THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Section ON Statistical Genetics

Genetic Linkage Analysis

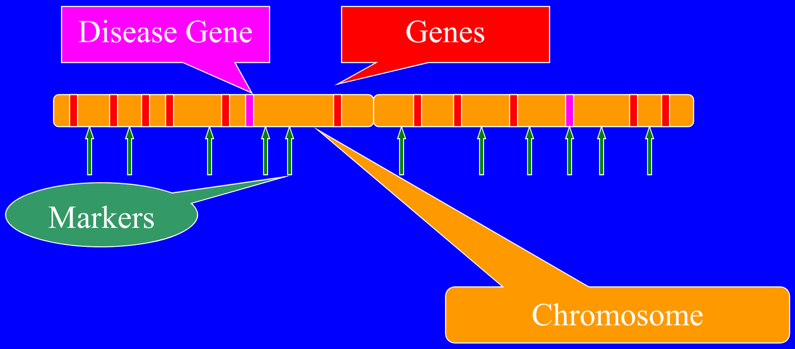
Hemant K Tiwari, Ph.D.

Section on Statistical Genetics

Department of Biostatistics

School of Public Health

How do we find genes implicated in disease or trait manifestation?



Markers: Microsatellites

- A microsatellite consists of a specific sequence of DNA bases or nucleotides which contains mono, di, tri, or tetra tandem repeats.

  For example,
  - AAAAAAAAAAA would be referred to as (A) 11
  - GTGTGTGTGTGT would be referred to as (GT) 6
  - CTGCTGCTGCTG would be referred to as (CTG) 4
  - ACTCACTCACTCACTC would be referred to as (ACTC) 4

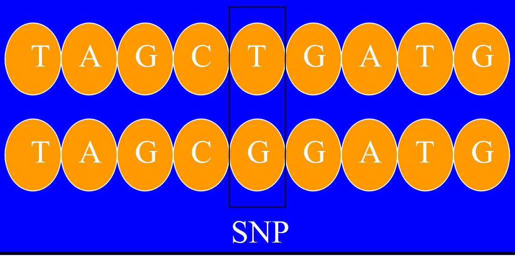
- In the literature they can also be called simple sequence repeats (SSR), short tandem repeats (STR), or variable number tandem repeats (VNTR). Alleles at a specific location (locus) can differ in the number of repeats. Microsastellites are inherited in a Mendelian fashion.

Single Nucleotide Polymorphisms (SNPs)

SNP: DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered.

On average, a SNP exists about every 100-300 base pairs

About 12 millions SNPs on a genome



HERITABILITY

Traits are familial if members of the same family share them, for whatever reason.

Traits are genetically heritable only if the similarity arises from shared alleles and genotypes.

To quantify the degree of heritability, one must distinguish between two sources of phenotypic variation:

Hereditary (i.e., Genetic) and environmental

Phenotype = heritable effect + environmental effect

V(phenotype) = V(hereditary) + V(envioronmental effect)

$$\sigma_p^2 = \sigma_h^2 + \sigma_r^2$$

HERITABILITY IN THE BROAD SENSE

- The proportion of the total phenotypic variance that is due to the hereditary variance =heritability = $h^2 = \dfrac{\sigma_h^2}{\sigma_h^2 + \sigma_r^2}$ (heritability in broad sense)

- Limitations of h^2 :
  1) not a fixed characteristic of a trait, but depends on the population in which it was measured and the set of environments in which that population grew;
  2) if genotype and environment interact to produce phenotype, no partition of variation can actually separate causes of variation;
  3) high h^2 does not mean that a trait cannot be affected by its environment.

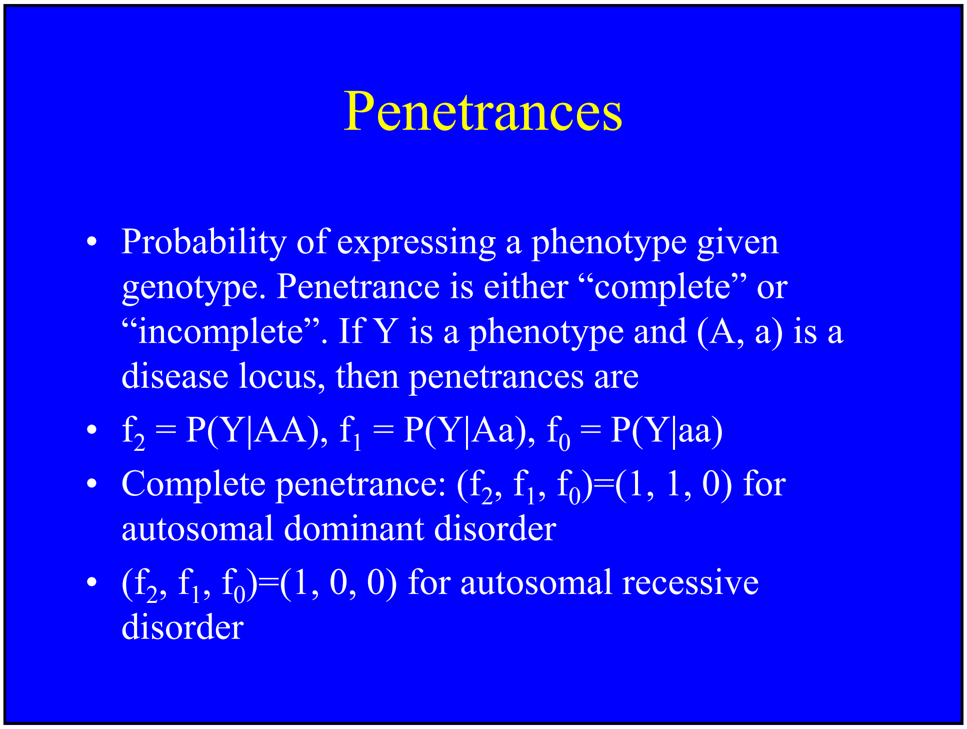
  Therefore h^2 has limited meaning and use in humans other than as a parameter to allow for familial correlations.

- **The genetic variance can be partitioned into the variance of additive genetic effects, of dominance (interactions between alleles at the same locus) genetic effects, and eppistatic (interactions between alleles at different loci) genetic effects.**
- **Thus, genetic variance can be broken into three parts: additive, dominance, and epistatic variances,**

  Thus we have $\boldsymbol{\sigma_h^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2}$
- $\boldsymbol{h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_d^2 + \sigma_i^2}}$ **and is called heritability in the narrow sense.**

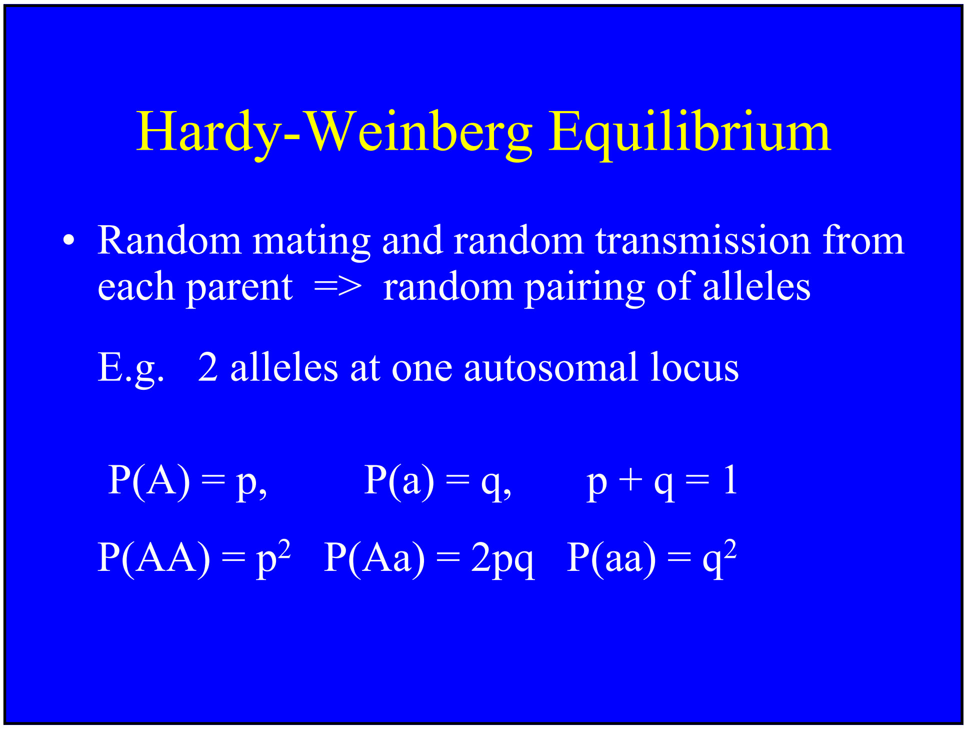
Estimation of heritability

- You can calculate using only trait information from familial data.
- see papers by:
  - Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 2008 Apr;9(4):255-66. doi: 10.1038/nrg2322. Epub 2008 Mar 4.
  - Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, et al. (2006) Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet 2(3): e41. doi:10.1371/journal.pgen.0020041



