Kinship Coefficients

Biostatistics 666

Today’s Lecture

- Genetic analyses require relationships to be specified

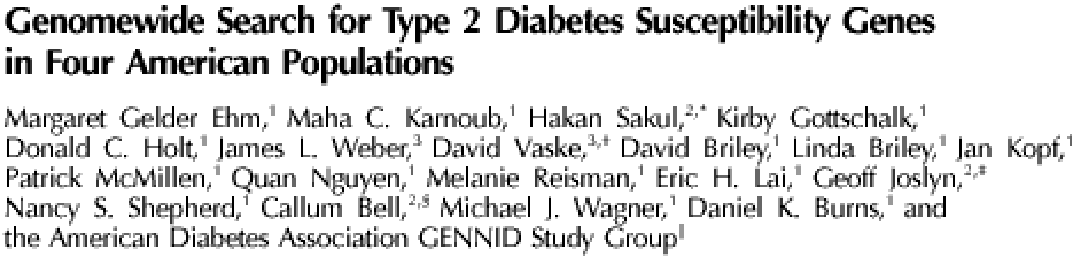
- Misspecified relationships lead to tests of inappropriate size
  - Inflated Type I error
  - Decreased power

- Kinship Coefficients

- Using data to verify genetic relationships

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[||]

Kinship Coefficients

- Summarize genetic similarity between pairs of individuals.

- Can be used to study relationship between genetic similarity and phenotypic similarity across individuals.

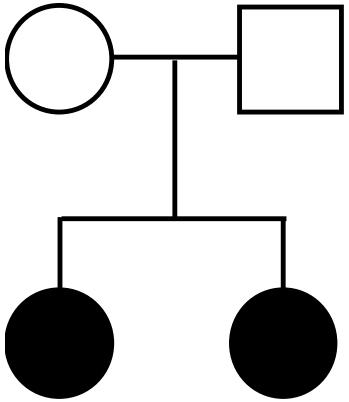
Kinship Coefficients – Definition

- Given two individuals
  - One with genes (g_i, g_j)
  - The other with genes (g_k, g_l)

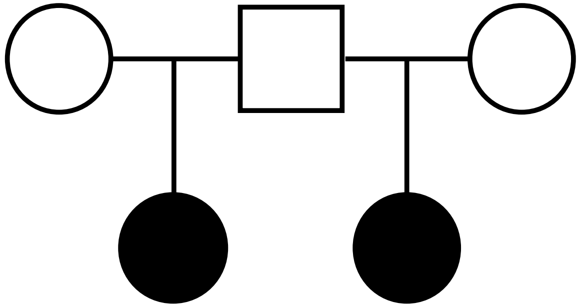
- The kinship coefficient is:
  - $\frac{1}{4}P(g_i \equiv g_k) + \frac{1}{4}P(g_i \equiv g_l) + \frac{1}{4}P(g_j \equiv g_k) + \frac{1}{4}P(g_j \equiv g_l)$
  - where "$\equiv$" represents identity by descent (IBD)

- Probability that alleles sampled at random from each individual are IBD

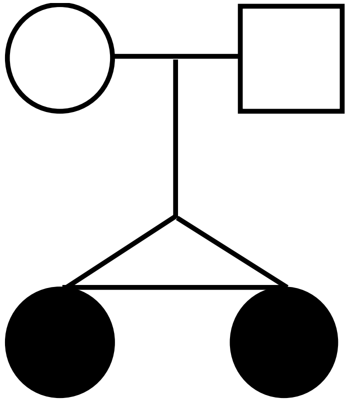
Some kinship coefficients



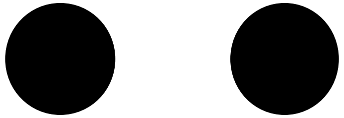
Siblings (φ=1/4)



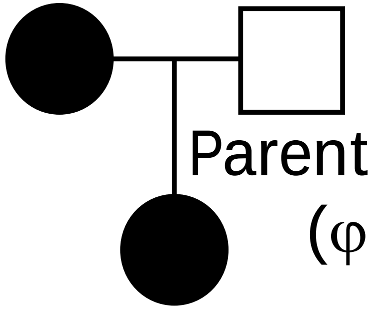
Half-Sibs (φ=1/8)



MZ Twins (φ=1/2)



Unrelated (φ=0)



Parent-Offspring
(φ=1/4)

What about other relatives?

