Introduction and Principles of Linkage Analysis

Julius van der Werf and Brian Kinghorn

Introduction and Principles of Linkage Analysis

Julius van der Werf and Brian Kinghorn

Segregation analysis

Segregation analysis by example

Figure 1 shows a simple pedigree in which all females have been genotyped for a 2-allele locus (A,a). Yellow (light colored) lines emanate from females, and red (dark colored) lines from males. Can we deduce the genotypes of the males?

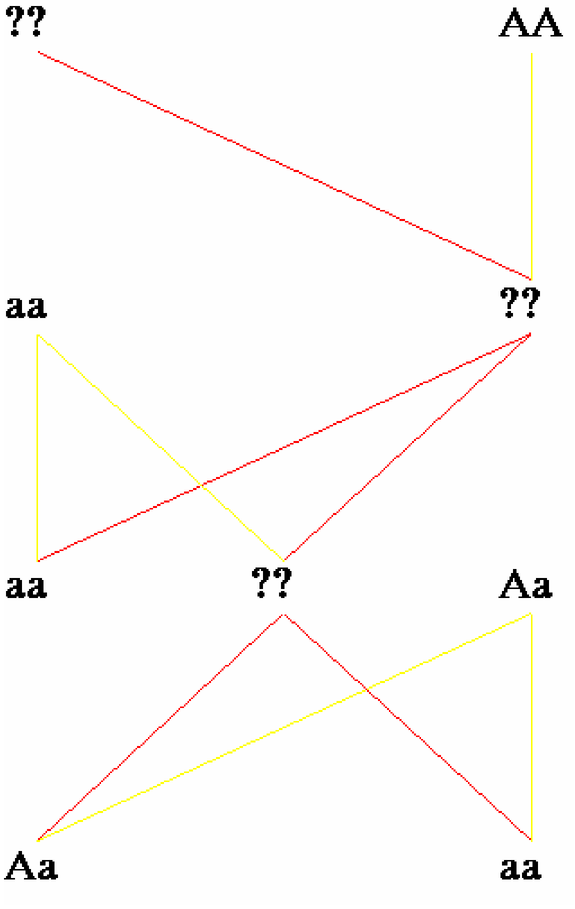


Figure 1. A simple pedigree in which females have been genotyped.

- The male in the second row is quite easy. Can you explain why he must be a heterozygote?

- The male in the first row is more difficult, and the answer depends on our prior assumption about gene frequency. This is taken as *p(A)* =.4 in Table 1.

- The male in the third row is even more difficult, but the answer in this case does not depend on gene frequency. Can you explain briefly why this is the case?

Table 1. Genotype probabilities for the males in Figure 1.

Male in row	p(aa)	p(Aa)	a(AA)
1	0.6	0.4	0
2	0	1	0
3	2/3	1/3	0

We might be able to deduce some results for simple cases like this, by relatively simple logic. But what about bigger examples in more realistic, complex pedigrees?

If we have 20 ungenotyped animals we have up to 3 to the power 20 (=3,486,784,401) 'possible' answers for the 2-allele, 3-genotype case. This makes solution by simple searching methods not feasible.

Most methods for doing this sort of thing make use of information from three sources:

1. Parents
2. Self
3. Mate(s) plus progeny,

... and use that information either recursively or iteratively over a number of cycles. Care has to be taken not to overuse information (double dipping) or mistreat loops in the pedigree (eg. as with inbreeding). At UNE, we currently use Richard Kerr's iterative method (Kerr and Kinghorn, 1996), coded in the program GENEPROB.

An example - the 'halothane gene' in pigs:

The halothane-sensitive variant of this gene causes porcine stress syndrome (PSS) and pale soft exudative (PSE) meat, as well as malignant hyperthermia on exposure to the gas halothane. It also gives higher lean percent. Breeding companies want a handle on this gene, but the DNA test is expensive.

In the example given here, the data set contains 4207 pigs in a complex pedigree structure. Money was spent genotyping 113 of these animals by DNA test at the Ryanodine receptor locus (the 'halothane' locus). Of these, 65 were normal homozygotes, 40 were heterozygotes and 8 carried two copies of the unfavourable allele.

After segregation analysis, an additional 1886 animals could be excluded form one genotype class and an additional 42 animals could be genotyped, both with 100 percent confidence. At the 90% confidence level (ie. probability of being any given genotype > 0.9) an additional 263 animals could be genotyped.

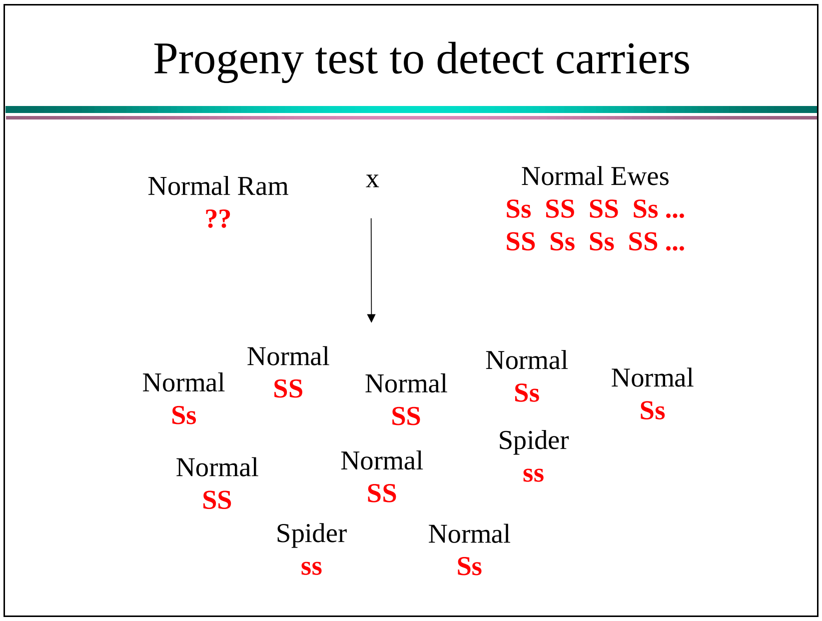
All animals had probabilities calculated for the halothane gene, and this information gave a basis for more accurate estimation of the effects of the halothane gene, using data from all pigs in the data set - not just those that were DNA tested.

