DEVELOPING GENETIC MAPS

We can determine θ*s* for pairs of genes or loci. Next question is how to place them in a genetic map; i.e., grouping loci into linkage groups and ordering sets of 3 or more loci.

1. Grouping loci into linkage groups

The question beyond determining linkage of any two loci is the ordering of the loci into a reasonable genetic map. Ordering can be done using the MLE of θ, determined via a two-point evaluation described before, in conjunction with the lod score (Z) or the probability value of linkage. Commonly used criteria, e.g., in the computer program MAPMAKER (Lander et al., 1987), are as follows:

if $\{\theta \leq 0.40$ and $Z \geq 3\}$ then loci *i* and *j* are grouped into the same linkage group.

Often, linkage groups are determined more iteratively; that is, after assembling groups based on the above criteria, the data can be reanalyzed with more stringent parameters, e.g. $\{\theta \leq 0.30$ and $Z \geq 5\}$. Groups staying together over a range of parameters are more likely to be truly linked.

2. Multiple locus ordering

Three loci can be ordered rather simplistically by looking at the two-locus recombination frequencies. Loci further apart will have larger recombination values.

e.g., $\hat{\theta}_{ab} = 0.10 \qquad \hat{\theta}_{ac} = 0.22 \qquad \hat{\theta}_{bc} = 0.30$

We might expect that the order is B-A-C, because B and C have the largest recombination fraction and define the ends; A fits in between.

The maximum likelihood approach, which is generally applicable to more than two loci, uses two-point recombination data and is calculated as follows (ignoring interference):

$$\ln L(x) = \sum_{i=1}^{l-1} n_{a_i a_j} \left[\theta_{a_i a_j} \ln \theta_{a_i a_j} + (1 - \theta_{a_i a_j}) \ln(1 - \theta_{a_i a_j}) \right]$$

where x = a linear order of l loci, e.g.1,2,...,l-1, l

$a_i a_j$ = pairs of loci, e.g. $a_1 a_2$ represents the first pair in the order (x) being evaluated; in this case, $a_1 a_2$ would represent loci 1 and 2. If the order we were evaluating was 3,4, ..., l, then $a_1 a_2$ would represent loci 3 and 4.

$j = i + 1$

n_{ij} = number of informative sample sizes for locus combination *i* and *j*

A likelihood is computed for each possible order; the one with the highest likelihood (i.e., closest to zero) is selected as the most likely order.

e.g. for our example above:

lnL (a-b-c) = [100(0.1 ln 0.1 + 0.9 ln 0.9)] + [100(0.3 ln 0.3 + 0.7 ln 0.7)] = -93.60

lnL (b-a-c) = - 87.20

lnL (a-c-b) = -113.78

Thus, order b-a-c is more likely than either of the other two orders.

MAPMAKER (Lander et al., 1987) is a computer program for multipoint analysis of any number of loci necessary for the development of an accurate genetic map. It computes the maximum likelihood map order for any number of loci. The correct map-order is the one that gives the lowest log-likelihood value.

3. Multilocus genetic maps (Lander and Green, 1987)(Liu, 1998)

Once θ is found, a new maximum likelihood estimation needs to be conducted to find the multilocus recombination values.

The basic problem, and the reason that two-locus recombinations are not the best estimate of a *multilocus* map, is that we cannot (in most cases) simply count recombinants between any two loci.

> For example, take the case of two codominant RFLP markers. Assume the parents are homozygous AB and ab. The two homologues in the F_1 will be AB/ab. We can then classify nine classes in the F_2. However, we won't be able to distinguish between the two types of double heterozygotes: double recombinant Ab/aB and the non-recombinant AB/ab.

If this is the case, then we can't tell the whether a recombination between A and C occurred in AB or in BC, but we can calculate expectations based on the 2-locus recombination ratios.

Thus, we need to do a genetic reconstruction to estimate the number of recombinants within the doubly heterozygous class, which can be done more effectively by looking at more than two loci. The best method to arrive at the MLE for each interval's recombination value is the EM algorithm. Newton-Raphson iterations can also be used, but for large numbers of loci, they are prohibitive in computation time.

