Human Linkage and Association Analysis

Thurs., September 15, 2016

- Model-Free (Nonparametric) Analysis I
  - IBD v.s. IBS
  - Sib Pair Analyses
  - Pedigree Analysis
  - Recurrence risk estimation

Arpana Agrawal: arpana@wustl.edu

Schedule

- Sept 15: IBD and Kp
- Sept 20: Inheritance vectors & variance components approach.
  - HW1 assigned
- Sept 22:
  - Data cleaning.
  - Linkage practical in Merlin;
  - **QUIZ**
  - HW2 assigned

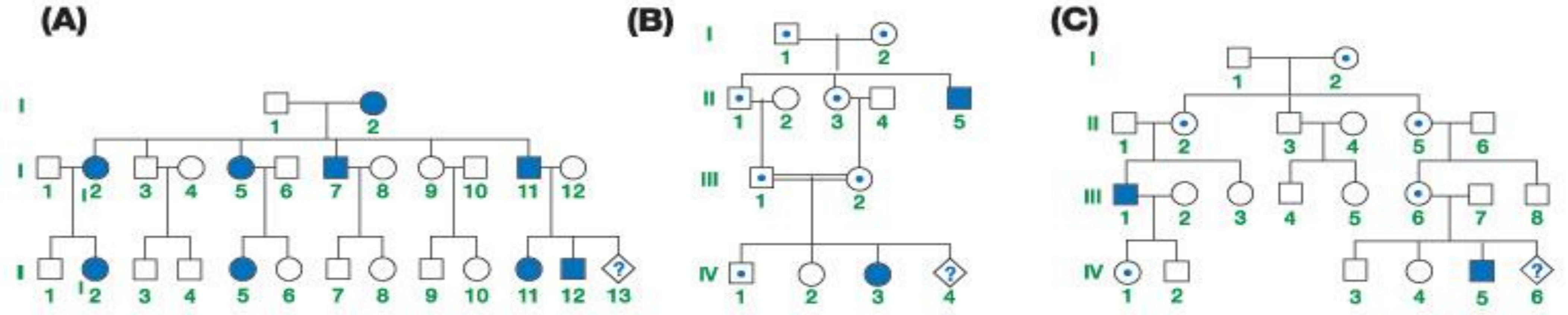
Motivation for Model-Free Linkage Analysis

- **Goal is to localize and identify genetic variants that regulate a trait of interest.**
- Linkage analysis is used to detect chromosomal regions that cosegregate with the trait of interest in pedigrees (i.e. localize).
- The classical method of linkage analysis requires parameters specifying a mode of inheritance:
  - gene frequencies
  - penetrance (likelihood of phenotype given genotype)
  - recombination fraction, θ
- Hence the term "parametric" ("model-based").
- Classical linkage analysis as applied to Mendelian traits may not be appropriate for complex traits.

Parameters

- **Recombination Fraction**:
- Probability that (an *odd number of)* crossovers occur between two loci and recombinant chromatids are produced.
- Genetic distance (m) is expected number of recombination events between 2 loci (for a single chromatid).
  - Haldane m = $-\frac{1}{2}[\ln(1-2\theta)]$
  - centiMorgan (cM) units, 5cM or θ=0.02.
  - 1cM is approx 1.05Mb (males) and 0.70Mb (females).
  - Problems?

- **Penetrance: Phenotype (x):Genotype (g)**
  - P (x) = ΣP(x|g ,c) P(g)
  - Eg. AD – 1:1:1:0 :: D/D:D/d:d/D:d/d
  - Problems?



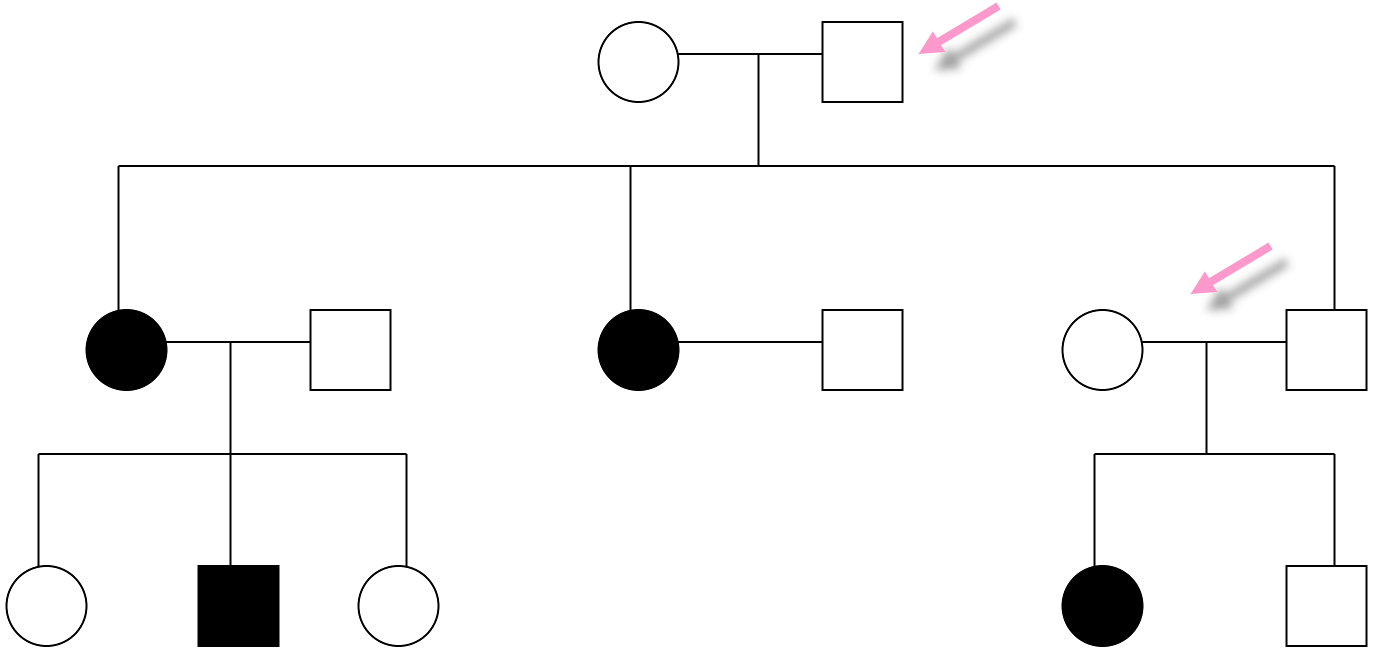
AD

AR

XR

Figure 4-2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Penetrance: Complete vs Incomplete



What is missing ?