To work out genotype probabilities in a large pedigree, ou need a computer program to do the job for you. Richard Kerr and Brian Kinghorn have written 'GENEPROB' for this. It works well and fast on large data sets, but does not fully account for inbreeding loops. Output is genotype probabilities (probabilities of being *AA*, *Aa* and *aa*, summing to 1) for each individual.

Another example - the spider syndrome in sheep:

The spider syndrome in Suffolk sheep is a recessive lethal condition. Development at the ends of the long bones is impaired and lambs end up on the ground with their legs splayed like a spider.

Classically, a progeny test is used to detect carriers, as illustrated in Figure 3. However, using segregation analysis, as in GENEPROB, information from all relatives can be used, such that some results are available at the time of making first breeding decisions - we do not really have to wait for the results from a progeny test.



Progeny test to detect carriers

Normal Ram
??

x

Normal Ewes
Ss SS SS Ss ...
SS Ss Ss SS ...

Normal
Ss

Normal
SS

Normal
SS

Normal
Ss

Normal
Ss

Spider
ss

Normal
SS

Normal
SS

Spider
ss

Normal
Ss

Figure 3. The classical progeny test approach to detecting carriers of deleterious recessive genes. The ram with a normal phenotype must be a carrier as he has some affected offspring.

In the example used here, a small data set containing 167 sheep included 6 spider lambs. These lambs were the progeny of just 3 rams and 3 ewes, which must be carriers. GENEPROB was run, and Figure 4 shows the resulting probabilities of being a carrier.

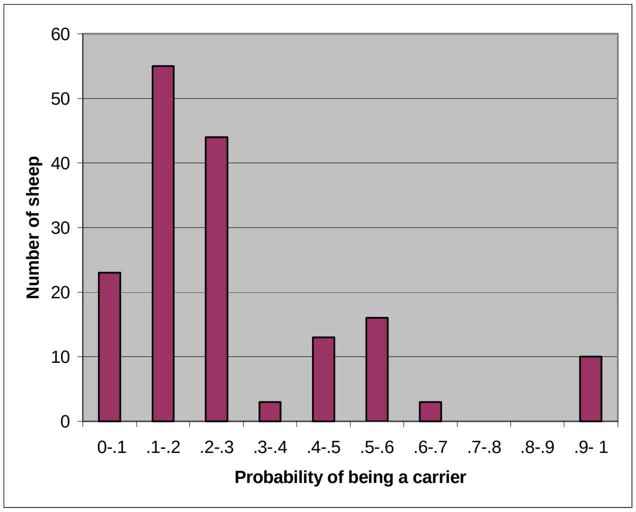


Figure 4. Frequency distribution of probability of being a spider carrier. This information comes from knowledge of the spider status of just six spider lambs.

These probabilities give breeders much more power to monitor and manipulate genes such as the spider gene. It is quite easy to include them in selection indices to provide appropriate penalty against animals probably carrying unwanted variants of such genes. TGRM can be used to help balance the speed of getting rid of such deleterious recessive genes [which is enhanced by mating to give higher incidence and thus more information in the medium term], and avoiding too much expression of the deleterious condition while doing so.

A DNA test for the spider gene has recently been developed, but this was not available for the current example. However, this example shows the power of segregation analysis in cases where such tests are not available, or too expensive.

It can be noted that DNA test results can be used to drive segregation analysis, to get genotype probabilities for all untested animals. Moreover, DNA test results and incidence information can both be used together for maximum power.

For conditions that have a DNA test available, there can still be some hard decisions to make because of the cost of testing. Strategies to decide on which animals to DNA test have been developed (Kinghorn, 1999). These involve an iterative approach, with batches of one or more animals DNA tested at each cycle. Segregation analysis is run between each cycle of DNA testing, with animals chosen for testing at each cycle in a manner that gives most improvement in genotype probabilities across the whole population.

References

Kinghorn, B.P. 1999. Use of segregation analysis to reduce genotyping costs. J. Anim. Breed. Genet. In Press.

Kinghorn, Kennedy and Smith. 1993. Genetics 134:351-360

Basics of Linkage and Gene Mapping

Julius van der Werf

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. 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 The recombinants can easily be identified among the phenotypes in the offspring of a *testcross* and the resulting phenotypes reflect the gamete frequency of the F1 parent.

A testcross is

F1	AaBb	x	aabb	parent 2	
Offspring	AaBb	Aabb	aaBb	aabb	

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%, then the genes are in linkage equilibrium. Hence, if we look at the gamete-frequencies, then we speak of linkage equilibrium if the

$$\text{freq}(AB) = \text{freq}\ (Ab) = \text{freq}\ (aB) = \text{freq}\ (ab).$$

And the amount of disequilibrium is measured as

$$D = \text{freq}(AB).\text{freq}(ab) - \text{freq}(Ab).\text{freq}(aB).$$

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, migration, or mutation, 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.

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.

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.

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:

$d = -\tfrac{1}{2}\ln(1-2r)$.

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

$r = \tfrac{1}{2}(1-e^{-2d})$

Kosambi's mapping function allows *some interference*:

$d = \tfrac{1}{4}\ln[(1+2r)/(1-2r)]$

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

$$r = \tfrac{1}{2}\frac{\exp(4d)-1}{\exp(4d)+1}$$

The different mapping functions are depicted in Figure 5.. Below 15 cM there is little difference between the different mapping functions, and we can safely assume that d = c.

