Replacing IBS with IBD: The MLS Method

Biostatistics 666

Lecture 15

Previous Lecture

- Analysis of Affected Relative Pairs
- Test for Increased Sharing at Marker
- Expected Amount of IBS Sharing

Previous Lecture: Expected IBS Sharing

- Calculated probability of IBS for each IBD state
- Probability of IBD state depends on relationship
  - Under the null hypothesis of no linkage

$$P(IBS = i \mid R) = \sum_{j=0}^{2} P(IBS = i \mid IBD = j) P(IBD = j \mid R)$$

Shortcomings of IBS Method

- All sharing is weighted equally
  - Sharing a rare allele
  - Sharing a common allele
  - Sharing homozygous genotype
  - Sharing heterozygous genotype
- Inefficient.
  - Data contains additional information that is being ignored.

Today

- A likelihood based approach
- Evaluate linkage in fully informative pairs
- An E-M algorithm for practical settings
- MLS method, Risch (1990)

Simple Case

- If IBD could be observed
- Each pair of individuals scored as
  - IBD=0
  - IBD=1
  - IBD=2
- Evaluate likelihood for null and alternative hypothesis

The Model

- 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

Sib Pair Likelihood (Fully Informative Data)

Under the null hypothesis:

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

Under the alternative hypothesis

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

Testing for Linkage

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

Commonly Used Test Statistics

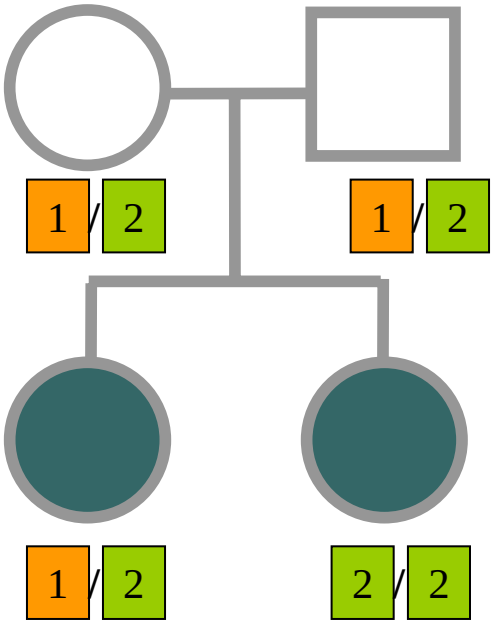
$$LOD = \log_{10} \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \tfrac{1}{4}, z_1 = \tfrac{1}{2}, z_2 = \tfrac{1}{4})}$$

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

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

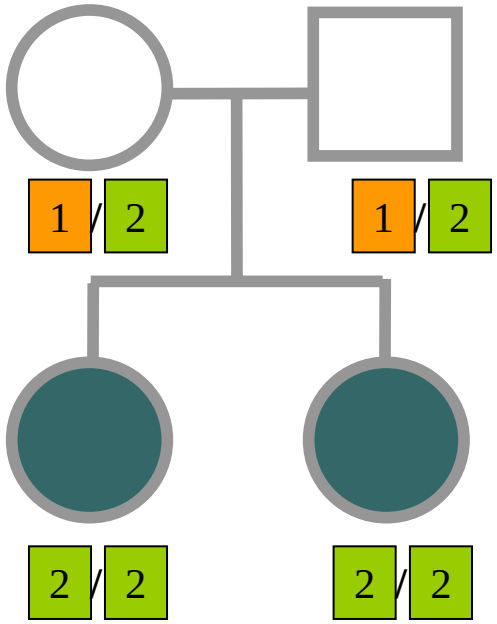
Example

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

- Also depends on parameters z_0, z_1, z_2
  - Probability of sharing 0, 1 and 2 alleles IBD
- Can incorporate partial information on IBD sharing
- For fully informative 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}

