Affected Sibling Pairs

Biostatistics 666

Today …

- Discussion of linkage analysis using affected sibling pairs

- Our exploration will include several components we have seen before:
  - A simple disease model
  - IBD sharing probabilities
  - Maximum likelihood
  - The E-M algorithm
  - A Hidden Markov model

- Linkage analysis with sibling pairs using Risch's Maximum LOD Score (MLS)

- Distribution of IBD in affected sibling pairs and Holman's "Possible Triangle Constraint"

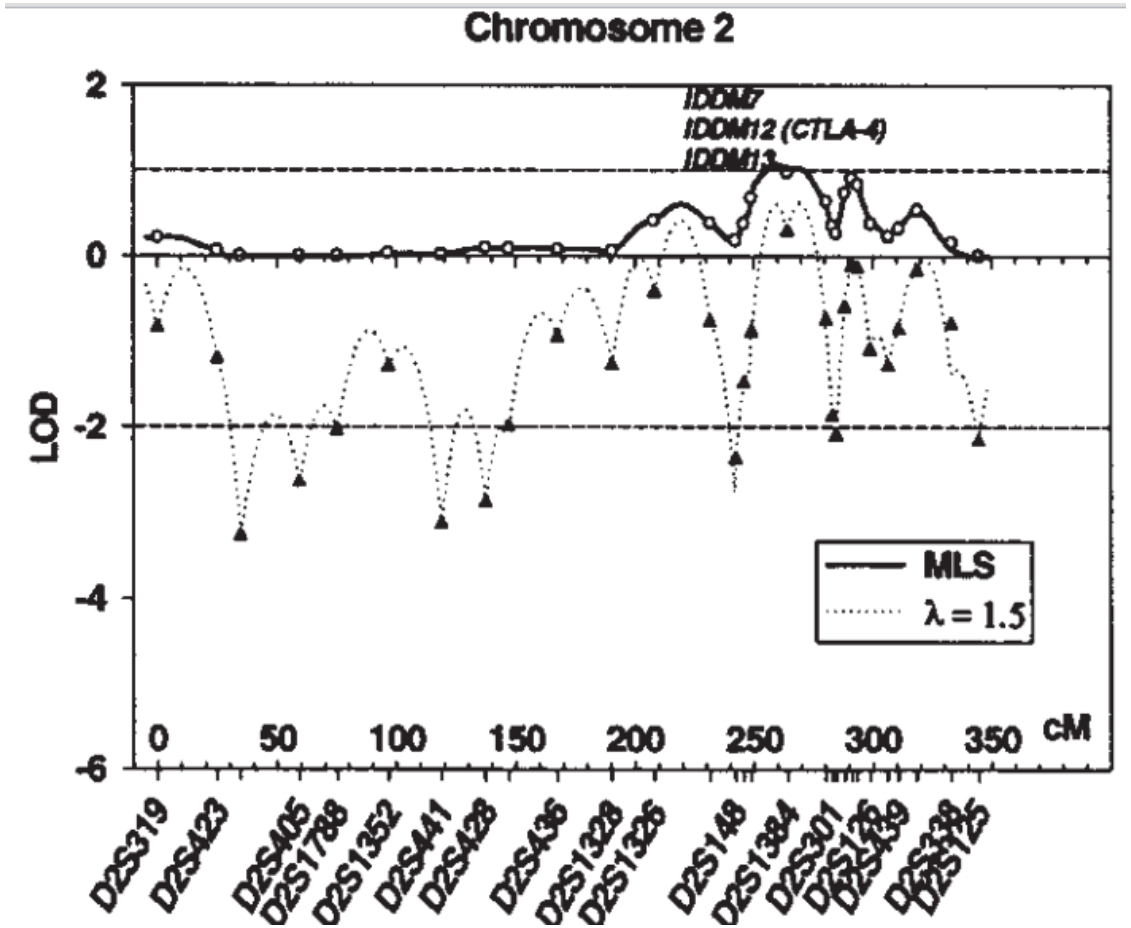
Examplar Linkage Study

- Concannon et al (1998) Nature Genetics, **19:**292-296

- Affected sibling pair study of type 1 diabetes
  - Chronic disease affecting ~1 in 250 children
  - 292 affected sibpairs for initial screen
  - 467 affected sibpairs for follow-up

- Highest LOD score reaches 34.2 near HLA on chr. 6
  - At this locus, chromosomes carried by affected sibs are identical in 73% of pairs.

Examplar Linkage Study Results



Concannon et al (1998) *Nature Genetics*, **19:**292-296

Single Locus Disease Model

- Allele frequencies *p* and *(1-p)*
- Penetrances f_{11}, f_{12}, f_{22}
- Useful in exploring behavior of linkage and association tests
  - We used similar constructs to explore genetic association test power
- Simplification of reality, ignores other loci and the environment

Using Penetrances

- Allele frequency p
- Genotype penetrances f_{11}, f_{12}, f_{22}

- Prevalence
  - K =

- Probability of genotype given disease
  - P(G = ij | D) =

Pairs of Individuals

- A genetic model can predict probability of sampling different affected relative pairs

- We will first consider some simple cases:
  - Unrelated individuals
  - Parent-offspring pairs
  - Monozygotic twins

- How much genetic material do these pairs share IBD?

Unrelated Individuals

- Probability of affected pair of unrelateds

$$
\begin{aligned}
P(a \text{ and } b \text{ affected}) &= \mathrm{P}(a \text{ affected})\mathrm{P}(b \text{ affected}) \\
&= \mathrm{P}(\text{affected})^2 \\
&= \left[p^2 f_{11} + 2p(1-p) f_{12} + (1-p)^2 f_{22}\right]^2 \\
&= K^2
\end{aligned}
$$

- For related individuals, probability that both affected is greater or equal

Monozygotic Twins

- Probability of affected pair of identical twins

$$
\begin{aligned}
P(MZ \text{ pair affected}) &= \sum_G P(G)P(a \text{ affected} \mid \mathrm{G})P(b \text{ affected} \mid G) \\
&= p^2 f_{11}^2 + 2p(1-p) f_{12}^2 + (1-p)^2 f_{22}^2 \\
&= K_{MZ} K \\
&= \lambda_{MZ} K^2
\end{aligned}
$$

- K_{MZ} is prevalence among MZ twins of an affected individual
  - It is always greater than or equal to *K*

- $\lambda_{MZ} = K_{MZ} / K$ is the increase in risk for MZ twins of an affected individual
  - For any single locus disease model, it is always greater than 1

