Pairs of Individuals: Simple Linkage Tests

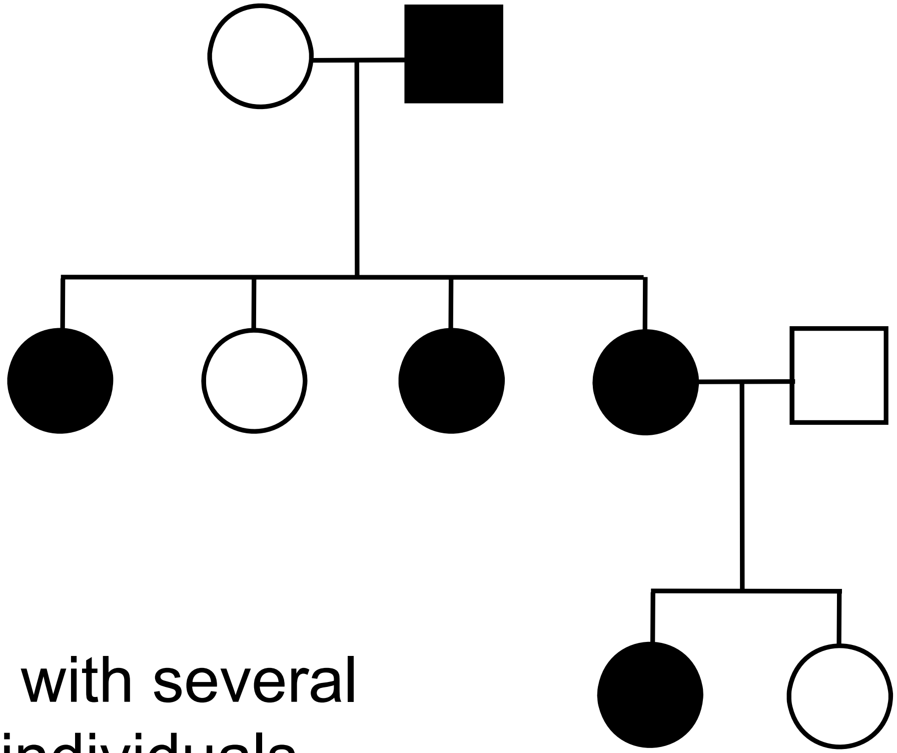
Biostatistics 666

Intuition for Linkage Analysis

- Millions of variations could potentially be involved
  - Costly to investigate each individually

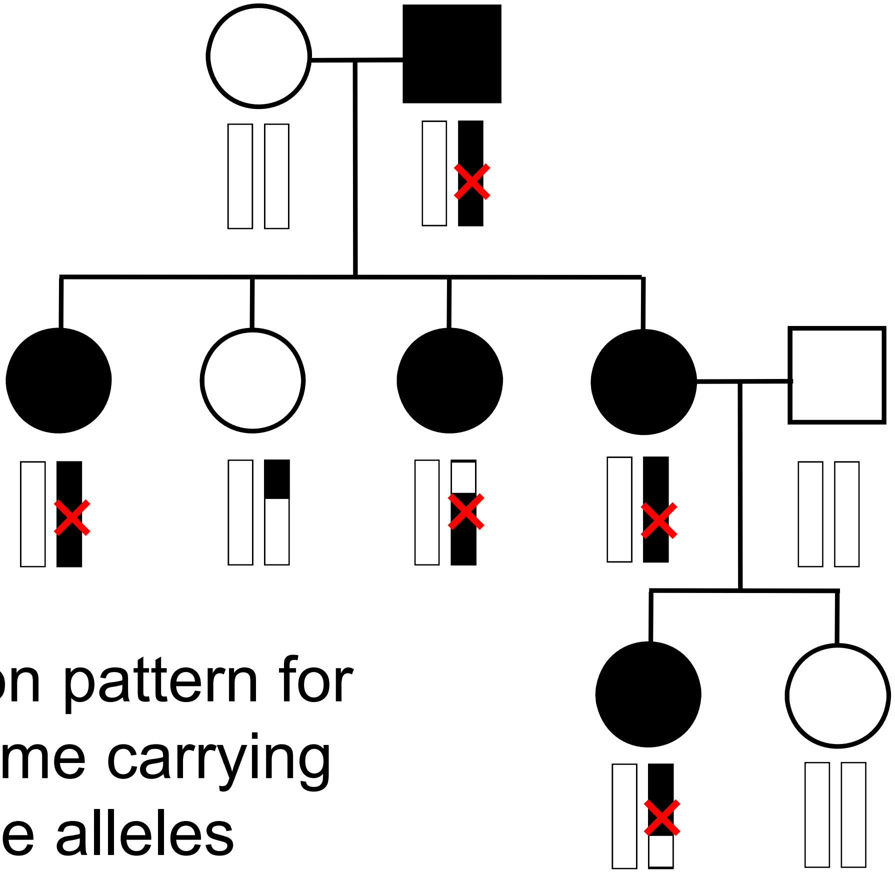
- Within families, variation is organized into a limited number of haplotypes
  - Sample modest number of markers to determine whether each stretch of chromosome is shared

Tracing Chromosomes



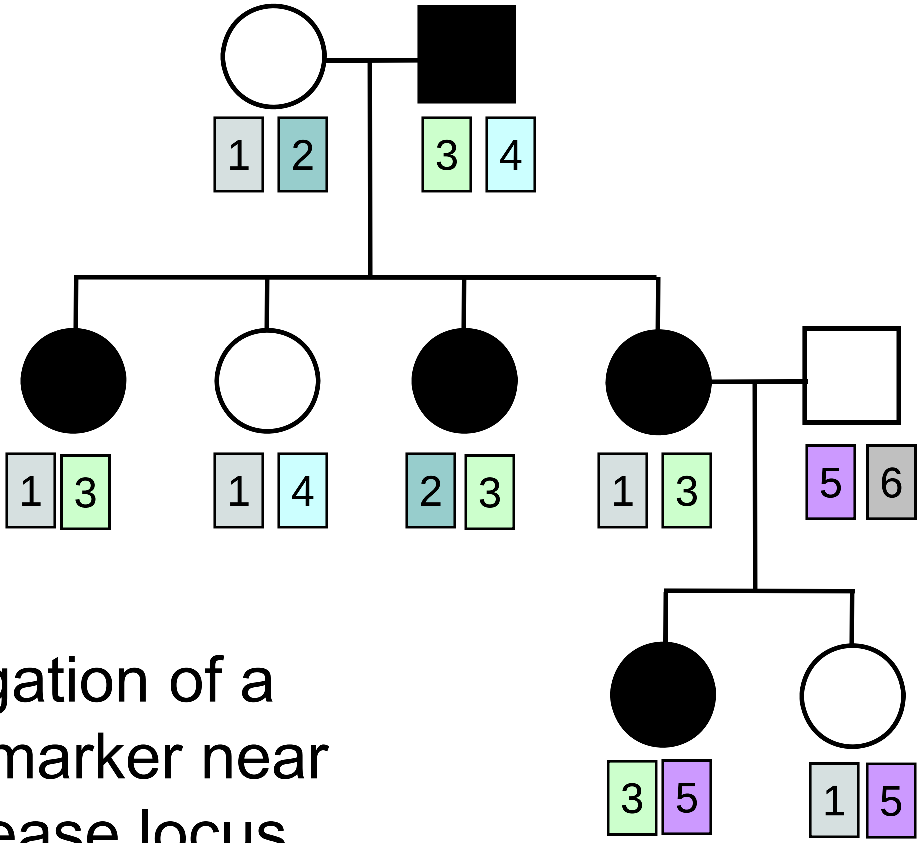
A pedigree with several affected individuals

