

# Lecture 10

## QTL Mapping 2: Outbred Populations

Bruce Walsh. July 2005. Asian Institute on Statistical Genetics

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

Before considering the variety of QTL mapping methods for outbred populations, some comments on the probability that an outbred family is informative are in order. A parent is **marker-informative** if it is a marker heterozygote, **QTL-informative** if it is a QTL heterozygote, and simply **informative** if it is both. Unless *both* the marker and QTL are highly polymorphic, most parents will not be informative. Given the need to maximize the fraction of marker-informative parents, classes of marker loci successfully used with inbred lines may not be optimal for outbred populations. For example, SNPs 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 and hence much more likely to yield marker-informative individuals.

**Table 10.1** Types of marker-informative matings.

---

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

Parents are different marker heterozygotes.

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

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

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

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

**Intercross:**  $M_i M_j \times M_i M_j$

Both parents are the same marker heterozygote.

Only homozygous offspring informative in distinguishing alternative parental alleles.

---

*Note:* Here  $i, j, k,$  and  $\ell$  index different marker alleles.

---

As shown in Table 10.1, 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.

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

$$\text{PIC} = 1 - \sum_{i=1}^n p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2 p_i^2 p_j^2 \leq \frac{(n-1)^2(n+1)}{n^3} \quad (10.1)$$

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 10.1) 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,

$$z_{ijk} = \mu + s_i + m_{ij} + e_{ijk} \quad (10.2)$$

where  $z_{ijk}$  denotes the phenotype of the  $k$ th 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  $\sigma_s^2$  (the between-sire variance),  $\sigma_m^2$  (the between-marker, within-sibship variance), and  $\sigma_e^2$  (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

$$F = \frac{\text{MS}_m}{\text{MS}_e} \quad (10.3)$$

Assuming normality, Equation 10.3 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 10.3 has  $N$  and  $N(n-2)$  degrees of freedom.

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

$$E(\text{MS}_m) = \sigma_e^2 + (n/2)\sigma_m^2 \quad \text{and} \quad E(\text{MS}_e) = \sigma_e^2 \quad (10.4)$$

where

$$\sigma_m^2 = \frac{E(\text{MS}_m) - E(\text{MS}_e)}{n/2} = (1-2c)^2 \frac{\sigma_A^2}{2} \quad (10.5)$$

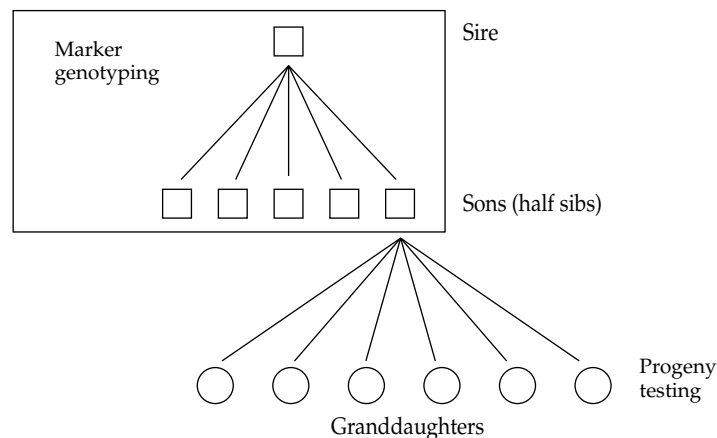
Thus, an estimate of the QTL effect (measured by its additive variance  $\sigma_A^2$ , scaled by the distance  $c$  between QTL and marker), can be obtained from the observed mean squares.

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  $\sigma_m^2$ . 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 sire is  $2p(1-p)$ . If this is small, even if  $a$  is large,  $\sigma_A^2$  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  $\sigma_m^2$ ), since here all families are informative, rather than the fraction  $2p(1-p)$  seen in 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 10.1), 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 characters in dairy cows, where the offspring are granddaughters of the original sires. The linear model for this design is

$$z_{ijkl} = \mu + g_i + m_{ij} + s_{ijk} + e_{ijkl} \quad (10.6)$$

where  $g_i$  is the effect of grandsire  $i$ ,  $m_{ij}$  is the effect of marker allele  $j$  ( $= 1, 2$ ) from the  $i$ th sire,  $s_{ijk}$  is the effect of son  $k$  carrying marker allele  $j$  from sire  $i$ , and  $e_{ijkl}$  is the residual for the  $l$ th 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 10.1** 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.

### 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. An alternative is to construct likelihood functions using the **variance components** associated with a QTL (or linked group of QTLs) in a genetic region of interest, rather than explicitly modeling all of the underlying genetic details. This approach allows for very general and complex pedigrees. The basic idea is to use marker information to compute the fraction of a genetic region of interest that is identical by descent between two individuals. Recall that two alleles are **identical by descent**, or **ibd**, if we can trace them back to a single copy in a common ancestor.

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

$$z_i = \mu + A_i + A_i^* + e_i \quad (10.7)$$

where  $\mu$  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  $\sigma_A^2$ ,  $\sigma_{A^*}^2$ , and  $\sigma_e^2$ . Here  $\sigma_A^2$  and  $\sigma_{A^*}^2$  correspond to the additive variances associated with the chromosomal region of interest and background QTLs in the remaining genome, respectively. We assume that none of these background QTLs are linked to the chromosome region of interest so that  $A$  and  $A^*$  are uncorrelated, and we further assume that the residual  $e$  is uncorrelated with  $A$  and  $A^*$ . Under these assumptions, the phenotypic variance is  $\sigma_A^2 + \sigma_{A^*}^2 + \sigma_e^2$ .

