Lecture 16

Major Genes, Polygenes, and QTLs

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

Notes from a short course taught May 2011 at University of Liege

Major and Minor Genes

While the machinery of quantitative genetics can blissfully function in complete ignorance of any of the underlying genetic details, we are certainly interested (at least one some level) in the genetic basis of trait variation. At the most practical level, if a gene of major effect is segregating in our population of interest, we would certainly like to be aware of this, as well as being able fine map it for future exploration/exploitation. Often no single locus contributes more than a fraction of the total genetic variation in a trait. In such cases, we often use the term **polygene** for one of these loci of small effect. The (still evolving) more modern use is to call these genes **quantitative trait loci**, or simply **QTL** or **QTLs** for short. Lecture 16 deals with statistical approaches for QTL mapping, while we focus here first on the genetic basis of quantitative trait variation, and then on methods for first detecting and then fine mapping genes of large effect. Included as a special case of the later are so-called **candidate genes**, loci for which we have biological information to suggest that they may contribution variation to the trait of interest.

Major Genes and Isoalleles

The honest truth is that we have very little understanding of the genetic basis of quantitative trait variation. But we have plenty of ideas and hypotheses. One of the simplest is the notion of **isoalleles**. Suppose we have a locus, say *polled*, which has an allele that gives cattle without horns. This is clearly an example of an allele with a dramatic effect. Might it also be the case that the polled locus also has other alleles with far less dramatic effects, say a 5-10 percent reduction in horn length? This is the idea of *isoalleles*, namely that a locus than can be detected because of it has mutant alleles with rather dramatic effects on phenotypes may also be segregating alleles with much subtler effects. While there is some evidence for such loci in *Drosophila*, as a whole, the support (to date) is only modest at best. If the isoallele model is correct, then genes known to have alleles showing a major effect on the character of interest are certainly candidate loci. Hence, a gene with a mutant allele giving a dramatic weight difference in mice would also be a candidate gene in livestock to search for alleles having less dramatic effetcs on body weight.

The general belief is that quantitative variation is more likely generated not by **structural** changes in genes (i.e., changes in the amino acid sequence) but rather by **regulatory** changes – changes in the amount and timing of a gene's product. Humans and chimps are roughly 99.9% identical in amino acid sequences, yet there are clearly fundamental differences. If most of the proteins are essentially identical between the species, much of the differences must reside in differences in the timing and amount of these otherwise identical products. Of course, the two blur in that **transcription regulators** (proteins that bind to specific sequences of DNA to modulate levels of mRNA transcription) can cause a regulatory change in any number of genes as a result of a structural change (i.e., an amino acid change that alters the DNA binding specificy) in the transcription factor.

We can classify regulatory changes affecting a gene as those that act in **cis** and those that act in **trans**. Cis-acting regulatory changes only act on the specific DNA molecular on which they reside, for example, mutants that effect promoter or enhancer strength of a gene, its ability to splice properly, etc. While cis-acting sequences can theoretically involve transcription units other than the gene target, most cis-regulatory mutants are expected to simply be changes within the target gene region itself. However, given that we are still rather ignorant of many features of fine-control of gene regulation, a cis-acting site could be many kilobases (or even megabases) away.

In contrast, trans-acting sequences make diffusable products (read RNAs and/or proteins) that can influence genes on other chromosomes. Transcription factors are an excellent example of this. The current data from microarray studies shows that trans-effects are much more common than cis-effects. Hence, if we can shows that (say) increasing the level of gene expression (measured by amount of mRNA) in gene X will improve our trait of interest, direct selection on gene X (say by marker-assisted selection) will have little effect if up-regulation is controlled by other genes acting in trans on X. We instead need to directly select on these trans-acting genes. Trans-acting factors are an example of a **modifier** – a gene that effects the phenotype produced by another gene. QTLs that influence the expression of another gene (or genes) are called **eQTs**, for **expression QTLs**.

Polygenic Mutation and the Mutational Variance, σ^2_m

While mutation rates for a gene have been traditionally measured by looking for gross changes in some phenotype, it is clear from the above discussion that mutations not only in the coding region of a gene, but also in any region of DNA that can influence its regulation, can potentially generate some quantitative variation. Hence, simply giving a mutation rate is not sufficient, as we must also account for the phenotypic effects of new mutants. The natural measure is the **polygenic mutational variance**, σ^2_m, which is the amount of new additive variation introduced to the trait each generation by mutation. A wide variety of studies in model systems suggests that $\sigma^2_m \simeq 10^{-3}\sigma^2_e$ is the typical order of magnitude for such variation. The mutational input of new variation is thus of considerable importance, even in short-term selection. For long-term selection, after the initial variation is exhausted, all further response is due to the effects of new mutation.

