GENETIC LINKAGE ANALYSIS

Reference: http://www.infobiogen.fr/services/chromcancer/IntroItems/LinkageShortID30031ES.html
Disclaimer: This document is a reviewed version of what can be found at the link mentioned above. Original author: Françoise Clerget-Darpoux.

Investigating the linked segregation of genes situated at different loci is a way of testing the independence of their transmission. This concept of independence is also reflected in the recombination fraction, θ, which is the percentage of the gametes transmitted by the parents to be recombined. If they are transmitted independently, there will be the same number of recombined gametes as there are parental gametes, and so $\theta = 1/2$. If they are not transmitted independently, then the parental gametes are transmitted preferentially compared to the recombined gametes, and $0 \leq \theta < 1/2$. In this case, there is said to be "linkage" between the two loci.

I RECOMBINATION FRACTION

Let us consider the case of two loci, A and B, with two co-dominant alleles at each of these loci, A_1, A_2 and B_1, B_2 respectively. Such an individual can produce four types of gamete:

A_1B_1
A_2B_1
A_1B_2
A_2B_2

Two situations are possible:

1) The loci A and B are on different chromosome pairs

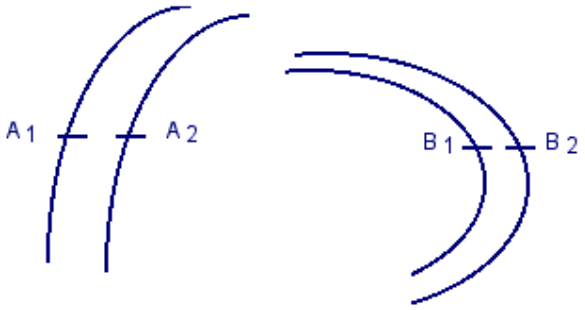

Figure 1

In this case, the four gametes all have the same probability: 1/4.

2) The loci A and B are on the same chromosome pairs

Here we have to distinguish between two possible situations: the alleles A_1 and B_1 may be on the same chromosome within the pair, in which case A_1 and B_1 are said to be "coupled"; or they may be on different chromosomes, in which case A_1 and B_1 are said to be in a state of "repulsion".

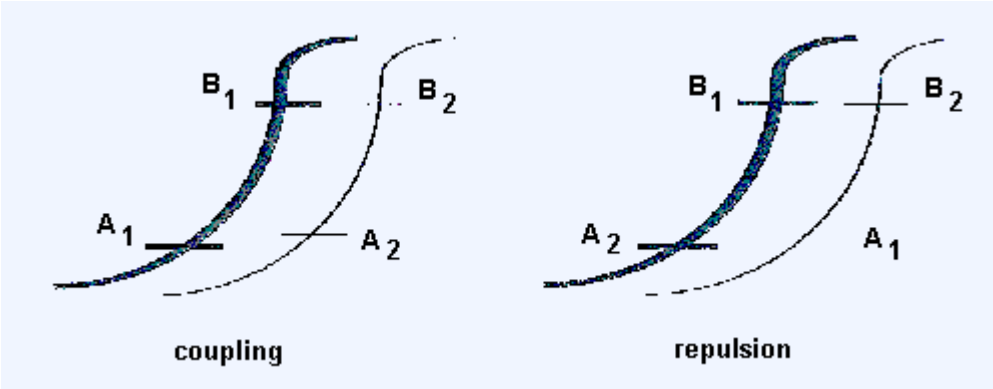

Figure 2

For instance, let us suppose that A_1 and B_1 are "coupled". Four types of gametes are still produced.

