# *Multipoint Analysis for Sibling Pairs*

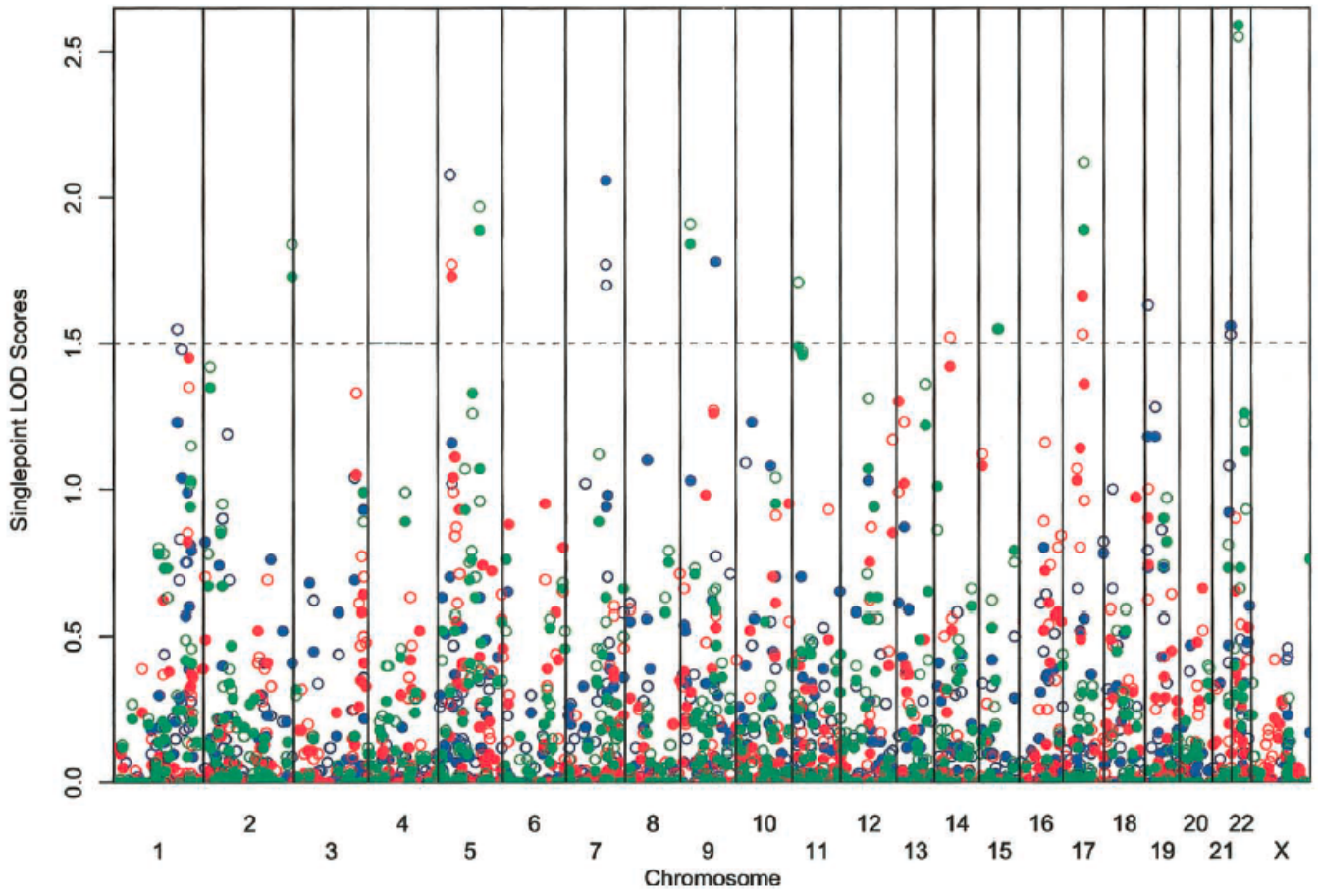
Biostatistics 666

# Example of a Linkage Study

- In the US, age-related macular degeneration (AMD) is the most common cause of blindness in the elderly.
- Disease is heterogeneous, with two common forms of severe disease, termed “wet” and “dry” disease.
- Linkage study examined 734 genetic markers (~1 per 5Mb) in 412 affected relative pairs.
- Evidence for linkage to several regions, including chr1 (~240 cM) and chr22 (~25 cM). We now know these correspond to CFH and TIMP3 susceptibility alleles.
- American Journal of Human Genetics (2004) **74:**482-494