- For any two related individuals i and j …
- … use a recursive algorithm allows calculation of kinship coefficient
- Algorithm requires an order for individuals in the pedigree where ancestors precede descendants
    - That is where for any i>j, i is not ancestor of j
    - Such an order always exists (e.g. the birth order!)

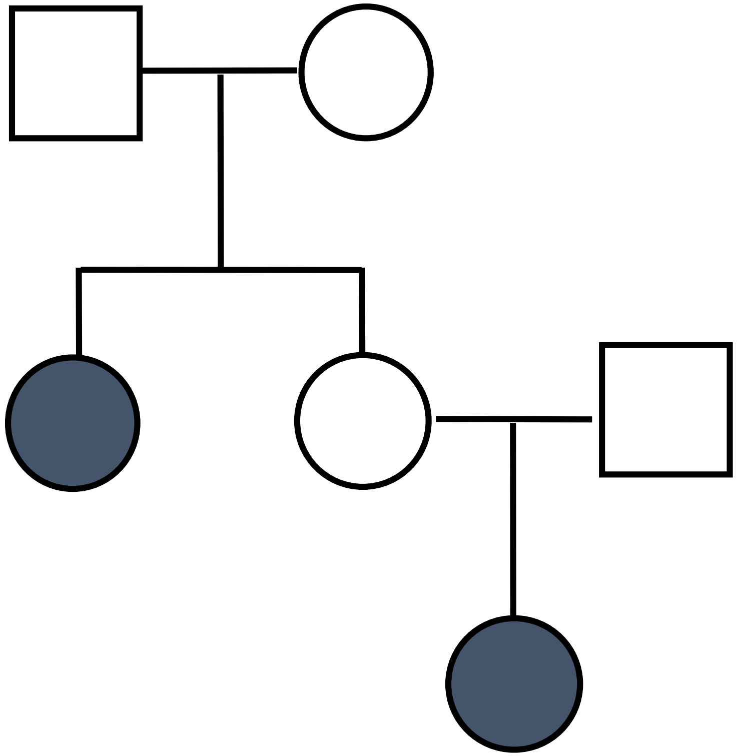
Computing Kinship Coefficients

- The recursive definition is then (for i ≥ j):

$$\varphi_{ij} = \begin{cases} 0 & i \text{ and } j \text{ are founders} \\ \tfrac{1}{2} & i = j, i \text{ is a founder} \\ \tfrac{1}{2}(\varphi_{mother(i)\,j} + \varphi_{father(i)\,j}) & i \neq j \\ \tfrac{1}{2}(1 + \varphi_{mother(i)\,father(i)}) & i = j \end{cases}$$

An example pedigree…

- Can you find …
- Suitable ordering for recursive calculation?
- Calculate kinship coefficient between shaded individuals?



Inbreeding Coefficients

- The kinship coefficient is related to the inbreeding coefficient

- If $\varphi_{ii} > 0.5$, individual *i* is inbred

- The inbreeding coeffient is $f_i = \varphi_{mother(i)father(i)} = 2(\varphi_{ii} - 0.5)$

- In most human populations, f_i is small – on the order of 0.001
  - Modifies probability of heterozygous genotypes to $2(1-f)p(1-p)$
  - Modifies probability of homozygous genotypes to $(1-f)p^2 + fp$

Verifying Relationships:
Strategy I - Allele Sharing Methods

- For each pair, summarize allele sharing across all markers
  - Specifically, average number of identical alleles at each marker pair
  - Number of alleles shared between two genotypes is the “identify-by-state”

- Compare observed values for each pair to expected values
  - Expected values derived by assessing all pairs with same putative relationship

