Model misspecification

Estimating allele frequencies in sibships

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

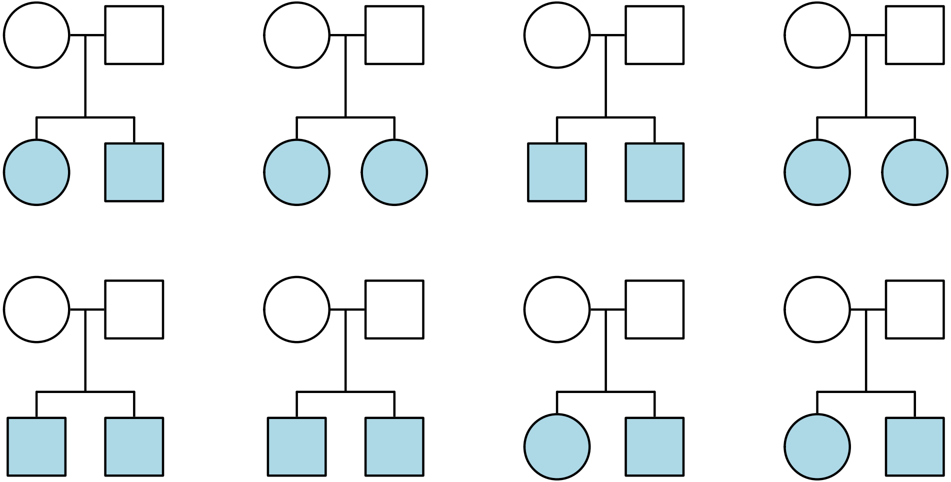
kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