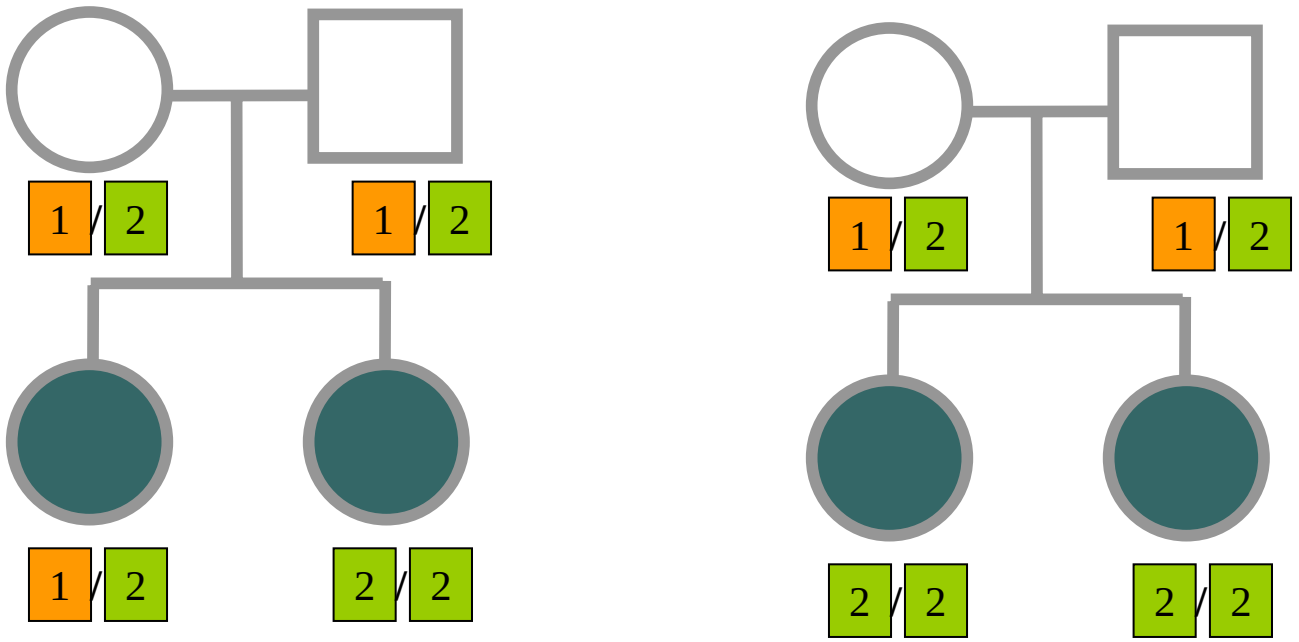
Likelihood and LOD Score

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

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

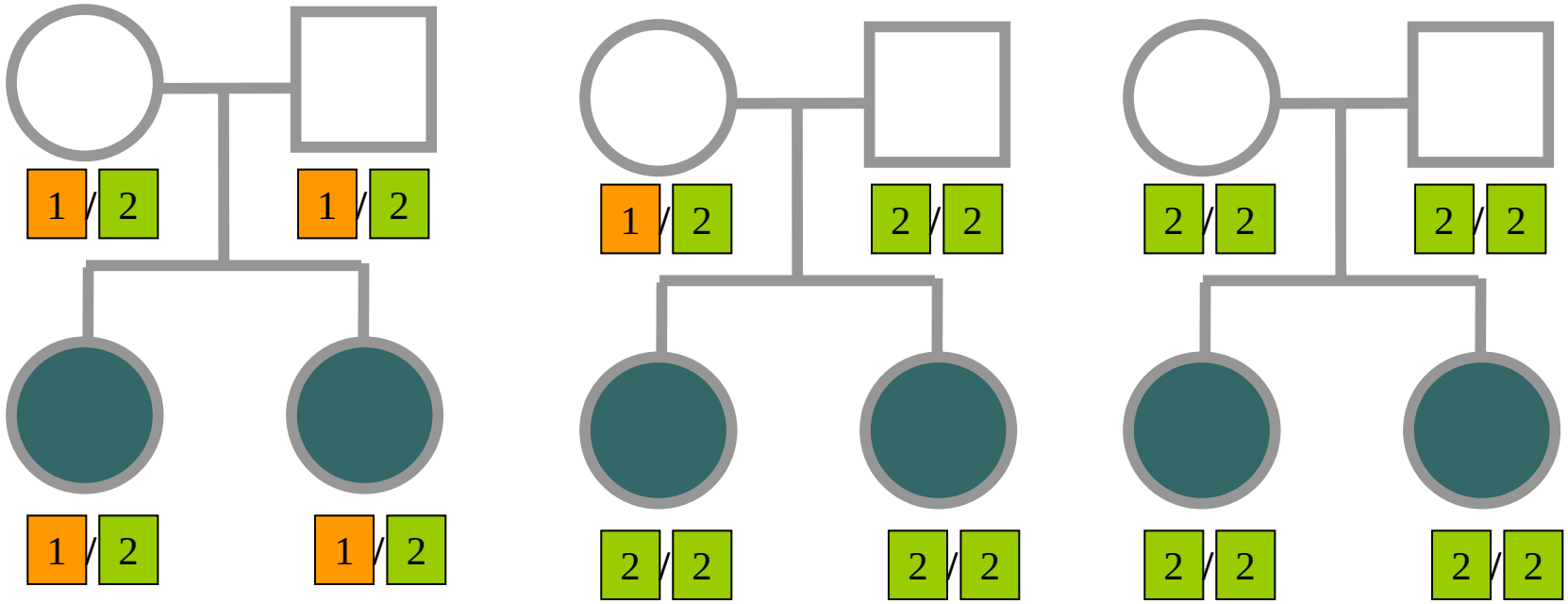
The MLS statistic is the LOD evaluated at the MLEs of z_0, z_1, z_2

Example: Scoring of w_{ij}