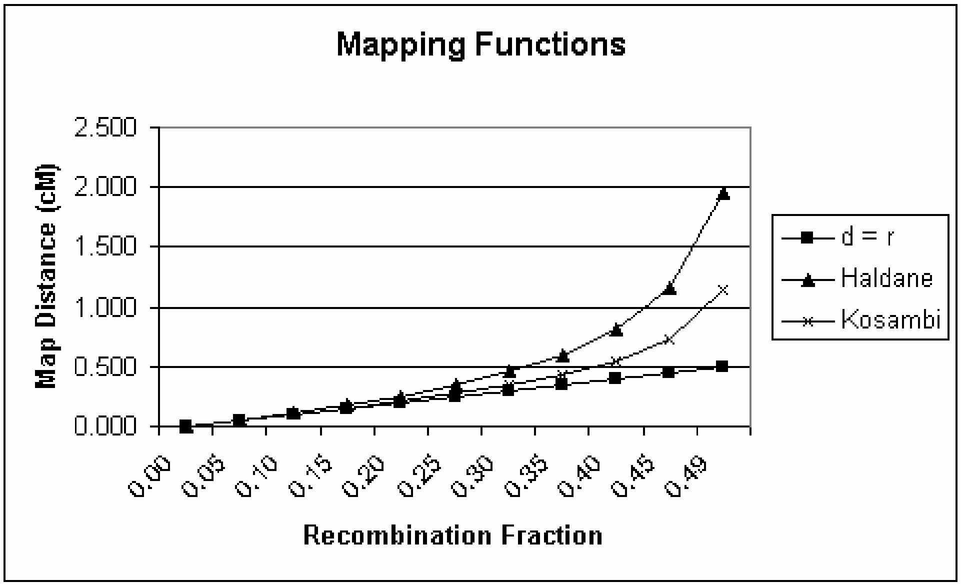


Figure 5: Map distance and recombination rate for 3 mapping functions

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 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 observing 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.

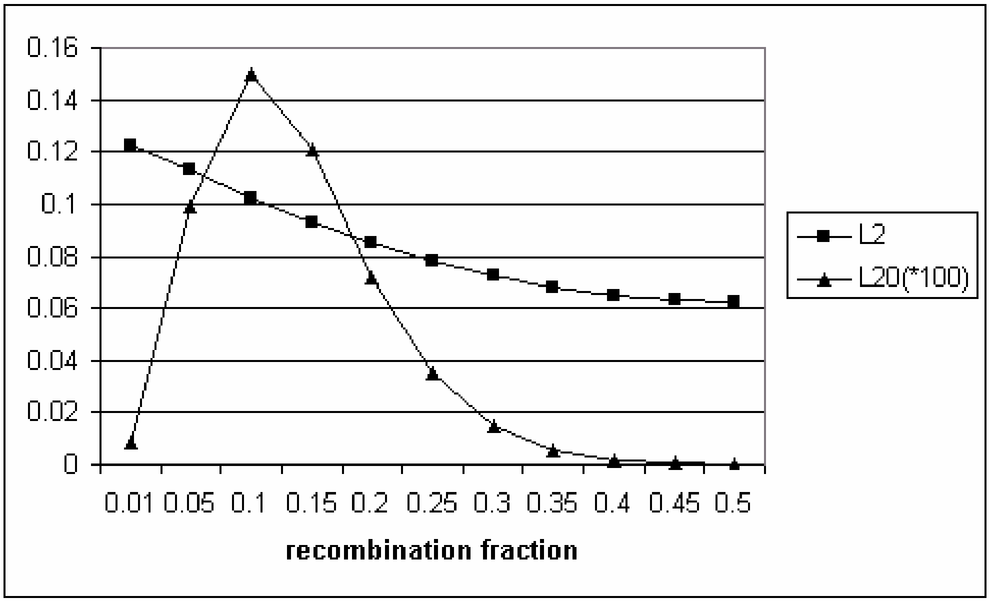


Figure 6. Likelihood patterns of linkage maps for examples with 2 (l2) and 20 (L20) progeny respectively.

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 implies 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.

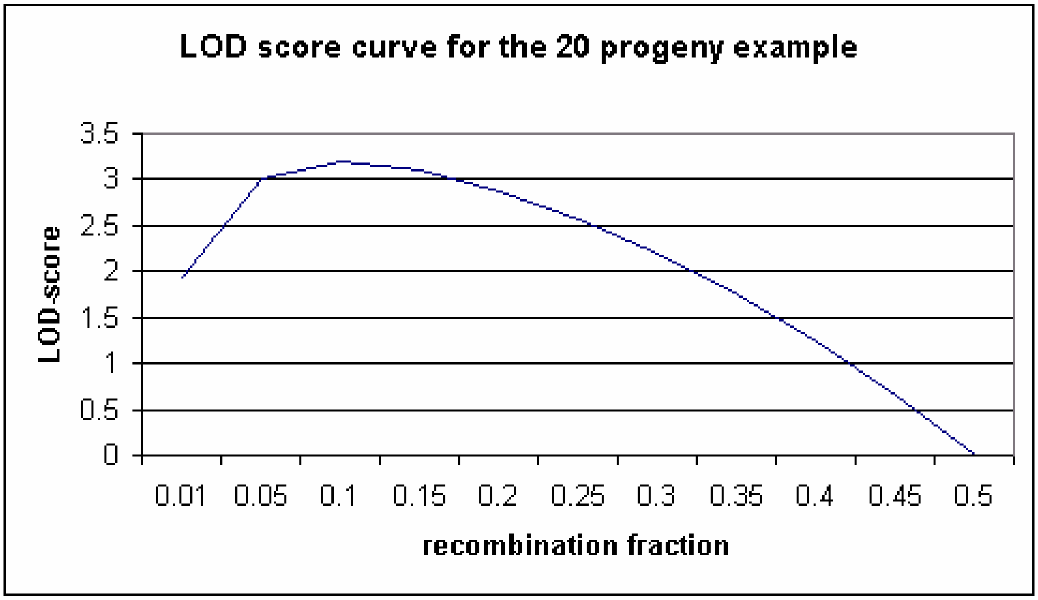


Figure 7: LOD score curve for the 20 progeny example.

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 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.