Simple Tests for Detecting Major Genes

Suppose our trait of interest shows strong resemblance between relatives, suggesting a genetic basis for some of the variation (provided shared environmental effects can be safely ignored). How do we determine if most of the genetic variation is due to segregation of alleles at a single locus (i.e., a major gene is present). The simplest indication would be phenotypes falling into a few discrete classes such as horns vs. no horns. However, even in such apparently simple cases, we are far from proving the involvement of major genes. Such differences could be largely environmental, such as exposure to a virus or heat stress. Further, even if the trait is binary (i.e., presence/absence), it may still have a very complicated gene basis, with no single gene accounting for more than just a small fraction of the probability that an individual shows the trait.

Keeping these caveats in mind, in the absence of clear breaks in the phenotypic distribution, another approach is see whether the trait distribution shows **multimodialty** — i.e., it has several distinct peaks, not just one. The logic for this observation is that if Q and q are alleles at some underlying locus influencing the trait variation, and $p_{XY}(z)$ is the distribution of trait (z) values for an individual with a XY genotype, then the total trait distribution can be written as a **mixture model**:

$$p(z) = p_{QQ}(z)\,\Pr(QQ) + p_{Qq}(z)\,\Pr(Qq) + p_{qq}(z)\,\Pr(qq) \qquad (16.1)$$

If each of the conditional distributions of trait value given genotype is itself a unimodial (for example, if $p(z|XY)$ is normal), then if the modes the conditional distributions are sufficient far apart (as might be expected with a major gene), then resulting mixture distribution could show several peaks. Even if multiple peaks are not obvious, maximum likelihood can be used to test for whether a mixture fits the data better than a single unimodial distribution, a point we return to shortly.

A second simple test for the presence of major genes is to look at the within-family variance. If a large number of genes of roughly equal effect are involved, then the distribution of trait values should largely be the same for each family, regardless of the genotype at any particular locus. On the other hand, if a major gene is involved, some families are expected to show much less variation than others. For example, a $QQ \times QQ$ family will have only QQ offspring. If the Q locus accounts for most of the genetic variation, the within-family variance in such families will be much smaller than in families where both Q and q are segregating. A simple test for heterogeneity of variances across families can be performed to see if such differences in variances across families are present.

Again, both of these simple tests (multimodality and heterogeneity of within-family variance) are simply *suggestive* of a major gene. If such signals are seen, then it is worth employing the more involved method of complex segregation analysis. As a lead-in to this method, we start with some background on mixture distributions.

Mixture Models

Mixture models appear widely in quantitative genetics, largely because we can decompose the total trait distribution into a weighted sum of conditional distributions over the various genotypes of interest. The basic structure of a mixture model is as follows. Assume the distribution of interest results from a weighted mixture of several underlying distributions. If there are $i = 1, \cdots, n$ underlying distributions, $p_1(z), \cdots, p_n(z)$, each with frequency $\Pr(i)$, the resulting probability density of an observed variable z is given by a generalization of Equation 16.1,

$$p(z) = \sum_{i=1}^{n} \Pr(i) \cdot p_i(z)$$

It is usually assumed that the underlying distributions are normals, so this becomes

$$p(z) = \sum_{i=1}^{n} \Pr(i) \cdot \varphi(z, \mu_i, \sigma_i^2) \tag{16.2}$$

where

$$\varphi(z, \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \, \exp\left[-\frac{(z-\mu_i)^2}{2\sigma_i^2} \right]$$

is the probability density function for a normally distributed random variable with mean μ_i and variance σ_i^2. Equation 16.2 has $3n-1$ parameters to estimate: the $n-1$ mixing proportions $\Pr(i)$, and the n means and n variances of the underlying distributions. It is usually assumed that all the variances are equal, reducing the number of unknown parameters to $2n$ ($n-1+n+1$). Various genetic hypotheses allow us to further specify and evaluate the structure of the mixing proportions.

