LD-based statistical phasing: models and algorithms

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

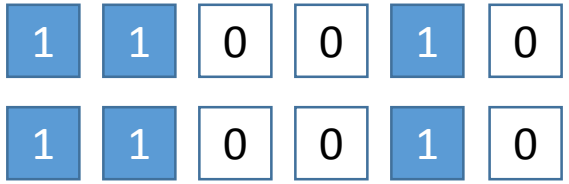
The haplotype phasing problem

- Assume ***M*** **bi-allelic** markers with two alleles coded as 0 and 1.
- A haplotype *h* defined over these markers is a vector of *M* values taken in {0, 1}.
- Genotypes defined over these markers is a vector of *M* values taken in {0, 1, 2}, denoted as *G*.
- A pair of haplotypes $H_i = \{h_{i1}, h_{i2}\}$ is said to be compatible with genotype G_i if for each marker $G_i = h_{i1} + h_{i2}$.
- Given the genotypes of ***N*** **unrelated** individuals, **G** = $(G_1,G_2,...,G_N)$, we want to infer their haplotype pairs **H** = $(H_1,H_2,...,H_N)$

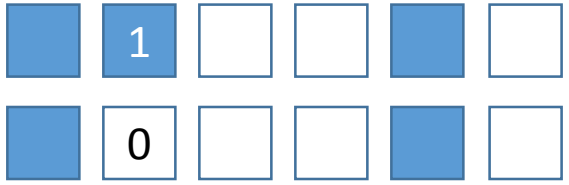
Intuition

Phasing is only needed when ≥ 2 makers are heterozygous, i.e., when ≥ 2 haplotype configurations are possible for an individual's genotypes.

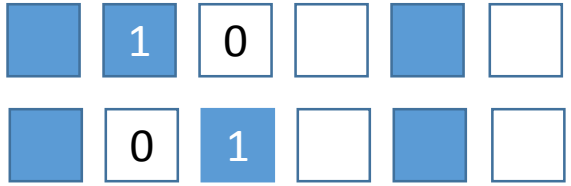
1



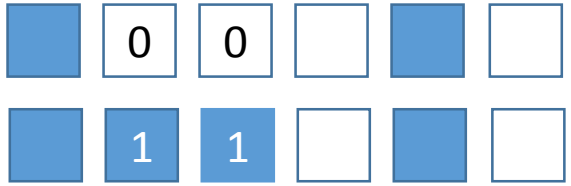
2



3

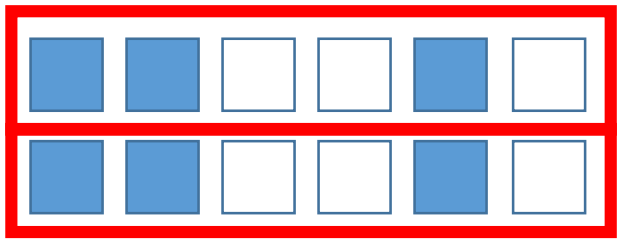


OR

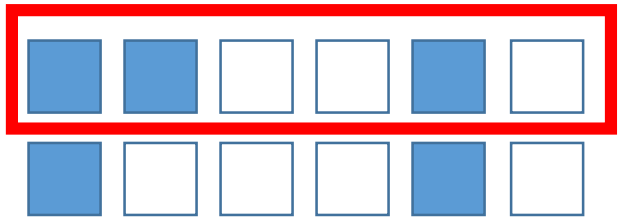


Intuition (cont.)

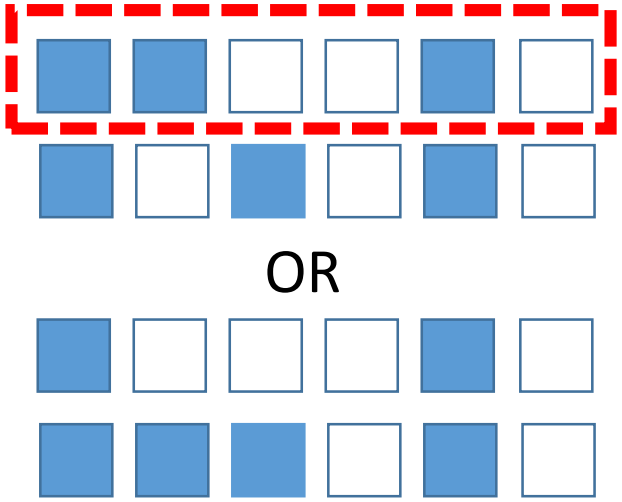
1



2



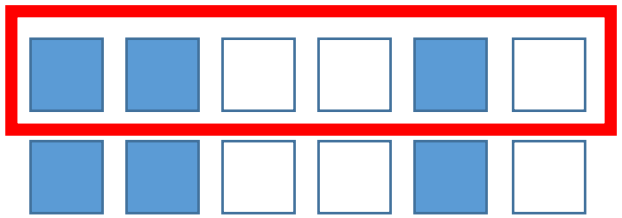
3



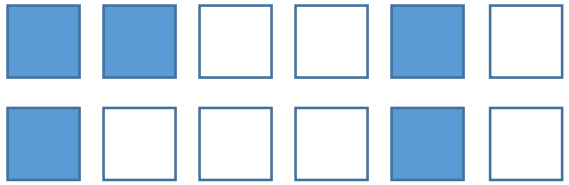
This one seems more likely, as we have seen this haplotype…

Intuition (cont.)

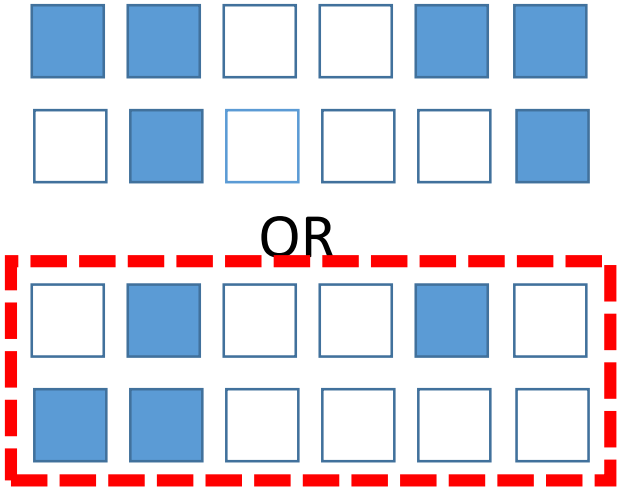
1



2



4



This one seems more likely, as the haplotypes are more similar to existing ones…

Intuition (cont.)

- In human genome, haplotype diversity is limited due to historical population bottleneck as well as relatively low recombination and mutation rates.
- We can infer an individual's haplotypes from other individuals' phased haplotypes at the same region.
  - If ≥ 2 individuals are likely IBD at that region, we can infer their haplotypes at markers where at least one individual is homozygous. (IBD-based phasing)
  - When IBD is less likely, we can still utilize "block-like" LD patterns in human genome (LD-based phasing)

IBD-based vs. LD-based phasing (cont.)

