Imputation

Shyam Gopalakrishnan

Summer Course 2025

Intended learning outcomes

1

Understand the motivation behind imputation

2

Understand the techniques used for phasing and imputation

3

Understand the use of genotype likelihoods with imputation

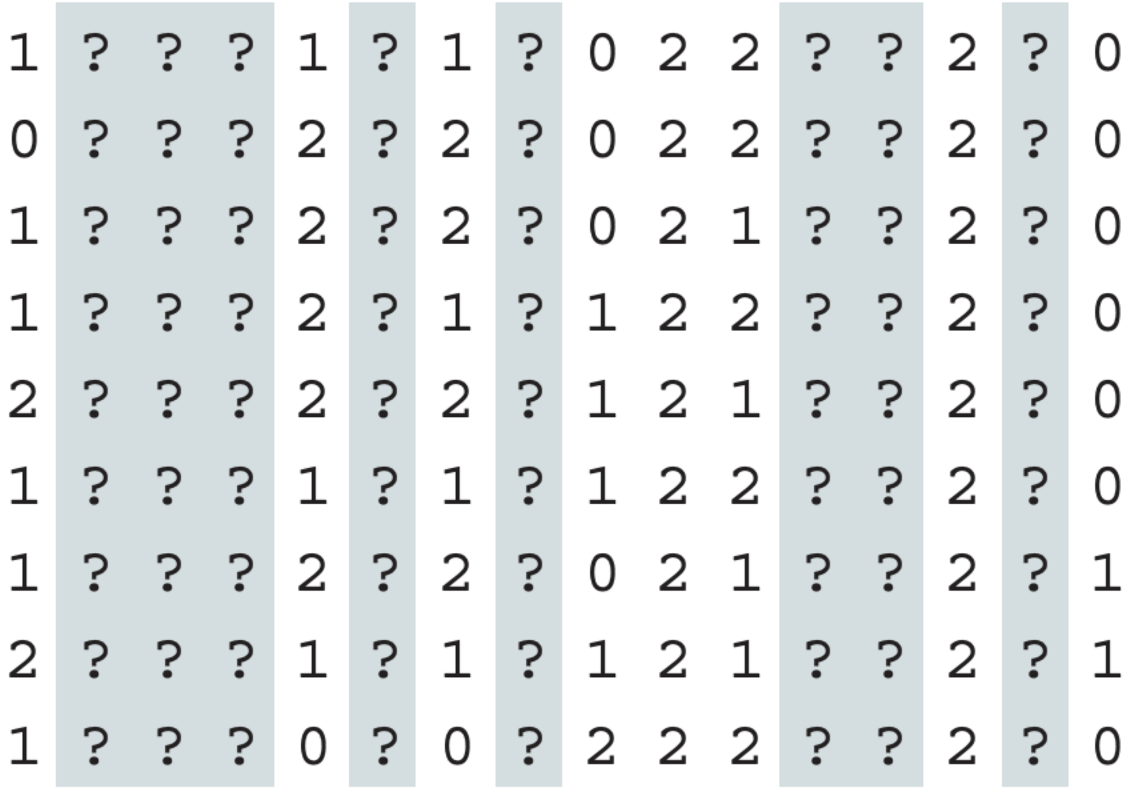
Overview for today

- Problem setting
- Motivation
- Background
- Phasing → Imputation
- Ideas behind imputation
- Exercises

Problem setting

You have sparse data for some samples at some markers – but you want to fill in the gaps.

Genotype data with missing data at untyped SNPs (grey question marks)

