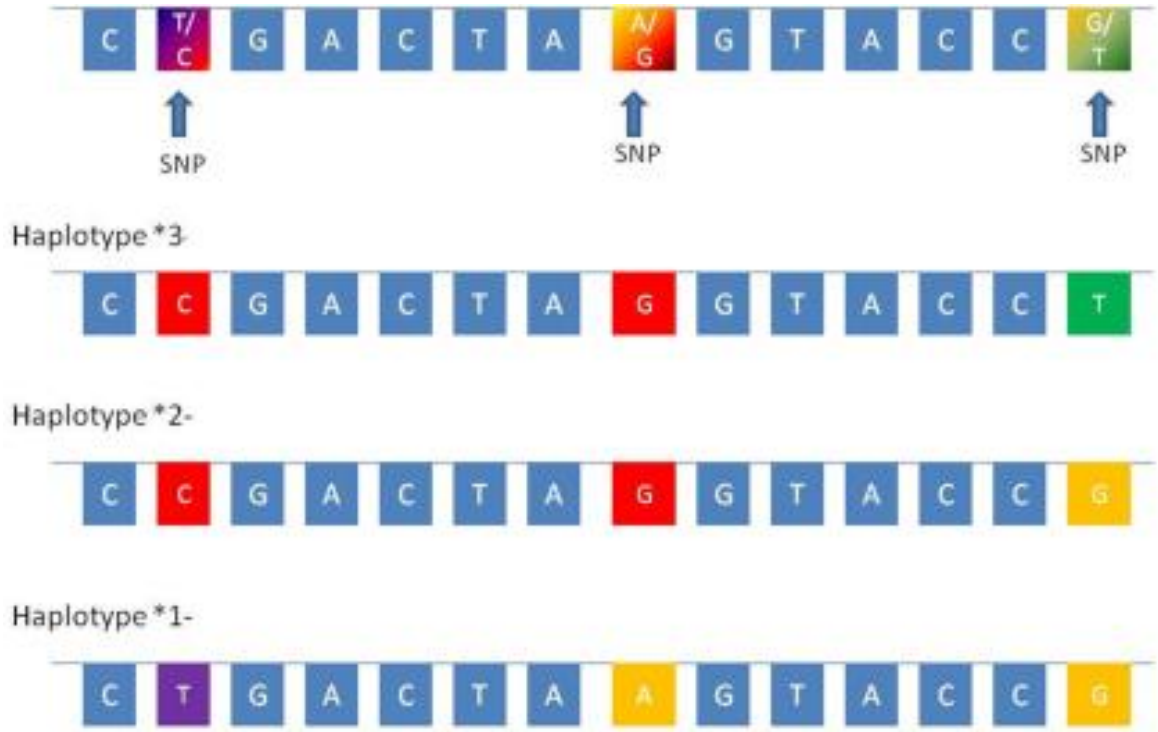
Detection of identity-by-descent (IBD) segments

Yuan Lin
Post-doc Research Associate
@ Dr. Kirk Wilhelmsen's Lab
Department of Genetics, School of Medicine
UNC at Chapel Hill

Definition

A IBD segment is a continuous chromosomal segment that delineates IBD haplotypes, which have identical alleles inherited without recombination from a common ancestor



Various applications

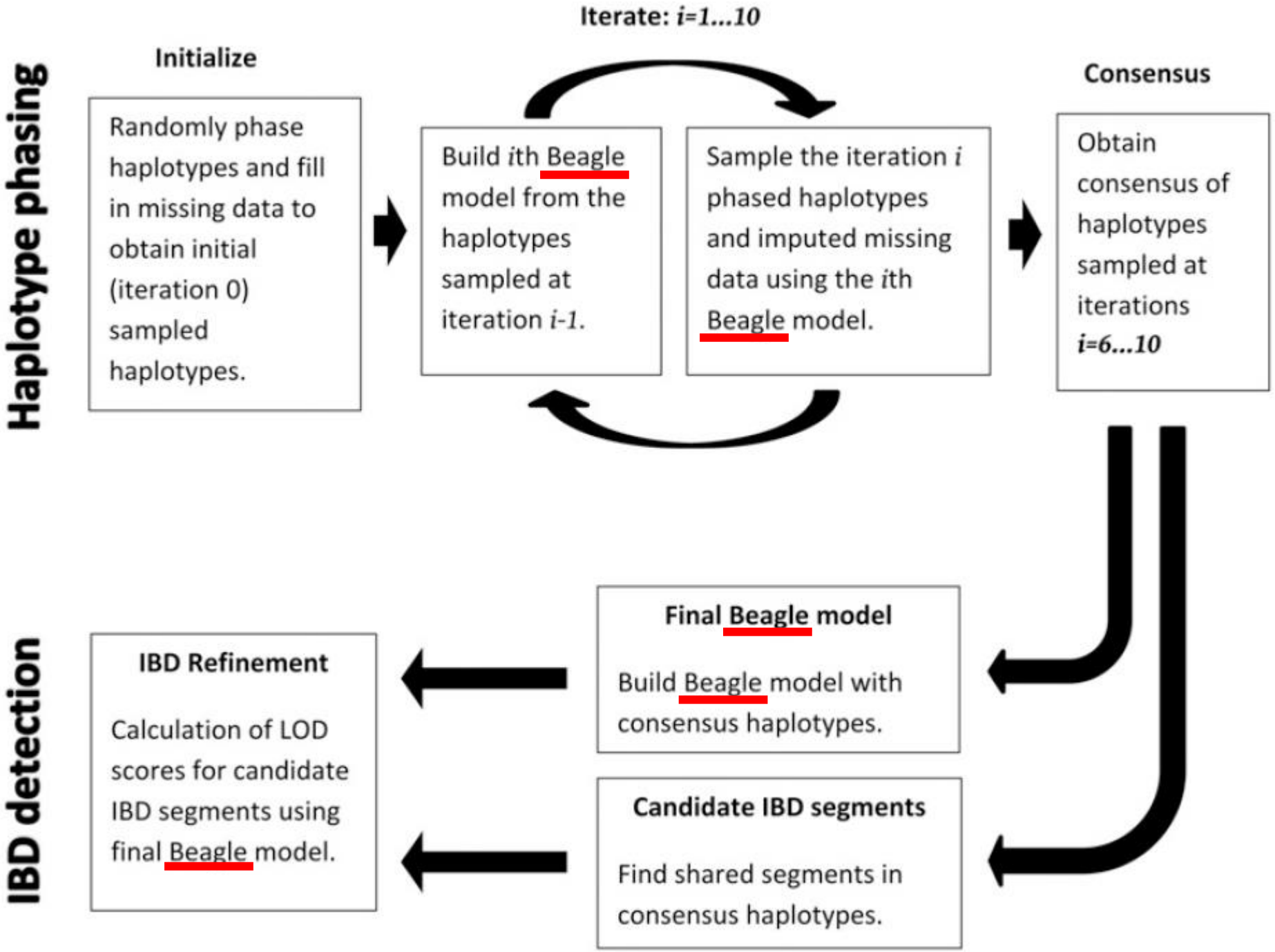
- IBD mapping
- Haplotype phasing and imputation
- Heritability analysis
- Inferring relationships and population structure
- Inferring signals of natural selection

…

Browning & Browning (2012)

Refined IBD

Browning & Browning (2013)

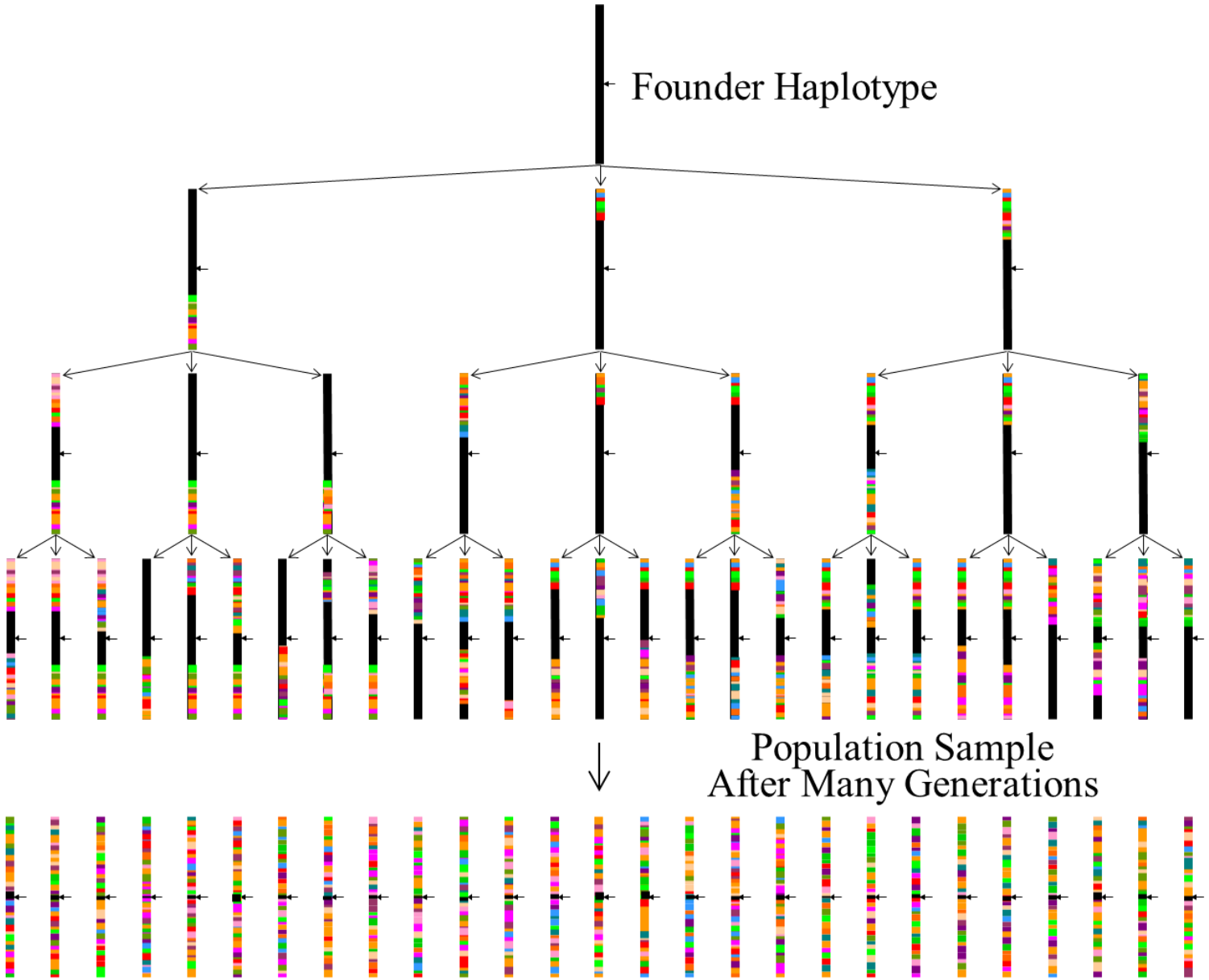


Outline

Why such a complicated process?

What is BEAGLE?

