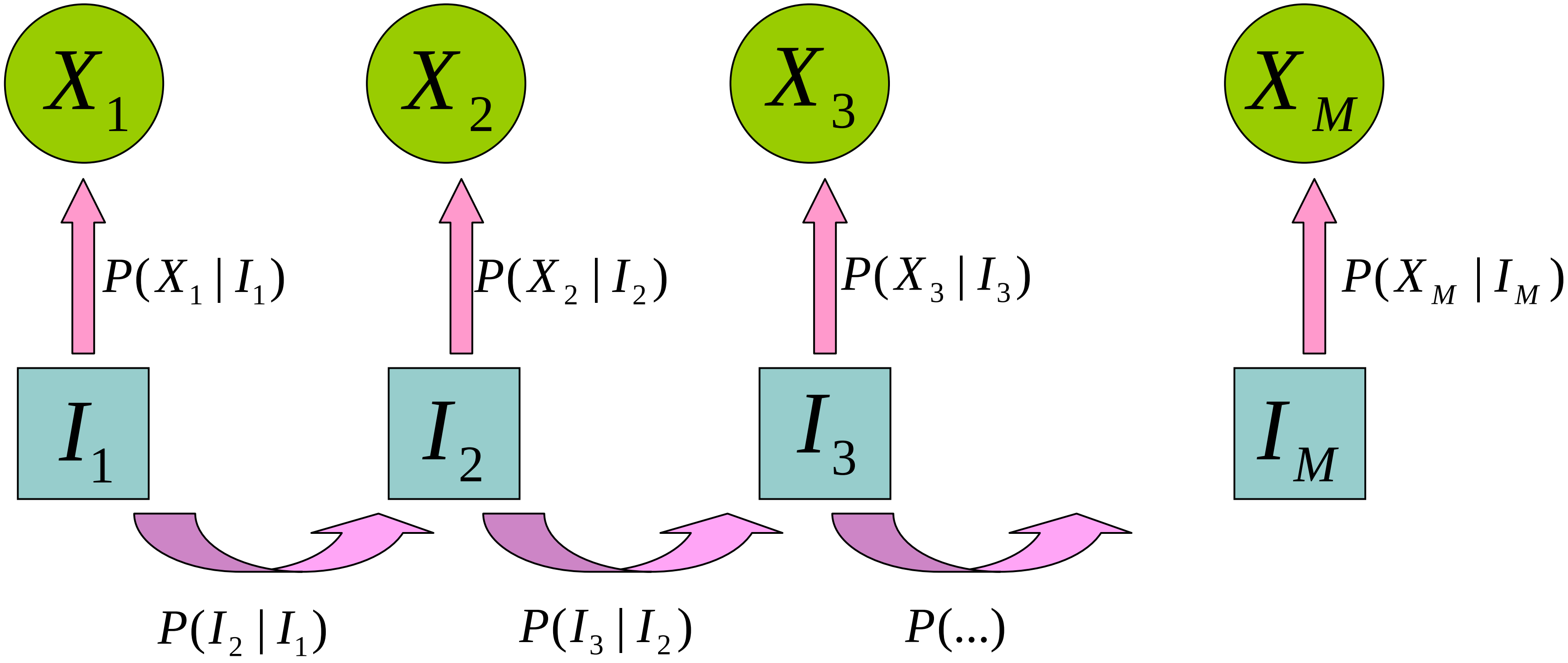
Checking Pairwise Relationships

Lecture 19

Biostatistics 666

Last Lecture: Markov Model for Multipoint Analysis



IBD states along the chromosome are modeled using a Markov Chain …

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.

- Combined with Bayes' Theorem allows us to estimate probability of IBD states at any marker.

Worked Example

- Consider two loci separated by θ = 0.1
- Each loci has two alleles, each with frequency .50

- If two siblings have the following genotypes:

	Sib1	Sib2
Marker A:	1/1	2/2
Marker B:	1/1	1/1

- What is the probability of IBD=2 at marker B when…
  - You consider marker B alone?
  - You consider both markers simultaneously?

Solution

I_1	I_2	$P(I_1)$	$P(I_2 \mid I_1)$	$P(X_1 \mid I_1)$	$P(X_2 \mid I_2)$	Prob
0	0					
0	1					
0	2					
1	0					
1	1					
1	2					
2	0					
2	1					
2	2					

Solution

I_1	I_2	$P(I_1)$	$P(I_2 \mid I_1)$	$P(X_1 \mid I_1)$	$P(X_2 \mid I_2)$	Prob
0	0	0.25	0.67	0.0625	0.0625	0.00066
0	1	0.25	0.30	0.0625	0.125	0.00058
0	2	0.25	0.03	0.0625	0.25	0.00013
1	0	0.5	0.15	0	0.0625	0.00000
1	1	0.5	0.70	0	0.125	0.00000
1	2	0.5	0.15	0	0.25	0.00000
2	0	0.25	0.03	0	0.0625	0.00000
2	1	0.25	0.30	0	0.125	0.00000
2	2	0.25	0.67	0	0.25	0.00000

Solution

- Taking into account all available genotype data…
  - $P(I_1 = 2) = 0.09$
  - $P(I_1 = 1) = 0.42$
  - $P(I_1 = 0) = 0.49$
- Considering only one marker, the corresponding probabilities would be 0.44, 0.44 and 0.11.
  - Quite a difference!, but which value do you expect to be more accurate?

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?

Extending the MLS Method ...

