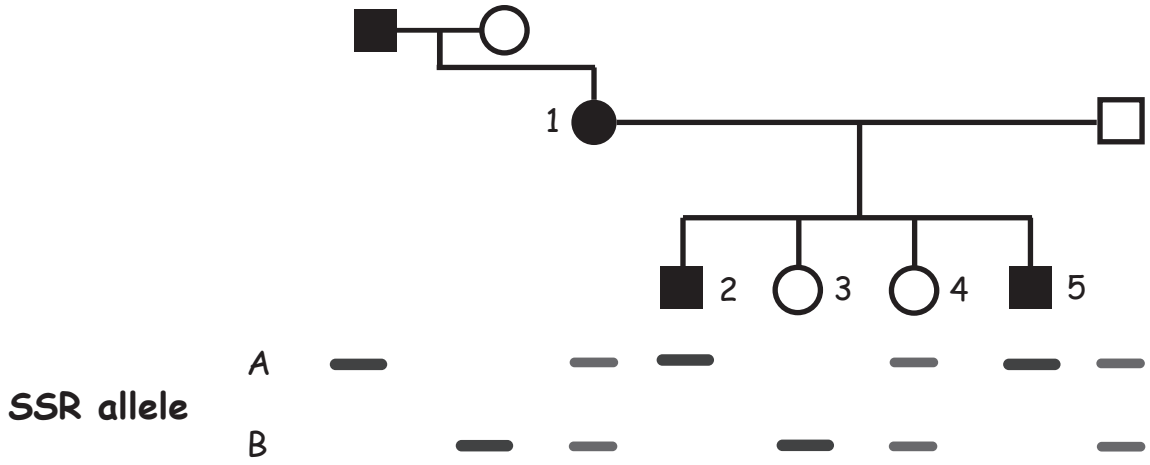
Lecture 10

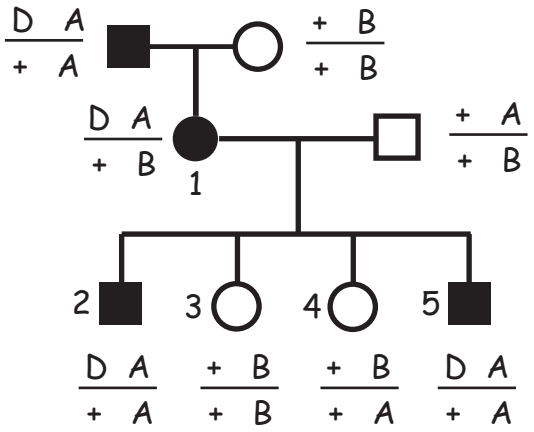
We will now consider how genetic linkage and genetic mapping experiments are done in humans. The basic principle of genetic mapping in humans is the same as that for other diploid organisms - genetic distance is determined as an inverse measure of the probability of meiotic crossovers in a given interval. However, in practice human linkage analysis differs from that of experimental organisms in that only a relatively small number of meiosis events (offspring) can be considered. The maximum is about 100 and these must occur in different families. For these reasons human linkage analysis requires many markers distributed across the genome.

The most common markers used for this purpose are DNA based genetic markers such as SNPs and SSRs. SSRs are frequently used because they are easy to detect and they often exist in many allelic forms (remember that for a marker to be useful in mapping it must be heterozygous). More than 20,000 SSRs have been identified. A set of about 400 highly polymorphic (many different alleles of different repeat lengths) that are spaced about every 10 cM are used for whole genome scans for linkage.

To see how a human pedigree can be evaluated for linkage, Consider the segregation of a dominant trait and a possibly linked SSR marker in the following pedigree.



The first step is to annotate the pedigree with the genotypes that we can deduce. We will use D to designate the allele for the dominant trait and + to designate the corresponding wild type allele. A and B will designate the two SSR alleles.



In this pedigree we are considering co-segregation of the allele for the dominant trait and the SSR allele during meiosis in the mother (individual 1) who is heterozygous for both the trait and the SSR. [We know she he inherited both D and SSR allele A from her father and the + and SSR allele B from her mother.]

We can see that among the woman's four children they inherit from their mother either D A or + B. Thus all four children inherited "parental" gamete genotypes from their mother. We would like to know the probability that D and the SSR marker are linked given the information from this family. This is exactly the type of probability calculation that Bayes theorem is useful for. In human genetics an alternative expression of Bayes theorem is used. To derive this expression consider the standard expression of Bayes Theorem for a condition X and its converse, not X ($\overline{X}$).

$$p(X|Y) = \frac{p(Y|X) \cdot p(X)}{p(Y)}$$

$$p(\overline{X}|Y) = \frac{p(Y|\overline{X}) \cdot p(\overline{X})}{p(Y)}$$

Rearranging we can equate the two different expressions for p(Y) :

$$\frac{p(Y|\overline{X}) \cdot p(\overline{X})}{p(\overline{X}|Y)} = p(Y) = \frac{p(Y|X) \cdot p(X)}{p(X|Y)}$$

Thus:

$$\frac{p(X|Y)}{p(\overline{X}|Y)} = \frac{p(Y|X)}{p(Y|\overline{X})} \cdot \frac{p(X)}{p(\overline{X})}$$

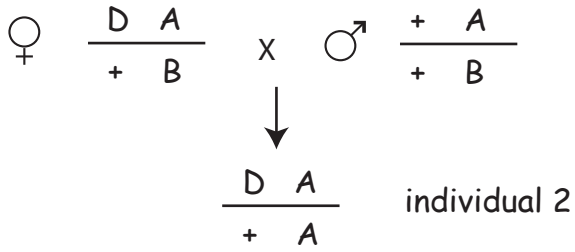
Posterior odds = Odds ratio · Prior odds

