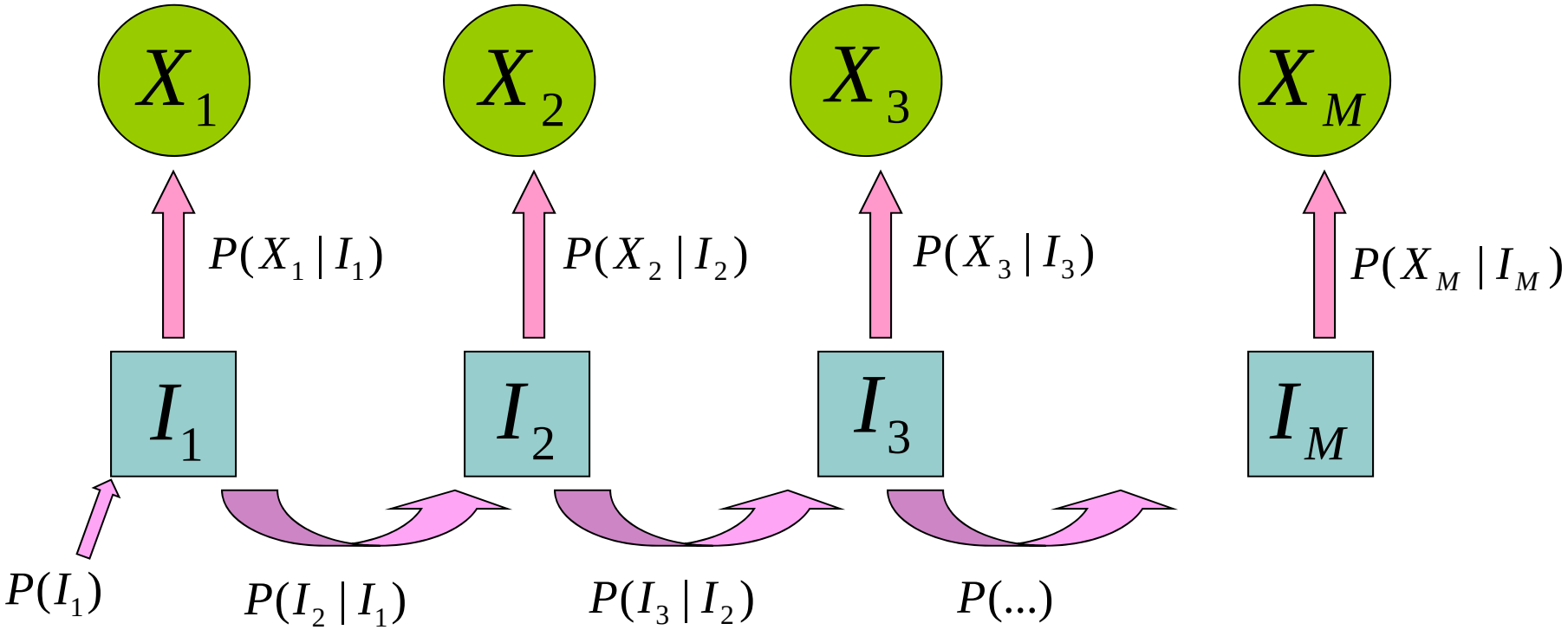
Relationship Checking

Biostatistics 666

Last Lecture

- Multipoint analysis
- Combining information across multiple markers for linkage analysis in sib pairs
- Markov Chain framework allows efficient computation

Ingredients of Multipoint Analysis



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

Last Lecture's 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?

Solution

I_1	I_2	$P(I_1)$	$P(I_2\|I_1)$	$P(X_1\|I_1)$	$P(X_2\|I_2)$	**P**
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	0.50	0.15	0.125	0.0625	**0.00058**
1	1	0.50	0.70	0.125	0.125	**0.00551**
1	2	0.50	0.15	0.125	0.25	**0.00231**
2	0	0.25	0.03	0.25	0.0625	**0.00013**
2	1	0.25	0.30	0.25	0.125	**0.00231**
2	2	0.25	0.67	0.25	0.25	**0.01051**

Solution

- Taking into account all available genotype data…
  - $P(I_1 = 2) = 0.57$
  - $P(I_1 = 1) = 0.37$
  - $P(I_1 = 0) = 0.06$
- Considering only one marker, the corresponding probabilities would be 0.44, 0.44 and 0.11.

