Chapter 5

Basics of Linkage and Gene Mapping

Julius van der Werf

Basics of Linkage and Gene Mapping..45
Linkage 45
Linkage disequilibrium...47
Mapping functions 48
Mapping of genetic markers 50
Estimation of the recombination fraction..50
Maximum likelihood estimation of linkage (recombination fraction).....................51
Testing for linkage: LOD scores..52
Design of mapping..54
References...54

Linkage

Two genes are said to be linked if they are located on the same chromosome.
We assume that different chromosomes segregate independently during meiosis. Therefore, for two genes located at different chromosomes, we may assume that their alleles also segregate independently. The chance that an allele at one locus co-inherits with an allele at another locus of the same parental origin is then 0.5 and such genes are unlinked.

	parent 1	AABB	x	aabb		parent 2
	F1	AaBb (100%)				
	F1-gametes	AB	Ab	aB	ab	
A and B are unlinked:	frequency (%)	25	25	25	25	
A and B linked:	e.g. frequency (%)	35	15	15	35	
A and B tightly linked	e.g. frequency (%)	48	2	2	48	

The chance that A/B or a/b co-inherit to the offspring is 0.5 in case the genes are unlinked. This chance increases if the genes are linked. We can observe a degree of linkage. The reason is that even if genes are located on the same chromosome, they have a chance of not inheriting as in the parental state. This is due to *recombination*. During meiosis, the chromosome often breaks and the rejoins with the homologue chromosome, such that new chromosomal combinations appear (indicated as *crossover*). In the example, the combination aB and Ab did not appear in the parental cells. These new combinations are the result of recombination, therefore indicated as *recombinants*.

In real life we can not observe gametes (at least, not the haplotypes), but the result from meiosis in an F1 can be checked in a *testcross*, which is a classical genetic test of linkage.
This is achieved by crossing an F1 back to the homozygote recessive parent. The recombinants can easily be identified among the phenotypes in the offspring of a *testcross*.

A testcross is

F1	AaBb		x	aabb	parent 2
Offspring	AaBb	Aabb	aaBb	aabb	

If the A and B alleles are dominant, the composition of the gamete produced by the F1 sire can be determined from the offspring's phenotype. In Drosophila, such linkage studies have been carried out during most of the 20th century. The further the distance between two genes, the more frequently there will be crossover, the higher the number of recombinations. Therefore, the recombination fraction is calculated from the proportion of recombinants in the gametes produced.

Recombination fraction = number of recombinants / total

Note that the combinations aB and Ab are not always the recombinants. If the F1 was made from a parental cross AAbb x aaBB, than the recombinant gametes would be AB and ab. Therefore, for each testcross, we have to determine how the alleles were joined in the parental generation. This is known as the *phase*. If AB and ab were joined in the parental gametes, the gene pairs are said to be in *coupling phase* (as in first example). Otherwise, as in the cross AAbb x aaBB, the gene pairs are in *repulsion phase*. (These terms can be somewhat arbitrary if there are no dominant or mutant alleles).

Example / exercise

In corn, the allele for coloured kernels (R) is dominant to the allele for colourless kernels (r) and the allele for green plant colour (Y) is dominant for the yellow plant colour (y). The R and Y genes are linked. Two different plants (plant 1 and plant 2) that were heterozygous for each trait were test crossed to plants that were homozygous for the recessive alleles. The phenotypes and the frequencies of the progeny from the test crosses are:

	Progeny of plant 1	Progeny of plant 2
Coloured kernels, green plants	12	45
Coloured kernels, yellow plant	155	5
Colourless kernels, green plants	115	3
Colourless kernels, yellow plant	18	27

- We can see that the frequency of offspring deviates from frequencies that would be expected if the genes were unlinked

- We can determine recombinant and non-recombinant progeny for each plant
- We can determine recombination frequencies for each plant
- If plant 1 and plant 2 were generated by crossing true-breeding plants (homozygous), we can write down the genotype of the parents of plant 1 and plant 2