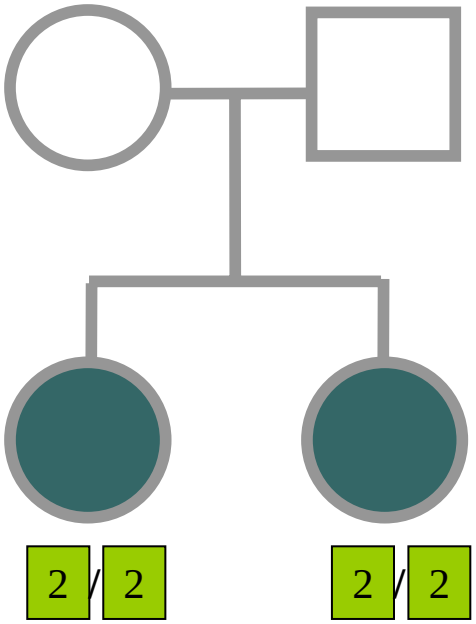
In this case, only one of the weights is non-zero for each family.

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

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

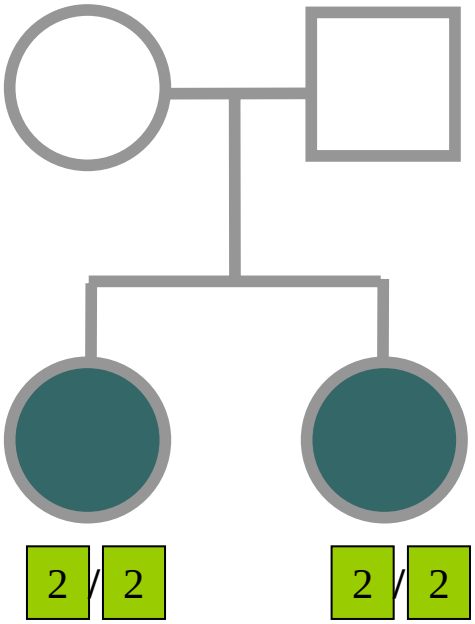
$$= \frac{P(IBD = j \mid ASP)P(Genotypes \mid IBD = j)}{L_i}$$