Tracing Chromosomes



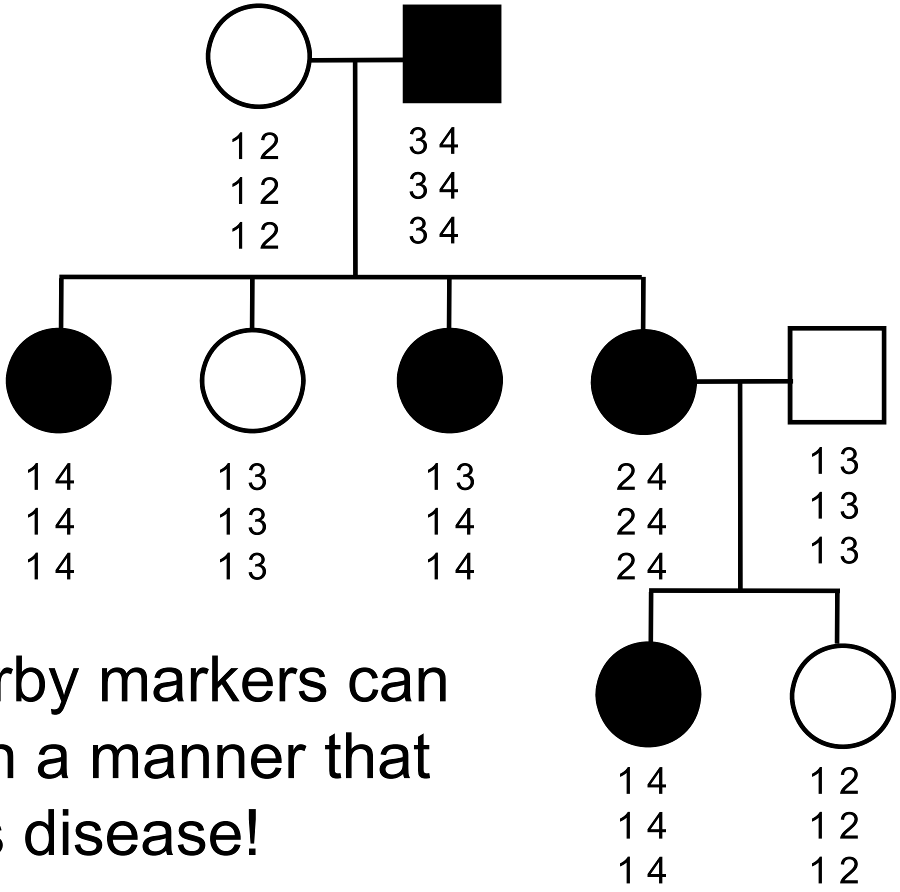
Segregation pattern for chromosome carrying disease alleles

Tracing Chromosomes



Segregation of a specific marker near the disease locus

Tracing Chromosomes



Multiple nearby markers can segregate in a manner that tracks disease!

Today ...

- Linkage analysis with sibling pairs

- Find markers that are near disease locus
  - Near means recombination fraction $\theta < \frac{1}{2}$

- Minimalist approach …

Bishop and Williamson (1990) Opening Line

"The availability of a large number of DNA markers has made possible mapping projects with the certainty that if:

(a) a major gene exists for a trait;

(b) the trait is reasonably homogeneous;

(c) there is sufficient family material available;

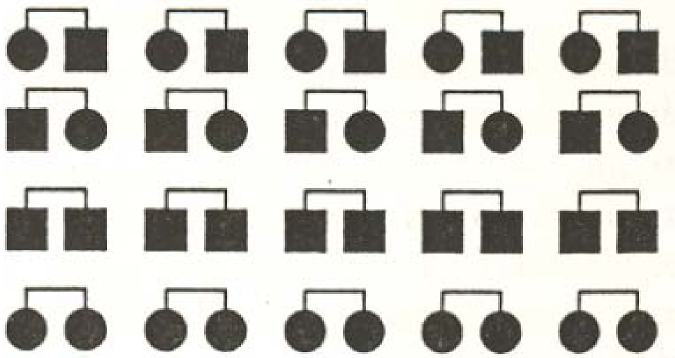
then a linked marker can be found."

Data for a Linkage Study: Minimalist Approach

- Pedigree
  - Two individuals of known relationship
- Observed Marker Genotypes
  - A single marker
- Phenotypes
  - Both individuals are affected

Allele Sharing Analysis

- Are affected pairs more similar than expected?
- Less powerful than analysis of larger pedigrees
- Does not require disease model to be specified



Consider
Autosomal Recessive Locus …

- For a collection of sibling pairs…

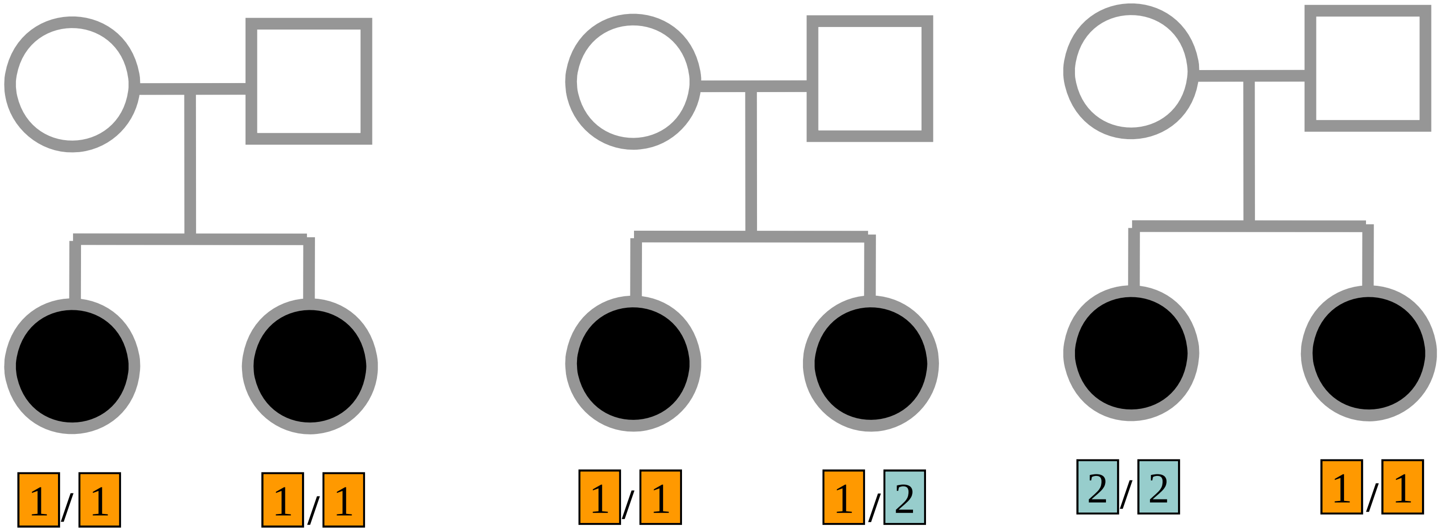
- What patterns of sharing do you expect at the disease locus?

- What patterns of sharing to you expect as you move away from the disease locus?