Parent Offspring Pairs

- Probability of affected parent-offspring pair

$$
\begin{aligned}
P &= P(\textit{parent} \text{ and } \textit{child} \text{ affected}) \\
&= \sum_{G_P}\sum_{G_O} P(G_P, G_O)\, f_{G_P}\, f_{G_O} \\
&= \sum_{i}\sum_{j}\sum_{k} P(i, j, k)\, f_{ij}\, f_{ik} \\
&= p^3 f_{11}^2 + (1-p)^3 f_{22}^2 + p(1-p) f_{12}^2 + 2p^2(1-p) f_{11} f_{12} + 2p(1-p)^2 f_{22} f_{12} \\
&= K_O K \\
&= \lambda_O K^2
\end{aligned}
$$

- K_o is the prevalence among offspring of an affected individual
- λ_o is the increase in risk for offspring of affected individuals, between 1 and λ_{MZ}

Point of Situation

- Probabilities of affected pairs for
  - Unrelated Individuals
  - Monozygotic Twins
  - Parent-Offspring Pairs

- Prevalences K_{MZ} and K_O among twins and offspring of affected individuals
- Relative risks λ_{MZ} and λ_O summarizing changes in risk

- How to predict K_R and λ_R for other types of relatives?

Recurrence Risks *vs* IBD

$$\lambda_{IBD=2} = \lambda_{MZ} = \frac{P(affected \mid IBD = 2 \text{ with affected relative})}{P(affected)}$$

$$\lambda_{IBD=1} = \lambda_{O} = \frac{P(affected \mid IBD = 1 \text{ with affected relative})}{P(affected)}$$

$$\lambda_{IBD=0} = 1 = \frac{P(affected \mid IBD = 0 \text{ with affected relative})}{P(affected)}$$

Affected Half-Siblings

- IBD sharing
  - 0 alleles with probability 50%
  - 1 allele with probability 50%

- This gives …

$$\lambda_H = \tfrac{1}{2}\lambda_O + \tfrac{1}{2} = \tfrac{1}{2}(\lambda_O + 1)$$

$$K_H = \tfrac{1}{2}K_O + \tfrac{1}{2}K = \tfrac{1}{2}(K_O + K)$$

Affected Sibpairs

- IBD sharing …
  - 0 alleles with probability 25%
  - 1 alleles with probability 50%
  - 2 alleles with probability 25%

- This gives …

$$\lambda_S = \tfrac{1}{4}\lambda_{MZ} + \tfrac{1}{2}\lambda_O + \tfrac{1}{4} = \tfrac{1}{4}(\lambda_{MZ} + 2\lambda_O + 1)$$

What does this have to do with linkage analysis?

- For a single locus model…
  - Siblings with IBD=0 are *like* unrelateds
  - Siblings with IBD=1 are *like* parent offspring pairs
  - Siblings with IBD=2 are *like* identical twins

- The genetic model parameters and the relative risks they imply allow us to calculate expected IBD probabilities at a disease locus …

- … and compare these to null expectations where $z_0 = ¼$, $z_1 = ½$, $z_2 = ¼$

Expected IBD sharing among affected siblings…
(at the disease locus!)