To use this equation to calculate the probability of linkage we will consider X = linked and $\overline{X}$ = not linked and Y represents the segregation data we obtained. Using the standard of $p > 0.95$ as the threshold for significance, posterior odds of linkage given the observed data should be ~ 20/1. The prior odds for linkage are about 1/50 thus for significance the odds ratio should be >1,000. The key calculation in linkage analysis is the odds ratio. By convention his is calculated as a LOD score (LOD means Log of the odds ratio).

$$LOD = \log_{10} \frac{P \text{ (observed segregation | loci completely linked)}}{P \text{ (observed segregation | loci unlinked)}}$$

A LOD score of > 3 is needed for publishable significance. (The log is used for simplicity - LOD scores from different families can be added to give a total score > 3.)

For our family shown we will calculate a LOD score for each of the 4 children.
Starting with the son with the trait (individual 2) we can write out his genotype and that of his parents as follows:



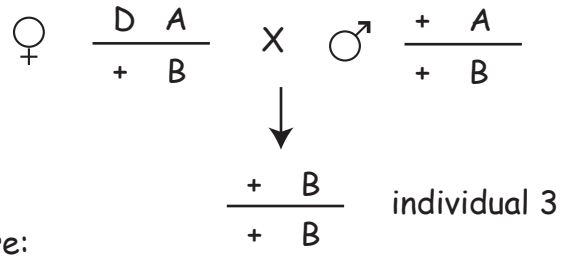
Next we calculate the two relevant conditional probabilities:

p(obtaining his genotype | complete linkage) = 1/2 • 1/2 = 1/4

p(obtaining his genotype | unlinked) = 1/2 • 1/4 = 1/8

Thus the LOD score for individual 2 is : LOD = $\log_{10} 2 = 0.3$

For individual 3 the relevant genotypes are:



And the two relevant conditional probabilities are:

p(obtaining this genotype | complete linkage) = 1/2 • 1/2 = 1/4

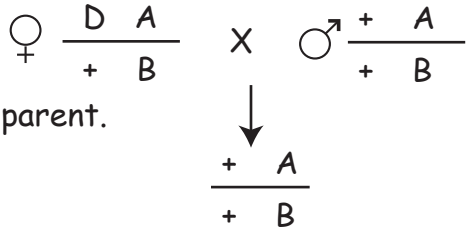
p(obtaining his genotype | unlinked) = 1/2 • 1/4 = 1/8

Thus the LOD score for individual 3 is : LOD = $\log_{10} 2 = 0.3$

Since $\log_{10}(a \cdot b) = \log_{10}(a) + \log_{10}(b)$, the aggregate LOD score for individuals 2 and 3 can be obtained by adding the individual scores and thus LOD = 0.6.

By inspection we can see that individual 5 is the same as individual 2 so individual 5 will contribute 0.3 to the total LOD score. Thus considering individuals 2, 3, and 5 the aggregate LOD = 0.9.

Now lets consider individual 4. The relevant genotypes are:



Note that we can't tell which SSR allele is inherited from which parent.

The two relevant conditional probabilities are:

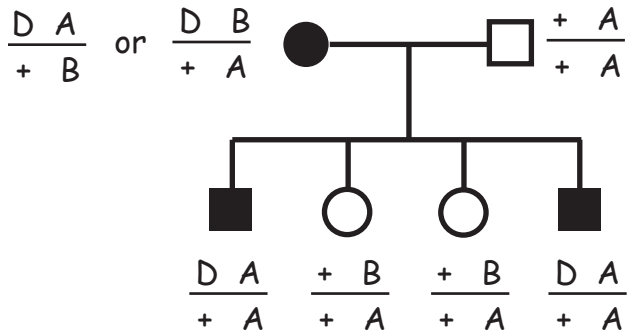
p(obtaining this genotype | complete linkage) = 1/2 • 1/2 = 1/4

p(obtaining his genotype | unlinked) = 1/2 • 1/4 = 1/4

The LOD score for individual 4 is : LOD = $\log_{10} 1 = 0$

Thus, individual 4 does not contribute to the total LOD score for this family. This type of situation is often called an uninformative meiosis. You can readily spot an uninformative individuals by asking the question: "Do I know exactly which alleles they inherited from each parent?". Thus the total LOD score for this family is 0.9. And at least four families with at least this much information would be needed to generate statistically significant evidence for linkage.

Now let's consider how to assess the information in a pedigree pedigree in which every meiosis is informative, but although we know that the mother is heterozygous for both markers (only one of her parents had the dominant trait and she has SSR alleles A and B) we don't have DNA from her parents so we don't know which of two possible phases for the relatinship between the dominant marker and the SSR alleles.



Since we don't know the phase of the mother we need to consider each phase with a probability of 1/2 in calculating the LOD score.

We will calculate the LOD score considering all four children in aggregate. Otherwise the the calculatio is similar as before.

P (data | completely linked) = 1/2(P if phase 1) + 1/2(P if phase 2)

$$= 1/2\ (1/2)^4 + 1/2\ (0)$$

P (data | loci unlinked) = $(1/4)^4$

$$\text{LOD} = \log_{10} \frac{(1/2)^5}{(1/4)^4} = \log_{10}(8) = 0.9$$

The net effect of the loss of phase data for this family is a decrease in the LOD score by a factor of $\log_{10} 2 = 0.3$. The total LOD score for a family can be quickly obtained by adding 0.3 for each informative meiosis, and subtracting 0.3 if the phase of the parent is unknown. Thus LOD = 1.2 – 0.3 = 0.9