Determining associations between genetic markers and QTL with single markers

Julius van der Werf and Brian Kinghorn

Detecting QTL segregation in a sire family

Consider a sire that is heterozygous for both genetic marker (M, m) and QTL (Q, q). We can actually determine his marker genotype from a DNA test, but his QTL genotype can only be postulated. Consider only 'informative progeny' that is, progeny form whom we know which of the marker alleles they obtained from their sire. This is true in Figure 8 where all dams are recessive mm:

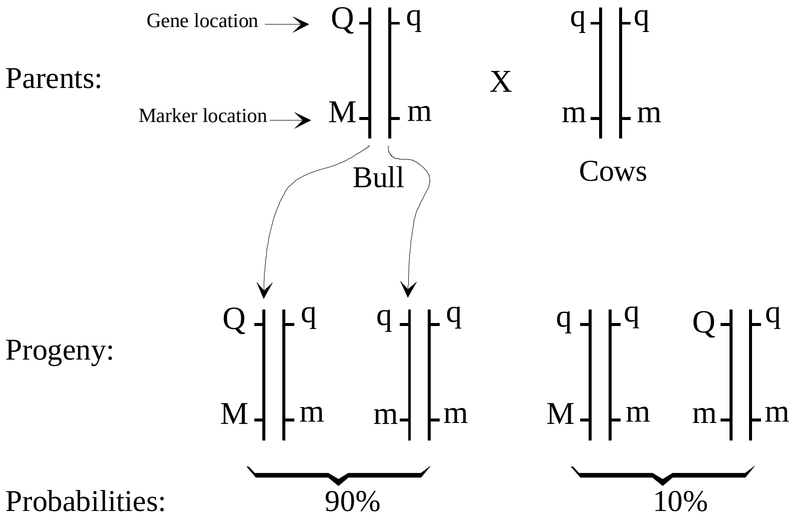


Figure 8: Example of QTL marker segregation

The next table shows what kind of progeny will be in the group, with their frequencies. What is important here:

– The *recombination rate* between marker allele and QTL allele (taken as 10% above). This is indicated by the symbol r.

– The difference between progeny receiving Q vs. progeny receiving q from their sire. This is equal to the allele-substitution effect with the symbol α.

Parental genotype:	M m	Q q		
	Possible gametes		**recombination?**	**gamete probability**
	M	Q	no	(1-r)/2
	m	Q	yes	r/2
	M	q	yes	r/2
	m	q	no	(1-r)/2

The need to treat each sire family separately

Note that, using this simple approach, these probabilities hold only within the progeny group of a given sire. In another family, the sire may have another QTL-allele associated with the M-allele. i.e. the sire maybe Mq/mQ.

Expected progeny means

For the given QTL effect we can now work out the expected progeny means of each marker group within the sire's progeny group:

Marker allele mean obtained from sire group	QTL allele obtained from sire	frequency	Expected of progeny
M	Q	$(1-r)/2$	$\mu + \alpha$
M	q	$r/2$	μ
m	Q	$r/2$	$\mu + \alpha$
m	q	$(1-r)/2$	μ

From the previous table we can work out the expected difference between the M-group and the m-group:

Mean of M-group:	$((1-r)(\mu + \alpha)/2 \quad + \quad r.\mu/2) \quad / 0.5 =$	$\mu + (1-r)\ \alpha$
Mean of m-group:	$(r.(\mu + \alpha)/2 \quad + \quad (1-r).\mu/2)/0.5 \ =$	$\mu + r\ \alpha$
Difference (D)		$(1-2r)\alpha$

Interpreting results

If there is no difference between the M-group and the m-group of progeny, then we have no evidence of a QTL. It could be that there is a QTL that is linked to the marker, but its recombination rate with the marker must be close to 0.5 (= unlinked), and/or its effect is small.

If we do find a progeny group difference (within the sire), it means that there is an association. However, we can't distinguish between a large QTL effect that is loosely linked (α high and r high) or a smaller effect that is tightly linked (α low and r low), as α and r cannot be untangled in $(1-2r)\alpha$.

The following Table shows some possible combinations of recombination rate (r) and QTL effect (α) giving rise to the same difference (D) between marker genotype groups within the sire's progeny.

Recombination rate M-Q (=r)	QTL-effect (Q-q substitution =α)	Mean of progeny[1] receiving M-allele	Mean of progeny[1] receiving m-allele	Marker allele contrast (D)
0	50	50	0	50
0.1	62.5	56.25	6.25	50
0.2	83	67	17	50
0.3	125	87.5	37.5	50
0.4	250	150	100	50
0.5	1000	500	500	0

[1] Mean is relative to the progeny receiving a q-allele from the sire.

The effect of QTL status in dams

Note that until now we have not worried what kind of alleles progeny would receive from their dam. In a way, this is not relevant for determining a marker-QTL association, as long as the dams are randomly distributed over the two marker groups of progeny (it is hard to be non-random here!). The dam population is relevant if we want to interpret the allele substitution effect (α). The possibilities are:

Dam population contributes	***The allele substitution effect (α) represents***
Only q-alleles	Difference between Qq and qq genotypes: $\mu_{qQ} - \mu_{qq}$
Only Q-alleles	Difference between QQ and Qq genotypes: $\mu_{QQ} - \mu_{qQ}$
Q in frequency p, q in frequency (1-p)	$p(\mu_{QQ} - \mu_{qQ}) + (1-p)(\mu_{Qq} - \mu_{qq})$

As in Falconer's book, we can define the following symbols for the means of the three possible genotype:

μ_{QQ}	+a
μ_{qQ}	d
μ_{qq}	-a

Where d is 'dominance' is the deviation of the heterozygote from the homozygote mean. In general, the Q-allele substitution effect can then be written as $\alpha = a + (q-p)d$. This shows that the QTL effect found in the sire's progeny is likely to be different when the sire is mated to another dam population (another breed), as the p (=freq(Q)) is likely to be different in this other population.

More powerful approaches

If we take account of the whole pedigree then we have some power to infer not just α under current conditions, but also a and d. This requires genotype information on sufficient animals (ideally on all animals) and use of statistical methods that we will cover later in the course.

Detecting QTL in designed experiments

Inbred parental lines

We have already seen in the Figure on the first page of this Chapter that there is a simple basis to infer QTL segregation in a cross that involves an ideal pattern of marker and QTL genotypes. The Figure actually refers to a backcross between inbred lines. The sire is heterozygous and the dams are all homozygotes. We can generally only be sure about such genotypes (at QTL and markers) if parental lines are fully inbred. In animal populations, this will be hard to achieve, of course. Using designed experiments helps to improve to get (most) of the right genotypes. For example, in a first cross of extreme breeds, it is more likely that major genes for the trait that is considered extreme is these breeds will be in a heterozygous state. We will now first look at some typical designs, and discuss the inferences that can be made with respect to genotype differences. Subsequently we consider outbred populations.