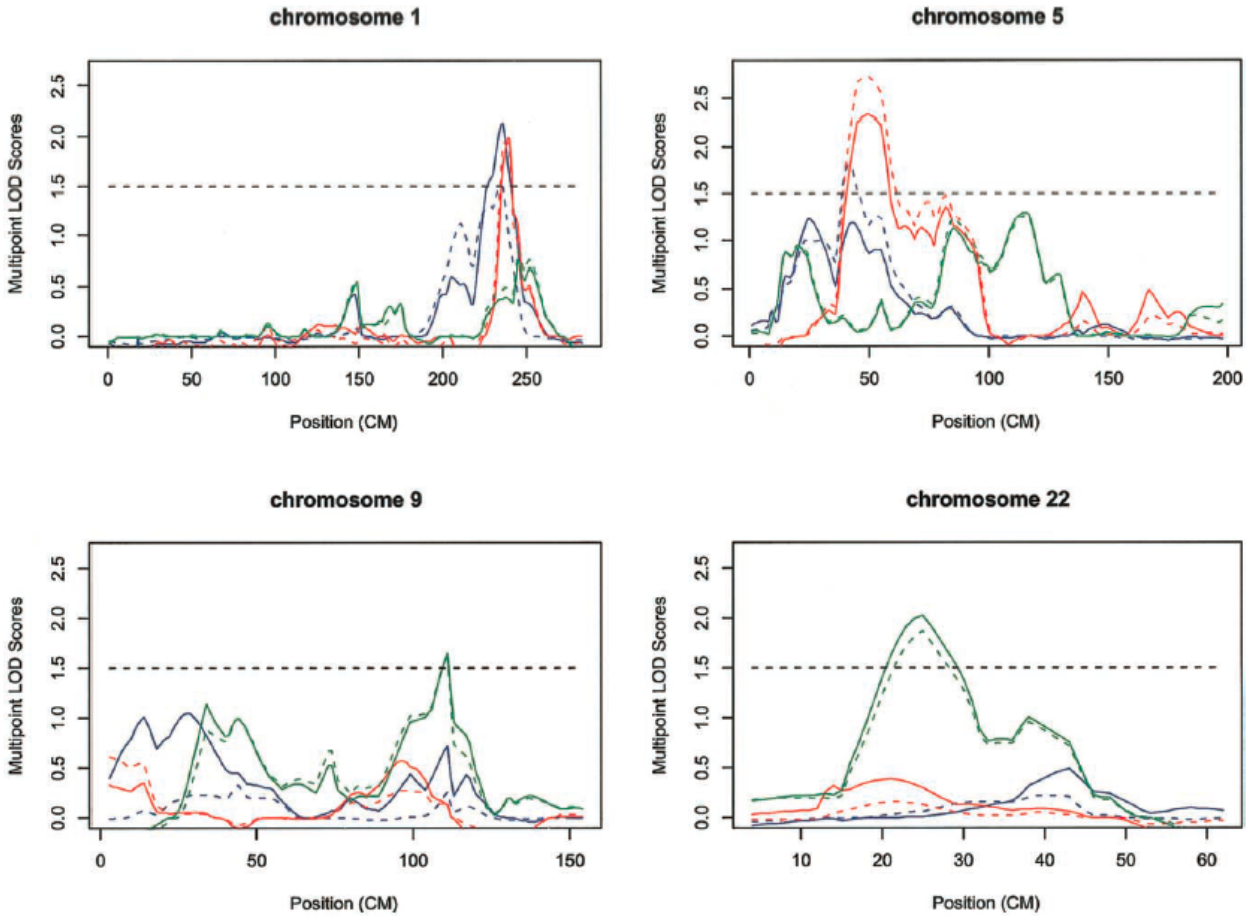
# AMD Linkage Study,

## Results of Marker by Marker Analysis




Colors: **All AMD**, **“Wet” Subtype**, **“Dry” Subtype**

# AMD Linkage Study,
## Results of Multipoint Analysis



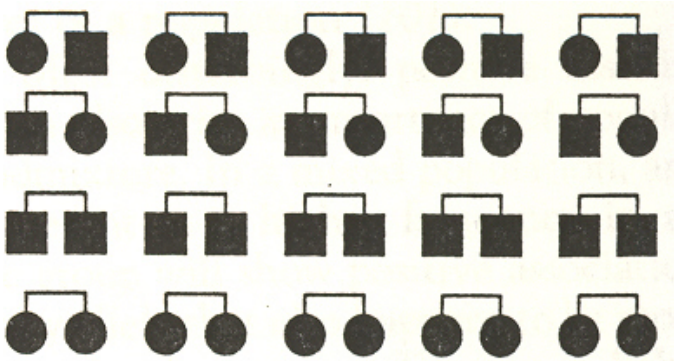

Colors: **All AMD**, **“Wet” Subtype**, **“Dry” Subtype**

## Previously ...

- Linkage analysis with sibling pairs
- Risch's Maximum LOD Score approach
- Holman's Possible Triangle Constraint
- Distribution of IBD in affected sibling pairs

# Affected Sib Pair Linkage Tests

- Consider affected sibling pairs
  - Pairs selected to have similar phenotypes …
  - … show increased similarity at loci that change disease risk
- Scan the genome and test whether pair genotypes are more similar than expected …



# Likelihood for a Single ASP

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

Risch (1990) defines

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

We only need proportionate $w_{ij}$

# $w_{ij}$ for single marker analyses

| 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.

# MLS Linkage Test

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

$$LOD = \log_{10} \prod_i \frac{z_0 w_{i0} + z_1 w_{i1} + z_2 w_{i2}}{\frac{1}{4} w_{i0} + \frac{1}{2} w_{i1} + \frac{1}{4} w_{i2}}$$

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

These MLEs can be calculated using an EM algorithm

# P(Affected Pair | IBD=j)

$$P(\text{Affected Pair}|\ \text{IBD} = 0) = (p^2 f_{11} + 2p(1-p)f_{12} + (1-p)^2 f_{22})^2 = K^2$$

$$P(\text{Affected Pair}|\ \text{IBD} = 1) = p^3 f_{11}^2 + 2p^2(1-p)f_{11}f_{12} + p(1-p)f_{12}^2 + 2p(1-p)^2 f_{12}f_{22} + (1-p)^3 f_{22}^2 = \lambda_O K^2$$

$$P(\text{Affected Pair}|\ \text{IBD} = 2) = p^2 f_{11}^2 + 2p(1-p)f_{12}^2 + (1-p)^2 f_{22}^2 = \lambda_{MZ} K^2$$

