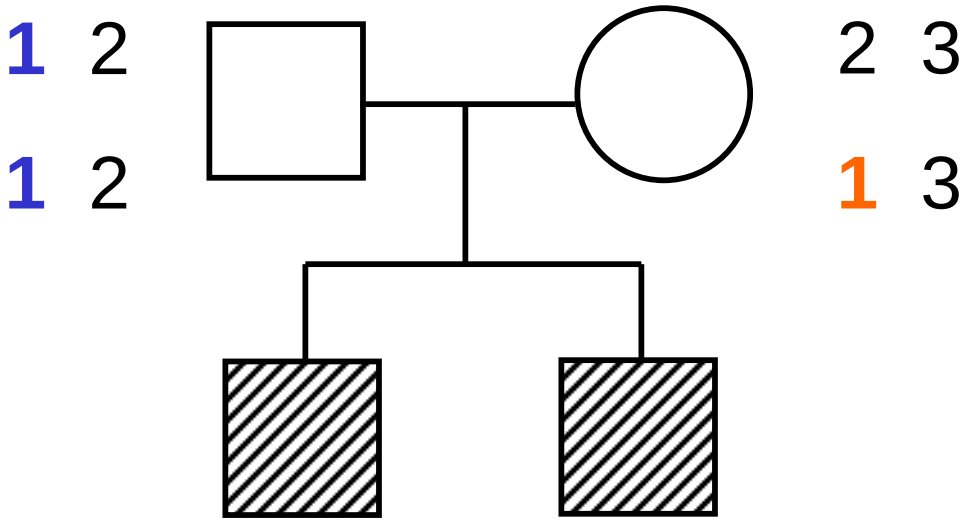
Estimation of allele frequencies with sibpair data

Karl W Broman
Department of Biostatistics
Johns Hopkins University

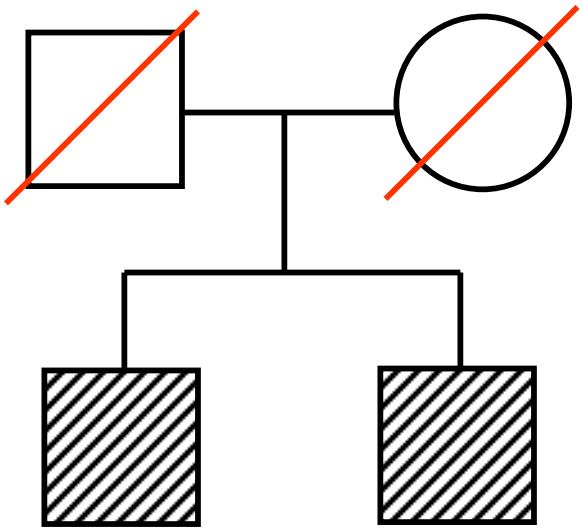
Identity by descent



1 3 ... 1 2

1 3 ... 1 2

Data



3	**4**	...	**4**	6
2	**6**	...	1	**6**
3	9	...	2	**3**
1	7	...	**1**	8
6	**6**	...	**6**	**6**
4	4	...	2	**4**
2	**8**	...	5	**8**
7	**7**	...	**7**	**7**
5	**5**	...	**5**	**5**
2	**6**	...	**2**	**6**
4	**6**	...	**4**	**6**

Goal

Estimate the allele frequencies at each of a set of linked genetic markers

Data on a set of (independent) sibling pairs; no parental genotypes

Model

- Hardy-Weinberg and linkage equilibrium
- Normal segregation (no disease gene)
- No interference in recombination
- No genotyping errors

Method 1

- Use one sibling from each pair

$$\text{var}\,\hat{p}^{(1)} = \frac{1}{2}\left(\frac{p(1-p)}{n}\right)$$

Method 2

- Use all individuals
- Ignore relationships

$$\text{var}\,\hat{p}^{(2)} = \frac{3}{8}\left(\frac{p(1-p)}{n}\right)$$

rel. eff. ≈ 1.33

Method 3

- Use all individuals
- Take account of their relationships
- One marker at a time

EM algorithm: weight each allele in 2nd sib by the probability it is not IBD with one of the 1st sib's alleles

Sib 1	Sib 2	Weight for g_{21}
1 1	1 1	$p_1/(1+p_1)$
1 1	1 2	$p_1/(1+p_1)$
1 2	1 3	$p_1/(1+p_1)$
1 2	1 1	$p_1/[2(1+p_1)]$

Method 4

- Use all individuals
- Take account of their relationships
- Use all linked markers together

EM algorithm w/ HMM technology:

- Weight each allele in 2nd sib as before
- Calculate probabilities (weights) condit'nal on genotype data for all markers
- Assuming no interference, the underlying IBD process follows a Markov chain

What we expect

- Methods 3, 4 better than methods 1, 2
- Method 3 improves with marker informativeness
- Method 4 improves with marker informativeness and marker density
- Method 4 better than method 3 when markers at very high density
- Relative efficiency = 1.50 (compared to method 1) is the best possible

Simulations

- 100 sib pairs
- 11 markers separated by d (= 0.1 – 15) cM
- Each marker has k alleles with frequencies (0.05, 0.10, 0.15, 0.20, p_4, …, p_k)
- p_4, …, p_k chosen to give specified heterozygosity (= 0.675 – 0.90)

$$\text{het} = 1 - \sum p_i^2$$

- Apply all methods to the first 4 alleles at each marker
- 2500 replicates per …

Ave relative efficiency

	Allele frequency			
Method	0.05	0.10	0.15	0.20
1		1.00		
2		1.33		
3	1.46	1.45	1.44	1.43
4	1.48	1.46	1.45	1.44

Unexpected but not surprising:

Relative efficiency depends on the allele frequency

Rel Eff: Method 3

	Allele frequency			
het	0.05	0.10	0.15	0.20
0.7	1.45	1.44	1.43	1.42
0.8	1.46	1.45	1.44	1.43
0.9	1.48	1.47	1.47	1.45

Rel Eff: Method 4

	Allele frequency			
d (cM)	0.05	0.10	0.15	0.20
0.1	1.50	1.49	1.48	1.48
1	1.49	1.48	1.47	1.46
5	1.48	1.46	1.44	1.44
10	1.47	1.45	1.43	1.42
(method 3)	1.46	1.45	1.44	1.43

For method 4:

- Marker heterozygosity had little effect
- Marker position had little effect

Summary

Method	Progr. time	CPU time	Rel. Eff.
1	2 min	1 msec	1.00
2	2 min	1 msec	1.33
3	1 morning	2 msec	1.45
4	1 afternoon	**2.5 sec**	1.46