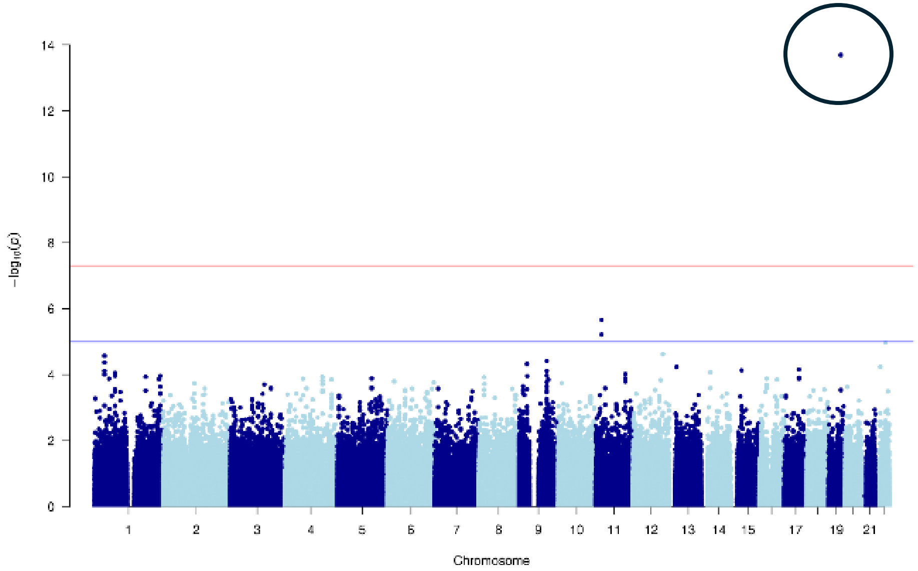


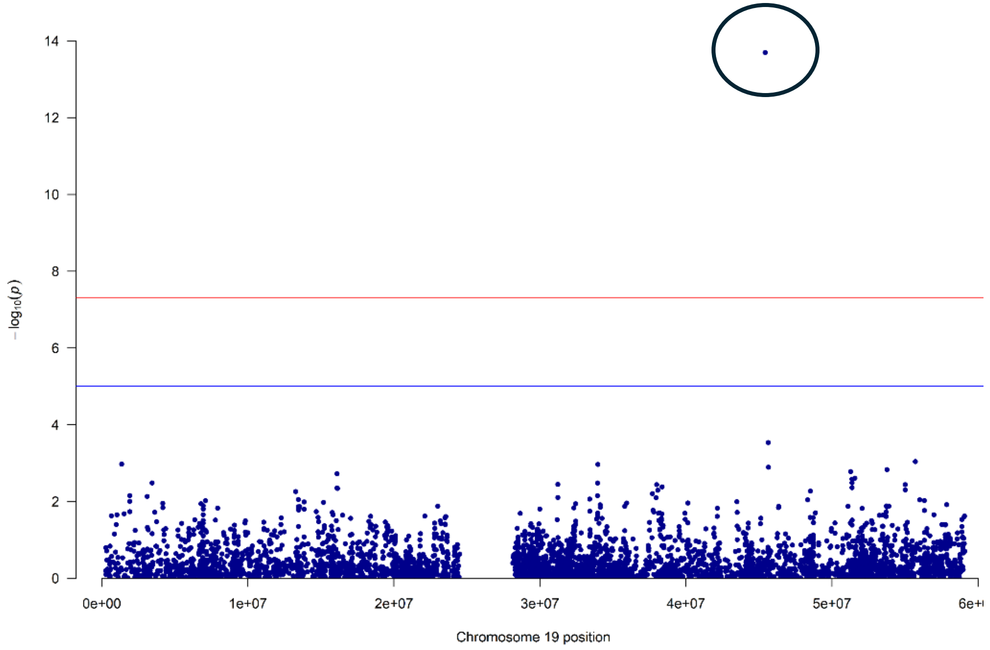
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Why impute?

- Many reasons to imput
  - Meta-analysis – combining results from multiple studies
  - Fine Mapping – better location of GWAS signals
  - Combining data from different chips
- Other less common uses
  - Sporadic missing data imputation
  - Correction of genotyping errors