- Accuracy and efficiency of IBD detection methods
- Models/heuristic approaches applied in Refined IBD
- Comparison among IBD detection methods



Familial IBD

Ancient IBD

Recent IBD

- IBD segments from a common ancestor *N* generations ago have the expected length of 1/(2*N*) Morgans (*M*).
  - Familial IBD segments > 10cM
  - A SNP is the smallest ancient IBD segment (IBS)

- Refined IBD targets at recent IBD segments
  - Common ancestry is relatively “recent” (<= 25 generations ago)
  - Genetic length is <=2cM
  - The extent of sharing exceeds background linkage disequilibrium
  - Population data (unknown individual relatedness)

Detection methods

- Probabilistic
  - Fit a probabilistic model (usually HMM) using the data
  - For each shared haplotype, calculate its (a) posterior probability of IBD or (b) likelihood ratio between IBD and non-IBD
  - w/o incorporating LD
  - *Example: BEAGLE IBD, Refined IBD*
- Non-probabilistic/deterministic
  - Report shared haplotype as IBD segments based on their (genetic) length (cM) or frequencies
  - *Example: GERMLINE, fast IBD, Refined IBD*

Accuracy

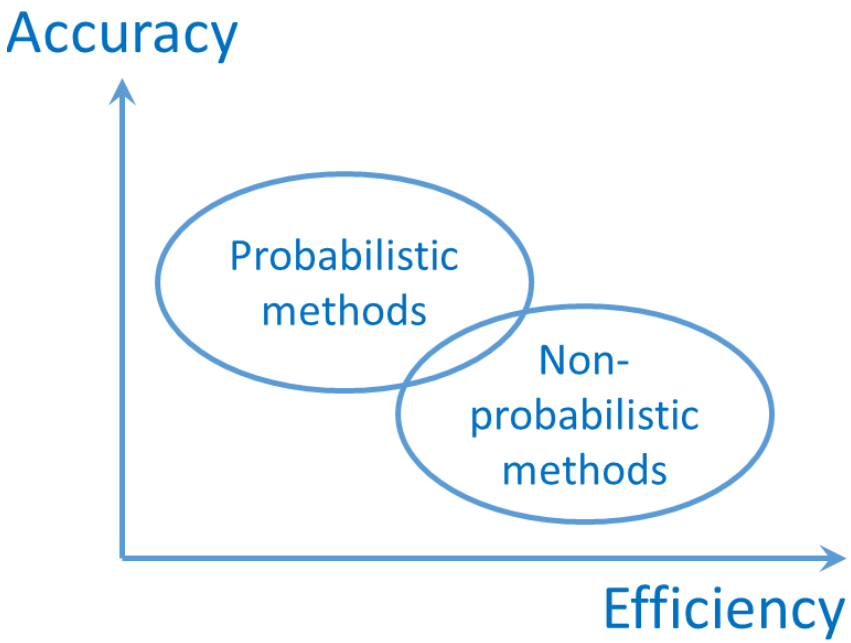
- Power / False negatives
  - Segment-level: % of true IBD segments included in an estimated IBD segments for the same pair of individuals
  - Marker-level: % of markers in a true IBD segment included in an estimated IBD segments for the same pair of individuals (underestimate the length)
- False positives
  - Segment-level: % of estimated IBD segments that do not cover a true underlying IBD segment
  - Marker-level: % of markers in an estimated IBD segment not contained in a true IBD segment (overestimate the length)

Efficiency

- Memory – usually not an issue
  - Divide long chromosomes into smaller pieces for analysis
  - Compare/analyze two or a subset of individuals at a time
  - Use hash tables
- Computation time

Factors that impact accuracy / efficiency

- Accuracy depends on accounting for
  - Linkage disequilibrium – false positives
  - Phasing errors – false negatives
  - Genotyping errors & missing data
- Computation time increases with
  - Sample size
  - Marker density
  - False positives



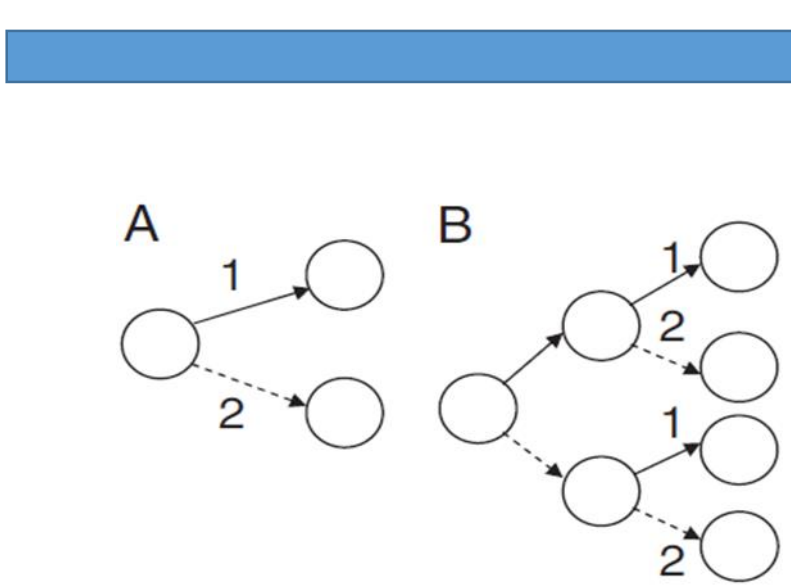
The story of BEAGLE

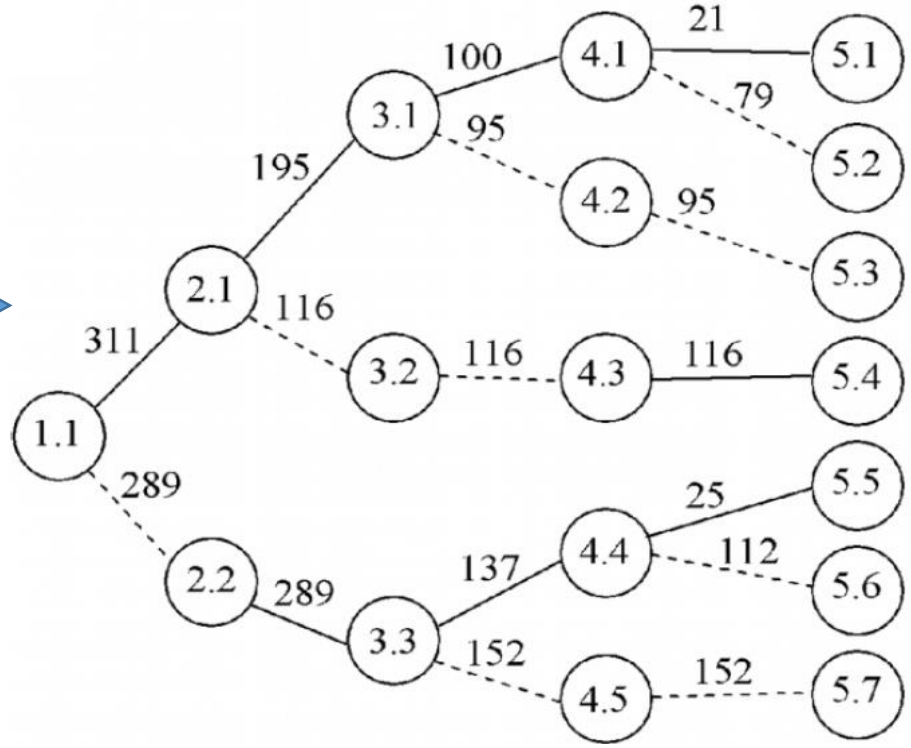
2006

2013

Haplotype frequency model in BEAGLE

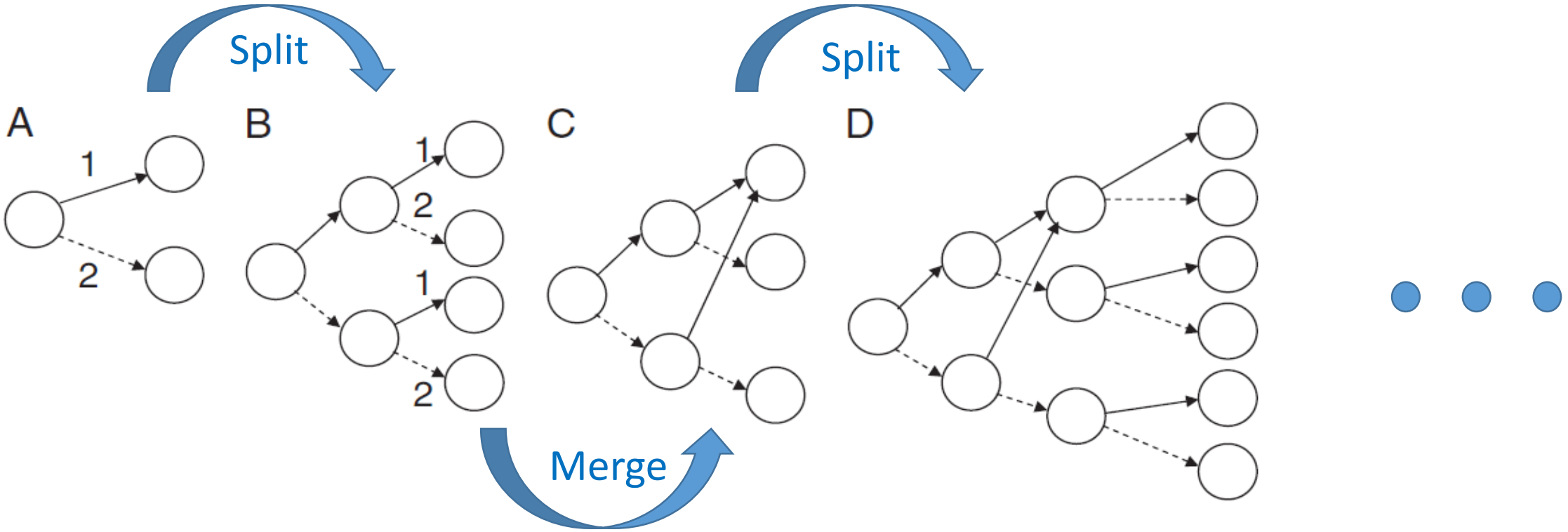
Haplotype	Total
1111	21
1112	79
1122	95
1221	116
2111	25
2112	112
2122	152





Browning (2006)

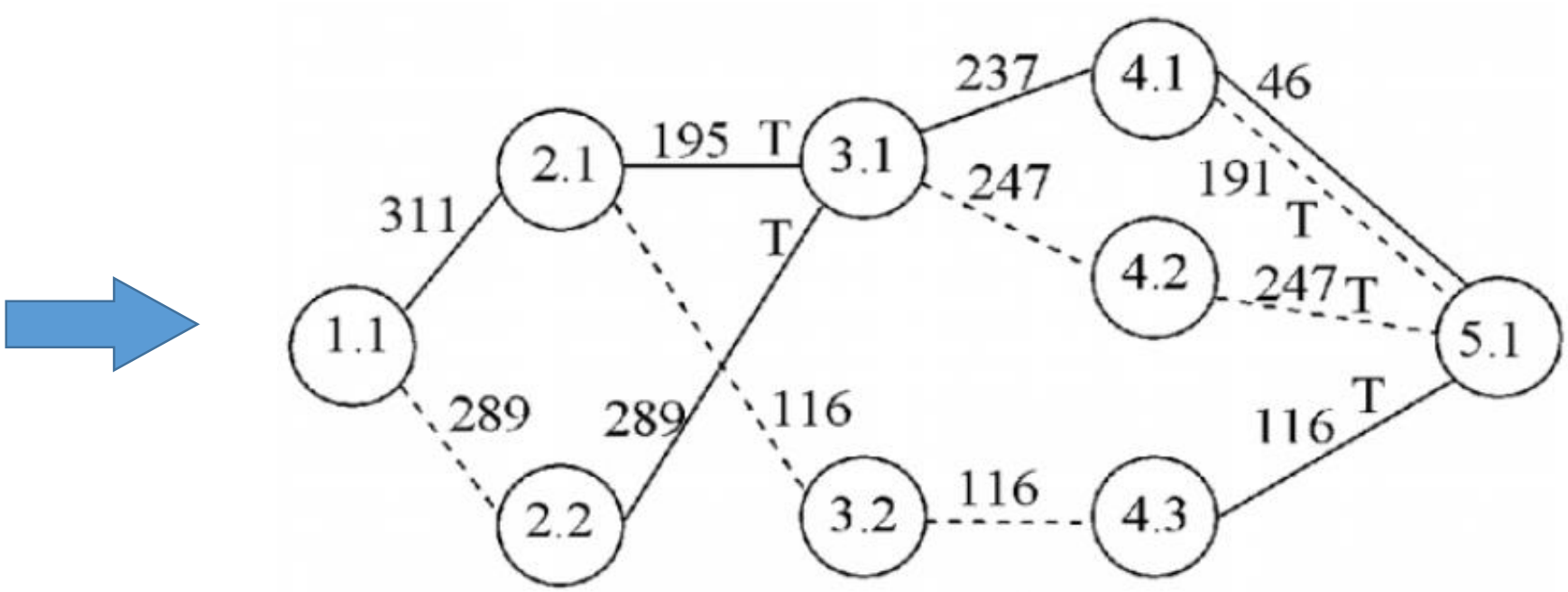
Haplotype frequency model in BEAGLE (cont.)