IBS Based Methods

- Sample of affected relative pairs
- Examine a marker of interest
- Count alleles shared for each pair
  - This includes both …
  - Chromosomes that are identical-by-descent
  - Chromosomes that simply carry identical alleles

Examples of IBS States

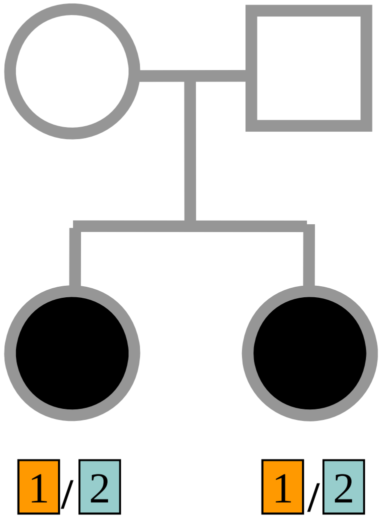


IBS = 2

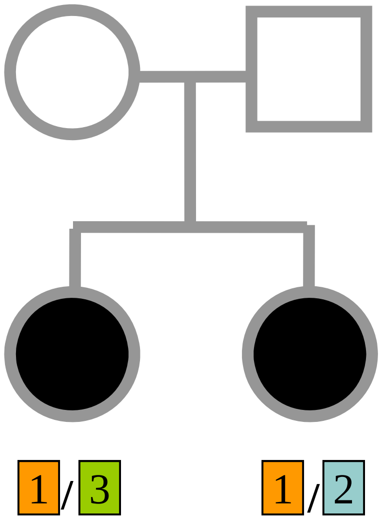
IBS = 1

IBS = 0

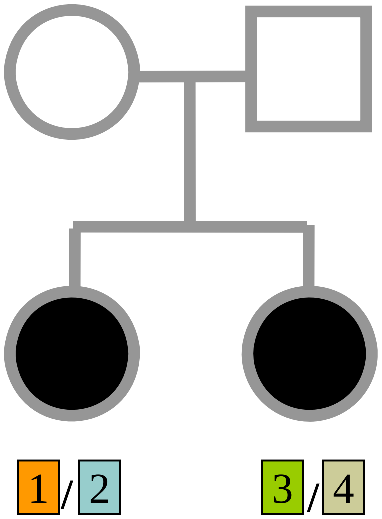
Examples of IBS States



IBS = 2



IBS = 1



IBS = 0

Evidence for Linkage

- Increased similarity in affected pairs

- Compared to:
  - Unselected pairs
  - Unaffected pairs
  - Discordant pairs
  - Expectations derived from allele frequencies

Possible Statistics

$$\chi^2_{2df} = \sum_i \frac{[N_{IBS=i} - E(N_{IBS=i})]^2}{E(N_{IBS=i})}$$

(general test, for sibling pairs)

$$\chi^2_{1df} = \frac{[N_{IBS=0} - E(N_{IBS=0})]^2}{E(N_{IBS=0})} + \frac{[N_{IBS>0} - E(N_{IBS>0})]^2}{E(N_{IBS>0})}$$

(grouping often preferable for other relatives)

- Assuming all counts are relatively large
- If counts are small, use binomial or trinomial distribution

Calculating Expected IBS

- For any relative pair, calculate:

1. Probability of IBD sharing
   - 0, 1 or 2 alleles

2. Conditional probability of IBS sharing
   - 0, 1, 2 alleles

3. IBS sharing >= IBD sharing
   - Why?

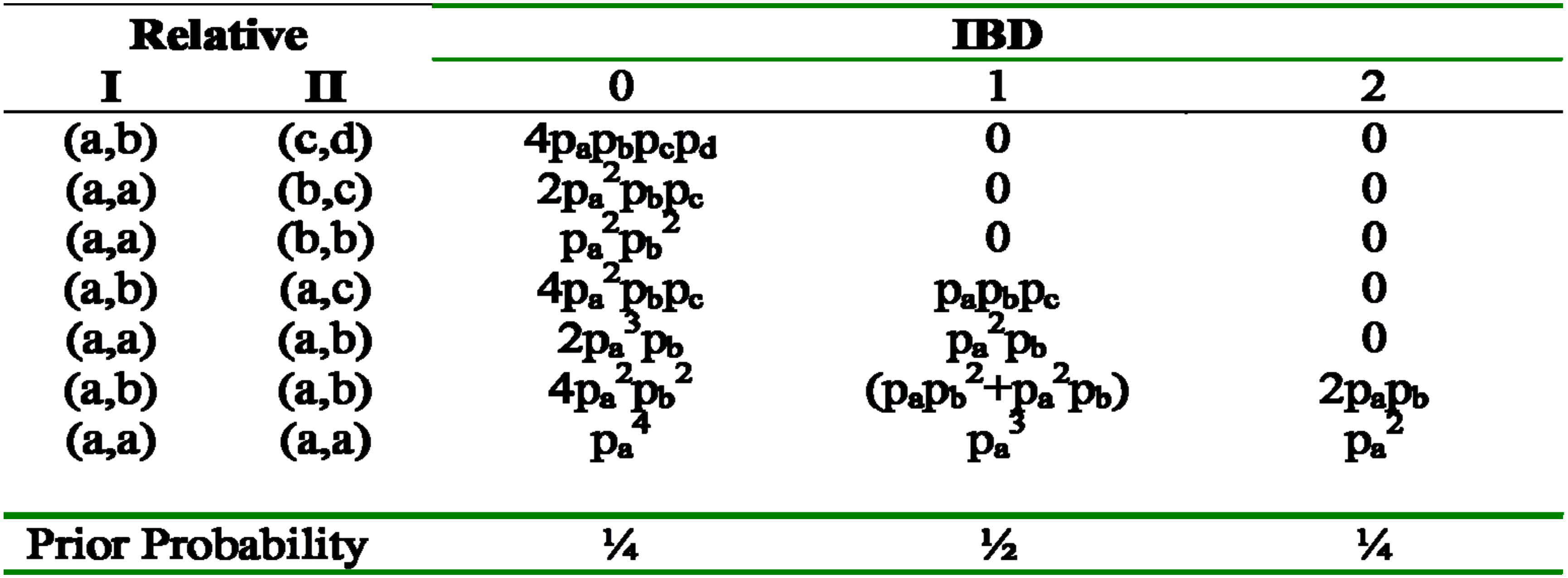
IBD

- The underlying sharing of chromosomes segregating within a family

- Siblings share 0, 1 or 2 alleles
  - Probabilities ¼, ½ and ¼

- Unilineal relatives share 0 or 1 alleles

P(Marker Genotype|IBD State)



Relative		IBD		
I	II	0	1	2
(a,b)	(c,d)	$4p_ap_bp_cp_d$	0	0
(a,a)	(b,c)	$2p_a^2p_bp_c$	0	0
(a,a)	(b,b)	$p_a^2p_b^2$	0	0
(a,b)	(a,c)	$4p_a^2p_bp_c$	$p_ap_bp_c$	0
(a,a)	(a,b)	$2p_a^3p_b$	$p_a^2p_b$	0
(a,b)	(a,b)	$4p_a^2p_b^2$	$(p_ap_b^2+p_a^2p_b)$	$2p_ap_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

Note: Assuming alleles unordered within genotypes

Example, Assuming Equal Allele Frequencies

	P(IBS=0)	P(IBS=1)	P(IBS=2)
2 alleles, IBD=0	.125	.500	.375
2 alleles, IBD=1	.000	.500	.500
3 alleles, IBD=0	.222	.592	.185
3 alleles, IBD=1	.000	.666	.333

IBS Probabilities

No. of Alleles	P(IBS=0)	P(IBS=1)	P(IBS=2)
2	.03	.37	.60
3	.05	.48	.47
4	.08	.51	.40
20	.21	.52	.27
∞	.25	.50	.25

Sibling IBS as a function of allele count, for marker with equally frequent alleles

Inference from Example

- IBS approaches IBD as number of alleles increases

- If linkage is being tested with chi-square test, how does the number of alleles (and marker informativeness) affect these two tests:
  - A test of whether $N_{IBS >= 1}$ increases?
  - A test of whether $N_{IBS > 1}$ increases?