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

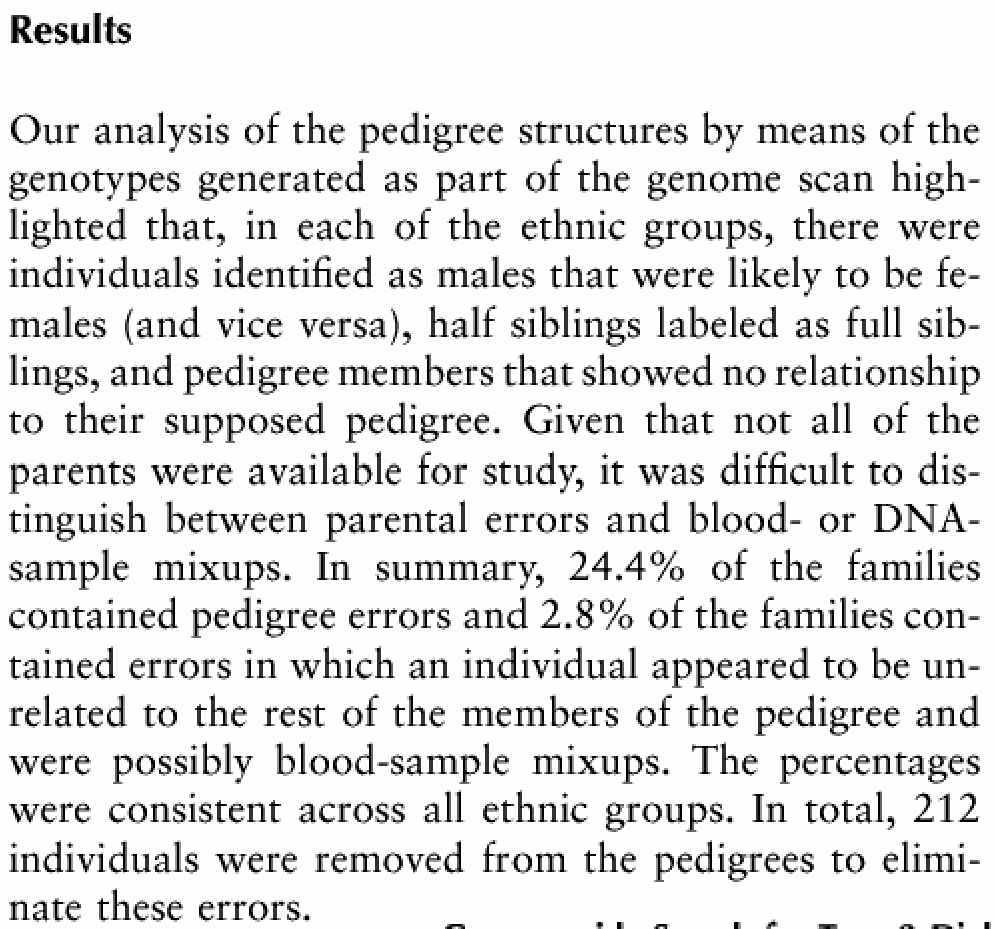
- We just change the definition for the “weights” given to each configuration!

Today …

- Checking accuracy of reported relationships
  - Why is this an important problem?
- Markov Chain for Different Relative Pairs
  - Likelihood approaches to relationship inference

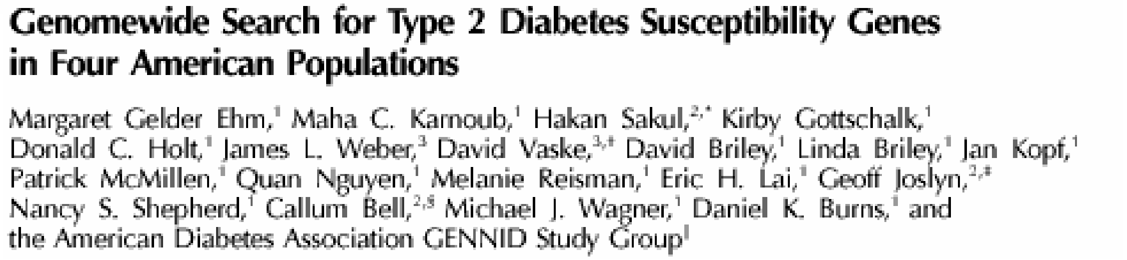
Verifying relationships is crucial

- Genetic analyses require relationships to be specified
- Misspecifying relationships can lead to tests of inappropriate size
  - Inflate Type I error
  - Decrease power



Results

Our analysis of the pedigree structures by means of the genotypes generated as part of the genome scan highlighted that, in each of the ethnic groups, there were individuals identified as males that were likely to be females (and vice versa), half siblings labeled as full siblings, and pedigree members that showed no relationship to their supposed pedigree. Given that not all of the parents were available for study, it was difficult to distinguish between parental errors and blood- or DNA-sample mixups. In summary, 24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree and were possibly blood-sample mixups. The percentages were consistent across all ethnic groups. In total, 212 individuals were removed from the pedigrees to eliminate these errors.



Genomewide Search for Type 2 Diabetes Susceptibility Genes in Four American Populations

Margaret Gelder Ehm,[1] Maha C. Karnoub,[1] Hakan Sakul,[2,*] Kirby Gottschalk,[1] Donald C. Holt,[1] James L. Weber,[3] David Vaske,[3,†] David Briley,[1] Linda Briley,[1] Jan Kopf,[1] Patrick McMillen,[1] Quan Nguyen,[1] Melanie Reisman,[1] Eric H. Lai,[1] Geoff Joslyn,[2,‡] Nancy S. Shepherd,[1] Callum Bell,[2,§] Michael J. Wagner,[1] Daniel K. Burns,[1] and the American Diabetes Association GENNID Study Group[‖]