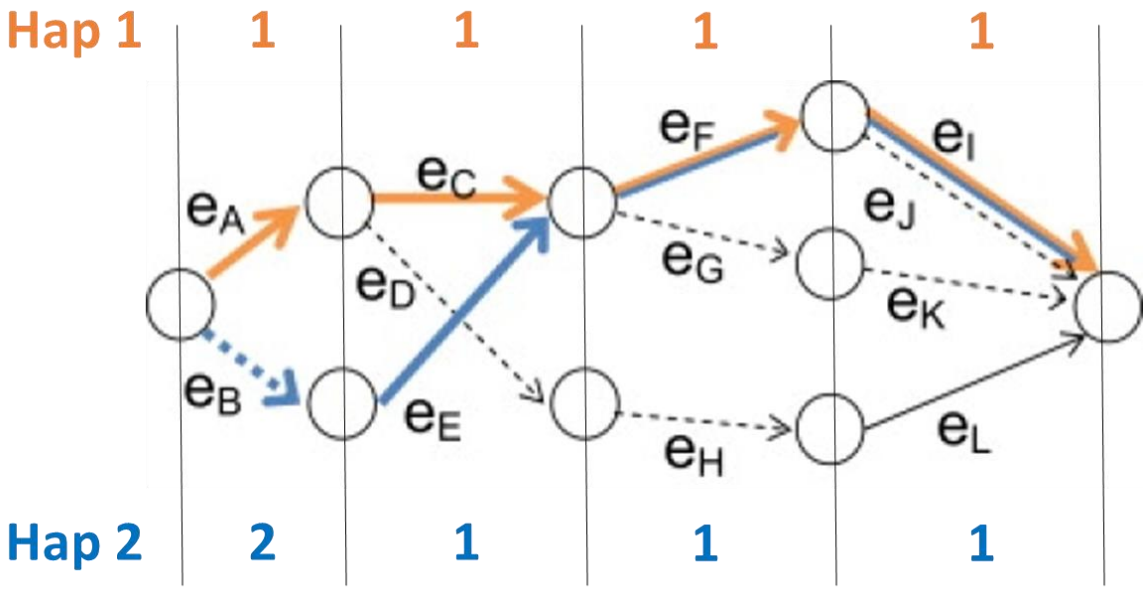
Browning & Browning (2007)

Haplotype frequency model in BEAGLE (cont.)

Haplotype	Total
1111	21
1112	79
1122	95
1221	116
2111	25
2112	112
2122	152



Haplotype frequency model in BEAGLE (cont.)



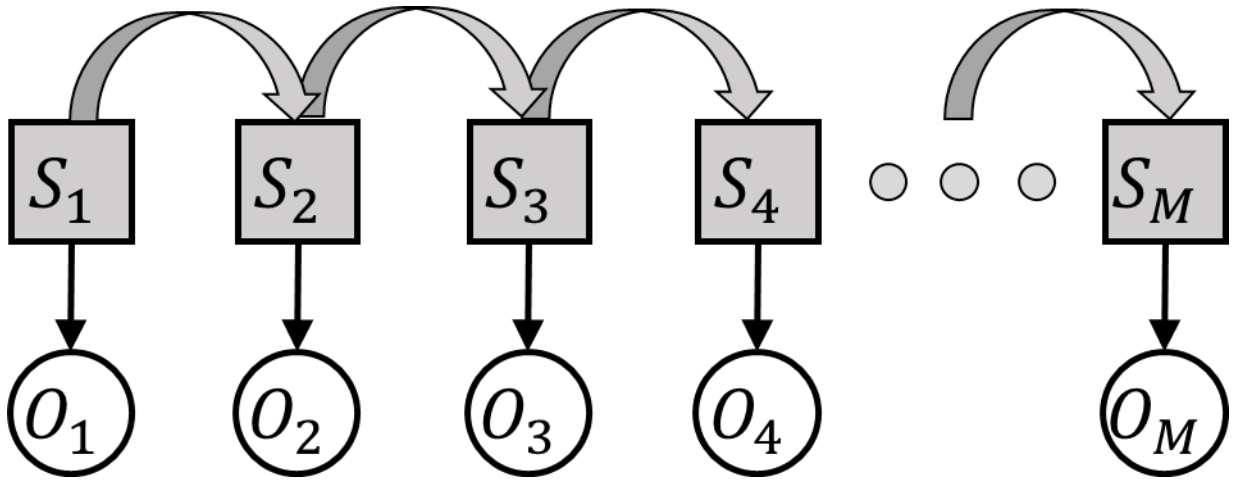
BEAGLE HMMs

- Haploid HMM (Browning & Browning, 2007)
- Diploid HMM (2007)
- IBD HMM (2009)
- Non-IBD HMM (2009)
- Unified HMM (2010)

Hidden Markov Models (HMMs) Recap

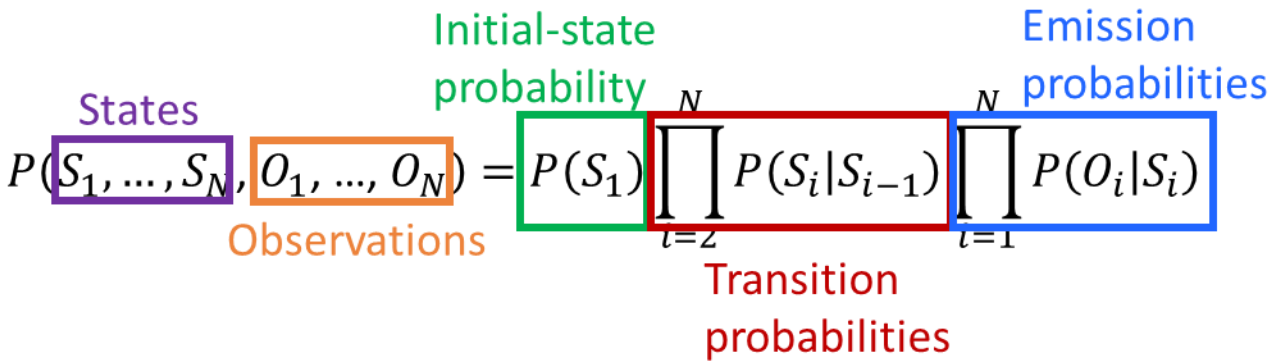
An HMM is defined by

- Hidden states $\{S_i\}$
- Observed values $\{O_i\}$
- Initial-state probability $P(S_1)$
- Emission probability $P(O_i|S_i)$
- State transition probability $P(S_i|S_{i-1})$



HMM Recap (cont.)

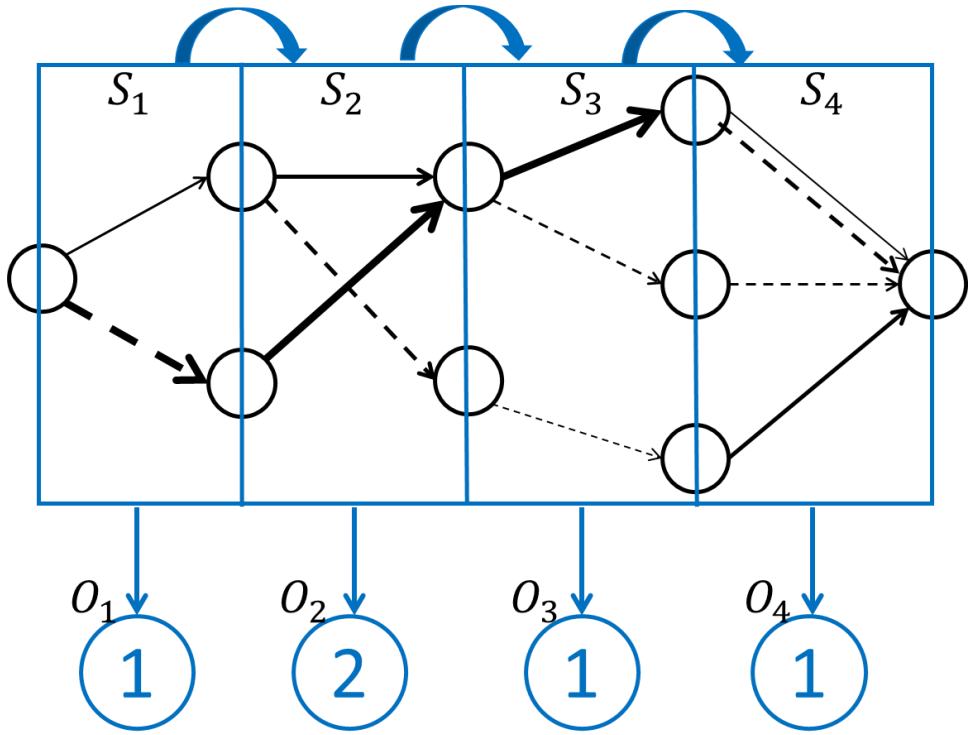
- The joint probability distribution defined by an HMM



- The forward-backward algorithms
  - What is the likelihood of a specific sequence of observations (given a specific HMM structure)?
- The Viterbi-algorithm
  - What is the most likely sequence of hidden states that generate the given sequence of observations?