Results of
Bishop and Williamson (1990)

- Effect size, P(IBS | Affected pair)
- Number of alleles at marker
- Different relationships
- Recombination fraction

More Alleles Increase Power

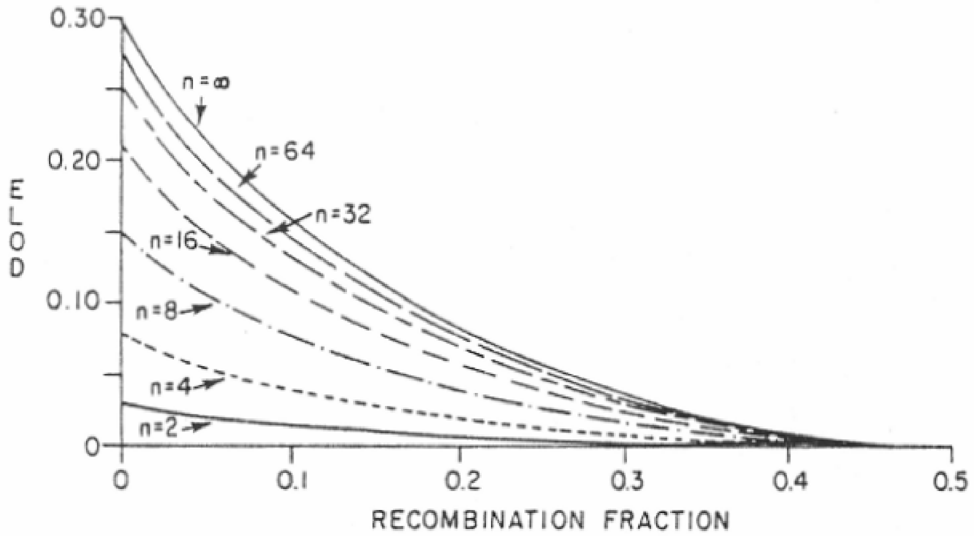


Figure 3 Variation in ELOD as a function of *n*, the number of alleles at the marker locus. All alleles are assumed to have frequency 1/*n*. This calculation is performed for the grandparent-grandchild relationship with a rare trait allele frequency.

Effect of Recombination Varies According to Relationship

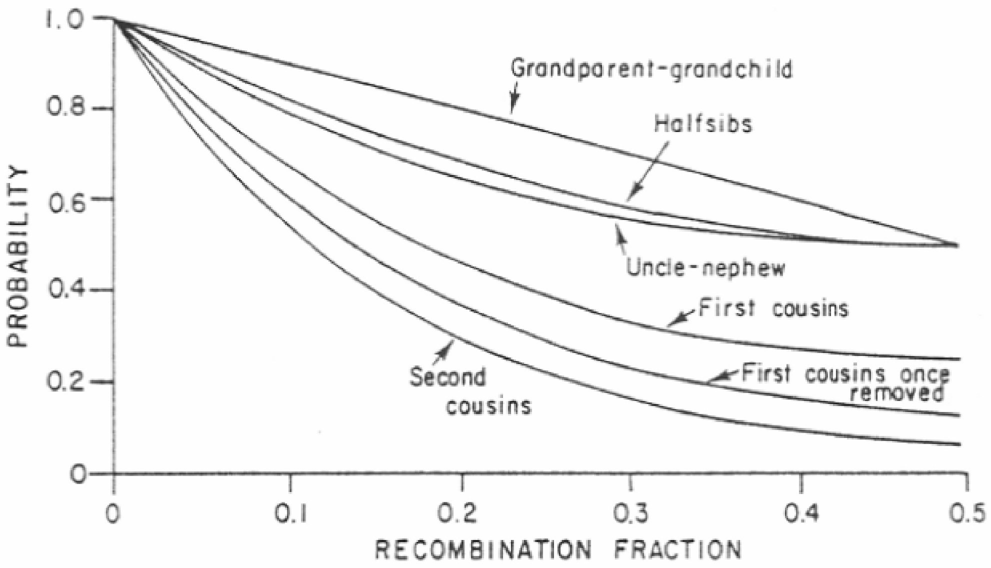


Figure 2 Probability of i.b.d. at a second linked locus conditional on i.b.d. at an index locus, as a function of the recombination fraction r between the loci, for specific genetic relationships. This function is $d_{11}(r)$ in the notation of table 1.

With no phenocopies, rare alleles are easier to map

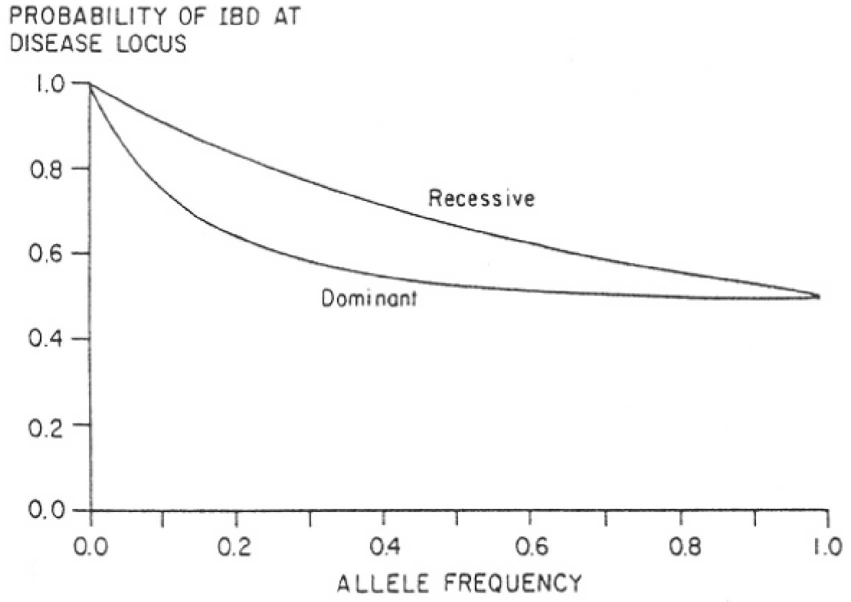


Figure 1 Probability of i.b.d. at a trait locus for two affected related individuals, as a function of the mode of inheritance of the trait. This figure is computed for the relationships with $\phi = .125$.