The Backcross Design

One way to maximise the probability of getting ‘ideal QTL genotypes’ is by making a backcross of inbred lines. Here the sire is a first cross between the lines and the Dams are purebred for one inbred line. The dams are all nicely homozygous and genetically identical to each other. The only things left to chance is that the two inbred lines are fixed for different alleles at both the QTL locus and the Marker locus. The Marker locus is no problem – we can tell pretty quickly from DNA test results whether the lines differ. However, for QTL loci, we can maximise the probability that the lines differ by choosing the lines appropriately – with large genetic distances and large differences for the key traits of interest. [Of course large genetic distances will also increase the chances of differences at market loci too.]

Earlier we showed that the difference in merit between progeny receiving M from the sire and those receiving m from the sire is $(1-2r)\alpha$, where $\alpha = a + (p-q)d$, and genotype effects are:

μ_{QQ}	+a
μ_{qQ}	d
μ_{qq}	-a

However, with inbred lines we have extra information – information that the QTL allele frequencies are 1 and 0 (or 0 and 1) in the inbred lines ***if*** the QTL is segregating.

Thus $\alpha = a + (p-q)d$ is:

$a + (1-0)d = a + d$ if the dam population is qq – going from qq to qQ adds a + d

$a + (0-1)d = a – d$ if the dam population is QQ – going from qQ to QQ adds a - d

Thus:

1. The effects of a allele substitution is either (a + d) if the dams hold the less favourable allele, or (a – d) if the dams hold the more favourable allele.

2. The differences between marker genotypes is either $(1-2r)(a+d)$ or $(1-2r)(a-d)$ accordingly.

3. If we make both backcrosses, we can get independent estimates of a and d.

The F_2 cross Design

Now both sire and dam lines are heterozygous, given that there is segregation at both loci:

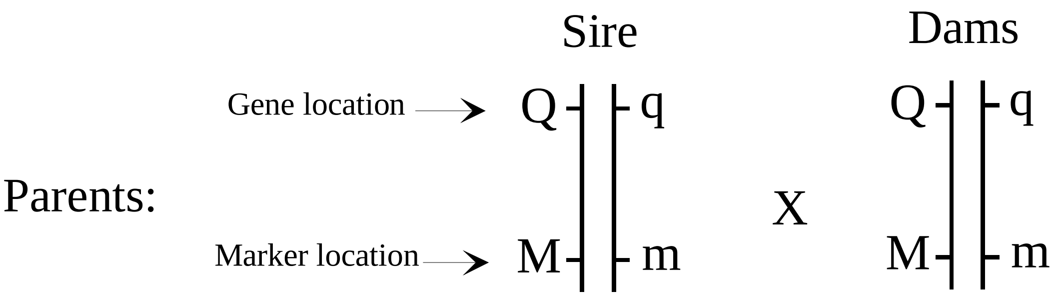


Figure 9: Paternal haplotypes when F1 parents are used to produce an F2 cross.

Given recombination fraction is r, we can work out gamete frequencies and progeny genotypes at both loci. The next table shows the genetic value (a, d or –a) and marker genotype (MM, Mm or mm) of the 16 possible 2-locus progeny genotypes:

Table 1.

	Eggs →	QM	qm	Qm	qM
Sperm ↓	Frequency	½(1-r)	½(1-r)	½r	½r
QM	½(1-r)	a MM	d Mm	a Mm	d MM
qm	½(1-r)	d mM	-a mm	d mm	-a mM
Qm	½r	a mM	d mm	a mm	d mM
qM	½r	d MM	-a Mm	d Mm	-a MM

Now we have three progeny groups, organised by marker genotype. By looking at the table above, we can derive the predicted frequency and merit for these:

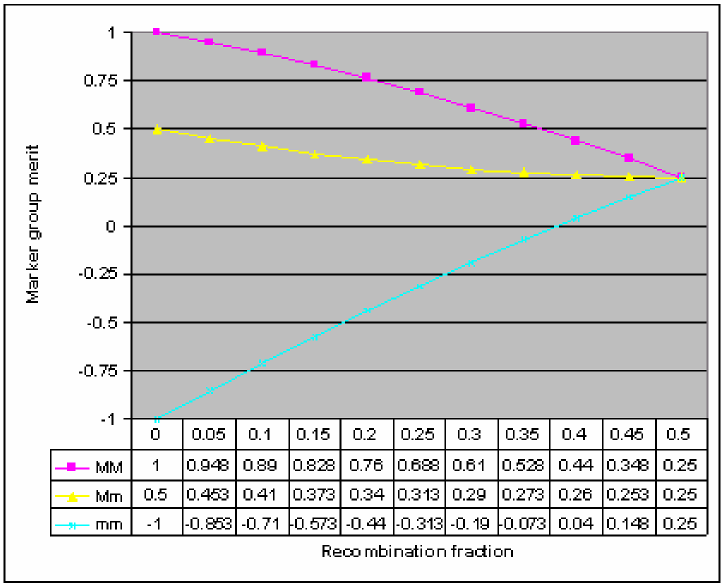
Marker genotype	Predicted frequency	Equals
MM	$(\frac{1}{2}(1-r))^2 + 2\frac{1}{2}r\frac{1}{2}(1-r) + (\frac{1}{2}r)^2$	$\frac{1}{4}$
Mm	$2[(\frac{1}{2}(1-r))^2 + 2\frac{1}{2}r\frac{1}{2}(1-r) + (\frac{1}{2}r)^2]$	$\frac{1}{2}$
mm	$(\frac{1}{2}(1-r))^2 + 2\frac{1}{2}r\frac{1}{2}(1-r) + (\frac{1}{2}r)^2$	$\frac{1}{4}$

