HST.161 Molecular Biology and Genetics in Modern Medicine
Fall 2007

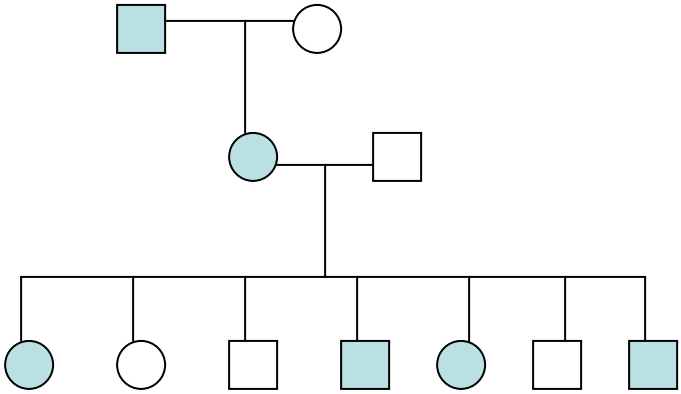
"LOD Demystified"

(v3 S Garg 2007)

What is LOD?

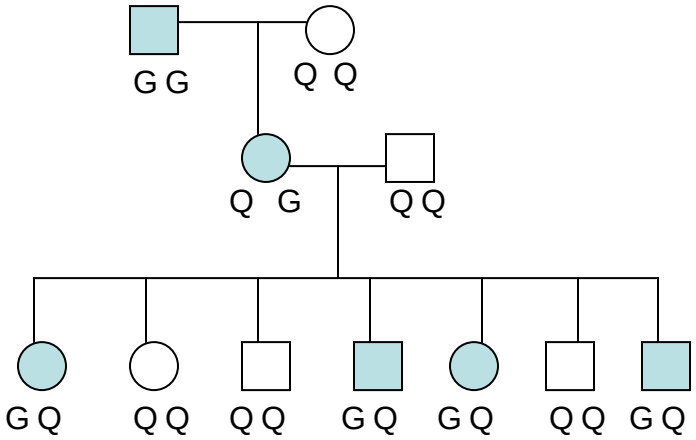
A LOD score is just a mathematical way to compare inheritance across families. It stands for "log odds ratio." Forget the word LOD for a minute, and let's see why we would need it and where it comes from. Definitely forget about theta until I bring it up, or I will smote thee down with great vengeance.

Let's say we observe the following family pedigree, where the shading just indicates a phenotype (Like excessive singing and nose-tweaking, if the individuals in question did not add a sprig of peppermint to their potion. Bonus points if you know where that reference is from).

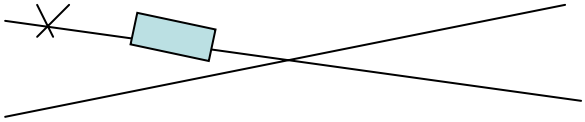


Now we can see that grandpa was the one who nose-tweaked a bunch. Now it is important to note that we can make a hypothesis that this is a genetically inherited disease because it is inherited in a Mendelian fashion. We aren't saying it definitely is or isn't, but that is precisely one idea which we want to test. We wish to find the gene responsible for nose-tweaking, as we have no idea why G-pa was such a weird old man.

How are we going to do this? Well we can do a couple of things. What we are going to do is find some DNA markers we can sequence on all of the folks in the family. These markers could be anything: # of repeats in an expansion, SNP's, a segment of sequence, etc., they are just markers. They may or may not have anything to do with the disease, that is precisely what we want to figure out. So let's take an example of such a marker (call it marker 1) and sequence it on each of these folks (marker 1 can have values G & Q):



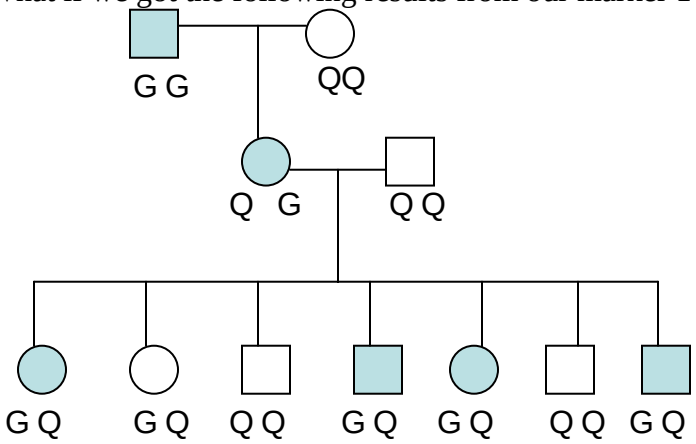
Now mom was GQ. That is, she had both of these markers present and she was hiccuppy.
Before we go any further, let's picture what the **actual** situation might look like on a chromosome:



Here we see a chromosome in mitosis (because it's duplicated). Let's say the rectangle is the giddiness gene, and the X is the location of marker 1. That means that usually a certain value for the marker X will associate with the diseased version of the gene since they are on the same chromosome. Remember, each offspring gets one chromosome, so if an offspring gets the marker it will get giddiness as well. When will this not be true? It won't be true when there is a recombination somewhere on the chromosome between X the marker and the rectangle box gene.