IBS Based Approach

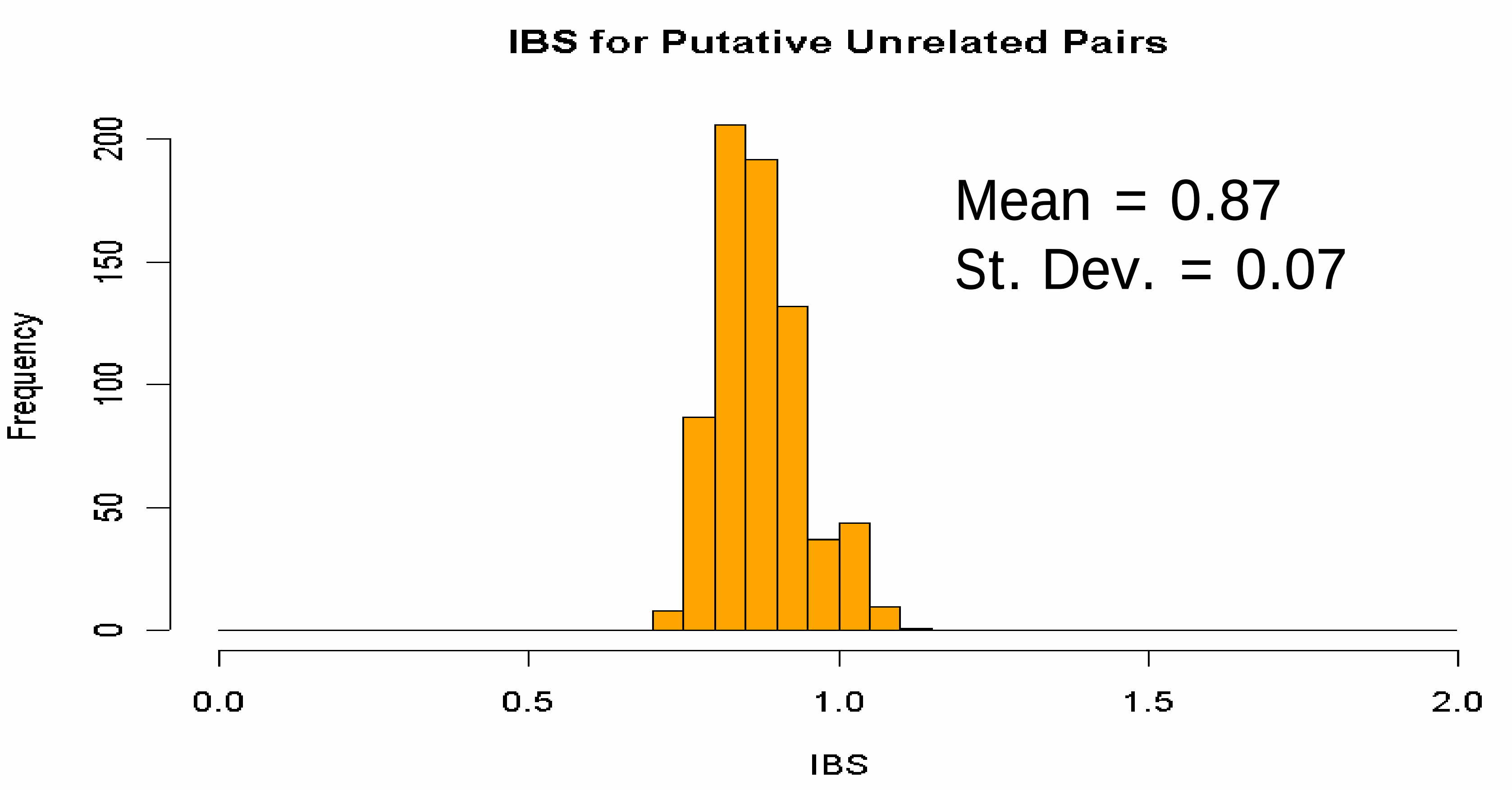
- Relative pairs will differ in terms of their genetic similarity …

- One way to contrast different types of relatives is to compare their overall similarity, for example, by:
  - Calculating the mean IBS sharing
  - Calculating the variance of IBS sharing

Example…

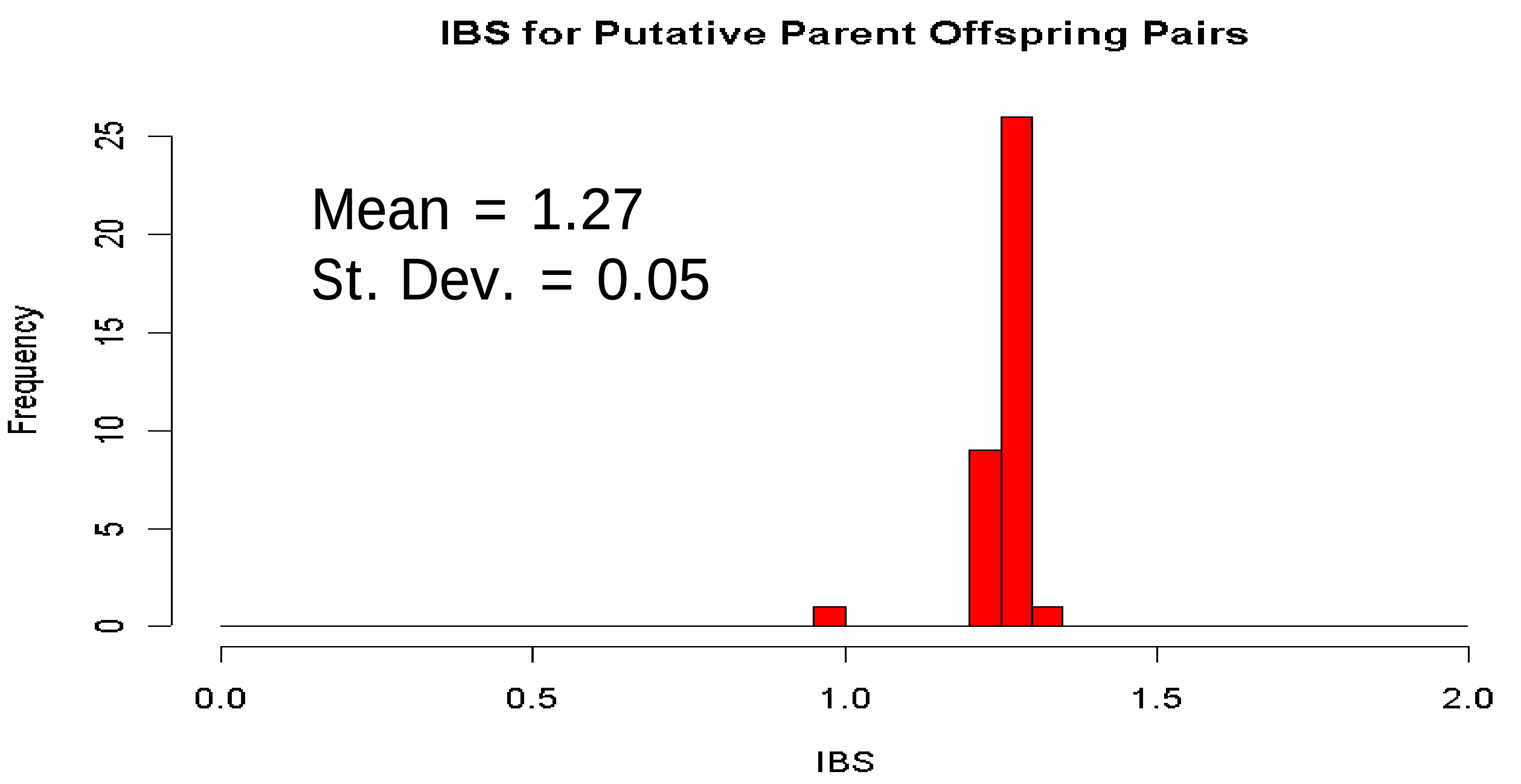
- ~800 marker genome scan
- Calculated IBS for each set of putative relationships…
  - Unrelated pairs
  - Sibling pairs
  - Parent-offspring pairs