Penetrances

- Probability of expressing a phenotype given genotype. Penetrance is either “complete” or “incomplete”. If Y is a phenotype and (A, a) is a disease locus, then penetrances are
- f_2 = P(Y|AA), f_1 = P(Y|Aa), f_0 = P(Y|aa)
- Complete penetrance: (f_2, f_1, f_0)=(1, 1, 0) for autosomal dominant disorder
- (f_2, f_1, f_0)=(1, 0, 0) for autosomal recessive disorder



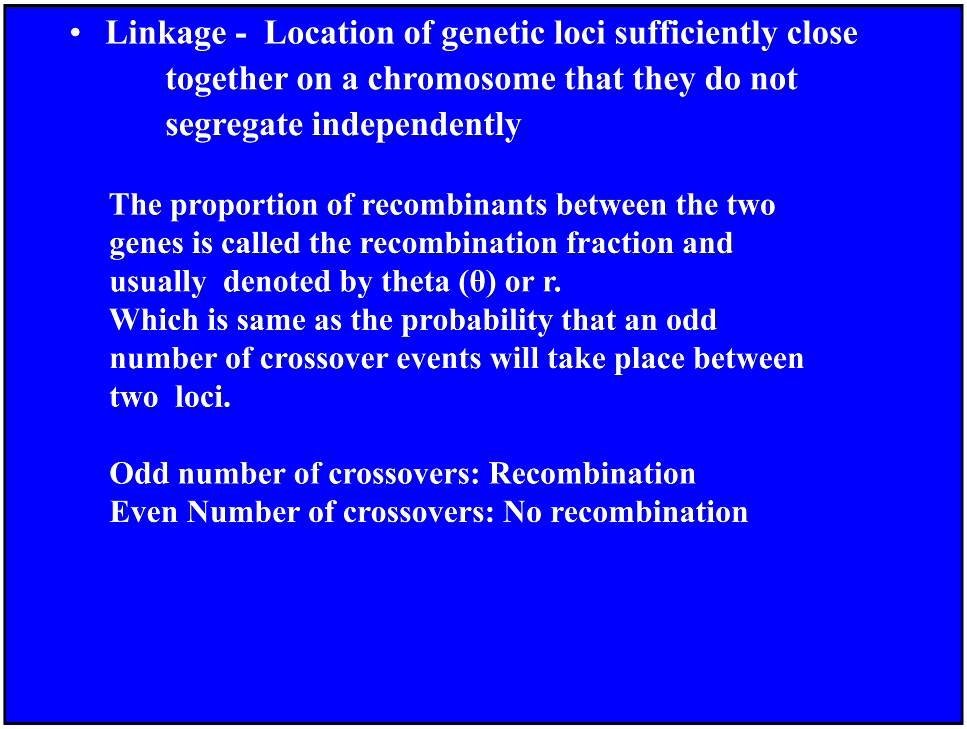
Hardy-Weinberg Equilibrium

- Random mating and random transmission from each parent => random pairing of alleles

E.g. 2 alleles at one autosomal locus

P(A) = p, P(a) = q, p + q = 1

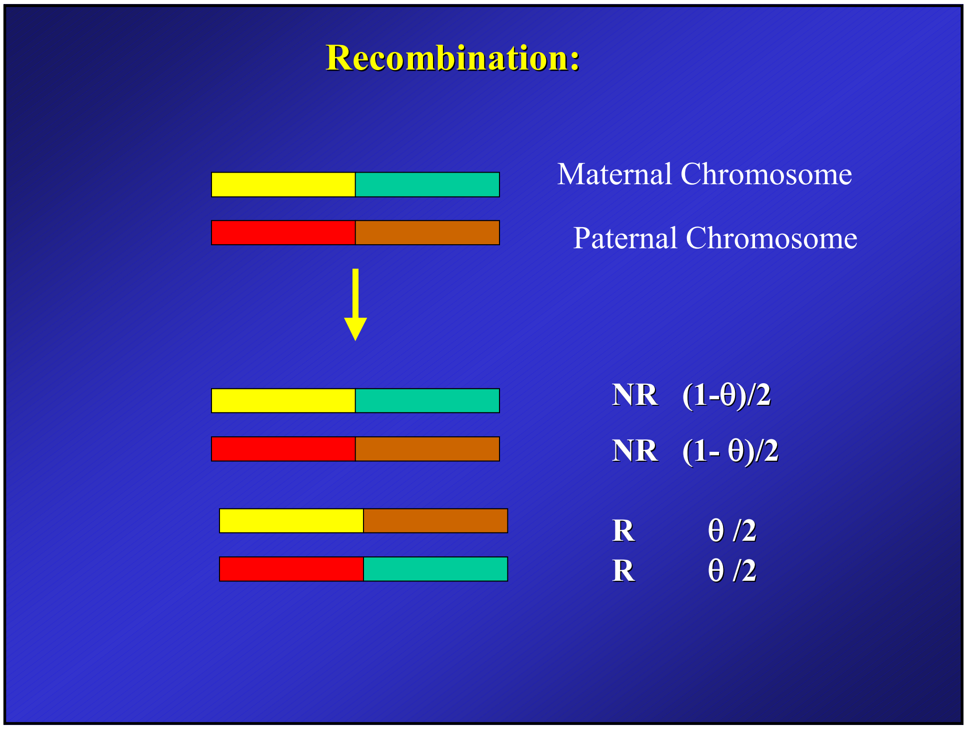
P(AA) = p^2 P(Aa) = 2pq P(aa) = q^2



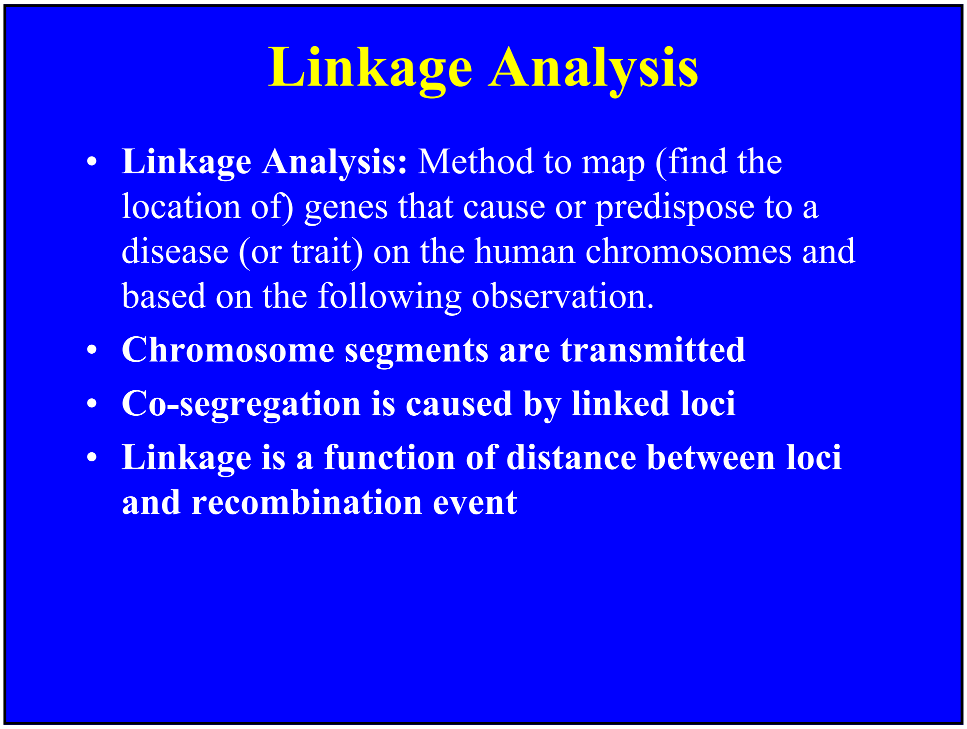
- **Linkage - Location of genetic loci sufficiently close together on a chromosome that they do not segregate independently**

The proportion of recombinants between the two genes is called the recombination fraction and usually denoted by theta (θ) or r.
Which is same as the probability that an odd number of crossover events will take place between two loci.