As an example of how likelihood functions are constructed, consider the situation for a random individual drawn from a population with a single segregating diallelic major locus. Indexing the three genotypes by i where $i =$ *QQ*, *Qq*, and *qq*, and assuming that individuals with major-locus genotype i are normally distributed with mean μ_i and common variance σ^2, the resulting likelihood for the jth individual is

$$\ell(z_j) = \Pr(QQ)\, p_{QQ}(z_j) + \Pr(Qq)\, p_{Qq}(z_j) + \Pr(qq)\, p_{qq}(z_j) \tag{16.3a}$$

$$= \Pr(QQ)\, \varphi(z_j, \mu_{QQ}, \sigma^2) + \Pr(Qq)\, \varphi(z_j, \mu_{Qq}, \sigma^2) + \Pr(qq)\, \varphi(z_j, \mu_{qq}, \sigma^2)$$

where z_j is the character value in the focal individual. For n random (unrelated) individuals, denoting the observed phenotypic values by $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, the overall likelihood is just the product of the n individual likelihoods,

$$\ell(\mathbf{z}) = \ell(z_1, z_2, \cdots, z_n) = \prod_{j=1}^{n} \ell(z_j) \tag{16.3b}$$

Assuming random mating, the Hardy-Weinberg principle describes the frequencies $\Pr(\cdot)$ of the major locus genotypes as a function of p, the frequency of one allele. This leaves five parameters to estimate — p, σ^2, μ_{QQ}, μ_{Qq}, and μ_{qq}.

An important issue in tests for major genes is model fitting, i.e., evaluating whether the full model is needed, or if some subset of the model gives essentially the same fit. For example, we might initially assume a mixture of two normals with different means and common variance, so that the full model has parameters μ_1, μ_2, σ^2, and p. Is the fit using these four parameters significantly better than the fit assuming a single underlying normal with parameters μ and σ^2? For large sample sizes, the **likelihood ratio** (LR) statistic test for whether the full model provides a better fit than a particular subset of the model is

$$\Lambda(\mathbf{z}) = -2\ln\left[\frac{\widehat{\ell}_r(\mathbf{z})}{\widehat{\ell}(\mathbf{z})}\right] = -2\left\{\ln\left[\widehat{\ell}_r(\mathbf{z})\right] - \ln\left[\widehat{\ell}(\mathbf{z})\right]\right\} \tag{16.4}$$

where $\widehat{\ell}(\mathbf{z})$ is the likelihood function evaluated at the MLE for the full model, and $\widehat{\ell}_r(\mathbf{z})$ is the maximum of the likelihood function for the restricted model under which r parameters of the full model are assigned fixed values. Under appropriate conditions, the LR test statistic is approximately distributed as χ^2_r, i.e., as a χ^2 distribution with r degrees of freedom.

We need to stress that for all its sophistication, the above likelihood test that an observed distribution is better fit by a mixing distribution than a single normal is again only *suggestive* of a major gene. For example, such a mixture distribution can result from individuals experiencing distinct environments. A more formal approach to using likelihood to demonstrate Mendelian inheritance of a major gene is the method of **Complex Segregation Analysis**, developed by human geneticists to search for disease genes.

Complex Segregation Analysis

The feature that distinguishes complex segregation analysis (CAS) from simply fitting a mixture distribution to data is that CAS requires a pedigree of individuals, and explicitly follows (and tests) for Mendelian transmission of genetic factors within the pedigree. By contrast, the above likelihood test for a better fit under a mixture model assumes a random sample of individuals from the population. Any feature that causes individuals to group into different conditional distributions for trait value given the group (such as different environments) will generate a mixture distribution. CAS goes a step further by explicitly testing for Mendelian transmission within the pedigree. Since we must follow all genes through a pedigree, CAS is much more complex, and computationally intense, than a simple mixture model test.

Likelihood Functions Assuming a Single Major Gene

We start by computing the likelihood for a single individual, then proceed to an entire family, and finally to the collection of all families in our sample. Assume that a single diallelic locus underlies the character and consider the jth offspring from the ith family, o_{ij}, which has father f_i and mother m_i (for notational ease, in the following we use f, m, and o_j, reminding the reader that these, of course, change as we change families). Denote the phenotypic value of this offspring by z_{ij}. Index the major-locus genotypes by g where $g = 1$ for *QQ*, $g = 2$ for *Qq*, and $g = 3$ for *qq*, with g_f, g_m, and g_{o_j} denoting the genotypes of the parents (father and mother) and their jth offspring. Phenotypic values for each major-locus genotype are assumed to be normally distributed with means μ_g and common variance σ^2. Finally, let $\Pr(g_o \,|\, g_f,\, g_m)$ be the probability that an offspring has genotype g_o given that its parents have genotypes g_f and g_m.

