Haplotyping

Biostatistics 666

Last Lecture

- Introduction to the E-M algorithm
- Approach for likelihood optimization
- Examples related to gene counting
  - Allele frequency estimation – recessive disorder
  - Allele frequency estimation – ABO blood group
  - Introduction to haplotype frequency estimation

Today: Haplotype Frequencies

- Evolution of haplotype estimation methods
  - Clark's *greedy* algorithm
  - Excoffier and Slatkin's E-M algorithm
  - Stephens et al's coalescent based algorithm
- Using estimated haplotypes in association studies

Useful Roles for Haplotypes

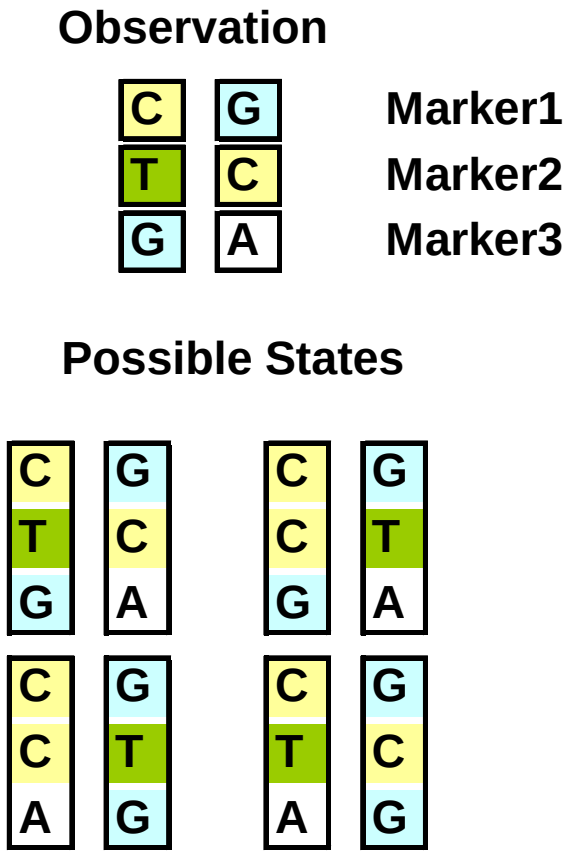
- Linkage disequilibrium studies
  - Summarize genetic variation
- Selecting markers to genotype
  - Identify haplotype tag SNPs
- Candidate gene association studies
  - Help interpret single marker associations
  - Capture the effect of ungenotyped alleles

The problem...

- Haplotypes are hard to measure directly
  - X-chromosome in males
  - Sperm typing
  - Hybrid cell lines
  - Other molecular techniques
- Often, statistical reconstruction required

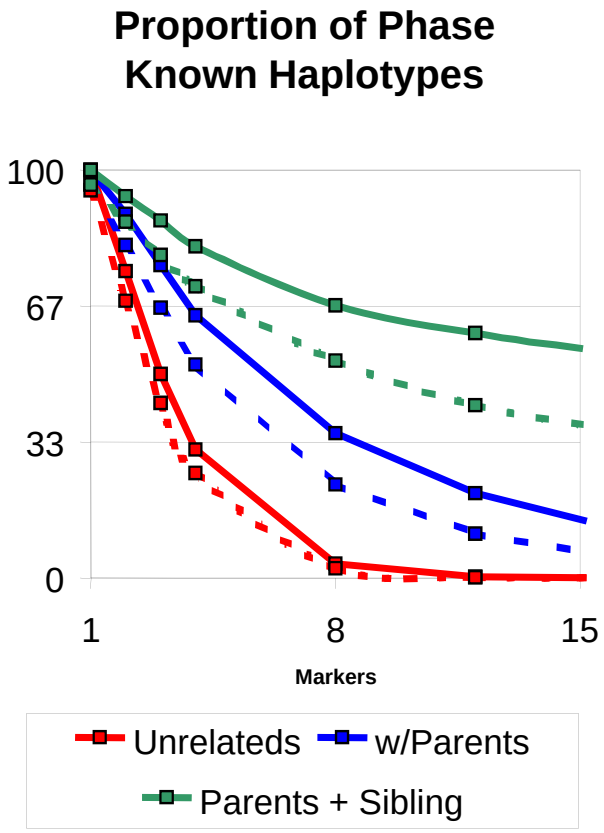
Typical Genotype Data

- Two alleles for each individual
  - Chromosome origin for each allele is unknown
- Multiple haplotype pairs can fit observed genotype



Use Information on Relatives

- Family information can help determine phase at many markers
- Still, many ambiguities might not be resolved
  - Problem more serious with larger numbers of markers
- Can you think of families where phase is ambiguous?



What if there are no relatives?