Odd number of crossovers: Recombination
Even Number of crossovers: No recombination

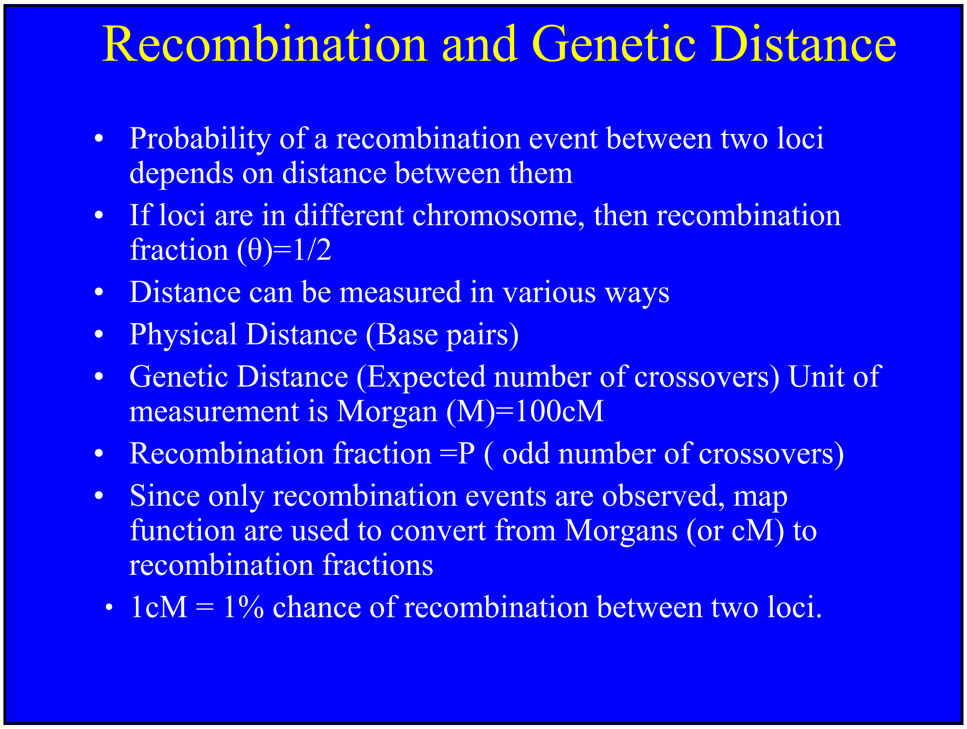


Recombination:



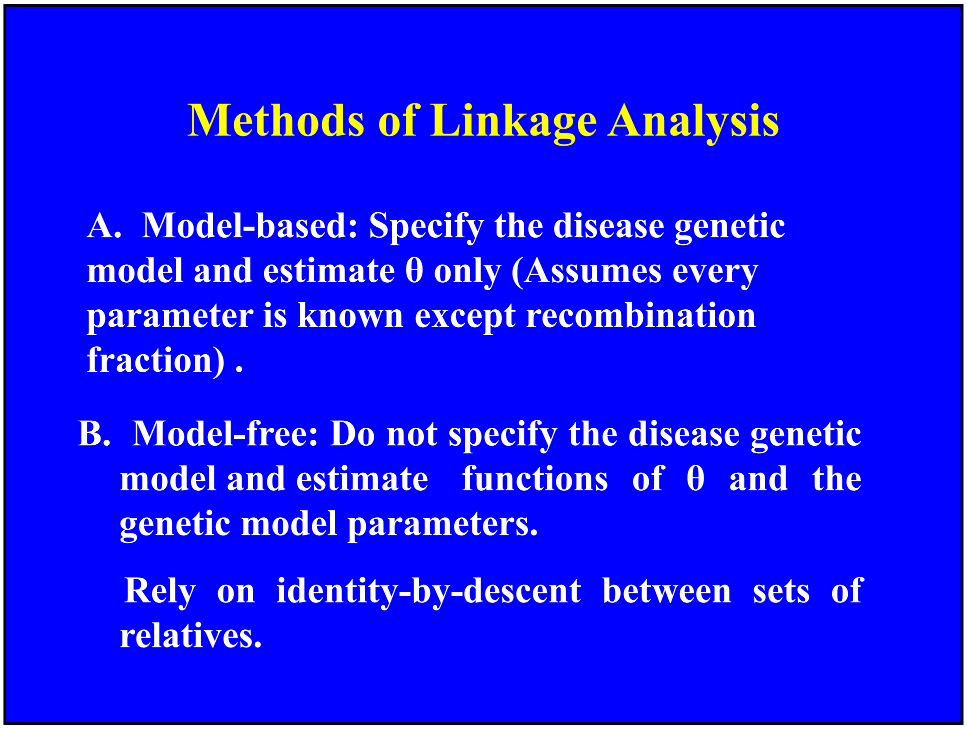
Linkage Analysis

- **Linkage Analysis:** Method to map (find the location of) genes that cause or predispose to a disease (or trait) on the human chromosomes and based on the following observation.
- **Chromosome segments are transmitted**
- **Co-segregation is caused by linked loci**
- **Linkage is a function of distance between loci and recombination event**



Recombination and Genetic Distance

- Probability of a recombination event between two loci depends on distance between them
- If loci are in different chromosome, then recombination fraction (θ)=1/2
- Distance can be measured in various ways
- Physical Distance (Base pairs)
- Genetic Distance (Expected number of crossovers) Unit of measurement is Morgan (M)=100cM
- Recombination fraction =P (odd number of crossovers)
- Since only recombination events are observed, map function are used to convert from Morgans (or cM) to recombination fractions
- 1cM = 1% chance of recombination between two loci.

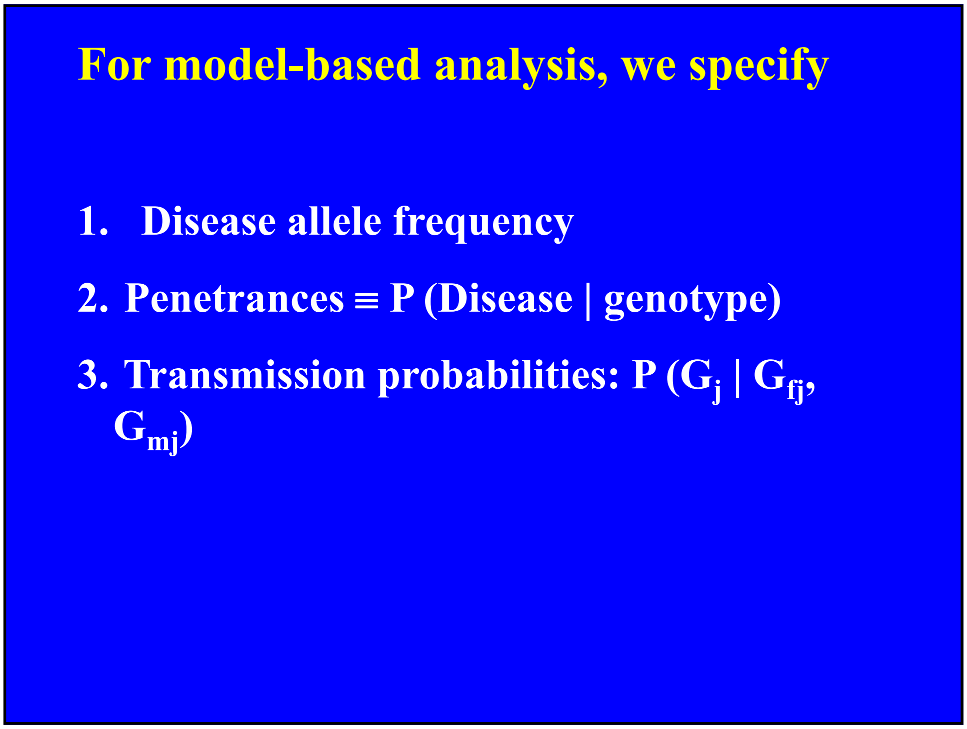


Methods of Linkage Analysis

A. Model-based: Specify the disease genetic model and estimate θ only (Assumes every parameter is known except recombination fraction) .