- IBD-based phasing methods mostly use heuristic rules. They are fast and accurate, but only at detectable IBD regions. They do not work well on a small sample of unrelated individuals from a large outbred population.
- LD-based phasing methods estimate the probabilities of possible haplotype configurations via statistical modeling of haplotype frequencies.
- There is a growing interest in combining LD-based statistical models and IBD-based heuristic rules to gain both accuracy and efficiency (e.g., Loh et al., 2016).

MLE on a multinomial model

- Find ***H*** that is consistent with ***G*** and maximizes

$$P(G|H) = \prod_{i=1}^{n} P(G_i|h_{i1}, h_{i2})$$

$$P(G_i|h_{i1}, h_{i2}) = \sum_{\substack{h_{i1},h_{i2}\in\mathcal{H} \\ h_{i1}+h_{i2}=G_i}} p(h_{i1}, h_{i2}) p(G_i|h_{i1}, h_{i2})$$

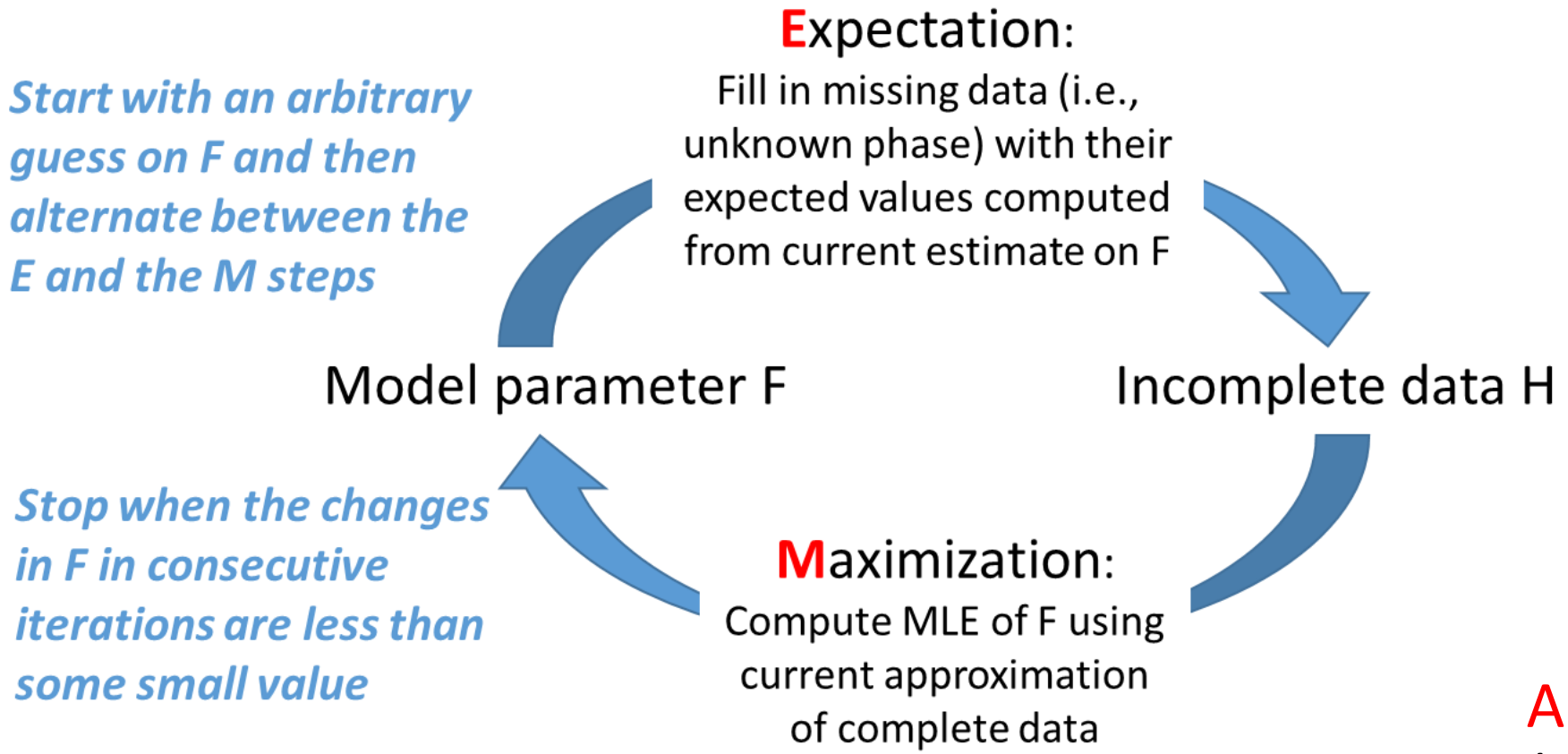
=1 (if no genotyping error)

$$= \begin{cases} F_{h_{i1}} F_{h_{i2}} & if\ h_{i1} = h_{i2} \\ 2F_{h_{i1}} F_{h_{i2}} & if\ h_{i1} \neq h_{i2} \end{cases}$$ (if HWE)

F : population haplotype frequency

$\mathcal{H}$: the set of all possible haplotypes; $|\mathcal{H}| = 2^{L-1}$ for *L* heterozygous SNPs

MLE on a multinomial model: implementation



An EM algorithm (e.g., Excoffier & Slatkin, 1995)

MLE on a multinomial model (cont.)

- EM stores the estimated *F* of every $h \in \mathcal{H}$ in computer memory. When applied to the multinomial model, it requires a memory size that grows exponentially with the number of SNPs.
- How can we bring down the computational cost?
  - Constrain the search space

  → Use a Bayesian approach (Stephens et al., 2001)
  - Improve the way we search for *H*

  → Use a better statistical model (Li & Stephens, 2003)

Bayesian haplotype inference

- Treat *H* as a model parameter too. Estimate its posterior distribution

$$P(H|G) = \frac{P(H,G)}{P(G)} = \frac{P(G|H)P(H)}{P(G)} \propto P(G|H)P(H) = P(H)$$

= 1 when H is consistent with G

Prior distribution of H

- There are two common ways to infer *H* using P(H|G)
  - Choose the *H* that maximizes P(H|G)
  - Sample a set of *H* ~ P(H|G) are calculate their consensus

Bayesian haplotype inference (cont.)

- P(H|G) is usually approximated by Markov chain Monte Carlo (MCMC) sampling, more specifically, a Gibbs sampler (Stephens et al., 2001).
- The sampling procedure starts by randomly generating a pair of haplotypes for each individual, consistent with the observed genotypes (e.g., randomly ordering alleles at each heterozygous site).
- Then these initial estimates are **iteratively refined**, **one individual after another** in each iteration. Each time the new pair of haplotypes for individual i is sampled from $\mathbf{P(H_i|H_{-i},G)}$, where H_{-i} is the most recent haplotype estimates of all other individuals.