Back to our pedigree. I hope you can look at it and make the hypothesis that "G" is the value that most closely segregates with the giddiness phenotype. So how can we test this hypothesis? Well, what are the odds that it is true? The odds that this hypothesis is true is the **ratio** between how many of these meioses we would see if marker and gene were linked (linked=on the same chromosome, do not segregate independently) and how many we would see if they weren't.
Think back to probability. The **odds** that we would see this data if the marker and gene were perfectly linked is 1*1*1*1*1*1*1, since there are seven kids and the odds are 100% for each if there is perfect linkage. What are the **odds** we would see this data if the marker and gene were not linked? Well sir it would be ½*1/2*1/2*1/2*1/2*1/2*1/2, since for each child there is a 50% chance they would inherit a particular allele anyway by chance. This makes the ratio of the odds our hypothesis is true to the odds it isn't 1^7/.5^7, or 128. That's pretty high odds that our hypothesis is true. We take the logarithm of this number for statistical reasons that are beyond the scope of this course (though I am happy to discuss them with any of you). Log128 = 2.1, which is the value for the LOD. It just so happens, also through statistics that are beyond the scope of this course, that we count LOD's greater than 3 to be significant. (it's somewhat arbitrary, just like $p<.05$ is basically a randomly chosen cutoff).

What if we got the following results from our marker 1 study?



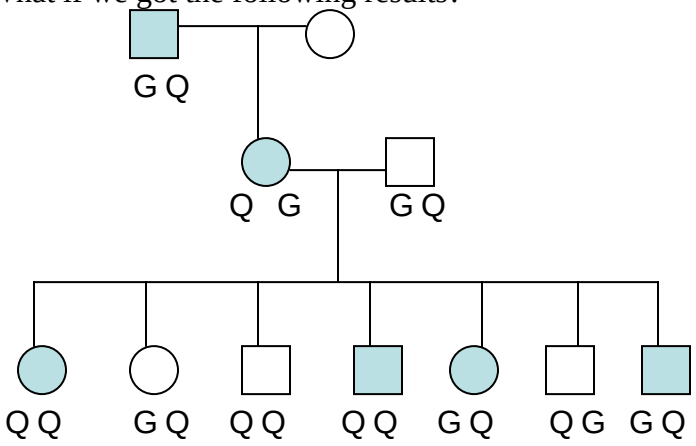
Here we can tell that every child in generation III who has a G got it from their mother in generation II, since dad was QQ (remember, they are just inheriting the marker, we don't know that this is associated with disease, that is exactly what we are trying to find out!!). We have a problem child. The second child has a G but no disease! We aren't dead in the water, because remember we could have had a recombination event. Again we must ask ourselves: What are the odds that marker 1 is near the gene or genetic element that causes giddiness and hiccupping?

Well what are the odds they are linked? 6 out of 7 children support this hypothesis. But in order for us to say there must have been a recombination, then we must hypothesize that marker X and the gene were not exactly the same. In the first case, marker X could very well have been inside the gene box. We couldn't tell. All we can tell is that they are linked.
So if they are not the same, we must make a hypothesis that there is some distance between them. We call this distance theta, ϑ. Theta is directly proportional to physical distance on the chromosome. ϑ of .01 is equivalent to a distance of 1 megabase (1 million bases, also the same thing as 1 centimorgan). A ϑ of .50 is 50 megabases. I will tell you if two things are 50 megabases apart on a chromosome they segregate as if they were on different chromosomes altogether, recombination is that likely between them.
What are the odds that our hypothesis is true now?
Well 6 out of 7 support our hypothesis. But now the odds that they would occur is no longer 1. Since we are hypothesizing some distance between marker and chromosome, there is some odds each child should NOT occur (as in, why aren't they a recombinant?). Let's hypothesize a distance of 5 centimorgans between marker and gene. Go through the exercise in your head: I claim that the 6 children who support the hypothesis each had a 95% chance of occurring now. The 1 oddball had a 5% chance of occurring. So the odds that the marker and disease are linked at a distance of 5 megabases is: (.95)^6*(.05)^1. The odds that they are not linked are still (.5)^7. So the ratio of these odds is (.037)/(.0078) = 4.74. The Log of this = .67, significantly lower.