- Rely on linkage disequilibrium
- Assume that population consists of small number of distinct haplotypes
- Haplotypes tend to be similar

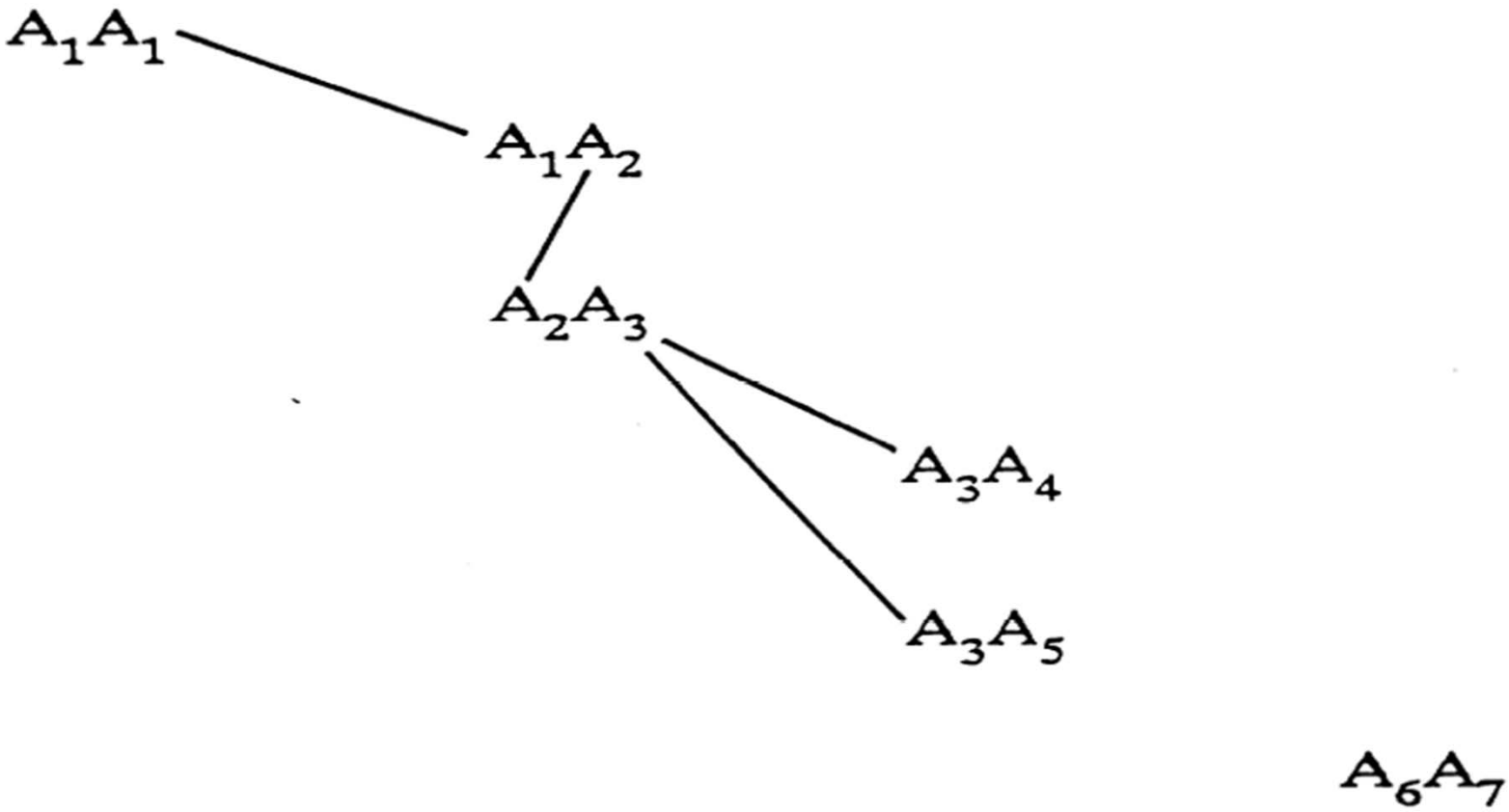
Clark's Haplotyping Algorithm

- Clark (1990) *Mol Biol Evol* **7**:111-122
- One of the first haplotyping algorithms
  - Computationally efficient
  - Very fast and widely used in 1990's
  - More accurate methods are now available
- Predictions from the coalescent can be used to model performance

Clark's Haplotyping Algorithm

- Find unambiguous individuals
  - What kinds of genotypes will these have?
  - Initialize a list of known haplotypes
- Resolve ambiguous individuals
  - If possible, use two haplotypes from list
  - Otherwise, use one known haplotype and augment list
- If unphased individuals remain
  - Assign phase randomly to one individual
  - Augment haplotype list and continue from previous step

