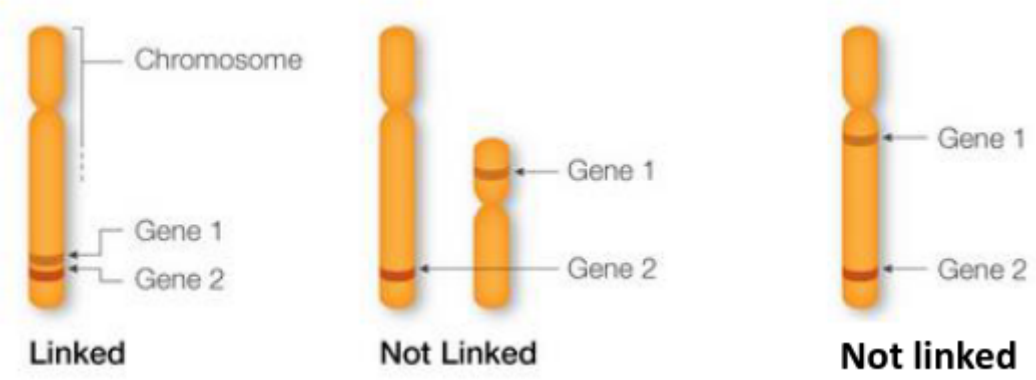
Linkage Analysis

The primary goal of linkage analysis is to determine the location (chromosome and region) of genes influencing a specific trait. We accomplish this by looking for evidence of co-inheritance of the trait with other genes or markers whose locations are known, and locating genes close to one another.



- Genes are markers that sit close together on a chromosome are called "linked" and are likely to be inherited together.
- Genes on separate chromosomes are never linked
- Genes that are far away from each other on a chromosome are likely to be separated during homologous recombination and are considered not linked

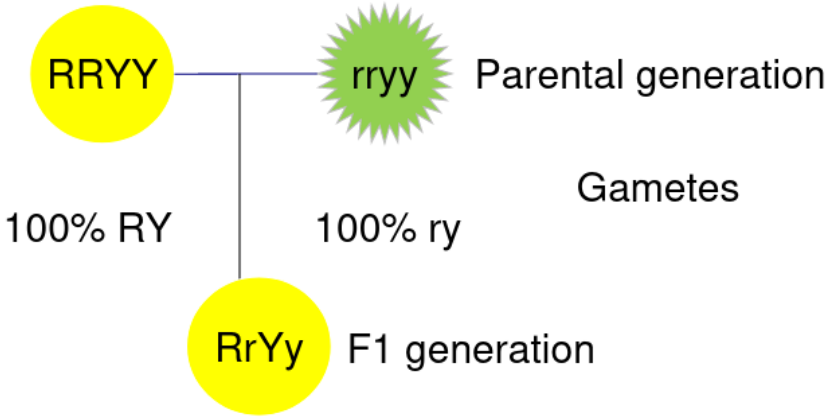
Linked Genes

Recall Mendel's 2nd Law: The Principle of Independent Assortment - Alleles of a gene pair assort independently of other gene pairs. The segregation of one pair of alleles in no way alters the segregation of another pair of alleles **EXCEPT when the genes are linked on a chromosome**.

A haploid genotype (haplotype) is a combination of alleles at multiple loci that are transmitted together on the same chromosome. It contains:

1. The alleles present at each locus (multi-locus genotype) such as Aa Bb
2. Which alleles are on the same chromosome, such as possible haplotypes Ab and aB or ab and AB

When traits are linked it means they are always inherited together. Consider the example:



The above represents 2 different phenotytpes on a pea plant, Y is color and R is whether its round or wrinkled.

If the color and shape are **unlinked** then we would consider each box of a punnet square to have equal likelihood:

	RY	Ry	rY	ry
RY	RRYY	RRYy	RrYY	RrYy
Ry	RRYy	RRyy	RrYy	Rryy
rY	RrYY	RrYy	rrYY	rrYy

How many phenotypes in F2?