Example – imputed vs. non-imputed signal



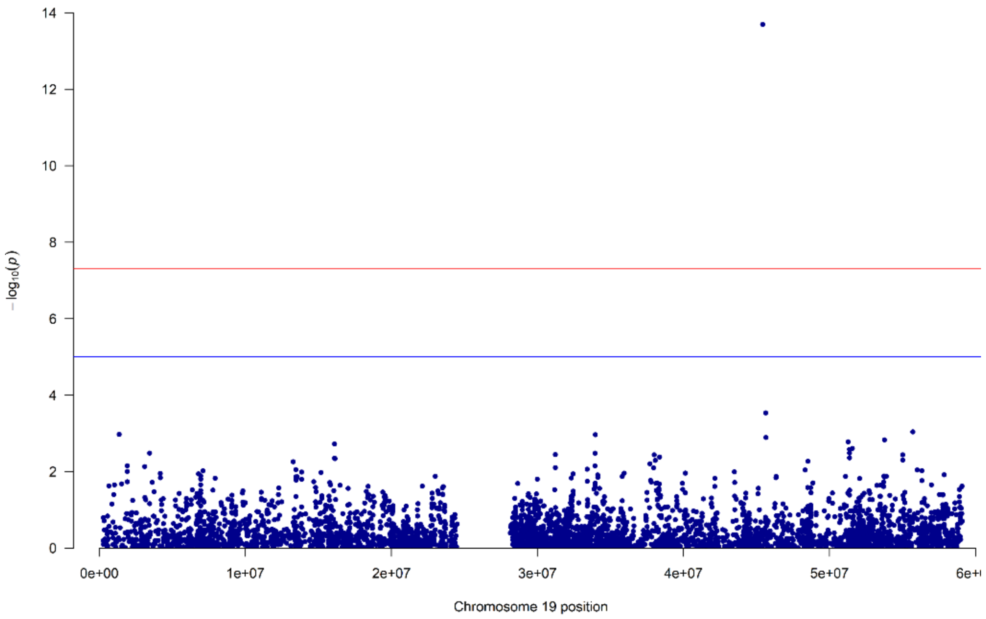


GWAS using only genotyped SNPs

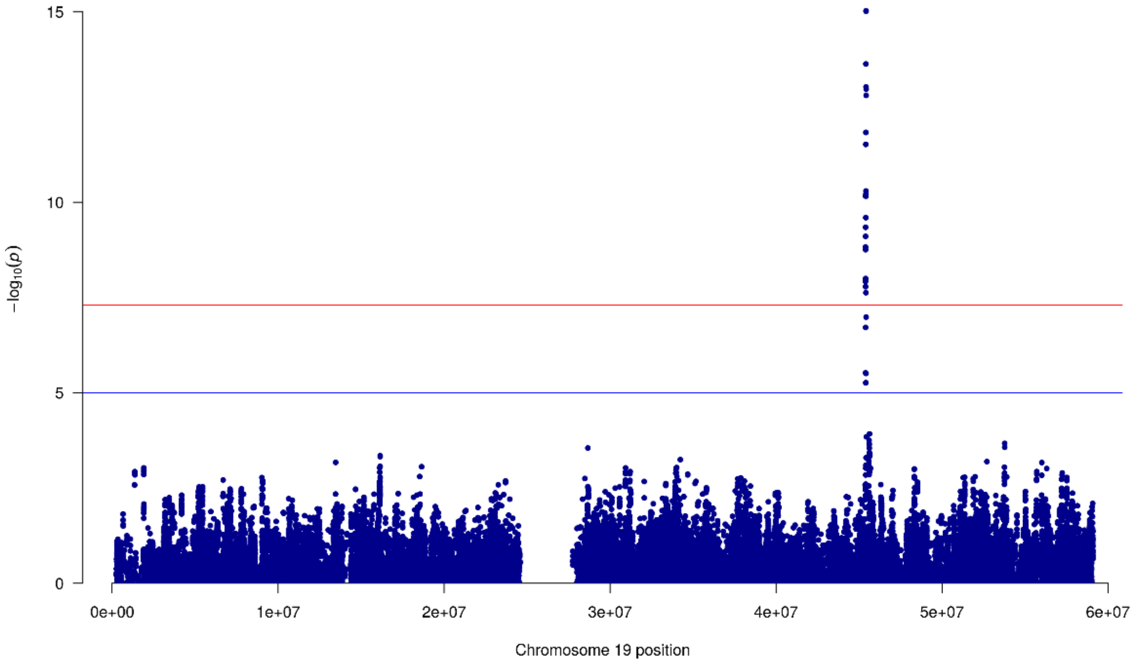
HT: Sarah Medland, QIMR

Example – imputed vs. non-imputed signal

Genotyped only



Post imputation



HT: Sarah Medland, QIMR

Why impute?

- Samples vs. depth tradeoff in sequencing studies
  - Low pass / low coverage sequencing quite common in humans
  - Several advantages to increasing sample size

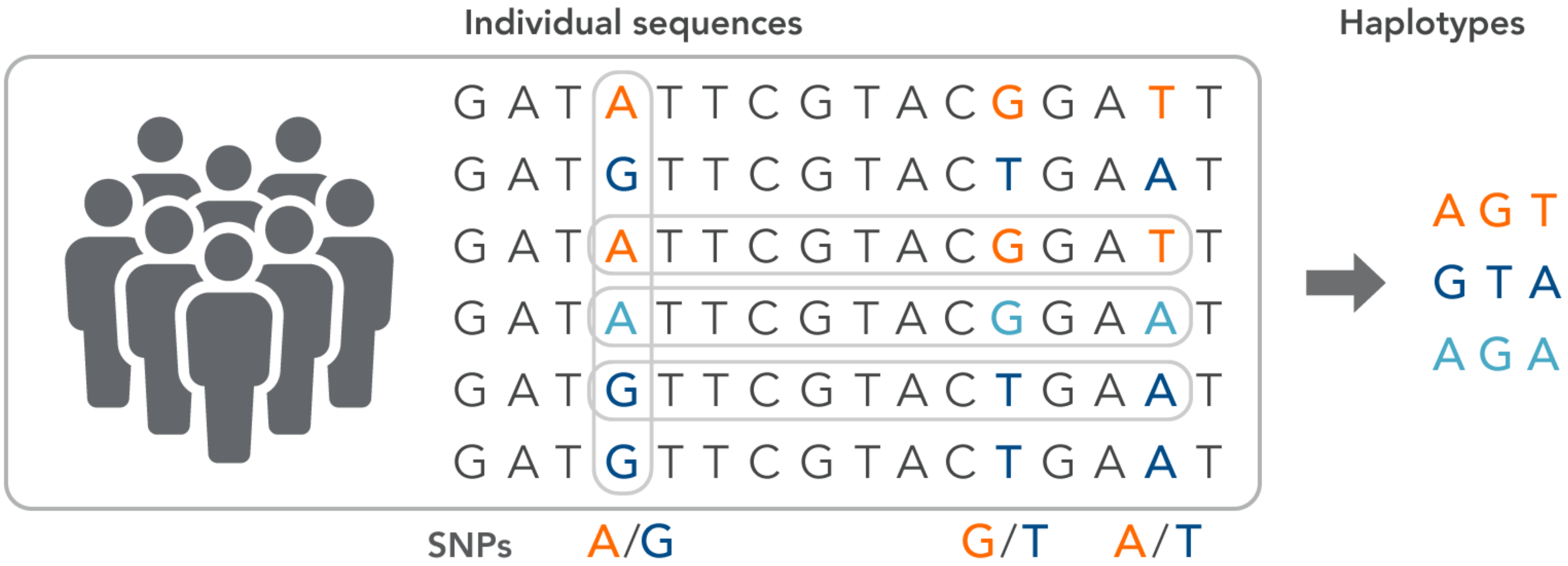
Why impute?

- Samples vs. depth tradeoff in sequencing studies
  - Low pass / low coverage sequencing quite common in humans
  - Several advantages to increasing sample size

- SNP chip data
  - Older studies
  - More focused set of markers – CVD, metabolic disorders etc.