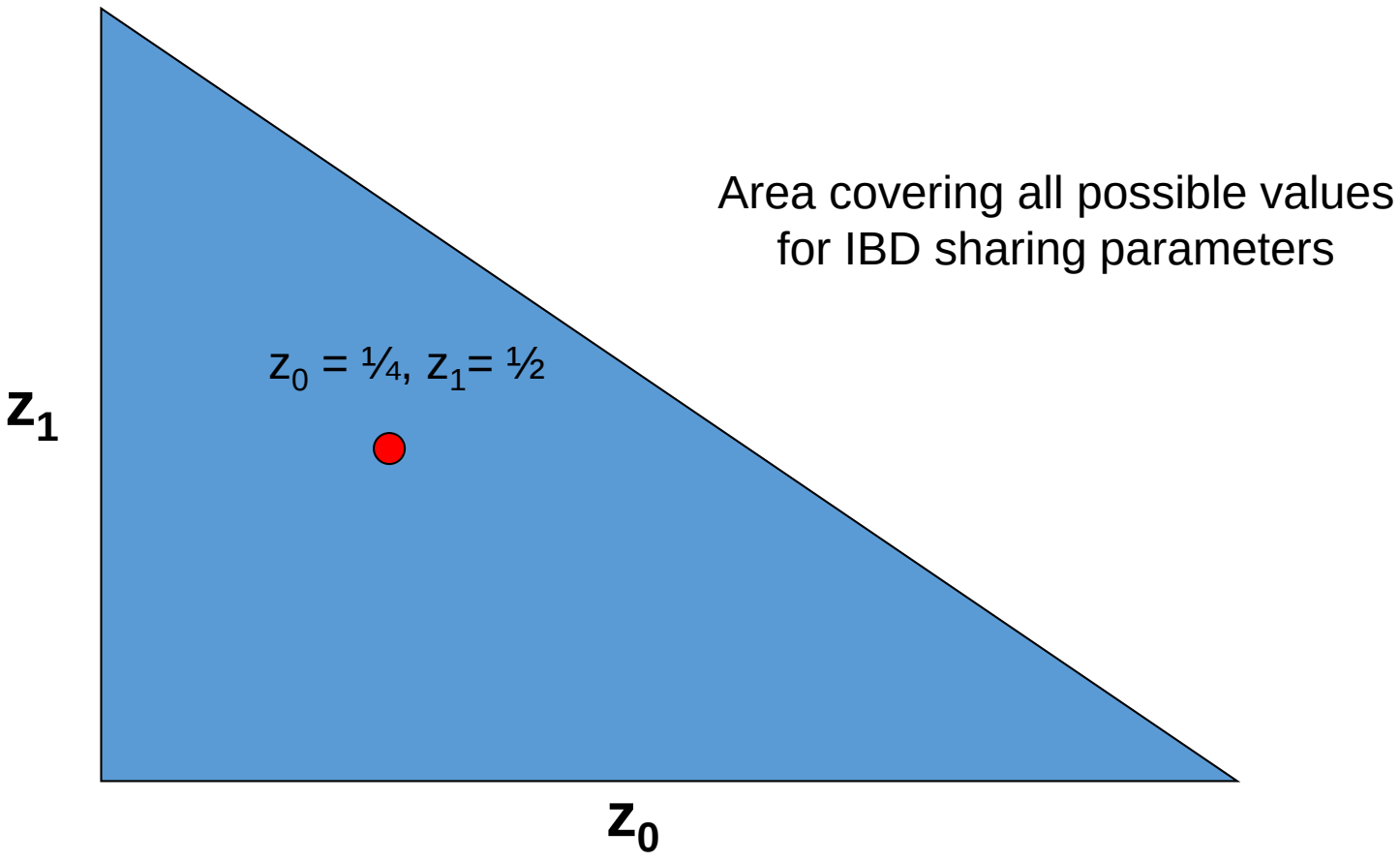
$$z_0 = 0.25\frac{1}{\lambda_s}$$

$$z_1 = 0.50\frac{\lambda_o}{\lambda_s}$$

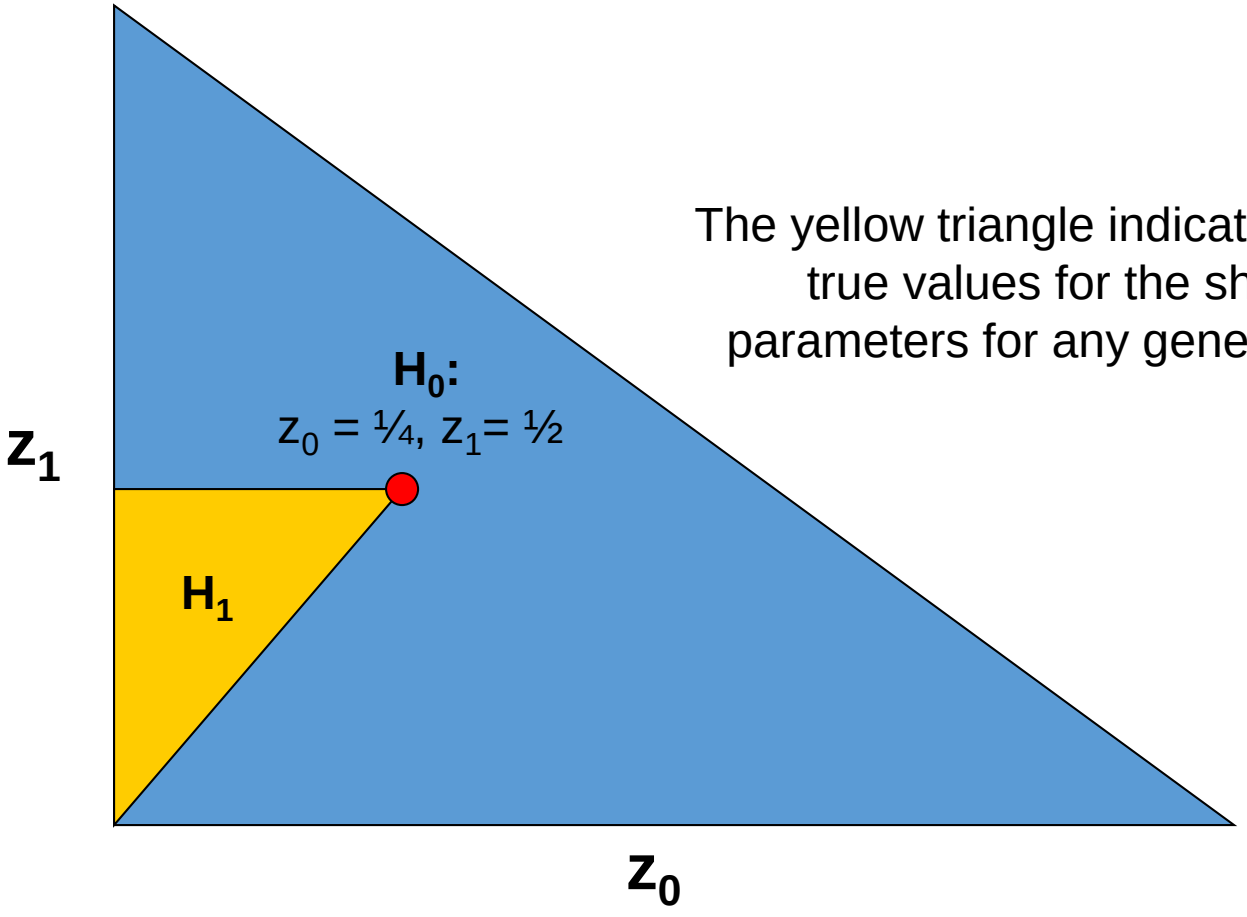
$$z_2 = 0.25\frac{\lambda_{MZ}}{\lambda_s}$$

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

Possible Triangle



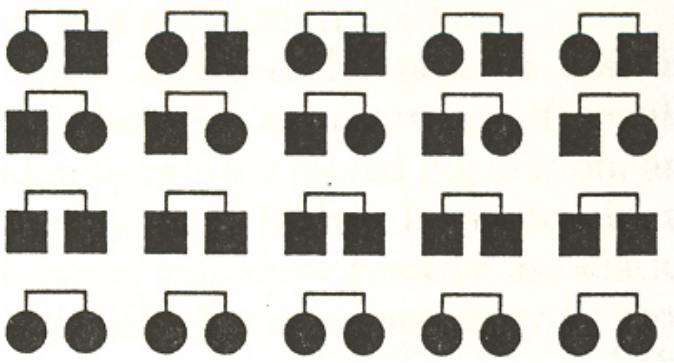
Possible Triangle



The yellow triangle indicates possible true values for the sharing parameters for any genetic model.

Intuition: Affected Sibpair Linkage Analyses

- Consider affected sibling pairs
- Consider one genetic marker at a time
- Are paired genotypes more similar than expected?