Haploid HMM induced from LHC

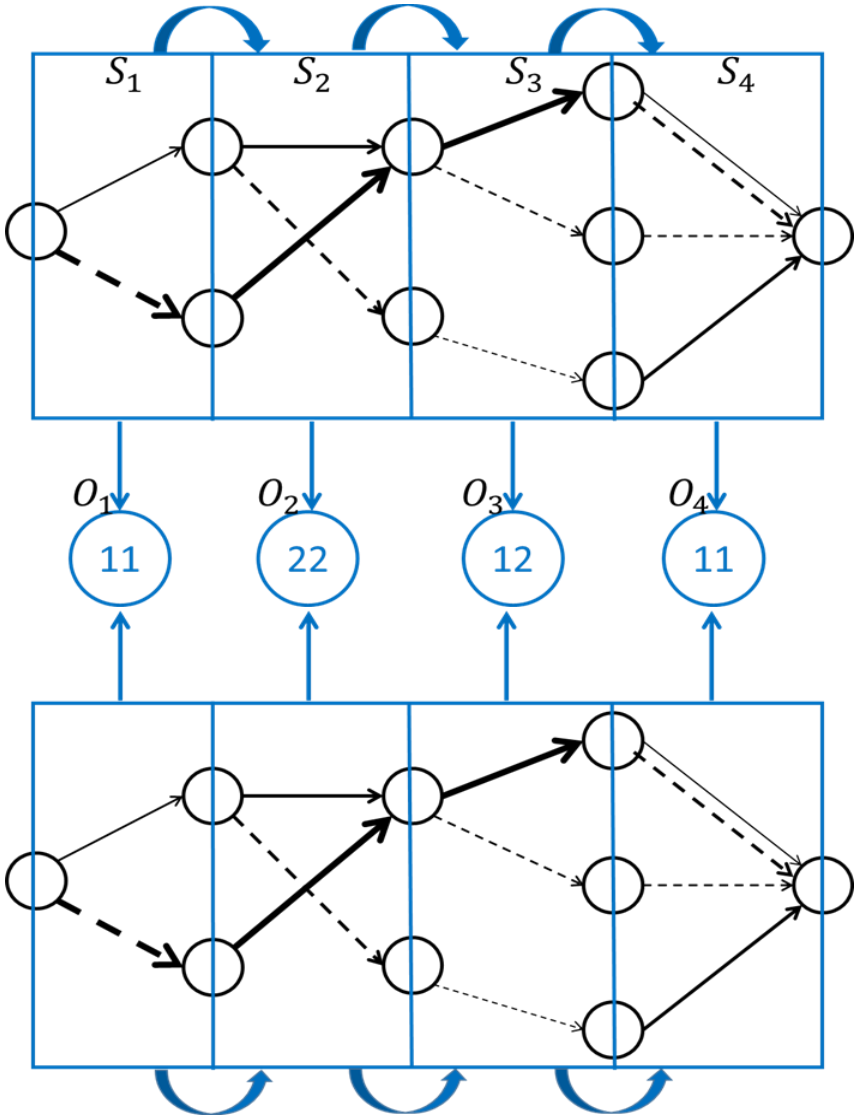
- States
  - Edges of LHC (i.e., local haplotype clusters)
- Observations
  - Alleles corresponding to edges
- Initial-state probabilities
  - $P(e) = n(e)/\sum$ if the parent node of e is the root, 0 otherwise
- Emission probabilities
  - Each state (edge) emits with probability 1 its label allele or missing allele, 0 otherwise
- State transition probabilities
  - $P(e2|e1) = n(e2)/n(e1)$ if the parent node of $e2$ is the child node of $e1$, 0 otherwise



Browning & Browning (2007)

Diploid HMM

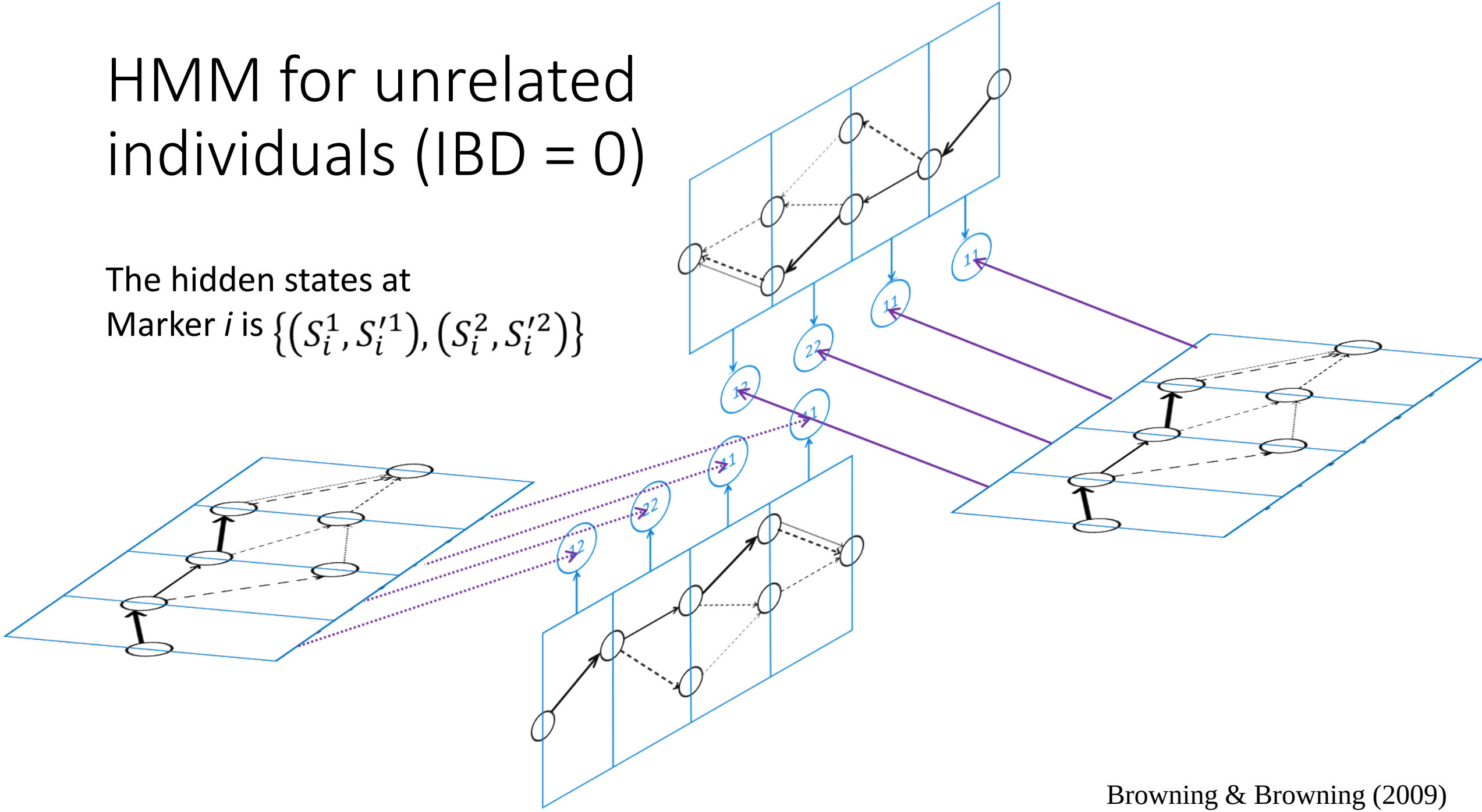
- States
  - Ordered pairs of edges from 2 copies of LHC
- Observations
  - Unordered genotypes of individuals
  - allow for genotype errors and missing data
- Emission probabilities
  - 1 for edge labels compatible with genotypes, including missing data, 0 otherwise
- Initial-state and Transition probabilities
  - Product of corresponding haploid HMM probabilities (assuming HW equilibrium)
  - $P(e1, e2) = P(e1)P(e2)$, $P((e3,e4)|(e1,e2)) = P(e3|e1)\,P(e4|e2)$



Browning & Browning (2007)

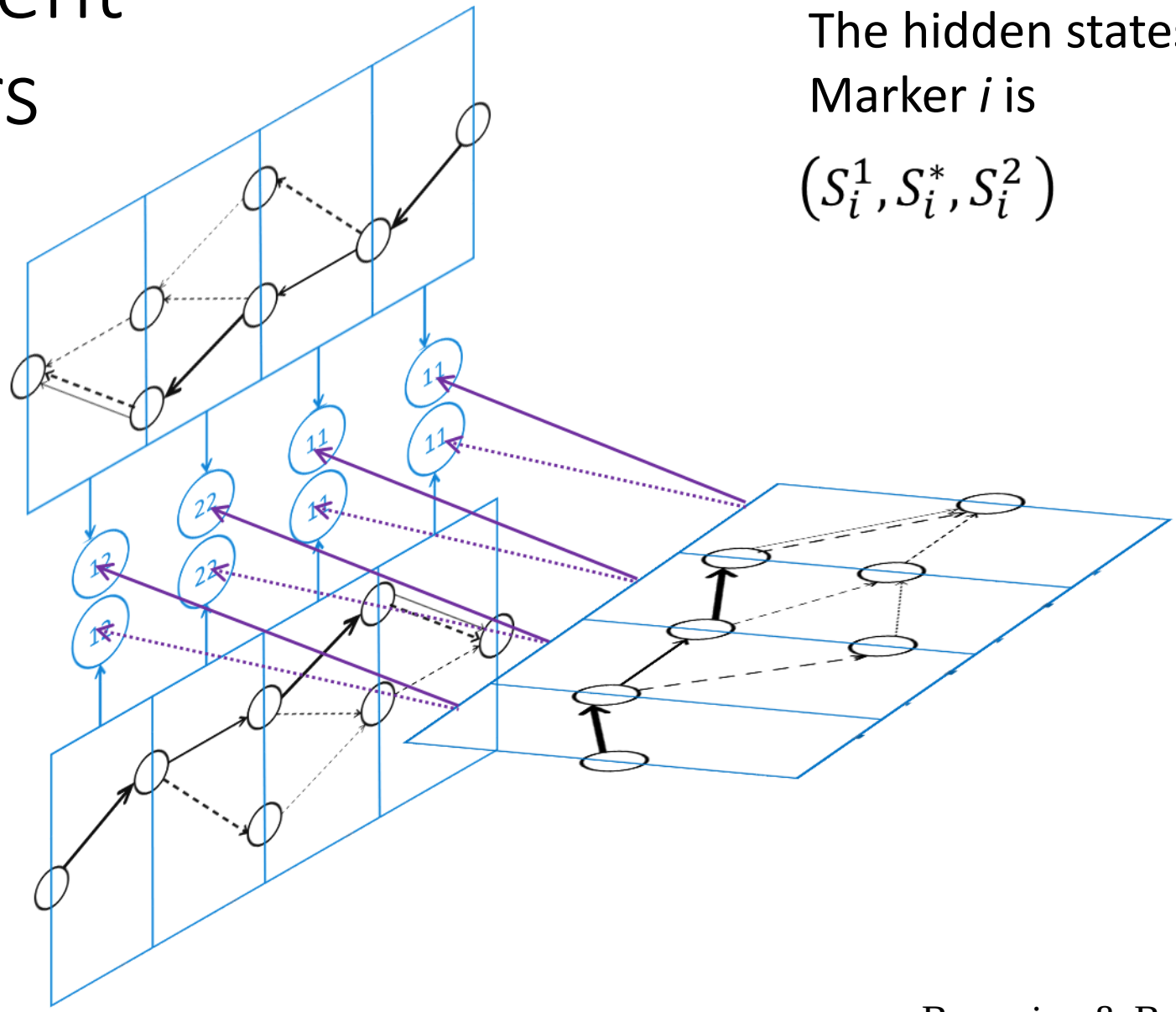
HMM for unrelated individuals (IBD = 0)

The hidden states at Marker *i* is $\{(S_i^1, S_i'^1), (S_i^2, S_i'^2)\}$



Browning & Browning (2009)

HMM for parent-offspring pairs (IBD = 1)



