Walsh Lecture 2
QTL and Assocation Mapping

Bruce Walsh. 25 Jan - 5 Feb 2016

Second Bangalore School on Population Genetics and Evolution

I: Overview and Using Inbred Lines

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) in 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 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 2.1). 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.

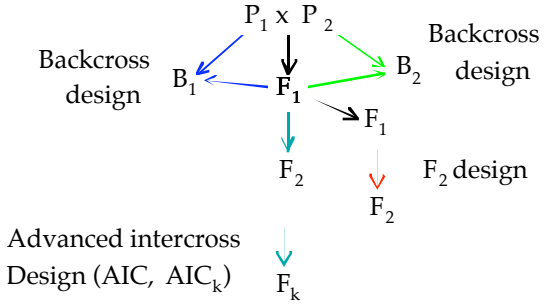

Figure 2.1. Various designs for QTL mapping starting with the F_1 between two lines. Backcross designs cross the F_1 back to either parent, F_2 designs cross the F_1s with each other, advanced intercross lines continue to intermate the F_2 for a number of generations.

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 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,

$$\Pr(Q_k \mid M_j) = \frac{\Pr(Q_k M_j)}{\Pr(M_j)} \tag{2.1}$$

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. 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 2.1 gives the F_2 conditional probabilities as

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

$$\Pr(QQ \mid Mm) = c(1-c), \ \Pr(Qq \mid Mm) = (1-c)^2 + c^2, \ \Pr(qq \mid Mm) = c(1-c) \tag{2.2}$$

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

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_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 2.1 gives

$$\begin{aligned}
\Pr(QQ \mid M_1M_1M_2M_2) &= \frac{(1-c_1)^2(1-c_2)^2}{(1-c_{12})^2} \\
\Pr(Qq \mid M_1M_1M_2M_2) &= \frac{2c_1c_2(1-c_1)(1-c_2)}{(1-c_{12})^2} \\
\Pr(qq \mid M_1M_1M_2M_2) &= \frac{c_1^2\,c_2^2}{(1-c_{12})^2}
\end{aligned} \tag{2.3}$$

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

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(\,Q_k \mid M_j\,) \tag{2.4}$$

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 2.2, the F_2 marker means become

$$(\mu_{MM} - \mu_{mm})/2 = a(1-2c) \tag{2.5a}$$

$$\frac{\mu_{Mm} - (\mu_{MM} + \mu_{mm})/2}{(\mu_{MM} - \mu_{mm})/2} = k(1-2c) \tag{2.5b}$$

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 2.3 (and its analog for $m_1m_1m_2m_2$ probabilities) implies

$$\frac{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}{2} \simeq a\,(1 - 2c_1\,c_2) \tag{2.6a}$$

where c_1 is the M_1-QTL recombination frequency. Equation 2.6a is essentially equal to a when the distance between flanking markers $c_{12} \leq 0.20$, as here $(1 - 2c_1\,c_2) \geq (1 - 2 \cdot 0.1^2) =, 0.98$. Thus, recalling from Equation 2.5a that $\mu_{M_1M_1} - \mu_{m_1m_1} = 2a(1 - 2c_1)$, we can obtain estimates of the recombination frequencies by substituting Equation 2.6a for a and rearranging to give

$$\begin{aligned} c_1 &= \frac{1}{2}\left(1 - \frac{\mu_{M_1M_1} - \mu_{m_1m_1}}{2a}\right) \\ &\simeq \frac{1}{2}\left(1 - \frac{\mu_{M_1M_1} - \mu_{m_1m_1}}{\mu_{M_1M_1M_2M_2} - \mu_{m_1m_1m_2m_2}}\right) \end{aligned} \tag{2.6b}$$

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},

$$z_{ik} = \mu + b_i + e_{ik} \tag{2.7a}$$

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

$$z_j = \mu + \sum_{i=1}^{n} b_i\,x_{ij} + e_j \tag{2.7b}$$

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 2.7a,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 2.5b). 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

$$z_j = \mu + \sum_{i}^{n_1} a_i \, x_{ij} + \sum_{k}^{n_2} b_k \, y_{kj} + \sum_{i}^{n_1} \sum_{k}^{n_2} d_{ik} \, x_{ij} \, y_{kj} + e_j \tag{2.7c}$$

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

$$\ell(z \mid M_j) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2) \, \Pr(Q_k \mid M_j) \tag{2.8}$$

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

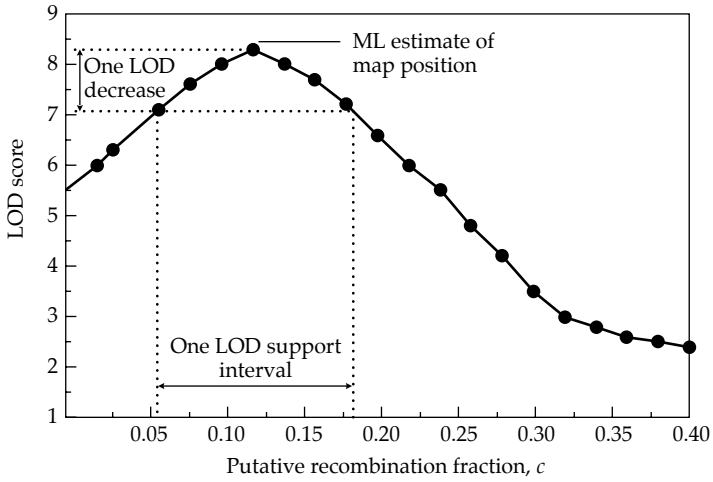

Figure 2.2 Hypothetical likelihood map for the marker-QTL recombination frequency c in a single-marker analysis. Points connected by straight lines are used to remind the reader that likelihood maps are computed by plotting the maximum of the likelihood function for each c value, usually done by considering 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.