Putative Unrelated Pairs



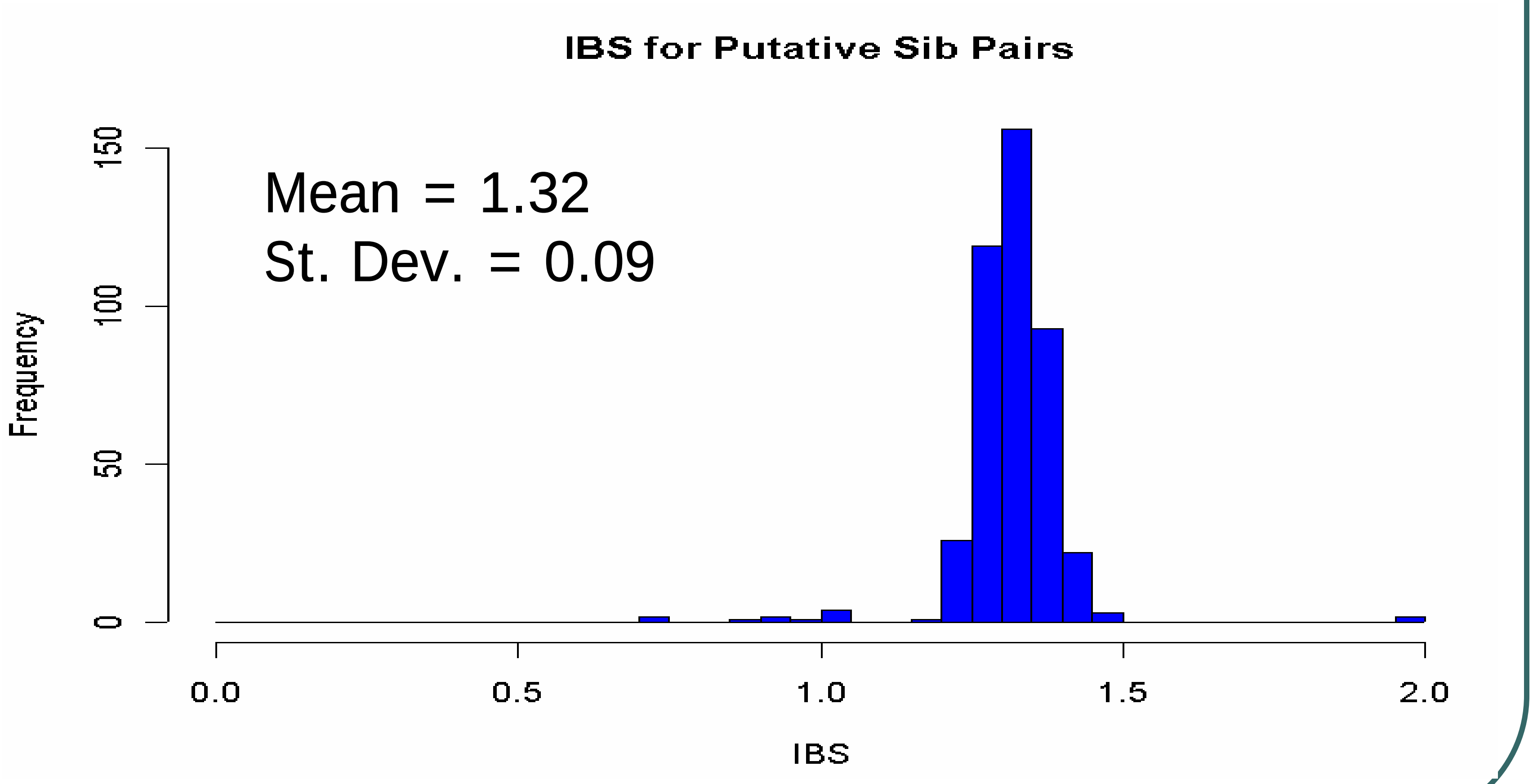
Mean = 0.87
St. Dev. = 0.07

Parent-Offspring Pairs

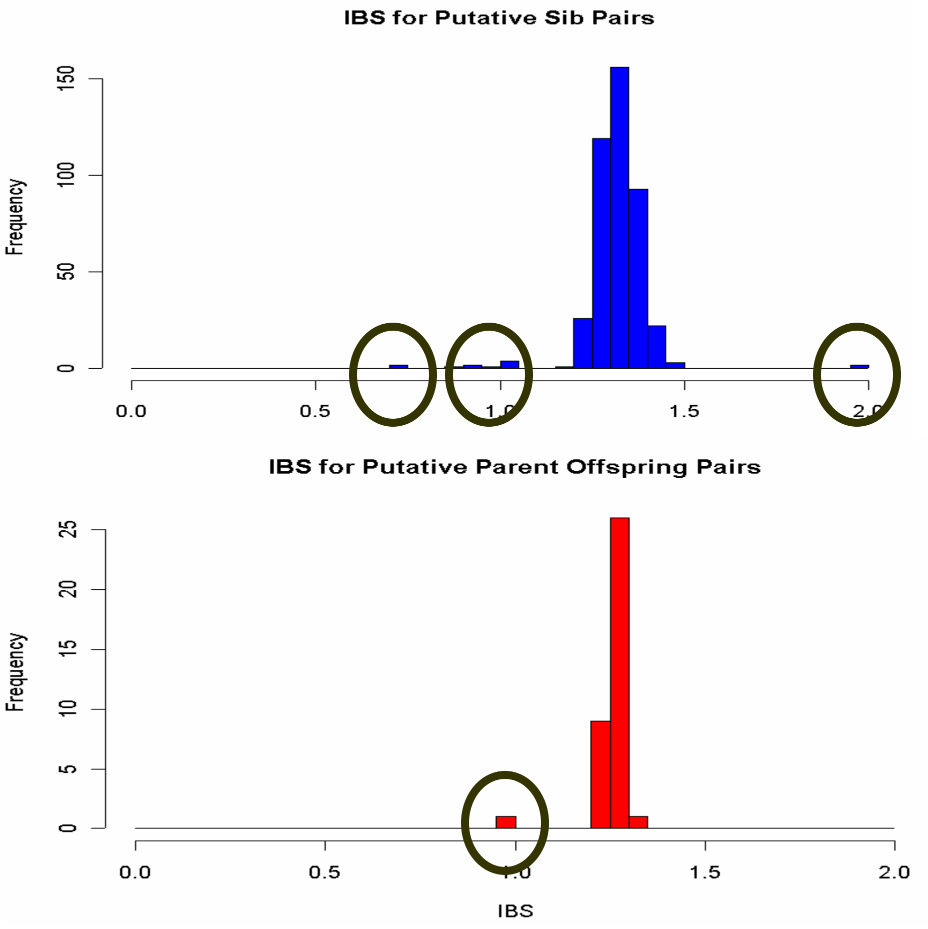


Mean = 1.27
St. Dev. = 0.05

Putative Sibling Pairs

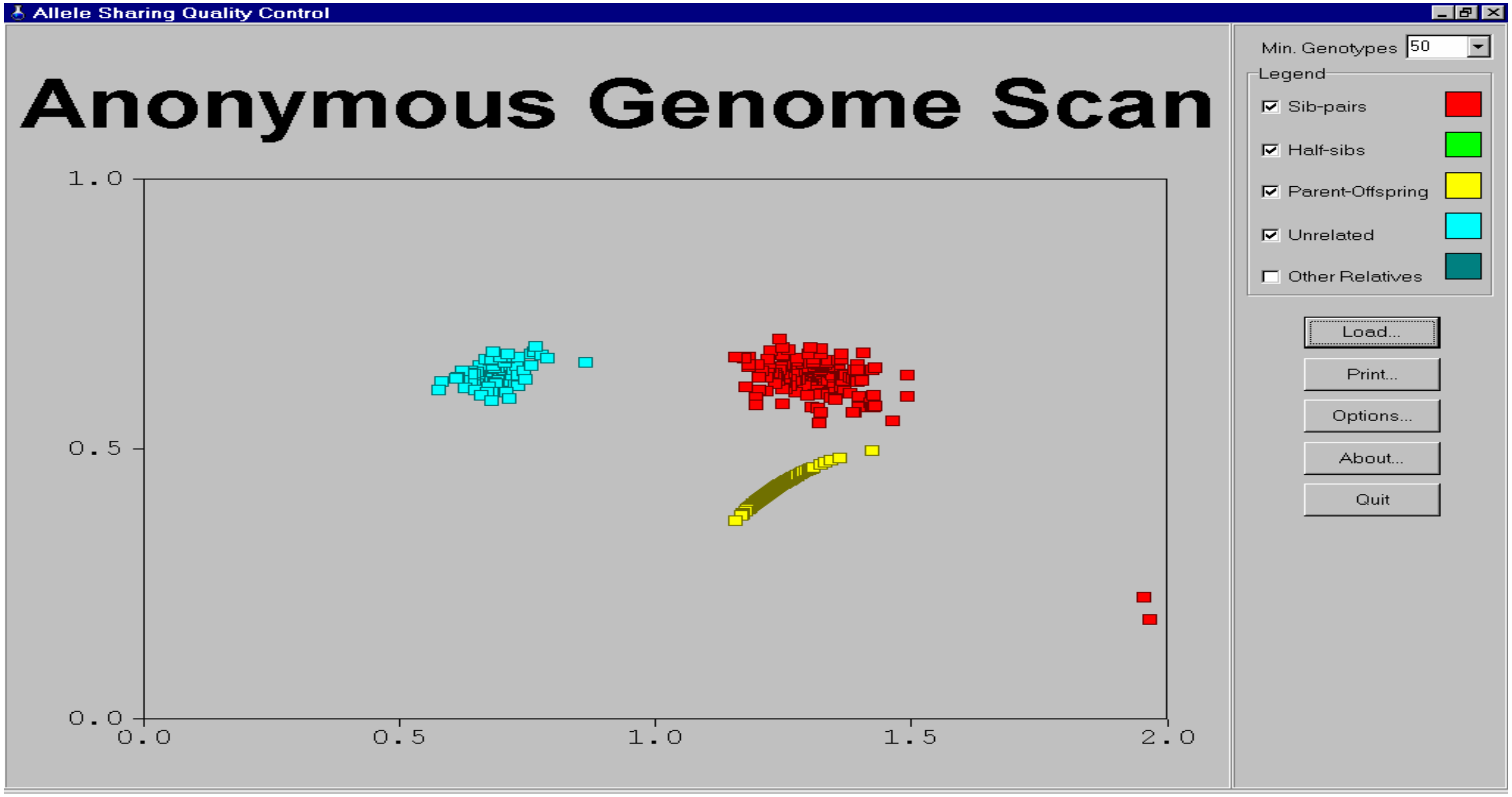


Problem Individuals Are Outliers



Circled pairs are likely misclassified

