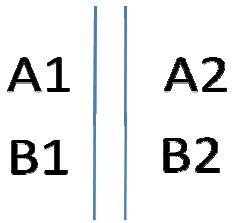
Linkage Introduction

Law of Independent Assortment

- Mendel's Second Law (Law of Independent Assortment) :
  - The segregation of the genes for one trait is independent of the segregation of genes for another trait, i.e., when genes segregate, they do so independently
- This law essentially states that during gamete formation, the segregation of one gene is independent of the other gene
- This "law" is frequently violated and is only true for loci/genes that are unlinked.

Recombination

- When a gamete is passed down, the chromosome inherited by an offspring from a parent is actually a mosaic of the parent's two chromosomes.
- Suppose we have two loci on the same chromosome, locus 1 and locus 2, where locus 1 has alleles A1 and A2, and locus 2 has alleles B1 and B2.
- In the example below, **phase** is known and is (A1,B1) and (A2,B2).
- If the genes are closely linked, a gamete is much more likely to contain (A1,B1) or (A2,B2), which are "non-recombinants."
- If there is recombination, a gamete will contain (A1, B2) or (A2,B1), but this is less likely if the loci are linked.



Recombination Fraction

- Two loci that are unlinked follow Mendel's Second Law, and all possible gametes for a parent are produced with equal frequency.
- When loci are physically located close to one another on a chromosome, there is a deviation from this relationship. This deviation is summarized by the recombination fraction.
- The recombination fraction is often denoted by θ where $0 \leqslant \theta \leqslant \frac{1}{2}$
- P(recombinant gamete)$= \theta$
- If $\theta < \frac{1}{2}$, the loci are said to be linked or in genetic linkage
- When loci are completely linked, $\theta = 0$
- Two loci are said to be unlinked if $\theta = \frac{1}{2}$.
- Note that if two loci are on different chromosome, then $\theta = \frac{1}{2}$.

Linkage in a simple genetic cross

- In the early 1900's, Bateson and Punnet conducted genetic studies using sweet peas. They studied two characters:
  - Petal color which has two alleles: P (purple) and p (red), where P is dominant.
  - Pollen grain shape has two alleles: L (elongated) and l (disc-shaped), where L is dominant

$$PPLL \times ppll$$
$$\downarrow$$
$$PpLl \qquad\qquad F1$$

- Plants in the F1 generation were intercrossed: PpLl X PpLl.
- According to Mendel's Second Law, during gamete formation, the segregation of one gene pair is independent of another gene pair.

Sweet Peas Linkage Example

F2	PL	Pl	pL	pl
PL	Purple/Long	Purple/Long	Purple/Long	Purple/Long
Pl	Purple/Long	Purple/Disc	Purple/Long	Purple/Disc
pL	Purple/Long	Purple/Long	Red/Long	Red/Long
pl	Purple/Long	Purple/Disc	Red/Long	Red/Disc

Sweet Peas Linkage Example

- The expected relative frequencies in the F2 generation if the genes segregated independently are

	Elongated	Disc-Shaped
Purple	9	3
Red	3	1

- The observed frequencies in 381 plants in the F2 generation where

	Elongated	Disc-Shaped
Purple	284	21
Red	21	55

- The observed data clearly do not fit what is expected under the model.
- The explanation: the petal color gene and the gene for pollen grain shape are linked.
- Let θ be the recombination fraction between the two genes. What is the probability of each possible plant type?

Sweet Peas Linkage Example

		$\frac{1}{2}(1-\theta)$	$\frac{1}{2}\theta$	$\frac{1}{2}\theta$	$\frac{1}{2}(1-\theta)$
		PL	Pl	pL	pl
$\frac{1}{2}(1-\theta)$	PL	Purple/Long	Purple/Long	Purple/Long	Purple/Long
$\frac{1}{2}\theta$	Pl	Purple/Long	Purple/Disc	Purple/Long	Purple/Disc
$\frac{1}{2}\theta$	pL	Purple/Long	Purple/Long	Red/Long	Red/Long
$\frac{1}{2}(1-\theta)$	pl	Purple/Long	Purple/Disc	Red/Long	Red/Disc

- P(red, disc-shaped)= $\frac{1}{4}(1-\theta)^2$
- P(red,elongated)= $\left(\frac{1}{2}\theta\right)\left(\frac{1}{2}\theta\right)+\left(\frac{1}{2}\theta\right)\left(\frac{1}{2}(1-\theta)\right)+\left(\frac{1}{2}(1-\theta)\right)\left(\frac{1}{2}(\theta)\right)$
- P(purple, disc-shaped) and P(purple, elongated) are calculated similarly.
- We can form a likelihood for the data that is a function of the recombination fraction θ. We can find the value of θ that maximizes this likelihood.
- Likelihood will follow a multinomial distribution.