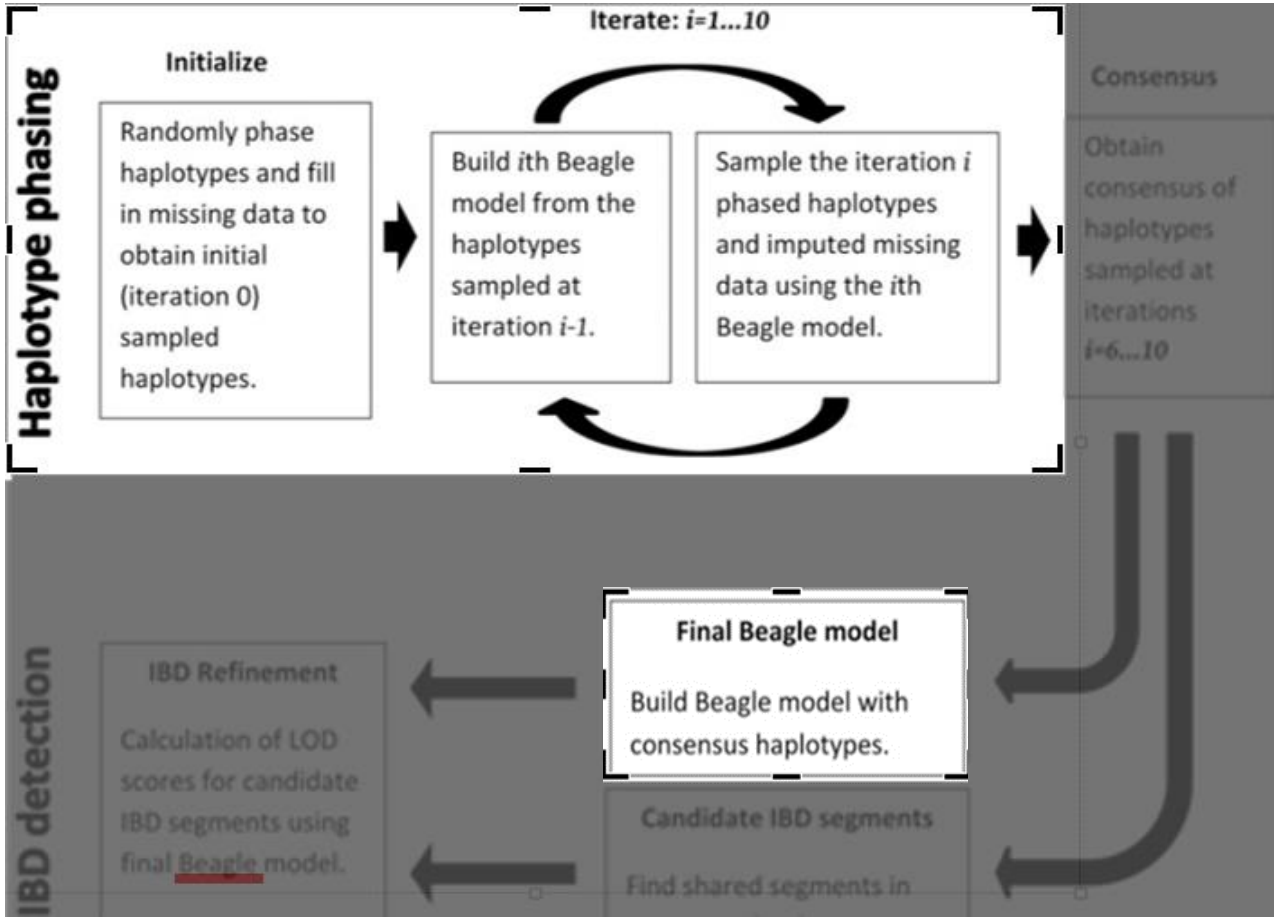
The hidden states at Marker *i* is

(S_i^1, S_i^*, S_i^2)

Browning & Browning (2009)

BEAGLE HMMs in Refined IBD

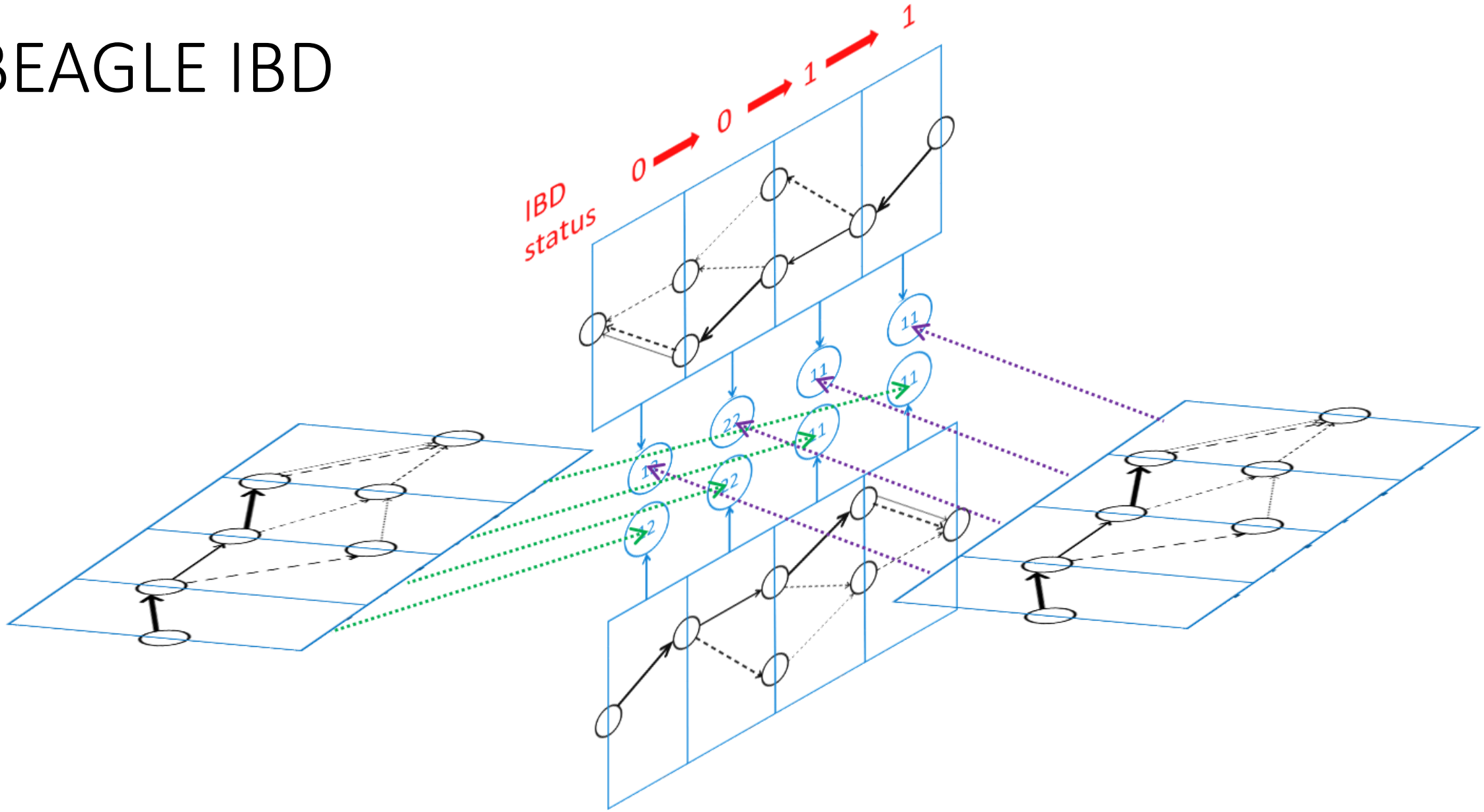
- Used for Phasing
  - Observations are genotypes
  - Alternate between fitting a BEAGLE HMM using sampled haplotypes (forward) and sampling new haplotypes based on the newest model (backward)
  - I0 iterations by default
  - Obtain consensus haplotypes



BEAGLE HMMs in Refined IBD (cont.)

- Used for IBD Detection
  - Observations are haplotypes
  - Build the haplotype frequency model (i.e., LHC) using consensus haplotypes
  - Estimate the IBD and non-IBD likelihood of a (trimmed) candidate IBD segment using IBD and non-IBD HMMs respectively
  - Calculate the LOD score, defined as the base 10 logarithm of the IBD likelihood divided by the non-IBD likelihood
  - Report final IBD segments whose LOD score exceeds a threshold

BEAGLE IBD

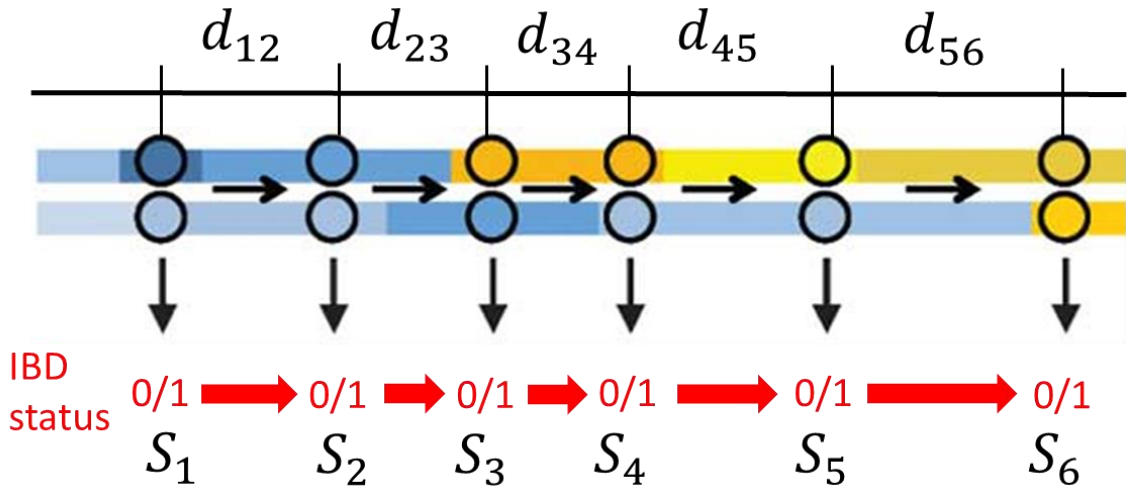


BEAGLE IBD

- Calculate posterior IBD probabilities using a single HMM that comprises two copies of diploid HMMs and one IBD model
- State at a given marker
  - Haplotype clusters at corresponding level (diploid HMM state)
  - IBD status for a pair of individuals at that marker (0/1)
- Observations at a given marker
  - Two individuals' genotype data and their all possible phasing
- Emission probabilities
  - 1 when phased haplotypes compatible with genotypes; 0 otherwise
- Initial-state probabilities
  - Product of corresponding initial-state probabilities from all haploid HMMs and the IBD model
- Transition probabilities
  - Multiply corresponding transition probabilities from haploid HMMs and the IBD model
  - The form of transition probabilities varies with the IBD status of the destination state (at a marker)
  - Incorporate genotype error when the destination state is IBD

Browning (2008), Browning & Browning (2010)

IBD model in BEAGLE IBD



- Model the changes in marker IBD status along a given chromosomal region as a Markov chain
  - States: Marker IBD status (0 for non-IBD and 1 for IBD)
  - Current state only depends on the previous state
  - Initial-state probabilities

$$P(S_1 = 1) = .0001$$
$$P(S_1 = 0) = 1 - P(S_1 = 1)$$

- Transition probabilities

$$P(S_i = 1 \rightarrow S_j = 1) = exp(-t_{10} d_{ij})$$
$$P(S_i = 0 \rightarrow S_j = 0) = exp(-t_{01} d_{ij})$$
$$P(S_i = 1 \rightarrow S_j = 0) = 1 - P(S_i = 0 \rightarrow S_j = 0)$$
$$P(S_i = 0 \rightarrow S_j = 1) = 1 - P(S_i = 0 \rightarrow S_j = 0)$$

  - Transition rate t_{10} = 1/cM, t_{01} = .0001/cM
  - d_{ij} is the genetic distance (cM) between two consecutive marker *i* and marker *j*

Browning (2008)