Additional Information in Standard Deviation of IBS Sharing



Allele Sharing Quality Control

Min. Genotypes 50

Legend

☑ Sib-pairs 🟥

☑ Half-sibs 🟩

☑ Parent-Offspring 🟨

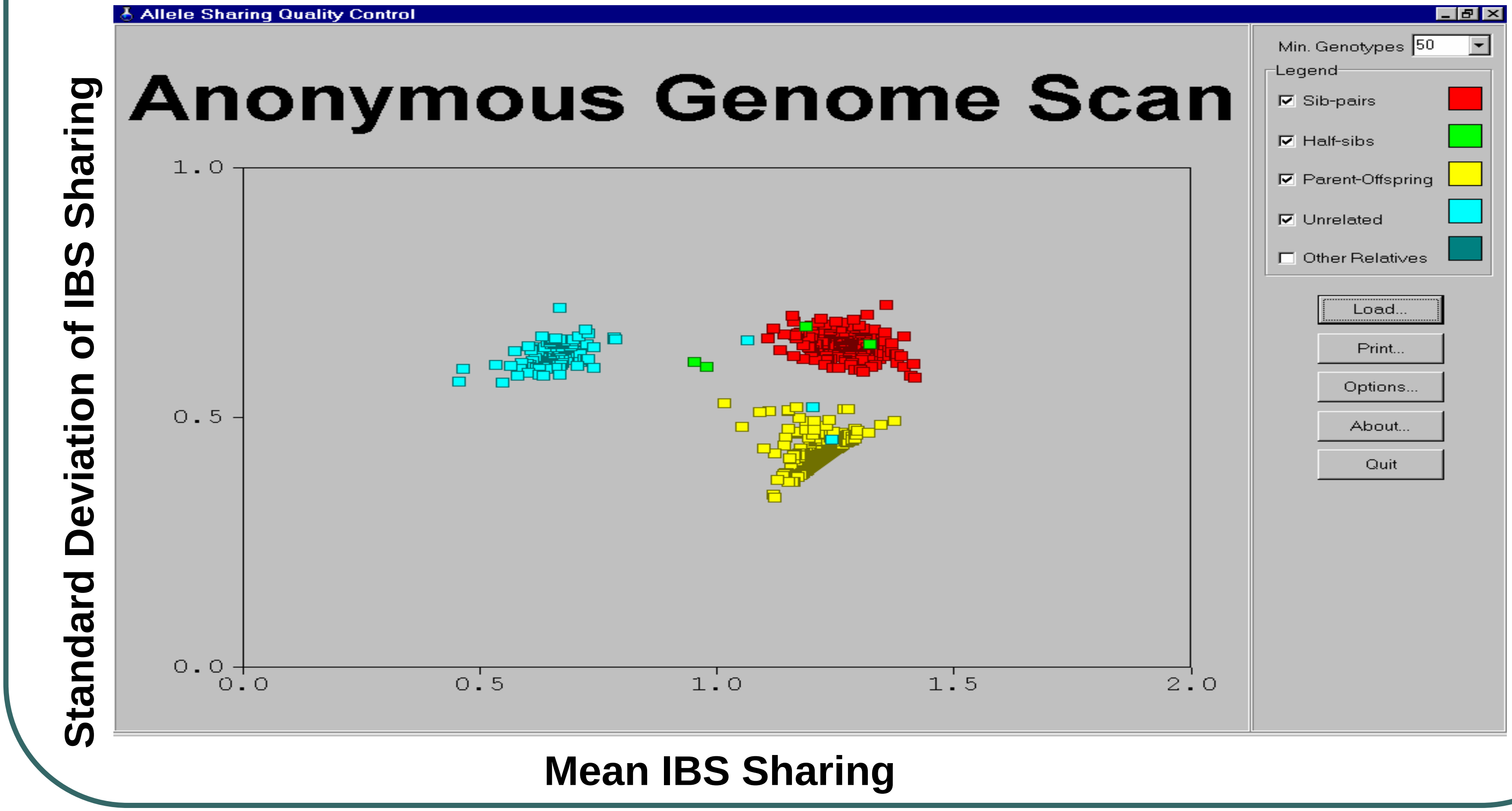
☑ Unrelated 🟦

☐ Other Relatives 🟦

Standard Deviation of IBS Sharing

Mean IBS Sharing

Additional Information in Standard Deviation of IBS Sharing



Min. Genotypes 50

Legend

☑ Sib-pairs 🟥

☑ Half-sibs 🟩

☑ Parent-Offspring 🟨

☑ Unrelated 🟦

□ Other Relatives 🟦

Problems with IBS Scores

- Inefficient
  - Ignore information on allele frequencies
  - Ignore correlations between neighboring markers

