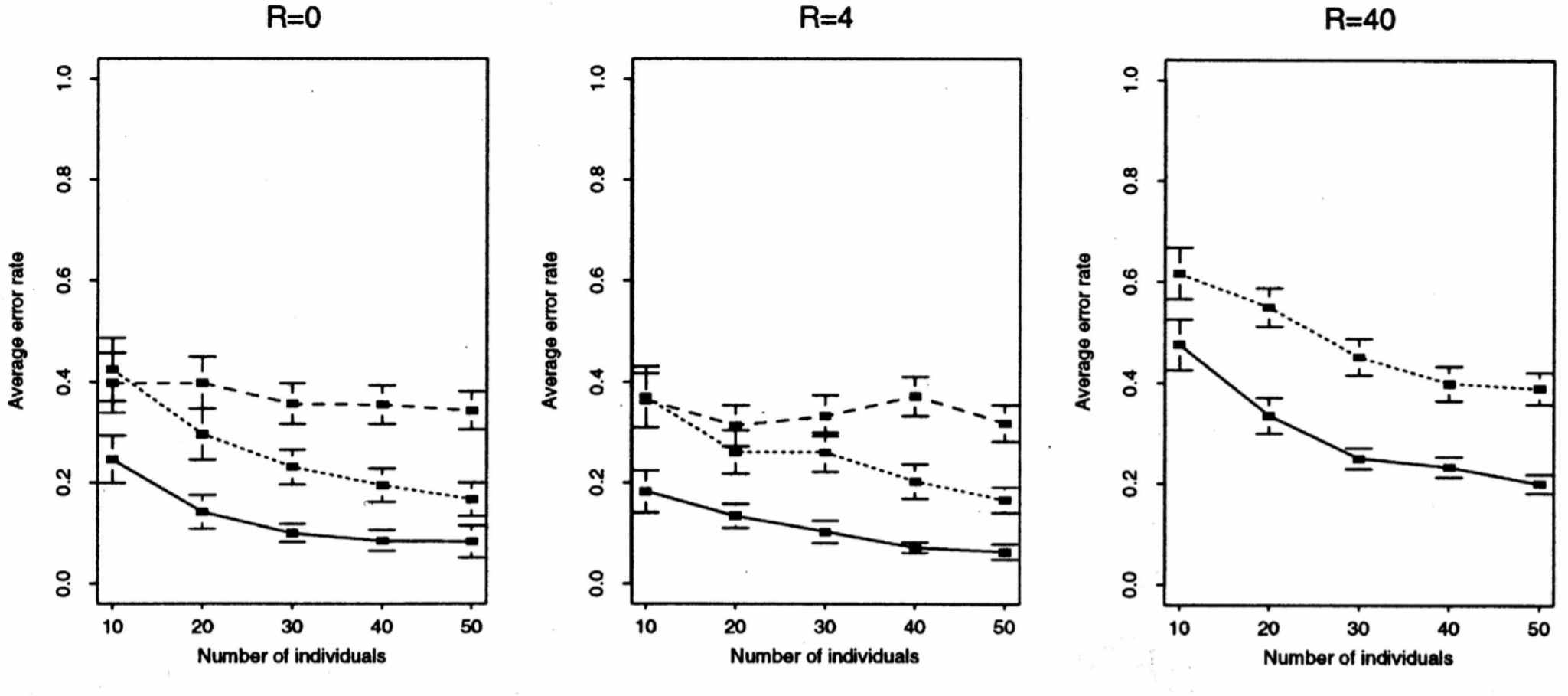
Advanced Haplotyping: Association Tests & Markov Models

Biostatistics 666

Previously …

- Evolution of Haplotype Estimation Methods
- Clark (1990) uses list of known haplotypes to resolve ambiguous individuals
- Excoffier and Slatkin (1995) propose an E-M algorithm that uses frequency information and allows for uncertainty in haplotype assignments
- Stephens et al. (2001) allow new haplotypes to be similar, but not identical, to previously seen haplotypes and use MCMC for gradually refining solution

Comparison of Three Haplotyping Algorithms



Clark's Method (- - - -), E-M algorithm (·····), Stephens et al (———)
Error Rate: Proportion of Ambiguous Individuals Phased Incorrectly

Stephens et al (2001)

Limitations

- All these methods work on relatively small regions of DNA

- In longer regions, all haplotypes are effectively unique and quite different from their most similar neighbor

Hypothesis Testing

- Often, haplotype frequencies are not final outcome.