Likelihood Based Linkage Test

- Depends on three parameters z_0, z_1, z_2
  - Probability of sharing 0, 1 and 2 alleles IBD
- Null likelihood uses z_0=¼, z_1=½, z_2=¼
- Alternative likelihood uses MLE for z_0, z_1, z_2
- Compare likelihoods with likelihood ratio test

Potential Sib-Pair Likelihood

Under the null hypothesis:

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

Under the alternative hypothesis

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

In real life…

- Markers are only partially informative
- IBD sharing is equivocal
  - Uncertainty can only be partly reduced by examining relatives
- Need an alternative likelihood
  - Should allow for partially informative data

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

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

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

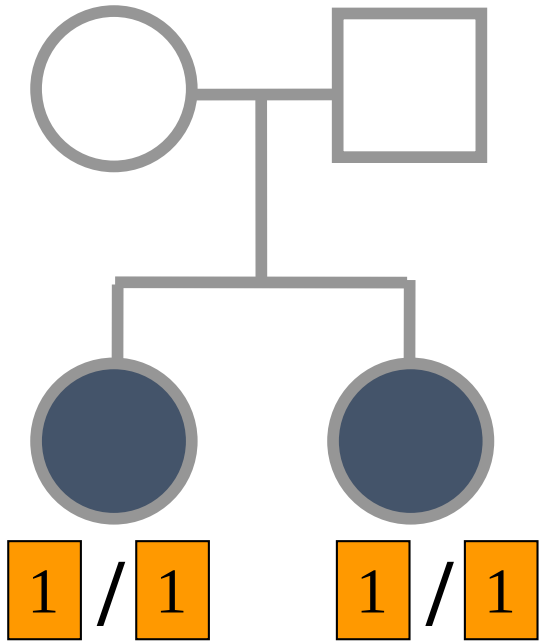
The MLS statistic is the LOD evaluated at the MLEs of z_0, z_1, z_2

w: 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

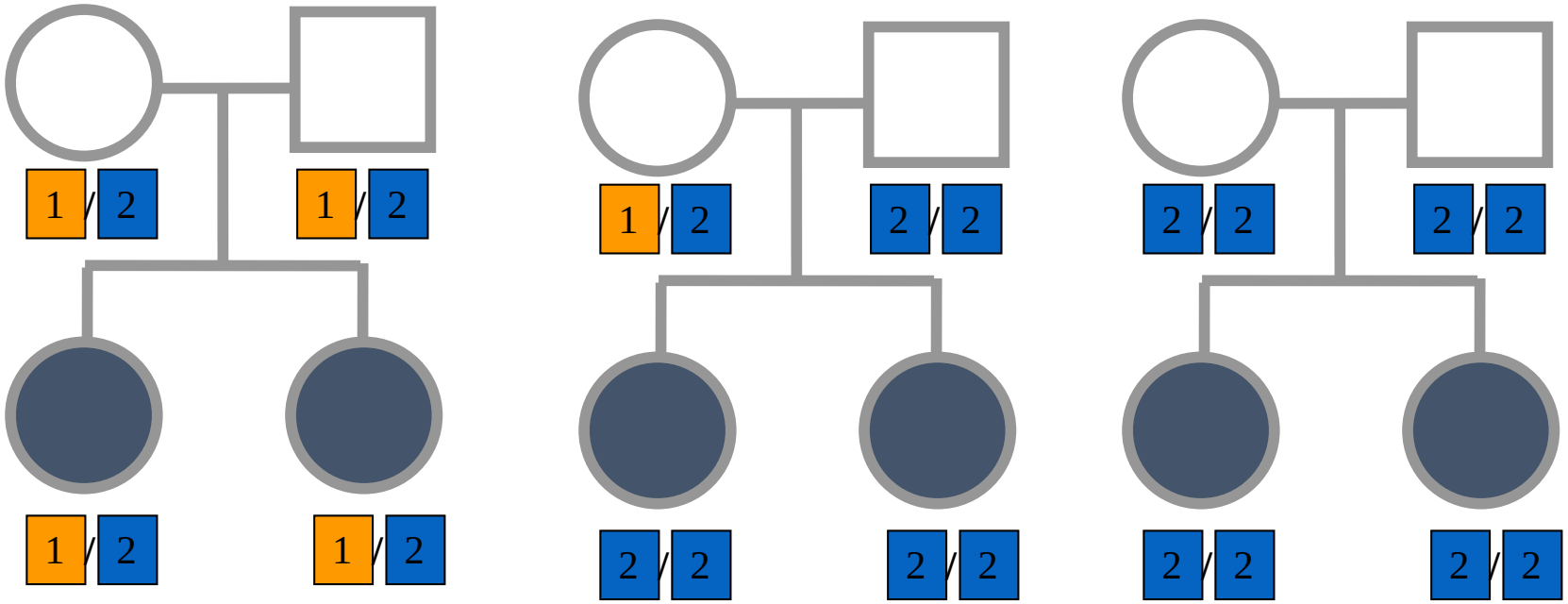
These probabilities apply to pair of individuals, when no other genotypes in the family are known.

Example scoring for w_{ij}



In this case, relative weights depend on allele frequency.