Conditioned on the parental genotypes, the likelihood for the ijth offspring is

$$\ell(z_{ij}\,|\,g_f,\,g_m) = \sum_{g_o=1}^{3} \Pr(g_o\,|\,g_f,\,g_m) \cdot \varphi(z_{ij}, \mu_{g_o}, \sigma^2) \tag{16.5a}$$

This conditional likelihood is a mixture model with mixing proportions given by Mendelian segregation. For example, if the father and mother have major-locus genotypes QQ and Qq, then $g_f = 1$ and $g_m = 2$, and

$$\begin{aligned}
\Pr(g_o = 3\,|\,g_f = 1,\, g_m = 2) &= \Pr(qq\,|\,g_f = QQ,\, g_m = Qq) = 0 \\
\Pr(g_o = 2\,|\,g_f = 1,\, g_m = 2) &= \Pr(Qq\,|\,g_f = QQ,\, g_m = Qq) = 1/2 \\
\Pr(g_o = 1\,|\,g_f = 1,\, g_m = 2) &= \Pr(QQ\,|\,g_f = QQ,\, g_m = Qq) = 1/2
\end{aligned} \tag{16.5b}$$

so that with these parents Equation 16.5a reduces to

$$\ell(z_{ij}\,|\,QQ, Qq) = \frac{1}{2} \cdot \varphi(z_{ij}, \mu_{QQ}, \sigma^2) + \frac{1}{2} \cdot \varphi(z_{ij}, \mu_{Qq}, \sigma^2) \tag{16.5c}$$

Conditioned on parental genotype values, each offspring in a family is independent, implying that the likelihood for a full-sib family of n_i offspring is the product of individual likelihoods, giving the conditional likelihood for the ith family as

$$\ell(z_{i\cdot}\,|\,g_f,\,g_m) = \prod_{j=1}^{n_i} \ell(z_{ij}\,|\,g_f,\,g_m) \tag{16.6a}$$

Since we do not know the genotypes of the parents, the unconditional likelihood for the ith family is obtained by summing over all nine possible pairs of parental genotypes,

$$\ell(z_{i\cdot}) = \sum_{g_f=1}^{3} \sum_{g_m=1}^{3} \ell(z_{i\cdot}\,|\,g_f,\,g_m)\,\text{freq}(g_f,\,g_m) \tag{16.6b}$$

Assuming the parents are chosen independently, freq$(g_f,\,g_m) = \Pr(g_f) \cdot \Pr(g_m)$. Further, if genotypes are in Hardy-Weinberg proportions, parental genotype frequencies are completely specified by the frequency p of allele Q, e.g.,

$$\begin{aligned}
\text{freq}(g_f = 1,\, g_m = 1) &= \text{freq}(g_f = QQ) \cdot \Pr(g_m = QQ) = p^2 \cdot p^2 \\
\text{freq}(g_f = 2,\, g_m = 1) &= \text{freq}(g_f = Qq) \cdot \Pr(g_m = QQ) = 2p(1-p) \cdot p^2, \text{ etc.}
\end{aligned}$$

If there are $n_g > 3$ major-locus genotypes (either because of multiple alleles at the major locus or because of several major loci), the appropriate likelihood has sums ranging over the n_g genotypes, and the transmission probabilities are modified to account for the assumed model. Likewise, if the parental phenotypic values (z_f, z_m) are known, these can also be incorporated into the likelihood. Since $\ell(z\,|\,g) = \varphi(z, \mu_g, \sigma^2)$, the probability that the genotype is g_i given the phenotype is z is

$$\Pr(g_i\,|\,z) = \frac{\Pr(g_i)\;\; \varphi(z, \mu_{g_i}, \sigma^2)}{\displaystyle\sum_{j=1}^{n_g} \Pr(g_j)\;\; \varphi(z, \mu_{g_j}, \sigma^2)} = \frac{\Pr(g_i)\;\; \varphi(z, \mu_{g_i}, \sigma^2)}{p(z)} \tag{16.7}$$

where $p(z)$ is the phenotypic density function for the entire population. Parental phenotypes are then incorporated by replacing $\Pr(g)$ by $\Pr(g\,|\,z)$. Equation 16.7 follows directly from Bayes' theorem (Lecture 1).