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

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

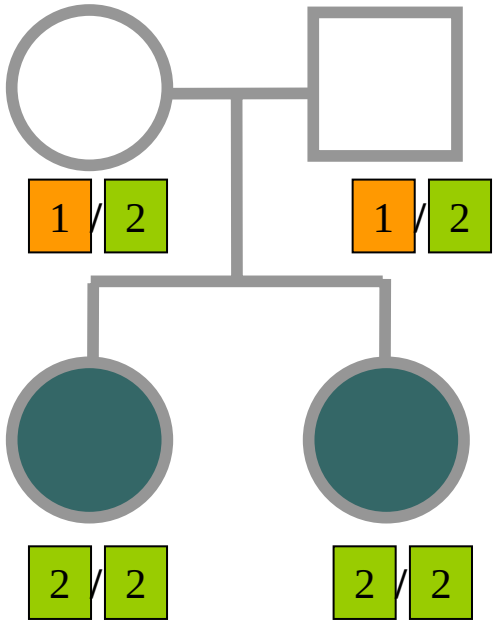
Example

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

PIC:
Measure of Marker Informativeness

- Probability that alleles of parent can be distinguished in offspring
  - Botstein et al, 1980.
    - Markers that could track dominant alleles
- Probability that parent is heterozygous and informative in relation to spouse

PIC – Definition

- In general:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2(p_i p_j)^2$$

- For *a* equally frequent alleles

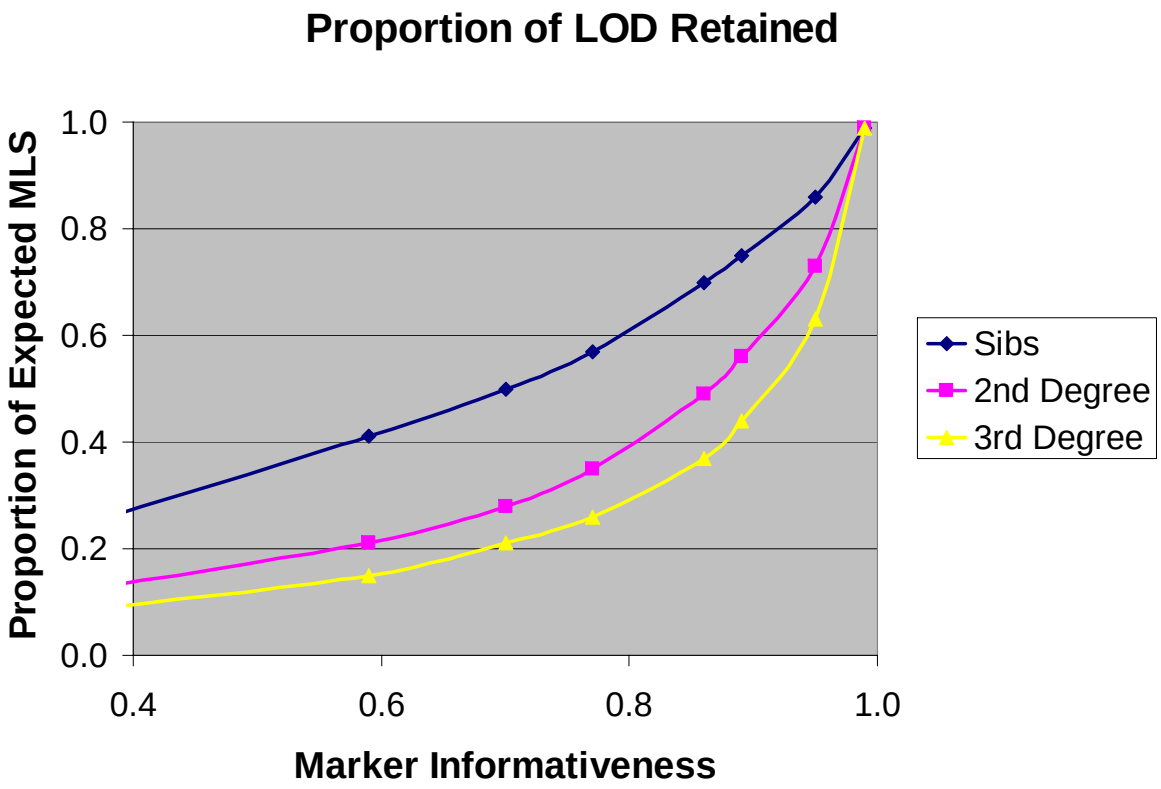
$$PIC = \frac{a-1}{a} - \frac{a-1}{a^3}$$