- 9/16 Round and Yellow
- 3/16 Round and green
- 3/16 wrinkled and Yellow
- 1/16 wrinkled and green

If the color and shape are **completely linked** then all Haploid gametes are RY or ry, that is to say R is always inherited with Y and r is always inherited with y:

	RY	Ry	rY	ry
RY	RRYY			RrYy
Ry				
rY				
ry	RrYy			rryy

How many phenotypes in F2?

- 3/4 Round and Yellow
- 0 Round and green
- 0 wrinkled and Yellow
- 1/4 wrinkled and green

Genes can also be **partially linked** so the likelihood of them being paired together more likely but not guaranteed.

	RY	Ry	rY	ry
RY	RRYY	RRYy	RrYY	RrYy
Ry	RRYy	RRyy	RrYy	Rryy
rY	RrYY	RrYy	rrYY	rrYy
ry	RrYy	Rryy	rrYy	rryy

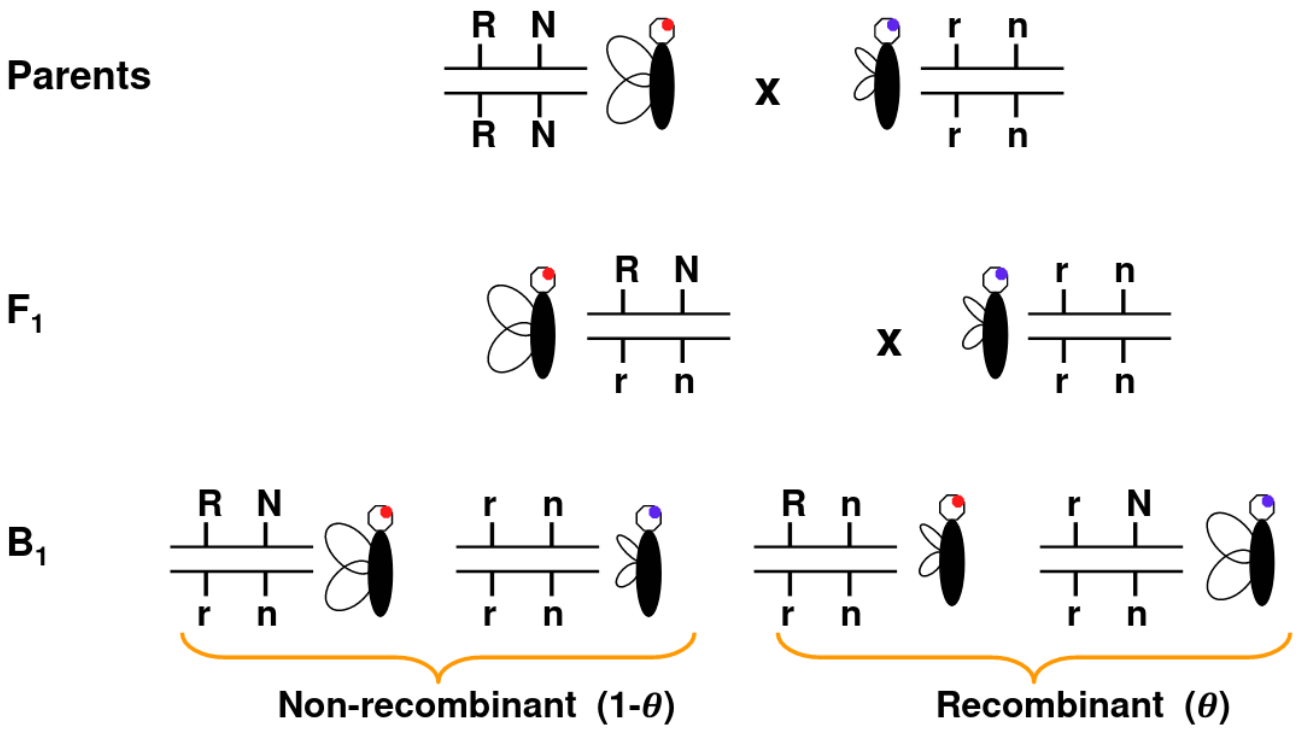
How many phenotypes in F2?