# P(Affected Sibling Pair), P(IBD = j | Affected Sibling Pair)

$$P(ASP) = \sum \mathrm{P(IBD = i)P(ASP|IBD = i)} = \left({}^{1}/_{4} + {}^{1}/_{2}\,\lambda_0 + {}^{1}/_{4}\,\lambda_{MZ}\right)K^2 = \lambda_{sib}K^2$$

$$P(IBD = j|ASP) = \frac{P(IBD = j)P(ASP|IBD = j)}{P(ASP)}$$

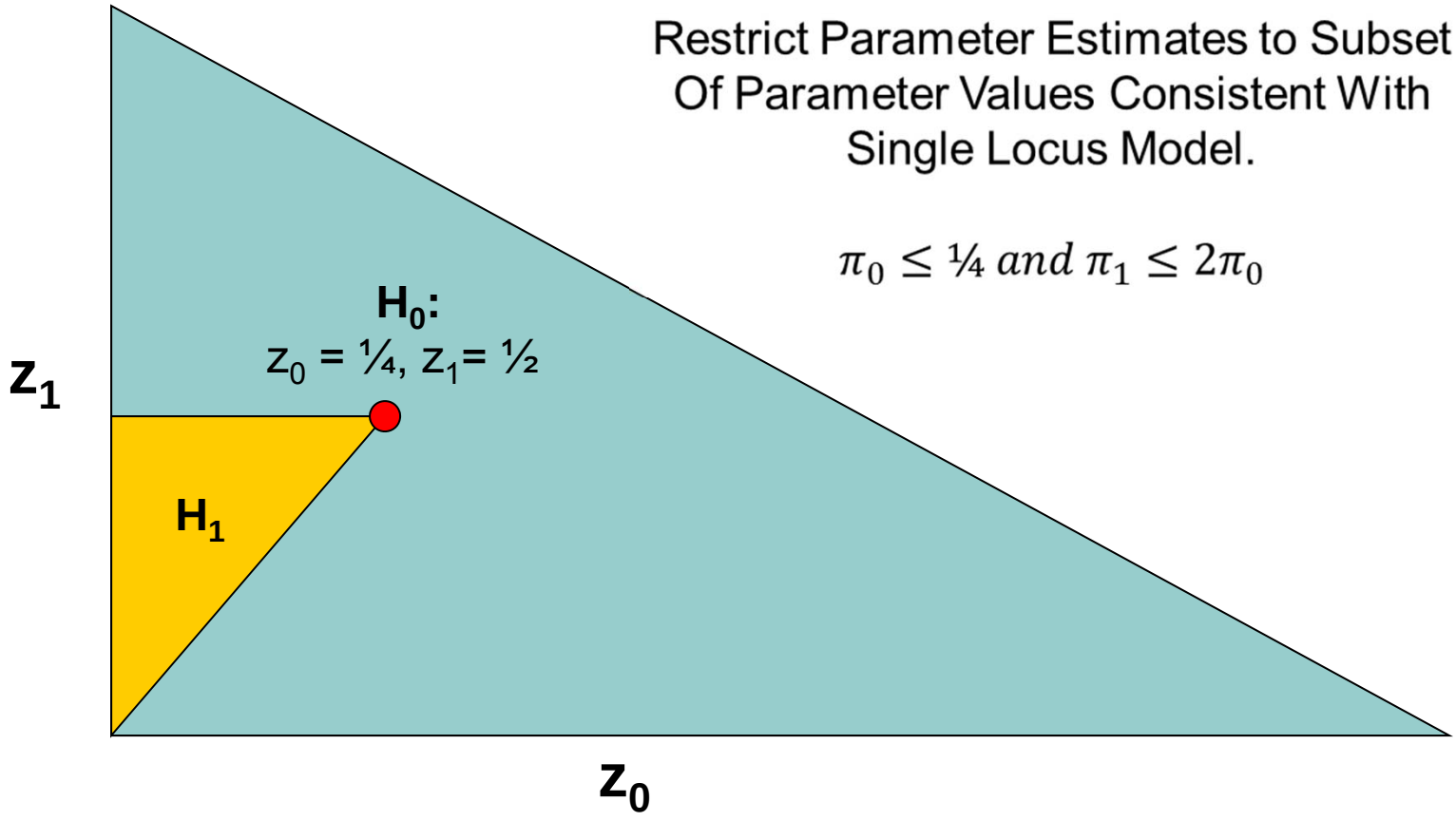
$$P(IBD = 0|ASP) = 0.25\,\frac{1}{\lambda_{sib}}$$