Example: Gibbs sampler applied in PHASE

Let $H^{(j)}$ be the approximated haplotype configuration at iteration j. The basic algorithm proceeds as follows:

1. Set t = 0 and randomly generate an initial set $H^{(t)}$ compatible with G.
2. Choose an individual i, uniformly and randomly, from all individuals with ≥ 1 possible haplotype configurations.
3. Sample $H_i^{(t+1)}$ from $\mathbf{P(H_i|G,H_{-i}^{(t)})}$, where H_{-i} is the set of haplotypes excluding those of individual i's.
4. Set $H_j^{(t+1)} = H_j^{(t)}$ for all $j \neq i$.
5. Set t = t + 1 and repeat from Step 2 until chain converges.

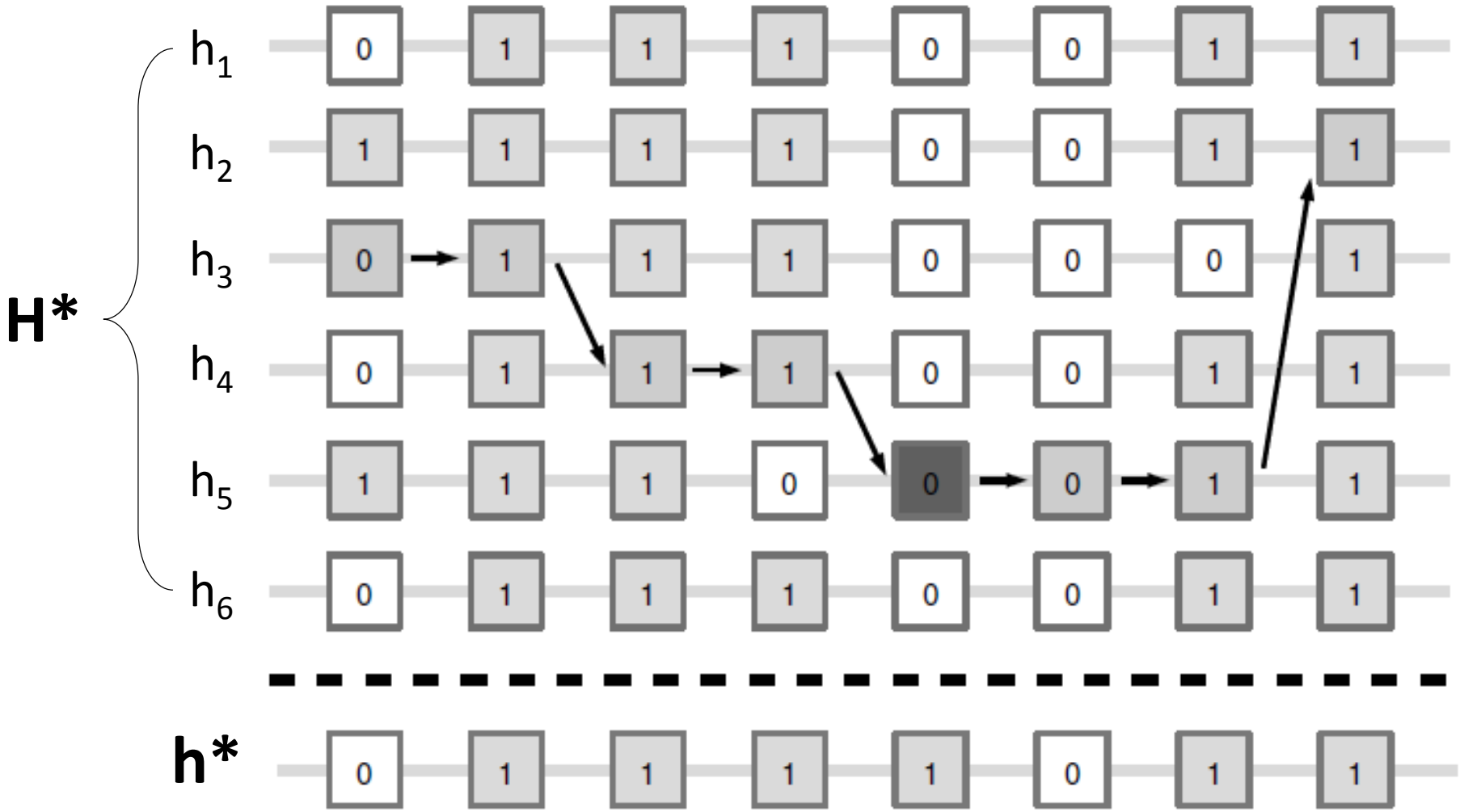
Bayesian haplotype inference: the core

- The problem boils down to fitting a posterior conditional distribution **P(h*|H*)** and using it for haplotype inference.
- h* is the haplotype to be inferred; H* is a set of reference haplotypes.
- Originally H* contains the haplotype estimates of all other individuals in the study sample (Stephens et al., 2001). More recent phasing algorithms modify H* in different ways.
- A major model of P(h*|H*) describes h* as an **imperfect mosaic** of the haplotypes in H* (Li & Stephens, 2003).