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

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

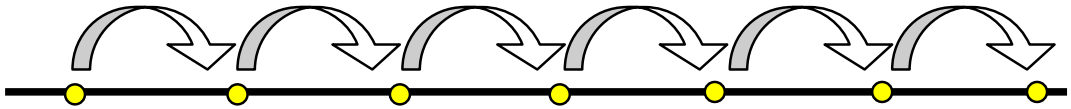
$$R_M(I_M) = 1$$

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

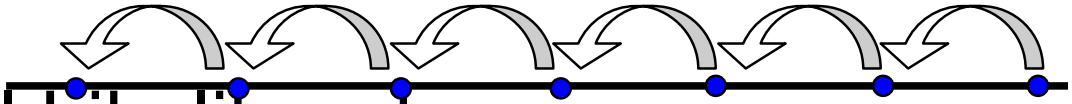
Pictorial Representation

- Single Marker

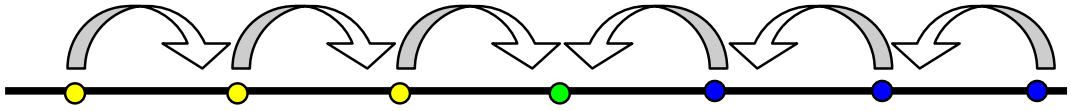
- Left Conditional



- Right Conditional



- Full Likelihood



Today's Lecture

- Verifying Genetic Relationships
- Multipoint Analysis for Different Relatives
  - Relationship changes transition probabilities
  - Relationship changes potential IBD states
- Approaches for Relationship Checking

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[†]

Verifying relationships is crucial

- Genetic analyses require relationships to be specified
- Misspecified relationships lead to tests of inappropriate size
  - Inflated Type I error
  - Decreased power

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

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

- Different relationships imply
  - Different $P(I_1)$
  - Different $P(I_i \mid I_{i-1})$

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

- For convenience, separate IBD=1 into maternal and paternal sharing states

- Possible inheritance patterns
  - (0,0) – no sharing
  - (1,0) – share maternal allele
  - (0,1) – share paternal allele
  - (1,1) – share both alleles

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 ?

$P(X_m \mid I_m)$

		IBD		
Sib	**CoSib**	(0,0)	(0,1) or (1,0)	(1,1)
(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

Note: Assuming unordered genotypes

Transition Matrix $P(I_i \mid I_{i-1})$ (Full Sibs)

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

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

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

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

Transition Matrix $P(I_i \mid I_{i-1})$ (Maternal Half Sibs)

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

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

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

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

Transition Matrix $P(I_i \mid I_{i-1})$ (Paternal Half Sibs)

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

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

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

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

Transition Matrix $P(I_i \mid I_{i-1})$ (Unrelated)

	(0,0)	(1,0)	(0,1)	(1,1)
(0,0)	1	0	0	0
(1,0)	0	0	0	0
(0,1)	0	0	0	0
(1,1)	0	0	0	0

Transition Matrix $P(I_i \mid I_{i-1})$ (MZ twins)

	(0,0)	(1,0)	(0,1)	(1,1)
(0,0)	0	0	0	0
(1,0)	0	0	0	0
(0,1)	0	0	0	0
(1,1)	0	0	0	1

Example I

- Consider genotypes for one marker
- Let X = (1/1, 1/1)
- Assume $p_1 = .5$

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

- How do results change with p_1?

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
  - θ = 0.0528, ψ = 0.10
  - θ = 0.5000, ψ = 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	.914	.085	.001
Half Sibs	.044	.872	.081
Unrelated	<.001	.059	.941

Simulations (θ=.2, M=50)

	Inferred R		
True R	**Full Sibs**	**Half Sibs**	**Unrelated**
Full Sibs	.948	.052	<.001
Half Sibs	.038	.899	.064
Unrelated	<.001	.062	.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

- Given some prior information on the expected frequency of each relative pair…
  - Alternative to simply maximizing $P(X|R=r)$