$$P(IBD = 1|ASP) = 0.50\,\frac{\lambda_O}{\lambda_{sib}}$$

$$P(IBD = 1|ASP) = 0.25\,\frac{\lambda_{MZ}}{\lambda_{sib}}$$

$$1 \le \lambda_O \le \lambda_{sib} \le \lambda_{MZ}$$

# Possible Triangle Constraint



# Important Limitation

- A major limitation of our approach so far is that it considers one marker at a time
- This may not allow us to extract all available information about IBD…

## Today ...

- Refresher on IBD probabilities
- Intuition behind multipoint calculations
- Framework for multipoint calculations
- Using a Markov Chain to speed analyses

# IBD Probabilities

- Number of alleles identical by descent
- For sibling pairs, must be:
  - 0
  - 1
  - 2
- Often, remains ambiguous given genotype

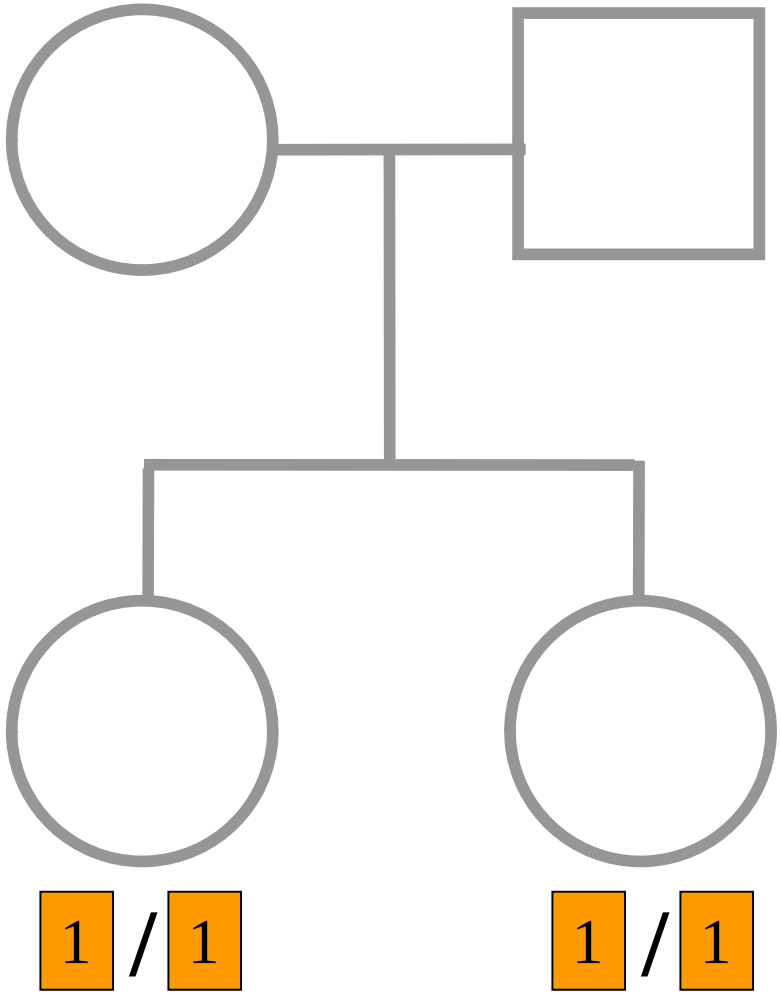
# Refresher ...
# Bayes Theorem for IBD Probabilities