The Li-Stephens model of P(h*|H*): idea

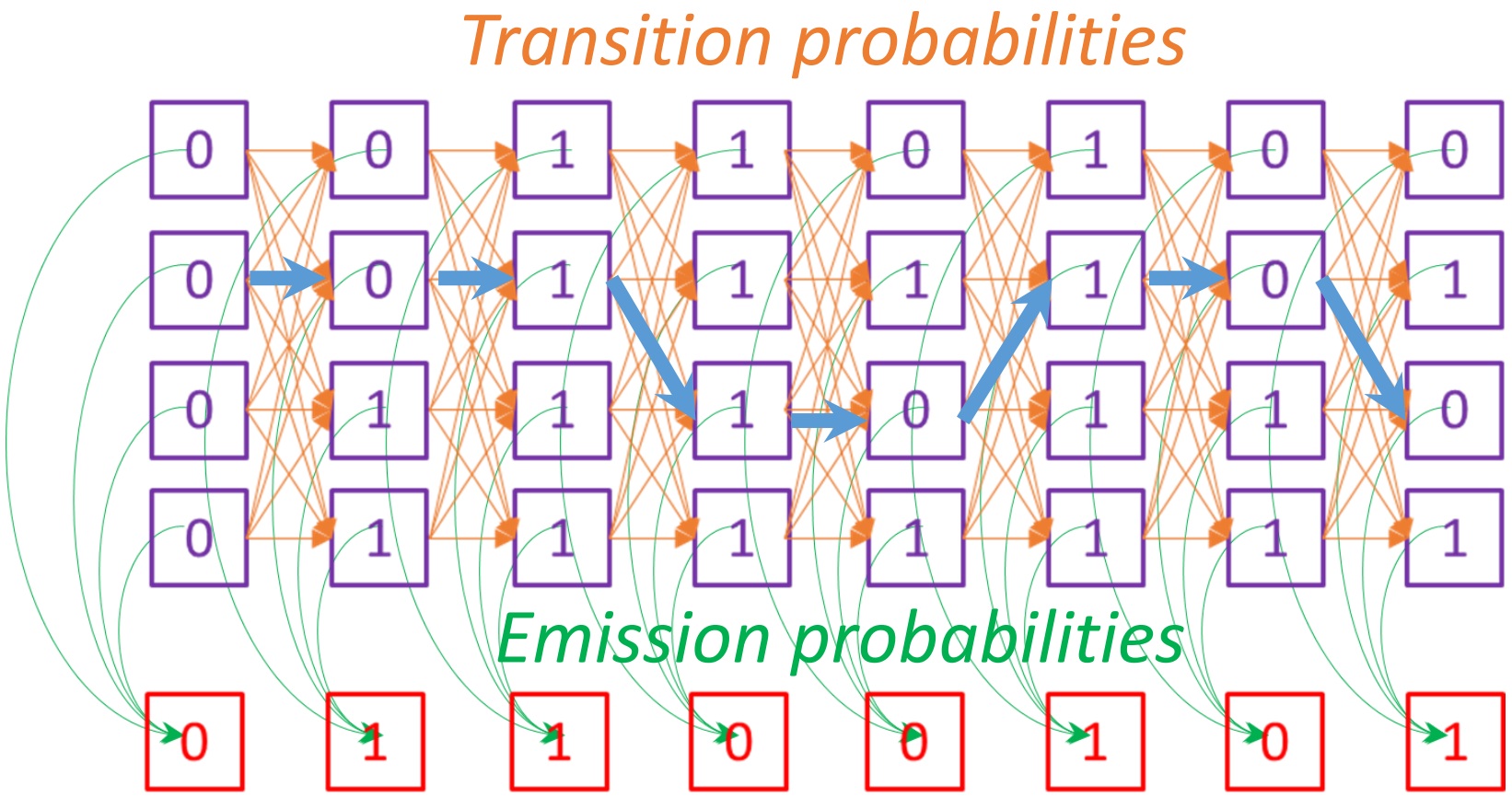
- Every haplotype observed today is an imperfect mosaic of a small number of ancestral haplotypes after many generations of recombination and mutation.
- Due to LD, even unrelated individuals have similar haplotypes over **short genomic regions**.
- The haplotypes of a specific individual are similar, or even identical to, those carried by other individuals in the sample, provided that the sample size is not too small and the mutation rate is not too high.

The model approximates a coalescent process



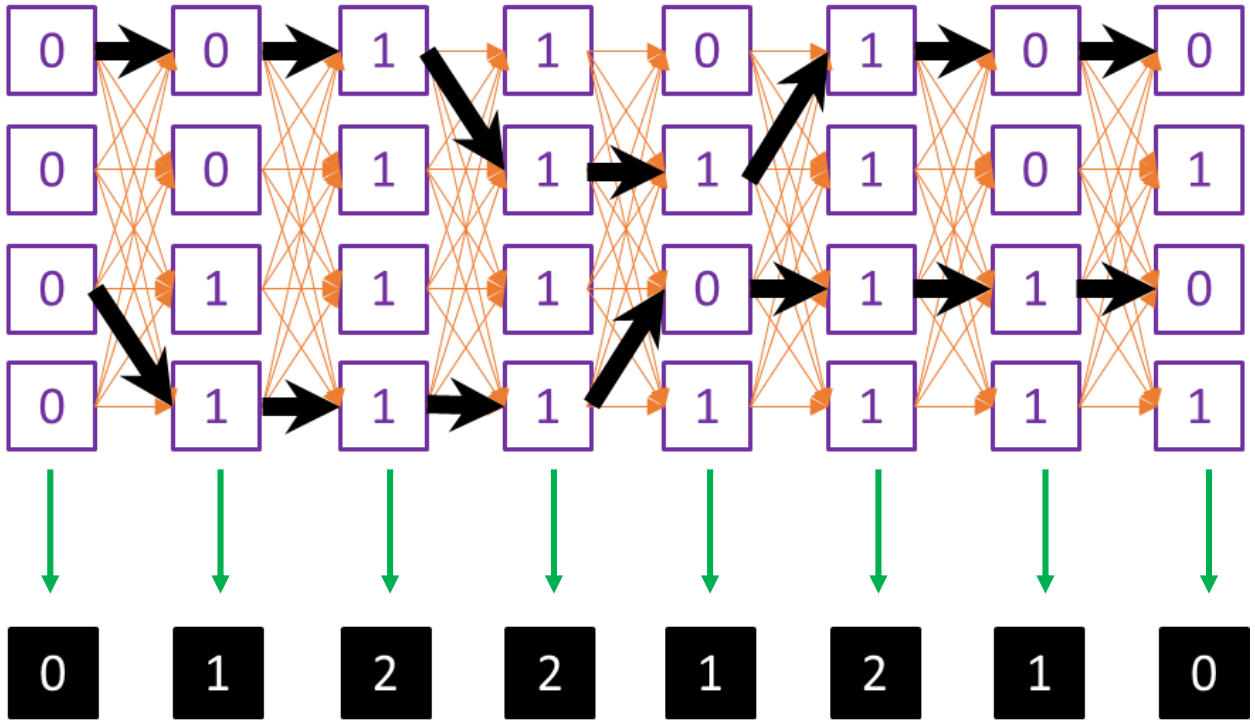
Every path indicates a possible h* sampled from P(h*|H*)

The model is fitted and used as an HMM



The diploid HMM

$$P(H_i|H^*, G_i) = P((h_{i1}, h_{i2})|H^*, G_i)$$
$$= P(h_{i1}|H^*, G_i)P(h_{i2}|H^*, G_i)$$
(assume HWE)



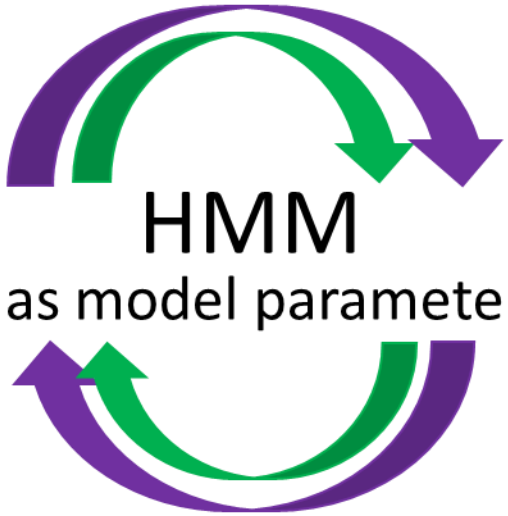
Independently sample two paths to get an ordered haplotype pair consistent with unordered genotype data

While hidden states are still haplotypes, observed values become genotypes {0,1,2}

Phasing with the diploid HMM

Initial haplotype estimates are randomly assigned and consistent with observed genotypes (e.g., randomly ordering alleles at heterozygous sites).