More examples for scoring: w_{ij}



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

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

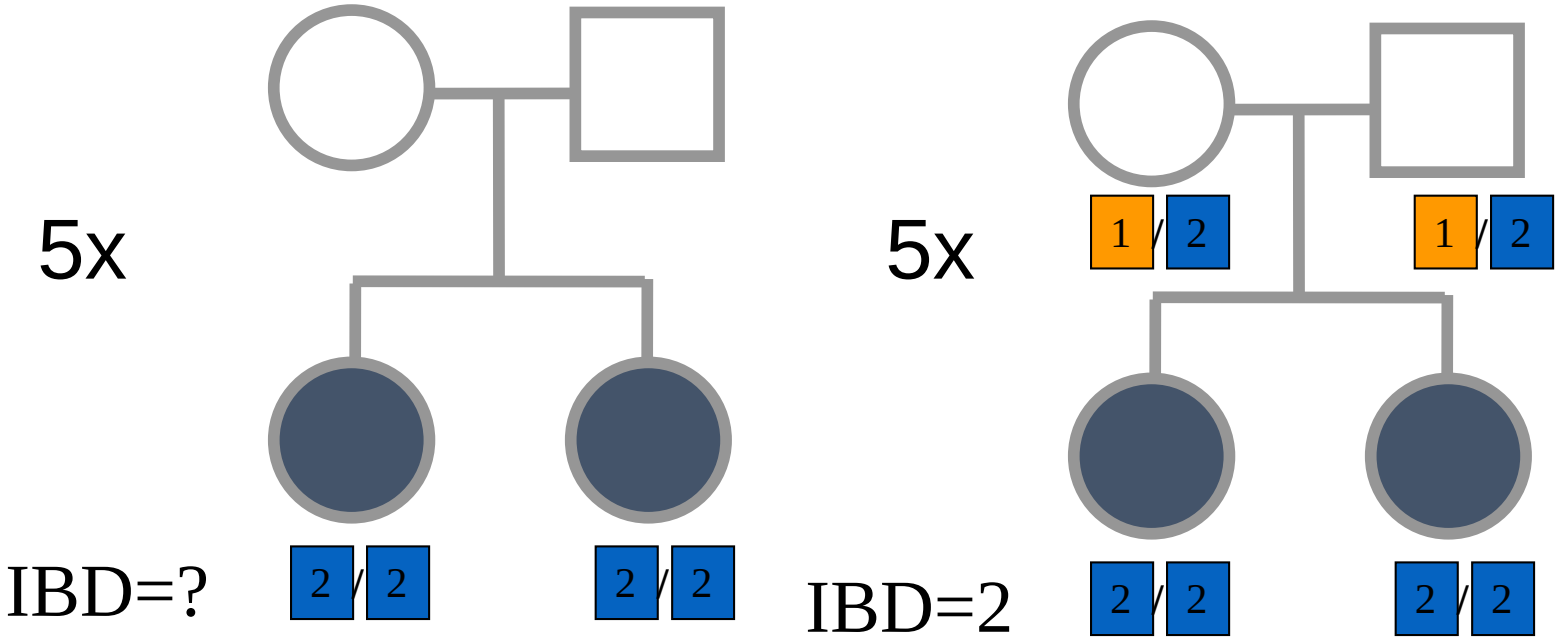
$$P(IBD = j \mid Genotypes) =$$

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

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

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

Example



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

Example of E-M Steps

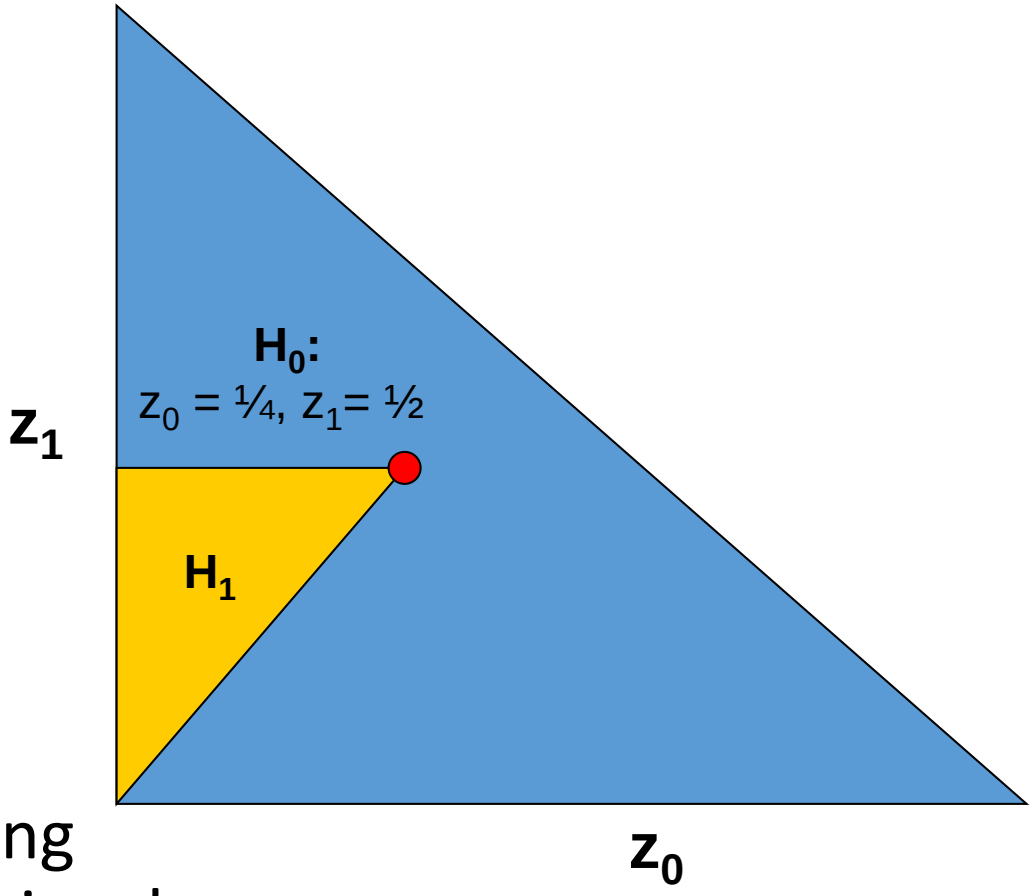
Parameters			Equivocal Families			Other			
z0	**z1**	**z2**	**IBD=0**	**IBD=1**	**IBD=2**	**IBD=2**	**LOD**	**LODi**	**LODu**
0.250	0.500	0.250	0.56	2.22	2.22	5	0.00	0.00	0.00
0.056	0.222	0.722	0.08	0.66	4.26	5	3.19	2.30	0.89
0.008	0.066	0.926	0.01	0.17	4.82	5	4.01	2.84	1.16
0.001	0.017	0.982	0.00	0.04	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

Point of Situation …

- Noted that affected siblings are more likely to share two alleles identical by descent
- Derived a likelihood based linkage test that compares sharing probabilities to null defaults
- Let’s examine these probabilities in more detail …