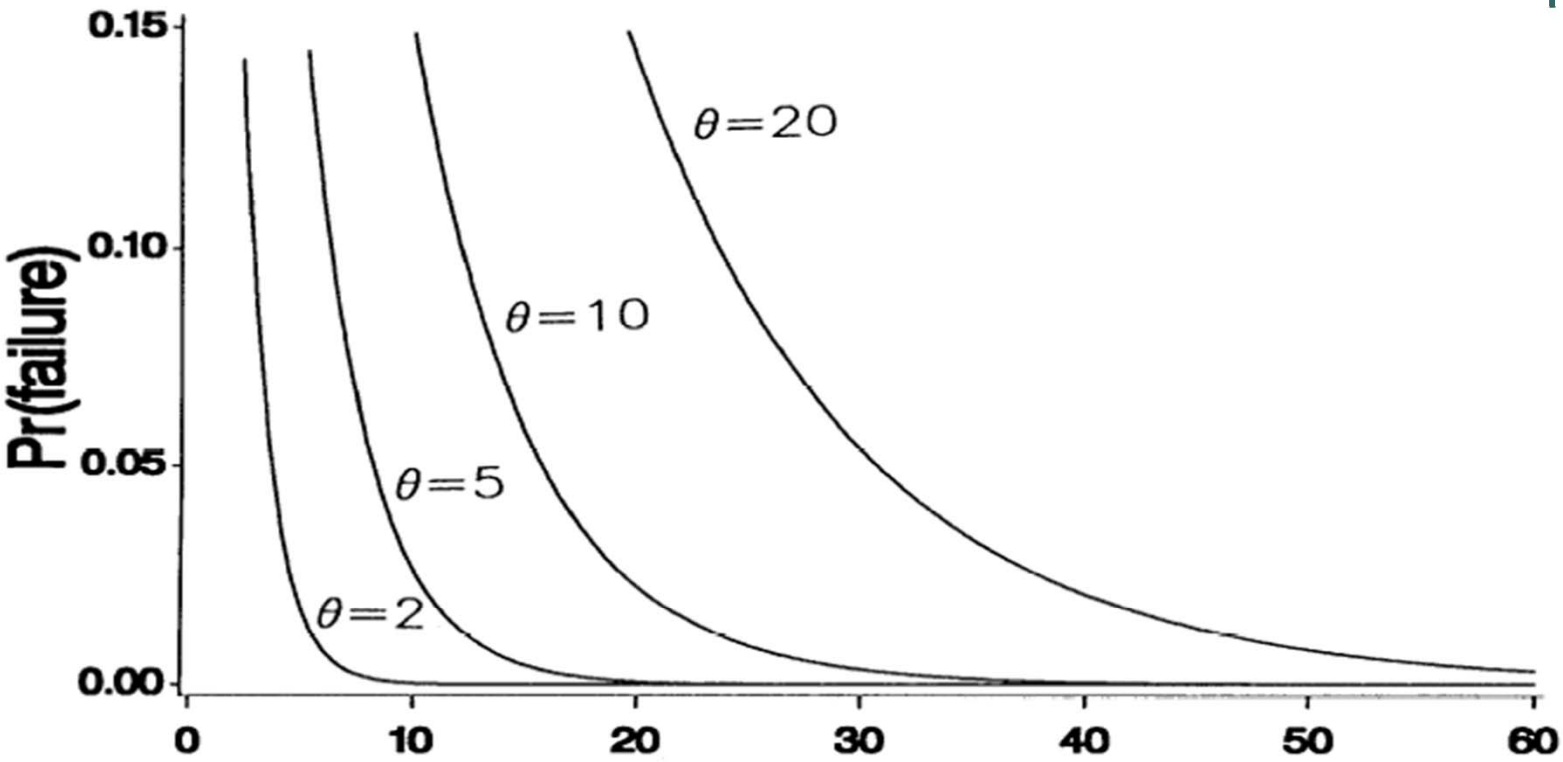
Chain of Inference (Clark, 1990)



Can The Algorithm Get Started?

- What kinds of genotypes do we need to get started?
- What kinds of haplotype pairs do we need to get started?
- What is the probability of these occurring?

Probability of Failing To Start



$$\Pr(\text{failure}) \approx \left[1 - \frac{1}{1+\theta} - \frac{\theta}{(1+\theta)^2}\right]^n$$

Distribution of Orphaned Alleles

Table 1

Fraction of Samples with Any Orphaned Alleles, and Average Frequency of Orphaned Alleles

	Sample Size (no. of individuals)			
θ	10	20	50	100
1.....	<0.001 (<0.001)	<0.001 (<0.001)	<0.001 (<0.001)	0.002 (<0.001)
2.....	0.008 (0.008)	<0.001 (<0.001)	<0.001 (<0.001)	0.003 (<0.001)
5.....	0.053 (0.205)	0.035 (0.008)	0.011 (0.001)	0.004 (<0.001)
10.....	0.204 (0.078)	0.102 (0.029)	0.037 (0.003)	0.011 (<0.001)
20.....	0.325 (0.118)	0.276 (0.092)	0.119 (0.018)	0.057 (0.001)

Distribution of Anomalous Matches

Fraction of Samples with Unresolved Anomalous Matches, and Frequency of Anomalous Haplotypes

θ	Sample Size (no. of individuals)			
	10	20	50	100
1	0.09 (0.07)	0.06 (0.04)	0.06 (0.06)	0.08 (0.09)
	0.07 (0.09)	0.06 (0.05)	0.07 (0.08)	0.08 (0.08)
2	0.15 (0.14)	0.14 (0.10)	0.12 (0.14)	0.15 (0.15)
	0.14 (0.11)	0.22 (0.14)	0.12 (0.13)	0.17 (0.17)
5	0.18 (0.22)	0.23 (0.00)	0.17 (0.23)	0.20 (0.34)
	0.10 (0.21)	0.23 (0.33)	0.20 (0.24)	0.26 (0.17)
10	0.00 (0.07)	0.17 (0.00)	0.13 (0.22)	0.20 (0.34)
	0.05 (0.13)	0.14 (0.31)	0.16 (0.28)	0.29 (0.30)
20	0.00 (0.20)	0.00 (0.00)	0.15 (0.24)	0.20 (0.34)
	0.02 (0.02)	0.02 (0.00)	0.15 (0.28)	0.35 (0.19)

Notes ...