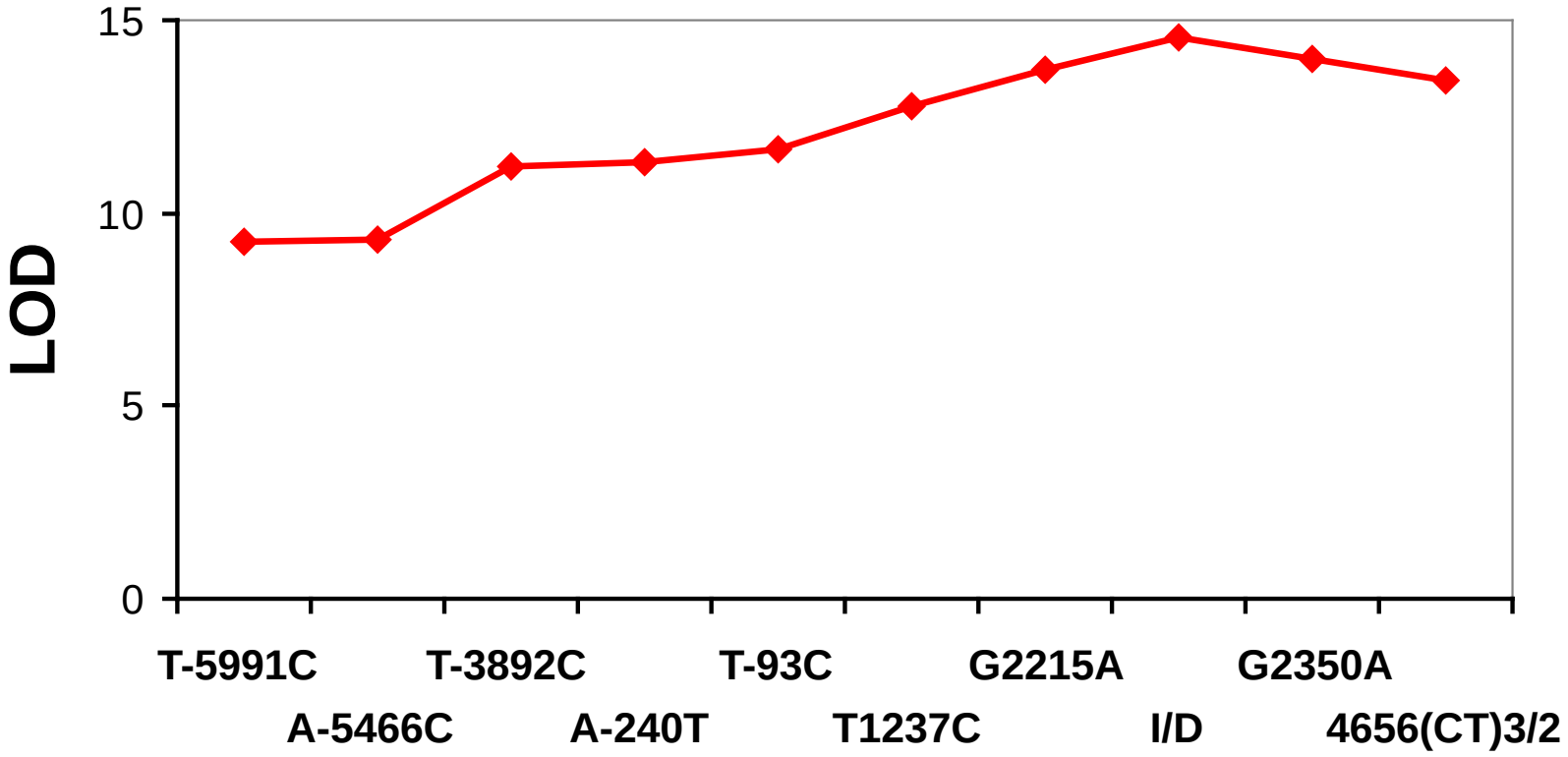
- For example, we may wish to compare two groups of individuals…
  - Are haplotypes similar in two populations?
  - Are haplotypes similar in patients and healthy controls?

Haplotype Association Tests

Why Do Haplotype Analysis?
ACE gene example

- Keavney et al (1998), Hum Mol Genet 7:1745-1751
- Studied a set of British individuals
- Measured angiotensin enzyme levels in each one
- Also measured 10 di-allelic polymorphisms
  - Markers span 26kb in angiotensin converting enzyme gene
  - Markers are common and in strong linkage disequilibrium

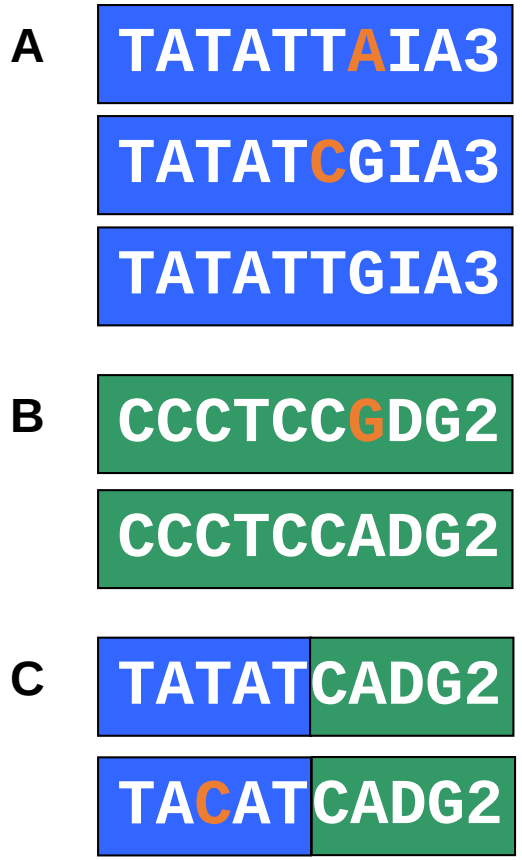
Single Marker Association Tests
ACE gene example



All markers examined show very strong evidence for association.