Assuming different families are unrelated, the total likelihood is the product of the individual likelihoods from the n_f families,

$$\ell(\mathbf{z}) = \prod_{i=1}^{n_f} \ell(z_{i\cdot}) \tag{16.8}$$

where $\ell(z_{i\cdot})$ is given by Equation 16.6b. Although there are numerous summation and product indices in this likelihood, there are only five unknown parameters: the three genotypic means, the common variance σ^2, and the major allele frequency p.

While the most obvious test for a major gene compares the full model with the restricted model of a single underlying normal, a much more robust approach is to treat the transmission probabilities $\Pr(g_o \,|\, g_f, g_m)$ as unknown parameters and base hypothesis tests on these. Above, we specified the transmission probabilities based on Mendelian assumptions of inheritance (e.g., Equation 16.5b), but we can also treat them as parameters to be estimated. This is most conveniently done by considering τ_x, the probability that genotype x transmits a Q allele. For a diallelic locus, there are three τ values to estimate, one for each genotype. From the definition of τ, the transmission probabilities can be expressed as

$$\begin{aligned}
\Pr(\,qq \,|\, g_f,\, g_m) &= (1-\tau_{g_f})\,(1-\tau_{g_m}) \\
\Pr(\,Qq \,|\, g_f,\, g_m) &= \tau_{g_f}\,(1-\tau_{g_m}) + \tau_{g_m}\,(1-\tau_{g_f}) \\
\Pr(\,QQ \,|\, g_f,\, g_m) &= \tau_{g_f}\,\tau_{g_m}
\end{aligned} \tag{16.9}$$

For example, Equations 16.5b become

$$\begin{aligned}
\Pr(qq \,|\, g_f = QQ,\, g_m = Qq) &= (1-\tau_{QQ})\,(1-\tau_{Qq}) \\
\Pr(Qq \,|\, g_f = QQ,\, g_m = Qq) &= \tau_{QQ}\,(1-\tau_{Qq}) + \tau_{Qq}\,(1-\tau_{QQ}) \\
\Pr(QQ \,|\, g_f = QQ,\, g_m = Qq) &= \tau_{QQ}\,\tau_{Qq}
\end{aligned} \tag{16.10}$$

so that with these parents, Equation 16.5c becomes

$$\begin{aligned}
\ell(z_{ij} \,|\, QQ, Qq) = \tau_{QQ}\,\tau_{Qq} \cdot \varphi(z_{ij}, \mu_{QQ}, \sigma^2) \\
+ \left[\,\tau_{QQ}\,(1-\tau_{Qq}) + \tau_{Qq}\,(1-\tau_{QQ})\,\right] \cdot \varphi(z_{ij}, \mu_{Qq}, \sigma^2) \\
+ (1-\tau_{QQ})\,(1-\tau_{Qq}) \cdot \varphi(z_{ij}, \mu_{qq}, \sigma^2)
\end{aligned}$$

Note that this likelihood reduces to Equation 16.5c using Mendelian segregation transmission probabilities ($\tau_{QQ} = 1$ and $\tau_{Qq} = 1/2$).

Three criteria must be satisfied for acceptance of a major-gene hypothesis: (1) a significantly better overall fit of a mixture model compared with a single normal, (2) failure to reject the hypothesis of Mendelian segregation ($\tau_{QQ} = 1$, $\tau_{Qq} = 1/2$, $\tau_{qq} = 0$), and (3) rejection of the hypothesis of equal transmission for all genotypes ($\tau_{QQ} = \tau_{Qq} = \tau_{qq}$). Criterion (1) reduces false positives due to polygenic background loci, while criteria (2) and (3) offer some robustness against nonnormality of the underlying distributions and resemblance due to common environmental effects. While incorporation of transmission-probability criteria into likelihood models decreases the possibility of a false positive, it does so at a cost of decreased power. Loss of power can be significant if the major gene is recessive.

The fact that not all families are expected to be segregating the major gene has important consequences for the optimal number and size of families for detecting a major gene. For a fixed number of individuals, highest power is generally obtained by examining a moderate number of families of moderate size, as opposed to many small families or a few large families. If a small number of large families is chosen, we run the risk that none of the families are segregating the gene. Conversely, with a large number of small families, while some are likely to have the gene segregating, power for detecting a major gene is reduced due to the small sample size in each segregating family.