Linkage disequilibrium

Linkage equilibrium and its opposite: linkage disequilibrium, are terms used for the chance of co-inheritance of alleles at different loci. Alleles that are in random association are said to be in linkage equilibrium. The chance of finding one allele at one locus is independent of finding another allele at another locus. In the previous example, suppose in the testcross progeny we observe the A allele. If the chance of finding either the B-allele or the b-allele is 50%, the genes are in linkage equilibrium. Hence, if we look at the gamete-frequencies, then we speak of linkage equilibrium if the

freq(AB) = freq (Ab) = freq (aB) = freq (ab).

And the amount of disequilibrium is measured as

D = freq(AB).freq(ab) – freq(Ab).freq(aB).

Linkage disequilibrium is somewhat a confusing term. It can be the result of physical linkage of genes. However, even if the genes are on different chromosomes, there can be linkage disequilibrium. This can be due to selection. If A and B both affect a characteristic positively, and the characteristic is selected for, than in the selected offspring there will be a negative association between A and B. This is also known as Bulmer effect, as Bulmer (1971) described it to (partly) explain loss of variation due to selection.
Linkage disequilibrium can also be the result of crossing or migration. If a new individual with AB gametes come into a population with ab gametes, then in the offspring there will be more AB and ab gametes if the genes are linked. However, after a number of generations, the number of AB and ab gametes will approach that of the recombinant aB and Ab gametes, indicating linkage equilibrium. If the linkage is closer, this process will take longer. But ultimately, even if the distance between two genes is less than 1 cM, genes will become in linkage equilibrium (with no selection).

Hence, linkage disequilibrium is due to
- recent migration or crossing
- selection
- recent mutation.

<u>Linkage disequilibrium is essential for mapping</u>.
We may expect full disequilibrium between linked genes within a family, as the number of recombinants is the result of one meiosis event. Similarly, the same disequilibrium exists between a cross of inbred lines, such as in the testcross example above.
However, in most other cases, at population level, genes are in linkage equilibrium. The important consequence is that if we find a particular allele at one gene (e.g. a

marker) we cannot say which allele at another gene (e.g. at a QTL) should be expected. However, such statements are possible within families or across all families in a population if it was a recent cross from inbred lines, as in such cases there is linkage disequilibrium.
Population-wide linkage disequilibrium exist in the case of selection, or with linked loci short after crossing or migration, or when two genes are so close that hardly any recombinations occur.

Mapping functions

The distance between two genes is determined by their recombination fraction. The map-units are Morgans. One Morgan is the distance over which, on average, one crossover occurs per meiosis.
When considering the mapping of more than two points on the genetic map, it would be very handy if the distances on the map were additive. However, recombination fractions themselves are not additive. Consider the loci A, B and C. The recombination fraction between A-C is not equal to the sum of the recombination fractions AB and BC.

Say, the distance A-B is r1, the distance B-C is r2, and the distance A-C = r12 depends on the existence of interference.

If the recombination between A and B (with probability r1) is independent from the event of recombination between B and C (with probability r2), we say that there is *no interference*.
In that case, the recombination between A and C is equal to r12 = r1 + r2 - 2*r1*r2.

Interference is the effect in which the occurrence of a crossover in a certain region reduces the probability of a crossover in the adjacent region.

The last term is a reflection of the double crossovers. If there is *complete interference* the event of a crossover in one region completely suppressed recombinations in adjacent regions.
In that case r12 = r1 + r2, i.e. the recombination fractions are additive.
Also within small distances, the term 2r1r2 may be ignored, and recombination fractions are nearly additive. More generally, double recombinants can not be ignored, and recombination fractions are not additive.

If distances were not additive, it would be necessary to redo a genetic map each time when new loci are discovered. To avoid this problem, the distances on the genetic map are mapped using a *mapping function*. A mapping function translates recombination frequencies between two loci into a map distance in cM.