Mapping disease genes

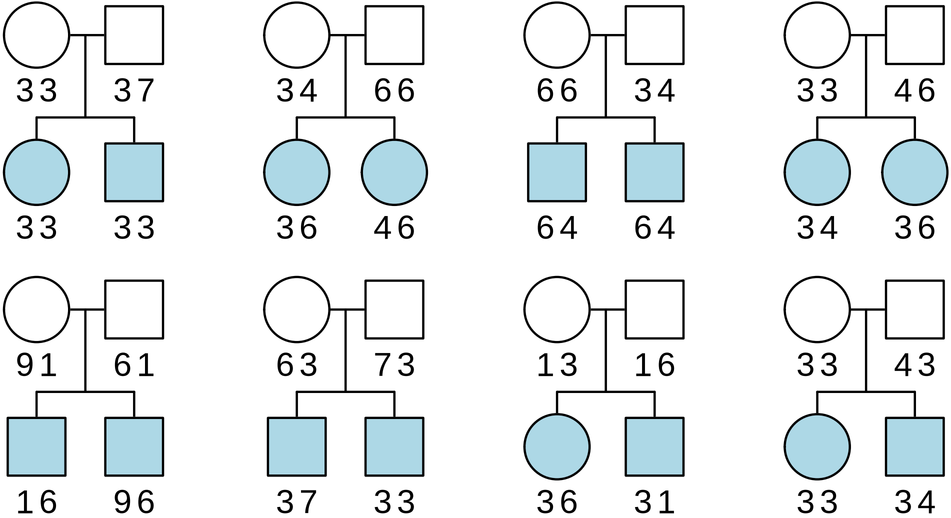
Look for genomic regions where

individuals with similar phenotypes
also have similar genotypes

Affected sib pairs



Affected sib pairs



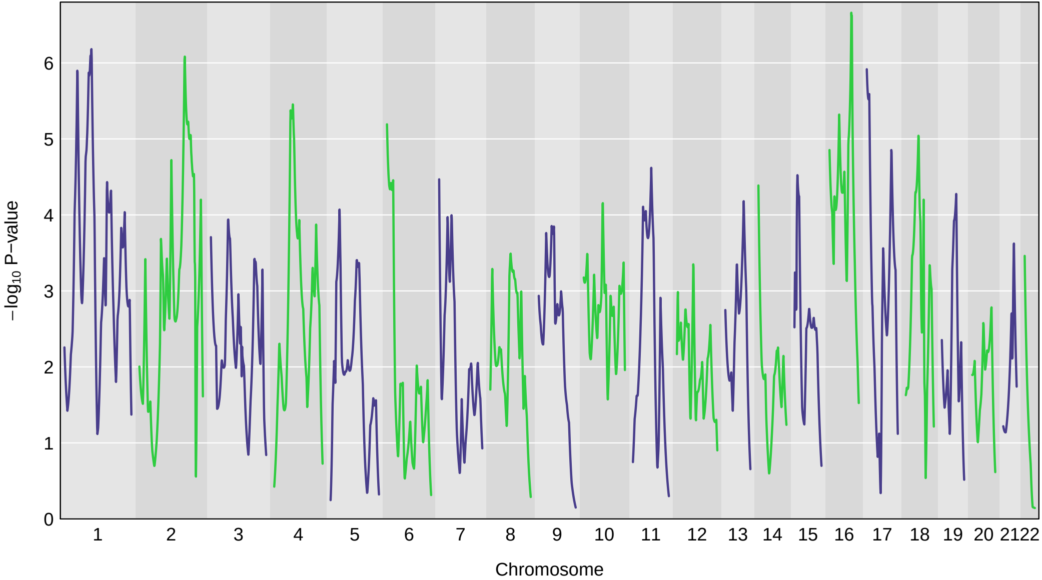
IBS vs IBD

IBS = identical by state
= same allele number

IBD = identical by descent
= copies of the same ancestral allele

non-inbred sibs are IBD = 0, 1, 2
with probability = 1/4, 1/2, 1/4

Prostate cancer genome scan

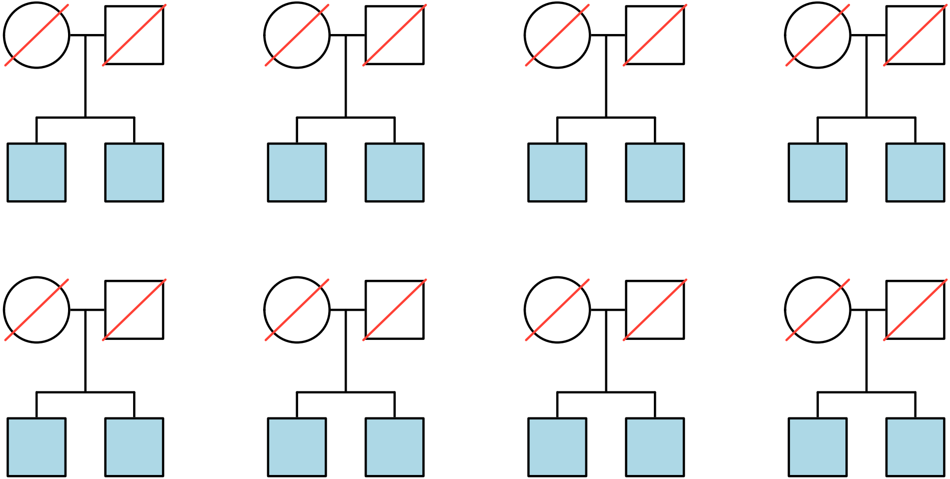


Lesson

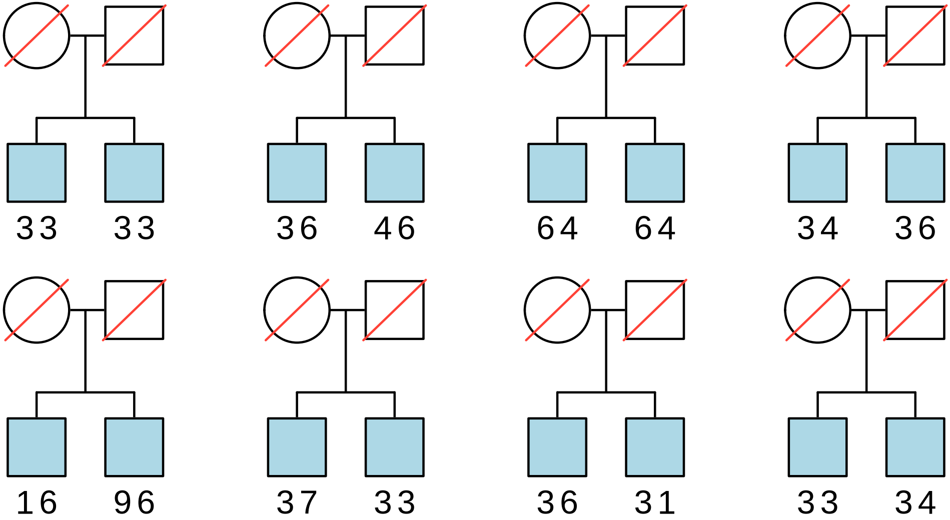
If it seems too good to be true,

it probably is.

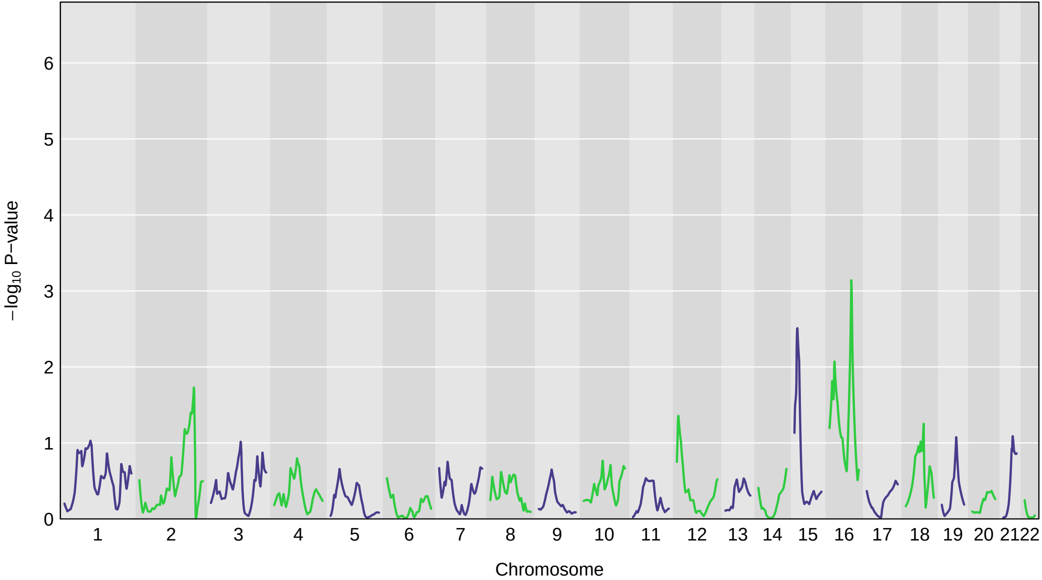
Prostate cancer pairs



Prostate cancer pairs



Prostate cancer genome scan – corrected



Estimating allele frequencies

Usually, you would use the founders in the pedigrees.
(assumed unrelated)

What if you only have sibships?

Estimating allele frequencies with sibpairs

Method 1: Use a random sibling from each

Method 2: Use everyone, ignoring relationships

Estimating allele frequencies with sibpairs

Method 1: Use a random sibling from each

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

Method 2: Use everyone, ignoring relationships

$$\text{var}(\hat{p}^{(2)}) = \textbf{?}$$

Estimating allele frequencies with sibpairs

Method 1: Use a random sibling from each

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

Method 2: Use everyone, ignoring relationships

$$\text{var}(\hat{p}^{(2)}) = (3/4) \left(\frac{p(1-p)}{2n}\right)$$

Estimating allele frequencies with sibpairs

Method 1: Use a random sibling from each

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

Method 2: Use everyone, ignoring relationships

$$\text{var}(\hat{p}^{(2)}) = (3/4) \left(\frac{p(1-p)}{2n}\right)$$

relative efficiency = 4/3 = 1.33

(best possible = 1.5)