Marker genotype	**Predicted merit**	Equals
MM	$\dfrac{(\frac{1}{2}(1-r))^2a + 2\frac{1}{2}r\frac{1}{2}(1-r)d + (\frac{1}{2}r)^2(-a)}{\frac{1}{4}}$	$(1-r)^2a + 2r(1-r)d + r^2(-a)$
Mm	$\dfrac{2[(\frac{1}{2}(1-r))^2d + 2\frac{1}{2}r\frac{1}{2}(1-r)(a-a) + (\frac{1}{2}r)^2d]}{\frac{1}{2}}$	$[(1-r)^2 + r^2]d$
mm	$\dfrac{(\frac{1}{2}(1-r))^2(-a) + 2\frac{1}{2}r\frac{1}{2}(1-r)d + (\frac{1}{2}r)^2a}{\frac{1}{4}}$	$(1-r)^2(-a) + 2r(1-r)d + r^2a$

This gives us some sensible predicted merits:

Marker genotype	$r = 0$	$r = \frac{1}{2}$
MM	a	$\frac{1}{4}a + \frac{1}{2}d - \frac{1}{4}a$
Mm	d	$\frac{1}{4}a + \frac{1}{2}d - \frac{1}{4}a$
mm	$-a$	$\frac{1}{4}a + \frac{1}{2}d - \frac{1}{4}a$

This is shown graphically below, with $a = 1$ and $d = \frac{1}{2}$ at the QTL. With no recombination, the marker groups reflect the true QTL genotypic merits. With full recombination ($r = \frac{1}{2}$) all marker groups are predicted to equal the population mean, which is $(p-q)a + 2pqd = \frac{1}{2}d$ – as $p = q = \frac{1}{2}$.



	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
MM	1	0.948	0.89	0.828	0.76	0.688	0.61	0.528	0.44	0.348	0.25
Mm	0.5	0.453	0.41	0.373	0.34	0.313	0.29	0.273	0.26	0.253	0.25
mm	-1	-0.853	-0.71	-0.573	-0.44	-0.313	-0.19	-0.073	0.04	0.148	0.25

Figure 10. Marker difference depending on recombination rate in an F2 cross

Non-inbred parental lines

The Backcross Design

If our parental lines are not inbred, there can be segregation at both QTL and marker loci in the parental lines. For the backcross design, the outcome is just as we found in Figure 8 – with a need to treat each family separately, if using simple analysis.

The F_2 cross Design

The big problem here is that progeny that are heterozygous for the marker locus are not informative (unless we have linked markers, more extensive pedigree information, and proper method, as will be described later in the course).

For an Mm progeny, we cannot tell if M came from the sire or the dam. However, for MM progeny, we can tell that allele M was inherited from each (and similarly for mm progeny), and if the parents are heterozygous then we have useful information. We are then left to contrast MM progeny and mm progeny.

Consider a sire of genotype QqMm (as shown in the diagram above). The distribution of progeny genotypes depends on the frequencies and phases of QTL and marker alleles in the population of dams. For example, assuming linkage equilibrium in the dam population, we can look at the distribution of progeny of marker genotype MM and mm. This is similar to Table 1, but with Mm and mM progeny excluded. the frequencies of Q and q are p and (1-p):

Table 2

	Eggs →	QM	qm	Qm	qM
Sperm ↓	Frequency within marker group→	p	(1-p)	p	(1-p)
QM	½(1-r)	a MM			d MM
qm	½(1-r)		-a mm	d mm	
Qm	½r		d mm	a mm	
qM	½r	d MM			-a MM

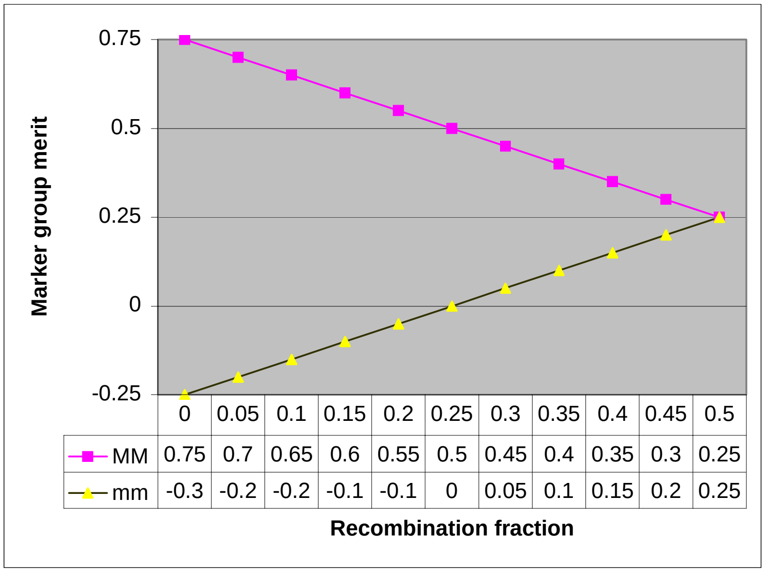
Under this assumption, the expectations of the marker group means are now:

Marker group	Expectation	Equals
MM	$\frac{\tfrac{1}{2}(1-r)pa + \tfrac{1}{2}r(1-p)(-a) + \tfrac{1}{2}rpd + \tfrac{1}{2}(1-r)(1-p)d}{\tfrac{1}{2}(1-r)p + \tfrac{1}{2}r(1-p) + \tfrac{1}{2}rp + \tfrac{1}{2}(1-r)(1-p)}$	$\frac{\tfrac{1}{2}(p-r).a + (r.p+\tfrac{1}{2}(1-p-r)).d}{\tfrac{1}{2}}$
mm	$\frac{\tfrac{1}{2}rpa + \tfrac{1}{2}(1-r)(1-p)(-a) + \tfrac{1}{2}r(1-p)d + \tfrac{1}{2}(1-r)pd}{\tfrac{1}{2}rp + \tfrac{1}{2}(1-r)(1-p) + \tfrac{1}{2}r(1-p) + \tfrac{1}{2}(1-r)p}$	$\frac{\tfrac{1}{2}(p+r-1).a + [\tfrac{1}{2}(r+p)-rp]d}{\tfrac{1}{2}}$

This is shown graphically below, with $a = 1$ and $d = ½$ at the QTL. With no recombination, the two marker groups no longer reflect the true QTL genotypic merits (as they did for inbred parental lines). This is because, even with no recombination, we do not know which maker allele is associated with which QTL allele in each dam. However, we can find this information for the sire, given sufficient progeny – not that it matters if the sire is heterozygous at both loci.