BEAGLE IBD – Transition probabilities

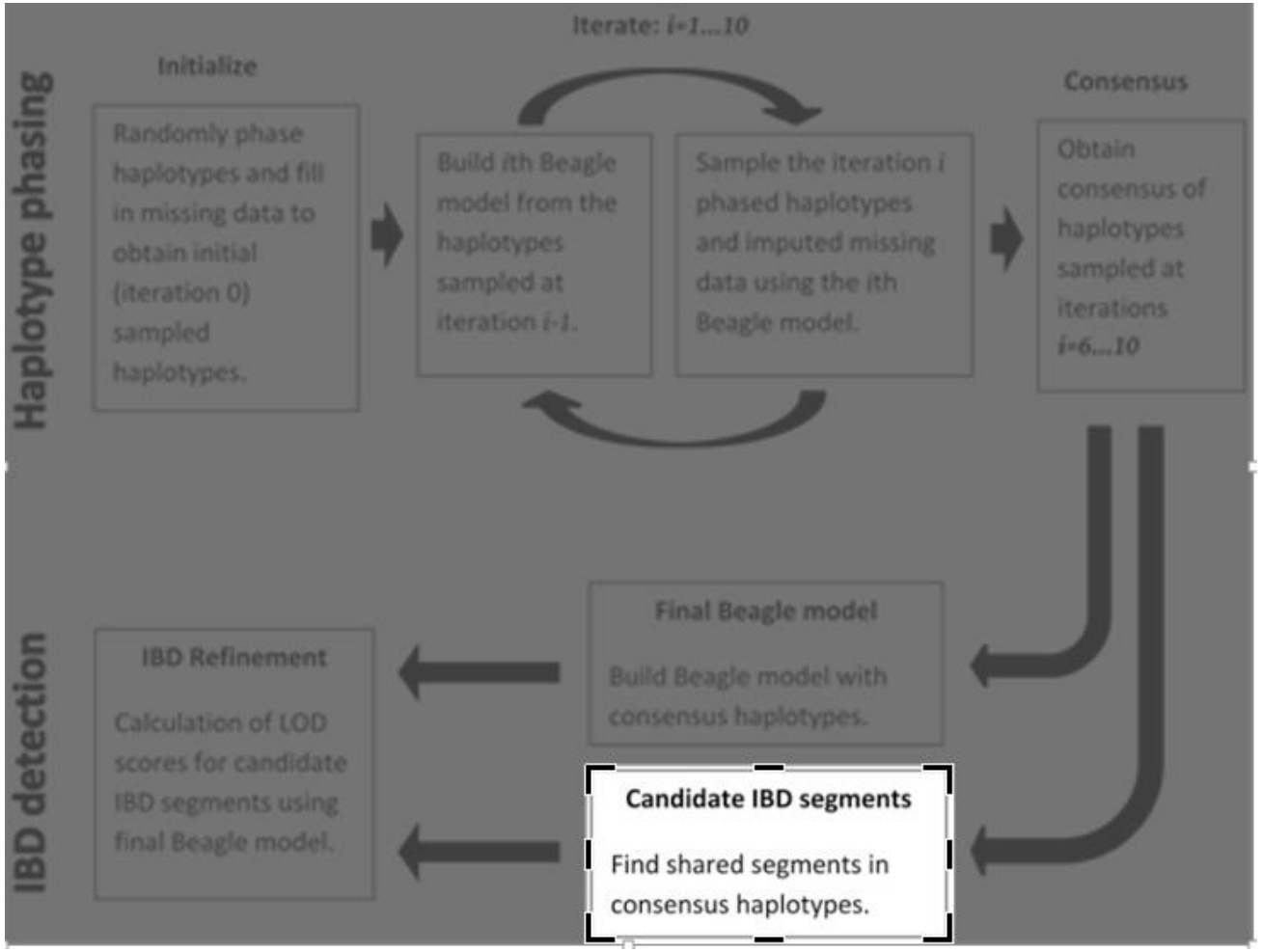
- If the destination state is non-IBD ($i' = 0$), the transitions of four haplotypes are conditionally independent
  - Overall transition probability (from State i to State i') = IBD model transition probability (S_{i0}) × P(e_1->e_1') × P(e_2->e_2') × P(e_3->e_3') × P(e_4->e_4')
- If the destination state is IBD ($i' = 1$), two of the four haploid HMMs are completely dependent at that marker and only one of their transition probabilities should go into the overall transition probability, but it is unclear which one.
  - Overall transition probability = IBD model transition probability (S_{i1}) × min(P(e_1->e_1'), P(e_3->e_3')) × P(e_2->e_2') × P(e_4->e_4')

BEAGLE IBD – Transition probabilities (cont.)

- If observed marker alleles of two IBD haplotypes are different, a genotype error is assumed
  - Overall transition probability = IBD model transition probability (S_{i1}) × min(P(e_1->e_1'), P(e_3->e_3')) × P(e_2->e_2') × P(e_4->e_4') × ε
- Otherwise, there is no genotype error
  - Overall transition probability = IBD model transition probability (S_{i1}) × min(P(e_1->e_1'), P(e_3->e_3')) × P(e_2->e_2') × P(e_4->e_4') × (1- ε)

GERMLINE as a pre-filter

- Refined IBD uses GERMLINE to identify (from the genome-wide data of all pairs of individuals) only individual pairs that are likely to have IBD in some genomic regions



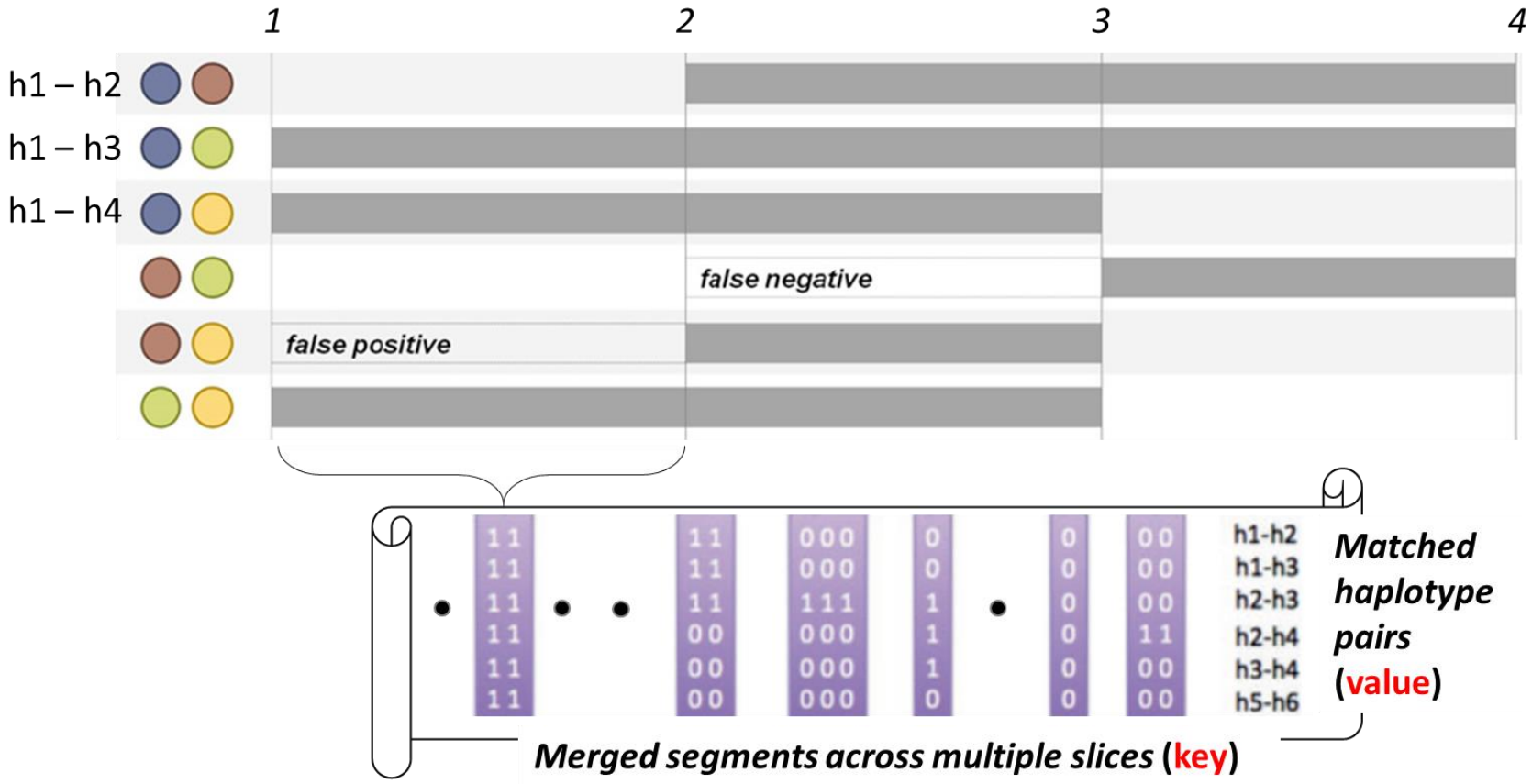
GERMLINE: slices

- Divide a chromosome into non-overlapping slices of equal width and search for (nearly) identical segments at each slice.
  - Much fewer distinct haplotypes in each slice than in the whole chromosome
- Allow for a small number of mismatches in each slice to accommodate genotype error and missing data
  - Choose an error rate that ensures the expected number of matching slices > 1

(Gusev et al., 2009)

GERMLINE: hash-table

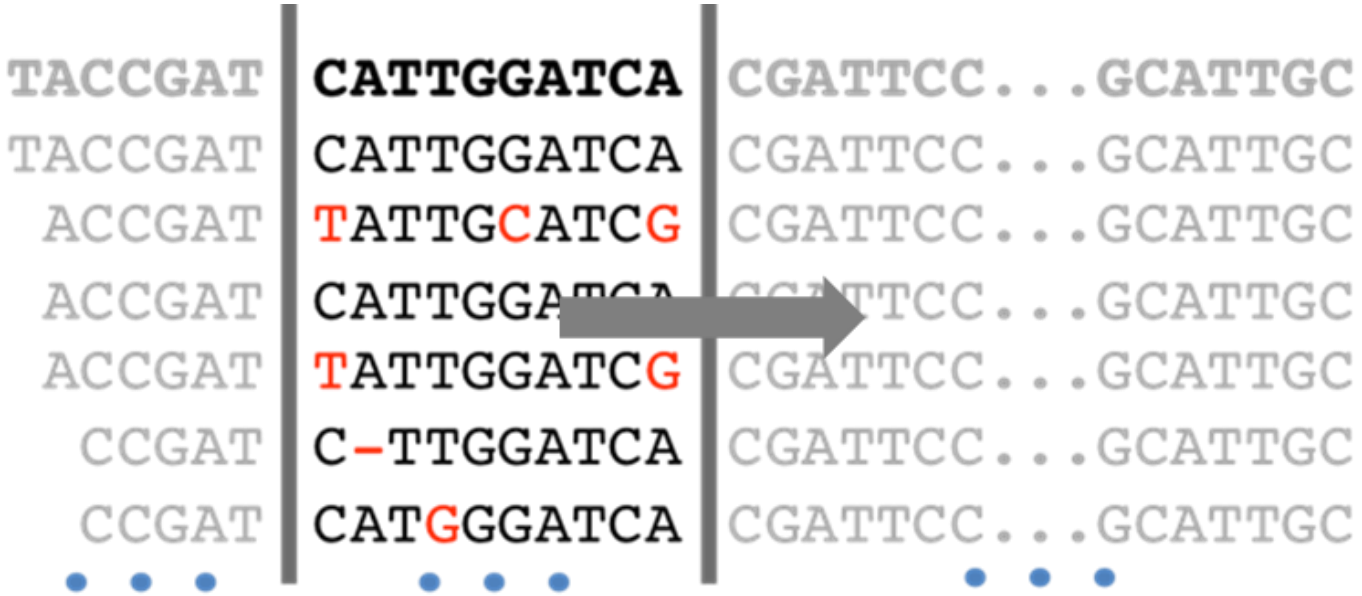
Matched (nearly identical) segments in consecutive slices and from the same pair of individuals are connected to form a longer segment



A hash table is created from the data to quickly identify all individuals (as value) who share a haplotype fragment (as key).