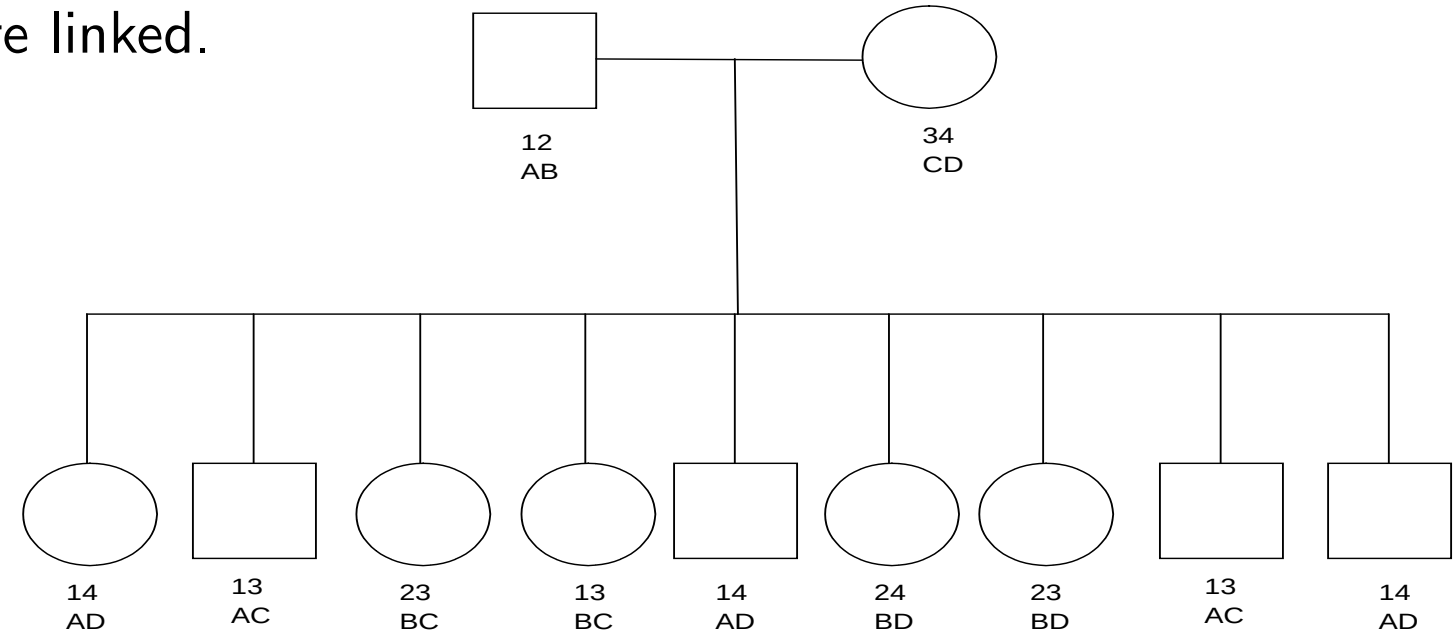
Parametric Linkage Analysis

Linkage Analysis

- Once aggregation and/or segregation studies established a genetic component for a phenotype of interest, parametric linkage analysis was the traditional approach used for Mendelian disease gene mapping since the 1970's
- Linkage analysis requires genetic marker data on pedigree.
- To illustrate linkage analysis, we will consider examples given by Suarez, B.K. and Cox, N.J. (1985)

Nuclear Family Example

- The figure below shows a large nuclear family segregating alleles from two loci: alleles at one of the loci are denoted by numbers while the alleles of the other are denoted by letters.
- Both of the parents are heterozygous at each locus and share no alleles in common, so the co-segregation of the alleles at the two loci can be unambiguously followed.
- We are interested in determining whether or not the two loci are linked.



Lod Scores

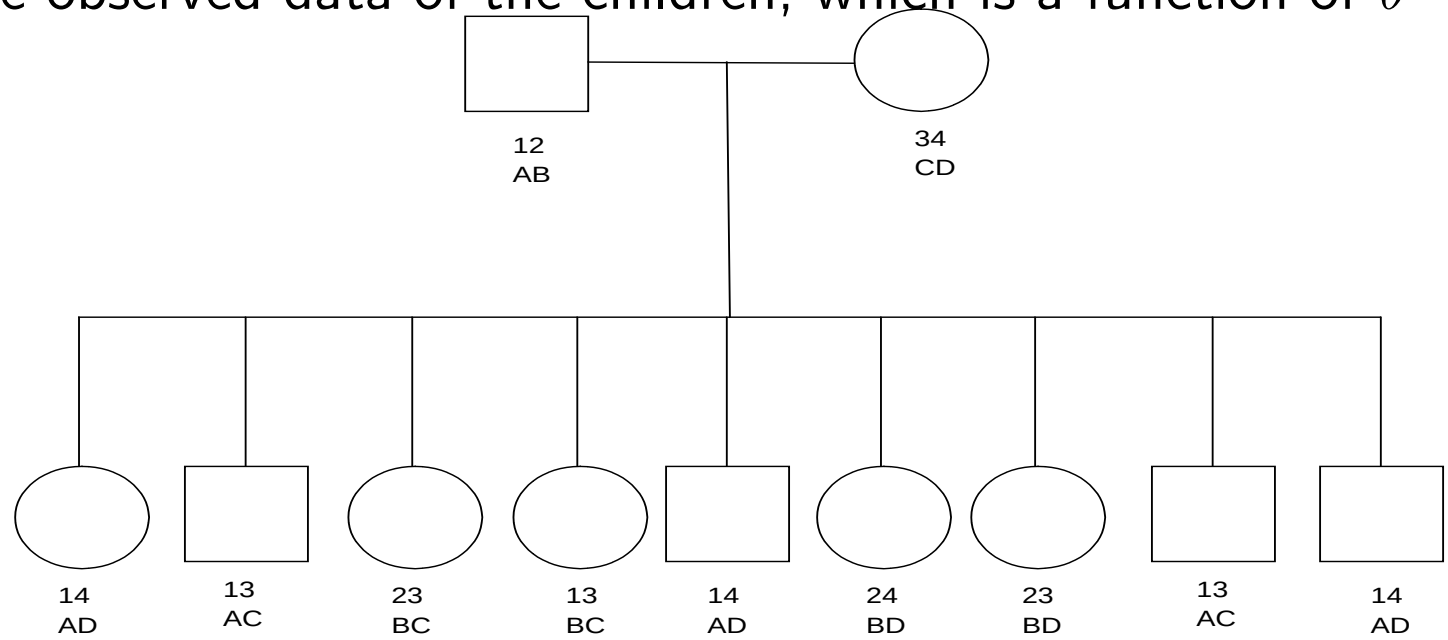
- **LOD scores** are calculated for recombination fraction θ values to determine if there is significant evidence for linkage
- For a given value of θ, the lod score is