In general, phenocopies decrease power

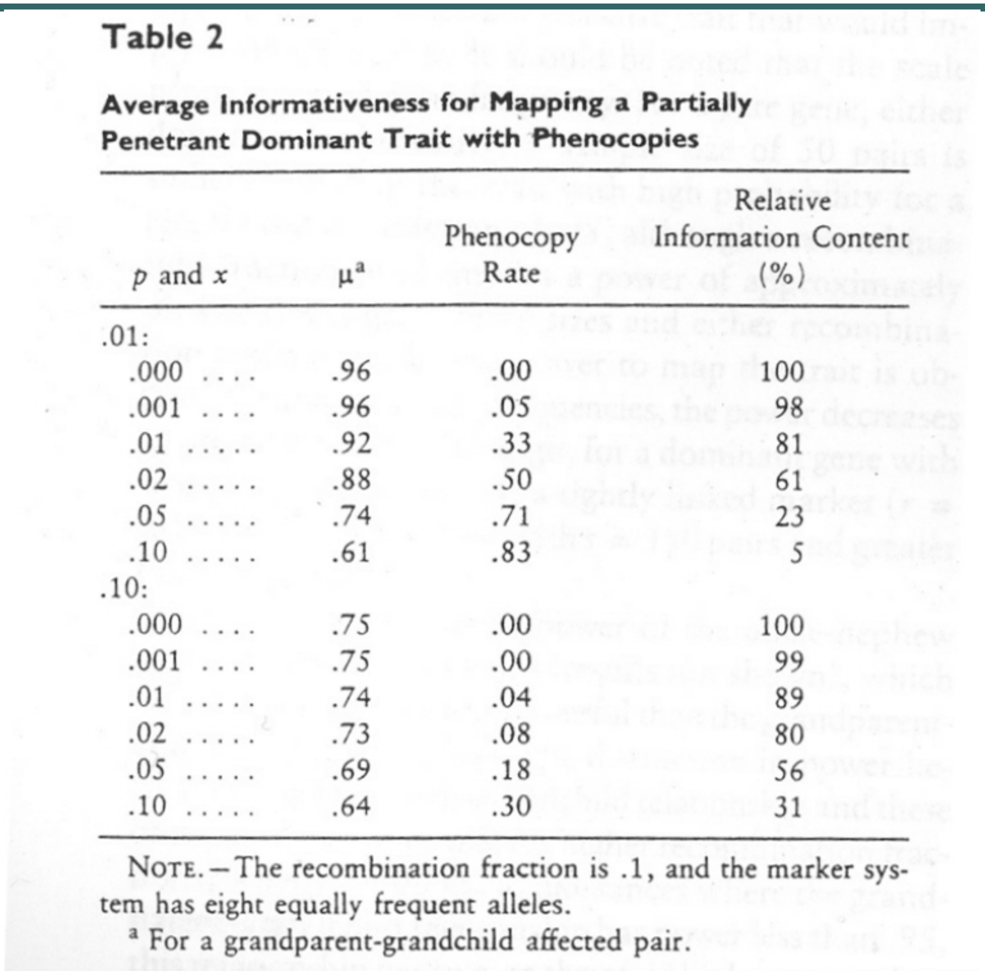


Table 2

Average Informativeness for Mapping a Partially Penetrant Dominant Trait with Phenocopies

p and *x*	μ[a]	Phenocopy Rate	Relative Information Content (%)
.01:			
.000	.96	.00	100
.001	.96	.05	98
.01	.92	.33	81
.02	.88	.50	61
.05	.74	.71	23
.10	.61	.83	5
.10:			
.000	.75	.00	100
.001	.75	.00	99
.01	.74	.04	89
.02	.73	.08	80
.05	.69	.18	56
.10	.64	.30	31

NOTE.—The recombination fraction is .1, and the marker system has eight equally frequent alleles.

[a] For a grandparent-grandchild affected pair.

Shortcomings of IBS Method

- All sharing is weighted equally
  - Sharing a rare allele
  - Sharing a common allele
  - Sharing homozygous genotype
  - Sharing heterozygous genotype

- Inefficient.

An Alternative, Likelihood Based Formulation

- Depends on three parameters z_0, z_1, z_2
  - Probability of sharing 0, 1 and 2 alleles IBD
- Under the null, determined by relationship
- Under the alternative, determined by genetic model

An Alternative, Likelihood Based Formulation

Under the null hypothesis:

$$L = \left(1/4\right)^{n_{IBD0}} \left(1/2\right)^{n_{IBD1}} \left(1/4\right)^{n_{IBD2}}$$

Under the alternative hypothesis

$$L = \left(\hat{z}_0\right)^{n_{IBD0}} \left(\hat{z}_1\right)^{n_{IBD1}} \left(\hat{z}_2\right)^{n_{IBD2}}$$

Maximum Likelihood Based Linkage Tests …

- Evaluate likelihood at null hypothesis
- Evaluate likelihood at MLE
- Compare alternatives using likelihood ratio test

Commonly Used Test Statistics

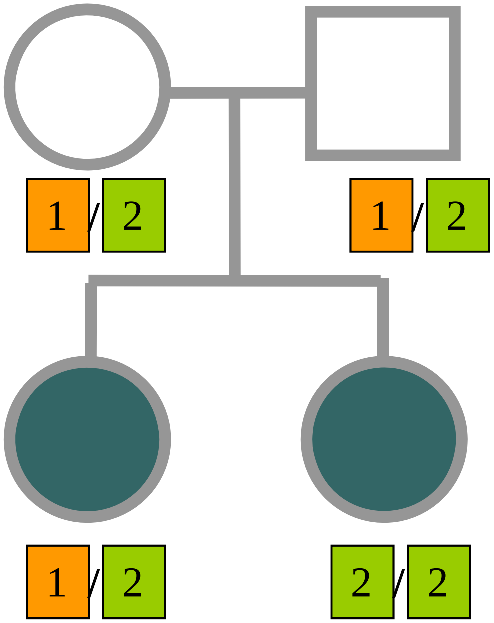
$$LOD = \log_{10} \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})}$$

$$\chi^2 = 2\ln \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})}$$

$$= 2\ln L(\hat{z}_0, \hat{z}_1, \hat{z}_2) - 2\ln L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})$$

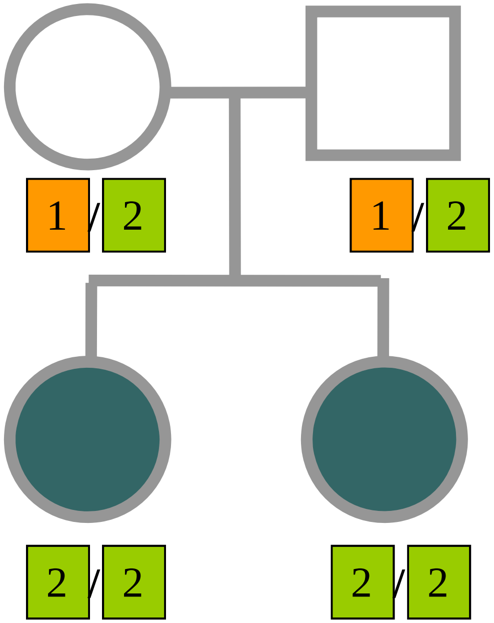
Example

5x



IBD=1

5x



IBD=2

Example

- Assume that 10 sib-pairs are examined
  - 5 share 2 alleles IBD
  - 5 share 1 allele IBD

- Calculate likelihood for null
- Calculate MLEs
- Calculate LOD score
- Evaluate LOD for each pair

In real life…

- Markers are only partially informative
- IBD sharing is equivocal
  - Some uncertainty removed by examining relatives
- Need an alternative likelihood
  - Should allow for partially informative data

Desirable Properties

- Models IBD probabilities z_0, z_1, z_2
  - Probability of sharing 0, 1 and 2 alleles IBD

- Uses partial information on IBD sharing

- For unambiguous data, equivalent to previous likelihood

For A Single Family

$$L_i = \sum_{j=0}^{2} P(IBD = j \mid ASP) P(Genotypes_i \mid IBD = j) = \sum_{j=0}^{2} z_j w_{ij}$$