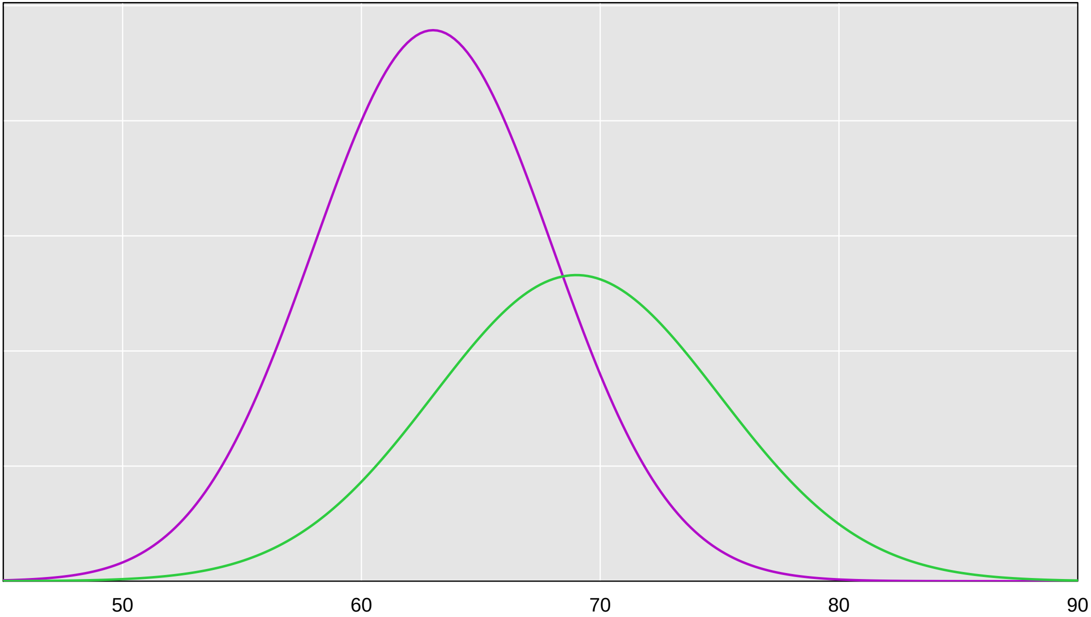
My favorite equations

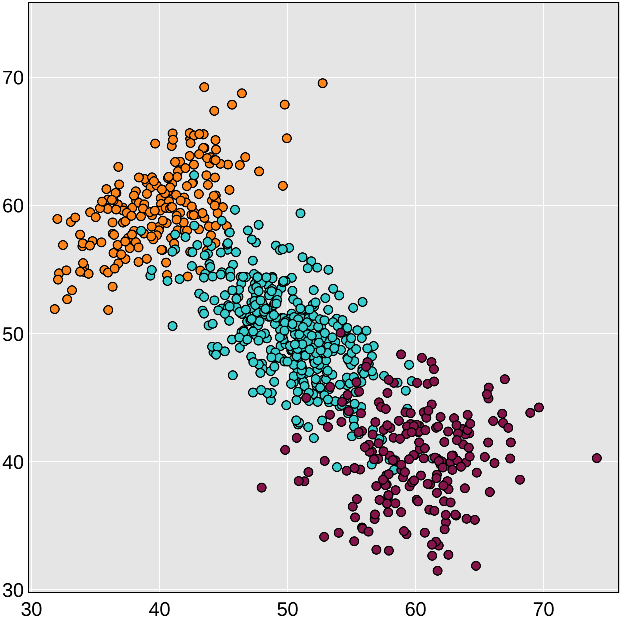
$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}[\mathrm{E}(X|Z)] \\
\mathrm{var}(X) &= \mathrm{E}[\mathrm{var}(X|Z)] + \mathrm{var}[\mathrm{E}(X|Z)] \\
\mathrm{cov}(X, Y) &= \mathrm{E}[\mathrm{cov}(X, Y|Z)] + \mathrm{cov}[\mathrm{E}(X|Z), \mathrm{E}(Y|Z)]
\end{aligned}$$