- PIC <= Heterozygosity

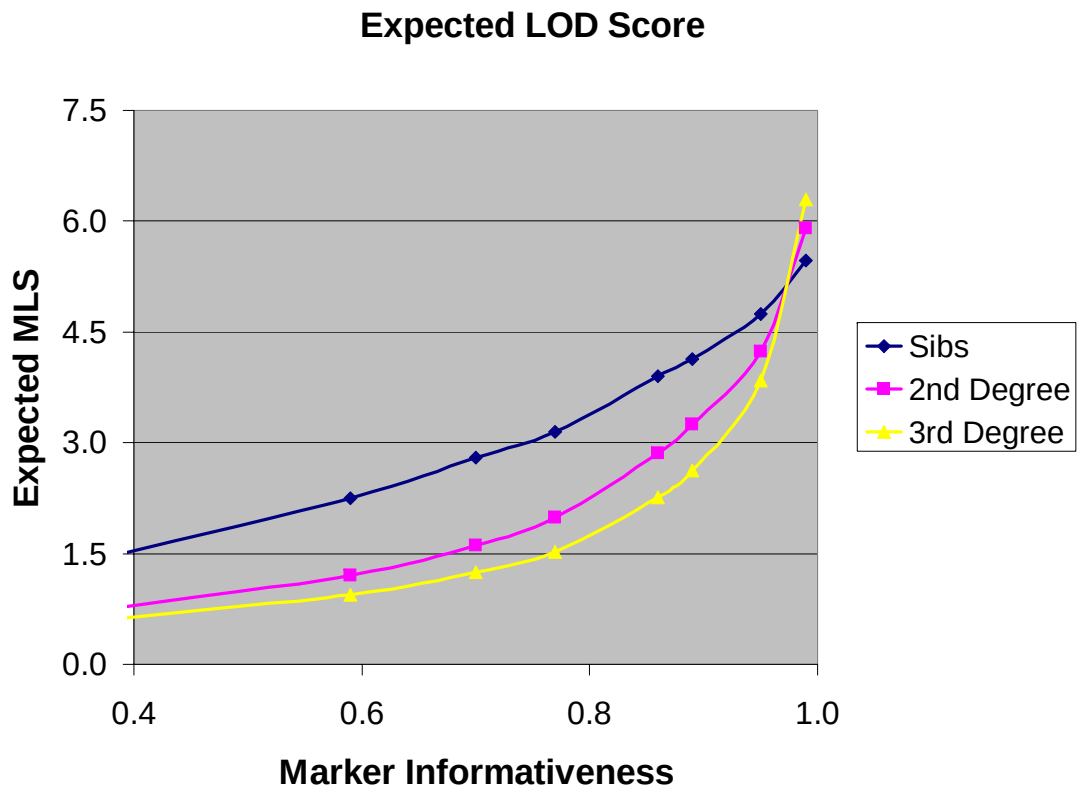
Some PICs and Heterozygosities

Alleles	PIC	H
2	0.38	0.50
3	0.59	0.67
4	0.70	0.75
5	0.77	0.80
8	0.86	0.88
10	0.89	0.90
20	0.95	0.95