Risch (1990) defines

$$w_{ij} = P(Genotypes_i \mid IBD = j)$$

We only need proportionate w_{ij}

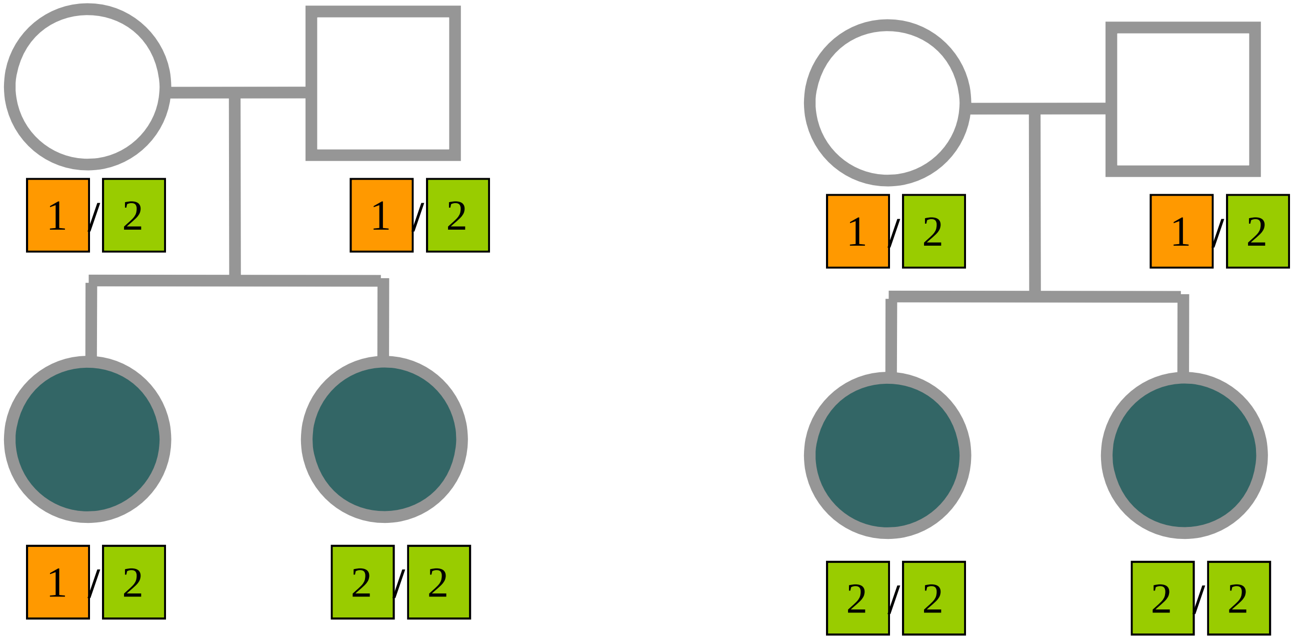
Likelihood and LOD Score

$$L(z_0, z_1, z_2) = \prod_i \sum_j z_j w_{ij}$$

$$LOD = \log_{10} \prod_i \frac{\hat{z}_0 w_{i0} + \hat{z}_1 w_{i1} + \hat{z}_2 w_{i2}}{\frac{1}{4} w_{i0} + \frac{1}{2} w_{i1} + \frac{1}{4} w_{i2}}$$

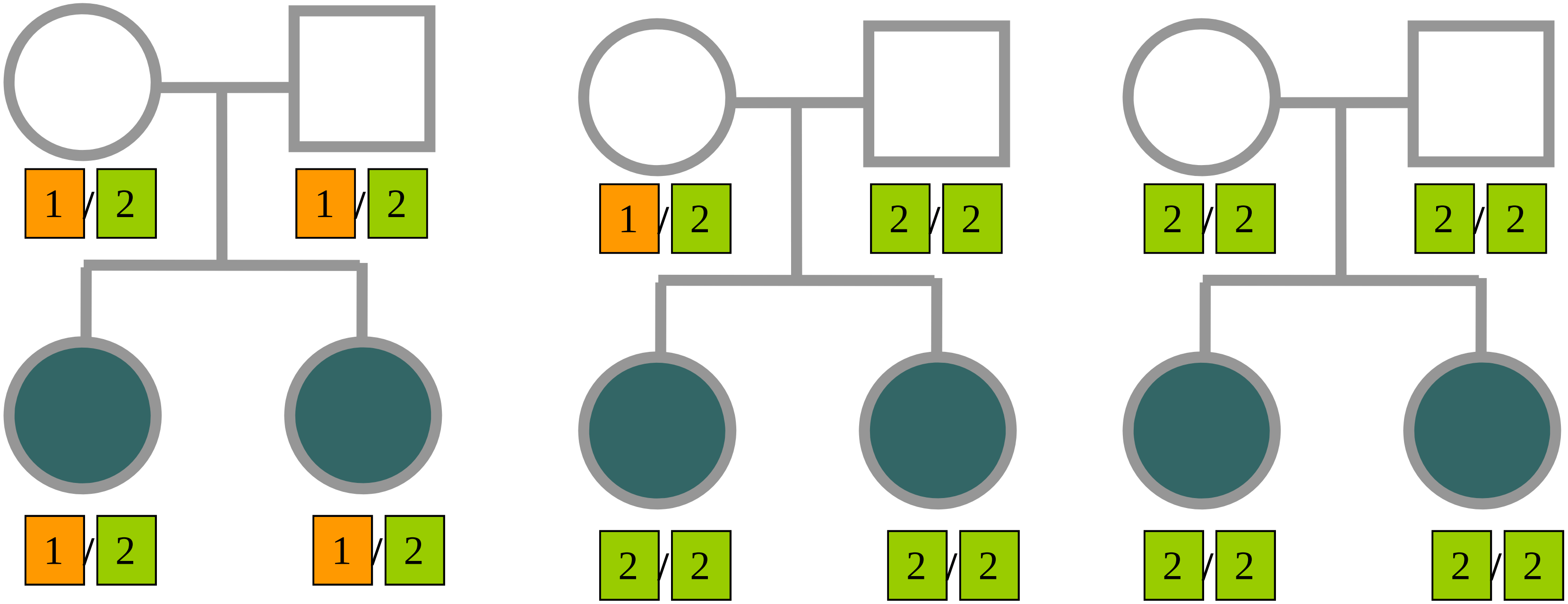
The MLS statistic is the LOD evaluated at the MLEs of z_0, z_1, z_2

Example: Scoring of w_{ij}



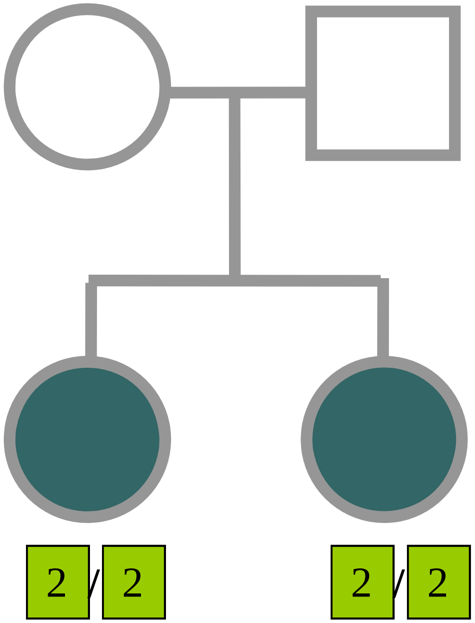
In this case, only one of the weights is non-zero for each family.

More interesting examples: w_{ij}



In these cases, multiple weights are non-zero (but equal) for each family.

More interesting examples: w_{ij}



In this case, relative weights depend on allele frequency.

How to maximize likelihood?