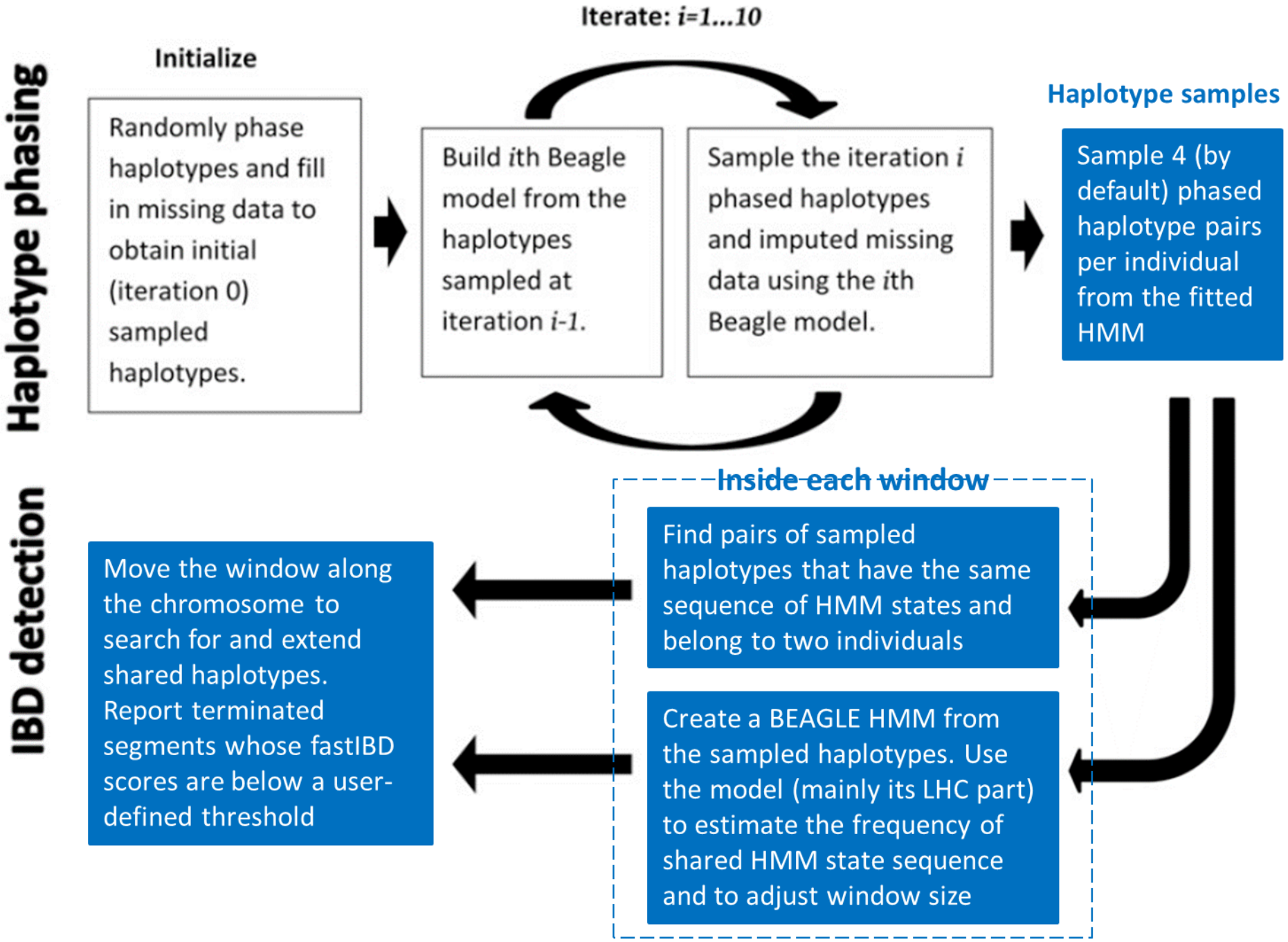
Sliding window instead of slices

- The problem of non-overlapping slices – false negatives
- Remedy: a sliding window



Fast IBD

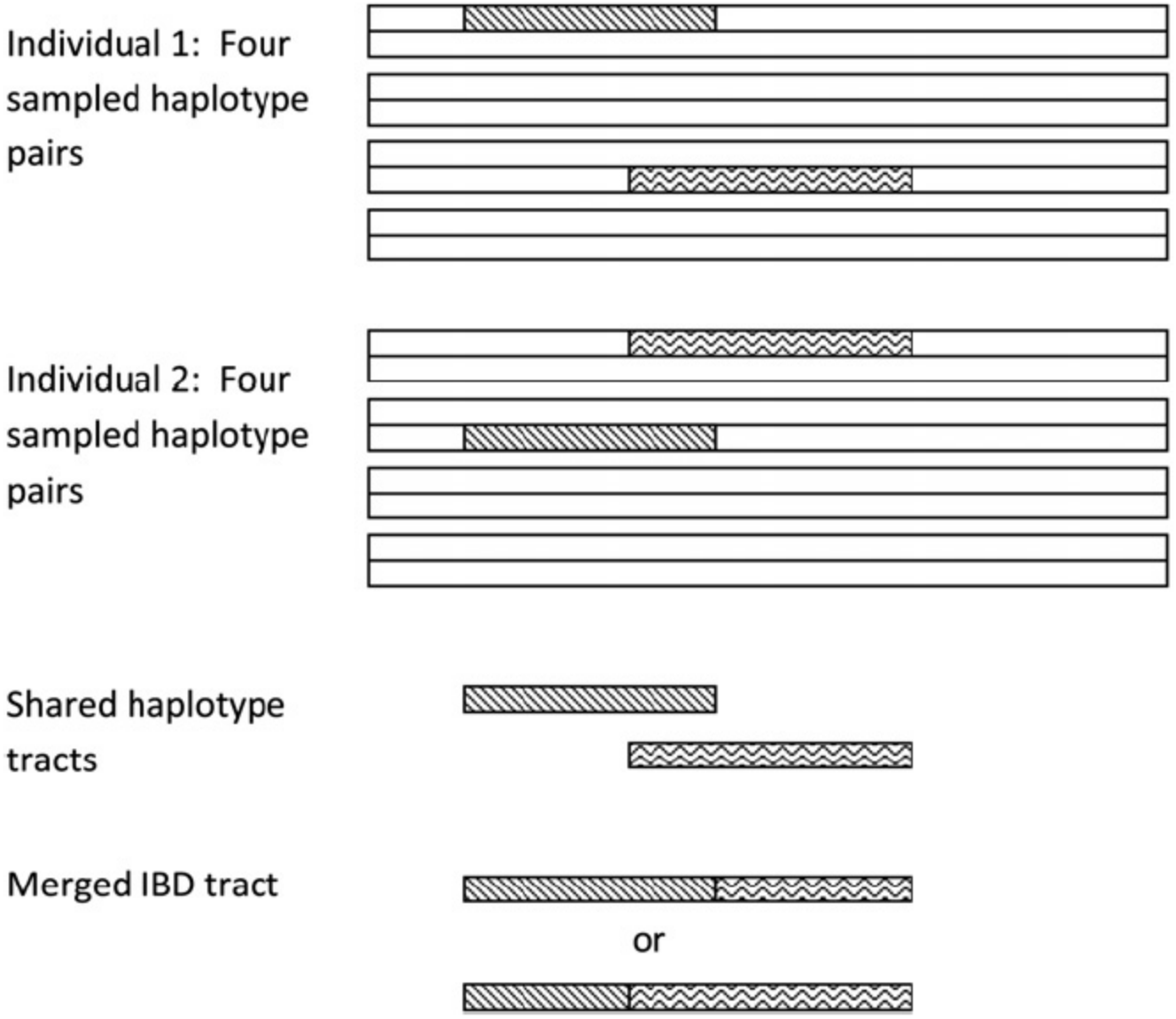
Browning & Browning (2011)



Fast IBD (cont.)

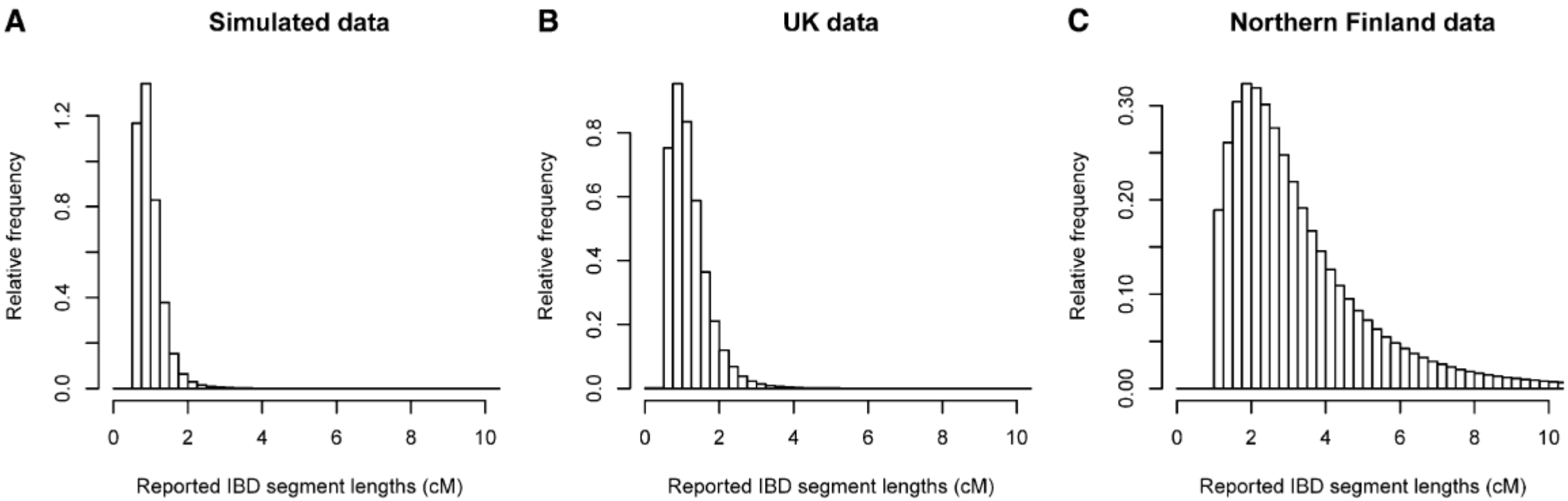
Example – Merging of Shared Haplotype Tracts

Four pairs of haplotypes have been sampled from individuals 1 and 2. Two shared haplotypes are found and merged into a single shared haplotype.