B. Model-free: Do not specify the disease genetic model and estimate functions of θ and the genetic model parameters.

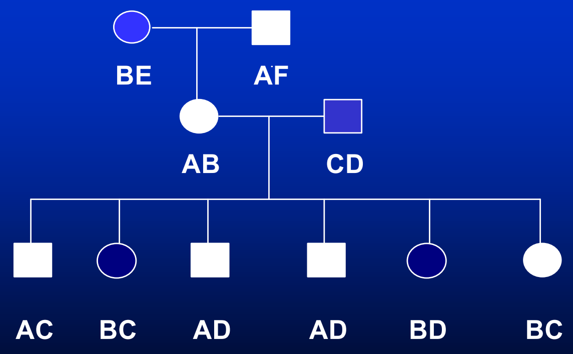
Rely on identity-by-descent between sets of relatives.



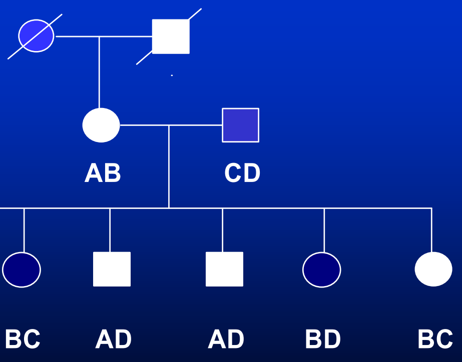
For model-based analysis, we specify

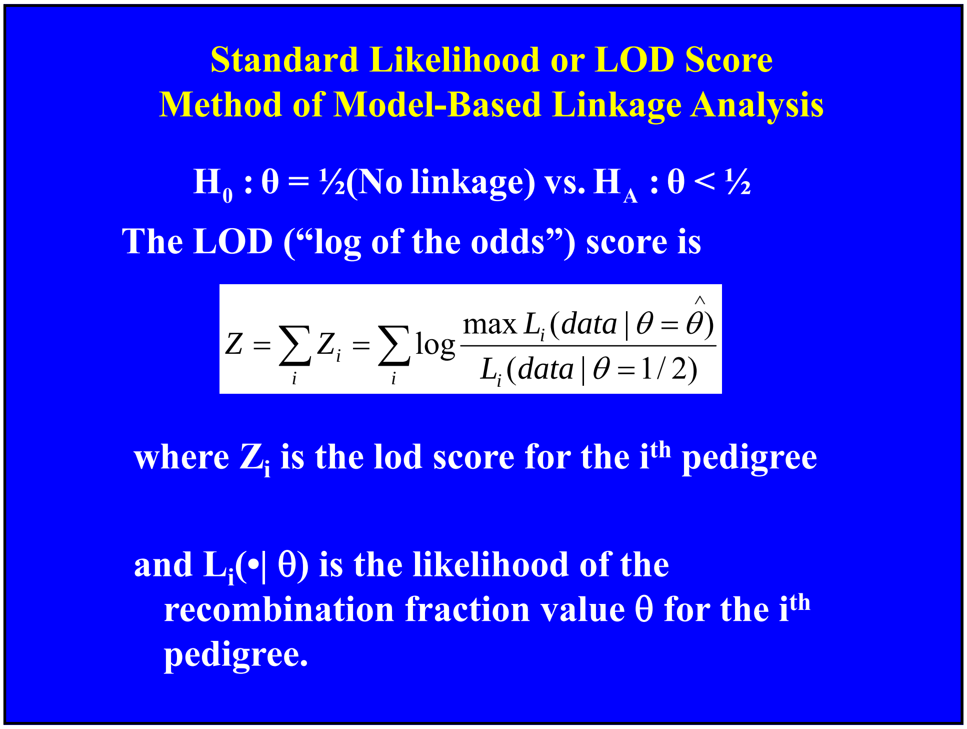
1. Disease allele frequency
2. Penetrances ≡ P (Disease | genotype)
3. Transmission probabilities: $P\ (G_j \mid G_{fj}, G_{mj})$

Phase Known Family



Phase Unknown Family





Standard Likelihood or LOD Score Method of Model-Based Linkage Analysis

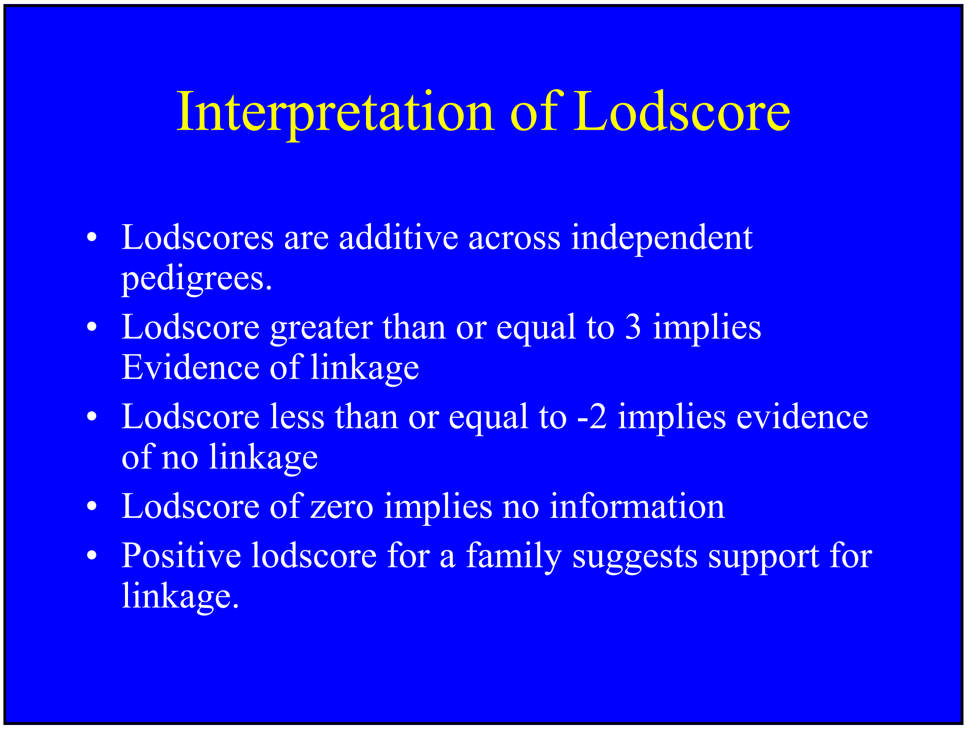
$H_0 : \theta = ½$(No linkage) vs. $H_A : \theta < ½$

The LOD ("log of the odds") score is

$$Z = \sum_i Z_i = \sum_i \log \frac{\max L_i(data \mid \theta = \hat{\theta})}{L_i(data \mid \theta = 1/2)}$$

where Z_i is the lod score for the ith pedigree

and $L_i(\bullet \mid \theta)$ is the likelihood of the recombination fraction value θ for the ith pedigree.



Interpretation of Lodscore

- Lodscores are additive across independent pedigrees.
- Lodscore greater than or equal to 3 implies Evidence of linkage
- Lodscore less than or equal to -2 implies evidence of no linkage
- Lodscore of zero implies no information
- Positive lodscore for a family suggests support for linkage.