$$P(IBD = i \mid X) = \frac{\mathrm{P}(\mathrm{IBD} = i, X)}{P(X)}$$

$$= \frac{P(IBD = i)P(X \mid IBD = i)}{P(X)}$$

$$= \frac{P(IBD = i)P(X \mid IBD = i)}{\sum_j P(IBD = j)P(X \mid IBD = j)}$$

# Worked Example




$$p_1 = 0.5$$

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

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

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

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

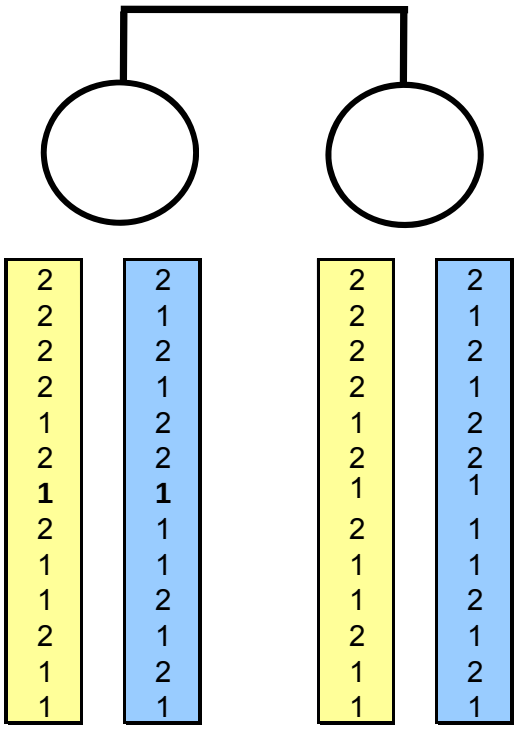
$$P(IBD=0 \mid X) = \frac{1/4\, p_1^4}{P(X)} = 1/9$$

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

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

# Intuition For Multipoint Analysis

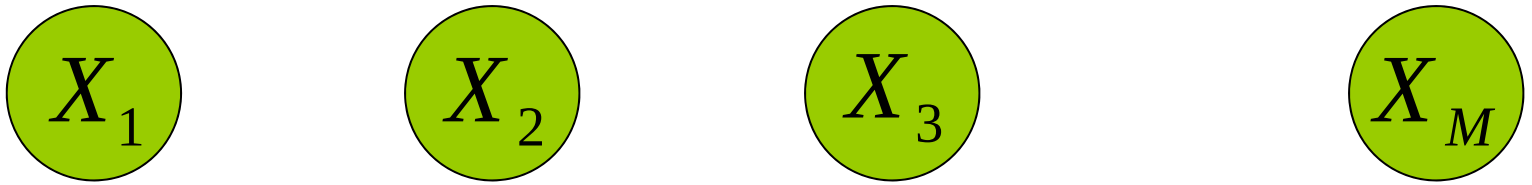
- IBD changes infrequently along the chromosome
- Neighboring markers can help resolve ambiguities about IBD sharing
- In the Risch approach, they might ensure that only one **w** is *effectively* non-zero