Everything is a mixture

Another fave

$$\begin{aligned} \mathrm{cov}(X, aY + bZ) &= a\,\mathrm{cov}(X, Y) + b\,\mathrm{cov}(X, Z) \\[1em] \text{thus}\quad \mathrm{var}(X + Y) &= \mathrm{cov}(X + Y, X + Y) \\ &= \mathrm{cov}(X + Y, X) + \mathrm{cov}(X + Y, Y) \\ &= \mathrm{cov}(X, X) + \mathrm{cov}(X, Y) + \mathrm{cov}(Y, X) + \mathrm{cov}(Y, Y) \\ &= \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X, Y) \end{aligned}$$


50
60
70
80
90


70
60
50
40
30
30
40
50
60
70

Back to that SE

Let X_i = number of 1 alleles in sib *i*.

We want $\text{var}(X_1 + X_2)$

And so really

$$
\begin{aligned}
\text{cov}(X_1, X_2) &= \text{E}[\text{cov}(X_1, X_2 | \text{IBD})] + \text{cov}[E(X_1|\text{IBD}), E(X_2|\text{IBD})] \\
&= \text{E}[\text{cov}(X_1, X_2 | \text{IBD})] \\
&= \sum_{k=0}^{2} \text{cov}(X_1, X_2 | \text{IBD} = k) \, \text{Pr}(\text{IBD} = k)
\end{aligned}
$$

Also

$$\begin{aligned}
\mathrm{cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \\
\mathrm{cov}(X, Y|Z) &= \mathrm{E}(XY|Z) - \mathrm{E}(X|Z)\mathrm{E}(Y|Z)
\end{aligned}$$

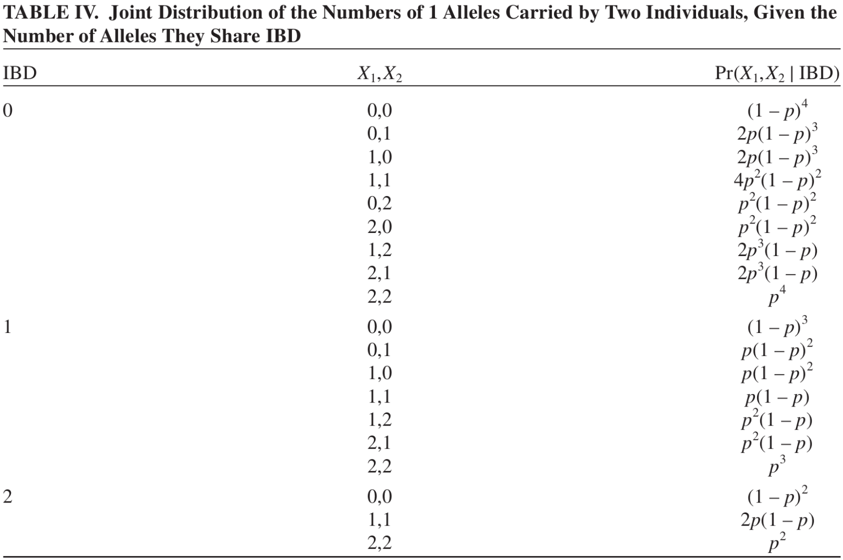


TABLE IV. Joint Distribution of the Numbers of 1 Alleles Carried by Two Individuals, Given the Number of Alleles They Share IBD