Calculating Lodscore (Phase unknown family)

Likelihood= L(θ) = P(data|θ)

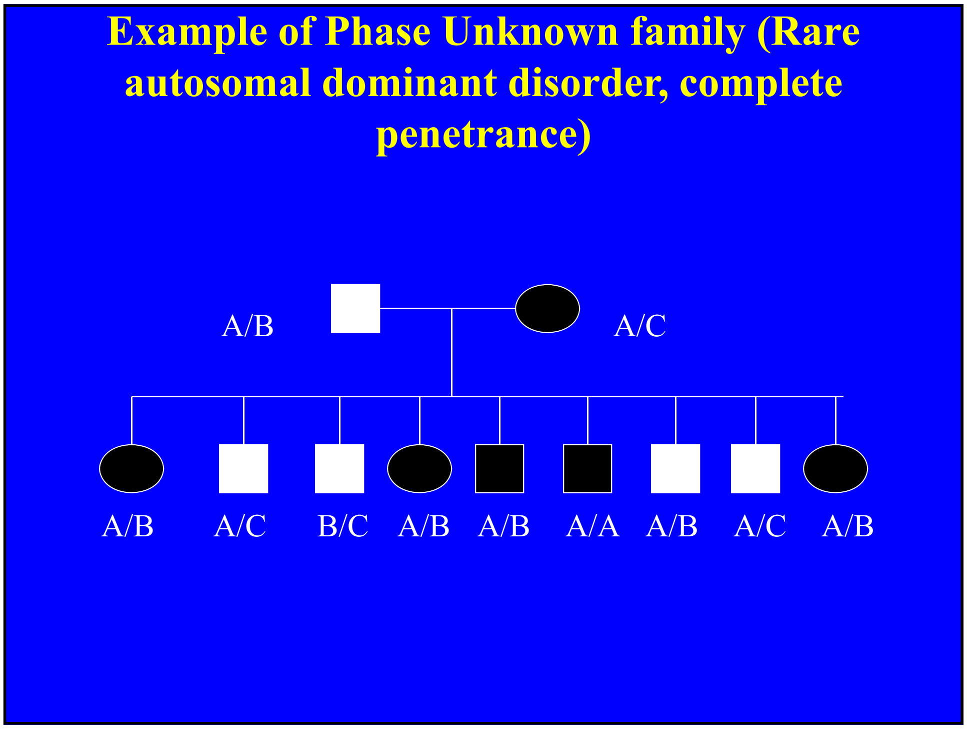
=1/2 [(θ) ^{k} (1-θ) $^{n-k}$

+ (θ) $^{n-k}$ (1-θ) ^{k}]

- k: No. of recombinants
- n: All meiosis

Lodscore (Phase unknown)

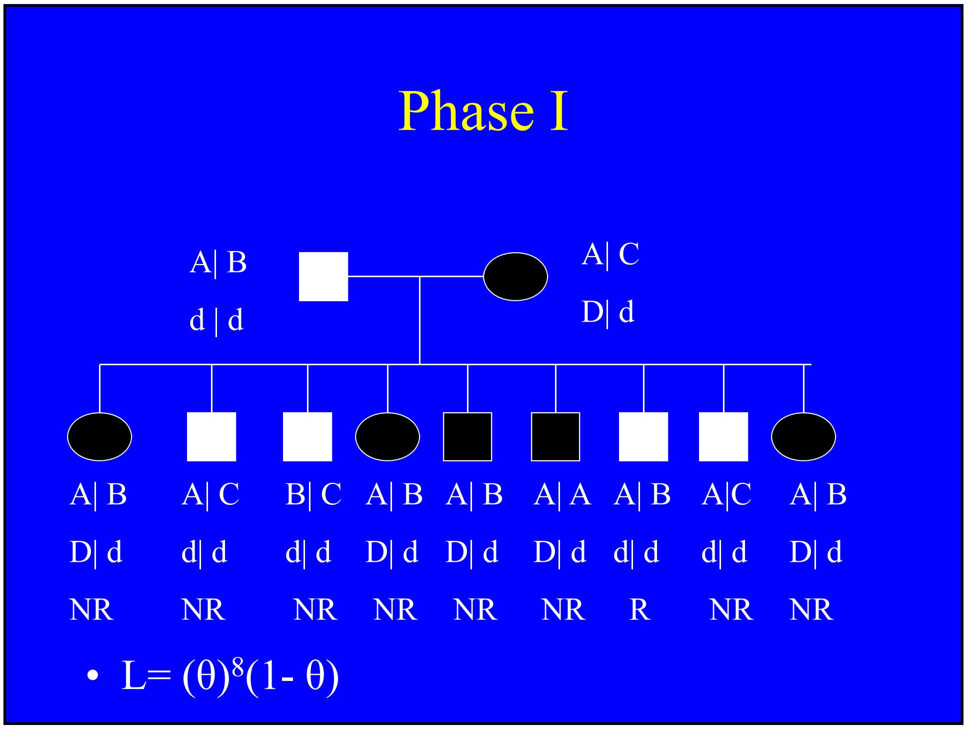
$$
\begin{aligned}
\boldsymbol{Z} = \boldsymbol{Lodscore} &= \log \frac{\max(\boldsymbol{L}(\boldsymbol{data} \mid \boldsymbol{\theta}))}{\boldsymbol{L}(\boldsymbol{data} \mid \boldsymbol{\theta} = 1/2)} \\
&= \log \frac{\max 1/2 \left[(\boldsymbol{\theta})^{\boldsymbol{k}} (1-\boldsymbol{\theta})^{\boldsymbol{n-k}}\right] + 1/2 \left[(\boldsymbol{\theta})^{\boldsymbol{n-k}} (1-\boldsymbol{\theta})^{\boldsymbol{k}}\right]}{(1/2)^{\boldsymbol{n}}} \\
&= \log 2^{\boldsymbol{n}-1} \left[\boldsymbol{\theta}^{\boldsymbol{k}} (1-\boldsymbol{\theta})^{\boldsymbol{n-k}} + \boldsymbol{\theta}^{\boldsymbol{n-k}} (1-\boldsymbol{\theta})^{\boldsymbol{k}}\right]
\end{aligned}
$$



Example of Phase Unknown family (Rare autosomal dominant disorder, complete penetrance)

A/B A/C

A/B A/C B/C A/B A/B A/A A/B A/C A/B



Phase I

A| B A| C

d | d D| d

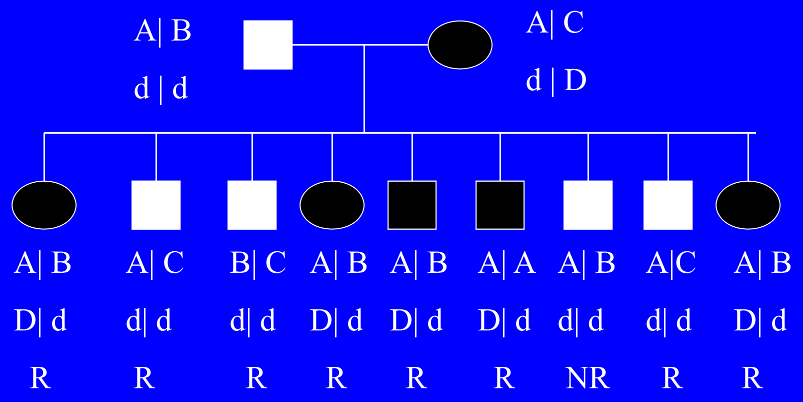
A| B A| C B| C A| B A| B A| A A| B A|C A| B

D| d d| d d| d D| d D| d D| d d| d d| d D| d

NR NR NR NR NR NR R NR NR

- L= $(\theta)^8(1- \theta)$

Phase II



A| B

d | d

A| C

d | D

A| B	A| C	B| C	A| B	A| B	A| A	A| B	A|C	A| B
D| d	d| d	d| d	D| d	D| d	D| d	d| d	d| d	D| d
R	R	R	R	R	R	NR	R	R

- L= $(\theta)(1- \theta)^8$

Lodscore

- Likelihood= $P(\text{data}|\theta)$

  =L(Phase I)+ L(Phase II)

  $=1/2[(\theta)^8(1- \theta)]+1/2\ [(\theta)(1- \theta)^8]$

- Lodscore= $\log\ [\max\ \{1/2[(\theta)^8(1- \theta))]+1/2\ [(\theta)(1- \theta)^8]\}/(1/2)^9\ ]$

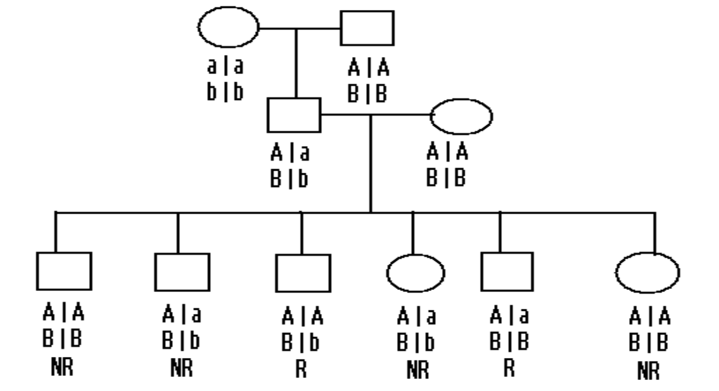
Calculating Lodscore (Phase known family)