$$log_{10} \frac{P(\text{observed data assuming recombination fraction is } \theta)}{P(\text{observed data assuming recombination fraction is } .5)}$$

- **LOD** stands for **L**og of **OD**ds
- Find the the value of θ that gives the maximum lod score
- Lod scores greater than 3 give evidence of linkage, and the null hypothesis of no linkage is rejected.
- How do you interpret a lod score equal to 3?
- Lod scores less than -2 give evidence that the loci are unlinked.

Nuclear Family Linkage Example

- Can calculate a lod score for the large nuclear family. We only observe the genotypes at the two loci so the phase is unknown. Possible phase for the parents:
  - 1A 2B 3C 4D
  - 1A 2B 3D 4C
  - 1B 2A 3C 4D
  - 1B 2A 3D 4C
- Given each parental phase type, can obtain the probability of the observed data of the children, which is a function of θ



Nuclear Family Linkage Example

Phase	1A,2B,3C,4D	1A,2B,3D,4C	1B,2A,3C,4D	1B,2A,3D,4C
Phase Probability	.25	.25	.25	.25
Offspring Probability	$\left(\frac{1}{2}\right)^{18}(1-\theta)^{16}\theta^2$	$\left(\frac{1}{2}\right)^{18}(1-\theta)^{9}\theta^9$	$\left(\frac{1}{2}\right)^{18}(1-\theta)^{9}\theta^9$	$\left(\frac{1}{2}\right)^{18}(1-\theta)^{2}\theta^{16}$