Take three loci, A–B–C, which have been found to be in this order.

If the true recombination value of AB is θ and the true value of BC is θ_2, then we can use maximum likelihood to estimate the true parameters from the observed two-locus recombinations.
The general idea is this:

1. Take an initial estimate, $\theta^{old} = (\theta_1, \theta_2, \ldots, \theta_{l-1})$, where l is the number of loci.
2. (E step) Using θ^{old} as if it were the true recombination fraction (i.e. each interval's recombination fraction is correct), compute the expected number of recombinant meioses for each interval.
3. (M step) Using this expected value as if it were true, compute the MLE θ^{new} for the recombination fraction.
4. Iterate E and M until $\theta^{new} \cong \theta^{old}$ (i.e. the likelihood converges to a maximum).

In our example:

1. The initial ("old") estimates are θ_{AB} and θ_{BC}, for recombination between AB and BC respectively.
2. The expectation step:
   Our first estimate of the true number of recombinations (t) in the intervals AB and BC. For interval AB:

   $t_{AB} = R_{AB} + a_1 R_{AC} + a_2(N_{AC}-R_{AC})$

   where R_{AB} is the number of recombinants observed
   N_{AC} is the number of meioses observed
   a_1 is the probability that a recombination between A and C occurred in AB:

$$a_1 = \frac{\theta_{AB}(1-\theta_{BC})}{\theta_{AB}(1-\theta_{BC}) + (1-\theta_{AB})\theta_{BC}}$$

This is the probability of recombination in AB times the lack of recombination in BC (that is the only way to have a recombination in AB and also in AC), divided by all the recombinations in AC, namely the probability of recombination in AB times the lack of recombination in BC plus the probability of a lack of recombination in AB times the probability of a recombination in BC.

a_2 is the probability that a recombination occurred in AB if AC is non-recombinant:

$$a_2 = \frac{\theta_{AB}\theta_{BC}}{\theta_{AB}\theta_{BC} + (1-\theta_{AB})(1-\theta_{BC})}$$

This is the probability of recombination in AB multiplied by the probability of recombination in BC (the only way to have a recombination in AB and have no apparent recombination in AC), divided by the probability of recombination in AB times the probability of recombination in BC plus the probability of no recombination in AB times the probability of no recombination in BC.

Thus, "t_{AB}" includes two components:

1. The number of recombinants that we observed between A and B = R_{AB}
2. The number of recombinants that we *expect* to observe between A and B, based on an estimate derived from the number of recombinants seen in AC and the computed two point recombination fractions θ_{AB} and θ_{BC}.

If the observed number equals the expected number, then our two-point estimates are the true values. However, if some recombinants are "hiding" in the double heterozygous class, as they almost certainly are, then we will observe a different number than the expectations based on the AC data.

and $t_{BC} = R_{BC} + (1-a_1)R_{AC} + a_2(N_{AC}-R_{AC})$

3. Maximization step:
with our new expectations for recombinations (t_1 and t_2), we can develop new recombination fractions:

$$\theta_{AB}^{new} = \frac{t_{AB}}{N_{AB} + N_{AC}} \quad \text{and} \quad \theta_{BC}^{new} = \frac{t_{BC}}{N_{BC} + N_{AC}}$$

We divide by the sum of AB observations and AC observations (which are equal if we have no missing data), since we included both our observed AB and our expected AB recombinants in "t_{AB}."

4. Iterate steps 2 and 3 until the new value of θ is approximately equal to the value from the previous step.

Example:
Say we have three loci, A-B-C, in that order. We want to construct the multilocus map for these markers.

Say we observe 10, 10, and 18 recombinants between AB, BC, and AC, respectively, in a population of 100 individuals.

Now, pretend our estimates of recombination between AB is 20% and between BC is 10%.

Then, we need to find the true value of recombination between AB.

t(AB) = 10 + (0.69)(18) + (0.027)(82) = 24.64

Therefore, the new estimate of θ for the interval AB = (24.64)/(100+100) = 0.123

Iterating gives new estimates of recombination for AB of ~11% and BC of ~9%.

This can be extended to more loci in a similar manner, by working down the linkage group, and including all data in the expression.