What are Complex Traits?

Simple answer

Disease does not conform to single gene dominant or recessive Mendelian laws

What are Complex Traits?

More complicated answers

- Existence of reduced penetrance or nearly Mendelian: [e.g. $(f_0, f_1, f_2) = (0, 0.2, 0.9)$] often caused by the
  - ✓ action of modifier genes
  - ✓ segregation of additional genes required for phenotypic expression
- Existence of phenocopies: [e.g. $(f_0, f_1, f_2) = (0.01, 0.2, 0.9)$]
- Locus heterogeneity: can the disease be caused by any one of the abnormal genes?
- Trait may result from the combined effects of several genes (oligogenic inheritance or epistasis)
- Trait may result from many genes (polygenic), each, on its own, of small effect
- Trait may result from gene-environment interaction (how do we find the environmental factors?)
- Ill defined phenotype: the definition may be too broad; may actually be a collection of several diseases. Eg. many psychiatric disorders could be hard to diagnose unambiguously.

If only we knew in what sense a complex disease were complex"!

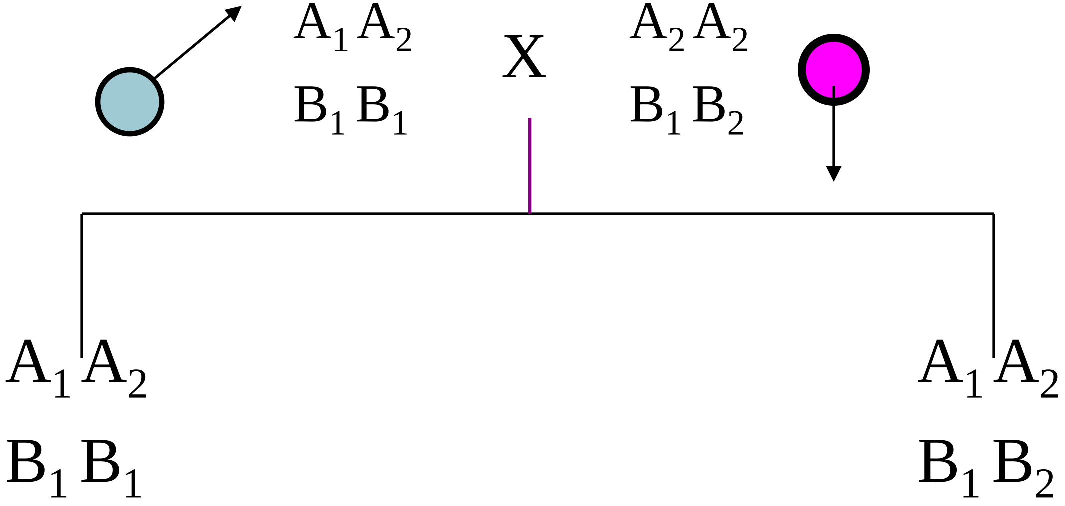
Examples of Simple and Complex Traits

- Simple
  - Cystic Fibrosis
  - Huntington’s Disease
  - Sickle Cell Anemia
  - Gaucher’s Disease
- Complex
  - Alzheimer’s Disease
  - Asthma
  - Depression
  - Diabetes
  - Heart Disease
  - Schizophrenia
  - Psoriasis

The Model-Free Approach: A simple idea, “sharing”

- Among sets of relatives, measure the relationship between
  - sharing of the disease trait and
  - sharing of marker alleles.
- Two types of sharing of alleles:
  - IBS – identity by state
  - IBD – identity by descent
- Two alleles of the same form (i.e., same DNA sequence) are said to be IBS.
- Two IBS alleles are said to be IBD if they are copies of the same ancestral allele.