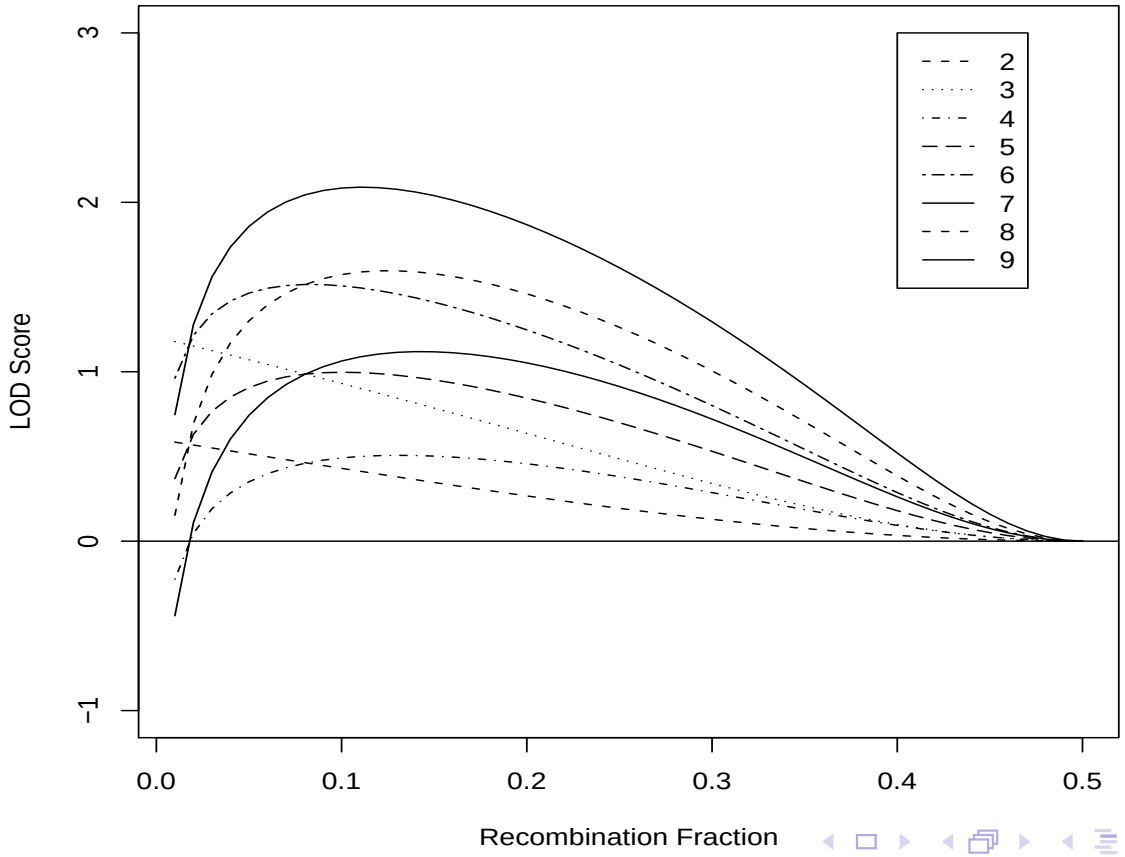
So, for $\theta = .1$ the lod score is

$$\frac{.25(.9)^{16}(.1)^2 + .5(.9)^9(.1)^9 + .25(.9)^2(.1)^{16}}{.25(.5)^{18} + (.5)^{19} + .25(.5)^{18}}$$

$$= 2.08$$

Nuclear Family Linkage Example LOD Score Graph

- For linkage analysis with nuclear families, data must be available on at least 2 offspring
- The figure below gives the lod score curves obtained for the large nuclear family according to the number of children included in the calculation.



Nuclear Family Linkage Example

- The previous figure illustrates how the lod score curve changes as more information becomes available
- The lod score is always 0 at $\theta = \frac{1}{2}$ since the odds ratio is 1
- The lod score calculated using the first 2 children and using the first 3 children steadily increases as $\theta \longrightarrow 0$
- With the addition of the fourth child, the lod score curve changes from its monotonically increasing from as $\theta \longrightarrow 0$ to one that increases as θ moves away from $\frac{1}{2}$
- Evidence for linkage becomes a little stronger with the addition of the fifth and sixth children, and decreases with the seventh child (due to an apparent maternal recombinant), and then increases with the remaining 2 children

Nuclear Family Linkage Example

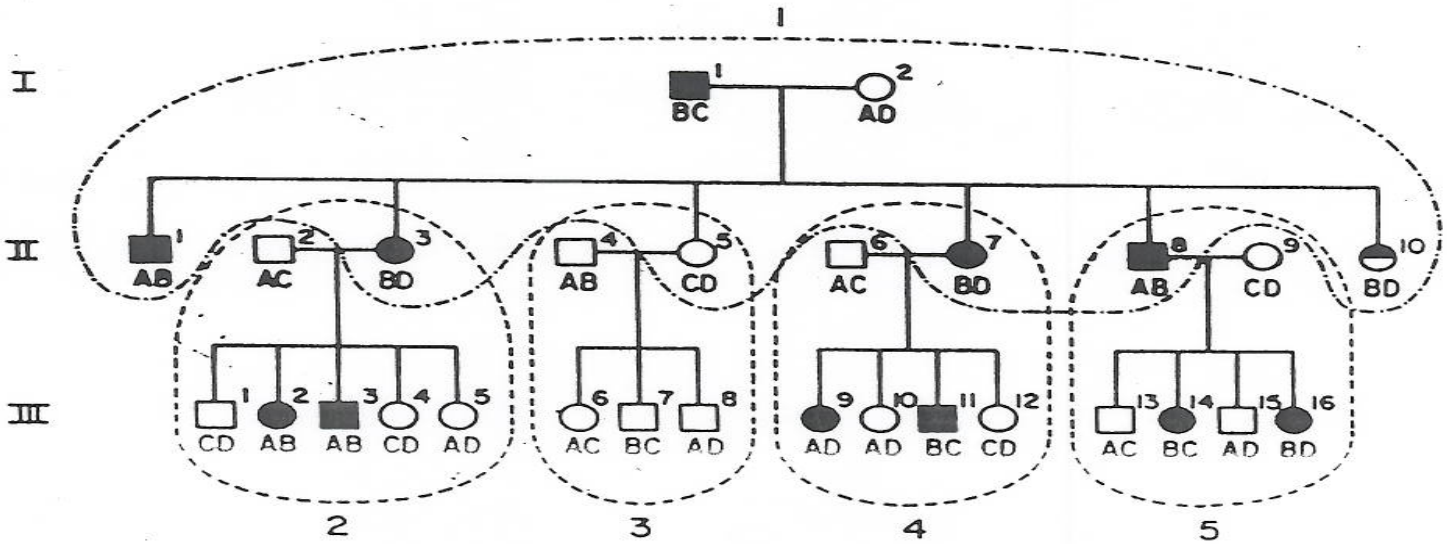
- This nuclear family provides moderate evidence that the 2 loci are linked. At $\hat{\theta} = 0.11$, the lod score curve reaches its maximum value of 2.09, indicating that the hypothesis of linkage with 11% recombination is about 123 times more likely than the hypothesis of no linkage
- Since the maximum lod score is in the range of -2 to 3, more families need to be sampled before a decision of $\theta = \frac{1}{2}$ or $\theta < \frac{1}{2}$ can be accepted or rejected.

