power) of detecting a QTL and the amount by which the estimated effect of a *detected* QTL overestimates it actual value. (Based on Beavis 1994.)

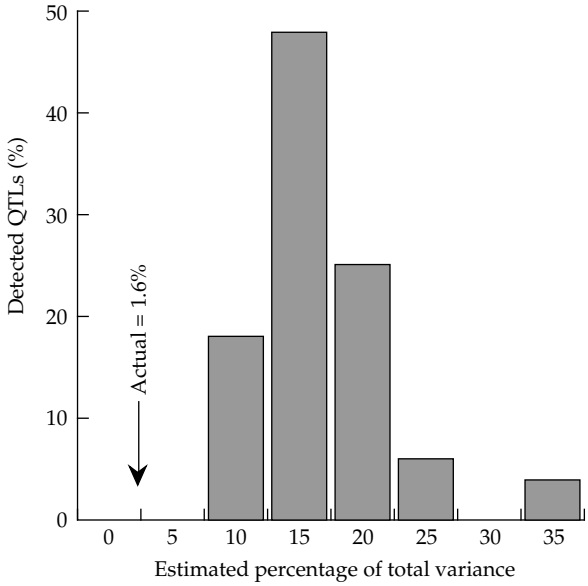

Figure 8.5 Distribution of the estimated effects of detected QTLs. Here 40 QTLs, each accounting for 1.58% of the variance, are assumed. Using 100 F_2 individuals, only 4% of such loci were detected. The average estimated fraction of total variation fraction accounted for by each detected QTLs was 16.3%, with the distribution of estimates skewed towards larger values. (From Beavis 1994.)

Dealing with Supersaturated Models: Model Selection

Consider the simple linear model allowing for epistasis (Equation 8.7c). Note that there are a very large number of potential terms to estimate. For example, Xu and Jia (2007, Genetics 175: 1955) searched for QTLs in 145 doubled-haploid lines of barley using $m = 127$ markers. The full linear model allowing for both main effects and pair-wise epistasis requires estimation of 127 marker-trait effects but $m(m-1)/2 = 8001$ potential epistatic effects. This is an example of a **supersaturated model**, with far more potential parameters to estimate than there are observations. One approach for dealing with such situations is to use **model selection**. Here one uses some optimality criteria and then searches over all possible models, either by having a computer run through all possible models (which is a huge space) or using probabilistic model-sampling approaches such as **stochastic search variable selection**. For each model examined (for example a model with, say, main effects for markers 1, 3, 5, 7 and an interaction between markers 3 and 6) one computes an optimality measure that rewards for better model fit while penalizing for the number of fitted parameters (the same fit with fewer parameters is always better), all the while adjusting for the sample size. A number of such model selection criteria have been proposed, AIC and BIC being the two most common that appear in the literature.

• **AIC, Akaike information criterion**: $AIC = 2p - 2\ln(L)$, where p are the number of fitted parameters and L is the value of the likelihood function at the solution. Assuming our model has normally-distributed residuals, $AIC = 2p + n(\ln[2\pi RSS/n] + 1)$ where RSS is the sum of squared residual errors and n is the number of observations.

• **BIC, Bayesian information criterion**: $BIC = p\ln(n) - 2\ln(L)$. For normally-distributed residuals, $BIC = p\ln(n) + n\ln(RSS/n)$. BIC is also referred to as the **Schwarz information criterion** (SIC).

The model with the lowest value of AIC or BIC is the "best" model, although it needs to be stressed that both are simply weighted goodness-of-fit measures, *not* formal statistical tests. Both AIC and BIC offer the same reward for model fit (higher likelihood L of the model implies a better fit and also a

lower AIC/BIC value). They differ in the penalty imposed by the number of fitted parameters, with BIC giving more parameters a greater penalty, one that increases with the number of observations. Both approaches have their advocates. My recommendation is to use both, and see how similar the best models are under these different criteria.

Bayesian Approaches

Bayesian statistics offers other approaches for dealing with both supersaturated models as well as models with high complexity. VERY briefly, the idea behind Bayesian statistics (as opposed the "frequentist" or "classical" statistics) is that what Bayesian statistics returns the full *distribution* of the possible values of a parameter given the data as opposed to a simple point estimate.