# Ingredients for Multipoint Model

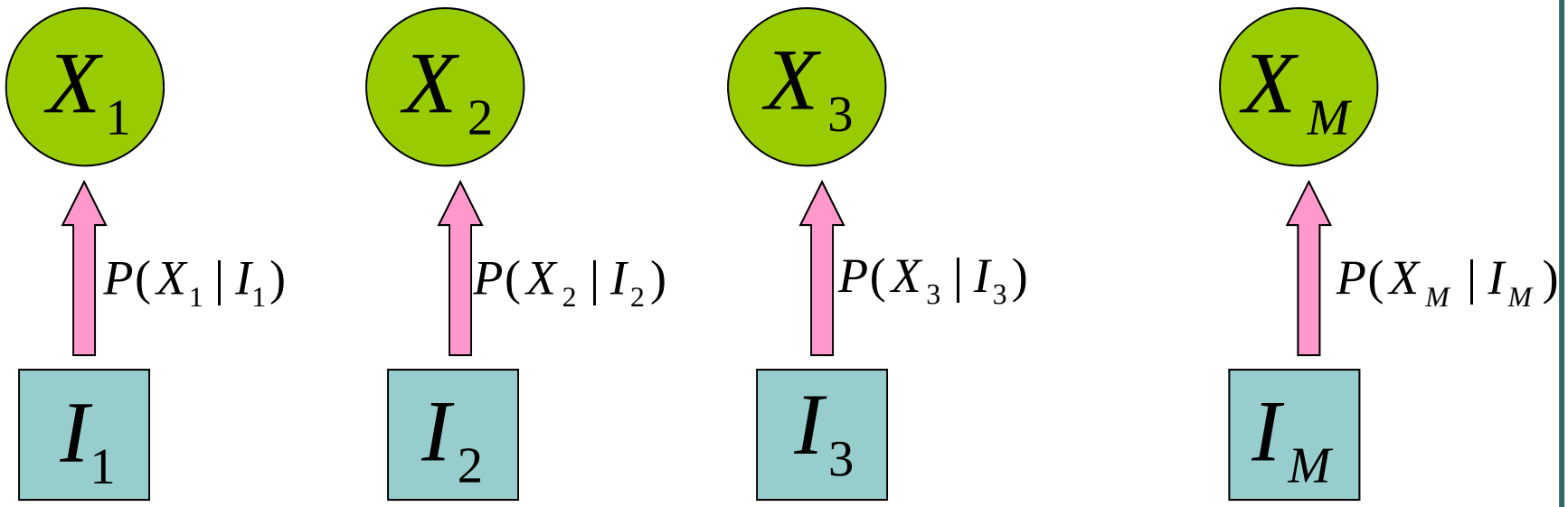
- Probability of observed genotypes at each marker conditional on IBD state
- Probability of changes in IBD state along chromosome
- Hidden Markov Model