Extended Family Linkage Example

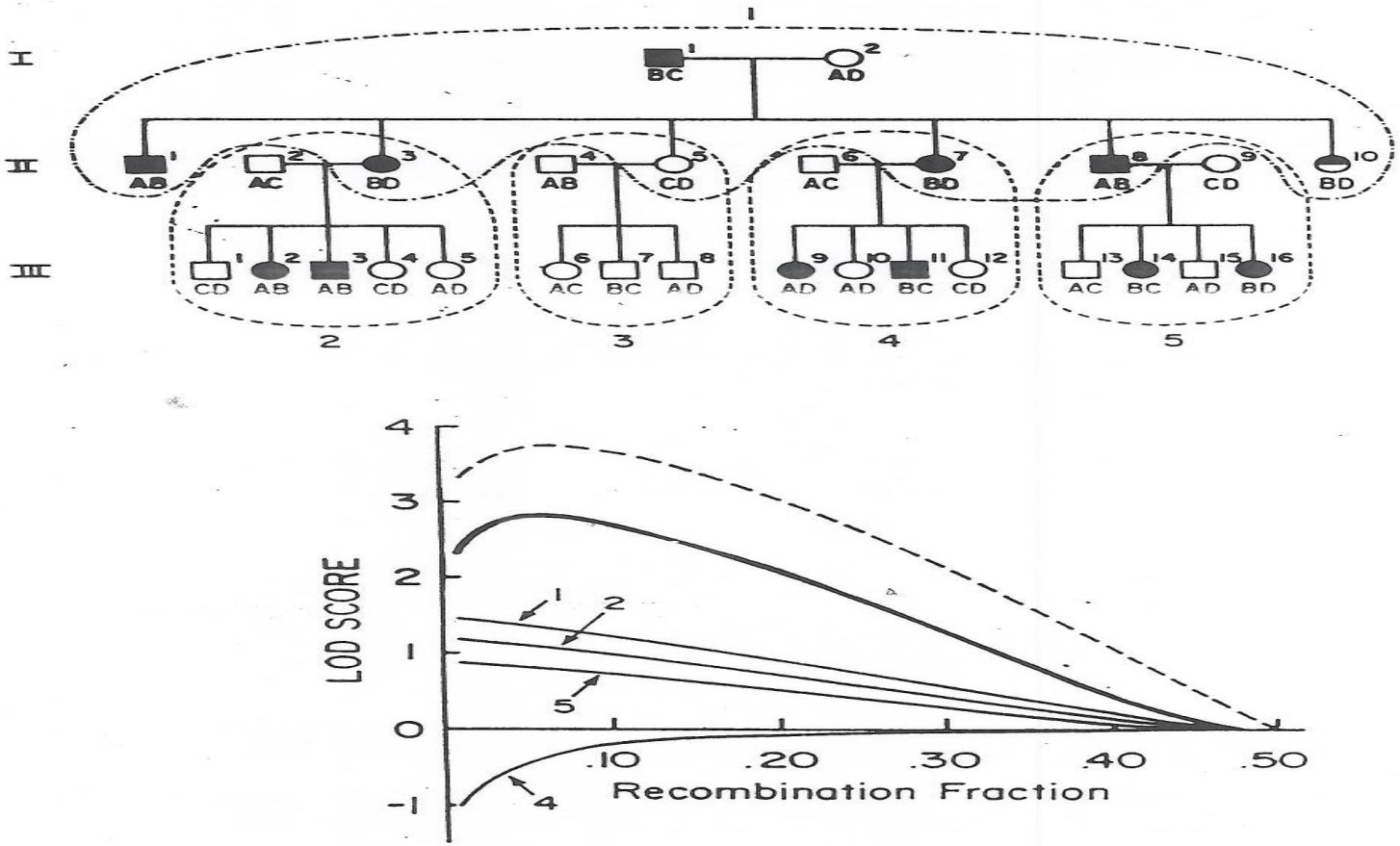
- Linkage analysis for co-dominant loci in straightforward and a decision in favor or against the hypothesis of linkage can usually be reached with a few informative families.
- In general, however, nuclear families are less efficient than extended 3-generation pedigrees because extended pedigrees provide more information regarding phase

Extended Family Linkage Example

- Consider the 28 member 3-generation pedigree below
- We would like to determine if the locus with available genotype data is linked to a disease locus for which we do not know the location.
- What are the possible genotypes for the individuals in the pedigree if the disease is caused by a single locus that is fully penetrant and dominant ?



