

MAPPING QUANTITATIVE TRAIT LOCI*

Introduction

So far in our examination of the inheritance and evolution of quantitative genetics, we've been satisfied with a purely statistical description of how the phenotypes of parents are related to the phenotypes of their offspring. We've made pretty good progress with that. We know how to partition the phenotypic variance into genetic and phenotypic components and how to partition the genetic variance into additive and dominance components. We know how to predict the degree of resemblance among relatives for any particular trait in terms of the genetic components of variance. We know how to predict how a trait will respond to natural selection.

That's not bad, but in the last fifteen or twenty years the emergence of molecular technologies that identify large numbers of Mendelian markers has led to a new possibility. It is sometimes possible to identify the chromosomal location, at least roughly, of a few genes that have a large effect on the expression of a trait by associating variation in the trait with genotypic differences at loci that happen to be closely linked to those genes. A locus identified in this way is referred to as a *quantitative trait locus*, and the name given to the approach is QTL mapping.

The basic ideas behind QTL mapping are actually very simple, although the implementation of those ideas can be quite complex. In broad outline, this is the approach:

- Produce a set of progeny of known parentage. One common design involves first crossing a single pair of "inbred" parents that differ in expression of the quantitative trait of interest and then crossing the F_1 s, either among themselves to produce F_2 s (or recombinant inbred lines) or backcrossing them to one or both parents.
- Construct a linkage map for the molecular markers you're using. Ideally, you'll have a large enough number of markers to cover virtually every part of the genome.¹

*Based on the discussion in Lynch, M., and B. Walsh, 1998. *Genetics and Analysis of Quantitative Traits*, Sinauer Associates, Sunderland, MA

¹We'll talk a little later about how many markers are required.

- Measure the phenotype and score the genotype at every marker locus of every individual in your progeny sample.
- Collate the data and analyze it in a computer package like **QTL Cartographer** to identify the position and effects of QTL associated with variation in the phenotypic trait you're interested in.

If that sounds like a lot of work, you're right. It is. But the results can be quite informative, because they allow you to say more about the genetic influences on the expression of the trait you're studying than a simple parent-offspring regression.

Thoday's Method²

Suppose there is a locus, Q , influencing the expression of a quantitative trait situated between two known marker loci, A and B .³ If we have inbred lines with different phenotypes, we can assume that one line has the genotype AQB/AQB and the other has the genotype aqb/aqb . The procedure for detecting the presence of Q is as follows:

1. Cross the inbred lines to form an F_1 . The genotype of all F_1 progeny will be AQB/aqb .
2. Intercross the F_1 's to form an F_2 and look at the progeny with recombinant genotypes, e.g., aB/ab .
3. If Q lies between A and B
 - (a) The phenotypes of progeny will fall into two distinct classes corresponding with the genotypes: aqB/aqb and aQB/aqb .⁴
 - (b) The recombination fraction between A and Q is related to the proportion of qq and Qq genotypes among the progeny.

²Primarily of historical interest, but it sets the stage for what is to follow.

³Of course, we don't know it's there when we start, but as we've done so many other times in this course, we'll assume that we know it's there and come back to how we find out where "there" is later.

⁴Actually there could be a third phenotypic class if there are two recombination events between a and b , i.e., aQB/aQb . Thoday's method assumes that the recombination fraction between A and B is small enough that double recombination events can be ignored, because if we don't ignore that possibility we must also admit that there will be some aqB/aQb genotypes that we can't distinguish from aQB/aqb genotypes.

Notice that in this last step we actually have a criterion for determining whether Q lies between A and B . Namely, if A and B are close enough in the linkage map that there is essentially no chance of double recombination between them, then we'll get the two phenotype classes referred to in recombinants between A and B . If Q lies outside this region,⁵ we'll get only one phenotypic class among progeny with recombinant genotypes.