Update the model using current haplotype estimates for each individual



Update each individual's haplotype estimates by sampling from the HMM conditional on the individual's genotypes

After a few "burn-in" iterations for the model to converge to a stationary distribution, several main iterations are conducted. Haplotype estimates from main iterations are used to obtain a pair of consensus haplotypes for each individual.

Model complexity and the implication

- Phasing N individuals over M markers using the above HMM has computational complexity **$O(N \cdot M \cdot |H^*|^2)$**
  - Computing $P(H_i | H^*, G_i)$ requires summing over all possible transitions between hidden states at adjacent markers ($|H^*|^2$ states per marker)
- Given a large dataset (huge N and/or M), how to construct H* is key to model performance
  - Including all study samples (2N haplotypes) into H* is usually unaffordable.
  - If a large reference set (>>N individuals) is available for selecting H*, little accuracy will be sacrificed by not including any sample (Delaneau et al., 2013)

Selected H*

- Select a subset of all haplotypes (external references and/or the latest estimates for the sample) as H*.
  - The size of the subset K (<<2N) is usually predefined.
- MaCH (Li et al., 2010) randomly choses K haplotypes for each individual per iteration
- IMPUTE2 (Howie et al., 2011), HAPI-UR (Williams et al., 2012) and SHAPEIT2 (Delaneau et al., 2013) choose K "closest" haplotypes for each individual per iteration.
  - The "closeness" of two haplotypes is measured by the **Hamming distance** between their corresponding {0,1} vectors

Selected H* (cont.)

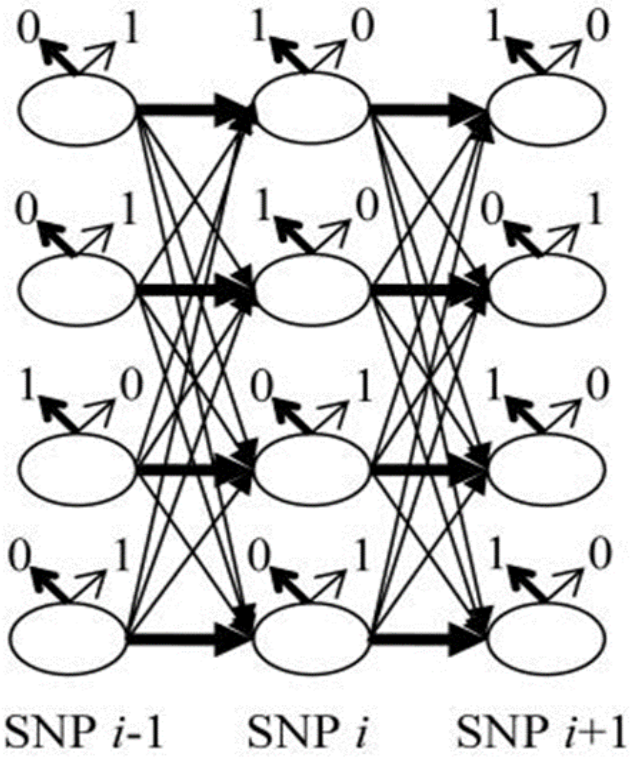
- The second selection approach has complexity $O(N^2)$, which can be a problem when N gets bigger (e.g., > 10,000).
- To improve efficiency, HAPI-UR exploits indexing and hash tables to quickly form HMM states. Its scale is between O(N) and $O(N^2)$.
- SHAPEIT3 (O'Connell et al., 2016) achieves O(NlogN) by
  - Grouping haplotypes into clusters of size M << 2N and select H* only from haplotypes in the same cluster
  - Stoping updating HMM states at segments where haplotype estimates have converged
  - Using a better parallelization scheme

Compressed H*

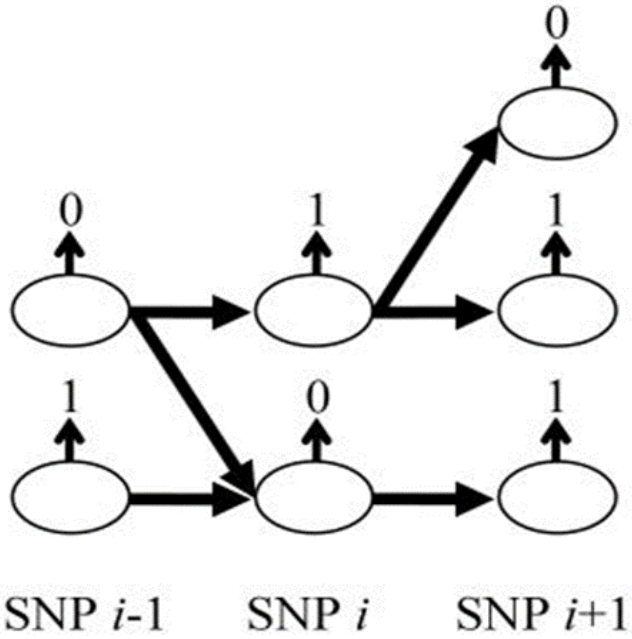
- A alternative strategy is to group locally similar haplotypes and create a compressed H*.
- fastPHASE (Scheet & Stephens, 2006) groups all haplotypes into a few clusters at each marker and use these clusters as HMM hidden states.
- BEAGLE (Browning & Browning, 2007) collapses all haplotypes into a graph structure and carry out HMM calculations on the graph. Locally it amounts to group all haplotypes into a few clusters at each marker.
- In fastPHASE HMM, the number of clusters (states), K, is predefined and held constant at each marker. BEAGLE HMM has variable K, which also makes it more parsimonious than the Li-Stephens model.

Compressed H* (cont.)

Li and Stephens framework



Browning model



- Each hidden state only emits one consistent allele (the other allele has 0 emission probability).
- Mutations are not explicitly modelled and only included when observed.