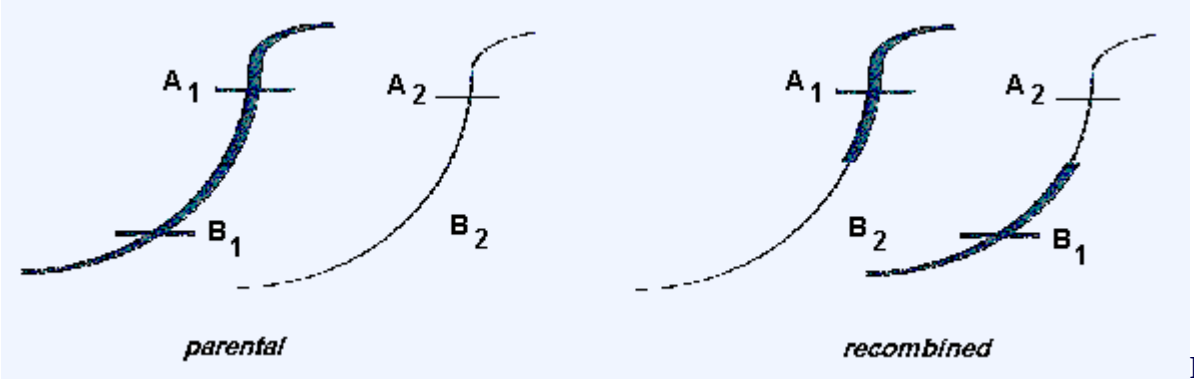

Figure 3

Gametes A_1B_1 and A_2B_2 are said to be "parental". In the offspring, as in the parents, A_1 is "coupled" with B_1 (and A_2 is "coupled" with B_2).

The gametes A_1B_2 and A_2 B_1 are therefore described as being "recombined". An uneven number of recombination or "crossing-over" phenomena have occurred between the A and B loci.

The proportion of recombined gametes amongst the gametes transmitted is known as the "recombination fraction".

θ= number of recombined gametes/number of gametes transmitted

Figure 4

Assuming that the crossing-over event for a pair of chromosomes follows Poisson's law, and knowing that a parental gamete has zero or an even number of crossings-over, whereas a recombined gamete has an odd number, we can show that the frequency of recombined gametes is always equal to or lower than that of the parental gametes and so

$$0 \leq \theta < 1/2$$

If $\theta = 1/2$, then all the gamete types have the same probability and the alleles at the loci A and B are transmitted independently. Loci A and B are therefore said not to exhibit genetic linkage. This is the

situation if A and B are on different pairs of chromosomes, and also if A and B are one the same pair, but at some distance from each other.

However, if $\theta < 1/2$, then the two loci are genetically linked.

For a couple of which the genotypes at the loci A and B are known, the probability of observing the genotypes of the offspring depends on the value of θ.

Let us assume the following crossing:

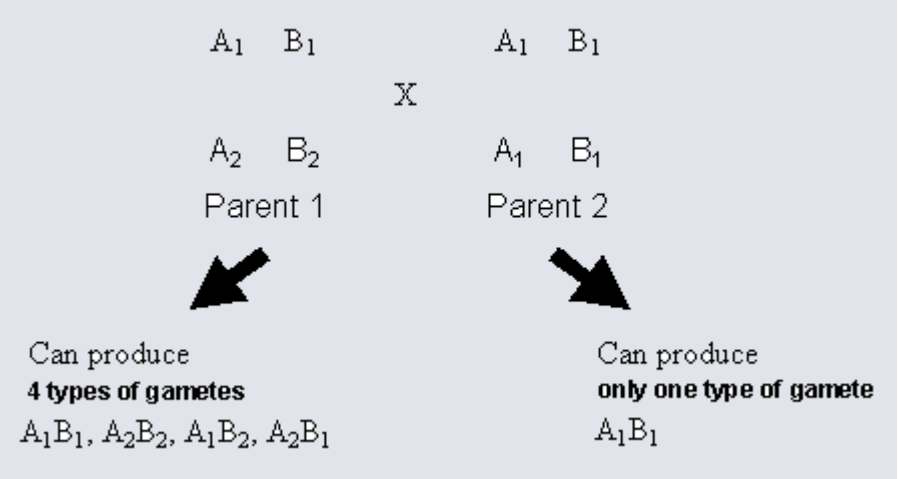

Figure 5

Therefore, such a couple can have 4 types of offspring

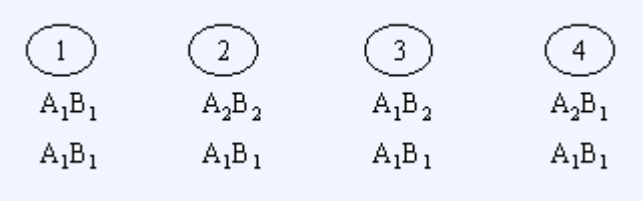

Figure 6

Assuming that there is gamete equilibrium at the A and B loci, in parent 1 there is a probability of 1/2 that alleles A_1 and B_1 will be coupled, and a probability of 1/2 that they will be in repulsion.

1) **A_1 and B_1 are coupled**, so the probability that parent (1) provides the gametes A_1B_1 and A_2B_2 is $(1-\theta)/2$ and the probability that this parent provides gametes A_1B_2 and A_2B_1 is $\theta/2$. The probability that the couple will have child of type (1) or (2) is $(1-\theta)/2$, and that of their having a type (3) or type (4) child is $\theta/2$.

The probability of finding n_1 children of type (1), n_2 of type (2), n_3 of type (3) and n_4 of type (4) is therefore

$$[(1-\theta)/2]^{n1+n2} \times (\theta/2)^{n3+n4}$$

2) **A_1 and B_1 are in a state of repulsion**, so the probability that parent (1) provides the gametes A_1B_2 and A_2B_1 is $(1-\theta)/2$ and the probability that this parent provides gametes A_1B_1 and A_2B_2 is $\theta/2$.

The probability of the previous observation is therefore:

$$(\theta/2)^{n1+n2} \times [(1-\theta)/2]^{n3+n4}$$

So in the end, with no additional information about the A_1 and B_1 phase , and assuming that the alleles at the A and B loci are in a state of coupling equilibrium, the probability of finding n_1, n_2, n_3 and n_4 children in categories (1), (2), (3), (4) is:

$$p(n_1,n_2,n_3,n_4/\theta) = 1/2\{[(1-\theta)/2]^{n1+n2} \times (\theta/2)^{n3+n4} + (\theta/2)^{n1+n2} \times [(1-\theta)/2]^{n3+n4}\}$$

So the likelihood of θ for an observation n1, n2, n3, n4 can be written :

$$L(\theta/n1,n2,n3,n4) = 1/2\ \{[(1-\theta)/2]^{n1+n2}\ (\theta/2)^{n3+n4} + (\theta/2)^{n1+n2}\ [(1-\theta)/2]^{n3+n4}\}$$

Special case: number of children n = 1

Regardless of the category to which this child belongs

$$L(\theta) = 1/2\ [(1-\theta)/2] + 1/2\ [\theta/2] = 1/4$$

The likelihood of this observation for the family does not depend on θ. We can say that such a family is not informative for θ.

Informative families

An "informative family" is a family for which the likelihood is a variable function of θ.

One essential condition for a family to be informative is, therefore, that it has more than one child. Furthermore, at least one of the parents must be heterozygotic.

Definition: if one of the parents is doubly heterozygotic and the other is

- A double homozygote, we have a backcross
- A single homozygote, we have a simple backcross
- A double heterozygote, we have a double intercross

II DEFINITION OF THE "LOD SCORE" OF A FAMILY

Take a family of which we know the genotypes at the A and B loci of each of the members.

Let L(θ) be the likelihood of a recombination fraction $0 \le \theta < 1/2$

L(1/2) be the likelihood of θ = 1/2, that is of independent segregation into A and B.

The lod score of the family in θ is:

$$Z(\theta) = \log_{10} [L(\theta)/L(1/2)]$$

Z can be taken to be a function of θ defined over the range [0,1/2].

Lod score of a sample of families

The likelihood of a value of θ for a sample of independent families is the product of the likelihoods of each family, and so the lod score of the whole sample will be the sum of the lod scores of each family.