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

More distant relationships

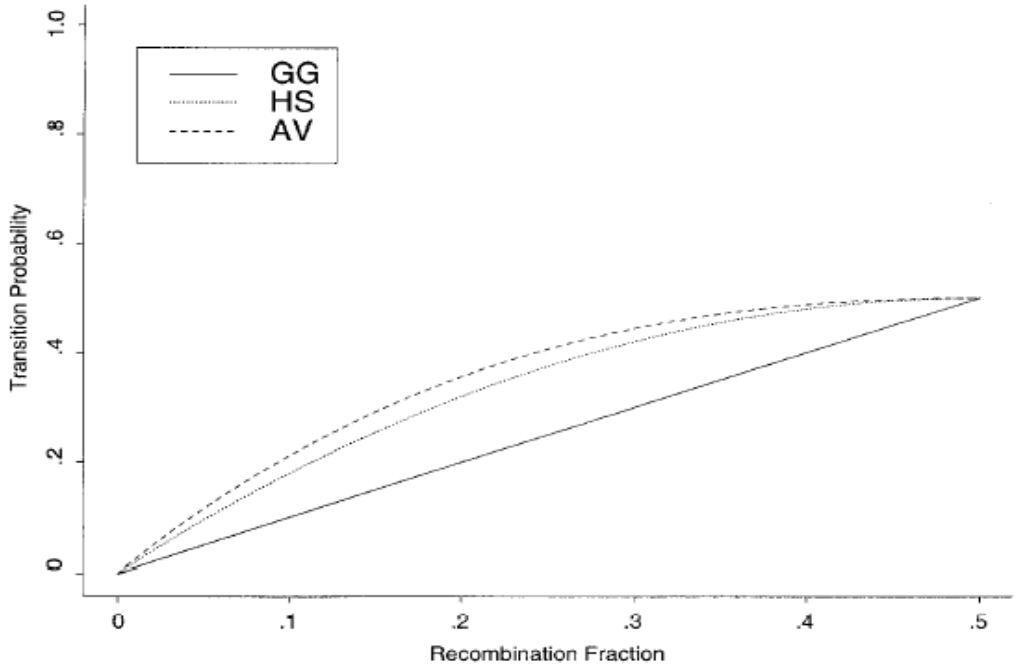


Figure 1 Autosomal transition probabilities for grandparent-grandchild (GG), half-sib (HS), and avuncular (AV) pairs. $P(I_{k+1} = 1 | I_k = 0) = P(I_{k+1} = 0 | I_k = 1)$ is shown. Note that $P(I_{k+1} = 0 | I_k = 0) = 1 - P(I_{k+1} = 1 | I_k = 0)$ and $P(I_{k+1} = 1 | I_k = 1) = 1 - P(I_{k+1} = 0 |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

Problem …

- Consider some genome scan data
  - 380 microsatellite markers
- Consider some pair of individuals
- Observed Sharing
  - Identical for 379/380 genotype pairs
- L(X|R=MZ Twins) = 0
  - L(X|R=Any other) > 0

Solution: Allow for Genotyping Errors

- Even a few errors can lead to misclassification
  - If likelihood formulation ignores errors
  - Need to update likelihood to allow errors
- ε – error rate parameter models difference between true genotypes G and observed genotypes X

$$P(X_i \mid I_i)$$

$$= \sum_{G_i} P(X_i \mid G_i, \varepsilon) P(G_i \mid I_i)$$

$$= (1-\varepsilon)^2 P(G_i \mid I_i) + \left[1-(1-\varepsilon)^2\right] P(X_{i1}) P(X_{i2})$$

Conclusions

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

Alternative Strategy I: Mendelian Inconsistencies

- Verify that observed genotypes are compatible with Mendelian segregation

- Common checks:
  - Does each putative offspring get one allele from each parent?
  - Is the set of genotypes in a putative sibship compatible?

Mendelian Checks

- If many markers exhibit incompatibilities, there may be a pedigree problem.
- Requires informative markers and relatively complete pedigrees
  - With only a sibling pair, any genotype pair fits
  - With bi-allelic markers, all genotypes fit any sibship
- Does not pinpoint source of error or suggest correct relationship

Alternative Strategy II: Allele Sharing Methods

- Summarize IBS sharing across all available markers
- Compare observed values for each pair to expected values
  - Expectations derived by examining other pairs with the same putative relationship

IBS Sharing Scores

- S_k – IBS score (0,1,2) for marker k

$$\bar{S} = \frac{\sum_k S_k}{n_{markers}} \qquad s^2 = \frac{\sum_k (S_k - \bar{S})^2}{n_{markers} - 1}$$