MAP DISTANCE

Assuming no interference for the moment:

Call the recombination fraction between AB = θ_{12} BC = θ_{23} AC = θ_{13}

$\theta_{13} = [(\theta_{12} - \theta_{12}\theta_{23}) + (\theta_{23} - \theta_{23}\theta_{12})] = \theta_{12} + \theta_{23} - 2\theta_{12}\theta_{23}$ [Trow's formula, 1913]

The recombination from AC includes all the recombination events in AB and BC *except* for those events that occurred in both segments, which then appear to be a non-recombination as far as AC is concerned. Thus, recombination between two loci is only recognized if an odd number of recombinations occurs! If we have 2 or 4 recombinations between A and C, we do not recognize their existence, even though they occurred. We are referring to the number of cross-overs that a particular chromatid experiences, so that chromatids that have 0, 2, 4 recombinations between two loci appear non-recombinant while those with 1, 3, 5 are recombinant.

The result of this is that the *recombination fractions are not additive*. (Consider the case where $\theta_{12} = 0.30$ and $\theta_{23} = 0.30$. θ_{13} cannot be 0.60–as we have previously shown that the recombination fraction cannot be more than 0.50.) We need to develop a function such that AB + BC = AC.

Mapping functions convert recombination fractions (θ) to map distances (x) that are additive.

Map distance is defined as the *expected* number of crossovers between two loci or one-half the expected number of chiasma (recombination nodules).

Map distance is measured in Morgans: 1 Morgan = 1 crossover per chomatid. One crossover between two loci is 1/2 crossover per chromatid or 50% recombination or 0.5 Morgans = 50 cM.

(Note that in the case of Sherman and Stack, they were able to directly estimate map distance because they could identify every crossover as a RN.)

When we observe segregating progeny, we cannot observe *cross-overs*, and double recombinations will only appear if we have many markers very close together so that double recombinations between any two contiguous markers are not possible.

As the distance between markers becomes greater, the probability of an even number of recombinations becomes greater as well, and thus, recombination fractions are not additive.

Map functions are necessary to convert from recombination fraction to additive map distances (Liu, 1998, p. 318-.)

1) Morgan's function: in the absence of multiple crossovers (i.e., complete interference):

$x = \theta$, where *x* is the map distance and θ is the recombination fraction.

This function also is a close approximation of map distances from recombination frequencies of loci that are closely linked.

2) Haldane's function: assumes no interference and that crossovers occur randomly (equally likely) at all points along the chromosome. This expression derives from Trow's formula (above):

$x = -\frac{1}{2}\ln(1-2\theta)$ which has an inverse of $\theta = \frac{1}{2}(1-e^{-2x})$; if θ = 0.22 (22%), then *x* = 29 cM

3) Kosambi's function: (Kosambi, 1944) a commonly used mapping function that includes interference considerations, such that interference is strongest closest to a recombination but diminishes as the distance increases:

$$x = \frac{1}{2}\tanh^{-1}(2\theta) = \frac{1}{4}\ln\frac{1+2\theta}{1-2\theta}$$ if θ = 0.22, *x* = 23.6 cM

Note that the problem is that interference is not even throughout the genome and that both positive and negative interference is possible, as shown in Sherman and Stack (1995).

4) Binomial function (Karlin): this function allows a specified value for the maximum number of recombinations, independently distributed, that can occur in a given interval, *N*.

$$x = \frac{1}{2}N\left[1-(1-2\theta)^{\frac{1}{N}}\right]$$

REVIEW: STEPS IN LINKAGE ANALYSIS AND MAP CONSTRUCTION

1. Test each marker (gene) for segregation distortion–i.e. deviation from expectations.

2. Test marker pairs for evidence of linkage (e.g. with a χ^2 test for independence).

3. Calculate the two-point recombination values between markers that show evidence of linkage.

    a. Derive expected gamete proportions in terms of θ.

    b. Derive expected proportions among the progeny in terms of θ based on the gametic array

    c. Develop a likelihood expression that includes both the observed numbers of individuals in each phenotype class and their expectations in terms of θ.

    d. Take the natural logarithm of the likelihood equation to make solving for θ easier–this is called the *support*.

    e. Solve for the maximum likelihood estimator (MLE) of θ by differentiating the log likelihood equation (which give the *scores*) and equating to 0.

    f. Calculate a standard error of the MLE. First calculate the *information* by taking the second derivative of the lnL equation. The square root of the inverse of the information is the SE.