III TEST FOR LINKAGE

Several methods have been proposed to detect linkage: the U scores, the sib pair test, the likelihood ratios, the lod score method. The lod score method is the one most commonly used at present.

The test procedure in the lod score method is sequential. Information, i.e. the number of families in the sample, is accumulated until it is possible to decide between the hypotheses H0 and H1 :

H0 : genetic independence; $\theta = 1/2$

and

H1: linkage of recombination fraction θ_1; $0 \leq \theta_1 < 1/2$

The lod score of the θ_1 sample

$$Z(\theta_1) = \log_{10} [L(\theta_1)/L(1/2)]$$

indicates the relative probabilities of finding that the sample is H1 or H0. Thus, a lod score of 3 means that the probability of finding that the sample is H1 is 1000 times greater than of finding that it is H0 ("lod = logarithm of the odds").

The decision thresholds of the test are usually set at -2 and +3, so that if:

$Z(\theta_1) \geq 3$ H0 is rejected, and linkage is accepted.

$Z(\theta_1) \leq -2$ linkage of θ_1 is rejected, and independence accepted.

$-2 < Z(\theta_1) < 3$ it is impossible to decide between H0 and H1. It is necessary to go on accumulating information.

For the thresholds chosen, -2 and +3, we can show that:

The first degree error, $\alpha < 10^{-3}$

The second degree error, $\beta < 10^{-2}$

The reliability, $1-\rho > 0.95 \; \forall \; \theta_1$

The power, $P(\theta) > 0.80 \; \forall \; \theta_1$ if the true value of $\theta < 0.10$

α = proba (H0 rejected /H0 true)
β = proba (H1 rejected /H1 true)
ρ = proba (H0 true /H1 concluded)

$P(\theta)$ = proba (H0 rejected / θ true value)

Figure 7

In fact, what is being tested is not a single value of θ_1 relative to $\theta = 1/2$, but a whole set of values between 0 and 1/2, with a step of various size (0.01 or 0.05).

If there is a value of θ_1 such that $Z(\theta_1) \geq 3$: linkage is concluded to exist.

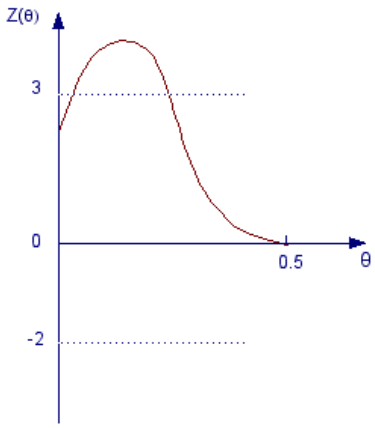

Figure 8

If there is a value of θ_1 such that

$$Z(\theta_1) = -2$$

The linkage is excluded for any $\theta \leq \theta_1$

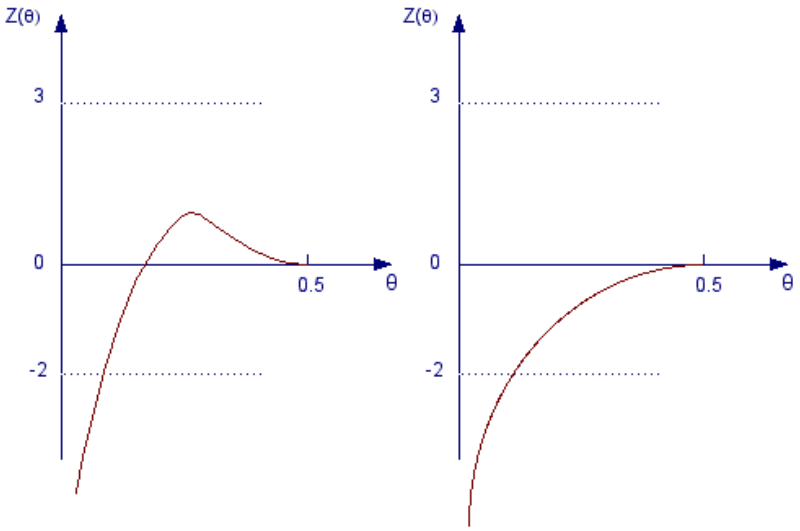

Figure 9

If $\forall\ \theta\ -2 < Z(\theta) < 3$, no conclusion can be drawn, the sample is not sufficiently informative.