Background

Genotypes vs. haplotypes



Individual sequences

GATATTCGTACGGATT
GATGTTCGTACTGAAT
GATATTCGTACGGATT
GATATTCGTACGGAAT
GATGTTCGTACTGAAT
GATGTTCGTACTGAAT

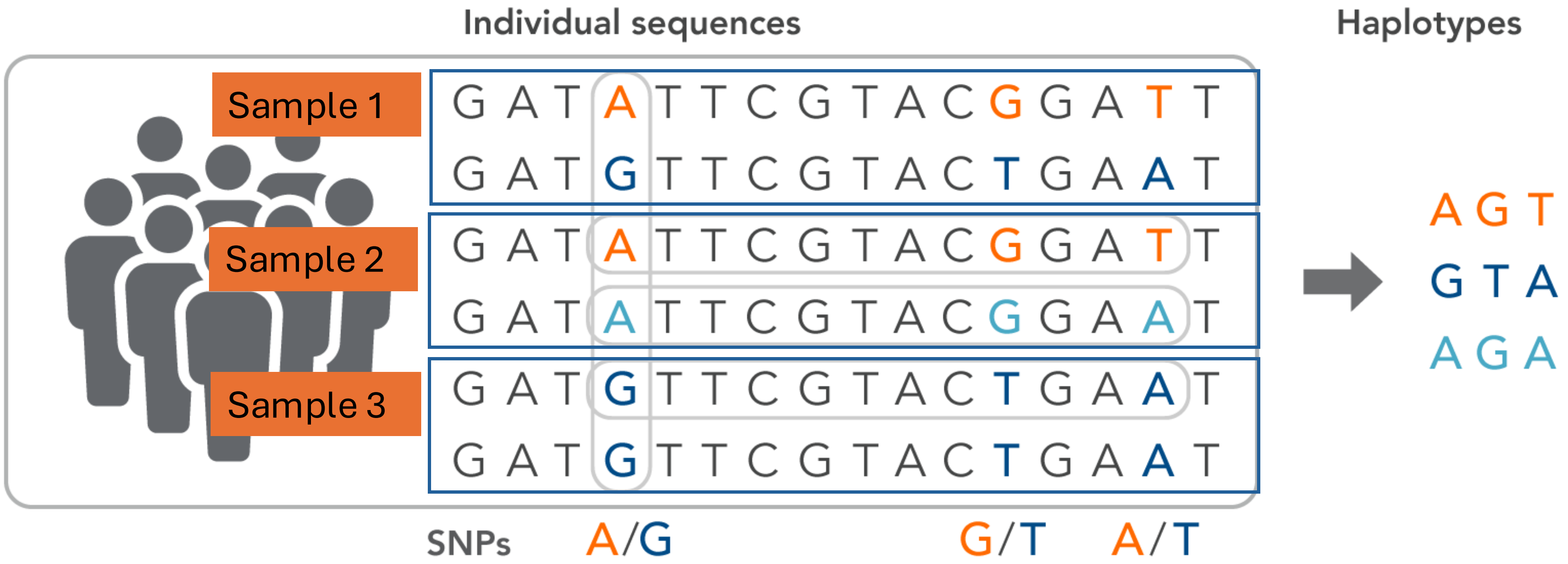
SNPs A/G G/T A/T

Haplotypes

AGT
GTA
AGA

Background

Genotypes vs. haplotypes



Individual sequences

Haplotypes

Sample 1

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

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

Sample 2

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

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

Sample 3