IBS Sharing Scores

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

$$\bar{S} = \frac{\sum_{k} S_k}{n_{markers}}$$

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

Could construct a Z-score

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

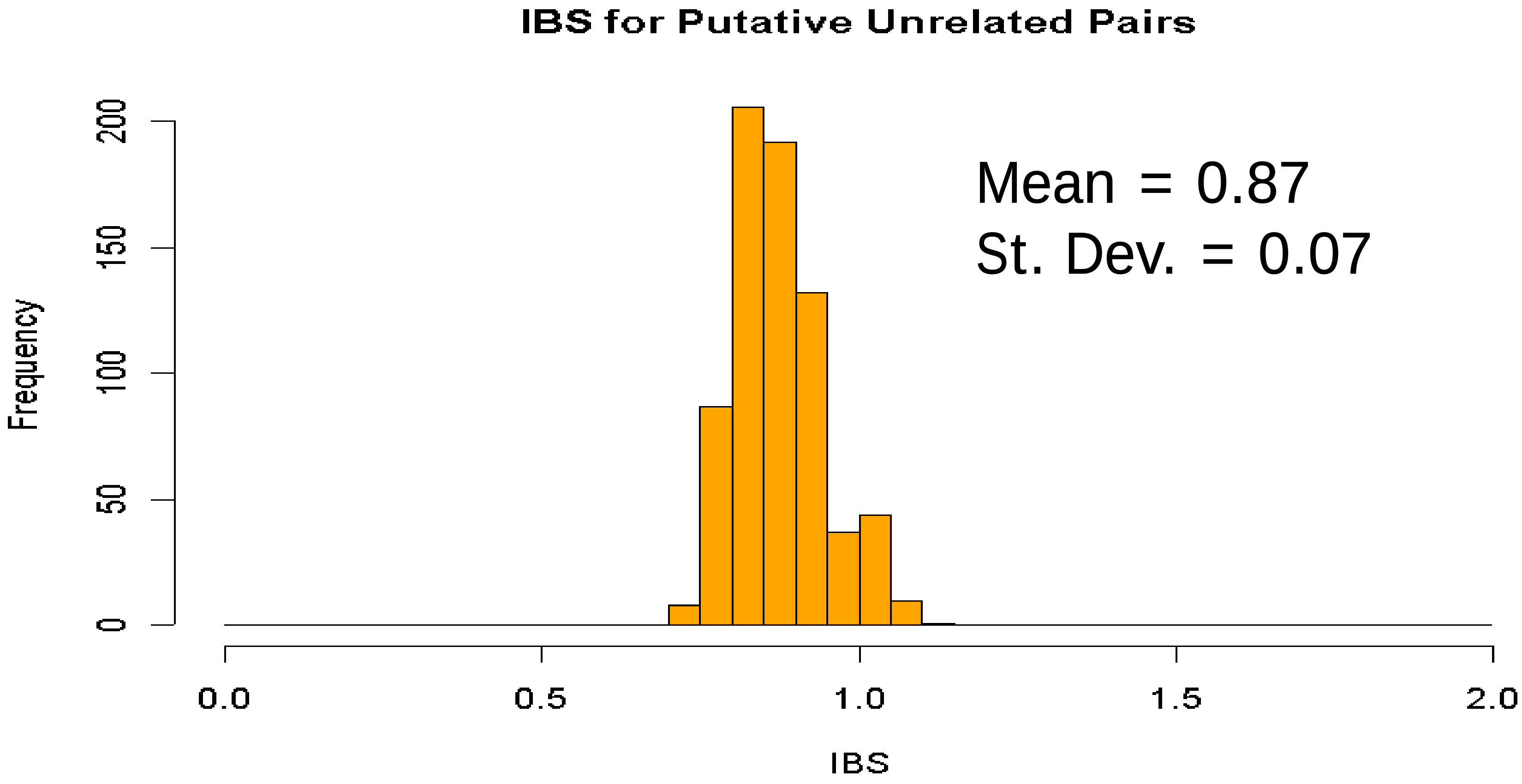
$$Z = \frac{\overline{S} - E(\overline{S} \mid R)}{\sqrt{Var(\overline{S} \mid R)}}$$

Example…

- ~800 marker genome scan

- Calculated IBS for each set of putative relationships…
  - Unrelated pairs
  - Sibling pairs
  - Parent-offspring pairs