A mapping function gives the relationship between the distance between two chromosomal locations on the genetic map (in centiMorgans, cM) and their recombination frequency.

Two properties of a good mapping function is that

- Distances are additive, i.e. the distance AC should be equal to AB + BC if the order is ABC
- A distance of more than 50 cM should translate into a recombination fraction of 50%.

In general, a mapping function depends on the interference assumed.

With *complete interference*, and within small distances, a mapping function is simply: distance (d) = r (recombination fraction).

With *no interference*, the *Haldane mapping function* is appropriate:

$$d = -\ ½ \ln(1-2r).$$

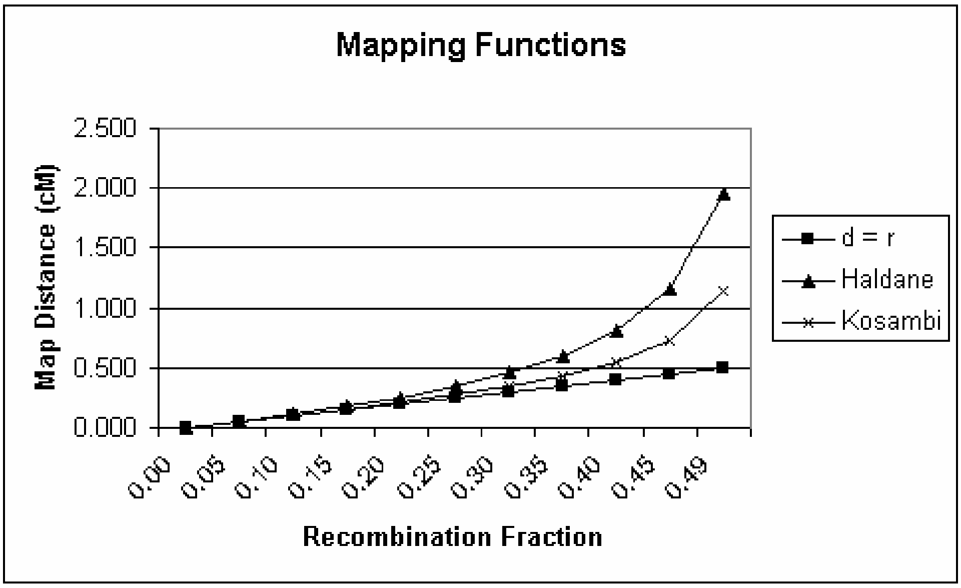
and given the map distance (d) the recombination fraction can be calculated as

$$r = ½ (1-e^{-2d})$$

Kosambi's mapping function allows *some interference*:

$$d = ¼ \ln[(1+2r)/(1-2r)]$$

The different mapping functions are depicted in Figure 1. Below 15 cM there is little difference between the different mapping functions, and we can safely assume that d = c.



Notes:
There is *no general relationship* between genetic distance and physical distance (in base pairs) The is a large variability between species for the average number of kilo base pairs (Kb) per centiMorgan. For humans this average is about 1000 kb per cM. Even within chromosomes there is variation, with some regions having less crossovers, and therefore more Kb per cM, than other.

The number of recombinations is not equal in the two sexes. It is usually lower in the heterogametic sex. In mammals, the female map is longer than the male map, as in females there are more recombinations for a certain stretch of DNA

Mapping of genetic markers

Genetic markers can be mapped relative to each other by

- Determining recombination fractions
- Using a mapping function

Such genetic mapping can only place markers on the genetic map, relative to each other. For a whole genome map, some markers need to be anchored to their physical position, using *in-situ* mapping. Several molecular techniques are available, e.g. FISH (Fluorescent In-Situ Hybridization)
Recombination fractions between genetic markers can be estimated from mapping experiments (as in a test cross). Since we can observe complete marker genotypes, we do not fully rely on such specific designs as in a testcross. However, some designs are more efficient for mapping than other designs, determining the percentage of meiosis observed that is actually informative