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 2.2), 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

$$\mathrm{LOD}(c) = -\log_{10}\left[\frac{\max\,\ell_r(\mathbf{z})}{\max\,\ell(\mathbf{z},c)}\right] = \frac{\mathrm{LR}(c)}{2\,\ln 10} \simeq \frac{\mathrm{LR}(c)}{4.61} \tag{2.9}$$

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 2.2), 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 2.9), 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. 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,

$$\sum_{k \neq i, i+1} b_k \cdot x_{kj} \tag{2.10a}$$

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

$$x_{kj} = \begin{cases} 1 & \text{if individual } j \text{ has marker genotype } M_k M_k \\ 0 & \text{if individual } j \text{ has marker genotype } M_k m_k \\ -1 & \text{if individual } j \text{ has marker genotype } m_k m_k \end{cases} \tag{2.10b}$$

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.

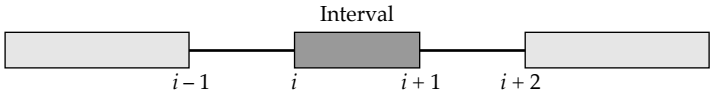

Figure 2.3 Suppose the interval being examined by CIM is between markers i and $i+1$. Addition of the adjacent markers $i-1$ and $i+2$ as cofactors absorbs the effects of any linked QTLs to the left of marker $i-1$ and to the right of marker $i+2$. Their inclusion, however, does not remove the effects 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 2.3). 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 2.3), 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 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 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 (i.e., declared to be significant). 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).

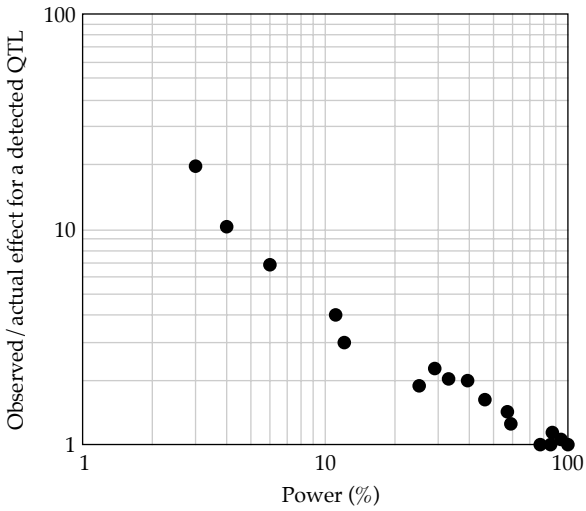

Figure 2.4 Relationship between the probability (power) of detecting a QTL and the amount by which the estimated effect of a *detected* QTL overestimates it actual value. (Based on Beavis 1994.)

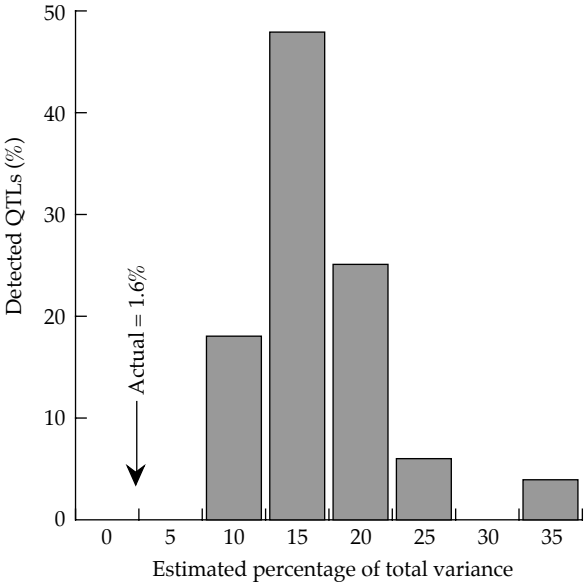

Figure 2.5 Distribution of the estimated effects of detected QTLs. Here 40 QTLs, each accounting for 1.58% of the variance, are assumed. Using 100 F_2 individuals, only 4% of such loci were detected. The average estimated fraction of total variation fraction accounted for by each detected QTLs was 16.3%, with the distribution of estimates skewed towards larger values. (From Beavis 1994.)

As shown in Figure 2.4, 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 2.5, 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.

Dealing with Supersaturated Models: Model Selection

Consider the simple linear model allowing for epistasis (Equation 2.7c). 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.

II: QTL Mapping Using Sets of Outbred Relatives

There are two fundamentally different approaches for mapping QTLs in outbred populations. The first is to look at marker-trait associations in sets of relatives (such as sibs). The second, **association mapping**, uses collections of unrelated individuals, but requires very dense markers. Both rely on linkage disequilibrium, which is guaranteed in a set of relatives but not certain in a collection of unrelated individuals. We consider these two approaches in turn.

The major difference between parents from inbred lines and those from 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.

Informative Matings

Before considering the variety of QTL mapping methods using sets of relatives from an outbred population, 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, **SNPs** (**single nucleotide polymorphisms**) are widely used in inbred lines, but these markers are typically diallelic and hence have modest polymorphism (at best). Microsatellite marker loci (STRs), on the other hand, are highly polymorphic (having a large number of alleles at intermediate frequencies) and hence much more likely to yield marker-informative individuals.