Could construct a Z-score

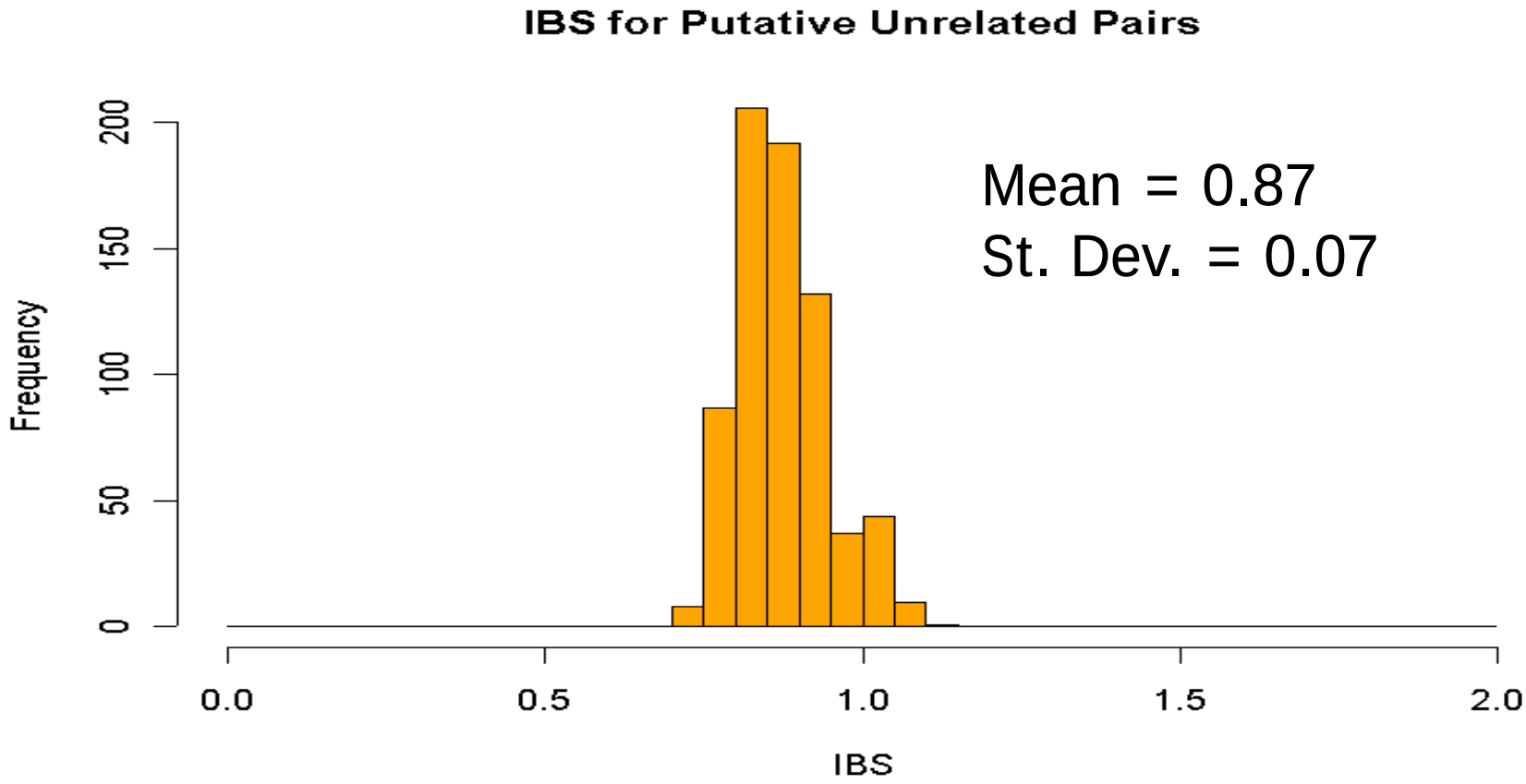
- Comparing observed IBS score to expected values within class of relatives