Example: IBD v.s. IBS

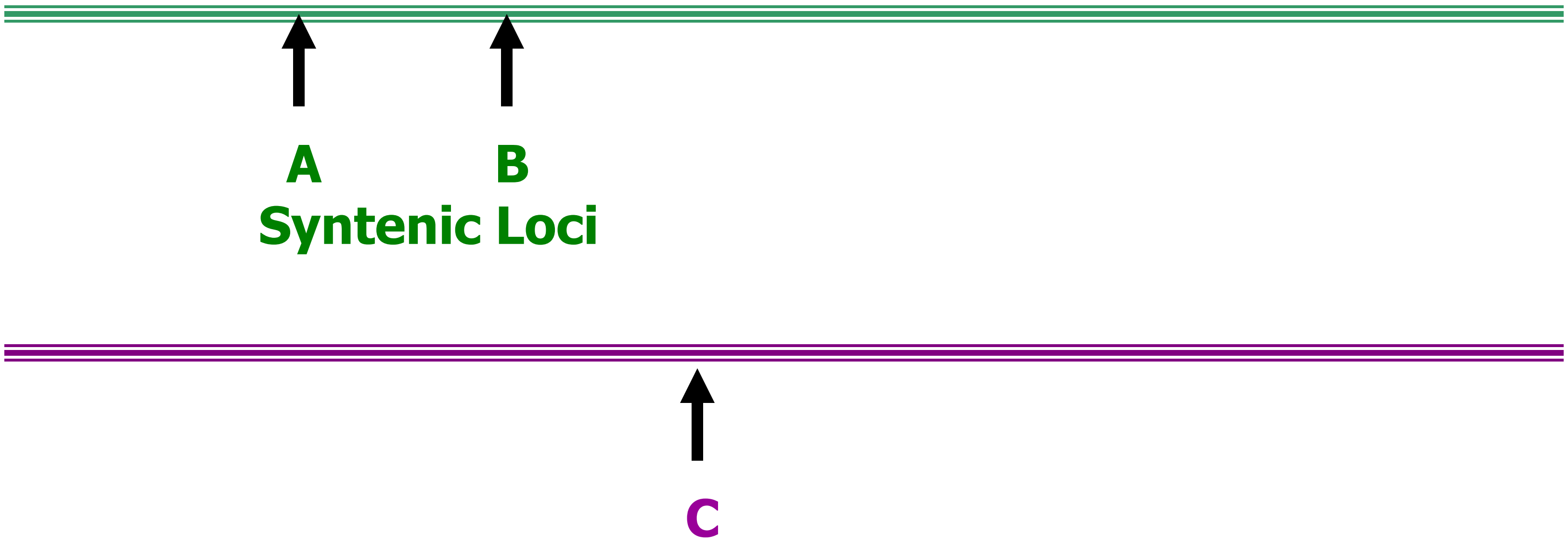


1. How many alleles are shared IBS and IBD at locus A?
2. How many alleles are shared IBS and IBD at locus B?
3. If there is tight linkage between the loci, can we infer anything more about locus A? Locus B?

The Previous Example:

- Illustrates the distinction between IBD and IBS
- Hints at the existence of a close relationship between IBD status and recombination events
- Implies that IBD is more directly relevant than IBS for linkage studies
- In general, IBD underlies phenotype similarities among relatives, i.e. relative of like phenotype have a greater probability of carrying genes IBD at loci contributing to the trait than <u>expected by pure chance</u>

Correlated IBD Values

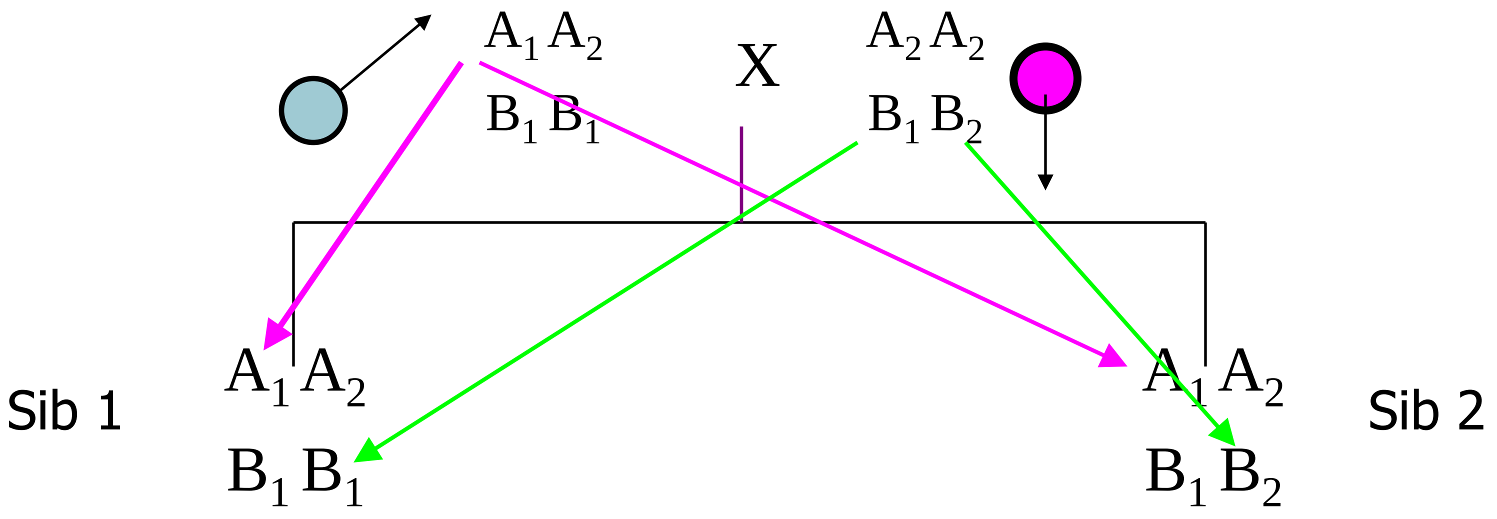


IBD values are correlated for loci A and B.