As shown in Table 2.1, 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** ($M_iM_j \times M_kM_\ell$) 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. $M_\ell-$ offspring. With a **backcross family** ($M_iM_j \times M_kM_k$), only the heterozygous parent can be examined for marker-trait associations. Finally, with an **intercross family** ($M_iM_j \times M_iM_j$), homozygous offspring (M_iM_i, M_jM_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.

Table 2.1 Types of marker-informative matings.

Fully informative:	$M_iM_j \times M_kM_\ell$
Parents are different marker heterozygotes.	
All offspring are informative in distinguishing alternative alleles from both parents.	
Backcross:	$M_iM_j \times M_kM_k$
One parent is a marker heterozygote, the other a marker homozygote.	
All offspring informative in distinguishing heterozygous parent's alternative alleles.	
Intercross:	$M_iM_j \times M_iM_j$
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.

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. Suppose there are n alleles at a marker locus, the ith of which has frequency p_i, then

$$\text{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\, p_i^2\, p_j^2 \leq \frac{(n-1)^2(n+1)}{n^3} \tag{2.11}$$

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 2.11) 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-sibs, where the basic linear model is a nested ANOVA, with marker effects nested within each sibship,

$$z_{ijk} = \mu + s_i + m_{ij} + e_{ijk} \tag{2.12}$$

Here 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 expected value 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

$$F = \frac{\text{MS}_m}{\text{MS}_e} \tag{2.13}$$

where MW_m and MS_e are the marker and error mean squares (from the ANOVA table). Assuming normality, Equation 2.13 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 2.13 has N and $N(n-2)$ degrees of freedom.

For a balanced design, the mean squares have expected values of

$$E(\,\text{MS}_m\,) = \sigma_e^2 + (n/2)\sigma_m^2 \qquad \text{and} \qquad E(\,\text{MS}_e\,) = \sigma_e^2 \tag{2.14}$$

where

$$\sigma_m^2 = \frac{E(\text{MS}_m) - E(\text{MS}_e)}{n/2} = (1-2c)^2\,\frac{\sigma_A^2}{2} \tag{2.15}$$

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 $\sigma_A^2 = 2a^2p(1-p)$. 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 parent 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 $2a$ (as opposed to σ_m^2), 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 2.6), 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 traits in dairy cows, where the offspring are granddaughters of the original sires. The linear model for this design is

$$z_{ijk\ell} = \mu + g_i + m_{ij} + s_{ijk} + e_{ijk\ell} \tag{2.16}$$

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.

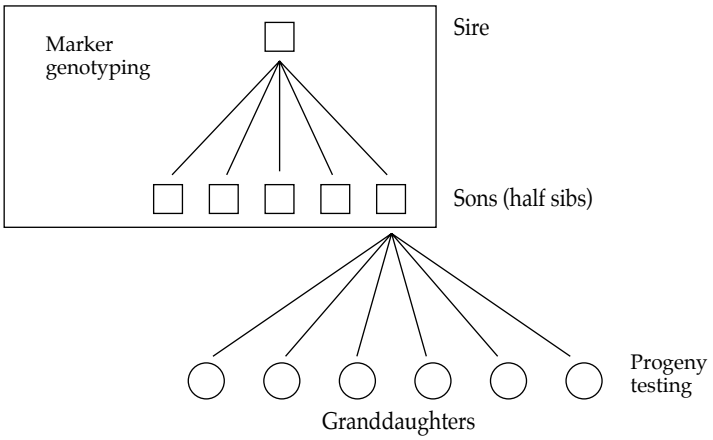

Figure 2.6 The granddaughter design of Weller et al. (1990). Here, each sire produces a number of half-sib sons that are scored for the marker genotypes. The character value for each son is determined by progeny testing, with the trait value being scored in a large number of daughters (again half-sibs) 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 multigenerational 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. In this **fixed-effects** analysis, one estimates a number of parameters, such as the QTL allele frequencies, means, and recombination fraction. Each of these unknowns is treated as a fixed effect to estimate. As a result, any changes in the genetic assumptions (for example, assuming three rather than two alleles at the QTL) require that the entire model has be reconstructed.

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 is a **random-effects** approach, as we replace estimation of all of the genetic means and frequencies with simply estimating the genetic variance associated with a region. Since we are uninterested in further details (such as the number of QTL alleles and their indivdiuals efects and frequecnies), a single model works for all situations. 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.

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

$$z_i = \mu + A_i + A_i^* + e_i \tag{2.17}$$

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 $\sigma^2_A + \sigma^2_{A^*} + \sigma^2_e$.

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

$$\sigma(z_i, z_j) = R_{ij}\,\sigma^2_A + 2\Theta_{ij}\,\sigma^2_{A^*} \tag{2.18}$$

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). 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,

$$\mathbf{V} = \mathbf{R}\,\sigma^2_A + \mathbf{A}\,\sigma^2_{A^*} + \mathbf{I}\,\sigma^2_e \tag{2.19a}$$

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

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases} \tag{2.19b}$$

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. When testing a particular region of interest, the marker information in that region allows us to estimate $\mathbf{R}$. As we move to a new region, new marker information is used to estimate $\mathbf{R}$ for this new region, but $\mathbf{A}$ remains unchanged as the pedigree is unaffected by the region under consideration. Thus, for any region, there is a unique $\mathbf{R}$, but the same $\mathbf{A}$ is used for all regions.

