

# MAPPING QUANTITATIVE TRAIT LOCI\*

## Thoday's Method<sup>1</sup>

Suppose there is a locus,  $Q$ , influencing the expression of a quantitative trait situated between two known marker loci,  $A$  and  $B$ . 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 phenotypes, 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  $aQB/aqb$ .<sup>2</sup>
  - (b) The recombination fraction between  $A$  and  $Q$  is related to the proportion of  $qq$  and  $Qq$  genotypes among the progeny.

---

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

<sup>1</sup>Primarily of historical interest, but it sets the stage for what is to follow.

<sup>2</sup>Actually 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.

## 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.<sup>3</sup>

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,  $r \approx m$ , so the map distance in centiMorgans is approximately equal to the recombination frequency expressed as a percent.

## 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

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

---

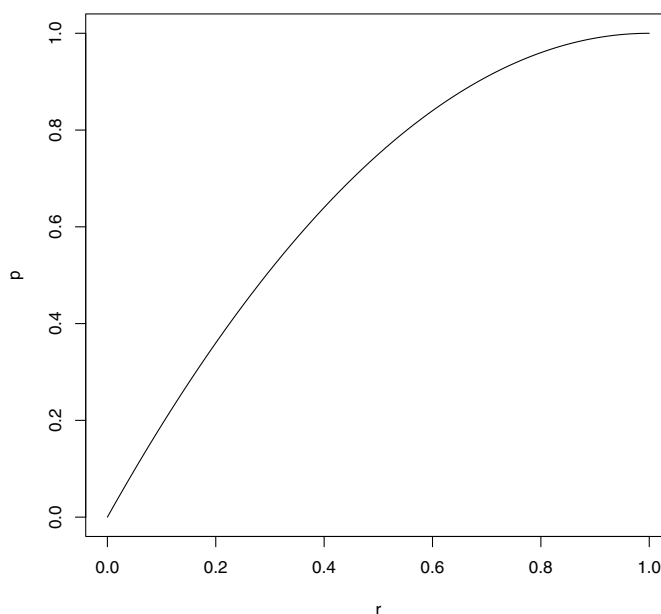
<sup>3</sup>In practice this isn't quite true. Interference may cause the recombination rate between  $A$  and  $C$  to differ from that predicted. That's not much of a problem since we can just add a little correction factor, but we'll ignore interference to keep things simple.

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

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

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 following graph, 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.<sup>4</sup> We start with two inbred lines  $M_1QM_2/M_1QM_2$  and  $m_1qm_2/m_1qm_2$ , make 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  $Q$ .<sup>5</sup> For example,

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

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

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

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.

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 and variance 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. What more could you ask for?

---

<sup>4</sup>Other breeding designs are possible, including backcrosses and recombinant inbred lines.

<sup>5</sup>You 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.

## The Caveats

Well,

1. 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*.
2. QTL mapping programs 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.
3. 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.
4. Most existing techniques identify QTL's important *in a particular cross*, but different crosses can identify different QTL's. Methods to analyze several progeny sets simultaneously are only now being developed.