IBD values are uncorrelated for loci A and C, and B and C

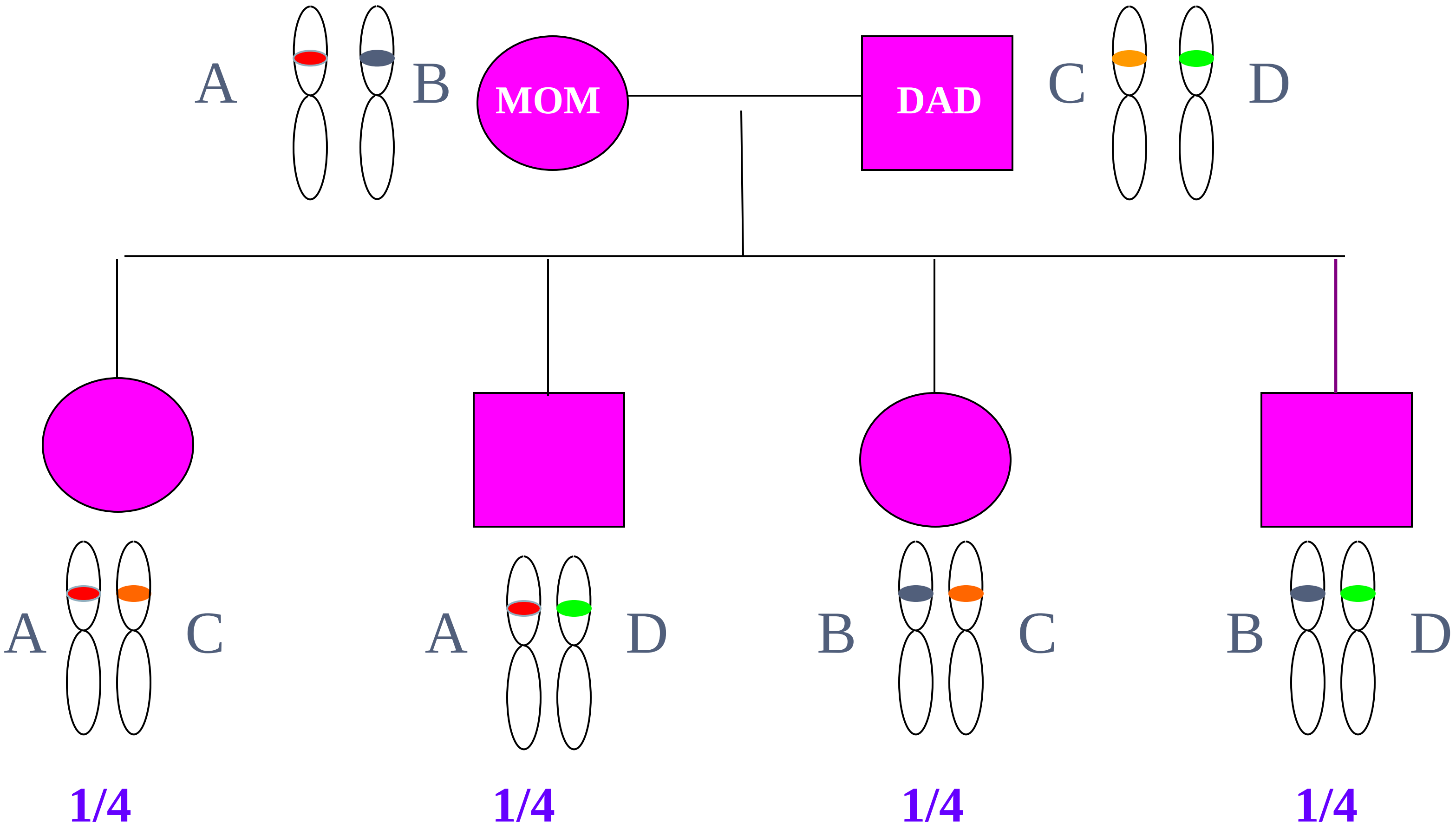
The closer the loci, the more highly correlated are the IBD values