Intuition: Possible Triangle Constraints

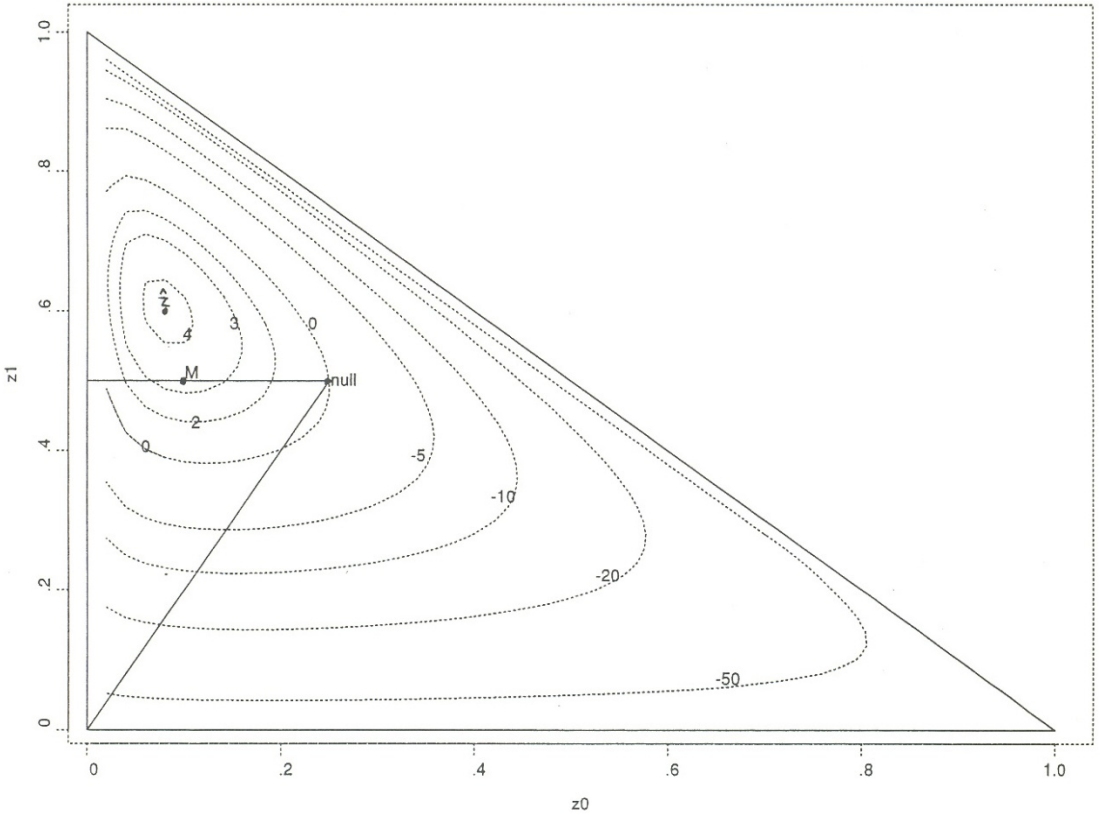
- Under the null
  - True parameter values are (¼, ½, ¼)
  - Estimates will wobble around this point
- Under the alternative
  - True parameter values are within triangle
  - Estimates will wobble around true point
- Holmans (1993) suggested we focus testing on searching for alternatives within the triangle
  - These suggest a disease gene



The possible triangle method

1. Estimate z_0, z_1, z_2 without restrictions
2. If estimate of $z_1 > ½$ then …
   a) Repeat estimation with $z_1 = ½$
   b) If this gives $z_0 > ¼$ then revert to null (MLS=0)
3. If estimates imply $2z_0 > z_1$ then …
   a) Repeat estimation with $z_1 = 2z_o$
   b) If this gives $z_0 > ¼$ then revert to null (MLS=0)
4. Otherwise, leave estimates unchanged.

Possible Triangle



Holman's Example:

IBD	Pairs
0	8
1	60
2	32

MLS = 4.22 (overall)
MLE = (0.08,0.60,0.32)

MLS = 3.35 (triangle)
MLE = (0.10,0.50,0.40)

MLS Combined With Possible Triangle

- Under null, true **z** is a corner of the triangle
  - Estimates will often lie outside triangle
  - Restriction to the triangle decreases MLS
  - MLS threshold for fixed type I error decreases

- Under alternative, true **z** is within triangle
  - Estimates will lie outside triangle less often
  - MLS decreases less
  - Overall, power should be increased

Example

- Type I error rate of 0.001
- LOD of 3.0 with unrestricted method
  - Risch (1990)
- LOD of 2.3 with possible triangle constraint
  - Holmans (1993)
  - For some cases, almost doubles power

Recommended Reading

- Holmans (1993)
  Asymptotic Properties of
  Affected-Sib-Pair Linkage Analysis
  Am J Hum Genet **52:**362-374

- Introduces possible triangle constraint
- Good review of MLS method

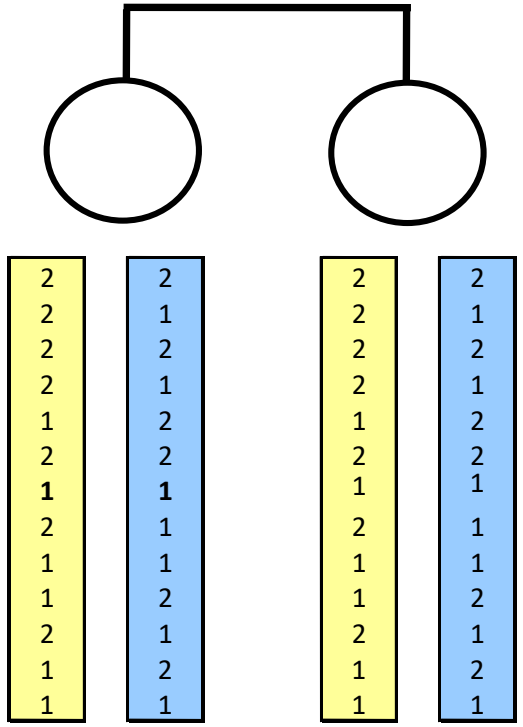
Recommended Reading

- 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

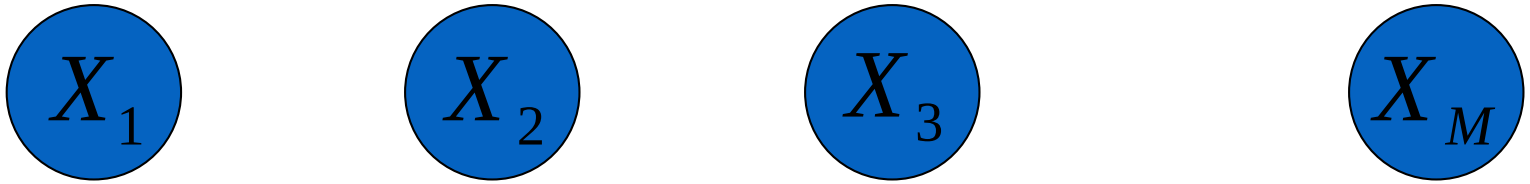
- Introduces MLS method for linkage analysis
  - Still, one of the best methods for analysis pair data
- Evaluates different sampling strategies
  - Results were later corrected by Risch (1992)

Intuition For Multipoint Analysis

- IBD changes infrequently along the chromosome
- Neighboring markers can help resolve ambiguities about IBD sharing
- In the Risch approach, they might ensure that only one **w** is *effectively* non-zero