Common-family Effects

Members of full-sib families usually share environmental effects, and likelihood functions accounting for these have been developed. Let the ith family have a common effect c_i, and assume that these effects are normally distributed among families with mean zero and variance σ_c^2. With this modification, the expected phenotypic value of an offspring with genotype g_o from family i is $\mu_{g_o} + c_i$. As before, we assume that the phenotypic values for each genotype (conditional on c_i) are normally distributed with variance σ^2, giving the conditional likelihood for the n_i offspring from this family as

$$\ell(z_{i\cdot} \,|\, g_f,\, g_m,\, c_i) = \prod_{j=1}^{n_i} \left[\sum_{g_{o_j}=1}^{3} \Pr(g_{o_j} \,|\, g_f,\, g_m) \cdot \varphi(z_{ij}, \mu_{g_{o_j}} + c_i, \sigma^2) \right] \tag{16.11}$$

Averaging over all possible values of the common-family effect c_i gives

$$\ell(z_{i\cdot} \,|\, g_f,\, g_m) = \int_{-\infty}^{\infty} \ell(z_{i\cdot} \,|\, g_f,\, g_m,\, c) \cdot \varphi(c, 0, \sigma_c^2)\, dc \tag{16.12}$$

Finally, using the above expression for $\ell(z_{i\cdot} \,|\, g_f,\, g_m)$, averaging over all possible parental genotypes gives the unconditional likelihood for this family (Equation 16.6b). Assuming the QTL genotypes are in Hardy-Weinberg proportions, the unconditional likelihood has six unknown parameters: the three genotypic means, the allele frequency p, and the variances σ^2 and σ_c^2. Assuming the n_f families in our pedigree are unrelated, the total likelihood is the product of the individual family likelihoods (Equation 16.8).

The likelihood for the ith family under the restricted model assuming common-family effects, but no major genes, is

$$\begin{aligned} \ell(z_{i\cdot}) &= \int_{-\infty}^{\infty} \ell(z_{i\cdot} \,|\, c) \cdot \varphi(c, 0, \sigma_c^2)\, dc \\ &= \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \varphi\left(z_{ij}, \mu + c, \sigma^2\right) \right] \cdot \varphi\left(c, 0, \sigma_c^2\right)\ dc \end{aligned} \tag{16.13}$$

A test for common-family effects but no major gene is given by the likelihood-ratio test using Equation 16.13 versus the likelihood function with $\sigma_c^2 = 0$. The latter is just the likelihood function assuming a single underlying normal. Likewise, the likelihood-ratio test for a major gene but no common-family effects uses the full likelihood and a restricted likelihood assuming $\sigma_c^2 = 0$.

Similar modifications to allow for a polygenic background (in addition to the major gene) have also been developed.

Genetic Maps and Candidate Genes

Suppose CAS, or some other approach, convinces us that a major gene is segregating in our population. The next step is to **map** or localize the gene with respect to a set of known molecular markers. Genetic localization of the major gene allows for more rapid introgression into other strains, for tests of the presence/absence of the gene, and for marker assisted selection.

The metric of genetic distance is whether recombination occurs between markers. This allows us to simply use recombination to map genes. **Physical mapping** of genes occurs by sequencing, allowing us to state the relationship between genes an markers in terms of actual DNA base pair differences.

Map Distances vs. Recombination Frequencies

Genetic map construction involves both the ordering of loci and the measurement of distance between them. Ideally, distances should be additive so that when new loci are added to the map, previously obtained distances do not need to be radically adjusted. Unfortunately, recombination frequencies are not additive and hence are inappropriate as distance measures. To illustrate, suppose that three loci are arranged in the order *A*, *B*, and *C* with recombination frequencies c_{AB}, c_{AC}, and c_{BC}. Each recombination frequency is the probability that an odd number of crossovers occurs between the markers, while $1 - c$ is the probability of an even number (including zero). There are two different ways to get an odd number of crossovers in the interval *A*–*C*: an odd number in *A*–*B* and an even number in *B*–*C*, or an even number in *A*–*B* and an odd number in *B*–*C*. If there is no **interference**, so that the presence of a crossover in one region has no effect on the frequency of crossovers in adjacent regions, these probabilities can be related as

$$c_{AC} = c_{AB}\,(1 - c_{BC}) + (1 - c_{AB})\,c_{BC} = c_{AB} + c_{BC} - 2c_{AB}\,c_{BC} \qquad (16.14)$$

This is **Trow's formula**. More generally, if the presence of a crossover in one region depresses the probability of a crossover in an adjacent region,

$$c_{AC} = c_{AB} + c_{BC} - 2(1 - \delta)c_{AB}\,c_{BC} \qquad (16.15)$$

where the **interference parameter** δ ranges from zero if crossovers are independent (no interference) to one if the presence of a crossover in one region completely suppresses crossovers in adjacent regions (complete interference).