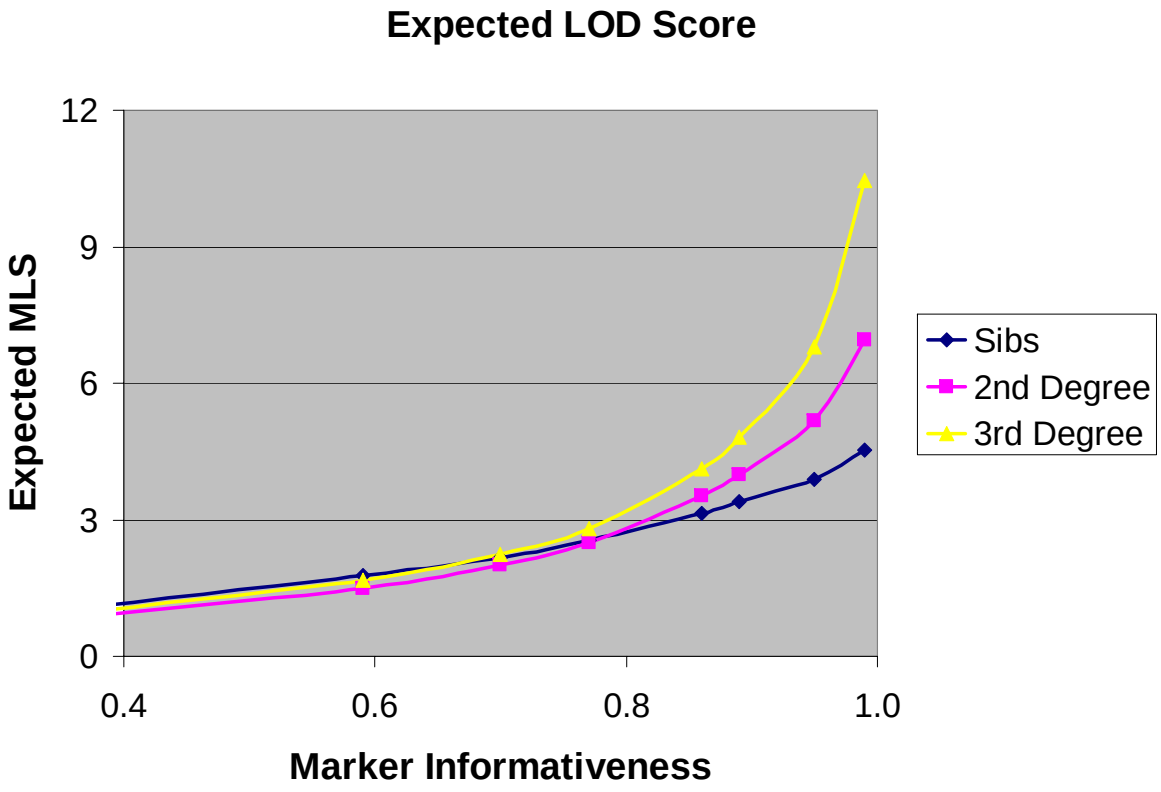
Marker Informativeness



Marker Informativeness
Gene of Modest Effect (λ_O=3)



Marker Informativeness
Gene of Larger Effect (λ_O=10)



Genotypes of Other Family Members

- Expected LOD score decreases
  - by < 33% if only sib-pairs are typed
  - by < 60% for second degree relatives
  - by < 70% for third degree relatives
- Genotyping effort decreases by
  - by 50% if only sib-pairs are typed
  - by 60% if only second degree relatives typed
  - by 75% if only third degree relatives typed

Quick Comment on Literature

- Greenwood and Schork (2004) suggested that uninformative families could bias MLS
- However, their results use a poor estimate for MLEs
- If an E-M algorithm is used, there is no problem

Today …

- Describe a likelihood model based on IBD sharing for pairs of individuals
- Model accommodates partially informative families
- Maximum LOD score can be calculated using an E-M algorithm

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)

Recommended Reading

- Risch (1992)
  - Corrections to
    Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs.
  - *Am J Hum Genet* **51:**673-675

- Evaluates utility of parental genotype data