Extended Family Linkage Example

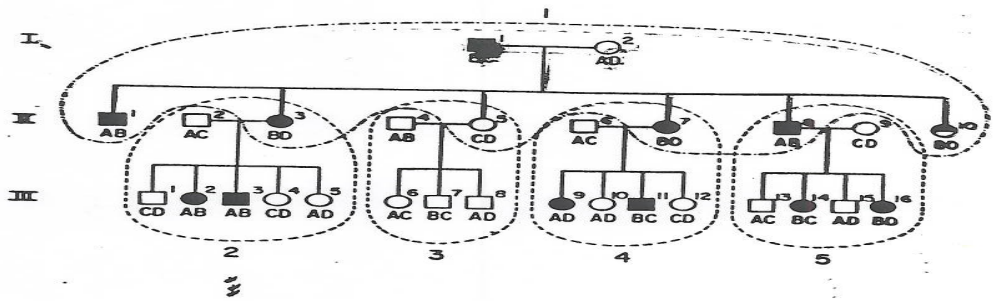


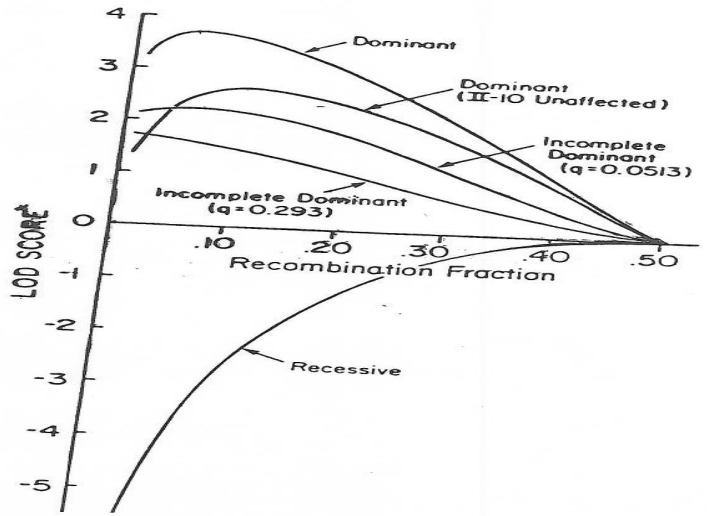
Extended Family Linkage Example

- Nuclear families 1, 2, and 5 provide evidence for linkage. The lod score curves monotonically increasing as $\theta \longrightarrow 0$ suggest that these families do not contain any recombinants. The different height of the lod score curves reflects that fact that larger nuclear families are more informative than smaller ones.
- Nuclear family 3 provides no information regarding linkage since neither partent is affected and at least one parent must be a double heterozygote to be informative.
- Nuclear family 4 provides slight evidence against the hypothesis of linkage.
- If the nuclear families were truly independent, then the lod scores could be summed, giving a maximum lod score of 2.81 at $\hat{\theta} = 0.05$.
- When analyzing the pedigree as a whole, the maximum is also at $\hat{\theta} = 0.05$ but with a lod score of 3.72.

Extended Family Linkage Example

- The plot below illustrates that misspecification of the mode of transmission of the disease affects the linkage analysis results.





For incompl dominant
$f_{AA} = 0$
$f_{Aa} = f_{aa} = 1/2$

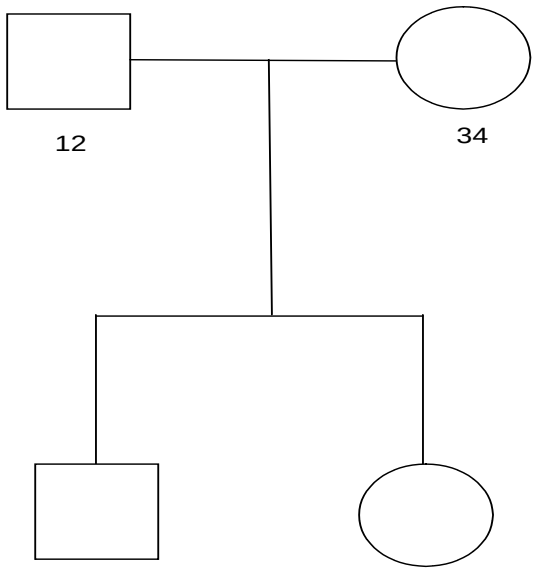
Nonparametric Linkage Analysis

Limitations of Parametric Linkage Analysis

- We previously discussed parametric linkage analysis
- Genetic model for the disease must be specified: allele frequency parameters and penetrance parameters
- Lod scores results are highly sensitive to the assumed mode of transmission of the disease, which will generally be uknown
- Nonparametric linkage analysis methods does not make any assumptions about the disease model

Sib Pair IBD Sharing Distribution

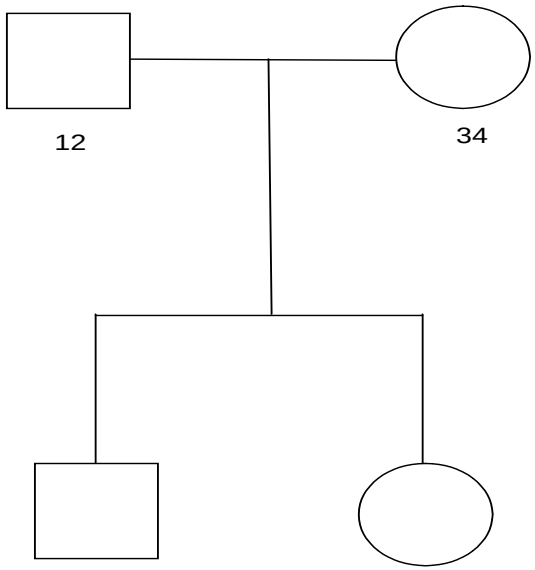
- Consider the nuclear family below with 2 siblings segregating alleles for a locus
- What is the probability of the siblings sharing 2, 1, or 0 alleles identical by descent (IBD)?