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

$$\ell(\,\mathbf{z}\,|\,\mu, \sigma^2_A, \sigma^2_{A^*}, \sigma^2_e\,) = \frac{1}{\sqrt{(2\pi)^n\,|\mathbf{V}|}}\,\exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\,\mathbf{V}^{-1}\,(\mathbf{z}-\boldsymbol{\mu})\right] \tag{2.20}$$

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

III: Introduction to Association Mapping

All of the above methods require known collections of relatives. These can be a problem to obtain, especially if very large pedigrees are needed. Further, fine-mapping can be difficult because we need (on average) 100 relatives to see a single recombination event between two markers separated by 1 cM. Hence, even if the QTL has a large effect, very large family sizes are still needed to obtain the number of recombinants required for fine mapping.

An alternative approach is **association mapping**. Here, one collects a large random set of individuals from a population and relies on linkage disequilibrium between very closely linked genes to do the mapping. Association mapping has its roots in human genetics, but it has become popular for mapping genes in plant populations as well. In plant breeding, one can use a large collection of cultivars as the mapping population, where each line represents a different 'individual' in our mapping population.

LD: Linkage disequilibrium

The key to association mapping is the presence of **linkage disequilibrium**. Consider two alleles A and B at two different loci, and consider the population frequency of gametes containing both A and

B, i.e., AB gametes. If A and B occur independently of each other in gametes, then the frequency p_{AB} of gametes containing AB is a simply the product of the allele frequencies p_A and p_B,

$$p_{AB} = p_A \cdot p_B$$

However, it is often the case that A and B have some statistical association in a population, and we measure the departure of gamete frequencies from this equilibrium value by the **linkage disequilibrium** D,

$$D_{AB} = p_{AB} - p_A \cdot p_B$$

If D is zero, then alleles A and B are uncorrelated in a population. If $D > 0$, then the AB gamete is more frequent than expected. If we start with a given level of disequilibrium, then after a single generation of recombination and random mating, the new amount of disequilibrium is just $(1-c)$ of the previous level, where c is the recombination frequency between the two loci. $D(t)$, the expected disequilibrium in generation t, is given by

$$D(t) = (1-c)^t \, D(0)$$

Thus, the time required to reduce any intitial disequilibrium by half is 1 generation for $c = 0.5$, 7 generations for $c = 0.1$, 68 generations for $c = 0.01$, and $0.69/c$ generations for smaller values of c.

The more formal term for LD is **gametic-phase disequilibrium**, as *unlinked* loci can still be statistically associated. Indeed, this can occur when our apparent single population is in reality a collection of partly isolated sub-populations. This notion of disequilibrium when loci are unlinked foreshadows caution when using LD to map genes. Indeed, dealing with spurious associations generated by population structure is a major concern when attempting to map genes, a point we return to shortly.

A variety of forces, such as genetic drift and selection, can generate LD. In particular, when a single new mutation arises, it is in complete LD with tightly-linked markers, as the only copy occurs in the particular haplotype in which it arose. Over time, the disequilibrium between this new mutation and background markers on the original haplotype decays, but for very tightly linked loci (c very small) this can take hundreds or thousands of generations. The key concept for association mapping is that apparently 'unrelated' individuals may still contain very small chromosomal segments that are IBD and hence by using sufficiently dense markers we may be able to detect an LD signal in this region

A critical issue when attempting to use LD to map genes in a population is the extent of LD in the genome of interest. For example, if LD typically extends 30 - 60 kb (kilobases = thousands of DNA base pairs) around any marker, then SNPs or STRs every 50 KB or so will provide complete coverage of the genome. However, if LD is typically much shorter, or much longer, than many more (or many fewer) markers are required. Genomes with large regions of LD offer higher power in mapping, but lower power in terms of resolution. If LD extends over (say) 150 kb, then a number of markers around functional site will provide a marker-trait association, offering higher power. However, dissecting this region further can be difficult if LD extends over large regions. Conversely, when LD extends over a very small region, power can be low (we have to have a very close marker), but resolution of the functional site is much higher.

D' and r^2: Measures of LD

LD is a function of allele frequencies, and thus a scaled measure, independent of allele frequencies, is desirable. A number of such measures have been proposed. One popular one is D', which ranges (in absolute value) from 0 to 1, where

$$D' = \frac{D}{D_{max}}, \qquad D_{max} = \begin{cases} \min\left(p_A \cdot [1-p_B], [1-p_A] \cdot p_B\right) & \text{if } D > 0 \\ \min\left(p_A \cdot p_B, [1-p_A] \cdot [1-p_B]\right) & \text{if } D < 0 \end{cases}$$

While D' is often reported, it is very unstable when one of the alleles is at very low frequency. A more useful measure is the squared correlation coefficient between alleles A and B,

$$r^2 = \frac{D^2}{p_A\, p_B\, (1 - p_A)\, (1 - p_B)}$$

This measure has several nice features. First, its expected value under a model with just drift and recombination is

$$E\left(r^2\right) = \frac{1}{4N_e c + 1}$$

where N_e is the effective population size. Second, if the sample size (number of gametes) is n, then $n \cdot r^2$ follows a χ^2_1, a chi-square distribution with one degree of freedom. Finally (and perhaps most importantly) r^2 is also less sensitive to rare alleles.