- Clark's Algorithm is extremely fast
- More likely to start with large sample
- Orphaned alleles and anomalous matches may occur
  - Solution with the least orphaned alleles is usually the one with the fewest anomalous matches

The E-M Haplotyping Algorithm

- Excoffier and Slatkin (1995)
  - *Mol Biol Evol* **12:**921-927
  - Provide a clear outline of how the algorithm can be applied to genetic data
- Combination of two strategies
  - E-M statistical algorithm for missing data
  - Counting algorithm for allele frequencies

E-M Algorithm For Haplotyping

1. “Guesstimate” haplotype frequencies
2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
3. Estimate frequency of each haplotype by counting
4. Repeat steps 2 and 3 until frequencies are stable

Expected Haplotype Counts

$$h_j = \text{haplotype } j$$

$$G(h_i, h_j) = \text{Unphased genotype corresponding to } h_i, h_j$$

$$n_G = \text{Number of genotypes of type G}$$

$$H \sim G = \text{Haplotype pairs compatible with G}$$

$$\mathrm{E}(n_{h_i}) = 2n_{G(h_i,h_i)} + \sum_{h_j \neq h_i} n_{G(h_i,h_j)} \frac{2\hat{p}_{h_i}\hat{p}_{h_j}}{\sum_{H \sim G(h_i,h_j)} P(H \mid \hat{p})}$$

Notation used in Excoffier and Slatkin paper

m	no. of observed phenotypes
$j = 1...m$	index for individual phenotypes
s_j	no. of heterozygous markers for phenotype j
c_j	haplotype pairs compatible for phenotype j
$P_j = \sum_{i=1}^{c_j} P(\text{haplotype pair } j) = \sum_{i=1}^{c_j} P(h_{ik} h_{il})$	probability of phenotype j
$L = \prod_{j=1}^{m} \left(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}$	Likelihood

For clarity, try replacing 'phenotype' with 'observed genotype combination'

Notation used in Excoffier and Slatkin paper

$p_1^{(g)}, p_2^{(g)} \ldots p_h^{(g)}$ frequency of haplotype h at round g

$$P_j^{(g)}(h_k h_l) = \begin{cases} p_k^{(g)} p_l^{(g)} \\ \left(p_k^{(g)}\right)^2 \end{cases}$$

probability of genotype $h_k h_l$ for heterozygotes and homozygotes

$$P_j^{(g)} = \sum_{i=1}^{c_j} P_j^{(g)}(h_{ik} h_{il})$$

probability of phenotype j

$$P^{(g)}(h_k h_l) = \frac{n_j}{n} \frac{P_j^{(g)}(h_k h_l)}{P_j^{(g)}}$$

fitted proportion for genotype $h_k h_l$

$$p_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j^{(g)}(h_k h_l)$$

allele frequencies for next iteration

For clarity, try replacing 'phenotype' with 'observed genotype combination'

Computational Cost (for SNPs)

- Consider sets of *m* unphased genotypes
  - Markers 1..*m*

For example, if m = 10

- If markers are bi-allelic
  - 2^m possible haplotypes **= 1024**
  - $2^{m-1}\ (2^m + 1)$ possible haplotype pairs **= 524,800**
  - 3^m distinct observed genotypes **= 59,049**
  - 2^{n-1} reconstructions for *n* heterozygous loci **= 512**

E-M Algorithm for Haplotyping

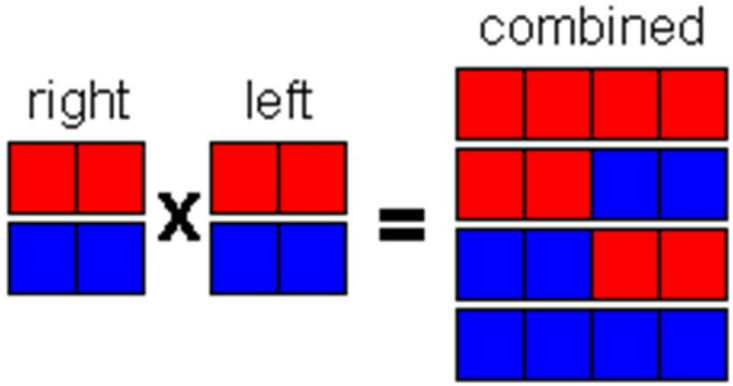
- Cost grows rapidly with number of markers
- Typically appropriate for < 25 SNPs
  - Fewer microsatellites
- More accurate than Clark's method
- Fully or partially phased individuals contribute most of the information