Sib Pair IBD Sharing Distribution

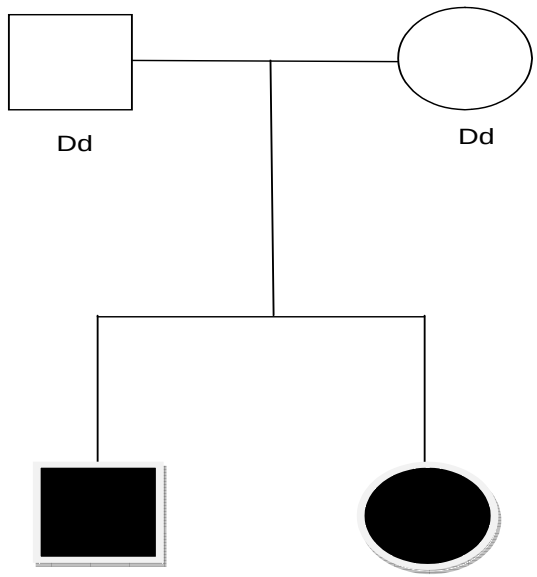
- Expected IBD Sharing

2	:	1	:	0
0.25	:	0.5	:	0.25



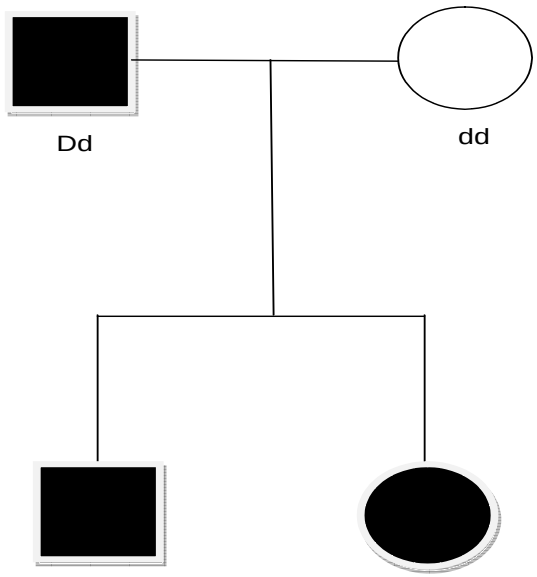
Affected Sib Pair Example

- Now consider a disease that is caused by a single locus.
- What would the allele sharing probabilities be for a sib pair at the disease locus?
- This depends on the mode of transmission of the disease. Assume for now that disease is caused by the D allele and D is recessive.



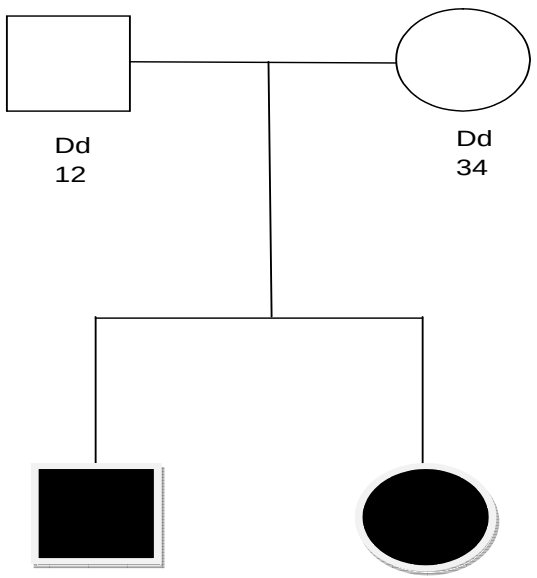
Affected Sib Pair Example

- Now assume that disease is caused by the D allele, and D is dominant.
- What would the allele sharing probabilities be for a sib pair at the disease locus?

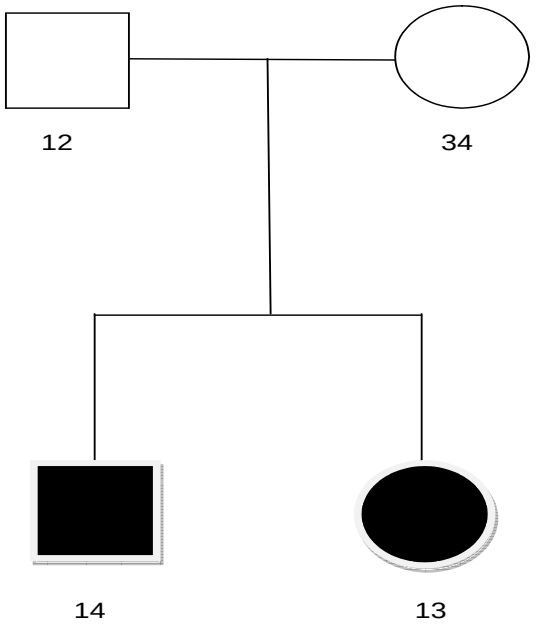


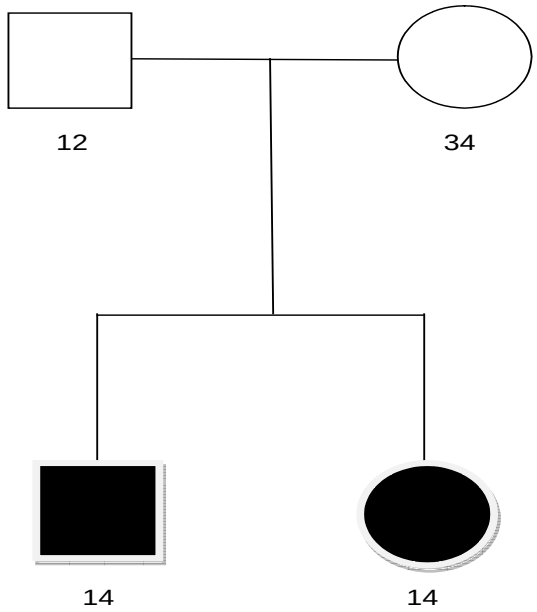
Affected Sib Pair Example

- The location of the disease gene is unknown and we would like to determine if the locus is linked to the disease gene.
- If the locus is linked to the disease gene, then the expected IBD probabilities of sharing 2, 1, and 0 alleles IBD for sibs at the disease gene will not be .25, .5, and .25, respectively, regardless of the mode of inheritance of the disease.