Example: IBD v.s. IBS

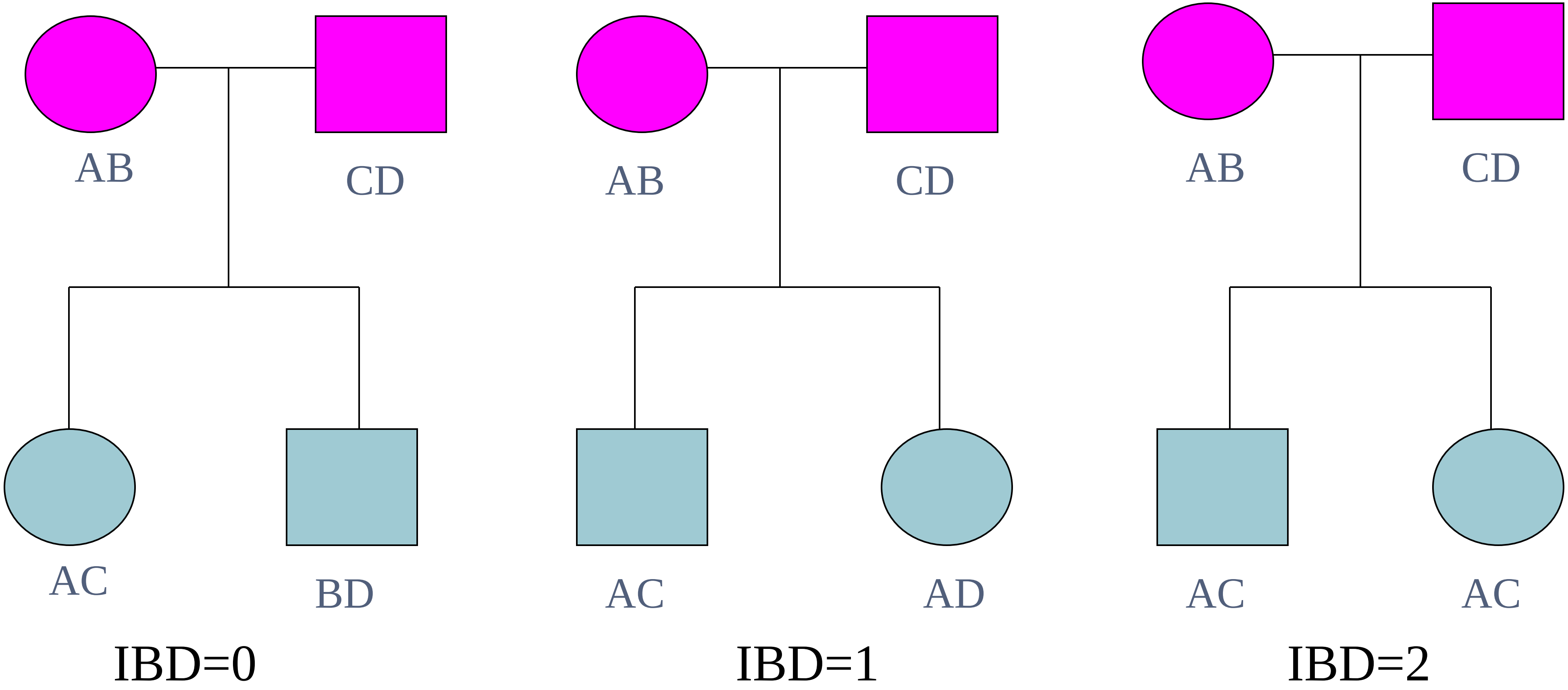


1. How many alleles are shared IBS and IBD at locus A?
2. How many alleles are shared IBS and IBD at locus B?
3. If there is tight linkage between the loci, can we infer anything more about locus A? Locus B?

Sibling Inheritance



IDENTITY BY DESCENT (IBD)
Siblings



IDENTITY BY DESCENT: Sibs

		Sib 1			
		AC	AD	BC	BD
Sib 2	AC	🟪	🟦	🟦	🟩
	AD	🟦	🟪	🟩	🟦
	BC	🟦	🟩	🟪	🟦
	BD	🟩	🟦	🟦	🟪

🟪 **4/16 = 1/4 sibs share BOTH parental alleles IBD = 2**

🟦 **8/16 = 1/2 sibs share ONE parental allele IBD = 1**

🟩 **4/16 = 1/4 sibs share NO parental alleles IBD = 0**

Kinship Coefficient

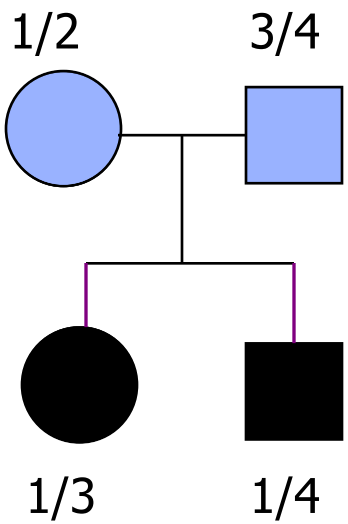
- Defined as the probability that a *randomly drawn allele at any locus of an individual* is IBD with a randomly drawn allele at the *same locus from another individual*.
- Equivalent to the inbreeding coefficient (F) of their offspring.
- $F = {}^{1}/_{2}P(IBD=2)+{}^{1}/_{4}\ P(IBD=1)+0P(IBD=0) = {}^{1}/_{2}\,\mathrm{E}(IBD) = {}^{1}/_{2}\pi$
- Π is *coefficient of relationship*: expected proportion of alleles shared IBD at a locus.
- Assumes *no* inbreeding.

Affected Sib Pair Analysis (ASP)

- Most popular form of linkage analysis based on IBD relationship
- Obtaining a sample of affected sib pairs is often easier than for other relationships
- Other types of relative pairs may be more powerful (depending on mode of inheritance) e.g. 1st cousins
- Occurrence of phenocopies can severely limit the power of ARP methods although the problem of incomplete penetrance is circumvented.

Test Statistics

- Assume n fully informative nuclear families with 2 affected siblings, i.e. an affected sibling pair (ASP)
- We observe the number of total pairs sharing 0, 1 or 2 alleles IBD, n_0, n_1, $n_{2,}$ respectively. Then n = n_0+ n_1+ $n_{2.}$
- Then (n_0, n_1, n_2) is a ‘realization’ from a multinomial distribution with parameters (p_0, p_1, p_2) where p_j = Pr(IBD = j), j = 0, 1, 2.
- Under H_0: no linkage, (p_0, p_1, p_2) = (1/4, 1/2, 1/4)
- There are a large number of statistical tests that can test the observed IBD distribution to the expected distribution under H_0



IBD=?

IBS=?

1. Pearson χ^2 Goodness-of-Fit Statistic

$$T_1 = \frac{(n_0 - e_0)^2}{e_0} + \frac{(n_1 - e_1)^2}{e_1} + \frac{(n_2 - e_2)^2}{e_2}$$

Where, e.g. for **full sibs**

e_0 = (1/4) n, e_1 = (1/2) n, e_2 = (1/4) n

Compare T_1 to a χ^2 with 2 df (two-tailed p)

Example

100 ASP:

12 pair IBD=0

54 pair IBD=1

34 pair IBD=2

$$T_1: \quad \frac{(12-25)^2}{25}+\frac{(54-50)^2}{50}+\frac{(34-25)^2}{25} \quad = 10.32, P \approx 0.0057$$

Reject the null hypothesis of no linkage !

2. Test for IBD=2
Based on, $n_{2,}$ the observed number ASP with IBD=2

$$X^2 = \frac{(n2\ - n/4)^2}{n/4}$$

Compare T_2 to a χ^2 with 1 df, one-tailed p

Example

100 ASP:

12 pair IBD=0

54 pair IBD=1

34 pair IBD=2

$$T_2: \frac{(34-25)^2}{100/4}$$

$$T_2: \frac{81}{25}$$

T_2=81/25 = 3.24

Using one-tailed (1df) test, p < .001

Reject the null hypothesis of no linkage!

3. 'Mean' Test
Based on the average number of marker alleles IBD

$$T3 = \frac{(\frac{1}{2}n_1 + n_2) - \frac{n}{2}}{\left(\frac{n}{8}\right)^{\frac{1}{2}}}$$

Dividing numerator and denominator by n; T_3 may be rewritten as:

$$= \sqrt{2n}\left(\hat{p}_1 + 2\hat{p}_2 - 1\right)$$

Testing whether the proportion, p, of alleles IBD is 1/2.