Haplotype Analysis
ACE gene example

- 3 ACE haplotype clades
  - Include all common haplotypes
  - >90% of all haplotypes
- Clade “B” = Clade “C”
  - Equal phenotypic effect
- Interpretation:
  - Functional variant on right
- Keavney et al (1998)



Introduction:
A Single Marker Association Test

- Simplest strategy to detect genetic association

- Compare frequencies of particular alleles, or genotypes, in set of cases and controls

- Typically, use contingency table tests…
  - Chi-squared Goodness-of-Fit Test
  - Cochran-Armitage Trend Test
  - Likelihood Ratio Test
  - Fisher's Exact Test

- … or regression based tests.
  - More flexible modeling of covariates

Construct Contingency Table

- Rows
  - One row for cases, another for controls
- Columns
  - One for each genotype
  - One for each allele
- Individual cells
  - Count of observations, with double counting for allele tests

Simple Association Study

	Genotype		
	1/1	1/2	2/2
Affecteds	$n_{a,11}$	$n_{a,12}$	$n_{a,22}$
Unaffecteds	$n_{u,11}$	$n_{u,12}$	$n_{u,22}$

Organize genotype counts in a simple table...

Notation

- Let index *i* iterate over rows
  - E.g. *i = 1* for affecteds, *i = 2* for unaffecteds
- Let index *j* iterate over columns
  - E.g. *j = 1* for genotype 1/1, j = 2 for genotype 2/2, etc.
- Let O_{ij} denote the observed counts in each cell
  - Let $O_{\bullet\bullet}$ denote the grand total
  - Let $O_{i\bullet}$ and $O_{\bullet j}$ denote the row and column totals
- Let E_{ij} denote the expected counts in each cell
  - $E_{ij} = O_{i\bullet} O_{\bullet j} / O_{\bullet\bullet}$

Goodness of Fit Tests

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

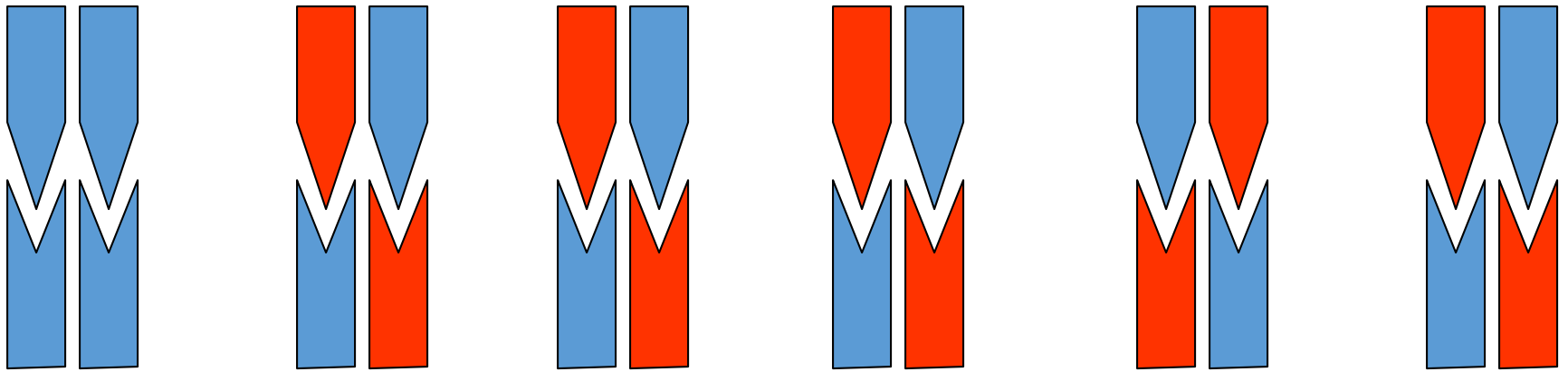
- If counts are large, compare statistic to chi-squared distribution
  - p = 0.05 threshold is 5.99 for 2 df (e.g. genotype test)
  - p = 0.05 threshold is 3.84 for 1 df (e.g. allele test)
- If counts are small, exact or permutation tests are better

Haplotype Association Test
A Simple Straw Man Approach

- Calculate haplotype frequencies in each group
- Find most likely haplotype for each individual
- Fill in contingency table to compare haplotypes in the two groups

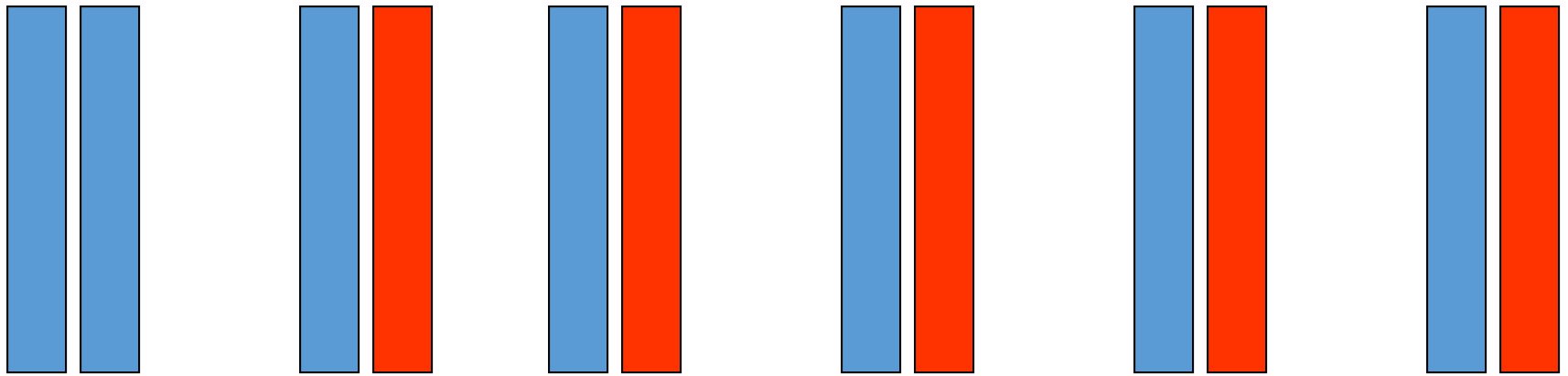
NOT RECOMMENDED!!!