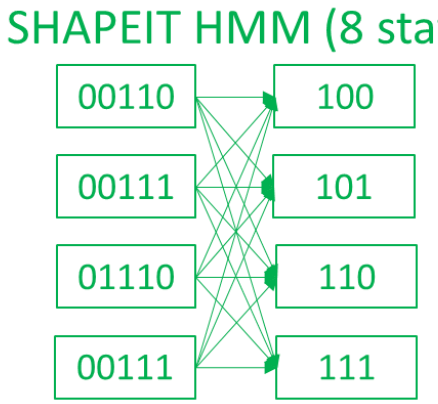
Sparse H*

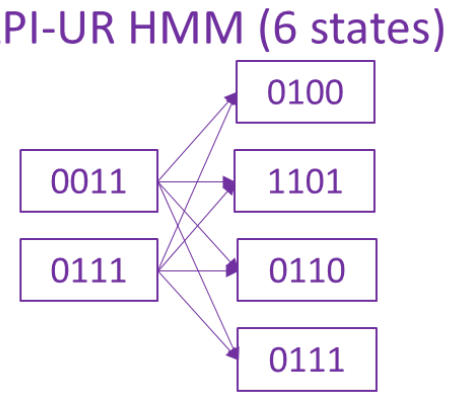
- HAPI-UR and SHAPEIT1&2 only collapse redundant haplotypes.
- They partition the chromosome into segments each with a few distinct haplotypes and use segment-specific instead of marker-specific hidden states.
- HAPI-UR has variable number of states at each segment. SHAPEIT1&2 predefine the number of states but the algorithm progressively prunes unlikely states and merges consecutive segments.
- As a result, their computational complexity is ~O(NM|H*|)

Sparse H* (cont.)

Haplotypes

00110100
00111101
01110110
01111111





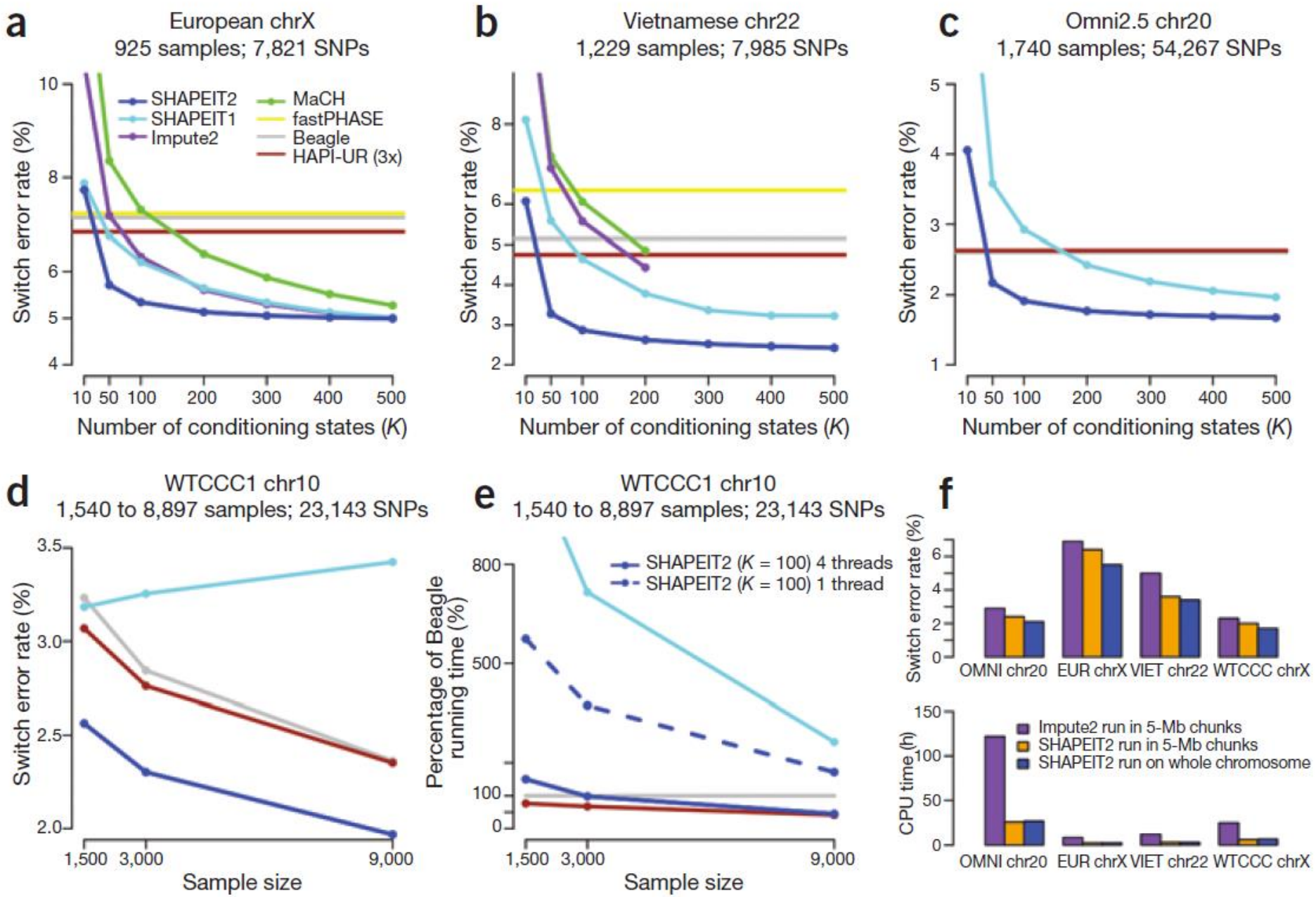
Performance

Switch error rate (Lin et al. 2002; Stephens and Donnelly 2003) is the proportion of successive pairs of heterozygote markers in an individual that are phased incorrectly with respect to each other.