- Likelihood= $L(\theta) = P(\text{data}|\theta)$
  $=(\theta)^{k}(1-\theta)^{n-k}$
- k: No. of recombinants
- n: All meiosis

Lodscore (Phase known family)

$$
\begin{aligned}
\boldsymbol{Z} = \boldsymbol{Lodscore} &= \log\frac{\max(\boldsymbol{L}(\boldsymbol{data}\,|\,\boldsymbol{\theta}))}{\boldsymbol{L}(\boldsymbol{data}\,|\,\boldsymbol{\theta}=1/2)} \\
&= \log\frac{\max(\boldsymbol{\theta})^{\boldsymbol{k}}(1-\boldsymbol{\theta})^{\boldsymbol{n}-\boldsymbol{k}})}{(1/2)^{\boldsymbol{n}}} \\
&= \log 2^{\boldsymbol{n}}\boldsymbol{\theta}^{\boldsymbol{k}}(1-\boldsymbol{\theta})^{\boldsymbol{n}-\boldsymbol{k}}
\end{aligned}
$$

Example: Lodscore of phase known family



- Lodscore = log $\{2^6[\theta^2(1- \theta)^4]$

MLE of $\theta = 2/6=1/3$

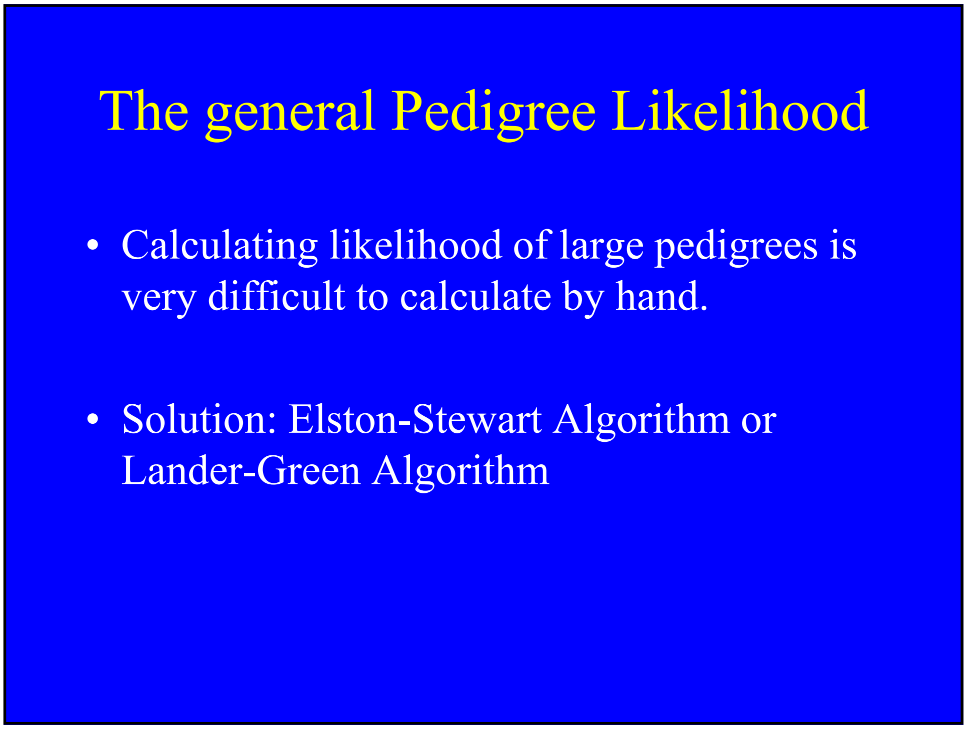
The general Pedigree Likelihood

Likelihood of the data

$$\propto \sum_{G_1} \ldots \sum_{G_n} \prod_{founder\ i} P(G_i) \prod_{nonfounders\ j} P(G_j \mid G_{f_j}, G_{m_j})$$

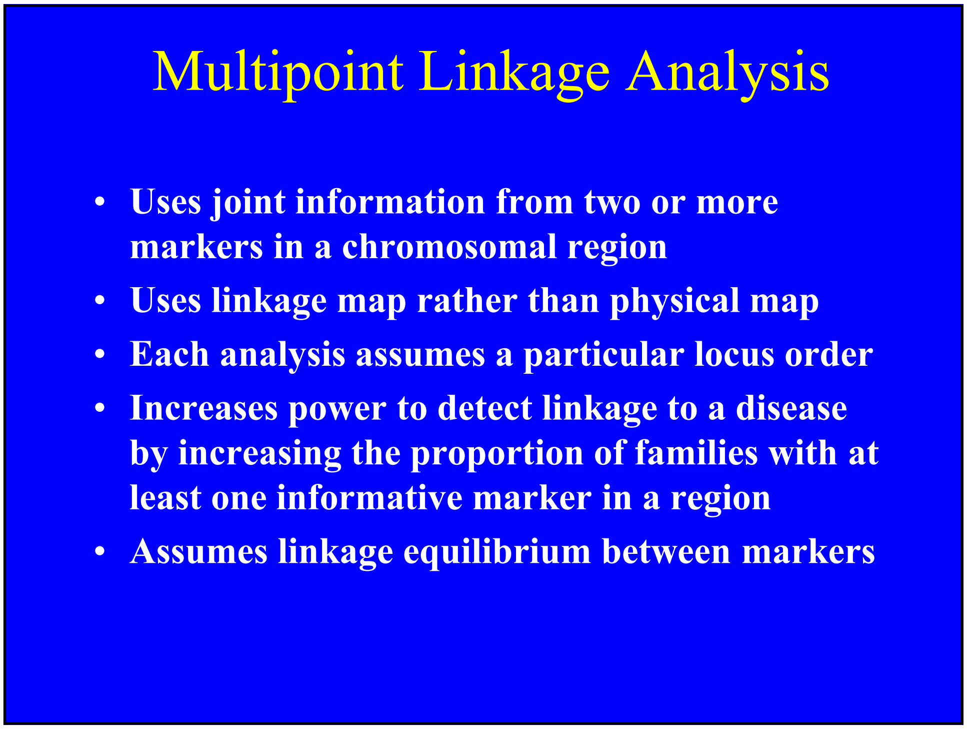
$$\times \prod_{observed\ l} P(Y_l \mid G_l)$$

$P(G_j \mid G_{f_j}, G_{m_j})$ *is* exp*ressed as a function of* 2 – *locus transmission probabilities*.



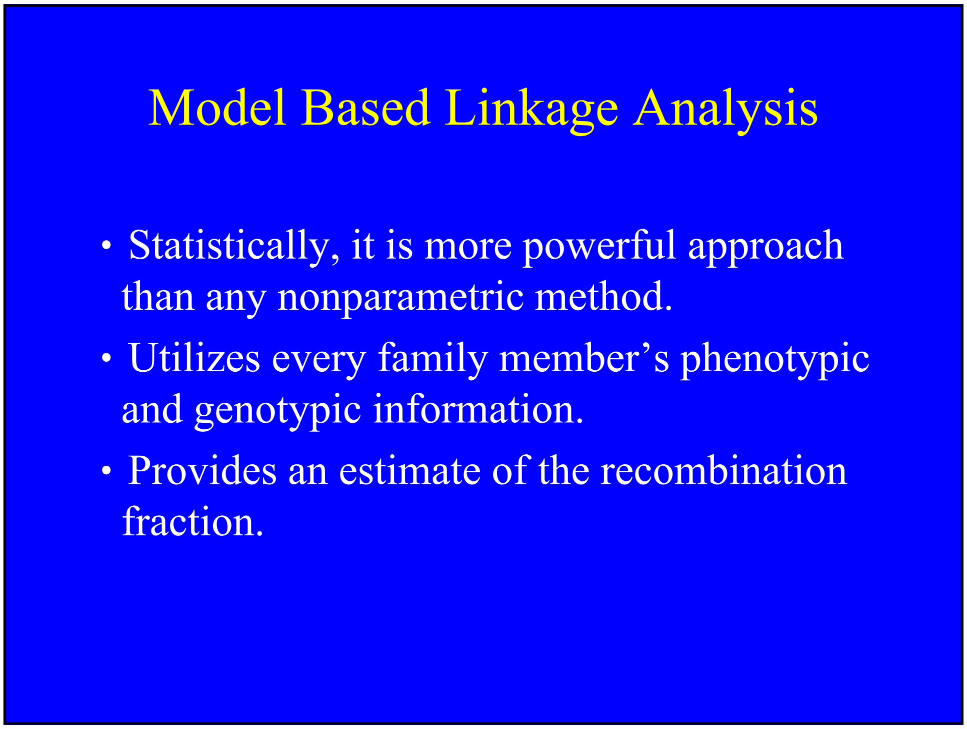
The general Pedigree Likelihood

- Calculating likelihood of large pedigrees is very difficult to calculate by hand.
- Solution: Elston-Stewart Algorithm or Lander-Green Algorithm