Estimation of the recombination fraction

Recombination fractions are estimated from the proportion of recombinant gametes. This is relatively easy to determine if we know

- Linkage phase in parents
- The haplotype of the gamete that was transmitted from parent to offspring

If the linkage phase is known in parents, we know can know which gametes are recombinants, and which ones are non-recombinant.
However, in practice, linkage phases are not always known. This is especially the case in animals, as it is hard to create inbred lines. And markers are often in linkage equilibrium, even across breeds.
If the linkage phase is not known, we can usually infer the parental linkage phase, as the number of recombinants is expected to be smaller than the number of non-recombinants. However, there is some chance that by chance there are more recombinants. Maximum Likelihood is used to determine the most likely phase, and therefore, to determine the most likely recombination fraction (see next section)

Information about the gamete that was received by an offspring depends on the genotypes on offspring, parents. If parents and offspring are all heterozygous (e.g. Aa), then we don't know which allele was paternal and which was maternal. If marker genotypes of parents are not heterozygous, we have no information about recombination events during their meiosis. For example, if the sire has genotype AB/Ab we cannot distinguish between recombinant gametes. However, if one parent is homozygous, it increases the chance of having informative meiosis on the other parent (think about a testcross, or see next example)

Maximum likelihood estimation of linkage (recombination fraction)

The likelihood is equal to the probability of observing a certain data set for given parameter values. In linkage studies, the most important parameter involved is recombination fraction. Other parameters can be population allele frequencies, but these are not needed if all parents are genotyped.
We use an example as described by Bovenhuis and Meuwissen (1996).
A sire with genotype AaBb and dam with genotype AABB are mated to produce offspring AABB.

We know for sure that the offspring received an AB gamete from both parents. However, we don't know whether this was a recombinant or a recombinant gamete. This depends on the phase in the sire. The dam produces an AB gamete with probability 1.

We have:

Sire's genotype	Probability	Probability of creating AB gamete
AB/ab	0.5	0.5*(1-r).
Ab/aB	0.5	0.5*r

r = recombination fraction

The probability (likelihood) for the parents and this offspring is then:
0.5*{0.5*(1-r)}+0.5*{0.5*r} = 0.25

The probability does not contain r, hence this offspring by itself does not provide information about the recombination fraction (r).

Now consider another offspring with genotype AABB.
We have then:

Sire's genotype	Probability	Probability of creating 2 AB gametes
AB/ab	0.5	$0.25*(1-r)^2$.
Ab/aB	0.5	$0.25*r^2$

r = recombination fraction

The probability (likelihood) for the parents and these two offspring is then:
$0.5*\{0.25*(1-r)^2\}+0.25*\{0.5*r^2\} = 0.125*\{(1-r)^2 + r^2\}$

Now the Likelihood is a function of the recombination fraction r. The maximum likelihood can be found with certain search routines. The value of r, which maximizes the Likelihood, is the ML estimate of r.

The small example is still not very informative, as we have only one kind of gamete in offspring. We can further expand the example by giving 20 offspring to these parents. In summary the data looks like:

Sire: AaBb
Dam AABB
20 Offspring: 9 AABB; 1 AaBB ; 1 AABb ; 9 AaBb

The dam always gives an AB gamete. The sire gives gametes AB, Ab, aB, ab in frequencies 0.45, 0.05, 0.05 and 0.45.