Observed Case Genotypes



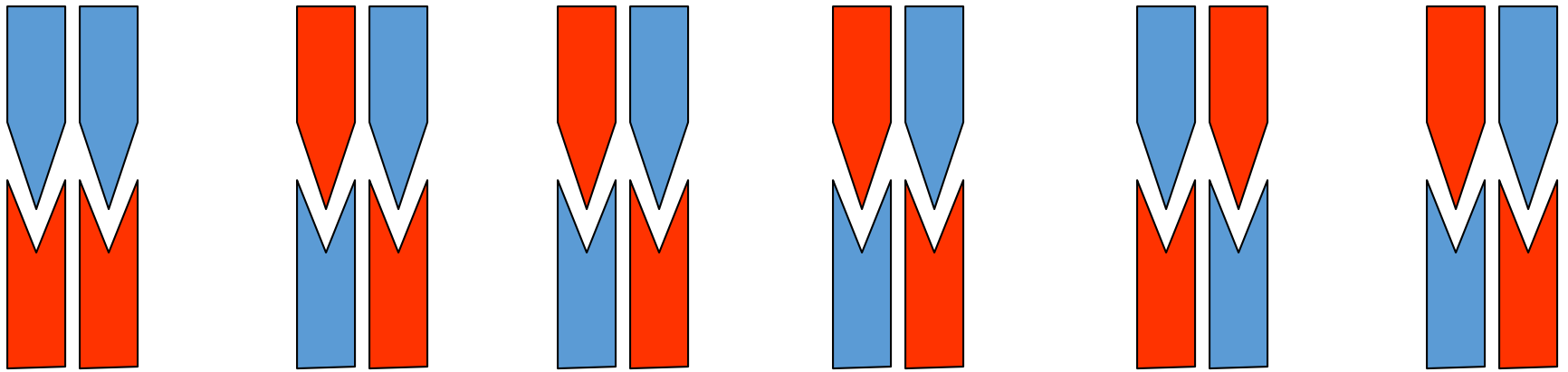
The phase reconstruction in the five ambiguous individuals will be driven by the haplotypes observed in individual 1 …

Inferred Case Haplotypes



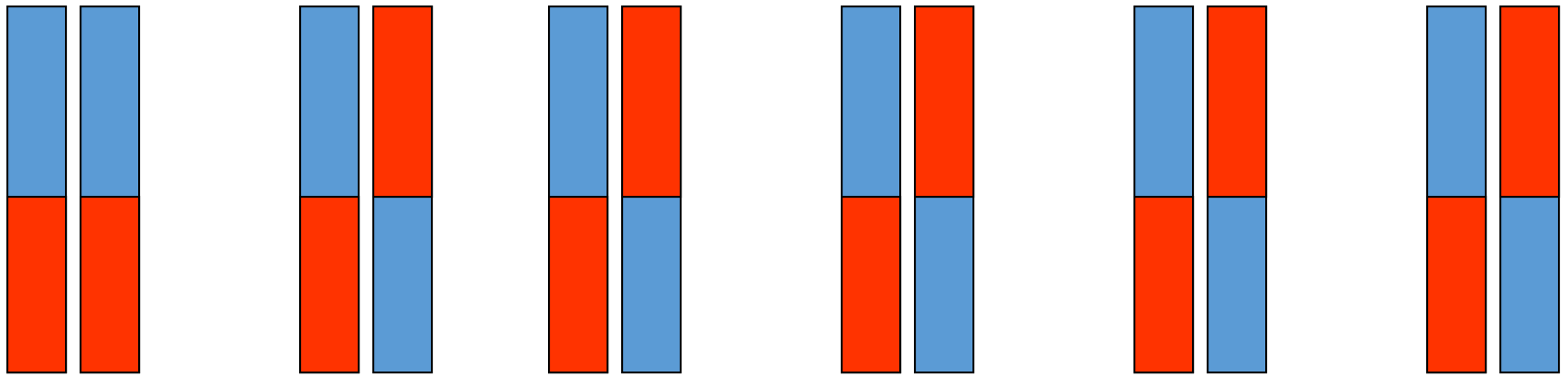
This kind of phenomenon will occur with nearly all population based haplotyping methods!