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

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

A G T

G T A

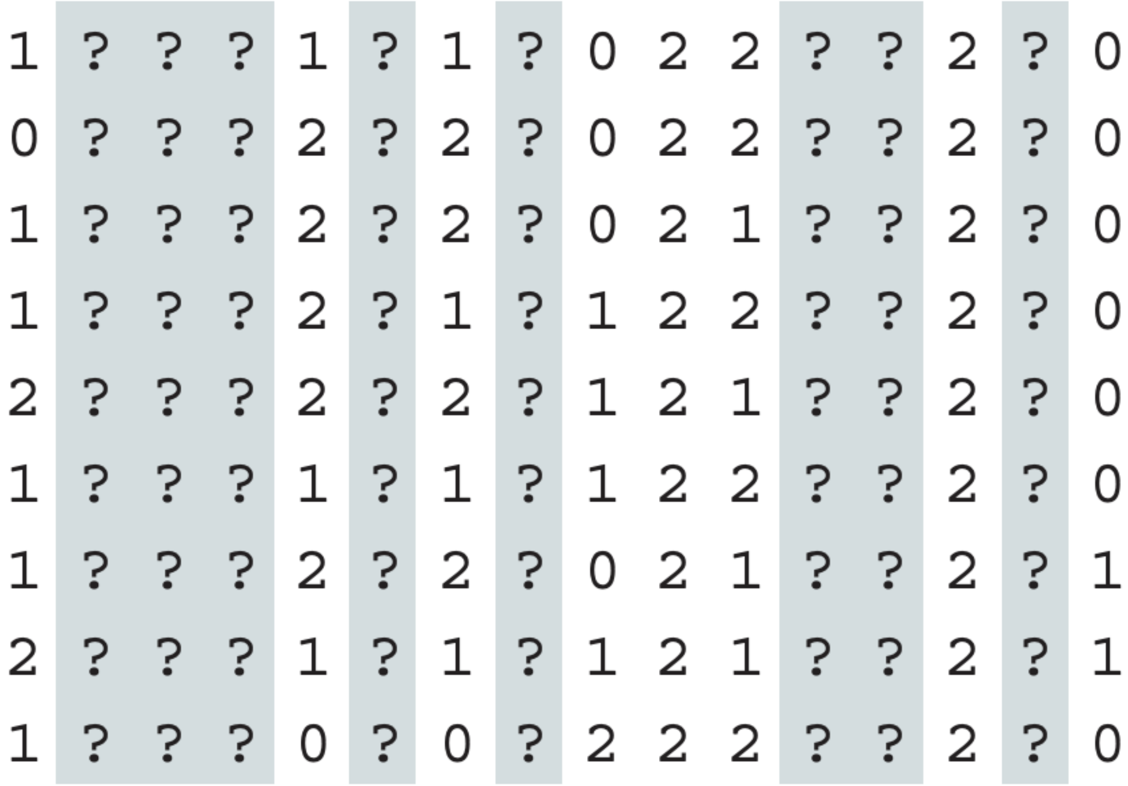
A G A

SNPs A/G G/T A/T

Genotypes			
Sample 1	A/G	G/T	A/T
Sample 2	A/A	G/G	A/T
Sample 3	G/G	T/T	A/A

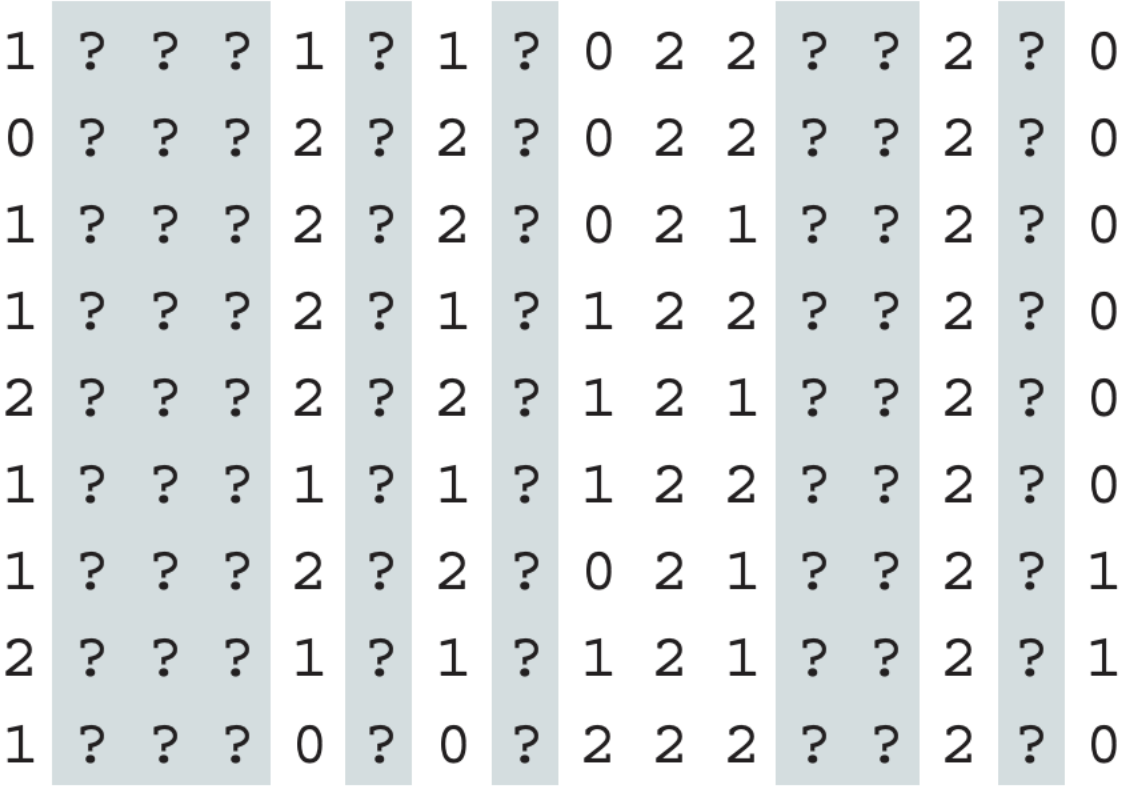
Problem setting revisited

You have sparse data for some samples at some markers – but you want to fill in the gaps.



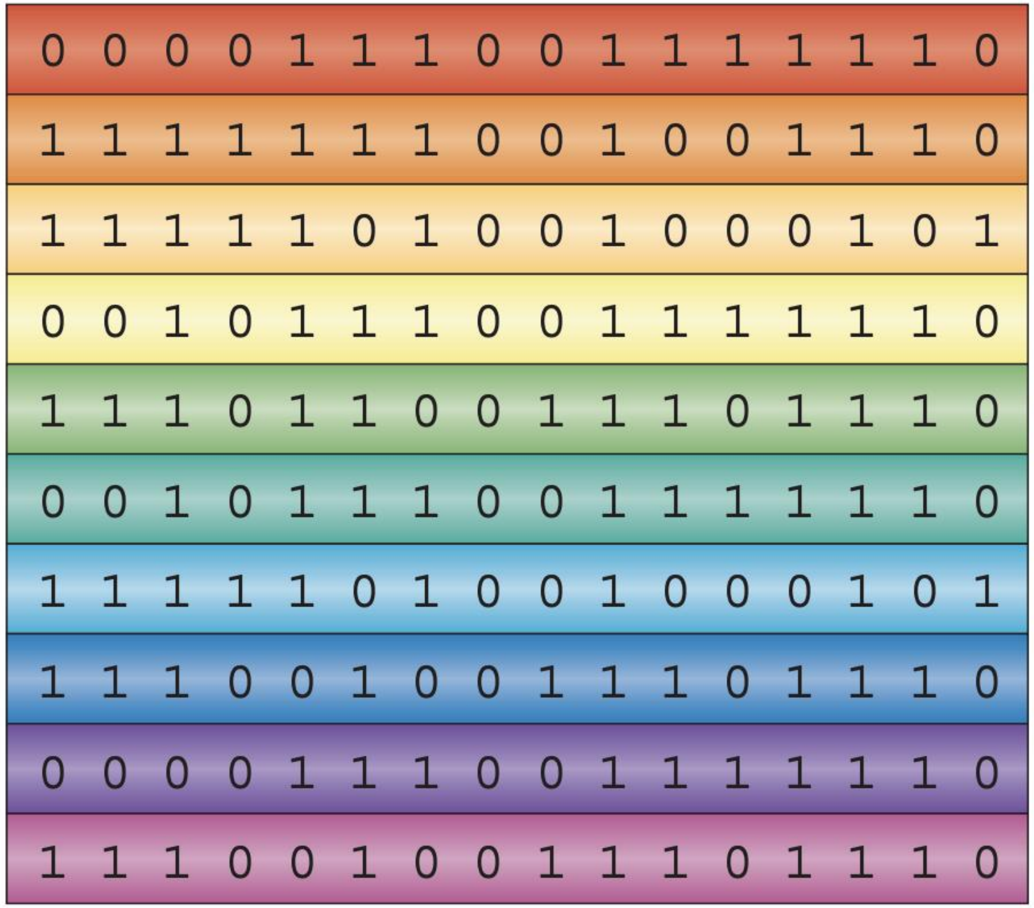
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Problem setting revisited – easy version