Asymptotically ~ N(0,1) under H_0: p = 1/2

Compare T_3 to a z with 1 df, one-tailed

Example

100 ASP:

12 pair IBD=0

54 pair IBD=1

34 pair IBD=2

T₃:

$$T_3: \quad \frac{\left(\frac{1}{2}\cdot 54 + 34\right) - \frac{100}{2}}{\left(\frac{100}{8}\right)^{\frac{1}{2}}} \approx 3.11 \qquad P=0.0009$$

Reject the null hypothesis of no linkage under either test.

Comments: Parametric vs Nonparametric

- Power to detect linkage with NP methods often depends on the true disease model.
  - Mean test is most powerful but all these tests assume fully informative sibs
  - Mean test is statistically equivalent to LOD score analysis under a recessive.
  - Others equivalences for allele-sharing tests with extended pedigrees have been shown by Whittemore [1996], Goring & Terwilliger [2000]
- Note that inherent in nonparametric analysis is the use of affecteds.
- **The concept of 'affecteds only' analysis implies that the trait phenotype of unaffecteds will not be used**.
  - This avoids incorrectly assigning a low-risk genotype to individuals in whom the mutation is not penetrant.

Affected Only Parametric Analysis

- Analyzing only affected individuals avoids (mistakenly) assigning a low-risk genotype to someone who is, in fact, at risk.
- In parametric analysis, one can explicitly specify the degree to which phenotypic information of unaffecteds will be used.
- Consider the following penetrance vectors:

Affected		Unaffected
(1,.4,0)	→	(0,0,1)
(.5,.5,0)	→	(.5,.5,1)
(.0001,.0001,0)	→	(.9999,.9999,1)

Assumptions

- Fully informative families at locus of interest
- Not the case in reality
- IBD sharing can be equivocal
- Need a method that uses partial & incomplete data
- Likelihood methods

The Likelihood (L) of Pedigree Data

Let n = the size of a human pedigree.

Let x_j = denote the phenotype of individual j.

Let g_j = denote the genotype of individual j.

Let $\vec{x} = (x_1, x_2, \cdots, x_n)$ be a vector of phenotypes.

Let $\vec{g} = (g_1, g_2, \cdots, g_n)$ be a vector of genotypes.

$$\begin{aligned} L(data) &= P(x_1, x_2, \cdots, x_n) \\ &= \Sigma_g P(\vec{x}, \vec{g}) \\ &= \Sigma_g P(\vec{x}|\vec{g})P(\vec{g}) \end{aligned}$$

In a simple pedigree, the pedigree likelihood above may be represented as

$$(1) \quad L = \Sigma_{g_{pa}}[P(x_{pa}|g_{pa})P(g_{pa})] \cdot \Sigma_{g_{ma}}[P(x_{ma}|g_{ma})P(g_{ma})] \cdot \Pi_{os} \Sigma_{g_{os}}[P(x_{os}|g_{os})P(g_{os}|g_{pa}, g_{ma})]$$

In the case of 2 possible phases in the parents, one can also write:

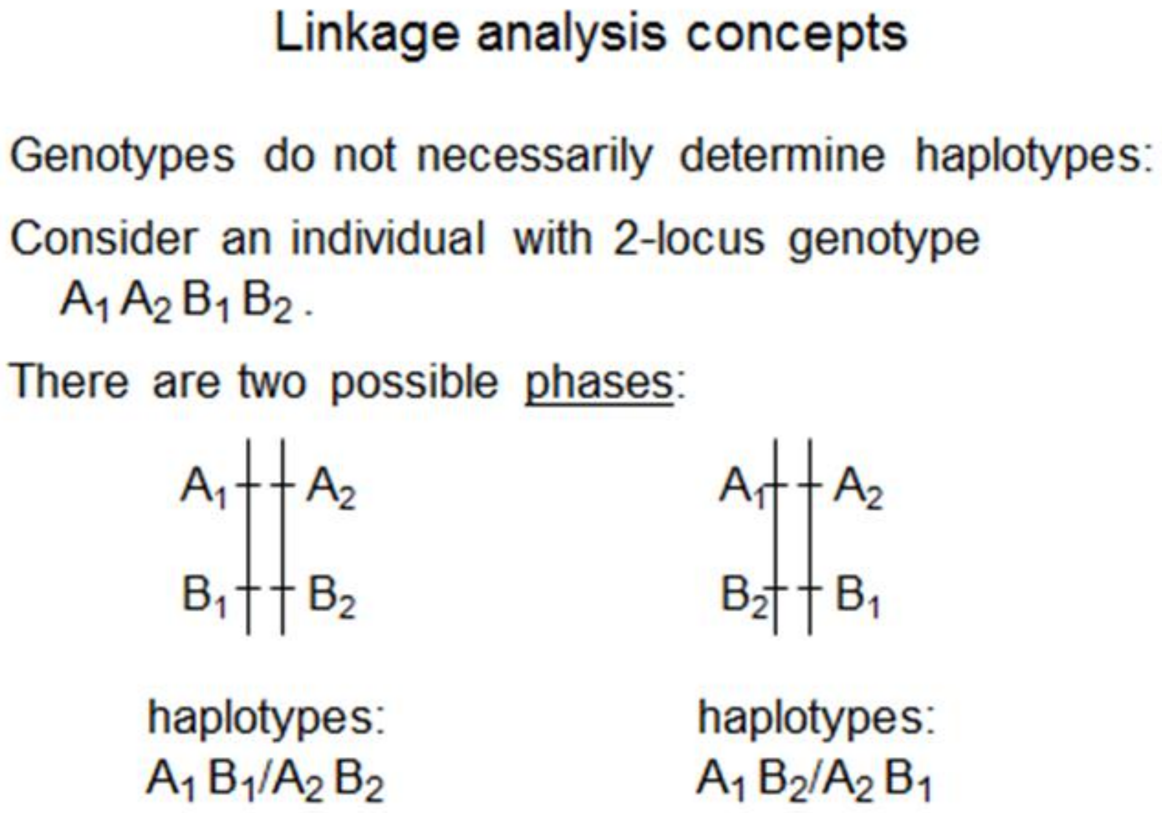
$$2. \quad L(data) = L(data|phaseI)P(phaseI) + L(data|phaseII)P(phaseII)$$

Use equation 1 to compute equation 2.