# Ingredients



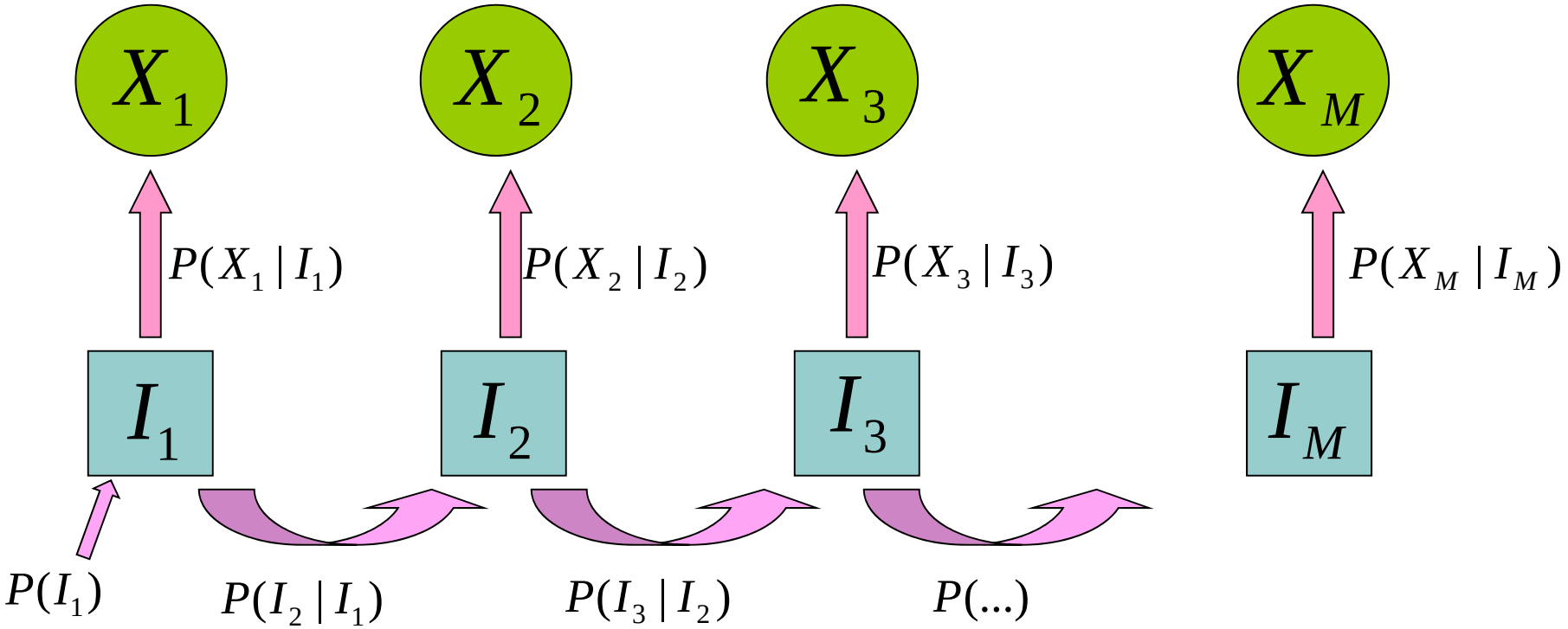

One ingredient will be the observed genotypes at each marker …

# Ingredients




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

# Ingredients




The final ingredient connects IBD states along the chromosome …

# The Likelihood of Marker Data

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

- General formulation, allows for any number of markers.
- Combined with Bayes' Theorem can estimate probability of each IBD state at any marker.
- This is not a linkage test yet!

# $P(X_m \mid I_m)$

| | | 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 |

# Question:
# What to do about missing data?

- What happens when some genotype data is unavailable?

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

- Depends on recombination fraction $\theta$
  - This is a measure of distance between two loci
  - Probability grand-parental origin of alleles changes between loci
- Leads to probability of a change in IBD:

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

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

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

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

## 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?

# The Likelihood of Marker Data

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

- General, but slow unless there are only a few markers.
- How do we speed things up?

# 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

# The Likelihood of Marker Data

$$L = \sum_{I_j} P(I_j)P(X_j \mid I_j)P(X_1 \ldots X_{j-1} \mid I_j)P(X_{j+1} \ldots X_M \mid I_j)$$

$$= \sum_{I_j} P(I_j)P(X_j \mid I_j)L_j(I_j)R_j(I_j)$$

- A different arrangement of the same likelihood
- The nested summations are now hidden inside the $L_j$ and $R_j$ functions…

# 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

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

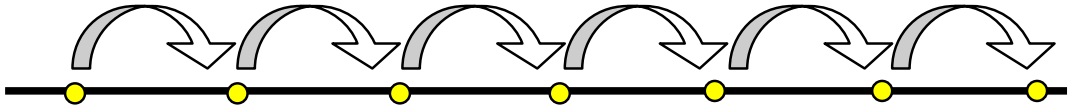
$$R_M(I_M) = 1$$

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

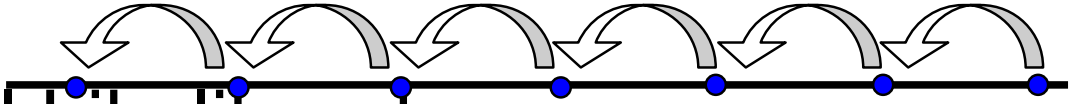
# Pictorial Representation

- Single Marker

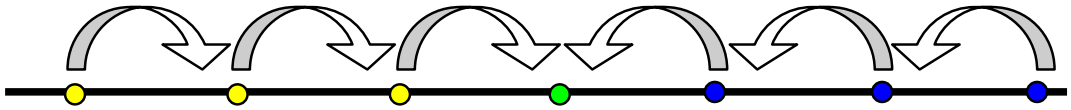
- Left Conditional



- Right Conditional



- Full Likelihood



# Extending the MLS Method ...

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

- 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?

## Today

- Efficient computational framework for multipoint analysis of sibling pairs