- … not too bad if large amounts of data available
  - Cannot distinguish some types of relatives

Strategy:

- Information we have:
  - X – observed genotypes at each marker
  - p – allele frequencies at each marker
  - θ - recombination fraction between consecutive markers

- P(X|R) for each possible relationship R
  - unrelated, half-sib, sib-pairs, MZ twins

Likelihood

- Sum over IBD states at each location

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

- Set of possible I changes with R

Notation

- R Hypothesized Relationship
- $I_k = (I_{km}, I_{kf})$ Allele sharing at locus *k*
- X_k Genotype pair at locus *k*

- $\alpha_k(j \mid R) = P(X_1, X_2, \ldots, X_{k-1}, I_k = j \mid R)$
  - Joint probability of data at first *k-1* markers and IBD vector $I_k=j$ at marker *k*

Details on I

- Possible inheritance patterns
  - (0,0) – no sharing
  - (1,0) – share maternal allele
  - (0,1) – share paternal allele
  - (1,1) – share both alleles
- For convenience, separate IBD=1 into maternal and paternal sharing states

Algorithm for Likelihood Calculation

$$\alpha_1(j \mid R) = P(I_1 = j \mid R)$$

$$\alpha_{k+1}(j \mid R) = \sum_i \alpha_k(i \mid R) P(X_k \mid I_k = i) t_k(i, j)$$

$$L = \sum_j \alpha_M(j \mid R) P(X_M \mid I_M = j)$$

Relationship between I and R

- Probability of I_1=(0,0), (1,0), (0,1) and (1,1):

MZ Twins	(0, 0, 0, 1)
Unrelated	?
Parent-Offspring	?
Full sibs	(¼, ¼, ¼, ¼)
Maternal half sibs	(½, ½, 0, 0)
Paternal half sibs	?

P(X|I) for pairs of individuals

GENOTYPE		**P(X_1,X_2|I) for I =**		
X_1	**X_2**	**(0,0)**	**(0,1) or (1,0)**	**(1,1)**
ii	*ii*	p_i^4	p_i^3	p_i^2
ii	*ij*	$2p_i^3 p_j$	$p_i^2 p_j$	*0*
ii	*jj*	$p_i^2 p_j^2$	*0*	*0*
ii	*jk*	$2p_i^2 p_j p_k$	*0*	*0*
ij	*ij*	$4p_i^2 p_j^2$	$p_i p_j (p_i + p_j)$	$2p_i p_j$
ij	*ik*	$4p_i^2 p_j p_k$	$p_i p_j p_k$	*0*
ij	*kl*	$4p_i p_j p_k p_l$	*0*	*0*

Transition Matrix
(Full Sibs)

$$\begin{array}{c|cccc} & (0,0) & (1,0) & (0,1) & (1,1) \\ \hline (0,0) & (1-\psi)^2 & (1-\psi)\psi & \psi(1-\psi) & \psi^2 \\ (1,0) & (1-\psi)\psi & (1-\psi)^2 & \psi^2 & \psi(1-\psi) \\ (0,1) & (1-\psi)\psi & \psi^2 & (1-\psi)^2 & (1-\psi)\psi \\ (1,1) & \psi^2 & (1-\psi)\psi & (1-\psi)\psi & (1-\psi)^2 \end{array}$$

$$r(i,j) = |i_1 - j_1| + |i_2 - j_2|$$

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

$$t(i,j) = \psi^{r(i,j)}(1-\psi)^{2-r(i,j)}$$

Transition Matrix (Maternal Half Sibs)

$$\begin{array}{c|cccc} & (0,0) & (1,0) & (0,1) & (1,1) \\ \hline (0,0) & (1-\psi) & \psi & 0 & 0 \\ (1,0) & \psi & (1-\psi) & 0 & 0 \\ (0,1) & 0 & 0 & 0 & 0 \\ (1,1) & 0 & 0 & 0 & 0 \end{array}$$

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

$$r(i, j) = | i_1 - j_1 |$$

$$t(i, j) = \psi^{r(i,j)}(1-\psi)^{1-r(i,j)}$$

Transition Matrix (Paternal Half Sibs)

$$\begin{array}{c|cccc} & (0,0) & (1,0) & (0,1) & (1,1) \\ \hline (0,0) & (1-\psi) & 0 & \psi & 0 \\ (1,0) & 0 & 0 & 0 & 0 \\ (0,1) & \psi & 0 & (1-\psi) & 0 \\ (1,1) & 0 & 0 & 0 & 0 \end{array}$$

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

$$r(i,j) = | i_2 - j_2 |$$

$$t(i,j) = \psi^{r(i,j)}(1-\psi)^{1-r(i,j)}$$

Transition Matrix (Unrelated)

$$\begin{array}{c|cccc} & (0,0) & (1,0) & (0,1) & (1,1) \\ \hline (0,0) & 1 & 0 & 0 & 0 \\ (1,0) & 0 & 0 & 0 & 0 \\ (0,1) & 0 & 0 & 0 & 0 \\ (1,1) & 0 & 0 & 0 & 0 \end{array}$$

Transition Matrix (MZ twins)

$$\begin{array}{c|cccc} & (0,0) & (1,0) & (0,1) & (1,1) \\ \hline (0,0) & 0 & 0 & 0 & 0 \\ (1,0) & 0 & 0 & 0 & 0 \\ (0,1) & 0 & 0 & 0 & 0 \\ (1,1) & 0 & 0 & 0 & 1 \end{array}$$

Example I

- Consider genotypes for one marker
- $X_1 = (1/1, 1/1)$
- Assume $p_1 = .2, .5$ or $.8$