With full recombination (r = ½) both marker groups are predicted to equal the population mean, which is (p-q)a + 2pqd = ½d – as p = q = ½.

Of course, results will differ when there is some linkage disequilibrium in the dam population.



	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
■ MM	0.75	0.7	0.65	0.6	0.55	0.5	0.45	0.4	0.35	0.3	0.25
▲ mm	-0.3	-0.2	-0.2	-0.1	-0.1	0	0.05	0.1	0.15	0.2	0.25

With no recombination, and linkage equilibrium in the dams, MM progeny have a probability p of being QQ (merit +a) and (1-p) of being Qq (merit d). This can be seen by inspection of Table 2. At p=½ in the graph above, this comes out at a mean merit of 0.75.

Unfortunately, we have not been able to get independent estimates of a and d. We can do better than this – if we have large full sib families then we have a basis to infer linkage phases in each dam (as we do for the sire in this example). This can lead us to independent estimates of a and d.

Outbred populations

We can also do better if we have more than one marker locus, a richer pedigree, and good analysis methods.

With more loci we can often get information about which allele is inherited from which parent – even when the parents and progeny are all heterozygous for the same alleles. We should cover that later.

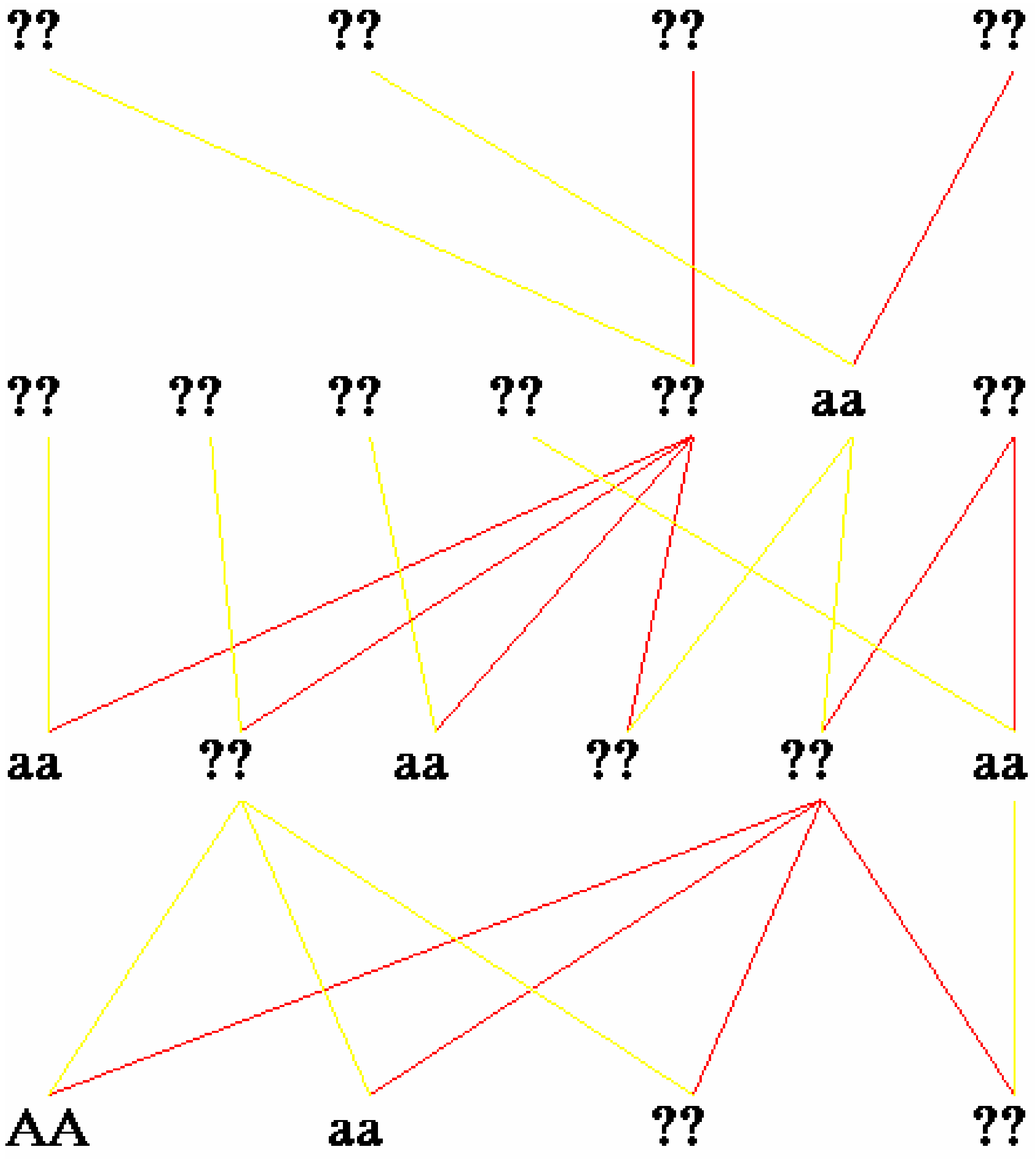
With richer pedigree and good analysis methods, we can infer the probabilities of being QQ Qq and qq for each animal in the pedigree.

Exercises

Segregation analysis

The object of this exercise is to use for large data sets as well as to get some feel for segregation analysis, if you are not already expert.

You can construct your own pedigree to work on, or alternatively adopt the example given in this diagram:



For this pedigree, the file EXAMPLE.DAT is the input to the program GENEPROB.EXE which does the analysis. Results go to the screen and to the file GENEPROB.PED. This result file can be read by the Pedigree Viewer for easy browsing of the results. You will be shown how to do this.