$$Z = \frac{\bar{S} - E(\bar{S} \mid R)}{\sqrt{Var(\bar{S} \mid R)}}$$

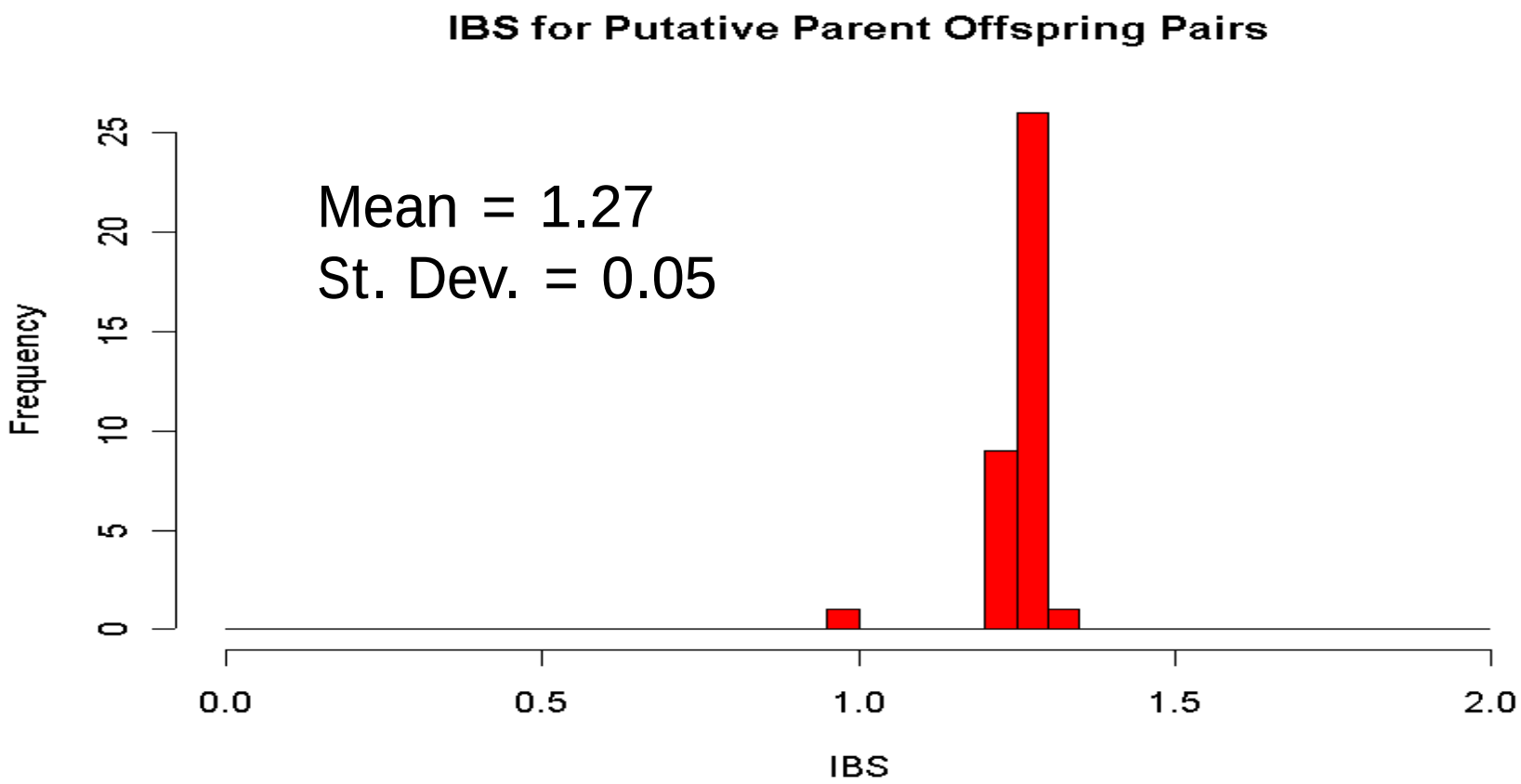
Example…

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