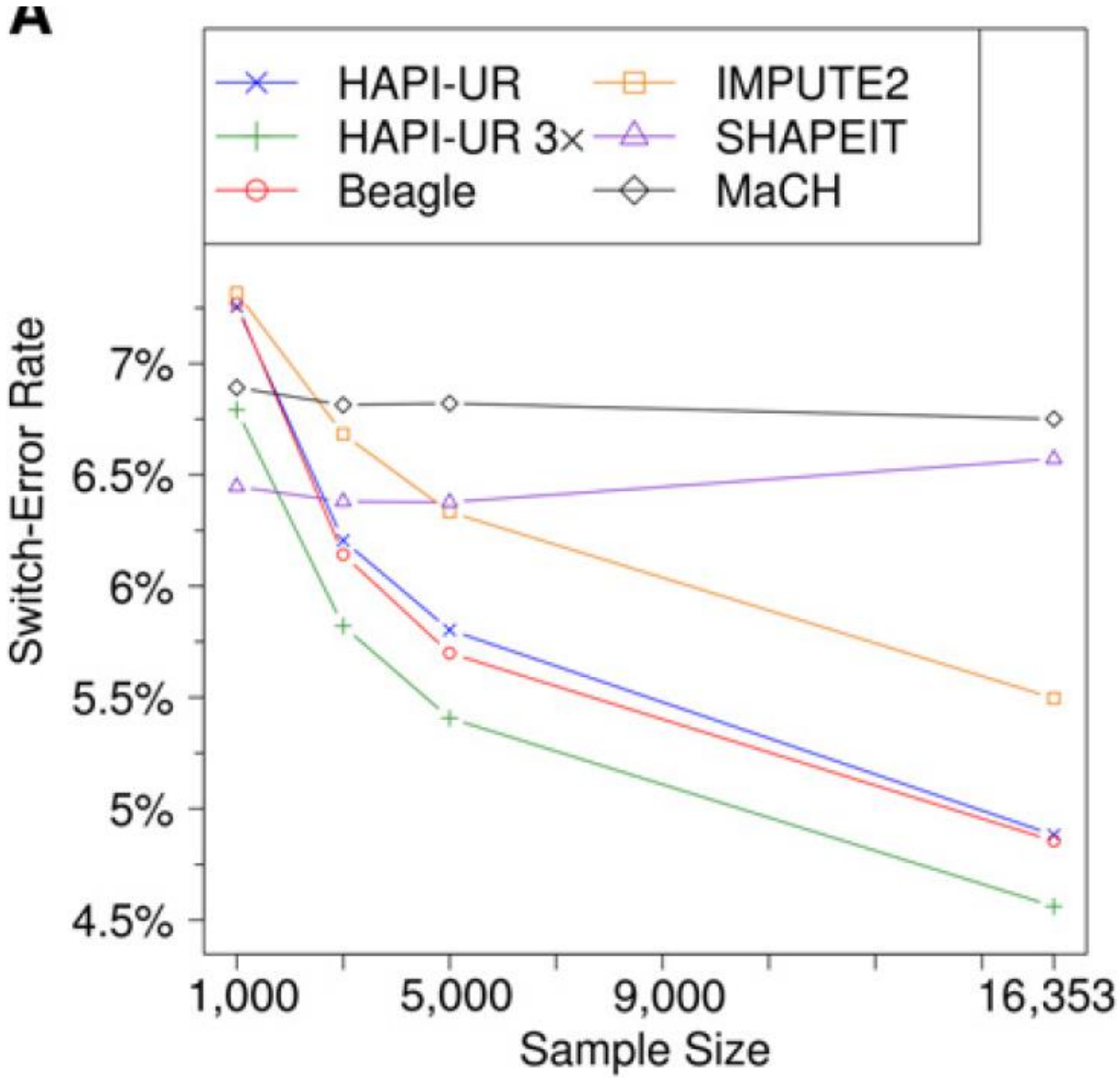
- The individual-specific H* selection strategy generally has higher accuracy, and the accuracy improves as the number and the closeness of haplotypes in H* increase.
- Exhibit 1: Switch error rate decreases with sample size.
- Exhibit 2: Switch error rate decreases with the number of states in the Li-Stephens-type HMM
- Exhibit 3: IMPUTE2 > Mach, SHAPEIT2 > SHAPIEIT3 > SHAPEIT1 regarding accuracy

Performance (cont.)

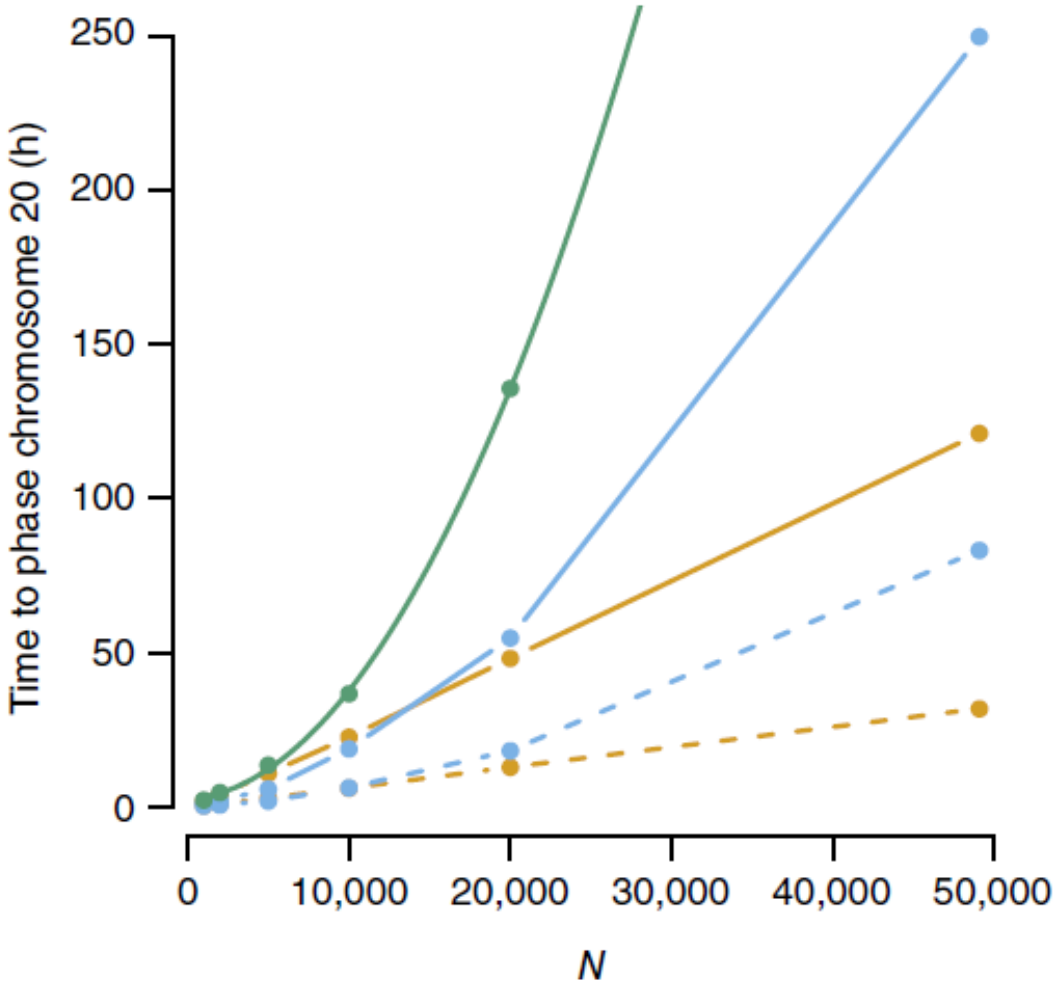
- BEAGLE and HAPI-UR both allow the number of states in their HMMs to vary with local LD structure in the data. There tend to be fewer states in regions with lower haplotype diversity.
- In addition, when they estimate a specific individual's haplotypes, both methods assign 0 probability to estimates that are inconsistent with that individual's genotypes.
- Thus, their HMMs are more parsimonious and take less time to fit, greatly reducing the entire running time. However, the efficiency comes with a cost to accuracy.