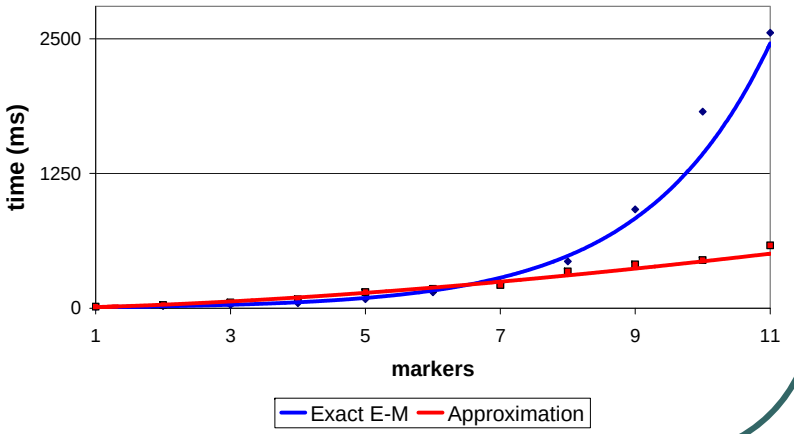
Enhancements to E-M

- List only haplotypes present in sample
- Gradually expand subset of markers under consideration, eliminating haplotypes with zero or low estimated frequency from consideration at each stage
  - SNPHAP [Clayton (2001)]
  - HAPLOTYPER [Qin et al. (2002)]

Divide-And-Conquer Approximation

- Number of potential haplotypes increases exponentially
  - Number of observed haplotypes does not
- Approximation
  - Successively divide marker set
  - Run E-M on each segment
  - Prune haplotype list as segments are ligated
- Computation order: ~ *m log m*
  - Exact E-M is order ~ 2^m





Further Refinements …

- More modern methods try to further improve haplotype estimation by favoring sets of similar haplotypes
- Stephens et al. (2001)
  - *Am J Hum Genet* **68**:978-89
- Genealogical approach…