Observed Control Genotypes



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…

Common Sense Rules for Haplotype Association Tests

- Never impute haplotypes in two samples separately
- Use maximum likelihood
  - Does not require imputing individual haplotypes
  - Likelihood statistic can allow for uncertainty
- If haplotypes imputed, treat cases and controls jointly
  - Schaid et al (2002) *Am J Hum Genet* **70**:425-34
  - Zaytkin et al (2002) *Hum Hered.* **53**:79-91

Likelihood Function for Haplotype Data

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

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

Likelihood Function for Haplotype Data

- 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

Likelihood Ratio Test For Difference in Haplotype Frequencies

- 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

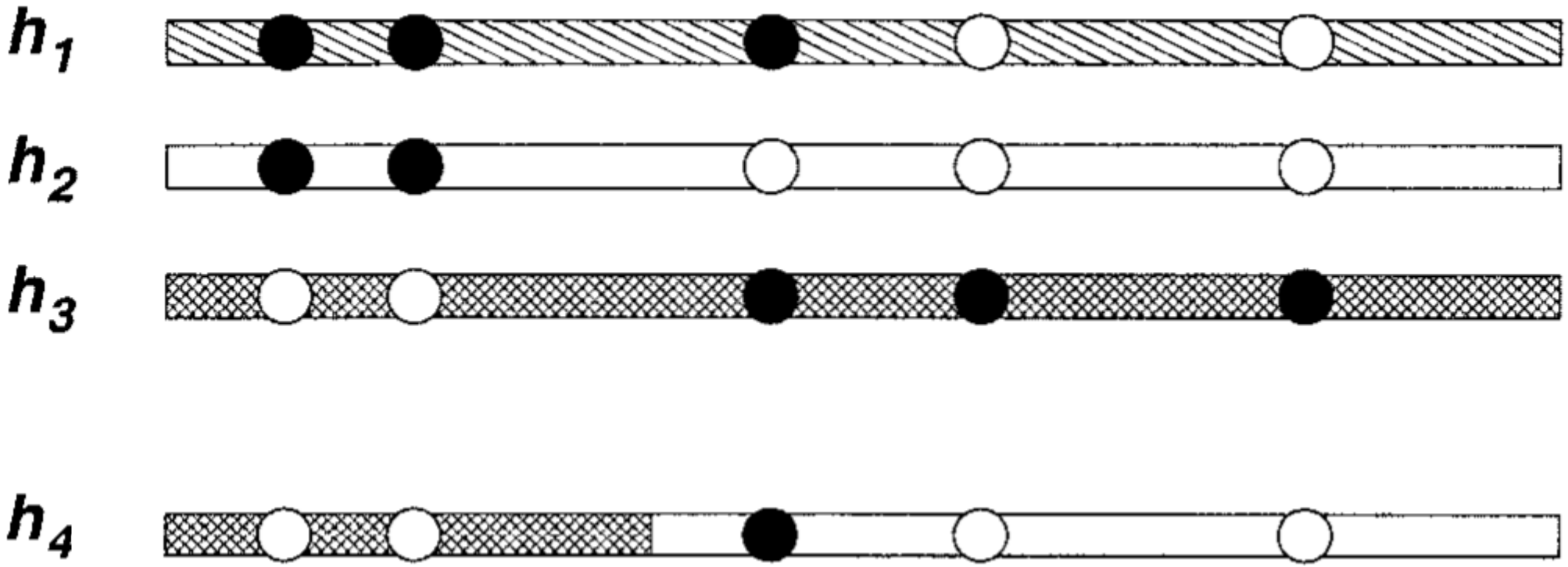
Significance in Small Samples

- In realistic sample sizes, it is hard to estimate the number of *df* accurately

- Instead, use a permutation approach to calculate empirical significance levels

Improved Haplotype Estimation

Haplotypes as Mosaics



Li and Stephens (2003)

Implementation

- Markov model is used to model each haplotype, conditional on all others
- At each position, we assume that the haplotype being modeled copies a template haplotype
- Each individual has two haplotypes, and therefore copies two template haplotypes
- We use MCMC, starting with a random solution and gradually updating one individual at a time as a mosaic of the others