EXAMPLE.DAT:

```
COMMENT: Example for segregation workshop
genefreq phenotypes(excl missing ones=9):
.5 3
phenotype then  g(f|u), u (aa Aa AA)  on columns ...
0   1  0  0
1   0  1  0
2   0  0  1
data format ...
 (3a5,i5)
data …(ID, sire, dam, phenotype (9 = missing)..  (Note that id's don't
have to be sequential as here)
    1    0    0    9
    2    0    0    9
    3    0    0    9
    4    0    0    9
    5    0    0    9
    6    0    0    9
    7    0    0    9
    8    0    0    9
    9    7    1    9
   10    8    2    0
   11    0    0    9
   12    9    3    0
   13    9    4    9
   14    9    5    0
   15    9   10    9
   16   11   10    9
   17   11    6    0
   18   16   13    2
   19   16   13    0
   20   16   13    9
   21   16   17    9
```

a. If necessary, copy files as instructed to your local hard or floppy disk.

b. Run GENEPROB from a DOS prompt and enter EXAMPLE.DAT as the input file. View the results in GENEPROB.PED using the Pedigree Viewer. Look at the declared genotypes (field 'Phen') and deduce the genotypes of as many of the ungenotyped animals as you can. View fields p(0) [probability of carrying zero A alleles], p(1) and p(2) to check your results.

c. Rerun after changing the prior estimate of gene frequency, and check that the changes in results are reasonable.

d. Rerun after making sensible changes to the penetrance values for the genotype/phenotype combinations (eg let there be some 'spillage' such that there is some finite probability that some genotypes are 'read' as belonging to the 'wrong' phenotypic class). You also can add extra phenotypes to represent, for example, dubious gel readings. Note that each of the three columns (aa Aa and AA genotypes) must add to unity no matter how many rows (phenotypic classes) they contain. Recall that each element in the array is the probability of observing phenotype *row* given knowledge that the genotype is (definitely) *col*.

Segregation problem:

The spider syndrome in Suffolk sheep is a recessive lethal condition. Development at the ends of the long bones is impaired and lambs end up on the ground with legs played like a spider.

Spider.dat contains 167 sheep numbered sequentially, together with sire and dam number (unknown parents are denoted 0), tag number, and phenotype: ‘1’ for normal and ‘2’ for spider syndrome. There are no unknown phenotypes in this case as the trait is so easy to score.

Run a segregation analysis to calculate the probability of each genotype (++ +S SS) for each individual in the data set. (NB: you will need to construct a suitable header for your version of spider.dat.)

Use Pedigree Viewer to examine your results.

Why is the probability of being SS [p(2)] always either 0 or 1?

Sort left to right on probability of being Ss [p(1)] and ‘Shade merit fields’. This will help to quickly identify likely carriers. Some animals are certain to be heterozygotes [p(1) = 1]. For each of these, deduce why this is so by inspecting the pedigree. Do this also for animals with values for p(1) which are high but less than unity. Can you always find good evidence to support this high probability of being a carrier?

Linkage analysis

1. There are several mutant eye colours seen in *Drosophila,* one of which is purple, while another mutation affects the length of the wings: vestigial giving a 'scrunched up' look. True-breeding wild red-eyed flies are $\boldsymbol{pr^+pr^+}$ and flies with purple eyes are ***prpr***. Pure breeding, normal-winged flies (wild type) are $\boldsymbol{vg^+vg^+}$ and vestigial-winged flies are ***vgvg***. The wild type is dominant

   Early last century, Morgan and his co-workers carried out many experiments with *Drosophila*. From one of these experiments involving the testcrossing of $\boldsymbol{pr^+pr}$ $\boldsymbol{vg^+vg}$ females to ***prpr vgvg*** males the following progeny were scored.

Phenotype	From female gametes	Observed
wild type	$pr^+\ vg^+$	157
normal, vestigial wing	$pr^+\ vg$	965
purple, normal	$pr\ vg^+$	1,067
purple, vestigial	$pr\ vg$	146
		2,335

- What does this tell us about the two genes?
- Identify the recombinant phenotypes in the progeny.
- Estimate the map distance between the two genes.
- What are the genotypes of the homozygous parental flies of the F_1 females used in the testcrosses? Are the alleles in the parents in coupling phase or repulsion phase?

Linkage Analysis by Maximum Likelihood

Verify that the LOD score for the 20 progeny example in the notes with no prior assumption about phase, should be equal to 2.9 for r = 0.10.

QTL analysis

Aim

Determine QTL-marker association of 4 marker loci, make inferences about effect and location of QTL, and determine whether the markers in this example can be used in a selection program.

Case study

A dairy bull is genotyped for 10 markers, and he was found to be heterozygous for 4 markers (A,B,C and D). In 100 of his 150 female offspring it was possible to determine which of the two marker alleles was obtained from the sire. Each of these female offspring produces a lactation record on milk production. The population average for milk production per lactation (corrected to a 305 days lactation period) is equal to 5000 Kg. The means for each group of offspring for the following paternal marker alleles was:

A1-	5025	C1-	4850
A2-	4975	C2-	5150
B1-	5200	D1-	5060
B2-	4800	D2-	4940

Based on linkage analysis, we know that marker A is located on chromosome 1, markers B and C are located on chromosome 4 and marker D is located on chromosome 19. The genetic distance between markers B and C has previously been estimated at 30cM with marker B at about 20 cM from the telomeric end.

Assume the group size of offspring for each marker allele was equal to 50. Also assume that the mean given are corrected for differences due to herd, age and season of calving etc. The within half-sib family standard deviation of milk production (for one lactation) is equal to 500kg.

- Test for each of the marker alleles whether there is a significant difference between the marker-haplotype groups.

You can use a t-test, assuming that the variance of the difference between two progeny group means is equal to $2\sigma^2/n$, where σ is equal to the within half sib family standard deviation, and n is the number of individuals in one marker-allele group.

- What does a significant group difference tell you about the existence of a putative QTL.
- Try to work out an expression where the difference between the marker-allele progeny groups is a function of the allele substitution effect and the recombination rate of QTL and marker
- Is it possible based on the information of a single marker to estimate the allele substitution effects of the QTL?
- What can you say about the location of the QTL when considering all information?
- Can you give an indication of which marker haplotype is associated with the positive QTL-allele?
- Describe how the current information could help in selection decisions in the breeding program, based on marker genotype information.