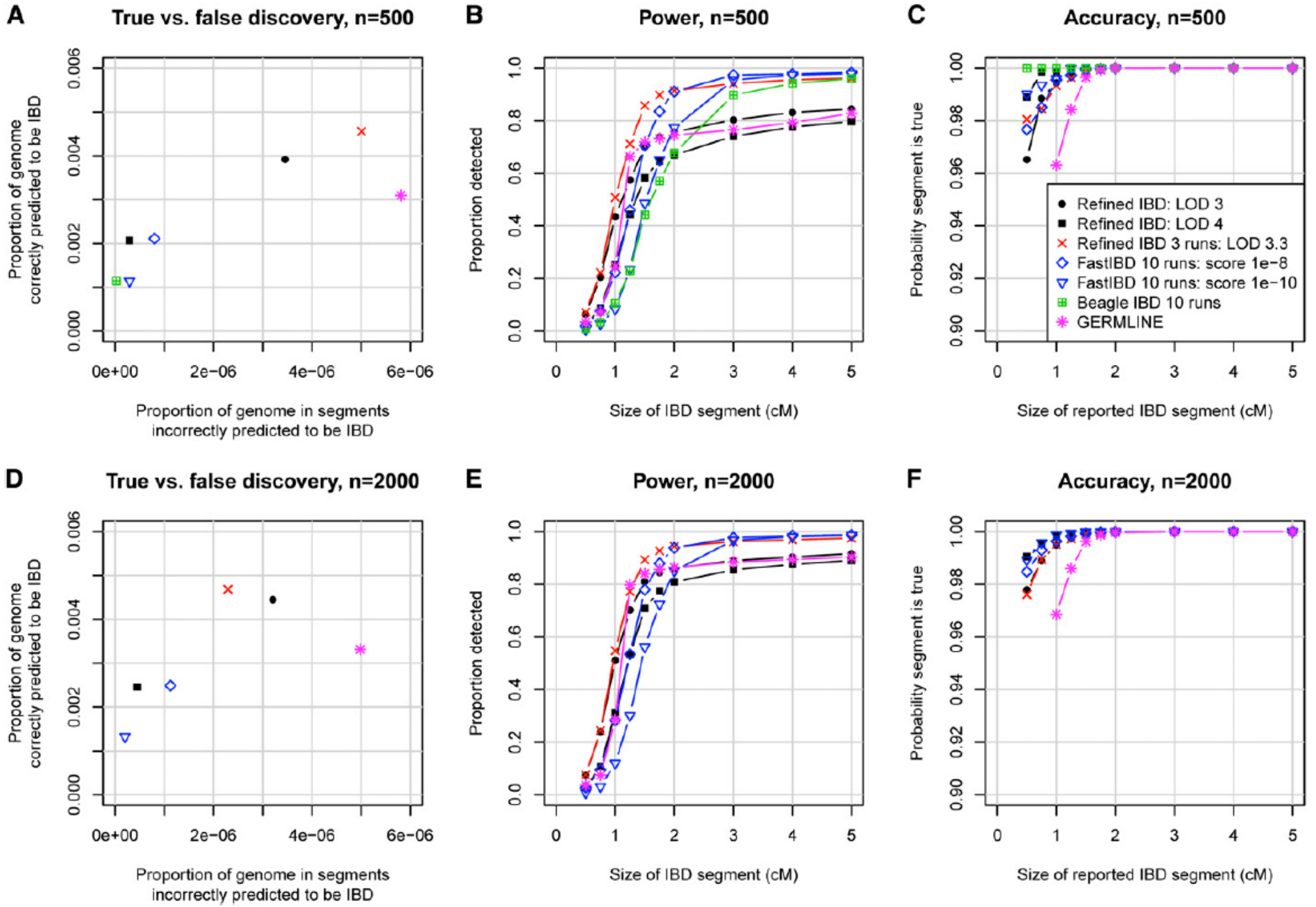


Data

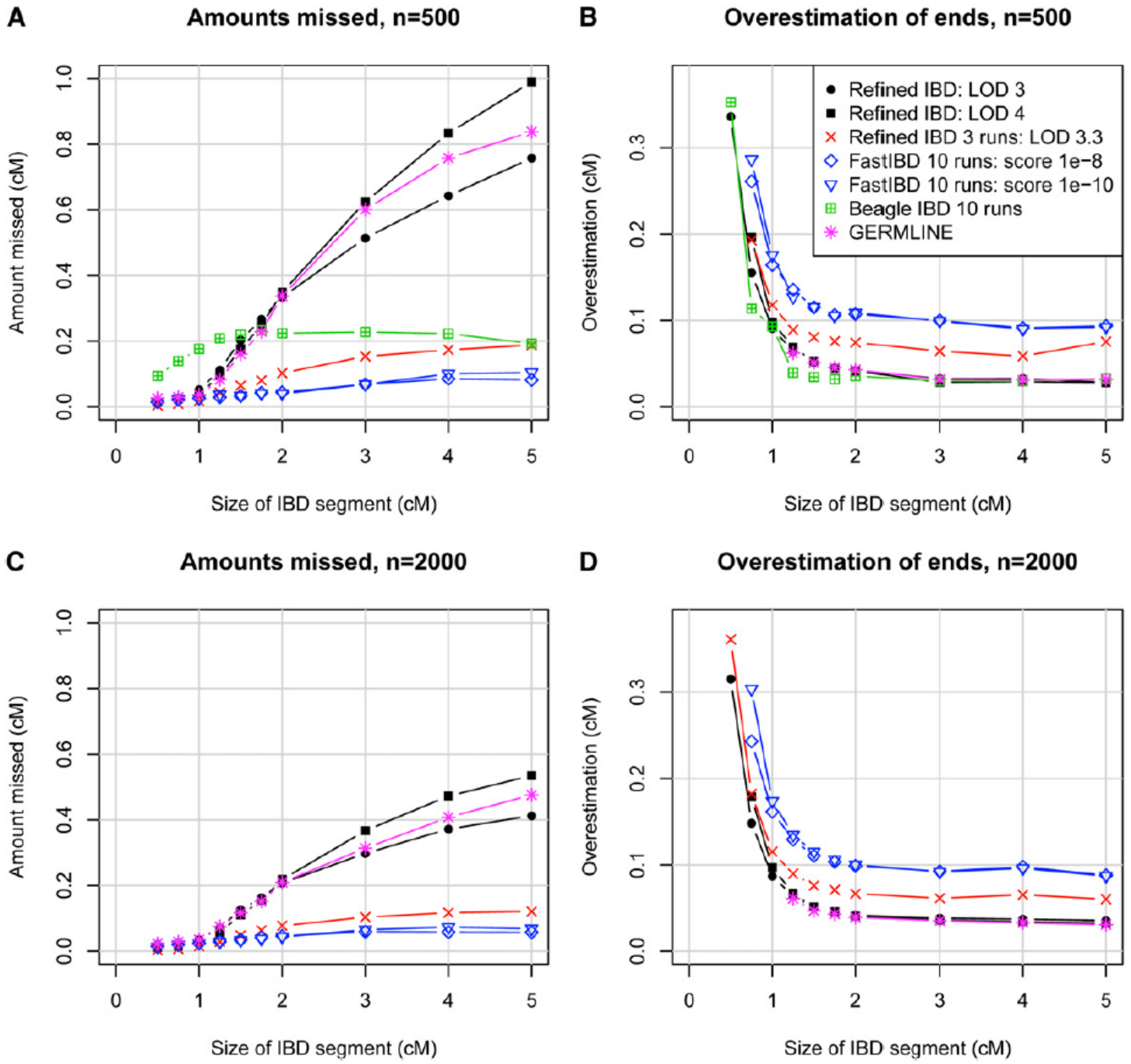
- Simulated data
  - Generated from a coalescent model
  - Attempt to simulate realistic effective population size
  - SNP array data and sequence data
- Real data
  - Wellcome Trust Case Control Consortium 2 data
  - Northern Finland Birth Cohort data


A
Simulated data
Relative frequency
0.0
0.4
0.8
1.2
0
2
4
6
8
10
Reported IBD segment lengths (cM)
B
UK data
Relative frequency
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
Reported IBD segment lengths (cM)
C
Northern Finland data
Relative frequency
0.00
0.10
0.20
0.30
0
2
4
6
8
10
Reported IBD segment lengths (cM)

Method Comparison

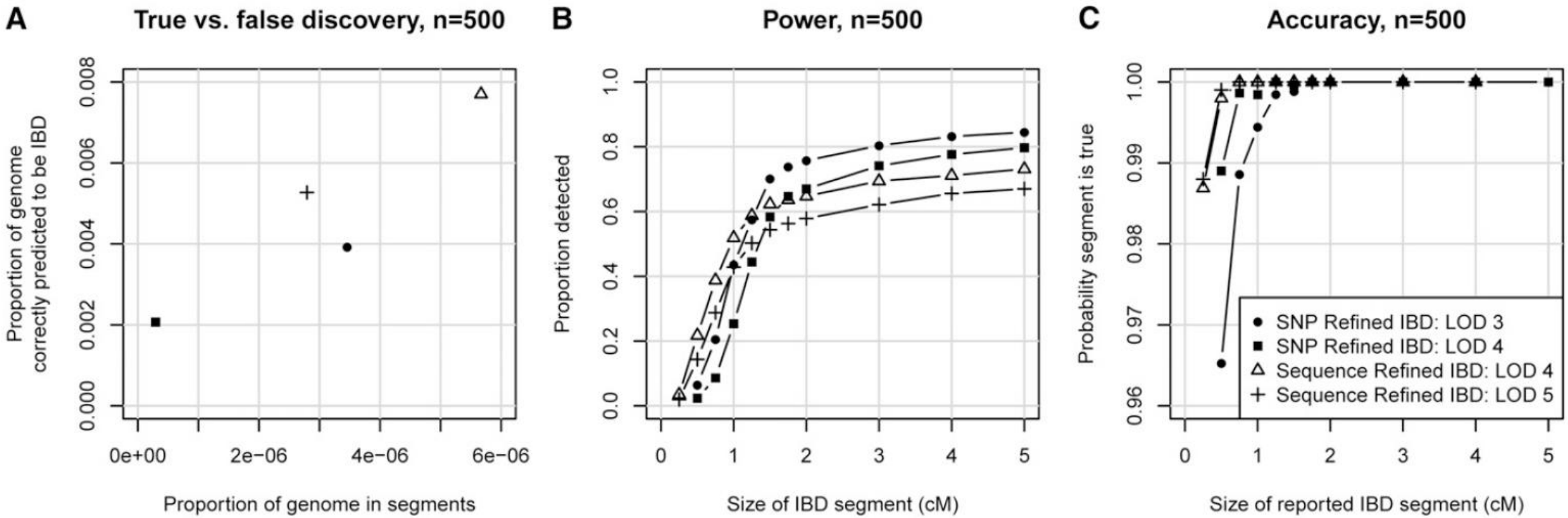


Method Comparison (cont.)



SNP array data vs. sequence data

- Rare variants
  - Rare variants are more informative for IBD detection than common variants.
  - Current IBD resolution may be restricted by the scarcity of rare variants in SNP array data. Analyzing sequence data that contain more rare variants should improve the power to detect short IBD segments.
  - However, extreme rare variants could be mutations since the most recent common ancestor. Their presence disrupts segment identity.
- Genotyping and phasing errors
  - Sequence data may have more genotype errors and phasing errors that affect the detection of long IBD segments


A
True vs. false discovery, n=500
Proportion of genome correctly predicted to be IBD
0.000
0.002
0.004
0.006
0.008
0e+00
2e−06
4e−06
6e−06
Proportion of genome in segments
B
Power, n=500
Proportion detected
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
Size of IBD segment (cM)
C
Accuracy, n=500
Probability segment is true
0.96
0.97
0.98
0.99
1.00
SNP Refined IBD: LOD 3
SNP Refined IBD: LOD 4
Sequence Refined IBD: LOD 4
Sequence Refined IBD: LOD 5
0
1
2
3
4
5
Size of reported IBD segment (cM)

Conclusion

- Refined IBD can efficiently determine pairwise IBD sharing in a large sample of thousands of individuals over the whole genome to a resolution of 0.5 – 1 cM with high power than existing methods and similar level of accuracy.
- Accurate and efficient detection algorithms can be created by using heuristic approaches to identify candidate IBD segments and using probabilistic models to refine the results.

References

- Browning, S. R., and B. L. Browning, 2012 Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. 46: 617–633.
- Browning, B. L., and S. R. Browning, 2011 A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88: 173–182.
- Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84: 210–223.
- Browning, B. L., and S. R. Browning, 2007 Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet. Epidemiol. 31: 365–375.
- Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084–1097.
- Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86: 526–539.
- Browning SR. 2006. Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* 78:903–13
- Browning SR. 2008. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* 178:2123–32
- Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler et al., 2009 Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19: 318–326.
- Gusev, A., E. E. Kenny, J. K. Lowe, J. Salit, R. Saxena et al., 2011 DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. 88: 706–717.