- If all families are informative
  - Use sample proportions of IBD=0, 1, 2
- If some families are uninformative
  - Use an E-M algorithm
  - At each stage generate complete dataset with fractional counts
  - Iterate until estimates of LOD and z parameters are stable

Assigning Partial Counts in E-M

$$P(IBD = j \mid Genotypes) =$$

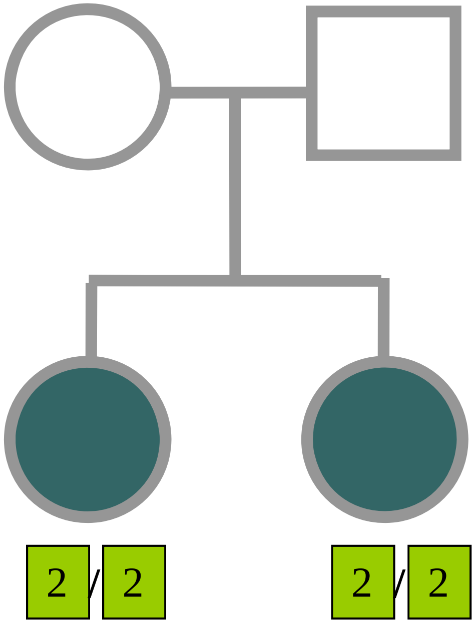
$$= \frac{P(IBD = j \mid ASP)P(Genotypes \mid IBD = j)}{L_i}$$

$$= \frac{P(IBD = j \mid ASP)P(Genotypes \mid IBD = j)}{\sum_{k=0}^{2} P(IBD = k \mid ASP)P(Genotypes \mid IBD = k)}$$

$$= \frac{z_j w_{ij}}{\sum_{k=0}^{2} z_k w_{ik}}$$

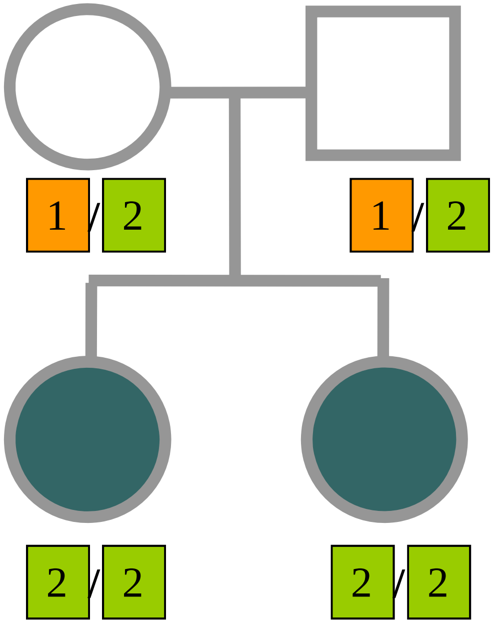
Example

5x



IBD=?

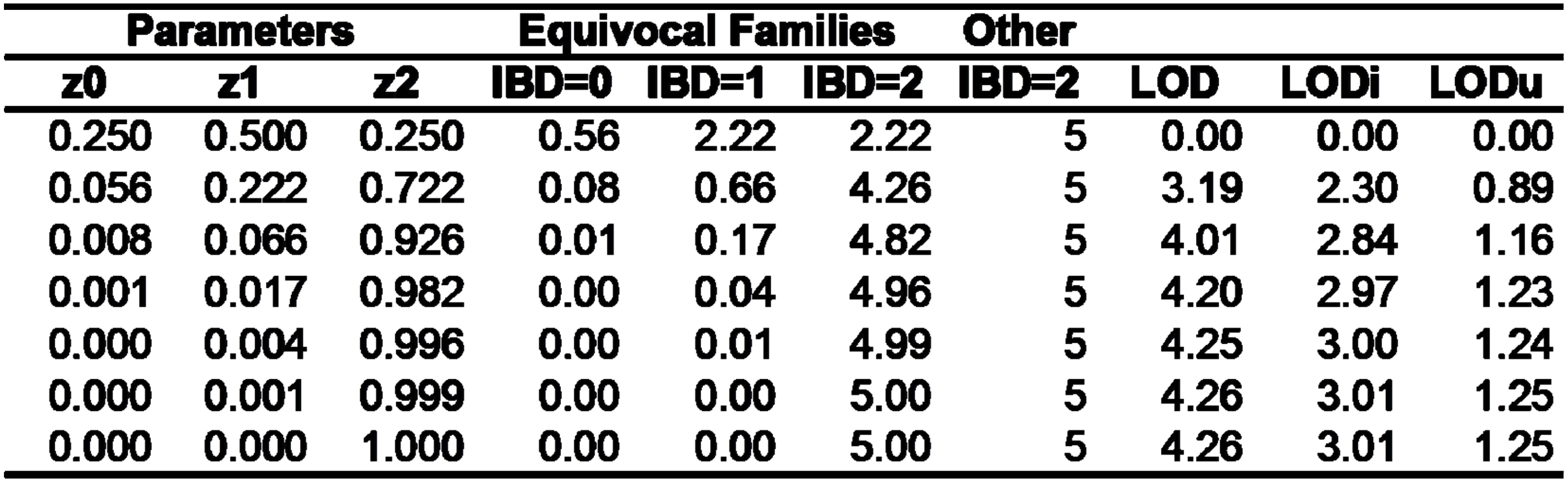
5x



IBD=2

Assume a bi-allelic marker where the two alleles have identical frequencies.

Example of E-M Steps

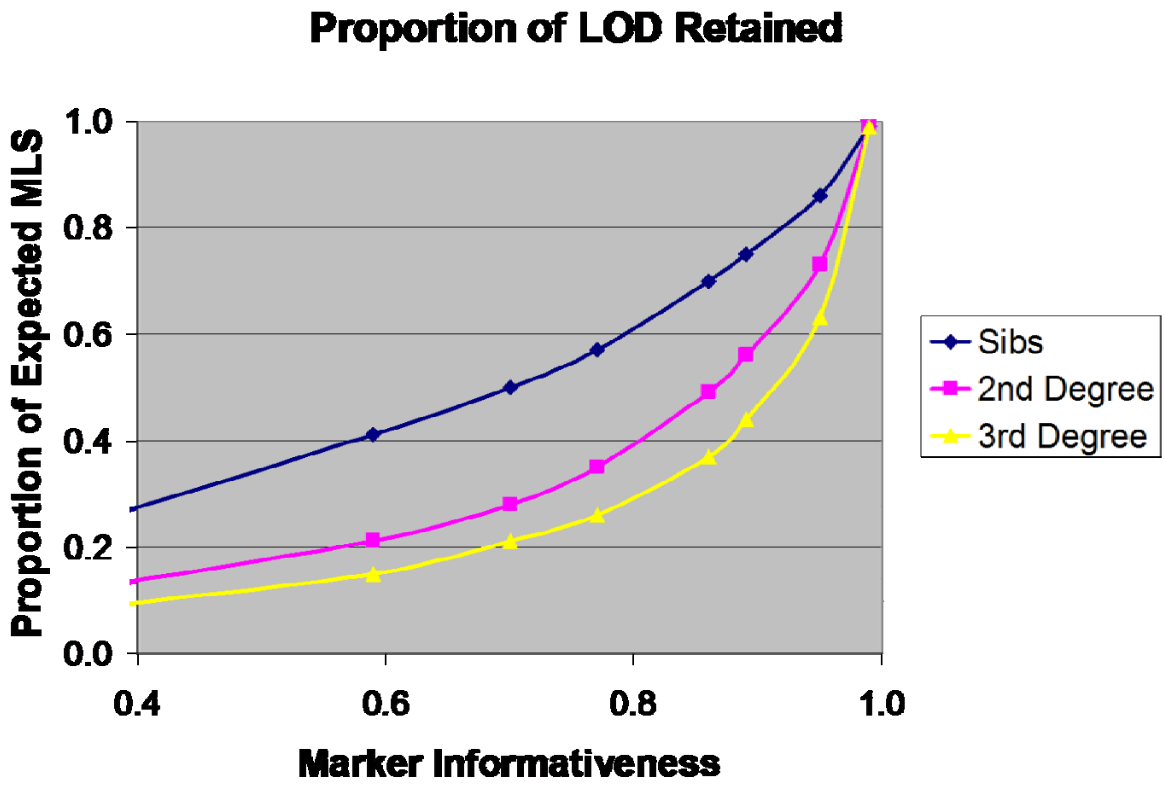


Parameters			Equivocal Families			Other			
z0	z1	z2	IBD=0	IBD=1	IBD=2	IBD=2	LOD	LODi	LODu
0.250	0.500	0.250	0.56	2.22	2.22	5	0.00	0.00	0.00
0.056	0.222	0.722	0.08	0.66	4.26	5	3.19	2.30	0.89
0.008	0.066	0.926	0.01	0.17	4.82	5	4.01	2.84	1.16
0.001	0.017	0.982	0.00	0.04	4.96	5	4.20	2.97	1.23
0.000	0.004	0.996	0.00	0.01	4.99	5	4.25	3.00	1.24
0.000	0.001	0.999	0.00	0.00	5.00	5	4.26	3.01	1.25
0.000	0.000	1.000	0.00	0.00	5.00	5	4.26	3.01	1.25

