Lecture 9

QTL and Association Mapping in Outbred Populations

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

Notes from a short course taught Jan-Feb 2012 at University of Uppsala

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.

QTL Mapping Using Sets of Relatives

The major difference between parents from inbred lines as opposed to 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 9.1, there are three kinds of marker-informative crosses. With a highly polymorphic marker, it may be possible to examine marker-trait associations for both parents. With a **fully** (marker) **informative family** ($M_iM_j \times M_kM_\ell$) 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_ℓ– 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 9.1 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

$$\text{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\, p_i^2\, p_j^2 \leq \frac{(n-1)^2(n+1)}{n^3} \tag{9.1}$$

which is the probability that one parent is a marker heterozygote and its mate has a *different* genotype (i.e., a backcross or fully informative family, but excluding intercross families). In this case, we can distinguish between the alternative marker alleles of the first parent in all offspring from this cross. The upper bound (given by the right hand side of Equation 9.1) 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,

$$z_{ijk} = \mu + s_i + m_{ij} + e_{ijk} \tag{9.2}$$

where z_{ijk} denotes the phenotype of the kth individual of marker genotype j from sibship i, s_i is the effect of sire i, m_{ij} is the effect of marker genotype j in sibship i (typically, $j = 1, 2$ for the alternative sire marker alleles), and e_{ijk} is the within-marker, within-sibship residual. It is assumed that s, m, and e have 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

$$F = \frac{\text{MS}_m}{\text{MS}_e} \tag{9.3}$$

where MW$_m$ and MS$_e$ are the marker and error mean squares (from the ANOVA table). Assuming normality, Equation 9.3 follows an F distribution under the null hypothesis that $\sigma^2_m = 0$. Assuming a balanced design with N sires, each with $n/2$ half-sibs in each marker class, Equation 9.3 has N and $N(n-2)$ degrees of freedom.

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

$$E(\,\mathrm{MS}_m\,) = \sigma^2_e + (n/2)\sigma^2_m \qquad \text{and} \qquad E(\,\mathrm{MS}_e\,) = \sigma^2_e \tag{9.4}$$

where

$$\sigma^2_m = \frac{E(\mathrm{MS}_m) - E(\mathrm{MS}_e)}{n/2} = (1-2c)^2\,\frac{\sigma^2_A}{2} \tag{9.5}$$

Thus, an estimate of the QTL effect (measured by its additive variance σ^2_A, 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^2_m. Consider a strictly additive diallelic QTL with allele frequency p, where $\sigma^2_A = 2a^2p(1-p)$. 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^2_A 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^2_m), since here all families are informative, rather that the fraction $2p(1-p)$ seen an outbred population.

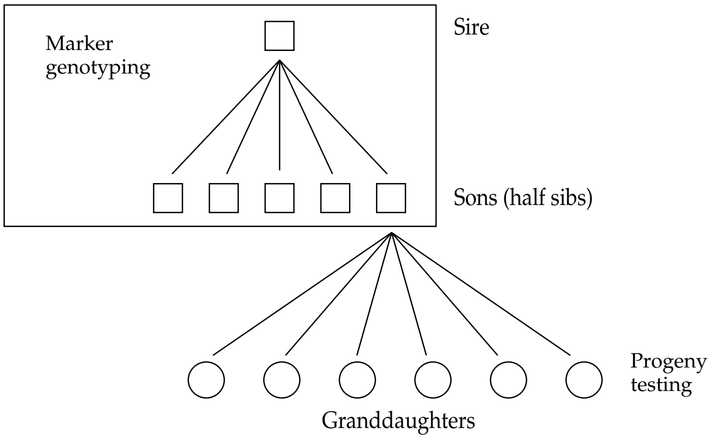

Figure 9.1 The granddaughter design of Weller et al. (1990). Here, each sire produces a number of half-sib sons that are scored for the marker genotypes. The character value for each son is determined by progeny testing, with the trait value being scored in a large number of daughters (again half-sibs) from each son.

One approach for increasing power is the **granddaughter design** (Weller et al. 1990), under which each sire produces a number of sons that are genotyped for sire marker alleles (Figure 9.1), 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

$$z_{ijk\ell} = \mu + g_i + m_{ij} + s_{ijk} + e_{ijk\ell} \tag{9.6}$$

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.

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

$$z_i = \mu + A_i + A_i^* + e_i \tag{9.7}$$

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^2_A, $\sigma^2_{A^*}$, and σ^2_e. Here σ^2_A and $\sigma^2_{A^*}$ 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

$$\sigma(z_i, z_j) = R_{ij}\,\sigma^2_A + 2\Theta_{ij}\,\sigma^2_{A^*} \tag{9.8}$$

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 **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,

$$\mathbf{V} = \mathbf{R}\,\sigma^2_A + \mathbf{A}\,\sigma^2_{A^*} + \mathbf{I}\,\sigma^2_e \tag{9.9a}$$

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

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases} \tag{9.9b}$$

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 $\mathbf{V}$,

$$\ell(\, \mathbf{z} \,|\, \mu, \sigma_A^2, \sigma_{A^*}^2, \sigma_e^2 \,) = \frac{1}{\sqrt{(2\pi)^n \, |\mathbf{V}|}} \, \exp\left[-\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{z} - \boldsymbol{\mu}) \right] \tag{9.10}$$