What the Genealogy Implies…

- Haplotypes are similar to each other…

Known Haplotypes

22544
22544
22544

33334
33334

23233

14234

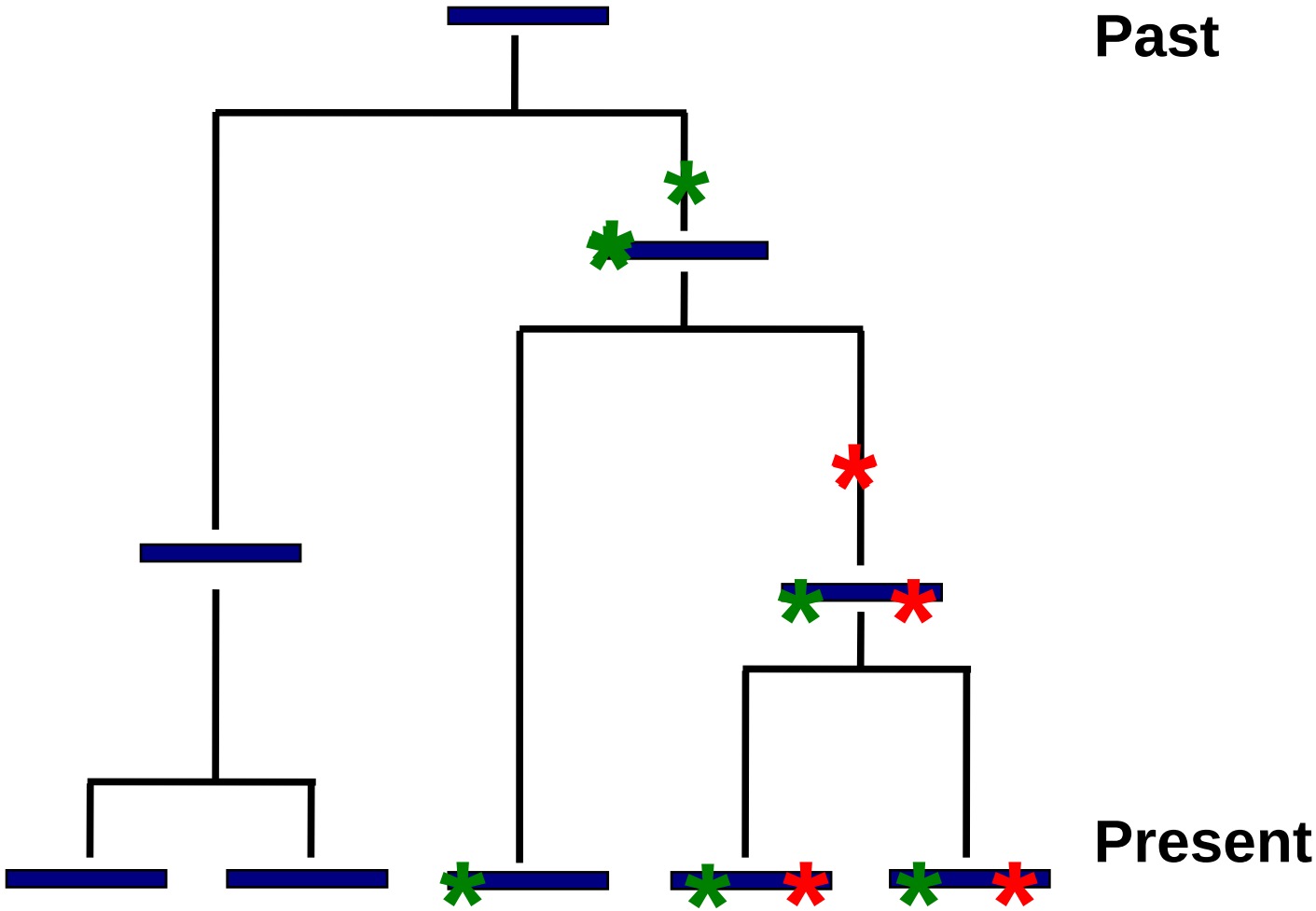
Individual 1

Genotype:
32344
23534

Individual 2:

Genotype:
32444
23434

Chromosome Genealogies



Method based on Gibbs sampler

- MCMC method
  - Stochastic, random procedure
  - Improves solution gradually

- Given initial set of haplotypes
- Sample haplotypes for one individual at a time, assuming other haplotypes are true
- Repeat a few million times…

Update Procedure I

- Pick individual U to update at random
- Calculate haplotype frequencies F in all other individuals
  - Since everyone is “phased”, this is done by counting
- Sample new haplotypes for U from conditional distribution of U’s haplotypes given F

Update Procedure I

- This procedure would produce an estimate of haplotype frequencies equivalent to those obtained by E-M …
- Stephens et al (2001) suggested an alternative estimate of F…

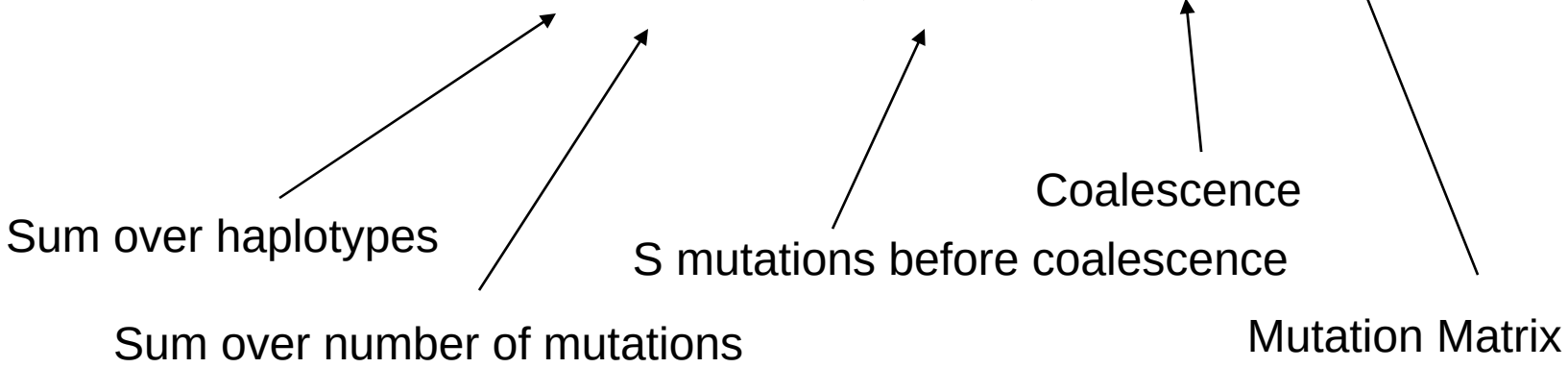
Update Procedure II

- Estimate F from the other individuals
- Construct F* to include haplotypes in F and also other similar haplotypes (possibly differing at a few sites)
- Update U's haplotypes conditional on F*

Stephens' Formula ...

- Pr*(h|H)* is the probability of observing haplotype *h* given previous set *H*

$$\Pr(h \mid H) = \sum_{\alpha} \sum_{S} \frac{n_\alpha}{n} \left(\frac{\theta}{n+\theta} \right)^S \frac{n}{n+\theta} \left(P^S \right)_{\alpha h}$$



Further Refinements

- This naïve strategy becomes impractical for very long haplotypes
  - List of haplotypes for each individual could become too long
- Instead, we can proceed by selecting a short segment of the haplotype to update at random

Hypothesis Testing

- Often, haplotype frequencies are not final outcome.
- For example, we may wish to compare two groups of individuals…
  - Are haplotypes similar in two populations?
  - Are haplotypes similar in patients and healthy controls?

Simplistic approach...