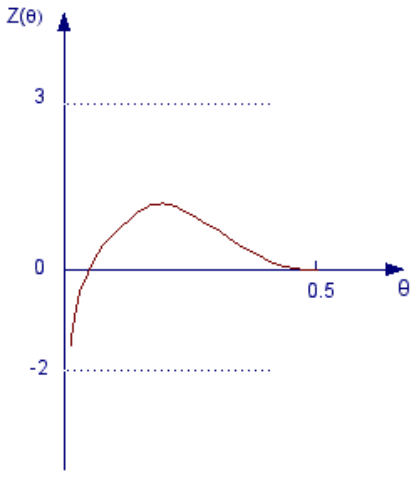

Figure 10

The proposed test has the advantage of being very simple, and of providing protection against falsely concluding linkage. However, some criticisms can be expressed, not only against the criteria chosen, but also against the entire principle of using a sequential procedure. The number of families typed is, indeed, rarely chosen in the light of the test results.

IV ESTIMATION OF THE RECOMBINATION FRACTION

If the test, on a sample of the family, has demonstrated linkage between the A and B loci, then one may want to estimate the recombination fraction for these loci.

The estimated value of θ is the value which maximizes the function of the lod score Z, and this is equivalent to taking the value of θ for which the probability of observing linkage in the sample is greatest.

V RECOMBINATION FRACTION FOR A DISEASE LOCUS AND A MARKER LOCUS

Let us assume we are dealing with a disease carried by a single gene, determined by an allele, g_0, located at a locus G (g_0 : harmful allele, G_0 : normal allele).

We would like to be able to situate locus G relative to a marker locus T, which is known to occupy a given locus on the genome. To do this, we can use families with one or several individuals affected and in which the genotype of each member of the family is known with regard to the marker T.

In order to be able to use the lod scores method described above, what is needed

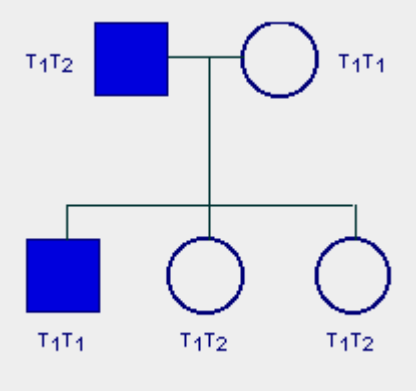

Figure 11

is to be able to extrapolate from the phenotype of the individuals (affected, not affected) to their genotype at locus G (or their genotypic probability at locus G). What we need to know is:

1. the frequency, g_0
2. the penetration vector f_1, f_2, f_3

f_1 = proba (affected /g_0g_0)

f_2 = proba (affected /g_0G_0)

f_3 = proba (affected /G_0G_0)

It will often happen that the information available for the marker is not also genotypic, but phenotypic in nature. Once again, all possible genotypes must be considered.

As a general rule, the information available about a family concerns the phenotype. To calculate the likelihood of θ, we must consider all the possible genotype configurations at each of the loci, for this family, writing the likelihood of θ for each configuration, weighting it by the probability of this configuration, and knowing the phenotypes of individuals in A and B.

Knowledge of the genetic parameters at each of the loci (gene frequency, penetration values) is therefore necessary before we can estimate θ.

It is obvious that calculating the lod scores, despite being simple in theory, is in fact a lengthy and tedious business; specific software have been designed for linkage analysis.

Analysis of gene linkage has made it possible to construct a gene map by locating the new polymorphisms relative to one other on the genome. The measurement used on the gene map is not the recombination fraction, which is not an additive unit of measurement, but the gene distance.