- <3/4 Round and Yellow
- >0 Round and green
- >0 wrinkled and Yellow
- <1/4 wrinkled and green

Recombination

Recombinant don't inherit the same chromosomes as their parents, while non-recombinant have chromosomes inherited by their parents. Proportion of recombinants gives an indication of the distance along the chromosome between the two loci; the closer they are together the less likely they are to combine.

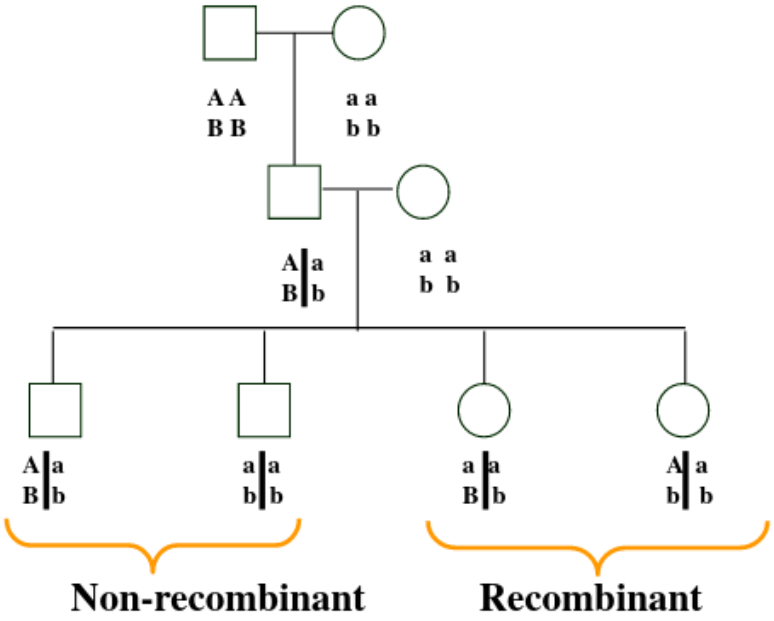
Consider Morgan's fruit fly experiment as an example:



Genotype	Observed	Expected	
Rr Nn	1339	710	nonrecombinant
rr nn	1195	710	nonrecombinant
Rr nn	151	710	recombinant
rr Nn	154	710	recombinant

Wing length and eye color are linked. So RN and rn should always be inherited together, but we can see the recombination Rn or rN in B1 that doesn't occur in parents.

When loci are on different chromosomes, we observe recombination 50% of the time.



The line in-between the allele means they are on the same chromosome, so the alleles on the same side are inherited together.

Overview of Recombination

The recombination fraction (theta) is the probability of recombination (i.e. the probability of an odd number of crossovers) between two loci. The further apart they are, the more likely they are to recombine and vice versa.

> □□ is related to how closely "linked" the two loci are, and ranges from 0 to .5

- □□ = 0 if:
  - The genetic marker is the polymorphism causing the disease
  - The marker is so close to the disease mutation that recombination can never occur
- □□ = .5 if:

    - There is a 50% chance that alleles at the two loci are inherited together. This happens when the two loci are
        - Very far apart on a chromosome
        - Located on two different chromosome

Linkage Analysis

Linkage analysis estimates the genetic distance between genetic markers or between genetic markers and a trait locus with recombination. This allows us to know where a trait locus is in the genome if we know:

1. The genetic distance between a disease locus and a marker
2. The location of the marker on the genome.

To perform a linkage analysis, estimate the recombination fraction between the disease locus and a marker locus and test:

H0: □□ = 1/2 -> not linked i.e segregating independently

H1: □□ < 1/2 -> linked, i.e. close together on the same chromosome

The most common test of linkage is **Logarithm of the Odds (LOD)**:

> LOD Score = log base 10 of the likelihood ratio = $\log_{10}(L(□□=□□_1)/L(□□=.5))$

This is a transformation of the usual likelihood ratio test. The probability of observing R recombinants and N non-recombinants, where the recombination fraction is □□, the binomial is as follows:

$$\binom{N+R}{R}\theta^R(1-\theta)^N$$

So the likelihood is:

$$L(\theta|R,N) \propto \theta^R(1-\theta)^N$$

We can ignore the constant because we will be working with a ratio of likelihoods and the constant cancels.