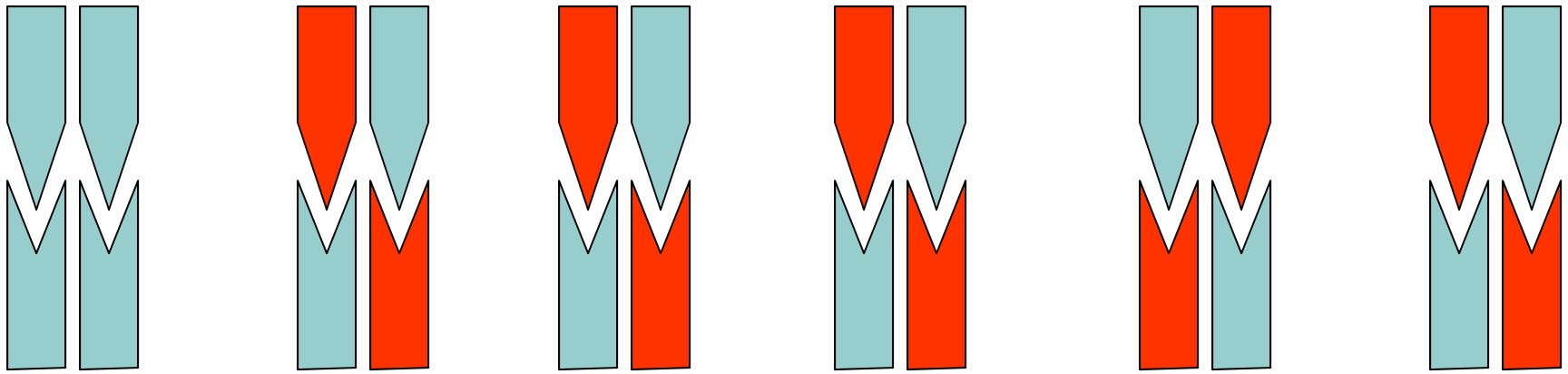
- Calculate haplotype frequencies in each group
- Find most likely haplotype for each individual
- Compare haplotype reconstructions in the two groups

Simplistic approach...

- Calculate haplotype frequencies in each group
- Find most likely haplotype for each individual
- Compare haplotype reconstructions in the two groups

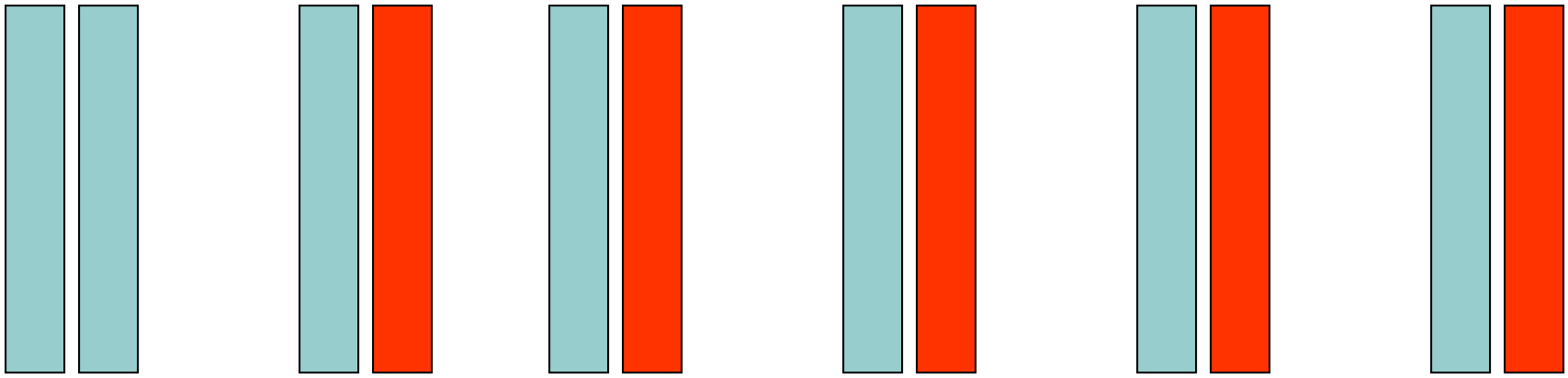
NOT RECOMMENDED!!!