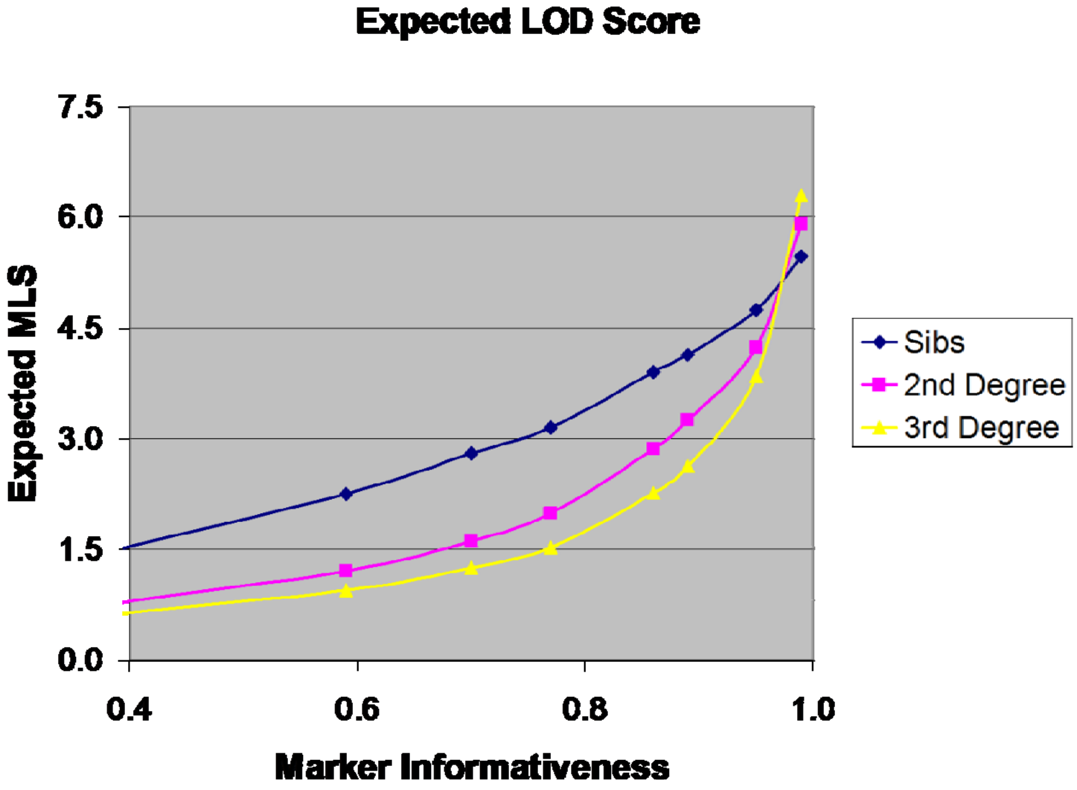
Properties of Pair Analyses Explored by Risch

- Effect of marker informativeness
- Effect of adding relative genotypes
- Size of genetic effect
- Degree of relationship

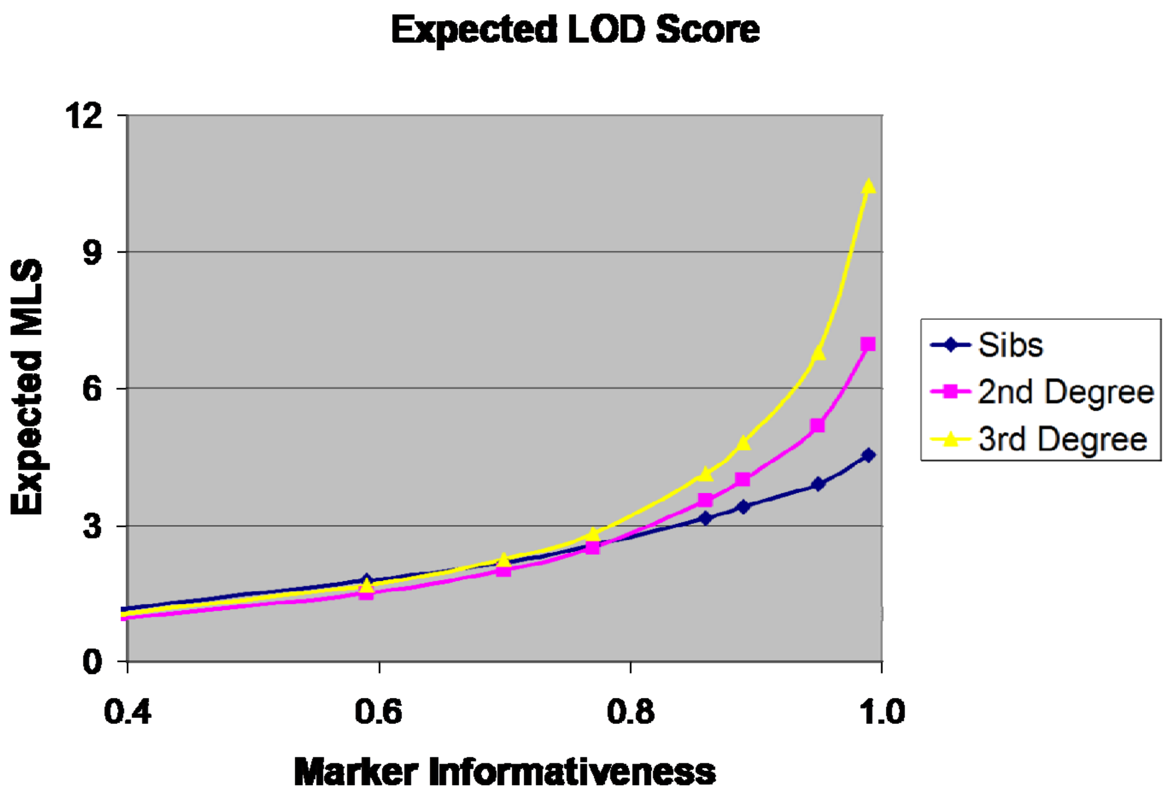
Marker Informativeness



Marker Informativeness
Gene of Modest Effect (λ_O=3)



Marker Informativeness
Gene of Larger Effect (λ_O=10)



Genotypes of Other Family Members

- Genotyping only pair decreseas LOD score by
  - Up to 33% if only sib-pairs are typed
  - Up to 60% for second degree relatives
  - Up to 70% for third degree relatives

- Genotyping effort decreases by
  - 50% if only sib-pairs are typed
  - 60% if only second degree relatives typed
  - 75% if only third degree relatives typed

Recommended Reading

- Bishop DT and Williamson JA (1990) *Am J Hum Genet* **46:**254-265

- Good introduction to linkage analysis in affected relative pairs, discusses
  - Marker choice
  - Recombination fraction
  - Disease model
  - Type of relative pair

Recommended Reading

- Risch (1990)
  - Linkage Strategies for Genetically Complex Traits. III. The Effect of Marker Polymorphism on Analysis of Affected Relative Pairs
  - *Am J Hum Genet* **46:**242-253
- Introduces MLS method for linkage analysis
  - Still, one of the best methods for analysis pair data
- Evaluates different sampling strategies
  - Results were later corrected by Risch (1992)