IBD	X_1,X_2	$\Pr(X_1,X_2 \mid \text{IBD})$
0	0,0	$(1-p)^4$
	0,1	$2p(1-p)^3$
	1,0	$2p(1-p)^3$
	1,1	$4p^2(1-p)^2$
	0,2	$p^2(1-p)^2$
	2,0	$p^2(1-p)^2$
	1,2	$2p^3(1-p)$
	2,1	$2p^3(1-p)$
	2,2	p^4
1	0,0	$(1-p)^3$
	0,1	$p(1-p)^2$
	1,0	$p(1-p)^2$
	1,1	$p(1-p)$
	1,2	$p^2(1-p)$
	2,1	$p^2(1-p)$
	2,2	p^3
2	0,0	$(1-p)^2$
	1,1	$2p(1-p)$
	2,2	p^2

Broman (2001) doi:10.1002/gepi.2

Lessons

- Omitting data is usually bad

- Crudely ignoring correlations can be good

  You might even be able to figure out the SE

Method 3

Account for relationships in the estimate

Missing data = IBD status for a sib pair at a marker

Use EM algorithm:

E step: estimate IBD status given allele frequencies

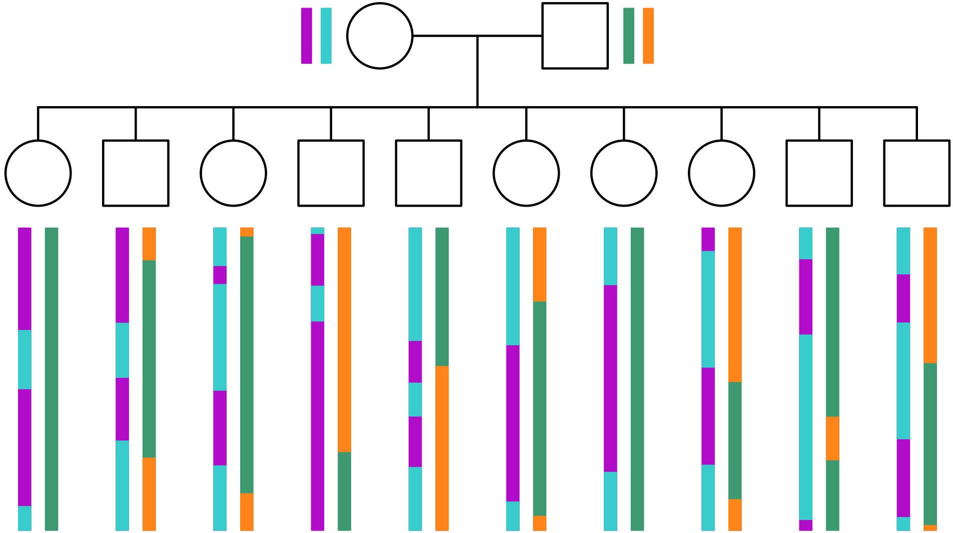
M step: estimate allele frequencies given IBD status

Method 4

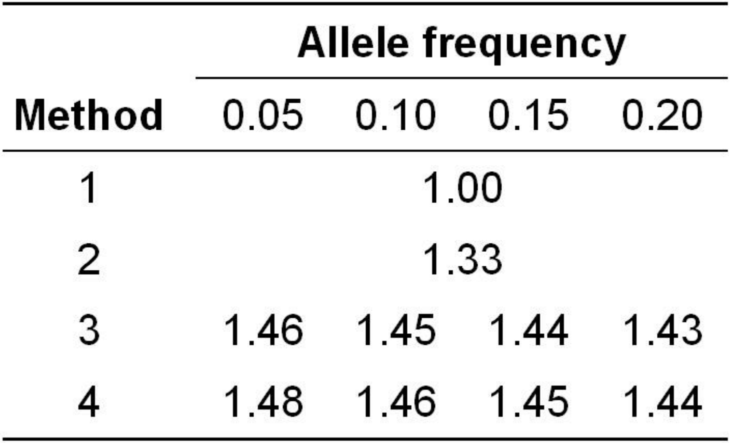
Make use of the multiple markers on a chromosome