Fine-mapping Major Genes Using LD

The simplest approach for using LD to map genes of large effect proceeds as follows. Suppose a disease allele is either present as a single copy (and hence associated with a single chromosomal haplotype) in the founder population or arose by mutation very shortly after the population was formed. Assume that there is no **allelic heterogeneity**, so that all disease-causing alleles in the population descend directly from a single original mutation, and consider a marker locus tightly linked to the disease locus. The probability that a disease-bearing chromosome has not experienced recombination between the **disease susceptibility** (DS) gene and marker after t generations is just $(1-c)^t \simeq e^{-c\,t}$, where c is the marker-DS recombination frequency. Suppose the disease is predominantly associated with a particular haplotype, which presumably represents the ancestral haplotype on which the DS mutant arose. Equating the probability of no recombination to the observed proportion π of disease-bearing chromosomes with this predominant haplotype gives $\pi = (1 - c)^t$, where t is the age of the mutation or the age of the founding population (whichever is more recent). Hence, one estimate of the recombination frequency is

$$c = 1 - \pi^{1/t} \tag{2.21}$$

The power for LD mapping comes from *historical recombinations*. If there are t generations back to the common ancestor, then in a sample of n individuals, we have nt generations of recombination. Conversely, if looking at sibs, each individual has just one generation of recombination. If c is the recombination rate, we expect ntc recombinants in our population sample versus nc recombinants in our sib sample.

Example

Hästbacka et al. (1992) examined the gene for diastrophic dysplasis (DTD), an autosomal recessive disease, in Finland. A total of 18 **multiplex** families (showing two or more affected individuals) allowed the gene to be localized to within 1.6 cM from a marker locus (*CSF1R*) using standard pedigree methods. To increase the resolution using pedigree methods requires significantly more multiplex families. Given the excellent public health system in Finland, however, it is likely that the investigators had already sampled most of the existing families. As a result, the authors turned to LD mapping.

While only multiplex families provide information under standard mapping procedures, this is not the case with LD mapping wherein single affected individuals can provide information. Using LD mapping thus allowed the sample size to increase by 59. A number of marker loci were examined, with the *CSF1R* locus showing the most striking correlation with DTD. The investigators were able to unambiguously determine the haplotypes of 152 DTD-bearing chromosomes and 123 normal chromosomes for the sampled individuals. Four alleles of the *CSF1R* marker gene were detected. The frequencies for these alleles among normal and DTD chromosomes were found to be:

	Chromosome type			
Allele	Normal		DTD	
1-1	4	3.3%	144	94.7%
1-2	28	22.7%	1	0.7%
2-1	7	5.7%	0	0%
2-2	84	68.3%	7	4.6%

Given that the majority of DTD-bearing chromosomes are associated with the rare 1-1 allele (present in only 3.3% of normal chromosomes), the authors suggested that all DTD-bearing chromosomes in the sample descended from a single ancestor carrying allele 1-1. Since 95% of all present DTD-bearing chromosomes are of this allele, $\pi = 0.95$. The current Finnish population traces back to around 2000 years to a small group of founders, which underwent around $t = 100$ generations of exponential growth. Using these estimates of π and t, Equation 2.21 gives an estimated recombination frequency between the *CSF1R* gene and the DTD gene as $c = 1 - (0.95)^{1/100} \simeq 0.00051$. Thus, the two genes are estimated to be separated by 0.05 cM, or about 50 kb (using the rough rule for humans that 1 cM = 10^6 bp). Subsequent cloning of this gene by Hästbacka et al. (1994) showed it be to 70 kb proximal to the *CSF1R* marker locus. Thus, LD mapping increased precision by about 34-fold over that possible using segregation within pedigrees (0.05 cM vs. 1.6 cM).

The Candidate-Gene Approach

There are a couple of general strategies for mapping genes influencing traits of interest: use of **anonymous markers** and use of **candidate gene** markers. QTL mapping approaches are an example of using anonymous markers: no assumptions are made as to what the genes might be and markers are chosen simply scan the whole genome for marker-trait associations. The second approach is the **candidate-gene approach**, wherein one has a target list of genes thought to be involved in the trait and markers are chosen within those genes. The key here is the search for alleles at these candidates that result in variation in the trait of interest. Thus, one typically chooses one (or a few) markers within candidate genes and searches for marker-trait associations. We use the term markers, as a (say) SNP marker used within a gene might not be the **causal site** of functional differences. However, if it is in LD with such a causal site, it will also show a marker-trait association. Candidate gene searches in human genetics have been very popular, but finding a marker-trait association at a candidate is not, by itself, evidence of a functional relationship between variation at that candidate and variation in the trait.

Population Structure and the Transmission/Disequilibrium Test

When considering genetic disorders, the frequency of a particular candidate (or marker) allele in affected (or **case**) individuals is often compared with the frequency of this allele in unaffected (or **control**) individuals. The problem with such **association studies** is that a disease-marker association can arise simply as a consequence of population structure, rather than as a consequence of linkage. Such **population stratification** occurs if the sample consists of a number of divergent populations (e.g., different ethnic groups) which differ in both candidate-gene frequencies and incidences of the disease. Population structure can severely compromise tests of candidate gene associations, as the following example illustrates.