Most times the assumption is that the 2 phases are equally probability, i.e.

$$P(phaseI) = P(phaseII) = 1/2$$

Phase



Linkage analysis concepts

Genotypes do not necessarily determine haplotypes:

Consider an individual with 2-locus genotype $A_1 A_2 B_1 B_2$.

There are two possible phases:

A_1
A_2
B_1
B_2
A_1
A_2
B_2
B_1

haplotypes: $A_1 B_1/A_2 B_2$

haplotypes: $A_1 B_2/A_2 B_1$

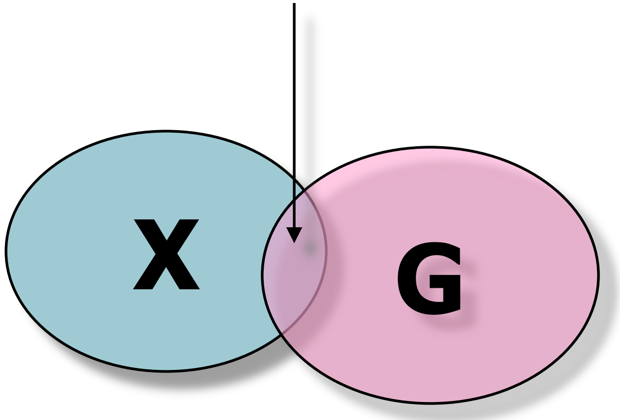
3 elements of pedigree data

- Probability of parental genotypes (*population*)
- Probability of phenotype given genotype (*penetrance*)
- Probability of offspring genotype given parental genotype (*transmission*)
- **Elston-Stewart algorithm**

$$P(x) = P(x|g)\,P(g)$$

(i) Prob of phenotype given genotype:

X ∩ G (The intersection of set X and G)



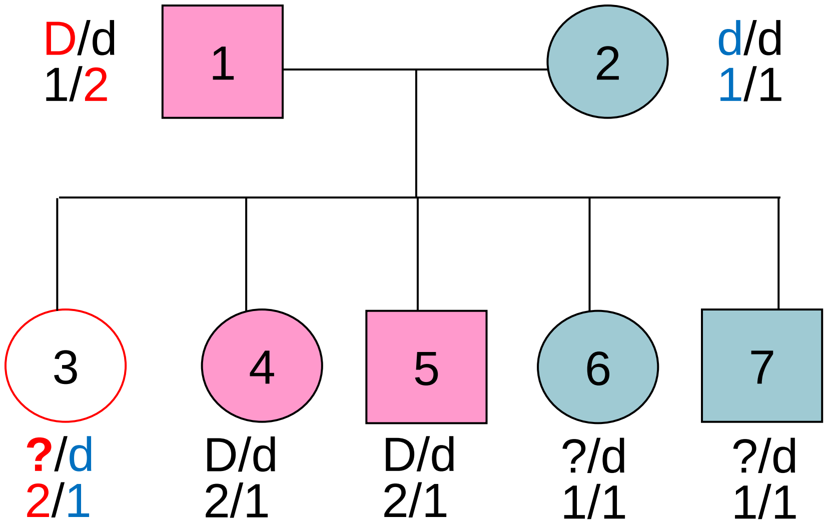
(ii) Probability of observing genotype itself

Bayes Rule

Essential for calculation of posterior probabilities;

$$P(A|B) = P(B|A)\, P(A) \,/\, P\,(B)$$

An “Affected Only” Pedigree Analysis (simplified)



Let $\vec{f} = (f_2, f_1, 0)$
Let $x_j = 0$ if unaffected
and $x_j = 1$ if affected

Assume mother is d/d

For DAD $(x_1 = 1)$:

Define Phase I $= \begin{array}{c|c} D & d \\ 1 & 2 \end{array}$ Phase II $= \begin{array}{c|c} D & d \\ 2 & 1 \end{array}$

Assume Phase I:

$$\begin{aligned} P(x_3 = 0) &= P(x_3 = 0|DD)P(DD) + P(x_3 = 0|Dd)P(Dd) + P(x_3 = 0|dd)P(dd) \\ &= (1 - f_2)\cdot 0 + (1 - f_1)\theta + (1 - 0)(1 - \theta) \\ &= (1 - f_1)\theta + (1 - \theta) \end{aligned}$$

If f_1 is very small then $P(x3 = 0) \approx 1$

$$\begin{aligned} P(x_4 = 1) &= P(x_4 = 1|DD)P(DD) + P(x_4 = 1|Dd)P(Dd) + P(x_4 = 1|dd)P(dd) \\ &= f_2 \cdot 0 + f_1\theta + 0\cdot(1 - \theta) \\ &= f_1\theta \end{aligned}$$

Simplified some more…

- P(DD)=0
- P(Dd)=2/1 (if Phase I) = θ
- P(Dd)=2/1 (if Phase II)= 1-θ
- P(dd)= 2/1 (if Phase I) = 1-θ
- P(dd)= 2/1 (if Phase II) = θ
- Likelihood of being affected if DD = f_2
- Likelihood of being unaffected if DD = $(1-f_2)$

Other ways to estimate IBD

- Because, in reality, IBD cannot be “called” unambiguously;
- Not fully informative;
- Some alternatives exist:
  - Recurrence risk (without markers)
  - Using PMTs, disease MOI and penetrance

Obtaining Evidence for Linkage without markers
Recurrence Risks and Risk Ratios for Pairs of Relatives

Risch, AJHG 1990

- GENERAL PRINCIPLE: Affected Relative Pairs should, on average, share MORE marker alleles IBD than expected by chance.
- IBD allele sharing at a marker for relatives is a function of:
  - the mode of inheritance (e.g. single loci, interacting loci, 100 loci?)
  - recombination fraction between the marker and the disease locus
- Examination of the recurrence risks and risk ratios for various ‘degrees’ of relatives of an affected individual, may indicate the underlying genetic model, i.e., the number of loci, whether there is epistasis or heterogeneity.