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

Haseman-Elston Regressions

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

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

$$Y_i = a + \beta \pi_{im} + e \tag{9.11}$$

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

$$\beta = -2\,(1 - 2c)^2 \, \sigma_A^2 \tag{9.12}$$

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

$$\beta = \begin{cases} -2\,(1 - 2c)\,\sigma_A^2 & \text{grandparent–grandchild;} \\ -2\,(1 - 2c)^2\,\sigma_A^2 & \text{half-sibs;} \\ -2\,(1 - 2c)^2\,(1 - c)\,\sigma_A^2 & \text{avuncular (aunt/uncle–nephew/niece).} \end{cases} \tag{9.13}$$

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

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

Affected Sib Pair Methods

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

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

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

$$T_2 = \frac{\widehat{p}_{22} - 1/4}{\sqrt{\dfrac{3}{16\,n_2}}} \tag{9.14a}$$

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

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

$$T_m = \sqrt{2\,n_2}\;(\,\widehat{p}_{21} + 2\,\widehat{p}_{22} - 1\,) \tag{9.14b}$$

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

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

$$MLS = \log_{10}\left[\prod_{i=0}^{2}\left(\frac{\widehat{p}_{2i}}{\pi_{2i}}\right)^{n_{2i}}\right] = \sum_{i=0}^{2} n_{2i}\,\log_{10}\left(\frac{\widehat{p}_{2i}}{\pi_{2i}}\right) \tag{9.15}$$

where π_{2i} is the probability that the pair of doubly affected sibs shares i alleles ibd in the absence of linkage to a QTL. (For full sibs, $\pi_{20} = \pi_{22} = 1/4$, $\pi_{21} = 1/2$.) The test statistic given by Equation 9.15 is referred to as the **maximum LOD score**, or **MLS**, with a score exceeding three being taken as significant evidence for linkage (Risch 1990b, Morton 1955b).

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

$$MLS = (1 - n_1)\log_{10}\left(\frac{1-\widehat{p}_1}{1/2}\right) + n_1\log_{10}\left(\frac{\widehat{p}_1}{1/2}\right) \tag{9.16}$$

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

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

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

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

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 that is starting to become popular with very high density SNP maps 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 (and several of our examples will be human), but it has recently 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 much 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

$$c = 1 - \pi^{1/t} \tag{9.17}$$

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. Thus, if c is the recombination rate, we expect ntc recombinants in our population sample and nc recombinants in our sib sample. Thus, if c is the recombination rate, we expect ntc recombinants in our population sample and 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 9.17 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 total 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

$$\chi^2_{td} = \frac{(T - NT)^2}{(T + NT)} \tag{9.18}$$

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 $\sigma(M, y) \neq 0$, i.e., there is non-zero covariance between the marker allele state and the phenotype y 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 association. Thus it is 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.

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 (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 $S/\lambda \sim \chi_1^2$ (or $S \sim \lambda\chi_!^2$), where the inflation factor λ is given by

$$\lambda \simeq 1 + nF_{st} \sum_k (f_k - g_k)^2 \tag{9.19}$$

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 9.19 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

$$y_i = \mu + \beta X_i + e_i \tag{9.20}$$

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

$$QT = \left(\frac{\widehat{\beta}}{\text{SE}(\widehat{\beta})} \right)^2 \sim \lambda\chi_1^2 \tag{9.21}$$

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_1^2 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

$$\widehat{\lambda} = \frac{\text{medium}(S_1, \cdots S_m)}{0.456} \tag{9.22}$$

where we have done m tests, and S_i is the test statistic value for the ith of these.

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

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 9.20. 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,

$$y = \mu + \sum_{k=1}^{n} \beta_k \, M_k + \sum_{j=1}^{m} \gamma_j \, b_j + e \tag{9.23}$$

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.

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

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

Here $\mathbf{y}$ is the vector of observations, typically the means of each of the lines measured over a number of environments. β 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). $\mathbf{u}$ 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. $\mathbf{X}$, $\mathbf{S}$, $\mathbf{Q}$, and $\mathbf{Z}$ are matrices determined by the experimental design.

Lecture 9 Problems

1. 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?

2. Suppose two marker loci (m and n) are being followed in a group of affected-sib pairs. Marker 1 has three alleles (m_1, m_2, and m_3) while marker two has two allles (n_1, n_2). For these alleles, the following share/not share numbers were found

Marker Allele	Share	Not Share
m_1	30	25
m_2	40	10
m_3	25	30
n_1	33	30
n_2	35	38

What do these data say about linkage (if any) to a disease gene?

Solutions to Lecture 9 Problems

1. 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

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

2.

$$MLS(m_1) = 30 \cdot \log_{10}\left(\frac{2 \cdot 30}{55}\right) + 25 \cdot \log_{10}\left(\frac{2 \cdot 25}{55}\right) = 0.099$$

$$MLS(m_2) = 50 \cdot \log_{10}\left(\frac{2 \cdot 50}{60}\right) + 10 \cdot \log_{10}\left(\frac{2 \cdot 10}{60}\right) = 6.23$$

$$MLS(m_3) = 25 \cdot \log_{10}\left(\frac{2 \cdot 25}{55}\right) + 30 \cdot \log_{10}\left(\frac{2 \cdot 30}{55}\right) = 0.099$$

$$MLS(n_1) = 33 \cdot \log_{10}\left(\frac{2 \cdot 33}{63}\right) + 30 \cdot \log_{10}\left(\frac{2 \cdot 30}{63}\right) = 0.030$$

$$MLS(n_2) = 35 \cdot \log_{10}\left(\frac{2 \cdot 35}{73}\right) + 38 \cdot \log_{10}\left(\frac{2 \cdot 38}{73}\right) = 0.027$$

Hence, only linkage to marker allele m_2 is significant. This suggests that the disease allele is in linkage disequilibrium with allele m_2