Segregation analysis gave evidence for a major gene for Type 2 diabetes mellitus segregating at high frequency in members of the Pima and Tohono O'odham tribes of southern Arizona. In an attempt to map this gene, Knowler et al. (1988) examined how the simple presence/absence of a particular haplotype, Gm^+, was associated with diabetes. Their sample showed the following associations:

Gm^+	Total subjects	% with Diabetes
Present	293	8%
Absent	4,627	29%

The resulting χ^2 value (61.6, 1 df) shows a highly significant negative association between the Gm^+ haplotype and diabetes, making it very tempting to suggest that this haplotype marks a candidate diabetes locus (either directly or by close linkage).

However, the presence/absence of this haplotype is also a very sensitive indicator of admixture with the Caucasian population. The frequency of Gm^+ is around 67% in Caucasians as compared to $< 1\%$ in full-heritage Pima and Tohono O'odham. When the authors restricted the analysis to such full-heritage adults (over age 35 to correct for age of onset), the association between haplotype and disease disappeared:

Gm^+	Total subjects	% with Diabetes
Present	17	59%
Absent	1,764	60%

Hence, the Gm^+ marker is a predictor of diabetes not because it is linked to genes influencing diabetes but rather because it serves as a predictor of whether individuals are from a specific subpopulation. Gm^+ individuals usually carry a significant fraction of genes of Caucasian extraction. Since a gene (or genes) increasing the risk of diabetes appears to be present at high frequency in individuals of full-blooded Pima/Tohono O'odham extraction, admixed individuals have a lower chance of carrying this gene (or genes).

The problem of population stratification can be overcome by employing tests that use family data, rather than data from unrelated individuals, to provide the case and control samples. This is done by considering the transmission (or lack thereof) of a parental marker allele to an affected offspring. Focusing on transmission within families controls for association generated entirely by population stratification and provides a direct test for linkage *provided* that a population-wide association between the marker and disease gene exists.

The **transmission/disequilibrium test**, or TDT, compares the number of times a marker allele is transmitted (T) versus not-transmitted (NT) from a marker heterozygote parent to affected offspring. Under the hypothesis of no linkage, these values should be equal, and the test statistic becomes

$$\chi^2_{td} = \frac{(T - NT)^2}{(T + NT)} \tag{2.22}$$

which follows a χ^2 distribution with one degree of freedom. How are T and NT determined? Consider an M/m parent with three affected offspring. If two of those offspring received this parent's M allele, while the third received m, we score this as two transmitted M, one not-transmitted M. Conversely, if we are following marker m instead, this is scored as one transmitted m, two not-transmitted m. As the following example shows, each marker allele is examined separately under the TDT.

Example: Mapping Type 1 Diabetes

Copeman et al. (1995) examined 21 microsatellite marker loci in 455 human families with Type 1 diabetes. One marker locus, *D2S152*, had three alleles, with one allele (denoted 228) showing a significant effect under the TDT. Parents heterozygous for this marker transmitted allele 228 to diabetic offspring 81 times, while transmitting alternative alleles only 45 times, giving

$$\chi^2 = \frac{(81 - 45)^2}{(81 + 45)} = 10.29$$

which has a corresponding P value of 0.001. As summarized below, the other two alleles (230 and 240) at this marker locus did not show a significant TD effect.

Allele	T	NT	χ^2	P
228	81	45	10.29	0.001
230	59	73	1.48	0.223
240	36	24	2.40	0.121

Hence, this marker is linked to a QTL influencing Type 1 diabetes, with allele 228 in (coupling) linkage disequilibrium with an allele that increases the risk for this disease.

Linkage vs. Association

The distinction between linkage and association is a subtle, yet critical, one. A marker allele M is *associated* with a trait if $\sigma(M, z) \neq 0$, i.e., there is non-zero covariance between the marker allele state and the phenotype z for the trait of interest. Such an association could arise because the marker is in linkage-disequilibrium with a linked QTL that influences the trait. However, it could also arise due to population structure, with the marker predicting subpopulation membership and subpopulation membership being predictive of trait value (as was the case for Gm^+ and diabetes). Thus, association DOES NOT imply linkage, and likewise linkage (by itself) does not imply association, as linkage disequilibrium is required.

The TDT is a joint test of BOTH linkage and associatio, specifically a test that the marker is linked AND in linkage disequilibrium with a QTL influencing the trait. Linkage within each family results in a non-random distribution of transmitted and non-transmitted marker alleles when the marker is linked to the QTL. However, population-level linkage disequilibrium (i.e., association) is also required for the TDT to be significant as we simply average over all families. If linkage phase varies randomly over families, the within-family signal will be randomized across families and hence no significant TDT signal is generated.

How is such population-level disequilibrium be generated? Think back to linkage disequilibrium mapping for major genes. If a disease is caused by a single mutation, then when that mutation arose it was in a particular background (or haplotype), and hence starts out in linkage disequilibrium with linked markers. Over time, recombination randomizes the association between even fairly tightly linked markers and the mutant allele. However, with very tightly linked markers, a very long time (potentially thousands of generations) is required to decay the initial disequilibrium. This same logic also holds true for QTLs, except now the effect of any particular mutation is small, so that larger sample sizes are required to detect this small association signal.