Genetic recombination and mapping functions

Genetic mapping is based on the idea that recombination is more likely between genes that are far apart on chromosomes than between genes that are close. If we have three genes A , B , and C arranged in that order on a chromosome, then

$$r_{AC} = r_{AB}(1 - r_{BC}) + (1 - r_{AB})r_{BC} \quad ,$$

where r_{AB} , r_{AC} , and r_{BC} are the recombination rates between A and B , A and C , and B and C , respectively.⁶

Haldane pointed out that this relationship implies another, namely that the probability that there are k recombination events between two loci m map units apart is given by the Poisson distribution:

$$p(m, k) = \frac{e^{-m}m^k}{k!} \quad .$$

Now to observe a recombination event between A and C requires that there be an odd number of recombination events between them (1, 3, 5, ...), i.e.,

$$\begin{aligned} r_{AC} &= \sum_{k=0}^{\infty} \frac{e^{-m}m^{(2k+1)}}{(2k+1)!} \\ &= \frac{1 - e^{-2m}}{2} \quad . \end{aligned}$$

This leads to a natural definition of map units as

$$m = -\ln(1 - 2r)/2 \quad .$$

m calculated in this way gives the map distance in Morgans ($1M$). Map distances are more commonly expressed as centiMorgans, where $100cM = 1M$. Notice that when r is small,

⁵Or if Q has only a small effect on expression of the trait we're studying.

⁶In practice this isn't quite true. Interference may cause the recombination rate between A and C to differ from that predicted. That's not much of a problem since we can just add a little correction factor, but we'll ignore interference to keep things simple.

$r \approx m$, so the map distance in centiMorgans is approximately equal to the recombination frequency expressed as a percent. There are several other mapping functions that can be chosen for an analysis. In particular, for analysis of real data investigators typically choose a mapping function that allows for interference in recombination. We don't have time to worry about those complications, so we'll use only the Haldane mapping function in our further discussions.

How many markers will you need?

If markers are randomly placed through the genome, then the average distance between a QTL and the closest marker is

$$E(m) = \frac{L}{2(n+1)} \quad ,$$

where L is the total map length and n is the number of markers employed. The upper 95% confidence limit for the distance is

$$\frac{L}{2} \left(1 - 0.05^{(1/n)}\right) \quad .$$

Since the human genome is 33M (3300cM), 110 random markers give an average distance of 14.9cM and an upper 95% confidence limit of 44.3cM, corresponding to recombination frequencies of 0.13 and 0.29, respectively. Since there are about 30,000 genes in the human genome, there are roughly 10 genes per centimorgan. So if your QTL is 44cm from the nearest marker, there are probably over 400 genes in the chromosomal segment you've identified.

If r_{MQ} is the recombination fraction between the nearest marker locus and the QTL of interest, the frequency of recombinant genotypes among F_2 progeny is $2r_{MQ}(1-r_{MQ})+r_{MQ}^2$. As you can see from the graph in Figure 1, there's a nearly linear relationship between recombination frequency and the frequency of recombinant phenotypes (p in the graph).

Analysis of an F_2 derived from inbred lines

An analysis of inbred lines uses the same basic design as Thoday, but takes advantage of more information.⁷ We start with two inbred lines M_1QM_2/M_1QM_2 and m_1qm_2/m_1qm_2 , make

⁷As I alluded to earlier, other breeding designs are possible, including backcrosses and recombinant inbred lines and analyses involving outbred parents. The principles are the same in every case, but the implementation is different.

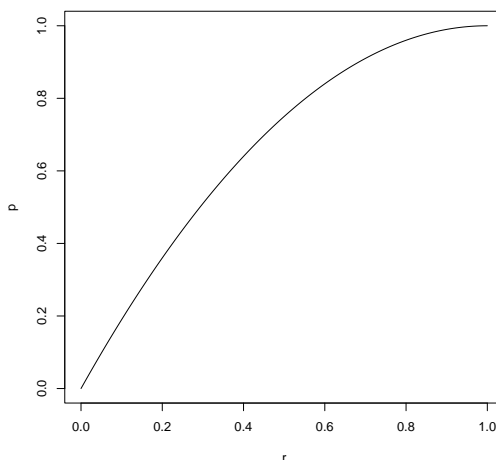


Figure 1: The relationship between recombination frequency, r , and the frequency of recombinant phenotypes, p , assuming a Haldane mapping function.