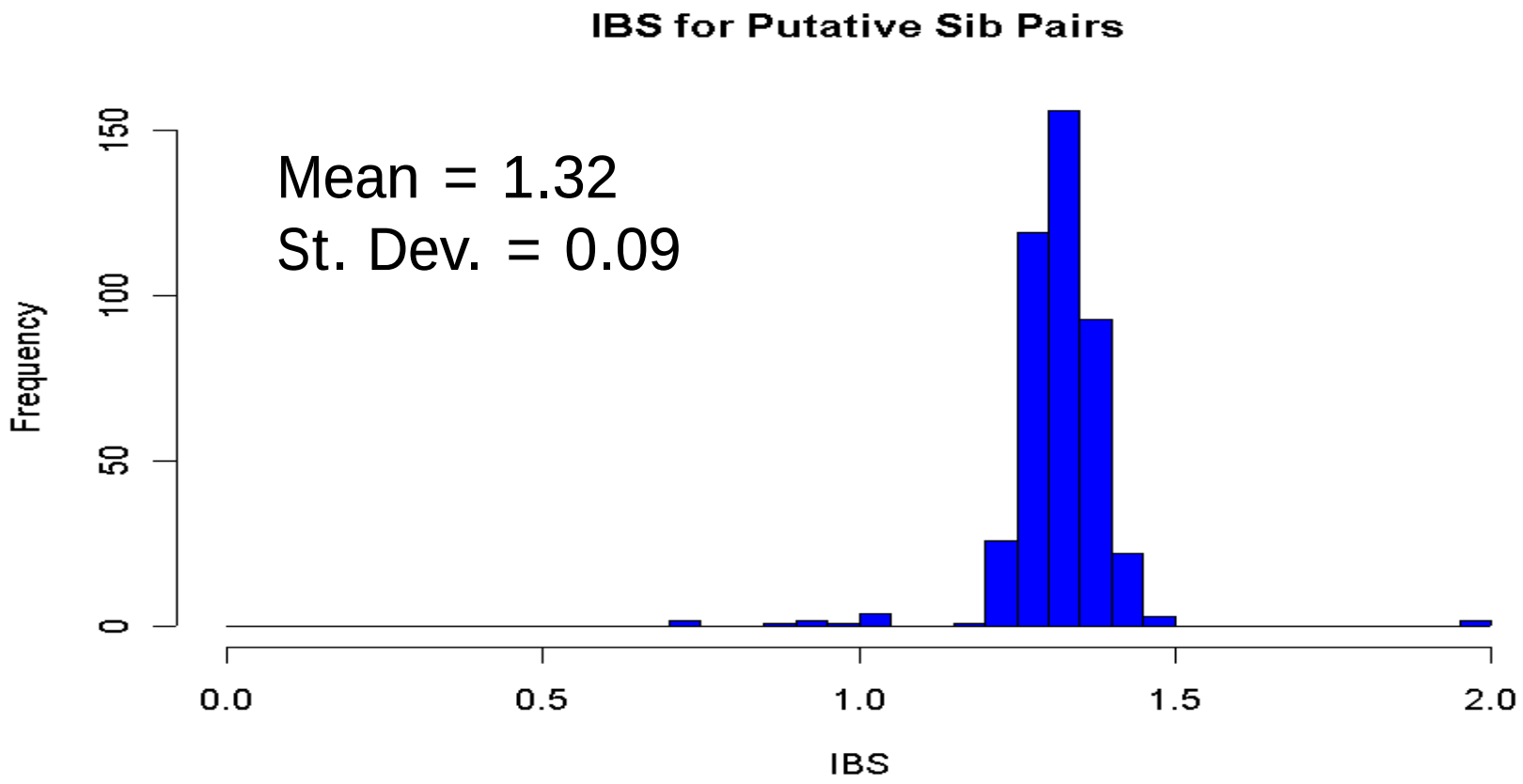
Putative Unrelated Pairs



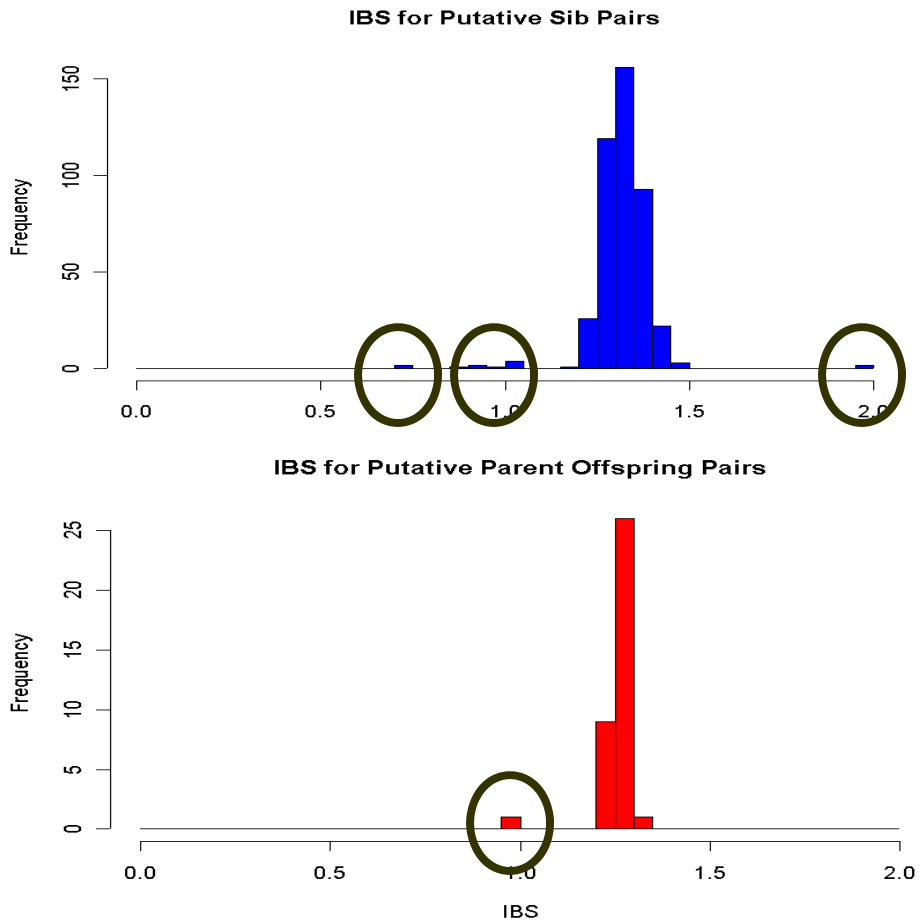
Parent-Offspring Pairs



Putative Sibling Pairs