1. Select a Sample to Update

Sample to be Updated

C G A A A C C C C C C G A C C T C A T G G
C G A G G T T T T T T T C T T T C A T G G

Current Haplotype Set

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

2. Find Matching Mosaic Pieces

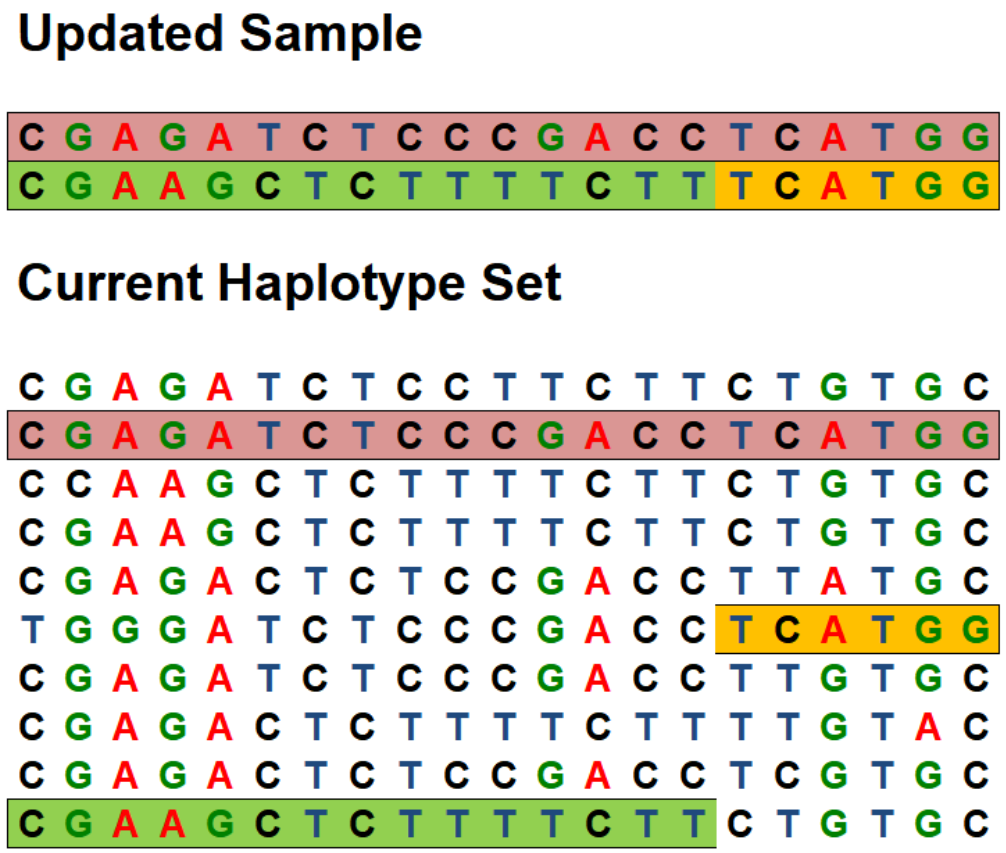
Sample to be Updated

C G A A A C C C C C C G A C C T C A T G G
C G A G G T T T T T T T C T T T C A T G G

Current Haplotype Set

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

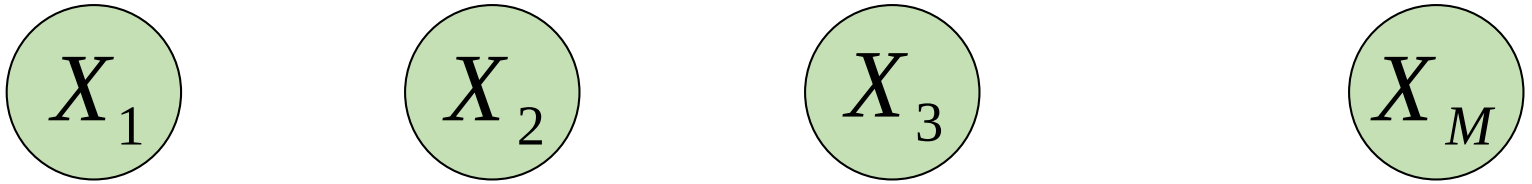
3. Update Haplotypes to Match Mosaic



How to Evaluate All Possible Configurations?

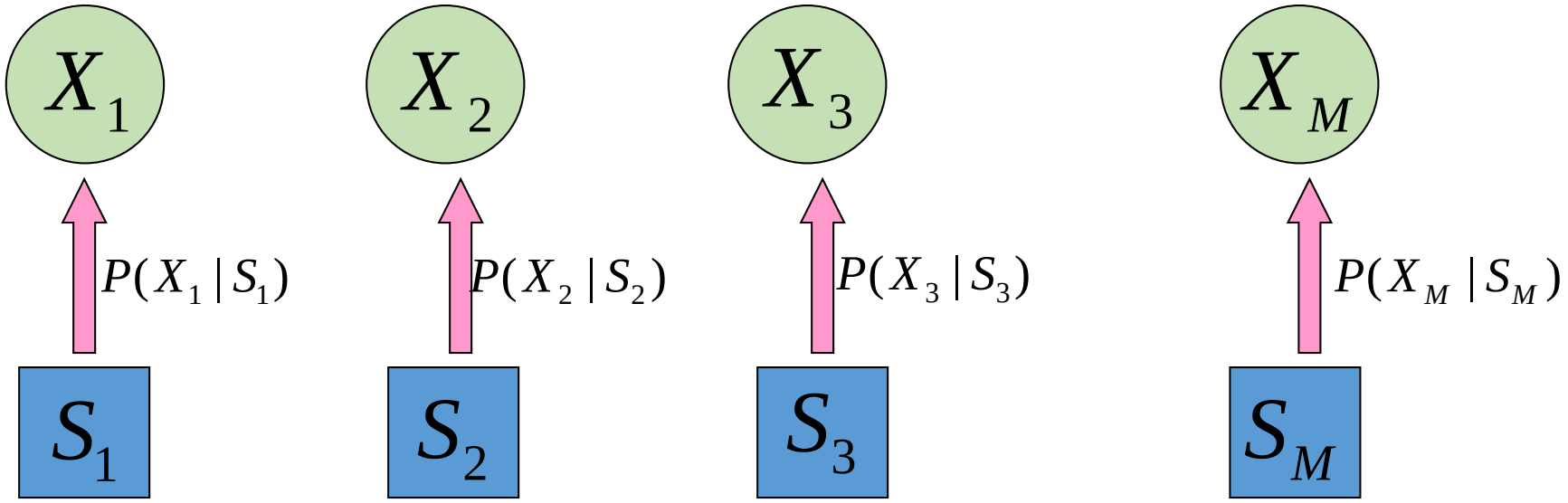
- We could imagine listing all possible mosaic states
- A mosaic state would specific template haplotype at each position
- We could compare mosaic states based on …
  - Number of template switches, favoring fewer switches
  - Number of mismatches between template and actual genotypes, favoring fewer mismatches
- One challenge is that the number of mosaic states is extremely large
  - With H potential templates and M genotyped sites, ~H^{2M} potential configurations