Affected Sib Pair Example





Affected Sib Pair Example

- The null hypothesis: locus is transmitted independently of the disease locus D/d.
- Under the null, the expected IBD sharing for sibs is

2	:	1	:	0
0.25	:	0.5	:	0.25

- Under the alternative, the locus is linked to the disease locus, and as a result, the IBD sharing probabilities do not follow the distribution specified under the null hypothesis.
- If the null is false, then you should see an increase in affected sibs sharing either 1 or 2 alleles IBD.
- For example if disease is caused by a rare dominant allele and the locus is tightly linked to the disease gene, then expected IBD sharing for sibs might be around

2	:	1	:	0
0.5	:	0.5	:	0

Affected Sib Pair Example

- More realistic scenario: marker is very close to locus which influences risk of disease in a more subtle manner (heterogeneity, epistasis, gene-environment interaction)

2	:	1	:	0
0.35	:	0.45	:	0.2

Model-Free Linkage Test

- The Pearson chi-squared goodness of fit test is a sample way of comparing the observed counts of sib pairs sharing 0, 1 and 2 alleles IBD with that expected under the null of no linkage.
- Let N be the number of affected sib pairs.
- Let n_i be the number of sib pairs that share i alleles IBD, where $i = 0$, 1, or 2.
- Under the null, what is the expected value of n_i for each i?
- Let the expected value of n_i under the null be E_{n_i}. The test statistic is:

$$X^2 = \sum_{i=0}^{2} \frac{(n_i - E_{n_i})^2}{E_{n_i}}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with 2 degrees of freedom

Extended Pedigrees

- Nonparametric linkage analysis can also be used for extended pedigrees, not just nuclear families with affected sib pairs
- Can calculate the expected IBD sharing for more distant relatives
- What is the expected IBD sharing probabilities for first cousins under the null?

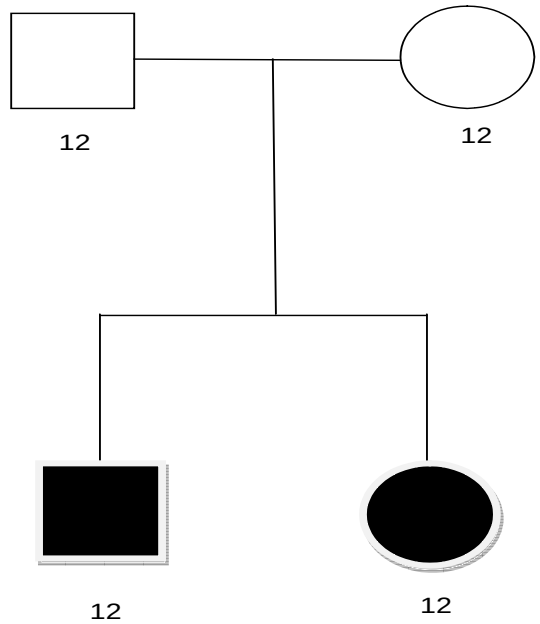
2	: 1	: 0
0	: 0.25	: 0.75

- What is the expected IBD sharing probabilities for second cousins under the null?

2	: 1	: 0
0	: 0.0625	: 0.9375

IBD Allele Sharing Uncertainty

- It may not be possible to determine exactly how many alleles a pair share IBD.
- In the example below, the affected sib pair could be sharing 2 or 0 alleles IBD, with each possibility having a probability of .5?



IBD Allele Sharing Uncertainty

- Methods to allow for this uncertainty developed, e.g., Kruglyak et al. (1996), Kong and Cox (1997).
- Multi-point method that incorporates the genotypes of nearby loci
- Obtain a probability distribution of IBD sharing at the locus being tested for linkage

Allele Sharing Statistics

- Allele sharing statistics S are often used for nonparametric linkage analysis. The general form of the statistics are

$$Z = \frac{S - \mu_0}{\sigma_0}$$

  where μ_0 and σ_0 are the expected value and variance of S, respectively, calculated under the null hypothesis. If a locus is not linked to a disease, Z will follow a standard Normal distribution.
- There are various types of allele sharing statistics
- S_{pairs} counts, for each pair of affected relatives, the number of alleles shared IBD, and then sums that counts over all pairs of affected relatives.
- If all affected individuals in a pedigree have a common ancestor in the pedigree, S_{all} is the number of alleles shared IBD by all affected relatives.

Allele Sharing Statistics

- S_{max} is the size of the largest group of related cases who all inherit the same allele IBD (high power for dominant disease alleles)
- McPeek (1999) showed that the optimal sharing statistic depends on the disease model