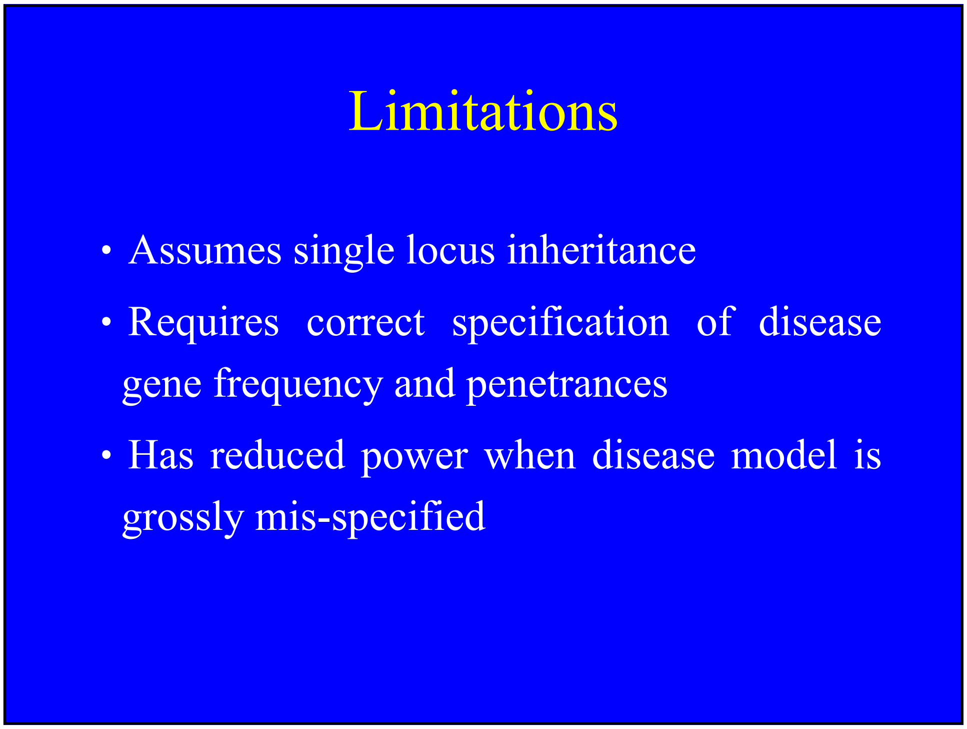
Multipoint Linkage Analysis

- **Uses joint information from two or more markers in a chromosomal region**
- **Uses linkage map rather than physical map**
- **Each analysis assumes a particular locus order**
- **Increases power to detect linkage to a disease by increasing the proportion of families with at least one informative marker in a region**
- **Assumes linkage equilibrium between markers**



Model Based Linkage Analysis

- Statistically, it is more powerful approach than any nonparametric method.
- Utilizes every family member’s phenotypic and genotypic information.
- Provides an estimate of the recombination fraction.



Limitations

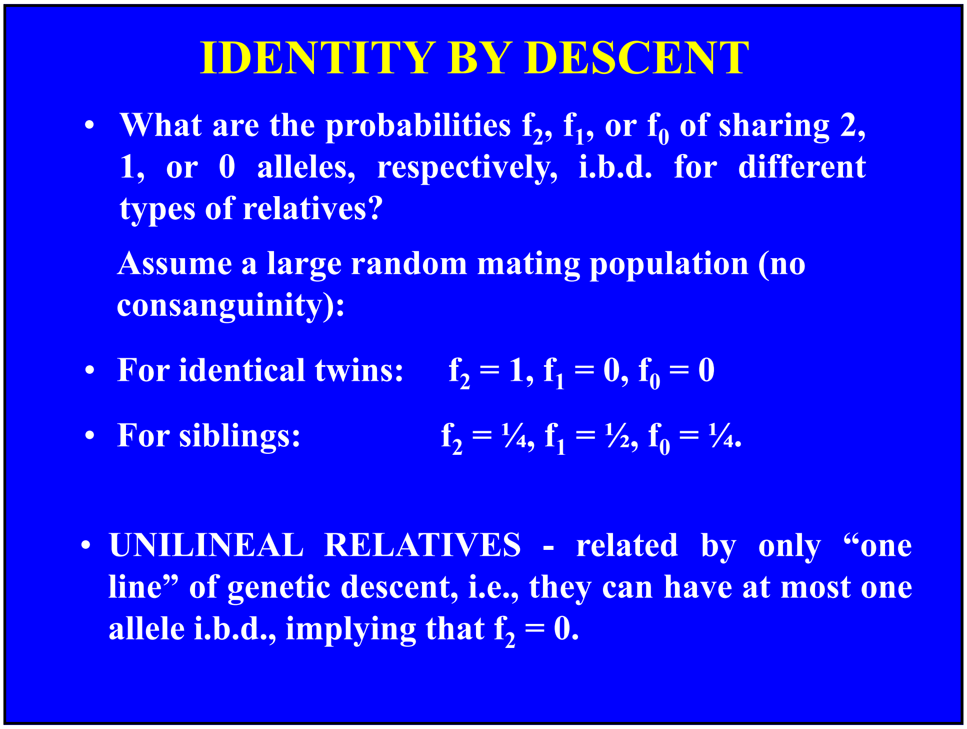
- Assumes single locus inheritance
- Requires correct specification of disease gene frequency and penetrances
- Has reduced power when disease model is grossly mis-specified

Many factors those can influence the lodscore

- Misspecification of disease inheritance model
- Misspecification of marker allele frequency
- Misspecification of penetrance values
- Misspecification of disease allele frequency

Model-Free Linkage Methods

- **Model-free linkage methods do not require specification of a genetic model for the trait of interest; that is, they do not require a precise knowledge of the mode of inheritance controlling the disease trait.**
- **Model-free linkage methods are typically computationally simple and rapid.**
- **Model-free linkage methods can be used as a first screen of multiple markers to identify promising linkage relationships. Such promising linkage relationships can subsequently be confirmed by consideration of other markers, by standard model-based analysis, by other methods, or a combination of approaches. Alternative approaches rely exclusively on model-free methods, particularly for the analysis of complex disorders, at this level of analysis.**



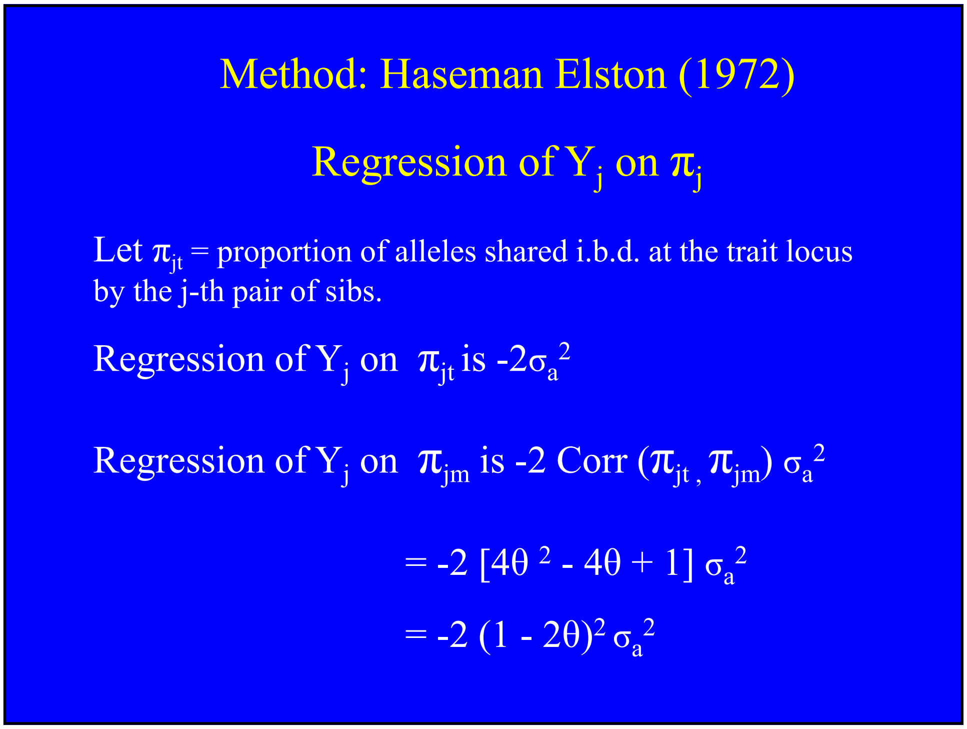
IDENTITY BY DESCENT

- What are the probabilities f_2, f_1, or f_0 of sharing 2, 1, or 0 alleles, respectively, i.b.d. for different types of relatives?