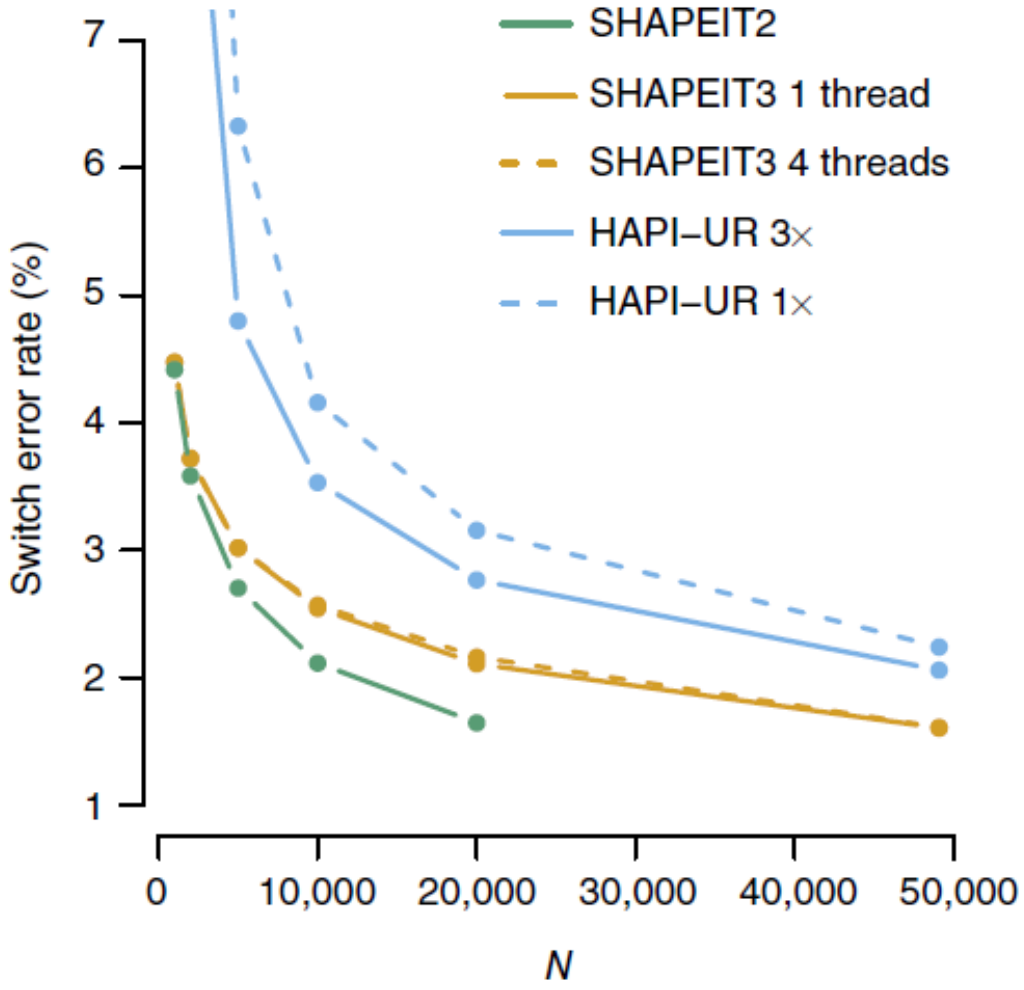


(Delaneau et al., 2013. Figure 1)

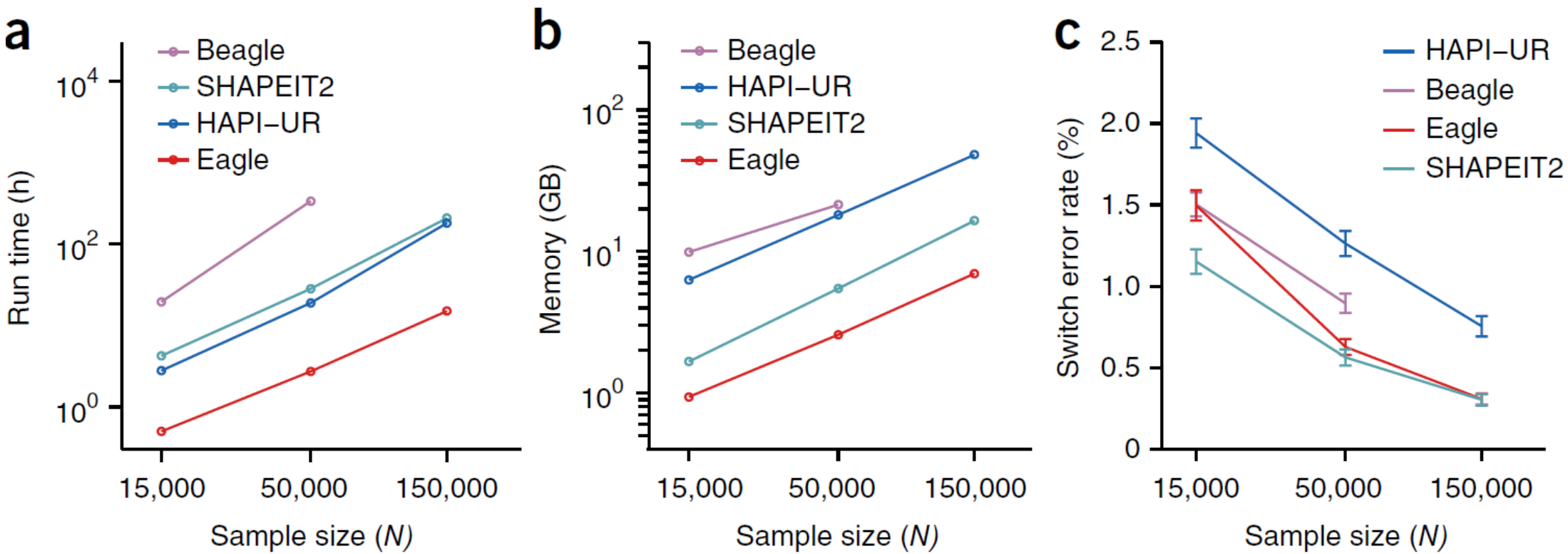


(Williams et al., 2012. Figure 3A)





Top: O'Connell et al., 2016. Figure 1
Left bottom:



(Loh et al., 2016. Figure 2)

Summary

- LD-based phasing methods fit some kind of HMMs using LD structure in the sample/reference data and infer sample haplotypes from the fitted model.
- They are accurate but computationally expensive. The most efficient method has complexity ~O(NlogNM)
- The design of the underlying HMM is the key to finding a balance between accuracy and efficiency.
- LD-base methods may not be suitable for phasing long-range haplotypes that cross recombination hotspots and contain rare variants.
  - The fitted HMM cannot well predict such haplotypes due to the large number of possible haplotype configurations over long regions and the poor representation of rare variants in H*.

References

- Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**, 1084–1097 (2007).
- Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2012).
- Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).
- Howie, Bryan, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics* 1.6 (2011): 457-470.
- Li, Yun, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* 34.8 (2010): 816-834.
- Loh P-R, Palamara PF, and Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* (2016).
- O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLOS Genetics* **10**, e1004234 (2014).
- O'Connell, Jared, et al. Haplotype estimation for biobank-scale data sets. *Nature genetics* (2016).
- Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**, 629–644 (2006).
- Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
- Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* **91**, 238–251 (2012).