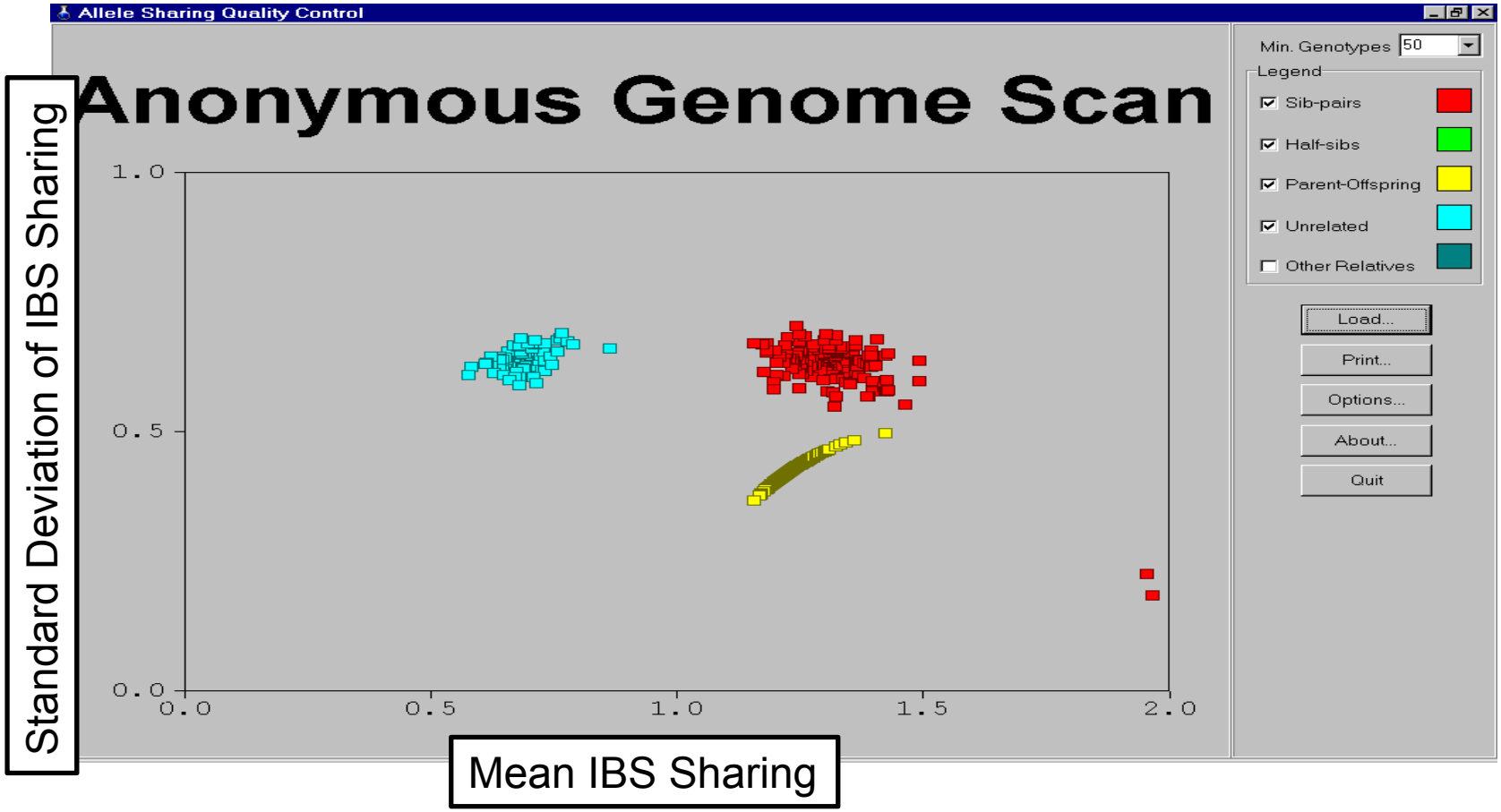


Problem Individuals Are Outliers

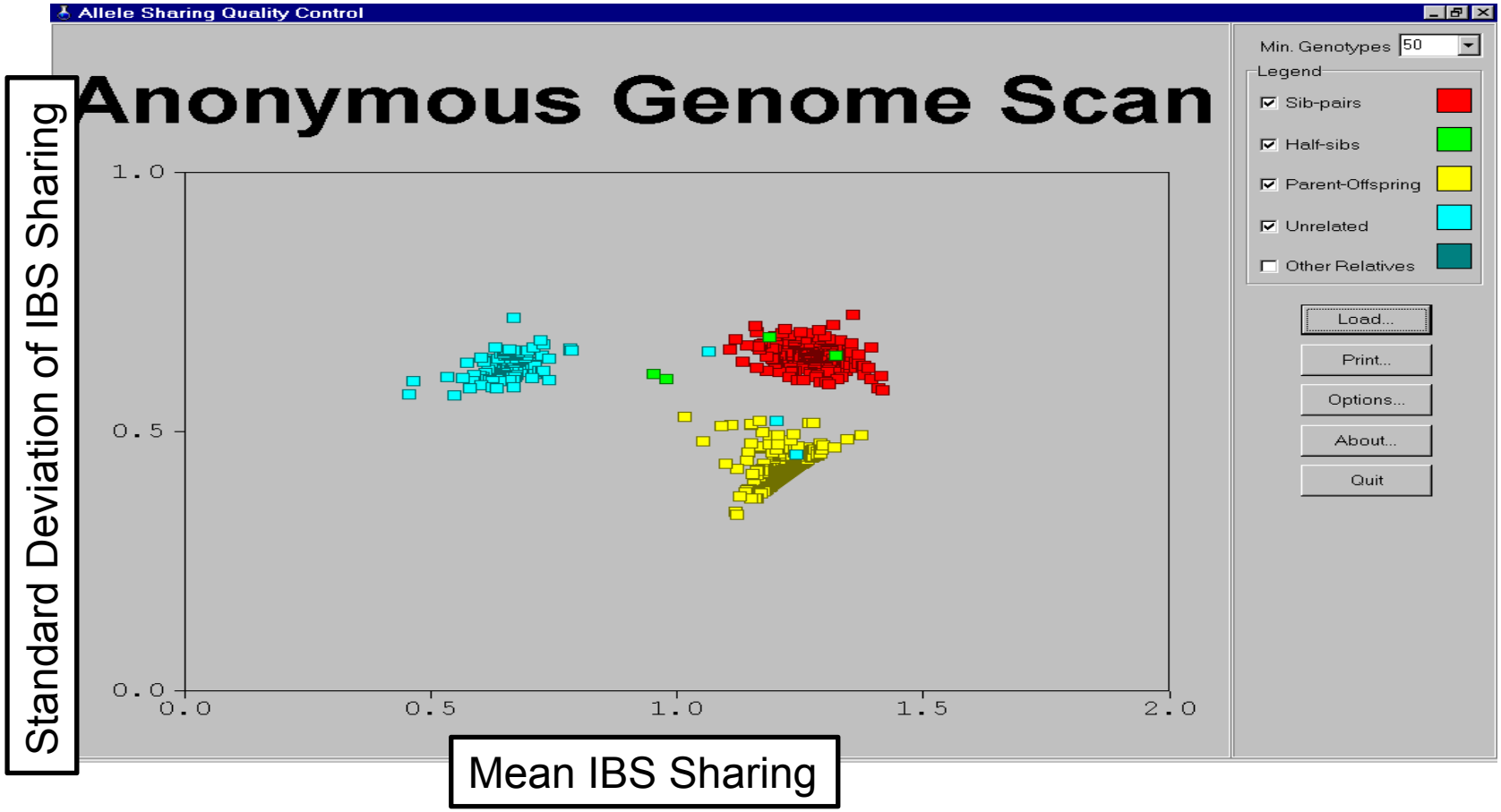


Circled pairs are likely misclassified

Additional Information in Variance of IBS Sharing



Additional Information in Variance of IBS Sharing



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

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