IV: Dense SNP Assocation Mapping

Typically association studies, especially in humans, have used the TDT experimental design, controlling for the confounding effects of population structure by using sets of known relatives. This is a very costly design, as sets of relatives can be hard to find, and often are expensive to track down and genotype/phenotype. In contrast, it is fairly straightforward to gather large sets of random individuals from populations. Often this is done using the **case/control** design, where an equal number of cases (showing the trait or disease) and controls (lacking the trait/disease) are chosen. More generally, for quantitative traits, one can sample from phenotypically extreme individuals (i.e., those, say, with high and low blood pressure). It is thus fairly easier to obtain very large samples using random, as opposed to related, individuals. If we are to map QTLs with small effects, such large sample sizes are indeed required. As mentioned, the problem is dealing with population substructure. Recently, several approaches have been developed to attempt to control for the potential of population structure without having to use sets of relatives as controls.

These developments have led to the notion of **dense SNP association mapping**. Dense in the

sense that a very large number (tens to hundreds of thousands) of SNPs are scored. By using a dense set of SNPs, we can find a SNP that is very tightly linked to any particular genomic location. For example, the human genetic map is roughly 3000 cM, so that if one uses 30,000 equally-spaced SNPs, each will be roughly 0.1 cM apart, and thus any particular region will be within 0.05 cM of a SNP. Under this design, we expect fairly high levels of linkage disequilibrium coupled with large sample sizes, the recipe needed to detect QTLs of small effects linked to a marker.

Kaplan and Morris (2001) have examined the effect of the age of the mutation on association studies. They conclude that disease (QTL) alleles at intermediate frequencies (presumably representing old mutations) tend to give the largest test statistic values at markers near the QTL. In contrast, QTL alleles at low frequencies (presumably representing fairly new mutations) often give large test statistic values scattered over markers within region, making localization of the QTL more uncertain. Presumably this occurs because older mutations have had time to for any initial disequilibrium to decay at all but the closest markers, while younger mutations have more stochastic levels of disequilibrium within a region around the QTL.

Why are SNPs used, as opposed to the more polymorphic STRs (microsatellite loci)? The reason is subtle, namely mutation rates. SNPs have very low mutation rates, and hence any decay in association between a SNP and linked QTL is entirely due to recombination. Such is not the case for STRs, which have high mutation rates (often around $1/1000$ to $1/250$ per generation). Hence, if (say) allelic state 12 was originally associated with the initial mutation, on some chromosomes this could mutate to state 11, and likewise to state 13 on others. Both events significantly diffuse any initial assocation signal, as mutation here is akin to recombination in decaying any initial disequilibrium.

How do we account for population structure? The basic idea is simple – if one looks at a number of markers not associated with the trait, each should still contain a signal for the common population structure. One can thus either correct for this common signal (**genomic control** and regression approaches), or else use the markers to predict subpopulation membership and look at association within each subpopulation (**structured association analysis**). If we have a number of subpopulations each in Hardy-Weinberg equilibrium, we can still see significant departures from HW when we ignore this structure and treat the sample as a single population. Hence, one crude test of structure is to look for consistent departures from HW across a large number of loci.

As an (important) aside, the genome is best thought of as having four (five in plants) components: The autosomes, the Y chromosome, the X chromosome, and mtDNA (and cpDNA in plants). For various reasons, such as mode of genetic transmission (Y only from males; mtDNA, cpDNA typically all from females) and differences in effective population size (Y smallest, X intermediate, autosomes largest), the population structure of these components may be different, and each should be examined separately.

Genomic Control

The notion of genomic control is due to Devlin and Roeder (1999). Their basic idea is as follows. Consider the case/control design. Here one would use standard 2 x 2 contingency tables to look at the level of association between SNP alleles in cases vs. controls. With no population structure, this test follows a χ^2 distribution with one degree of freedom. However, when population structure is present, the test statistic now follows a *scaled* χ^2, so that if S is the value of the test statistic, then $S/\lambda \sim \chi^2_1$ (or $S \sim \lambda\chi^2_!$), where the inflation factor λ is given by

$$\lambda \simeq 1 + nF_{st}\sum_k (f_k - g_k)^2 \tag{2.23}$$

where we assume n cases and n controls, f_k (g_k) is the fraction of cases (controls) in the k-th subpopulation, and F_{st} is a measure of the population structure. A critical feature of Equation 2.23 is that the departure of the test from a χ^2 *increases* as sample size (n) increases. Thus, things can be worst for larger sample sizes unless we correct for population structure.

Likewise, moving from case-control to quantitative traits, our measure of association might be

a regression, with the phenotype of the ith individual predicted by

$$y_i = \mu + \beta X_i + e_i \tag{2.24}$$

where X_i is the number of copies of the SNP marker allele (0,1, or 2) in individual i. The test for a trait-SNP associate is a test for a significant value of the regression slope β. Devlin, Roeder and Wasserman (2001) showed that using the OLS estimates for β ($\widehat{\beta}$) and its variance [SE$^2(\widehat{\beta})$], that the test statistic

$$QT = \left(\frac{\widehat{\beta}}{\mathsf{SE}(\widehat{\beta})} \right)^2 \sim \lambda \chi^2_1 \tag{2.25}$$

and hence has the same population-structure inflation factor as for case-control studies.

Genomic control attempts to estimate the value of λ using all of the SNP association tests, as most of these have no effect (attributable to linkage) on the trait, but rather simply represent effects of any population structure. One approach to estimating the inflation factor λ is to note that the mean of a χ^2_1 is 1, so that the mean value of the tests estimates λ. The problem with this approach is that this estimator is not a particular robust, as a few extreme test values (which might be expected for SNPs with linkage association with the trait) can significantly inflate the mean. Rather, Devlin et al. suggest at an estimator more robust to extreme values is given by using the medium (i.e., 50% value) of all tests, namely

$$\widehat{\lambda} = \frac{\mathrm{medium}(S_1, \cdots S_m)}{0.456} \tag{2.26}$$

where we have done m tests, and S_i is the test statistic value for the ith of these. The estimated value of λ (the **inflation factor**) is widely used as a test for improved model fit, with a model with $\lambda \leq 1.05$ considered to be well-adjusted for population structure effects.