1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Reference set of haplotypes, for example, HapMap



0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

Group discussion

Take 10 minutes to discuss how you would impute the missing genoyptes in the sample, given the reference haplotypes.

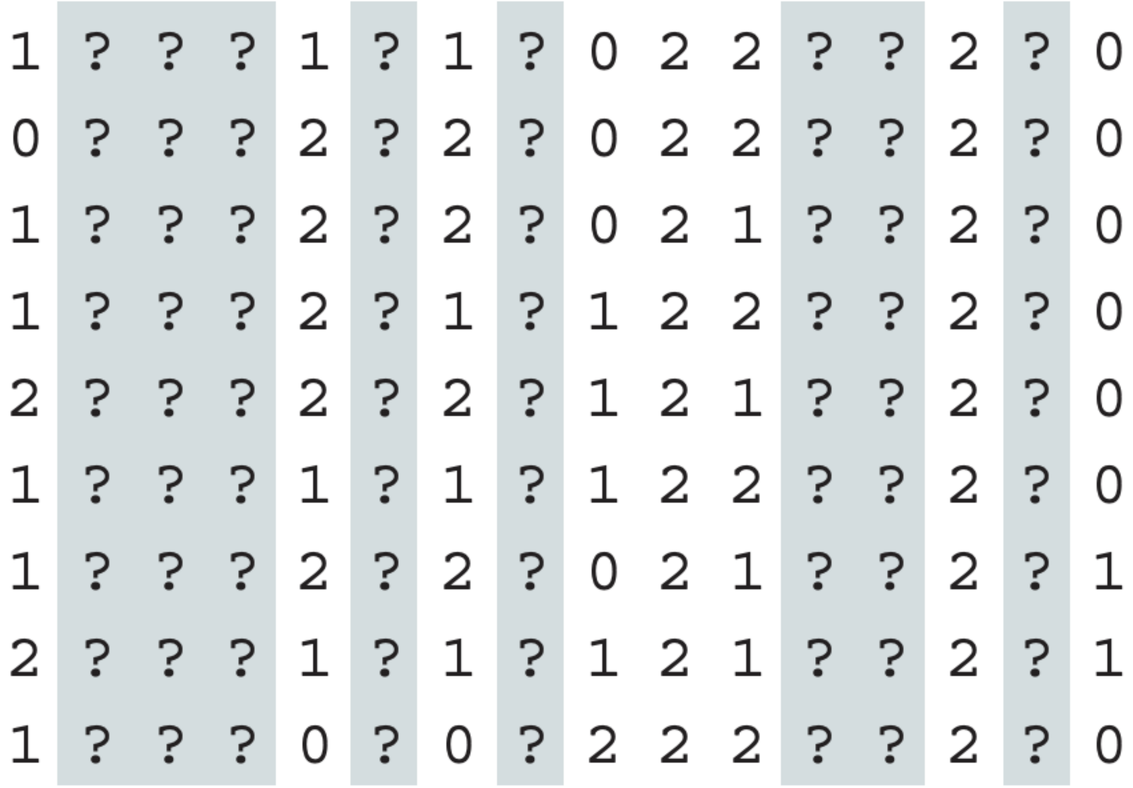
Group discussion

Take 10 minutes to discuss how you would impute the missing genoyptes in the sample, given the reference haplotypes.

- Consider how you would get the missing genotypes?
- Is there intermediate information that would make the problem easier to solve?
  - What information – if you knew about the target samples – would make imputation easier?
- How do things like mutation and recombination play a role?