- Markers along chromosome give improved info about IBD status
- Again, an EM algorithm:
  - Estimate IBD along chromosome given allele frequencies
  - Re-estimate allele frequencies using IBD information

Siblings' chromosomes



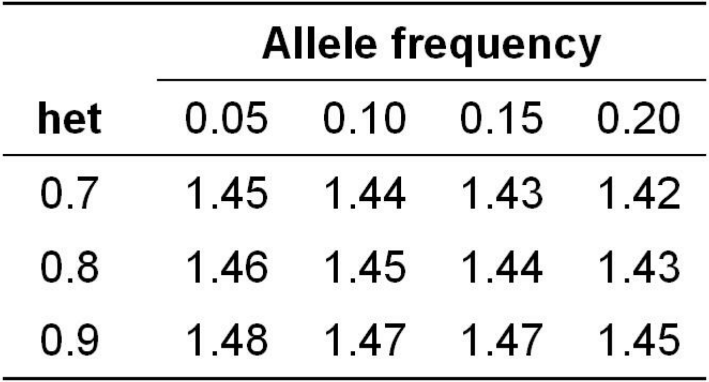
Average relative efficiency



	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

Broman (2001) doi:10.1002/gepi.2

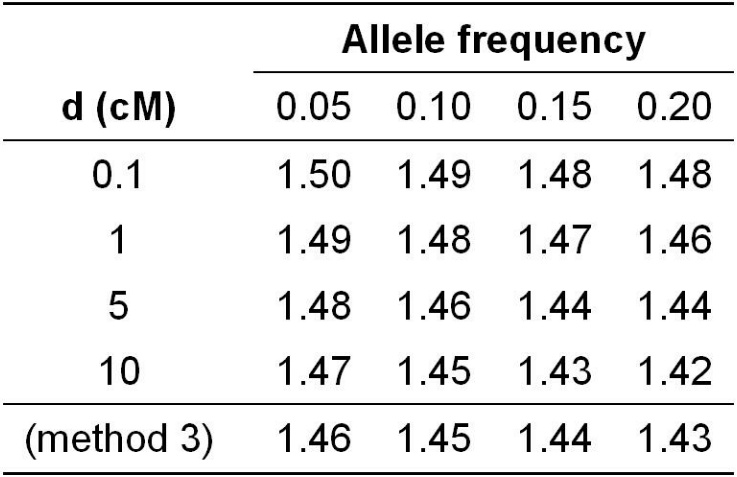
Method 3



	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

Broman (2001) doi:10.1002/gepi.2

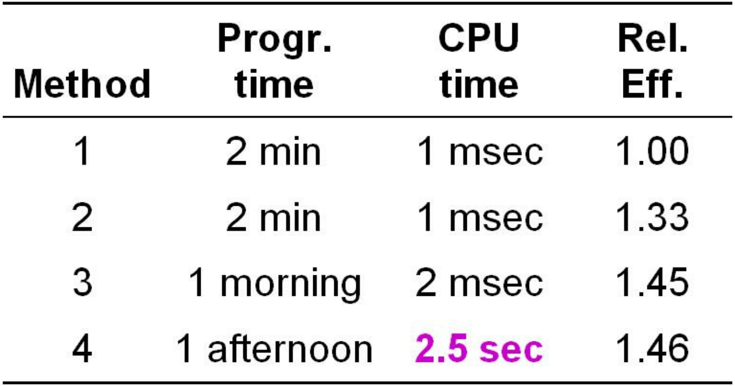
Method 4



	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

Broman (2001) doi:10.1002/gepi.2

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

One last thing

- Turns out, I made a mistake in Method 3

  Mary Sara McPeek (U Chicago) spotted it

- Fixed problem, re-ran simulations, and...

One last thing

- Turns out, I made a mistake in Method 3

  Mary Sara McPeek (U Chicago) spotted it

- Fixed problem, re-ran simulations, and...

  the correct MLE was worse than my mistaken estimate