Thus, in the absence of very strong interference, recombination frequencies can only be considered to be additive if they are small enough that the product $2c_{AB}c_{BC}$ can be ignored. This is not surprising given that the recombination frequency measures only a part of all recombinant events (those that result in an odd number of crossovers). A map distance m, on the other hand, attempts to measure the total number of crossovers (both odd and even) between two markers. This is a naturally additive measure, as the number of crossovers between *A* and *C* equals the number of crossovers between *A* and *B* plus the number of crossovers between *B* and *C*.

A number of **mapping functions** attempt to estimate the number of cross-overs (m) from the observed recombination frequency (c). The simplest, derived by Haldane (1919), assumes that crossovers occur randomly and independently over the entire chromosome, i.e., no interference. Let $p(m,k)$ be the probability of k crossovers between two loci m map units apart. Under the assumptions of this model, Haldane showed that $p(m,k)$ follows a Poisson distribution, so that the observed fraction of gametes containing an odd number of crossovers is

$$c = \sum_{k=0}^{\infty} p(m, 2k+1) = e^{-m} \sum_{k=0}^{\infty} \frac{m^{2k+1}}{(2k+1)!} = \frac{1 - e^{-2m}}{2} \qquad (16.16)$$

where m is the expected number of crossovers. Rearranging, we obtain Haldane's mapping function, which yields the (Haldane) map distance m as a function of the observed recombination frequency c,

$$m = -\frac{\ln(1 - 2c)}{2} \qquad (16.17)$$

For small c, $m \simeq c$, while for large m, c approaches 1/2. Map distance is usually reported in units of **Morgans** (after T. H. Morgan, who first postulated a chromosomal basis for the existence of linkage groups) or as **centiMorgans** (cM), where 100 cM = 1 Morgan. For example, a Haldane map distance of 10 cM ($m = 0.1$) corresponds to a recombination frequency of $c = (1 - e^{-0.2})/2 \simeq 0.16$.

Although Haldane's mapping function is frequently used, several other functions allow for the possibility of crossover interference in adjacent sites. For example, geneticists often use Kosambi's (1944) mapping function, which allows for modest interference,

$$m = \frac{1}{4}\ \ln\left(\frac{1+2c}{1-2c}\right) \tag{16.18}$$

Lecture 16 Problems

1. Suppose the observed recombination frequencies c between three loci (A, B, C) are as follows:

$$c_{AB} = 0.20, \quad c_{BC} = 0.05, \quad c_{AC} = 0.17$$

a: What is the gene order (i.e., which locus is in the middle?)

b: Compute the Haldane map distances between these loci.

2. a: Suppose in a small closed population there is a allele segregating that causes dogs to have red hair. You know from pedigree studies that this trait first appeared 10 generations ago. Suppose you have the following trait-marker information:

Marker locus

allele	Freq. in normal	Freq in red chromosomes
1	36%	96%
2	64%	4%

Marker locus 2

allele	Freq. in normal	Freq in red chromosomes
1	16%	12%
2	30%	35%
3	54%	53%

a: Is marker locus 1 linked to this gene? What about Marker locus 2?

b: What is the estimated recombination frequency between the linked marker(s) and this gene?

Solutions to Lecture 16 Problems

1. a: map order: A ——— C —- B

b: $m = -\ln(1-2c)/2$, so that $m_{AB} = 0.26$, $m_{AC} = 0.21$, and $m_{BC} = 0.05$. Notice that while the recombination frequencies are not additive (i.e., $c_{AB} \neq c_{AC} + c_{BC}$), while the Haldane distances are.

2. a: Locus 1 is linked, locus 2 is unlinked.

b: The mutation appears to have arisen in the allele 1 background, so that $\pi = 0.96$. From Equation 16.19,

$$c = 1 - \pi^{1/t} = 1 - 0.96^{1/10} = 0.0041$$