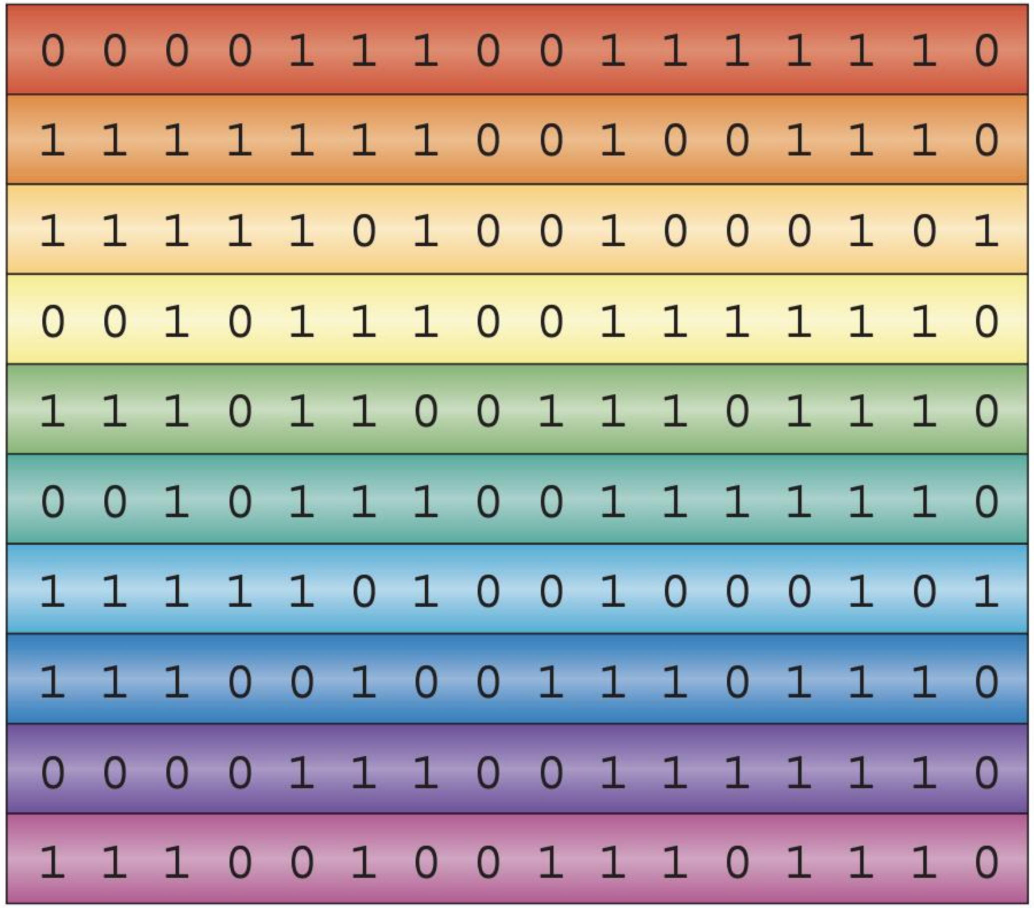
Discuss solutions

Imputation – changing the problem definition



1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

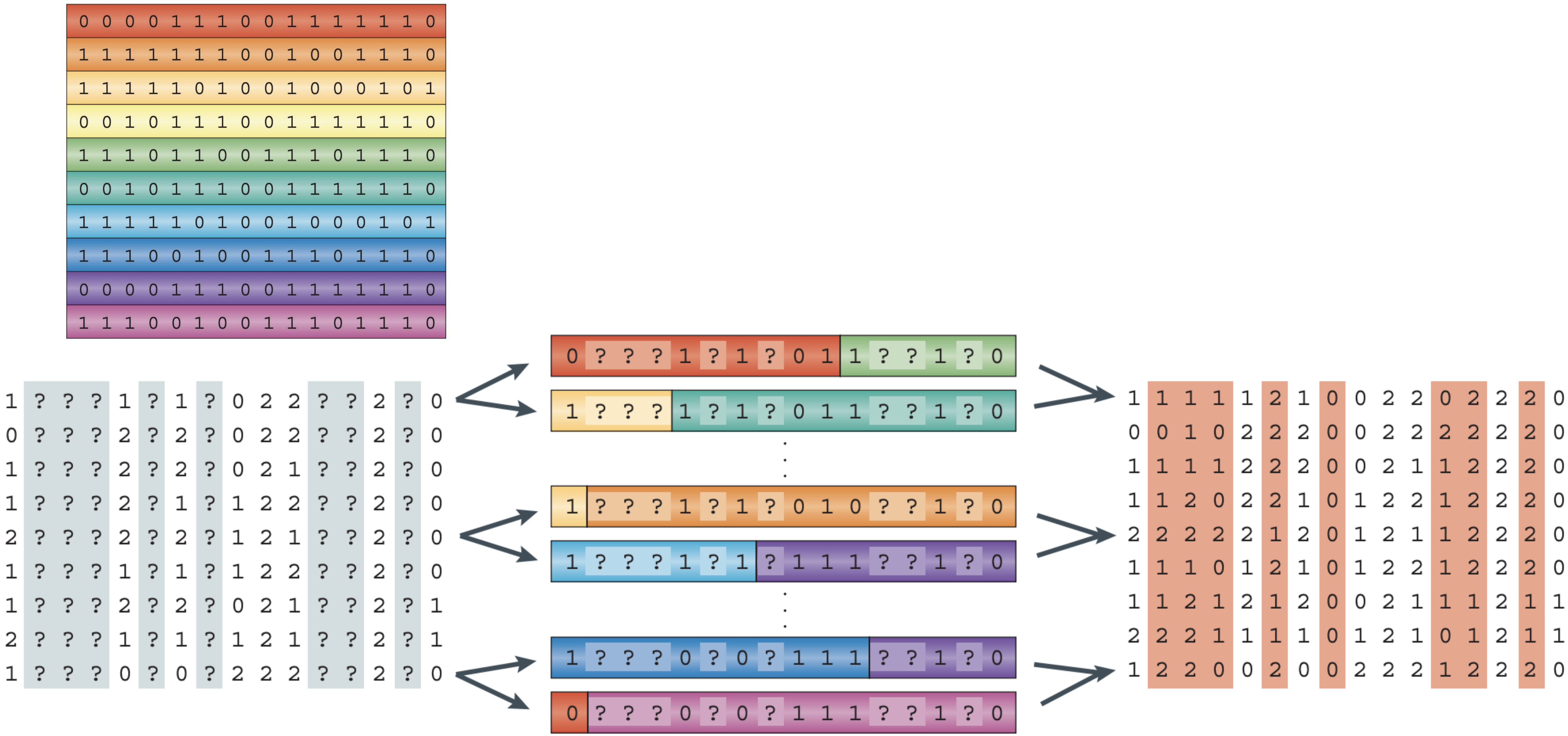
Reference set of haplotypes, for example, HapMap