When we know the exact number of recombinants and non-recombinants:

$$LOD(\theta_1) = \log_{10}\left(\frac{L(\theta = \theta_1)}{L(\theta = 0.5)}\right) = \log_{10}\left(\frac{\theta_1^R(1-\theta_1)^N}{0.5^{R+N}}\right)$$

The MLE of □□ is □□_hat = R/(R+N) which is the value of □□ that maximizes the LOD score and likelihood function.

- LOD score > 3 -> linkage
- LOD score (for a particular q value) < -2 -> no linkage
- When the LOD scores is between 3 and -2, results are inconclusive. We might want to obtain additional individuals, additional families or utilize additional markers

Factors Influencing Linkage Analysis

- Penetrance - the probability of expressing the disease given a specific genotype.
  - Age dependent penetrance is also common in some diseases, such as Huntingtons.
  - Ex. reduced penetrance could be P(Disease | DD or Dd) < 1.0 for Autosomal Dominant
- Genetic Heterogeneity - multiple genes which mutations cause the same phenotype
  - When heterogenetiy exists LOD score over families may not show evidence of linkage, but significant linkage may occur within a subset
  - Ex. there are 7 genes identified for fmilial Parkingson's disease; some dominant, some recessive.

Overview of Linkage Analysis

Parametric linkage analysis assesses linkage between a marker and a locus

- Need to specify a model for the inheritance of the disease
- Need to specify risk allele frequency
- Need to specify penetrances

Advantages

- Most powerful approach when the model is correctly specified
- It utilizes every family member's phenotypic and genotypic information
- It provides a statistical test for linkage and for genetic heterogeneity

Disadvantages

- Poor power if the genetic model is misspecified
- Unaffected individuals may provide little information if penetrance is low
- Can be difficult to recruit large families

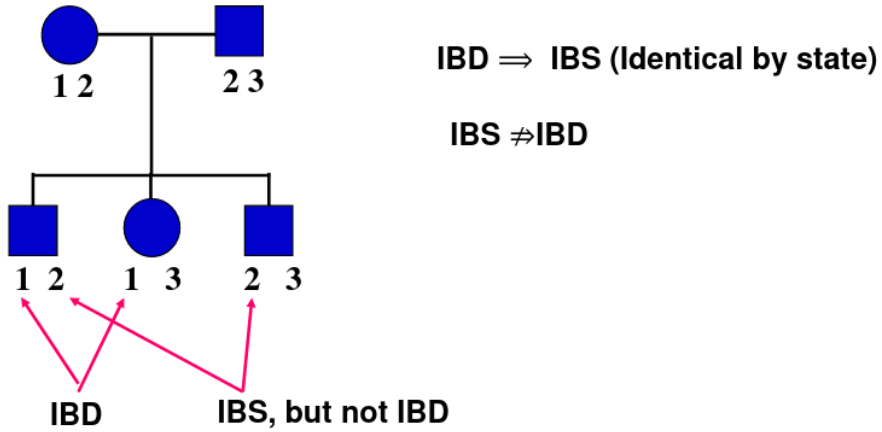
Multi-point Linkage Analysis

Multi-point analysis incorporates multiple markers into the likelihood computation. It computes the likelihood that a disease is located at a certain position on a chromosome. The null hypothesis is that the disease locus is not on the chromosome. It has more information and thus is more powerful.

Affected Relative Pairs

Based on sets of affected pairs, no need to specify a model. This is a non-parametric tests of linkage analysis. Easier to collect but need a larger sample.

Alleles that are copies of the same allele from a common ancestor are called "identical-by-descent" or IBD for short. IBS is "Identity-by-state". All IBD are IBS but the opposite is not true.



For the bottom rightmost "3" in the example definitely comes from the father, so it is shared IBD.