Suppose we have a data vector x and wish to estimate the mean μ. Classic theory returns a **point estimate** (say) 5 for the mean, as well as presenting some measure of uncertainty (or confidence) for this estimate. A bayesian statistician would take some **prior** information on the mean $p(\mu)$ together with the likelihood for the mean given the data $L(x \mid \mu)$ and compute a **posterior** distribution $p(\mu \mid x) = C * L(x \mid \mu) * p(\mu)$, the product of the likelihood function times the prior (C is a constant ensuring the posterior integrates to one, and hence is a proper probability distribution). In essence, Bayesian analysis in an extension of likelihood analysis. It has several advantages. First, (assuming the prior is correct), it is *exact* for any sample size, as compared to the large-sample approximations required by likelihood analysis. Second, it surprisingly turns out that what appears to be a much more complicated approach is actually often computationally easier than likelihood for very complex models. Why? The key is **Markov chain Monte Carlo**, **MCMC**, methods that allow one to take a complex posterior and fairly easily generate random samples from this distribution. One can then have their favorable computer draw millions of samples, essentially giving us the complete distribution. Further, this method is often efficient even over very high dimensional spaces.

We briefly mention two powerful uses of Bayesian methods for QTL mapping: **model averaging** and **shrinkage estimates**. Suppose we are trying to map QTLs. As we change the number of QTLs assumed in our data, the estimates of their effects and positions change. One could deal with this with model selection using AIC/BIC to find the "best" model. Bayesian model averaging replaces this "point" estimate of the best model with a weighted average over all possible models drawn from the posterior distribution. The weight is based on the likelihood of the model. For example, it might be that the AIC value for a model with 4 QTLs is very slightly smaller than the AIC value for a model with 5 QTLs. Model averaging would weight the relative strength of the 4 versus 5 (and other) models, presenting a final average over all models.

Shrinkage estimates are another approach for dealing with supersaturated models (such as the hunt for epistatic interactions). Rather than adding interaction terms one at a time, a shrinkage method starts with all interactions included, and then shrinks most back to zero. Since under a Bayesian analysis, any effect is random (i.e., there are no fixed effects – constants to be estimated – but rather everything is a random variable), one can assume the effect for (say) interaction ij is drawn from a normal distribution with mean zero and variance σ^2_{ij}. Further, the interaction-specific variances are themselves random variables drawn from a hyperparameter distribution, such as an inverse chi-square. One then estimates the hyperparameters and uses these to predict the variances, with effects with small variances shrinking back to zero, and effects with large variances remaining in the model. Sounds complicated, but generally it works well.

Lecture 8 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. Recall for a single-marker analysis that $E[MM - mm] = 2a(1 - 2c)$ and suppose the trait variance in each marker class mean is σ^2/n where n is the number of individuals in the marker class. Assuming a standard two-sided normal test for a null hypothesis of no mean difference, the sample size n (number of individuals in each marker class) to have power $1 - \beta$ using a test of significant α is (LW Appendix 5)

$$n = \frac{1}{(a/\sigma)^2(1-2c)^2}\left(z_{(1-\alpha/2)} + z_{(1-\beta)}\right)^2$$

where $z_{(x)}$ satisfies $\Pr(U \leq z_{(x)}) = x$ where U is the unit normal. Hence, for $\alpha = 0.05$, $z_{(1-\alpha/2)} = 1.96$ as $\Pr(U \leq 1.96) = 0.975$.

a: Compute the required F_2 sample size (recall for an F_2 that for n individuals, $n/4$ are expected to be in marker class MM) to have 80% power (note that $z_{(0.8)} = 0.84$) to detect a QTL whose effects are $a/\sigma = 1, 0.1$ and 0.05 for both complete ($c = 0$) and modest linkage ($c = 0.2$ linkage).

Solutions to Lecture 8 Problems

1. a. Yes. The marker means are significantly different.

   b:

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

   Hence, $a \geq 2.85$

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

   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. Here

$$n = 4n_{MM} = 4\,\frac{(1.96 + 0.84)^2}{(a/\sigma)^2(1 - 2c)^2} = \frac{31.36}{(a/\sigma)^2(1 - 2c)^2}$$

   The factor of four coming from the fact that the expression for sample size is for a particular homozygote marker class, which is just $1/4$ of the total F_2 sample (as for n F_2, $n/4$ are MM, $n/4$ are mm and $n/2$ are Mm). Hence

a/σ	c	n
1	0	32
1	0.2	88
0.1	0	3136
0.1	0.2	8711
0.05	0	12544
0.05	0.2	34844