Observed Case Genotypes



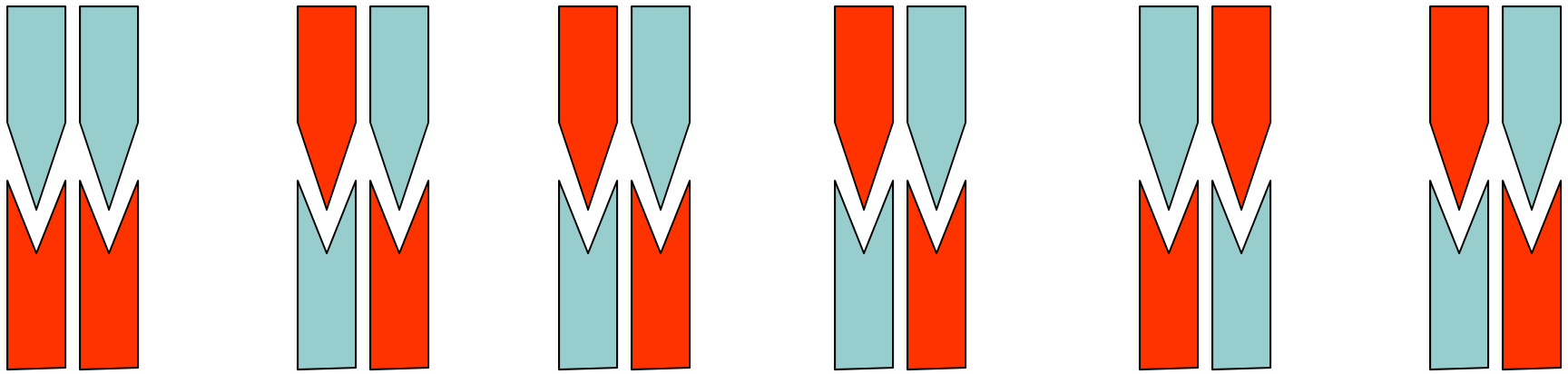
The phase reconstruction in the five ambiguous individuals will be driven by the haplotypes observed in individual 1 …

Inferred Case Haplotypes



This kind of phenomenon will occur with nearly all population based haplotyping methods!

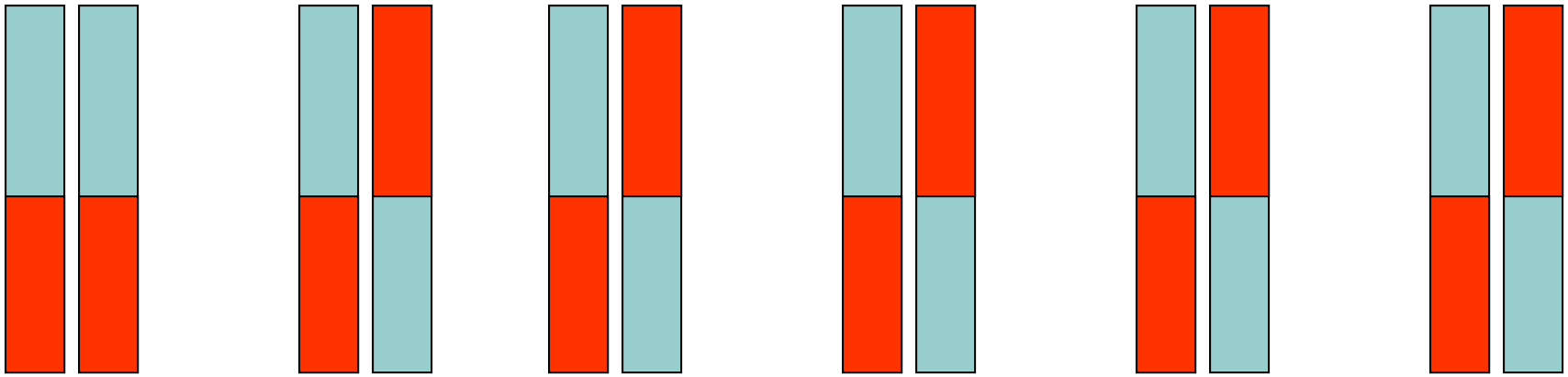
Observed Control Genotypes



1 2 3 4 5 6

Note these are identical, except for the single homozygous individual …

Inferred Control Haplotypes



Ooops… The difference in a single genotype in the original data has been greatly amplified by estimating haplotypes…

Hypothesis Testing II

- Never impute case and control haplotypes separately
- Instead, consider both groups together …
  - Schaid et al (2002) *Am J Hum Genet* **70**:425-34
  - Zaytkin et al (2002) *Hum Hered.* **53**:79-91
- Another alternative is to use maximum likelihood

Hypothesis Testing III

- Estimated haplotype frequencies, imply a likelihood for the observed genotypes

$$L = \prod_{i} \sum_{H \sim G_i} P(H)$$

Hypothesis Testing III

- Estimated haplotype frequencies, imply a likelihood for the observed genotypes

$$L = \prod_{i} \sum_{H \sim G_i} P(H)$$

individuals

possible haplotype pairs, conditional on genotype

haplotype pair frequency

Hypothesis Testing III

- Calculate 3 likelihoods:
  - Maximum likelihood for combined sample, L_A
  - Maximum likelihood for control sample, L_B
  - Maximum likelihood for case sample, L_C

$$2\ln\left(\frac{L_B L_C}{L_A}\right) \sim \chi^2_{df}$$

df corresponds to number of non-zero haplotype frequencies in large samples

Significance in Small Samples

- In realistic sample sizes, it is hard to estimate the number of *df* accurately
- Instead, permute case and control labels randomly

Final thoughts…

- Compare alternative reconstructions
  - Change input order
  - Change random seeds
  - Change starting values
- When analyzing case-control studies
  - Randomize case-control labels

Summary

- Describe principles underlying haplotype estimation in unrelated individuals
  - Heuristic algorithms
  - The E-M algorithm
  - Genealogical approach