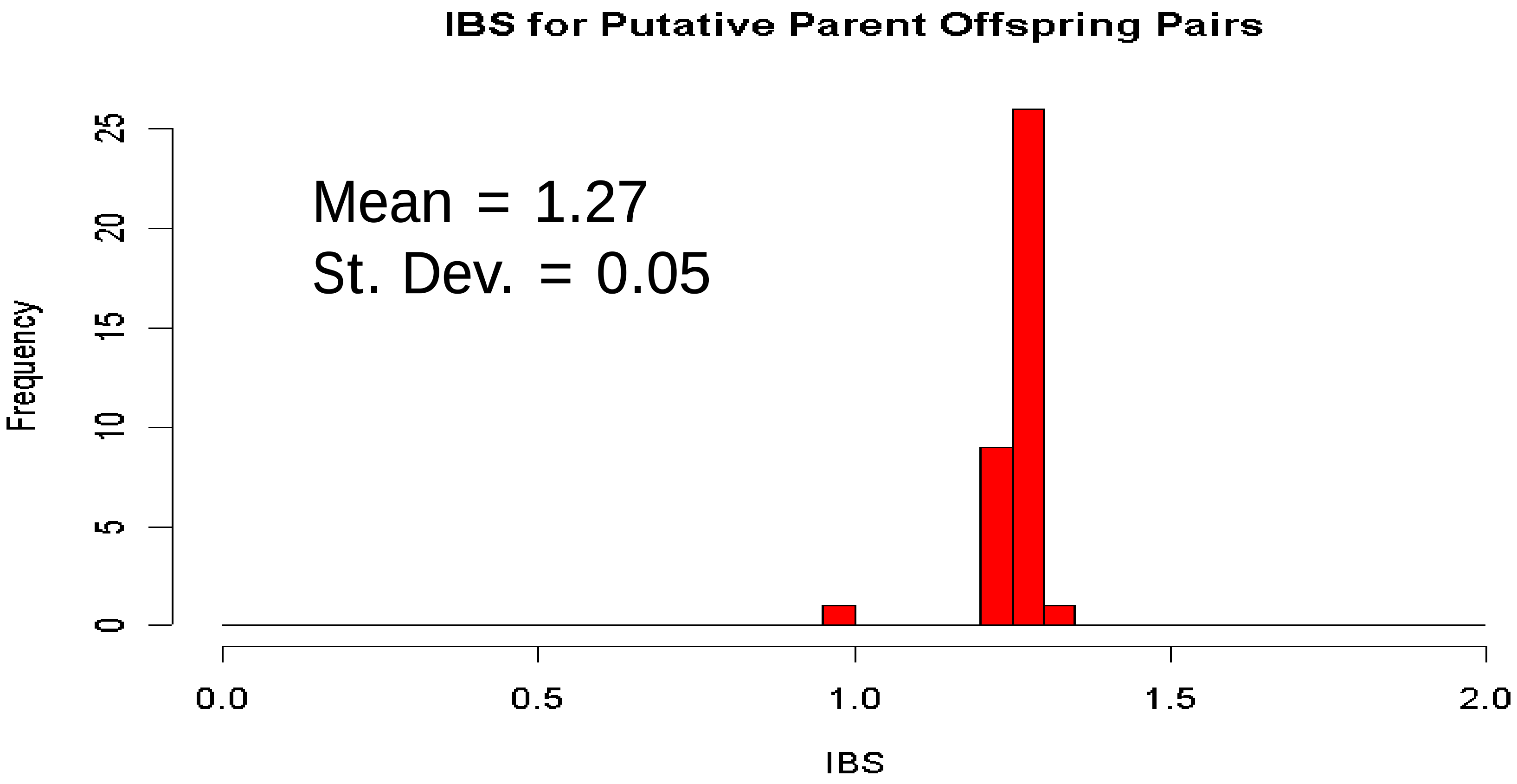
Putative Unrelated Pairs



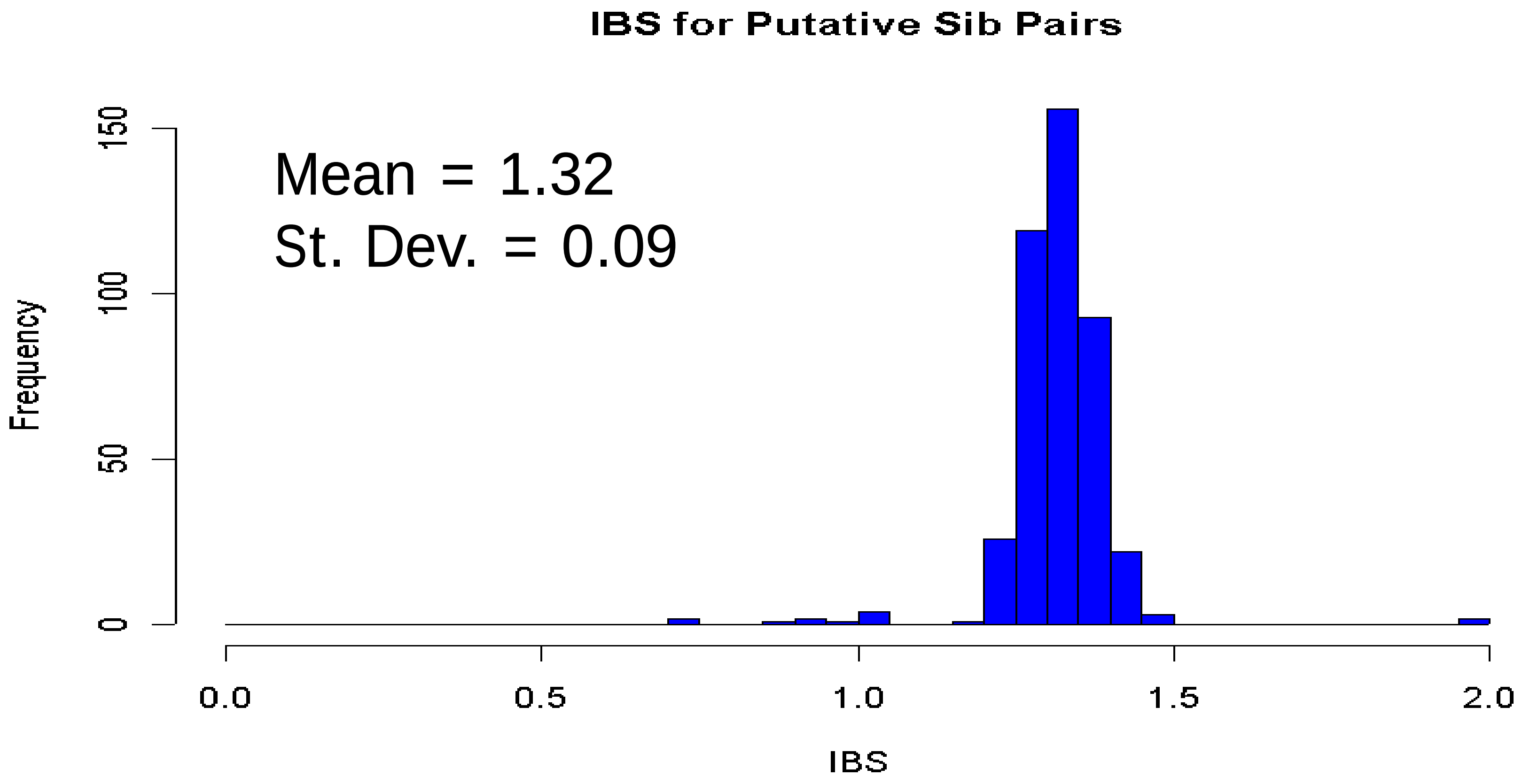
IBS for Putative Unrelated Pairs

Mean = 0.87
St. Dev. = 0.07

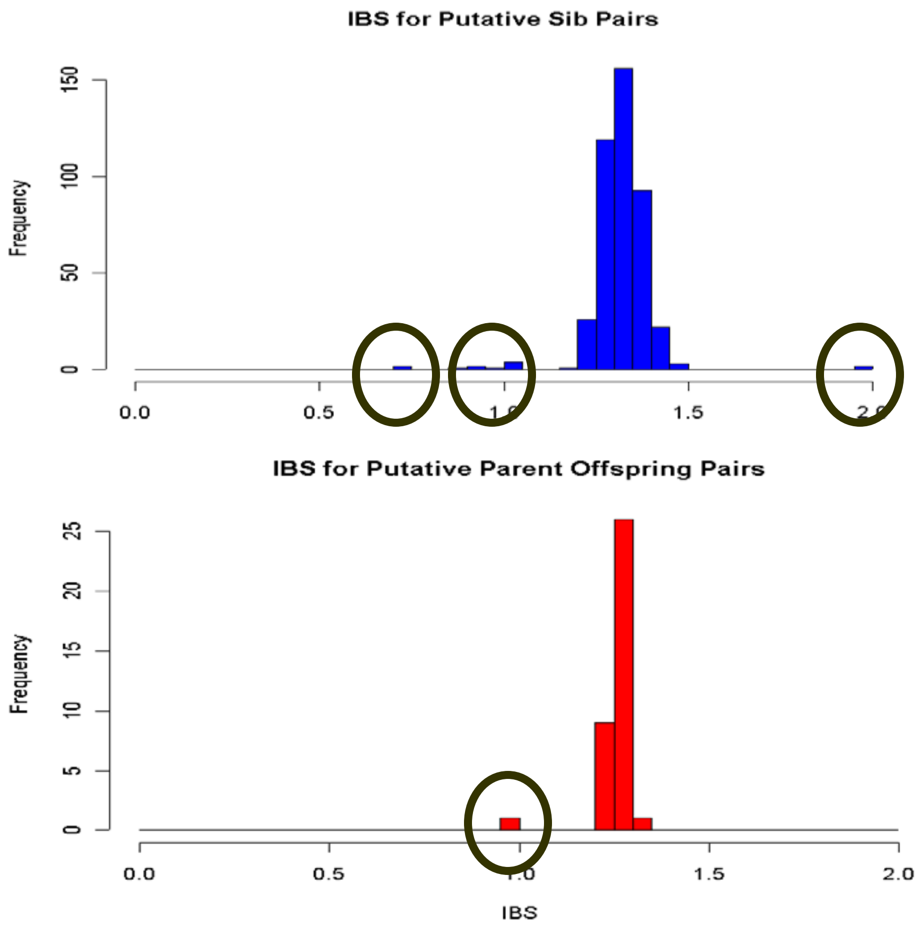
Parent-Offspring Pairs



Putative Sibling Pairs



Problem Individuals Are Outliers



Circled pairs
are likely
misclassified

Problems with IBS Scores

- Inefficient
    - Ignores information on allele frequencies
    - Ignores correlations between neighboring markers
- … work well if large amounts of data available
    - Cannot distinguish some types of relatives

Verifying Relationships: Strategy I - Likelihood Based Methods

- When evaluating sharing, take allele frequency into account
  - Place greater importance in sharing of rare alleles
  - Recognize that sharing of common alleles can occur by chance

- Choice of parameters to maximize and constraints on underlying probabilities

P (X_m | IBD)

		IBD		
Sib	**CoSib**	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