Wasn't it arbitrary that we chose ϑ of .05? What if we chose ϑ of .10, hypothesizing that the marker and gene were further apart? Yes it is arbitrary, and we can do exactly that. I

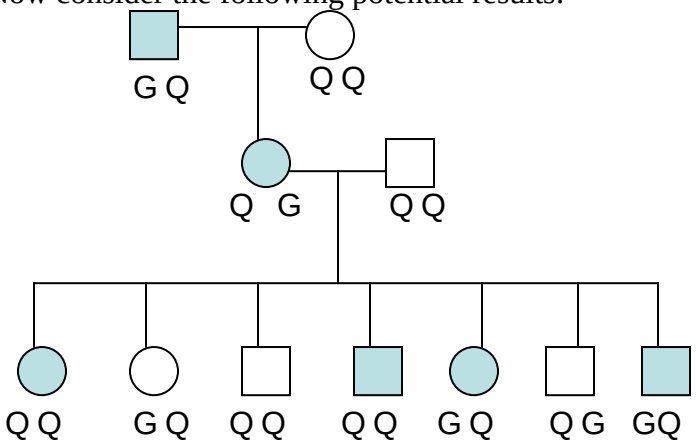
leave it as an exercise to you to show yourself that with ϑ of .10, LOD is .83. For ϑ of .40, LOD is .36. You can see that there is in fact an ideal ϑ at which the LOD will be the highest, and this is the case. This is how we can use this information to determine how far apart the marker and gene are actually, and this is how we can map two things to each other. Essentially, keep in mind, all we are doing is measuring recombination frequencies between two things are using that to determine the genetic distance between them.

What if we got the following results?



Now can you even tell if it is G or Q associating with disease? You can't at all. Which one is it? You don't know. You can't even make a hypothesis, because G or Q could be inherited from either the mother or father in generation II. We can't assign a specific allele to disease. We call these **uninformative meioses**, because we can't use them for anything. Anytime you can't reasonably assign an allele with disease and make a hypothesis, they are uninformative meioses.

Now consider the following potential results:



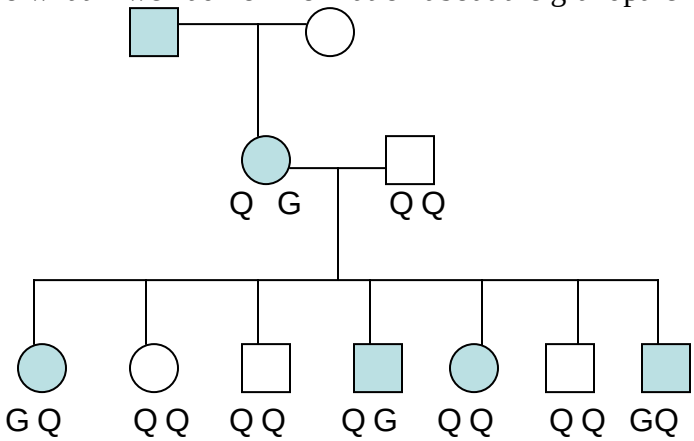
Now here we can definitely say that each child that got a G got it from his mother's side, which is the affected side. Thus we can hypothesize that G associates with disease. We can at least make the hypothesis.

But check it out: that hypothesis doesn't really hold water. If G causes disease, there are 4 recombinants! Yikes! No matter what you plug in for theta, this won't look very significant. This is just not a good marker. It's probably on a different chromosome, and these are expected results for that situation. In other words, any marker at random will likely give results like this, since most likely a given marker is not linked to the disease and is in fact on a different chromosome.

If you hypothesize that Q on the mother's side causes disease, then all of the offspring are informative. There are now 3 recombinants. Again, not a good marker! All the meiosis are informative. You can tell anyone with a G in generation III got the Q from their father. Likewise anyone with 2 Q's got one from each parent. This is not an unreasonable hypothesis. The value Q for the marker on the mother's side could be part of a totally different allele than the value Q on the father's side, so it's possible that the mother's side of the family and the father's side just have different versions of an allele. You would hypothesize in one case that Q is on an allele containing disease, and in the other allele the same Q is there but the disease is not.

What is this business about phase?

So what if we had no information about the grandparents?



Here it is equally likely that G or Q from the mother is the allele associated with disease. Recall that Q in the mother is not necessarily the same allele as Q in the father, so we can't just say that for sure it's G that is associated with disease. All we know is affected individuals. There is either one recombinant if G is associated with the disease causing allele and 6 non recombinants, or there are six recombinants and one non recombinant if Q from the mother is associating with the disease. We can't distinguish these, so we call this situation "**phase unknown**." You should realize that you expect a lower LOD because there is more uncertainty in the problem. The LOD is calculated by taking the average of both possible arrangements of recombinants and non-recombinants, though for practical purposes a good estimate is to just subtract .3 for a phase unknown problem. If you are astute, you just realized the example we did above was actually both a bad marker and phase unknown.
This covers the basics of LOD.