4. Calculate a LOD score for each pair of markers for which you estimated θ.

5. Group markers into linkage groups based on θ and LOD–starting with $\theta<0.4$ and LOD>3.0 and iterating until clear groups have been developed.

6. Determine the best order of markers in the group using maximum likelihood.

7. Develop a multipoint genetic map by genetic reconstruction–that is, use data from adjacent markers (via an EM algorithm, for example) to calculate a new, better value for recombination between each locus.

8. Convert recombination fractions into map distances using some type of map function (e.g. Kosambi or Haldane)

REVIEW: EXPECTATIONS OF GAMETES AND PROGENY PHENOTYPES

Gametes from AaBb heterozygote (= backcross progeny phenotype classes).

	Gamete Frequencies				
	AB	Ab	aB	ab	Total
No linkage	¼	¼	¼	¼	1
Complete linkage, coupling	½	0	0	½	1
Complete linkage, repulsion	0	½	½	0	1
Recombination θ, coupling	½ (1-θ)	½ θ	½θ	½ (1-θ)	1
Recombination θ, repulsion	½ θ	½ (1-θ)	½ (1-θ)	½ θ	1

Progeny classes in an F_2 (AaBb x AaBb) can be determined by multiplying out from above (see next table):

Coupling		Male Gametes			
		AB	Ab	aB	ab
	AB	AB*AB	AB*Ab	AB*aB	AB*ab
	Ab	etc.			
	aB			½ θ * ½θ = ¼ θ^2	½ θ * ½ (1-θ) = ¼ θ (1-θ)
Female Gametes	ab			½ θ * ½ (1-θ) = ¼ θ (1-θ)	½ (1-θ) * ½ (1-θ) = ¼ $(1-\theta)^2$

From the above Punnett square, we can derive progeny frequencies:

a. For example, aabb can only arise through joining male 'ab' and female 'ab' : ½ (1-θ) * ½ (1-θ)= ¼ $(1-\theta)^2$

b. but aaB- can arise three ways: male 'aB' and female 'aB'; male 'aB' and female 'ab'; or male 'ab' and female 'aB'. Thus, the expectations of aaB- progeny is the sum of these: ¼ θ^2 + 2[¼θ(1-θ)] = ¼ θ(2-θ)

These are included in the table below:

	Progeny Frequencies in F_2				
	A-B-	A-bb	aaB-	aabb	Total
No linkage	9/16	3/16	3/16	1/16	1
Complete linkage, coupling	¾	0	0	¼	1
Complete linkage, repulsion	½	¼	¼	0	1
Recombination θ, coupling	¼ $(3-2\theta+\theta^2)$	¼ (θ2-)θ	¼ (θ2-θ)	¼ $(1-\theta)^2$	1
Recombination θ, repulsion	¼ $(2+\theta)^2$	¼ $(1-\theta)^2$	¼ $(1-\theta)^2$	¼ θ^2	1

For the coupling example:

	Progeny Frequencies in F_2				
	A-B-	A-bb	aaB-	aabb	Total
Observed numbers	187	35	37	31	290
Expected frequencies	¼ $(3-2\theta+\theta^2)$	¼ θ (2-θ)	¼ θ (2-θ)	¼ $(1-\theta)^2$	1

The ln-likelihood expression:

$$\ln L = \text{Constant} + 187 \ln (¼ (3-2\theta+\theta^2) + (35+37) \ln (¼\theta (2-\theta)) + 31 \ln (¼ (1-)\theta^2)$$

$$\frac{d \ln L}{d\theta} = 187\frac{2(\theta - 1)}{3 - 2\theta + \theta^2} + 72\frac{2(1-\theta)}{\theta(2-\theta)} + 31\frac{-2}{(1-\theta)} = 0$$

Solve for θ (lots of math)

$\theta = 0.3047$