Example I

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

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

- How do results change with p_1?

Likelihood

$$L = \prod_{m=1}^{M} \sum_{k=0}^{2} P\left(G_{im}, G_{jm} \middle| IBD = k\right) P(IBD = k | relationship)$$

- Likelihood above assumes markers are independent
  - With smaller amounts of data, important to model recombination

- With large amounts of data, this works well

- Maximize probability of IBD=0, IBD=1, IBD=2
  - Or, often, just P(IBD=1) = $2\Phi_{ij}$ and P(IBD=0) = 1 - 2 Φ_{ij}

Simulations (M=50, 10 cM apart)

	Inferred R		
True R	**Full Sibs**	**Half Sibs**	**Unrelated**
Full Sibs	.914	.085	.001
Half Sibs	.044	.872	.081
Unrelated	<.001	.059	.941

Simulations (M=400, 10 cM apart)

	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

Weaknesses with likelihood approach…

- One weakness is that the approach is sensitive to genotyping error

- Consider some genome scan data
  - 380 microsatellite markers

- Observed Sharing
  - Identical for 379/380 genotype pairs

- L(G|R=MZ Twins) = 0
  - L(G|R=Any other) > 0

- How to resolve?

Solution: Allow for Genotyping Errors

- If likelihood ignores errors, even a few errors can lead to misclassification
    - Need to update likelihood to allow errors

- Introduce a distinction between true genotypes *G* and observed genotypes *X*
    - An error rate parameter, say ε, models the difference between the two

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

Weaknesses with likelihood approach…

- Another weakness is that the approach is sensitive to allele frequency estimates
- How can we make sure that we have chosen the right allele frequencies?
- Manichaikul et al (2010) proposed focusing on marker pairs that have configuration (a/a, b/b) or (a/b, a/b)
  - The ratio of these two configurations does not depend on allele frequencies!
  - However, it will depend on the ratio of P(IBD=1) to P(IBD=0)

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