- Calculate P(X|R) for each relationship
  - MZ twin, Full Sibs, Half-Sibs, Unrelated

Example II

- Consider genotypes for 2 markers
  - $X_1 = (1/1, 2/2)$
  - $X_2 = (1/1, 2/2)$
- Assume $p_1=p_2=½$
- Assume
  - $\theta = 0.0528, \psi = 0.10$
  - $\theta = 0.5000, \psi = 0.50$
- Calculate $P(X|R)$ for each relationship

Simulations (θ=.1, M=50)

	Inferred R		
True R	Full Sibs	Half Sibs	Unrelated
Full Sibs	0.914	0.085	0.001
Half Sibs	0.044	0.872	0.081
Unrelated	<.001	0.059	0.941

Simulations (θ=.2, M=50)

	Inferred R		
True R	Full Sibs	Half Sibs	Unrelated
Full Sibs	0.948	0.052	<.001
Half Sibs	0.038	0.899	0.064
Unrelated	<.001	0.062	0.938

Simulations (θ=.1, M=400)

	Inferred R		
True R	Full Sibs	Half Sibs	Unrelated
Full Sibs	1.000	<.001	<.001
Half Sibs	<.001	1.000	<.001
Unrelated	<.001	<.001	1.000

Bayesian Approach

- Alternative to simply maximizing *P*(*X*|*R*=*r*)

- Incorporates prior information on the expected frequency of each relative pair…

$$P(R = r \mid X) = \frac{Prior(R)P(X \mid R = r)}{\sum_{R} Prior(R)P(X \mid R)}$$

More distant relationships

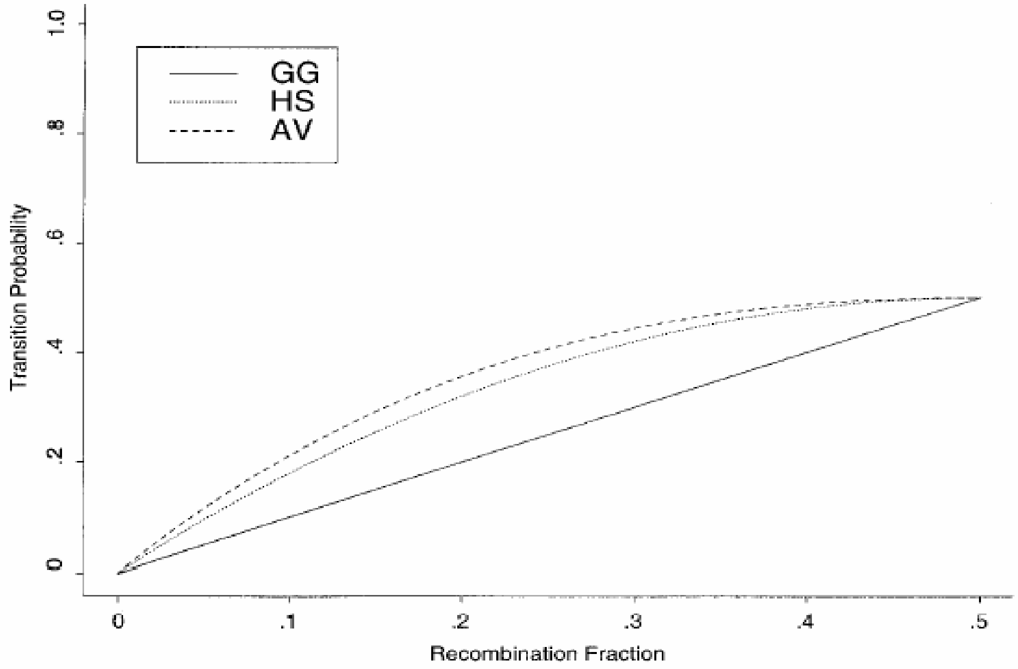


Figure 1 Autosomal transition probabilities for grandparent-grandchild (GG), half-sib (HS), and avuncular (AV) pairs. $P(I_{k+1} = 1 | I_k = 0) = P(I_{k+1} = 0 | I_k = 1)$ is shown. Note that $P(I_{k+1} = 0 | I_k = 0) = 1 - P(I_{k+1} = 1 | I_k = 0)$ and $P(I_{k+1} = 1 | I_k = 1) = 1 - P(I_{k+1} = 0 | I_k = 1)$.

Problem …

- Consider some genome scan data
  - 380 microsatellite markers
- Consider some pair of individuals
  - Putative siblings
- Observed Sharing
  - Identical for 379/380 genotype pairs
- $L(G|R=\text{MZ Twins}) = 0$
  - $L(G|R=\text{Any other}) > 0$

Solution:
Allow for Genotyping Errors

- A small proportion of errors could lead to misclassification
  - Allow for possibly erroneous genotypes
- ε – error rate parameter

$$
\begin{aligned}
&P(X_i \mid I_i) \\
&\qquad = \sum_{G_i} P(X_i \mid G_i, \varepsilon) P(G_i \mid I_i) \\
&\qquad = (1-\varepsilon)^2 P(G_i = X_i \mid I_i) + \left[1-(1-\varepsilon)^2\right] P(G_{i1} = X_{i1}) P(G_{i2} = X_{i2})
\end{aligned}
$$

Conclusions

- Likelihood approach provides reliable manner to infer relationships
- Can incorporate multiple linked markers
  - Some distant relationships can only be discerned by likelihood approach

Today

- Checking of Relationships for Pairs of Individuals

- Multipoint algorithm for calculating likelihoods for genotype data

Recommended Reading

- Boehnke and Cox (1997), *Am J Hum Genet* **61:**423-429
- Optional
  - Broman and Weber (1998), *AJHG* 63:1563-4
  - McPeek and Sun (2000), *AJHG* 66:1076-94
  - Epstein et al. (2000), *AJHG* 67:1219-31