  Assume a large random mating population (no consanguinity):

- For identical twins: $f_2 = 1$, $f_1 = 0$, $f_0 = 0$
- For siblings: $f_2 = ¼$, $f_1 = ½$, $f_0 = ¼$.

- UNILINEAL RELATIVES - related by only “one line” of genetic descent, i.e., they can have at most one allele i.b.d., implying that $f_2 = 0$.



Method: Haseman Elston (1972)

Regression of Y_j on π_j

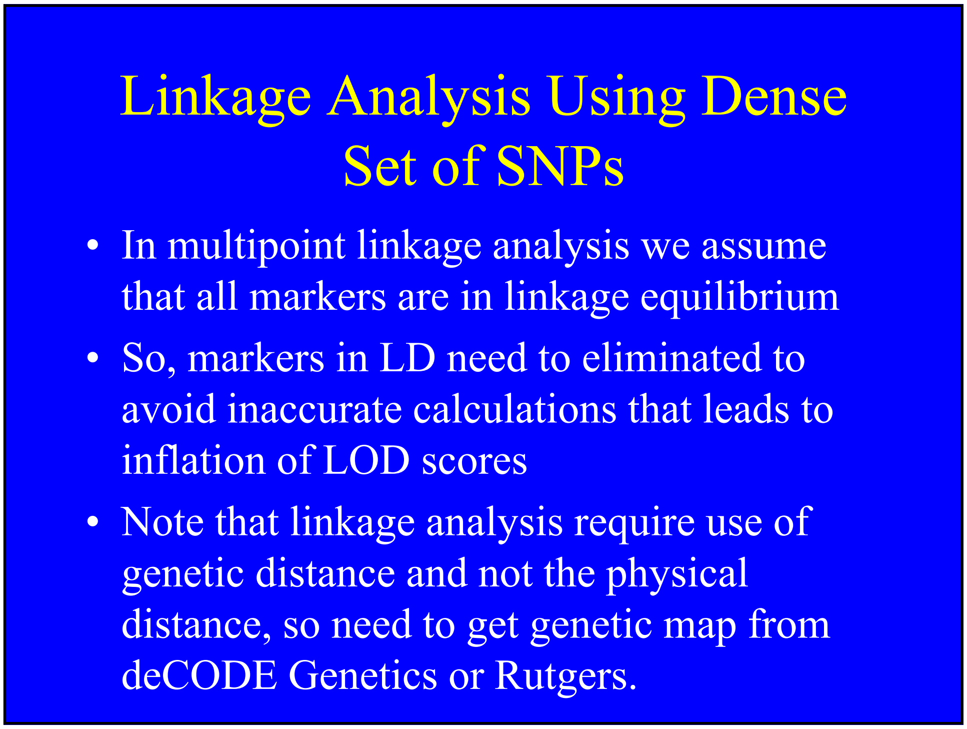
Let π_{jt} = proportion of alleles shared i.b.d. at the trait locus by the j-th pair of sibs.

Regression of Y_j on π_{jt} is $-2\sigma_a^2$

Regression of Y_j on π_{jm} is $-2 \text{ Corr } (\pi_{jt}, \pi_{jm}) \sigma_a^2$

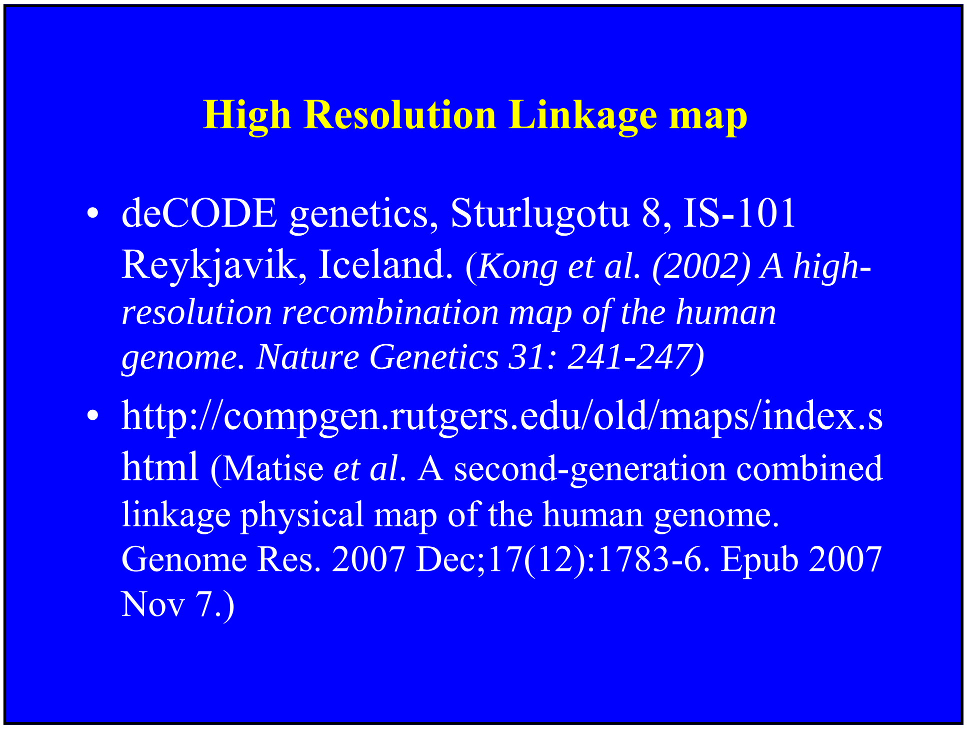
$$= -2\,[4\theta^2 - 4\theta + 1]\,\sigma_a^2$$

$$= -2\,(1 - 2\theta)^2\,\sigma_a^2$$



Linkage Analysis Using Dense Set of SNPs

- In multipoint linkage analysis we assume that all markers are in linkage equilibrium
- So, markers in LD need to eliminated to avoid inaccurate calculations that leads to inflation of LOD scores
- Note that linkage analysis require use of genetic distance and not the physical distance, so need to get genetic map from deCODE Genetics or Rutgers.



High Resolution Linkage map

- deCODE genetics, Sturlugotu 8, IS-101 Reykjavik, Iceland. (*Kong et al. (2002) A high-resolution recombination map of the human genome. Nature Genetics 31: 241-247)*
- http://compgen.rutgers.edu/old/maps/index.shtml (Matise *et al*. A second-generation combined linkage physical map of the human genome. Genome Res. 2007 Dec;17(12):1783-6. Epub 2007 Nov 7.)

Steps in Linkage Analysis Using Dense Set of SNPs

- Calculate Allele frequency of each marker
- Perform Hardy-Wineberg Equilibrium Test
- Mendelian Test
- Remove markers with LD
- Use appropriate linkage program to find gene locus for your trait

Software

- LINKAGE (Lathrop et al., 1984)
- FASTLINK (Schaffer et al., 1994)
- VITESSE (O'Connell & Weeks, 1995)
- GENEHUNTER (Kruglyak et al., 1996)
- S.A.G.E. (Elston et al., 2004)
- MERLIN (Abecasis, 2000)
- ALEGRO (Gudbjartsson et al., 2000, 2005)
- SOLAR (Almasy et al., 1998)
- SNP HiTLink (Fukuda et al., 2009)