The logic behind this test is that if an allele is "causing" a disease, and the disease is not common, two affected relatives will most probably share the allele IBD. This means the surrounding region will most probably be IBD, and we can look for regions which are more likely to hold the disease locus.

Consider the below chart with the expected frequency of relative pairs showing sharing 0,1, or 2 alleles IBD under no linkage:

	Alleles shared IBD		
Relative Pair	**0**	**1**	**2**
Monozygotic twins	0	0	1.00
Siblings	0.25	0.50	0.25
Parent-child	0	1.00	0
Grandparent-Grandchild	0.50	0.50	0
Half-siblings	0.50	0.50	0
Avuncular	0.50	0.50	0
Cousins	0.75	0.25	0
Double first cousins	0.56	0.38	0.06

Twins always share their alleles. Children always inherit 1 allele IBD with their parents, making the relationship useless for linkage analysis. Siblings can give the most information.

The process of linkage analysis through affected relative pairs is as follows:

1. Collect a sample of pairs of affected relatives
2. Genotype some markers and estimate IBD
3. At multiple locations across the genome test whether the pairs share more alleles IBD than would be expected by chance.

Tests for affected sibling pairs:

- Goodness of Fit
- Mean Sharing Test (not covered here)
- Nonparamtric Linkage Analysis Score Test (NPL), best for affected general relatives (not covered here)

Goodness of Fit

n_i = number of sibling pairs sharing i alleles IBD

N = total number of sibling pairs = $n_0 + n_1 + n_2$

IBD	Observed	Expected
0	n_0	$N/4$
1	n_1	$N/2$
2	n_2	$N/4$

Perform χ^2 goodness of fit test (or a Likelihood ratio test):

$$\chi^2 = \frac{4(n_0 - N/4)^2 + 2(n_1 - N/2)^2 + 4(n_2 - N/4)^2}{N}$$

Goodness of fit statistic follows a χ^2 distribution with **2** degrees of freedom (number of categories -1)

Linkage Measured By LOD Score

LOD Score = log_10(Likelihood Ratio)

Likelihood Ratio Test (LRT) = 2 * ln(Likelihood Ratio) ~ chi^2

Therefor: chi^2 = 2 * ln(10) * log_10(likelihood ratio) = 4.6 * LOD Score

> LOD Score = chi^2 test value / 4.6, with 1 df

LOD score ~ z^2 / 4.6 or t^2 / 4.6

Quantitative Trait Locus Mapping

Relatives who have similar trait values should have higher than expected levels of sharing genetic material near the genes that influence those traits (a greater IBD).

The Haseman-Elston (HE) was the first approach to QTL in humans.

- Let Y_1j and Y_2j be the phenotypes of siblings 1 and 2 of the jth subpair
- Y_j^D = (Y_1j - Y_2j)^2 -> the difference in sib phenotypes
- pi_j the proportion of alleles shared IBD by the jth sibpair at the marker of interest. Could be 0, .5, or 1 when IBD is known with certainty. Otherwise it is a range from 0 to 1.

- To test for linkage perform the linear regression of:

$$Y_j^D \, on \, \pi_j: \quad E\left[Y_j^D \middle| \pi_j\right] = \alpha + \beta\pi_j$$

It can be proven that:

$$\beta = -2(1-2\theta)^2\sigma_a^2 \qquad \text{and}$$
$$\alpha = \sigma_e^2 + 2(\theta^2 + (1-\theta)^2)\sigma_a^2$$

Where sigma_squared_a is the genetic variance explained by the locus and sigma_squared_e is the envoirnmental variance of the trait.

Under the null hypothesis Ho: $\theta = 0.5$

$$\beta = -2(1-2\theta)^2\sigma_a^2 = 0$$

Under the alternative hypothesis that the marker and trait are linked, $0 \leq \theta < 0.5$ and therefore

$$\beta = -2(1-2\theta)^2\sigma_a^2 < 0$$

A negative slope implies linkage, as relatives with similar trait values have small squared differences and high IBD sharing

---

Revision #9
Created 19 September 2022 22:05:40 by Elkip
Updated 22 September 2022 16:33:47 by Elkip