Hidden Markov Model Ingredients



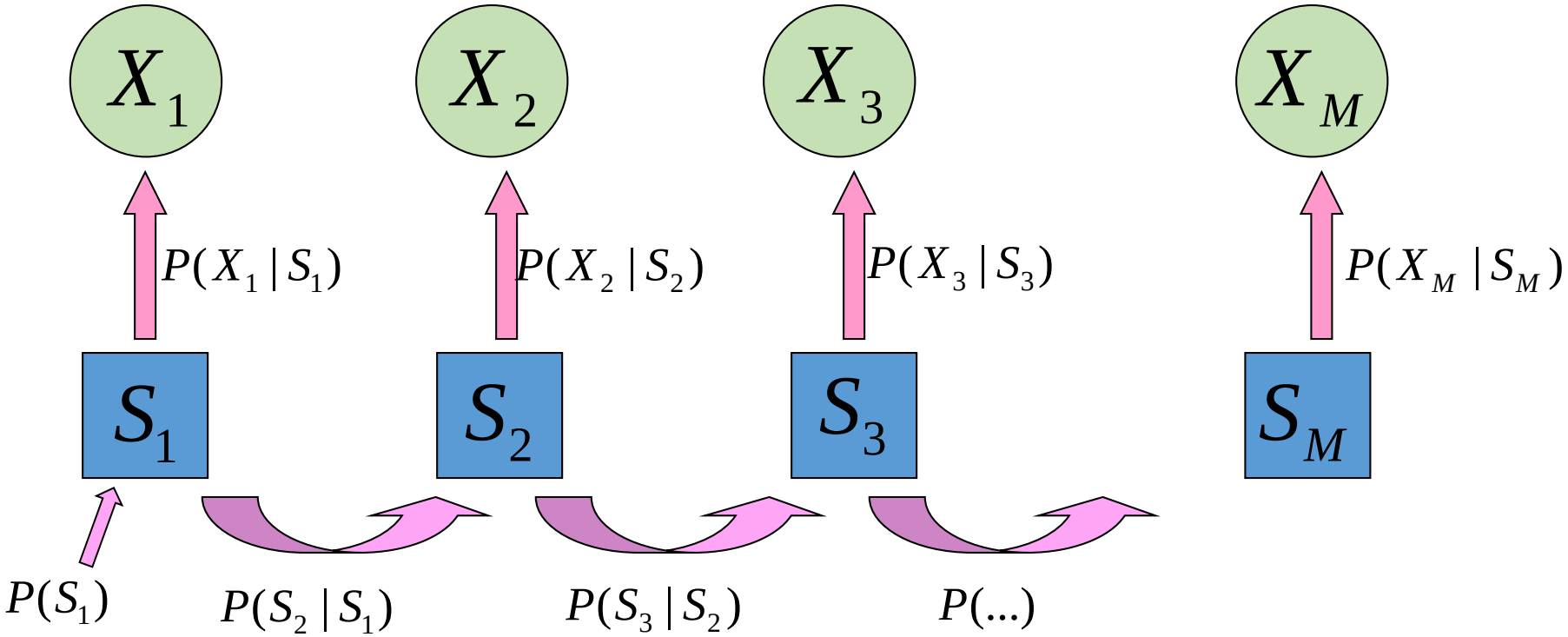
One ingredient will be the observed genotypes at each marker …

Hidden Markov Model Ingredients



Another ingredient will be the choice of template at each position …

Hidden Markov Model Ingredients



The final ingredient connects mosaic states as we move along the chromosome

Likelihood for Specific Mosaic State

$$L(S_1, S_2, \ldots, S_M) = P(S_1)\prod_{i=2}^{M} P(S_i \mid S_{i-1})\prod_{i=1}^{M} P(X_i \mid S_i)$$

- Likelihood accounts for template switches and mismatches
- To update haplotypes, we choose among most likely configurations
- Each mosaic configuration implies a specific set of haplotypes

Summing Over All Potential Mosaics

$$L = \sum_{S_1} \sum_{S_2} \cdots \sum_{S_M} P(S_1) \prod_{i=2}^{M} P(S_i \mid S_{i-1}) \prod_{i=1}^{M} P(X_i \mid S_i)$$

- General formulation, allows for any number of markers.
- Easy to write and understand (hopefully!) but challenging to compute
- Challenge: How to compute this efficiently?

A Markov Model

- Re-organize the computation, 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

Left-Chain Probabilities

$$L_m(S_m) = P(X_1,\ldots,X_{m-1} \mid S_m)$$
$$= \sum_{I_{m-1}} L_{m-1}(S_{m-1}) P(X_{m-1} \mid S_{m-1}) P(S_{m-1} \mid S_m)$$
$$L_1(S_1) = 1$$

- Proceed one marker at a time.

- Computation cost increases linearly with number of markers.

Right-Chain Probabilities

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

- Proceed one marker at a time.

- Computation cost increases linearly with number of markers.

The Likelihood of Marker Data

$$L = \sum_{I_j} P(S_j)P(X_j \mid S_j)P(X_1 \ldots X_{j-1} \mid S_j)P(X_{j+1} \ldots X_M \mid S_j)$$

$$= \sum_{I_j} P(S_j)P(X_j \mid S_j)L_j(S_j)R_j(S_j)$$

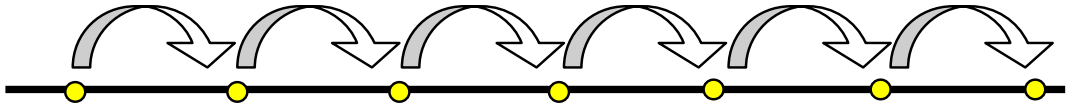
- A different arrangement of the same likelihood