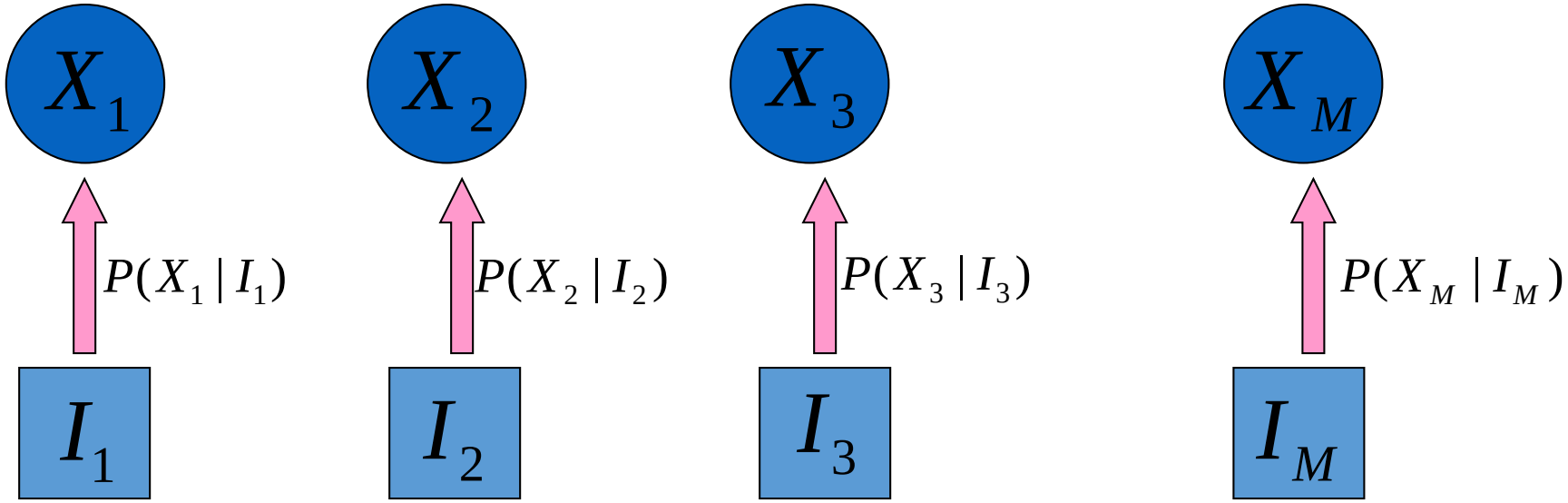


Ingredients for a multipoint model…



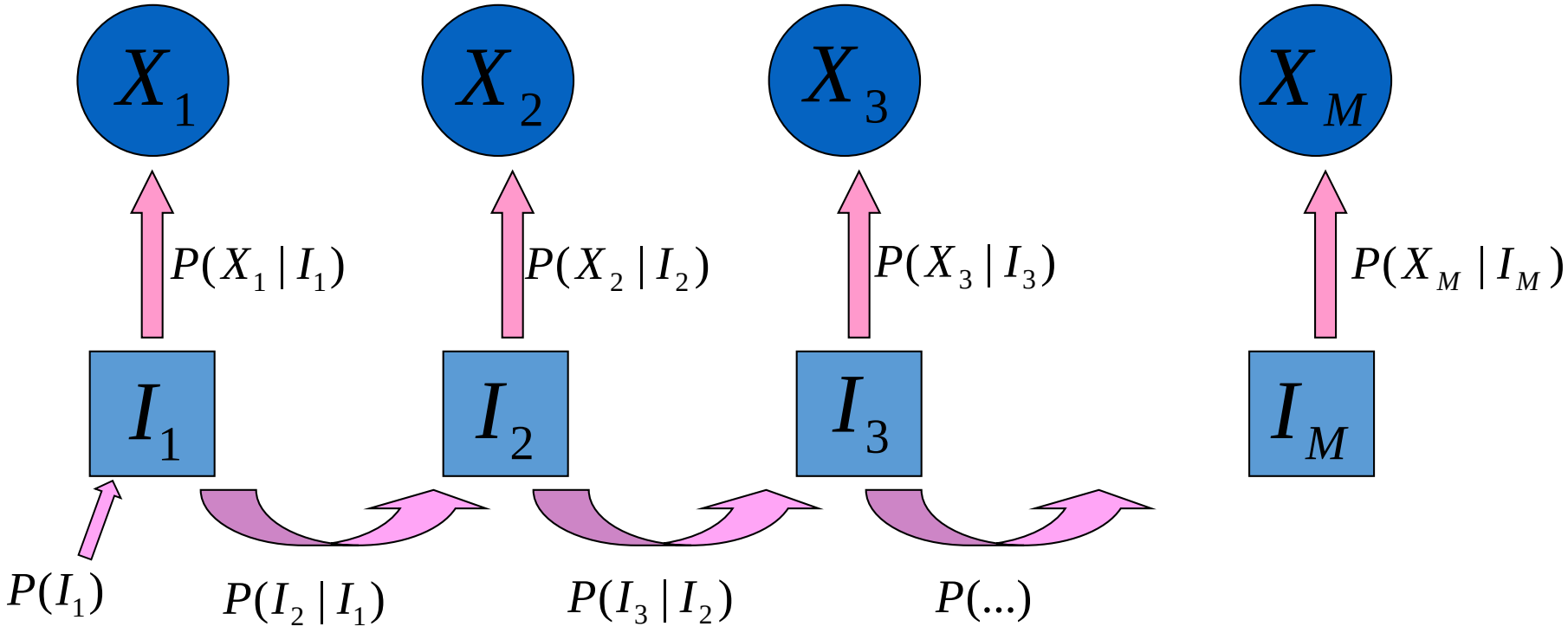
One ingredient will be the observed genotypes at each marker …

Ingredients for a multipoint model…



Another ingredient will be the possible IBD states at each marker …

Ingredients for a multipoint model…



The final ingredient connects IBD states along the chromosome …

The Likelihood of Marker Data

$$L = \sum_{I_1} \sum_{I_2} \ldots \sum_{I_M} P(I_1) \prod_{i=2}^{M} P(I_i \mid I_{i-1}) \prod_{i=1}^{M} P(X_i \mid I_i)$$

- General formulation, allows for any number of markers.

- Combined with Bayes' Theorem can estimate probability of each IBD state at any marker.