an F_1 , intercross them, and score the phenotype and marker genotype of each individual. Analysis of the data is based on calculating the frequency of each genotype at the Q locus as a function of the genotype at the marker loci and the recombination fractions between the marker loci and Q .⁸ For example,

$$\begin{aligned} P(M_1QM_2/M_1QM_2) &= ((1 - r_{1Q})(1 - r_{Q2})/2)^2 \\ P(M_1QM_2/M_1qM_2) &= 2((1 - r_{1Q})(1 - r_{Q2})/2)(r_{1Q}r_{Q2}/2) \\ P(M_1qM_2/M_1qM_2) &= (r_{1Q}r_{Q2}/2)^2 \quad . \end{aligned}$$

Because the frequency of $M_1M_2/M_1M_2 = ((1 - r_{12})/2)^2$, we can use Bayes' Theorem to write the conditional probabilities of getting each genotype as

$$\begin{aligned} P(QQ|M_1M_2/M_1M_2) &= \frac{(1-r_{1Q})^2(1-r_{Q2})^2}{(1-r_{12})^2} \\ P(Qq|M_1M_2/M_1M_2) &= \frac{2r_{1Q}r_{Q2}(1-r_{1Q})(1-r_{Q2})}{(1-r_{12})^2} \\ P(qq|M_1M_2/M_1M_2) &= \frac{r_{1Q}^2r_{Q2}^2}{(1-r_{12})^2} \quad . \end{aligned}$$

Clearly, if we wanted to we could right down similar expressions for the nine remaining marker genotype classes, but we'll stop here. You get the point.⁹

⁸You should be getting used to the idea now that we always assume we know something we don't and then backcalculate from what we do know to what we'd like to know.

⁹I should say, I *hope* you get the point.

Now that we've got this we can write down the likelihood of getting our data, namely

$$L(x|M_j) = \sum_{k=1}^N \phi(x|\mu_{Q_k}, \sigma^2)P(Q_k|M_k) \quad ,$$

where N is the number of QTL genotypes considered, $\phi(x|\mu_{Q_k}, \sigma^2)$ is the probability of getting phenotype x given the mean phenotype, μ_{Q_k} , and variance, σ^2 , associated with Q_k , and $P(Q_k|M_k)$ is the probability of getting Q_k given the observed marker genotype. Fortunately, we don't have to do any of these calculations, all we do is to ask our good friend (QTL Cartographer) to do the calculations for us. It will scan the genome, and tell us how many QTL loci we are likely to have, where they are located relative to our known markers, and what the additive and dominance effects of the alleles are.

The Caveats

That's wonderful, isn't it? We have to do a little more work than for a traditional quantitative genetic analysis, i.e., we have to do a bunch of molecular genotyping in addition to all of the measurements we'd do for a quantitative genetic experiment anyway, but we now know how many genes are involved in the expression of our trait, where they are in the genetic map, and what their additive and dominance effects are. We can even tell something about how alleles at the different loci interact with one another. What more could you ask for? Well, there are a few things about QTL analyses to keep in mind.

- As currently implemented, QTL mapping procedures assume that the distribution of trait values around the genotype mean is normal, *with the same variance for all QTL genotypes*.¹⁰
- QTL mapping programs often estimate the effects of each locus individually. It's not at all easy to search simultaneously for the joint effects of two QTL loci, although it's not too hard to look at the combined effects of QTL loci first identified individually. Composite interval mapping, in which additional markers are included as cofactors in the analysis, partially addresses this limitation. Multiple interval mapping looks at several QTLs simultaneously and shows some promise, but as you may be able to imagine it's pretty hard to search for more than a few QTLs simultaneously.
- If some loci in the "high" line have "low" effects and vice versa, the effects of both loci (and possibly other loci) may be masked.

¹⁰I *know* you picked up on that when I said that the phenotypic variance associated with each QTL genotype was σ^2 . You were just too polite to point it out and interrupt me.

- Most existing techniques identify QTL's important *in a particular cross*, but different crosses can identify different QTL's. Even the same cross may reveal different QTL's if the measurements are done in different environments. Methods to analyze several progeny sets simultaneously are only now being developed.

Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.