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

$$\sigma(z_i, z_j) = R_{ij} \sigma_A^2 + 2\Theta_{ij} \sigma_{A^*}^2 \quad (10.8)$$

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  $\mathbf{z}$  of observations on  $n$  individuals, the associated covariance matrix  $\mathbf{V}$  can be expressed as contributions from the region of interest, from background QTLs, and from residual effects,

$$\mathbf{V} = \mathbf{R} \sigma_A^2 + \mathbf{A} \sigma_{A^*}^2 + \mathbf{I} \sigma_e^2 \quad (10.9a)$$

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

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \hat{R}_{ij} & \text{for } i \neq j \end{cases}, \quad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases} \quad (10.9b)$$

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.

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

$$\ell(\mathbf{z} | \mu, \sigma_A^2, \sigma_{A^*}^2, \sigma_e^2) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp \left[ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right] \quad (10.10)$$

This likelihood has four unknown parameters ( $\mu$ ,  $\sigma_A^2$ ,  $\sigma_{A^*}^2$ , and  $\sigma_e^2$ ). A significant  $\sigma_A^2$  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  $\pi_{im}$  of alleles ibd at the marker of interest,

$$Y_i = a + \beta\pi_{im} + e \quad (10.11)$$

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

$$\beta = -2(1 - 2c)^2 \sigma_A^2 \quad (10.12)$$

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  $\sigma_A^2$ . The expected slopes for other pairs of relatives are

$$\beta = \begin{cases} -2(1 - 2c) \sigma_A^2 & \text{grandparent-grandchild;} \\ -2(1 - 2c)^2 \sigma_A^2 & \text{half-sibs;} \\ -2(1 - 2c)^2(1 - c) \sigma_A^2 & \text{avuncular (aunt/uncle-nephew/niece).} \end{cases} \quad (10.13)$$

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 ( $\sigma_A^2$ ) are confounded and cannot be separately estimated from the regression slope  $\beta$ . 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  $\sigma_A^2$  is possible by extending the H-E regression by using ibd status of two (or more) linked marker loci to estimate  $\pi_{jt}$ .

## 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  $\hat{p}_{ij}$  has mean  $p_{ij}$  and variance  $p_{ij}(1-p_{ij})/n_i$ . One ASP test is based on  $\hat{p}_{22}$ , the observed frequency of doubly affected pairs that have two marker alleles ibd. Under the assumption of no linkage,  $\hat{p}_{22}$  has mean  $1/4$  (as full sibs have a 25% chance of sharing both alleles ibd) and variance  $(1/4)(1-1/4)/n_2 = 3/(16n_2)$ , suggesting the test

$$T_2 = \frac{\hat{p}_{22} - 1/4}{\sqrt{\frac{3}{16n_2}}} \quad (10.14a)$$

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,  $p_{i1} + 2p_{i2}$ . 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

$$T_m = \sqrt{2n_2} (\hat{p}_{21} + 2\hat{p}_{22} - 1) \quad (10.14b)$$

which again for large samples is approximately distributed as a unit normal and is a one-sided test, as  $p_{21} + 2p_{22} > 1$  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  $\hat{p}_{2i} = n_{2i}/n_2$ . The LOD score for the test of no linkage becomes

$$MLS = \log_{10} \left[ \prod_{i=0}^2 \left( \frac{\hat{p}_{2i}}{\pi_{2i}} \right)^{n_{2i}} \right] = \sum_{i=0}^2 n_{2i} \log_{10} \left( \frac{\hat{p}_{2i}}{\pi_{2i}} \right) \quad (10.15)$$

where  $\pi_{2i}$  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 10.15 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

$$MLS = (1 - n_1) \log_{10} \left( \frac{1 - \hat{p}_1}{1/2} \right) + n_1 \log_{10} \left( \frac{\hat{p}_1}{1/2} \right) \quad (10.16)$$

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

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

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

Thus, the first marker shows significant evidence of linkage, while the second does not. Translating these LOD scores into LR values (the latter being distributed as a  $\chi^2$  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 know collections of relatives. This 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 maps is **Association mapping**. Here, one collects a large random set of individuals from the population and relies on linkage disequilibrium between very closely linked genes to do the mapping.

### 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 the 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  $(1 - c)^t \simeq e^{-ct}$ , 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  $\pi$  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

$$c = 1 - \pi^{1/t} \tag{10.17}$$

### Example

Hästbacka et al. (1992) examined the gene for diastrophic dysplasia (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:

Allele	Chromosome type			
	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  $\pi$  and  $t$ , Equation 10.17 gives an estimated recombination frequency between the *CSF1R* gene and the DTD gene as  $c = 1 - (0.95)^{1/100} \approx 0.00051$ . 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 to be 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).

### 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  $\chi^2$  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

$$\chi_{td}^2 = \frac{(T - NT)^2}{(T + NT)} \quad (10.18)$$

which follows a  $\chi^2$  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	<i>T</i>	<i>NT</i>	$\chi^2$	<i>P</i>
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.

## Lecture 10 Problems

1. Consider an outbred population. The allele frequencies at a marker are  $\text{freq}(M) = 0.3$  and  $\text{freq}(m) = 0.7$ , while the allele frequencies at a QTL are  $\text{freq}(Q) = 0.1$  and  $\text{freq}(q) = 0.9$ .
  - 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 equilibrium)?
  - 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 alleles ( $n_1, n_2$ ). 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 10 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.  $\text{Prob}(\text{Not informative}) = 1 - 0.0756 = 0.9244$ .  $\text{Prob}(n \text{ individuals all not informative}) = 0.9244^n$ .  $\text{Prob}(\text{at least one informative individual}) = 1 - 0.9244^n$ . Solve for  $n$  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  $m_2$