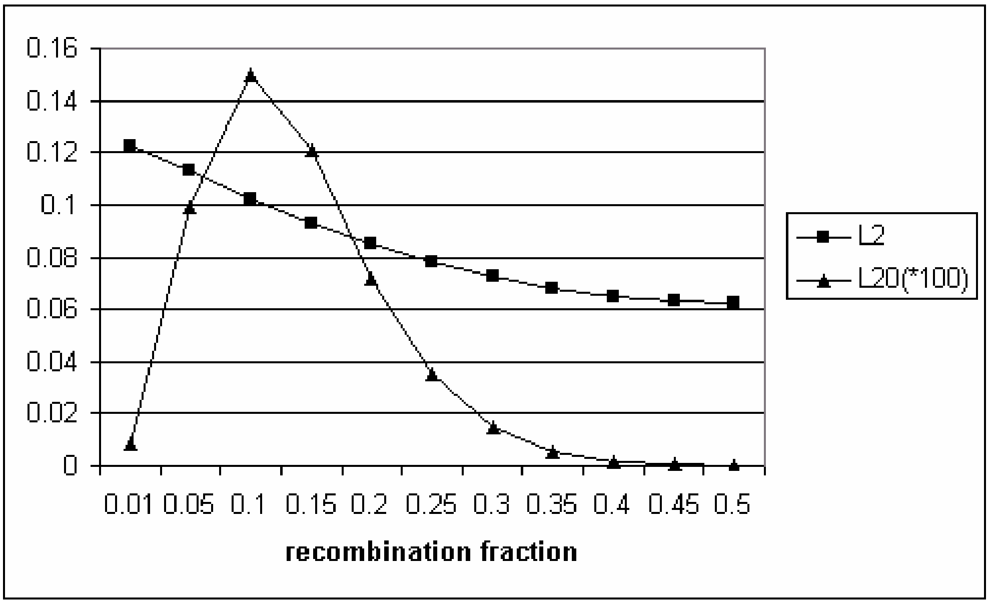
The data shows clearly that AB and ab are parental haplotypes (non-recombinant) and Ab and aB are recombinants.
The probability of obs4rving a certain number of recombinants can be calculated using the binomial distribution. The probability of observing 18 non-recombinants and 2 recombinants is equal to

$$\binom{20}{2}(1-r)^{18}.r^{2} \qquad [5.1]$$

This is equal to the likelihood. Note that we have now assumed known phase. Strictly, we should also consider the other possibility, i.e. that the phase in the sire was Ab/aB. This would give an additional term to the likelihood equal to

$$\binom{20}{2}(1-r)^{2}.r^{18}$$

However, this term is always very small compared to the previous, and therefore, in such cases it would not matter that much for the likelihood value whether or not if phase was assume known or not, as there is such overwhelming evidence from the data.
The next figure plots the likelihood against recombination fraction for the example with 2 (L2) offspring and for the example for 20 offspring (L20, multiplied by 100). The first term in formula [5.1] is ignored, as this term is constant and not dependent on recombination fraction.



Testing for linkage: LOD scores

Besides estimating the most likely recombination fraction, we usually also want to test those estimates statistically. In particular we want to test whether or not two loci are really linked. Therefore, the statistical test to perform is the likelihood versus a certain recombination fraction (r) vs the likelihood of no linkage (r=0.5).
Different likelihoods are usually compared by taking the ratio of the likelihood.
In this case:

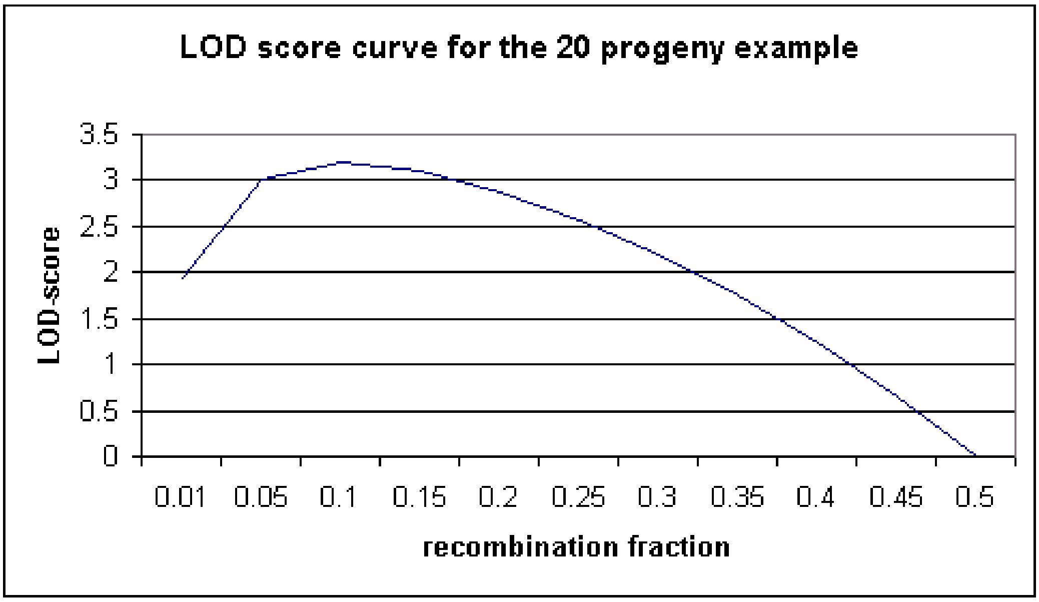
$$\frac{Likelihood\,(r = \hat{r})}{Likelihood\,(r = 0.5)}$$

The ^{10}log ratio of this likelihood ratio is indicated by LOD-score (abbreviation of log-of-odds) (Morton, 1955)
A LOD-score above 3 is generally used a critical value. A LOD-score>3 imply that the null-hypothesis (r = 0.5) is rejected. This value implies a ratio of likelihoods of 1000 to 1.
This seems like a very stringent criterion. However, it accounts for the prior probability of linkage. Due to the finite number of chromosomes, there is a reasonable probability (5% in humans with 23 chromosome pairs) that two random loci are linked (see Morton, (1955) for more detail)
Morton (1955) suggested that LOD scores from data from additional families, or from additional progeny within a family, could be added to the original LOD score.

The LOD score for the example, for a particular r-value can be written as

Z(r) = (n - nrec).log((1 - r) + nrec.log(r) – n.log(0.5)

Where n is the number of progeny and nrec the number of recombinants.



Note that these LOD scores assume the phase in the sire to be known. For r = 0.1 the LOD score is equal to 3.2.The LOD score would be somewhat lower if the phase was assumed unknown. You may want to check for yourself that that would give a LOD score of 2.9.

There is a lot of software written for linkage analysis and marker mapping. A well-known program is CRI-map. It gives LOD scores, estimates of recombination fractions, and marker maps (based on Kosambi’s function) for possibly many families, and many markers.

Design of mapping

Marker maps can be made from genotyping certain families that are genotyped for a series of markers. To construct the marker map for livestock species, most labs have used DNA from certain reference families. However, there are no strict rules for creating a reference families. A few comments can be made about efficiency of mapping.

- The amount of information available for mapping is based on the number of informative meiosis.

- An efficient design minimizes the number of genotypings for a given number of informative meioses.

From the previous we already noticed that informative meiosis depend on the number of marker alleles and hetero/homo-zygosity of parents. Some suggestions are:

- Full sibs families are better than half sib families, as the number of genotypings is lower for the same number of informative meiosis.

- It is better to use more families as two parents might have such genotypes at certain markers that they will never produce informative meioses.

References

Bovenhuis, H. and T.H.E. Meuwissen. 1996. Detection and mapping of quantitative trait loci. Animal Genetics and Breeding Unit. UNE, Armidale, Australia. ISBN 1 86389 323 7

Bulmer, M.G. 1971. The effect of selection on genetic variability. Amer. Nat. 105:201.

Morton, N.E. 1955. Sequential tests for the detection of linkage. American Journal of Human Genetics. 7:277-318.