Relationship of the Risk Ratios, K_R, Between Relative Pairs

- Let T=1 if an individual from the population is affected, T=0 if unaffected
- Let K**=**P(T=1), the population prevalence
- Note also: K=E(T) and T is dichotomous
- Let T_1 and T_2 be the disease status for related individuals 1 and 2, respectively.
- Let the **recurrence risk** be defined as $K_R = P(T_2=1|T_1=1)$
- Note: $P(T_1=1 \text{ and } T_2=1) = P(T_2=1|T_1=1)\ P(T_1=1)= K_R\, K$

Risk Ratios, K_R, Between Relative Pairs (continued)

$$Let\ \ \lambda_R \equiv \frac{K_R}{K} \quad \left[\begin{array}{c}\text{this is the increase in risk to}\\ \text{a relative of an affected person}\\ \text{over the population prevalence}\end{array}\right]$$

Eg. For Type I Diabetes, K~.4%=.004

For siblings, K_s=.06

Thus,

$$\lambda_s = \frac{K_s}{K} = \frac{.06}{.004} = 15$$

(Davies et al., 1994)

Probabilities of Sharing IBD for Affected Pairs

Let:

$\alpha_{R_i} =$ P(IBD=i at an arbitrary locus for 2 relatives of type R)

$Z_{R_i} =$ P(IBD=i |2 relatives are affected)

Then $Z_{R_0} = \dfrac{\alpha_{R_0}}{\lambda_R}$

$$Z_{R_1} = \alpha_{R_1} \frac{\lambda_0}{\lambda_R}$$

$$Z_{R_2} = \alpha_{R_2} \frac{\lambda_M}{\lambda_R}$$

[M=MZ twins]

The assumption underlying these formulae is that θ=0 between trait and marker loci.

Derivation

$$Z_{R_0} = P(IBD = 0 \mid ARP) = \frac{P(ARP \mid IBD = 0)P(IBD = 0)}{P(ARP)}$$

$$= \frac{(K \times K)\alpha_{R_0}}{K_R \times K} = \frac{\alpha_{R_0}}{\lambda_R}$$

0.25 for sibs

$$Z_{R_2} = P(IBD = 2 \mid ARP) = \frac{P(ARP \mid IBD = 2)P(IBD = 2)}{P(ARP)}$$

$$= \frac{(K \times K_{MZ})\alpha_{R_2}}{K_R \times K} = \alpha_{R_2} \times \frac{\lambda_{MZ}}{\lambda_R}$$

$$Z_{R_1} = 1 - Z_{R_0} - Z_{R_2}$$

*Simplified version on page 123

$1 \leq \lambda_o \leq \lambda_s \leq \lambda_{MZ}$ for any genetic model

An Example

Suppose $\lambda_{P/O} = \lambda_S = 1.0625$

$\lambda_{MZ} = 1.125$

What are the probabilities that an affected sib pair share 0, 1, or 2 alleles IBD at a disease locus?

$Z_0 = .25/1.0625$; $Z_2 = .25*(1.125/1.0625)$

Answer: 0.235, 0.5, and 0.265, respectively

NB: Even if the marker were totally informative and tightly linked to a disease locus, huge sample sizes are necessary to detect a perturbation from (0.25, 0.5, 0.25)

Risk Ratios, K_R, Between Relative Pairs (continued)

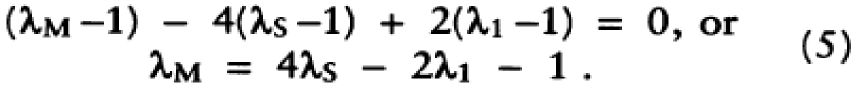
How can these help us to understand the genetic underpinnings of a disease?

By examining the familial recurrence patterns, an “assessment of the number of loci and the relevant parameters for linkage can be made…”

E.g. If the underlying mode of inheritance is a <u>single locus</u> disease, then

$$\lambda_1 - 1 = 2(\lambda_{\underline{2}} - 1) = 4(\lambda_3 - 1)$$

where 1= first degree relative
2= second degree relative
3= third degree relative



$$(\lambda_M - 1) - 4(\lambda_S - 1) + 2(\lambda_1 - 1) = 0, \text{ or} \qquad (5)$$
$$\lambda_M = 4\lambda_S - 2\lambda_1 - 1\,.$$

And $(\mathbf{K}_{\mathbf{MZ}}\text{-}1)=4\mathbf{K}_{\mathbf{S}} + 2\mathbf{K}_{\mathbf{1}}\text{-}1$

Note: Risch developed formulae for more complex made of inheritance (e.g. 2-locus models, multiplicative and additive models, heterogeneity models). He applied it to a schizophrenia data set and showed the risk ratios were not compatible with a single-locus.

Summary

- Why non-parametric?
  - Penetrance, recombination fraction
- IBD vs IBS – which is more useful?
- Likelihood of pedigree data
- Using recurrence risk to guess at IBD
  - NEXT week:
    - Using PMT to guess at IBD
    - Inheritance vectors
    - Variance components analysis