- The nested summations are now hidden inside the Lj and Rj functions…

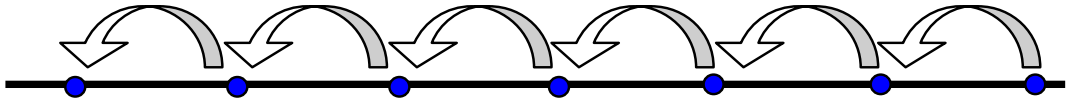
Pictorial Representation

- Single Marker

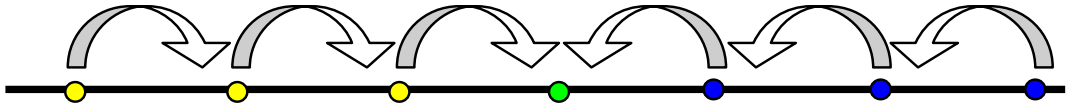
- Left Conditional



- Right Conditional



- Full Likelihood



Question:
What to do about missing data?

- What happens when some genotype data is unavailable?

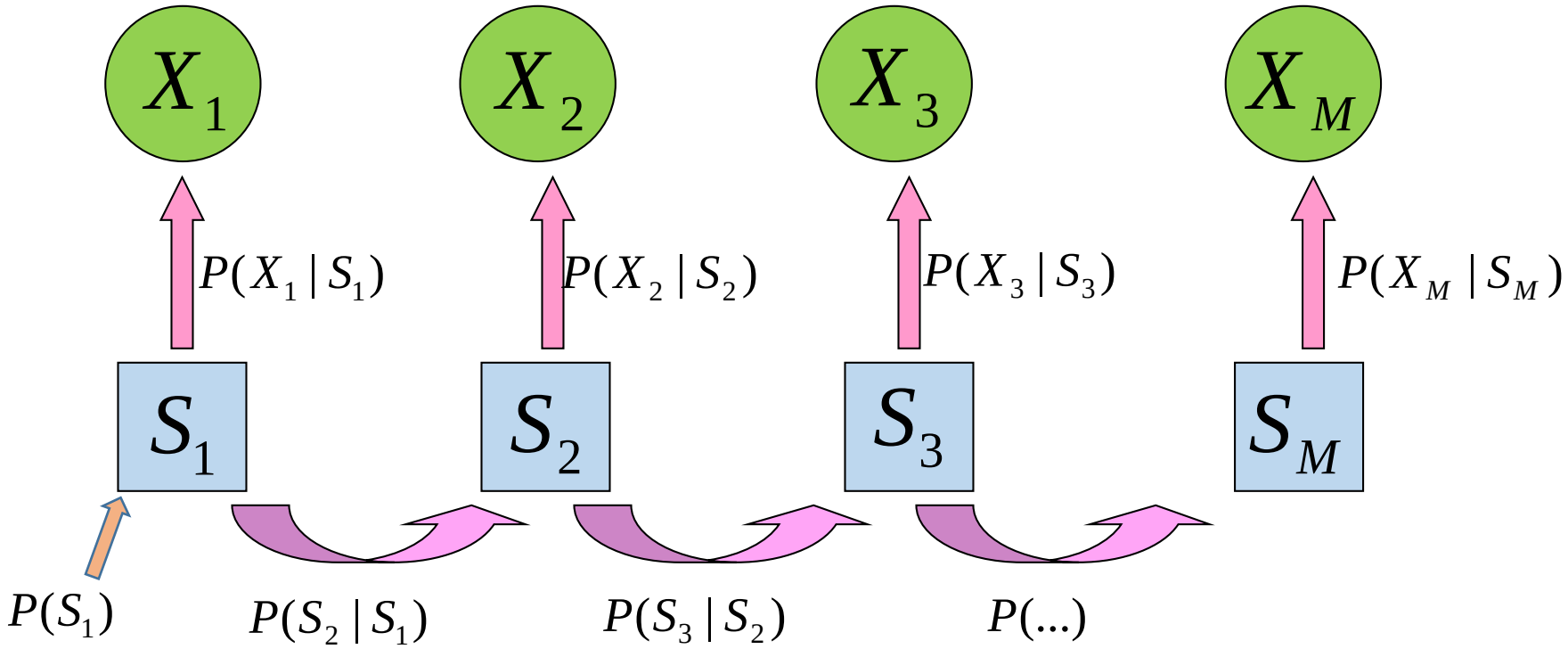
Some Assessments of the Model

Quality of haplotypes and missing genotypes estimates

			Dataset mimicking HapMap CEU			Dataset mimicking HapMap YRI		
Method	**# Iterations**	**Computation time**	**# Errors**	**# Flips**	**# Perfect**	**# Errors**	**# Flips**	**# Perfect**
MaCH	20	~2 min	11.6	216	26.5	17.9	256	22.6
	60	~5 min	10.8	200	28.4	16.6	232	24.1
	200	~15 min	10.6	192	29.1	16.3	222	25.1
	1,000	~1.4 hr	10.6	182	29.3	16.3	218	25.5
	3,000	~3.9 hr	10.5	178	29.7	16.1	214	25.7

Li et al (2010) Assessment of 100 x 1Mb regions

Markov Model



The final ingredient connects template states along the chromosome …

Today

- Efficient computational framework for modeling haplotype mosaics

Recommended Reading

- Chen and Abecasis (2007) Family based association tests for genome wide association scans. *Am J Hum Genet* **81:**913-926

- Li et al (2010) Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**:816-834