$P(X_m \mid I_m)$

		IBD		
Sib	**CoSib**	0	1	2
(a,b)	(c,d)	$4p_a p_b p_c p_d$	0	0
(a,a)	(b,c)	$2p_a^2 p_b p_c$	0	0
(a,a)	(b,b)	$p_a^2 p_b^2$	0	0
(a,b)	(a,c)	$4p_a^2 p_b p_c$	$p_a p_b p_c$	0
(a,a)	(a,b)	$2p_a^3 p_b$	$p_a^2 p_b$	0
(a,b)	(a,b)	$4p_a^2 p_b^2$	$(p_a p_b^2 + p_a^2 p_b)$	$2p_a p_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

$P(I_{m+1} \mid I_m)$

- Depends on recombination fraction θ
  - This is a measure of distance between two loci
  - Probability grand-parental origin of alleles changes between loci

		IBD State at m + 1		
		0	1	2
IBD state at marker m	0	$(1-\psi)^2$	$2\psi(1-\psi)$	ψ^2
	1	$\psi(1-\psi)$	$(1-\psi)^2+\psi^2$	$\psi(1-\psi)$
	2	ψ^2	$2\psi(1-\psi)$	$(1-\psi)^2$

$$\psi = 2\theta(1-\theta)$$

The Likelihood of Marker Data

$$L = \sum_{I_1} \sum_{I_2} \ldots \sum_{I_M} P(I_1) \prod_{i=2}^{M} P(I_i \mid I_{i-1}) \prod_{i=1}^{M} P(X_i \mid I_i)$$

- General, but slow unless there are only a few markers.

- How do we speed things up?

Example

- Consider two loci separated by $\theta = 0.1$
- Each loci has two alleles, each with frequency .50
- If two siblings are homozygous for the first allele at both loci, what is the probability that IBD = 2 at the first locus?

A Markov Model

- Re-organize the computation slightly, to avoid evaluating nested sum directly

- Three components:
  - Probability considering a single location
  - Probability including left flanking markers
  - Probability including right flanking markers

- Scale of computation increases linearly with number of markers

Left-Chain Probabilities

$$L_m(I_m) = P(X_1, \ldots, X_{m-1} \mid I_m)$$
$$= \sum_{I_{m-1}} L_{m-1}(I_{m-1}) P(X_{m-1} \mid I_{m-1}) P(I_{m-1} \mid I_m)$$
$$L_1(I_1) = 1$$

- Proceed one marker at a time.
- Computation cost increases linearly with number of markers.

Right-Chain Probabilities

$$R_m(I_m) = P(X_{m+1},\ldots,X_M \mid I_m)$$
$$= \sum_{I_{m+1}} R_{m+1}(I_{m+1}) P(X_{m+1} \mid I_{m+1}) P(I_{m+1} \mid I_m)$$
$$R_M(I_M) = 1$$

- Proceed one marker at a time.

- Computation cost increases linearly with number of markers.

Extending the MLS Method …

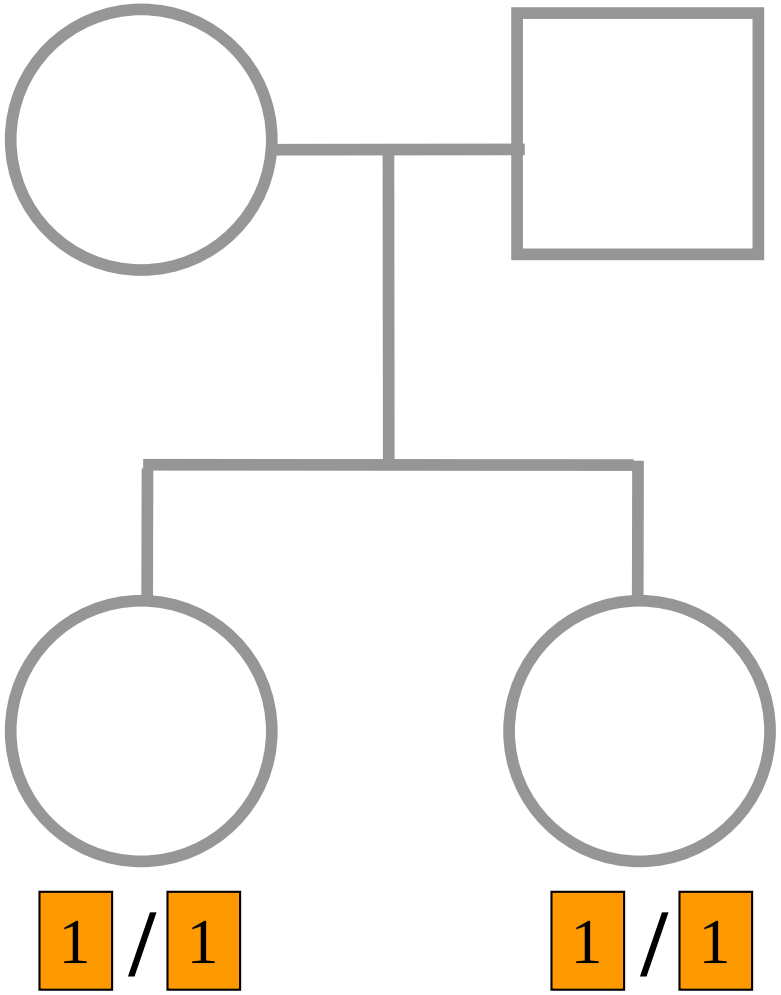
$$
\begin{aligned}
w_j &= P(X_j \mid I_j) P(X_1 \ldots X_{j-1} \mid I_j) P(X_{j+1} \ldots X_M \mid I_j) \\
&= P(X_j \mid I_j) L_j(I_j) R_j(I_j)
\end{aligned}
$$

- We just change the definition for the “weights” given to each configuration!

Possible Further Extensions

- Modeling error
  - What components might have to change?
- Modeling other types of relatives
  - What components might have to change?
- Modeling larger pedigrees
  - What components might have to change?

Worked Example



$$p_1 = 0.5$$

$$w_0 = P(X \mid IBD = 0) = p_1^4 = 1/16$$

$$w_1 = P(X \mid IBD = 1) = p_1^3 = 1/8$$

$$w_2 = P(X \mid IBD = 2) = p_1^2 = 1/4$$

If $z_0 = 0.25, z_1 = 0.50, z_2 = 0.25$, then

$$P(X) = 1/4\, p_1^4 + 1/2\, p_1^3 + 1/4\, p_1^2 = 9/64$$

$$P(IBD = 0 \mid X) = \frac{1/4\, p_1^4}{P(X)} = 1/9$$

$$P(IBD = 1 \mid X) = \frac{1/2\, p_1^3}{P(X)} = 4/9$$

$$P(IBD = 2 \mid X) = \frac{1/4\, p_1^2}{P(X)} = 4/9$$