The advantage of genomic control is that it is simple to apply and uses information from a large number of tests. Its major disadvantage is that it does not *re-order* the p values of marker-trait association tests, rather it simply rescales them. Thus, one could imagine a situation where if population structure is fully take into account, that the p values would not only change, but the rank order of p values over different markers would also change. This does not occur under genomic control.

Structured Assocation Analysis

An alternative approach to correcting for population structure was offered by Pritchard and Rosenberg (1999), who suggested using marker information to first assign individuals into subpopulations, and then once these assignments are made to perform associations tests within each. Under the assumption that each of k subpopulations is itself in Hardy-Weinberg, Pritchard and Rosenberg developed a MCMC Bayesian classifier to assign individuals into k subpopulations. One then adjusts the number of potential subpopulations until optimal model fit is obtained. This is an example of a **latent variable** approach, where each observation has some unobservable value (here class membership), that if we knew this value, the analysis would be significantly easier. (The same situation arises for the mixture models discussed above).

The program **Structure** from Pritchard's group is often applied, especially when one is using a collection of lines, to examine how much internal structure exists. The problem with Structure (and other such classification approaches) is that they can be relatively unstable. One suggested approach is to do a number of structure runs for each sample, then compute some genetic distance measure and see how well the different runs cluster on a phylogenetic tree constructed using this distance measure. If runs assuming (say) five groups tend to cluster together, one might have confidence in the outcome. If they tend to be distributed over the other samples, some of the assumptions of structure (such as HW within each group) are likely violated.

Regression Approaches/Principal Components

A third approach for correcting for structure is simply to include a number of markers, outside of the SNP of interest, in a regression analysis. These markers would absorb the effects of structure, so that the regression slope of the trait on the SNP is now a partial regression, giving the effect on the trait of changing the SNP while holding the effects of population structure constant.

We can regression the SNP on the trait in a couple of ways. First, we could use simply use the number of copies X of a SNP allele (0,1,2), giving a βX term, as in Equation 2.24. This type of analysis only looks at the additive value of the SNP-trait association. If strong dominance is of concern, then each genotype should be coded separately. Suppose we are considering a particular marker, with n genotypes (typically for a SNP, $n = 3$, both homozygotes and the heterozygote). Let β_k be the effect of marker genotype k on predicting the phenotype. We will also add to the regression m markers scattered throughout the genome, with γ_j being the effect of marker j on predicting the phenotype. The resulting regression,

$$y = \mu + \sum_{k=1}^{n} \beta_k M_k + \sum_{j=1}^{m} \gamma_j b_j + e \tag{2.27}$$

has two components. The effect of the second sum (the γ) is to account for population structure on phenotype, while a significant effect of the target SNP marker is indicated at least one β significantly different from zero. How might these marker cofactors be selected? The answer is quite straightforward: select those that show the greatest departure from HW as initial candidates and then let a model-selection approach (such as stepwise regression) choose among these.

A very similar approach is based on the **principal components** (PCs) of the marker covariance matrix. PCs extract a subset of all the variables that accounts for a significant fraction of the total variance (these are the k leading eigenvectors corresponding to the k largest eigenvalues of the covariance matrix). The last term in Equation 2.27 is replaced by the first m PCs, which predicts the group membership and then adjusts for any group-specific effects. For both this approach, and Equation 2.27, markers should be unlinked to the SNPs being tested, leading to the so-called **LOCO** (leave one chromosome out) approach, where the analysis are done one chromosome at a time. A separate regression is performed for all of the markers on a given chromosome, with the weights for the second sum (either markers or PCs) based on markers on other chromosomes, and these weights are the same for all of the SNP regressions on the target chromosome are scanned.

Structure plus Kinship Methods

Association mapping in plants offer occurs by first taking a large collection of lines, some closely related, others more distantly related. Thus, in addition to this collection being a series of subpopulations (derivatives from a number of founding lines), there can also be additional structure within each subpopulation (groups of more closely related lines within any particular lineage/subpopulation). These associations may be partly accounted for if we have extensive data on the origins of these various lines. However, often such information in lacking, in which case we need to estimate this using molecular markers. Yu et al in 2006 proposed a mixed-model that allows for these different levels of association: membership in different subpopulations and differences in relatedness (**kinship**) among lines within a given subpopulation. The basic structure of their model (in matrix form) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\alpha + \mathbf{Q}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{2.28}$$

Here $\mathbf{y}$ is the vector of observations, typically the means of each of the lines measured over a number of environments. $\boldsymbol{\beta}$ are various fixed-effects (such as correcting for common environmental factors), α is the vector of interest, namely the SNP-trait associations. $\mathbf{v}$ is a vector indicating the effect of the line coming from a particular subpopulation (these are estimated using a program like Structure, or from PCs). $\mathbf{u}$ is the vector of background polygenic effects, which can be correlated due to having relatives *within* a given line (estimated from marker information), and $\mathbf{e}$ is the residual error.

Lecture 2 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).

3. 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 individual is marker-informative?

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

c: What is the probability that a random individual 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 2 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

3.

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

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