0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

Imputation – changing the problem definition

Reference set of haplotypes, for example, HapMap



0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

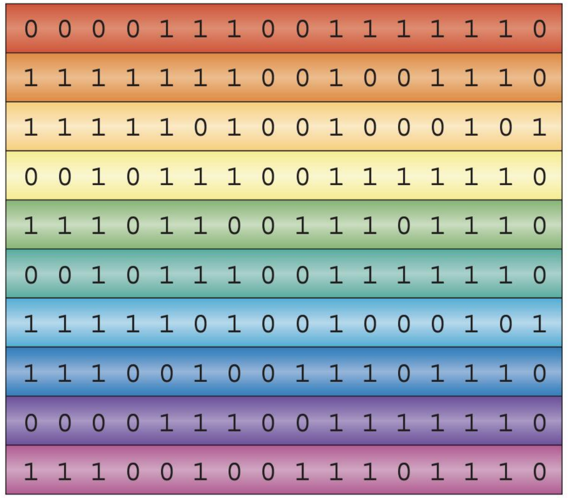
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
⋮															
1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0
1	?	?	?	1	?	1	?	1	1	1	?	?	1	?	0
⋮															
1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0
0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

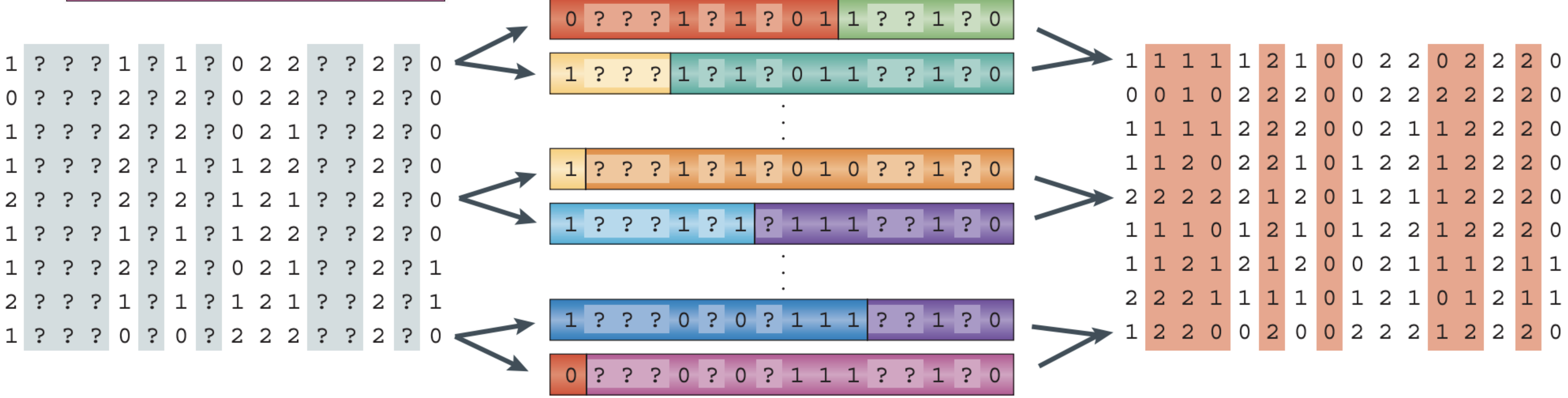
Imputation – changing the problem definition

Reference set of haplotypes, for example, HapMap



0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

Problem now changed from imputation (filling missing genotypes) to phasing (finding haplotypes given genotypes)



1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
⋮															
1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0
1	?	?	?	1	?	1	?	1	1	1	?	?	1	?	0
⋮															
1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0
0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

Phasing – from genotypes to haplotypes

Given a set of reference haplotypes H, we need to sample a mosaic of haplotypes that are consistent with the observed genoytpes G.

Phasing – from genotypes to haplotypes

Given a set of reference haplotypes H, we need to sample a mosaic of haplotypes that are consistent with the observed genoytpes G.

$$
\begin{aligned}
P(G|H) &= \prod_i P(G_i|H) \\
P(G_i|H) &= \sum_{h_1}\sum_{h_2} P(G_i|H=\{h_1,h_2\})P(H=\{h_1,h_2\}))
\end{aligned}
$$

Phasing – from genotypes to haplotypes

Given a set of reference haplotypes H, we need to sample a mosaic of haplotypes that are consistent with the observed genoytpes G.

Extending to sequencing data D – including genotype likelihoods

$$\begin{aligned} P(D|H) &= \prod_i P(D_i|H) \\ P(D_i|H) &= \sum_{h_1}\sum_{h_2} P(D_i|G_i)P(G_i|H=\{h_1,h_2\})P(H=\{h_1,h_2\})) \end{aligned}$$

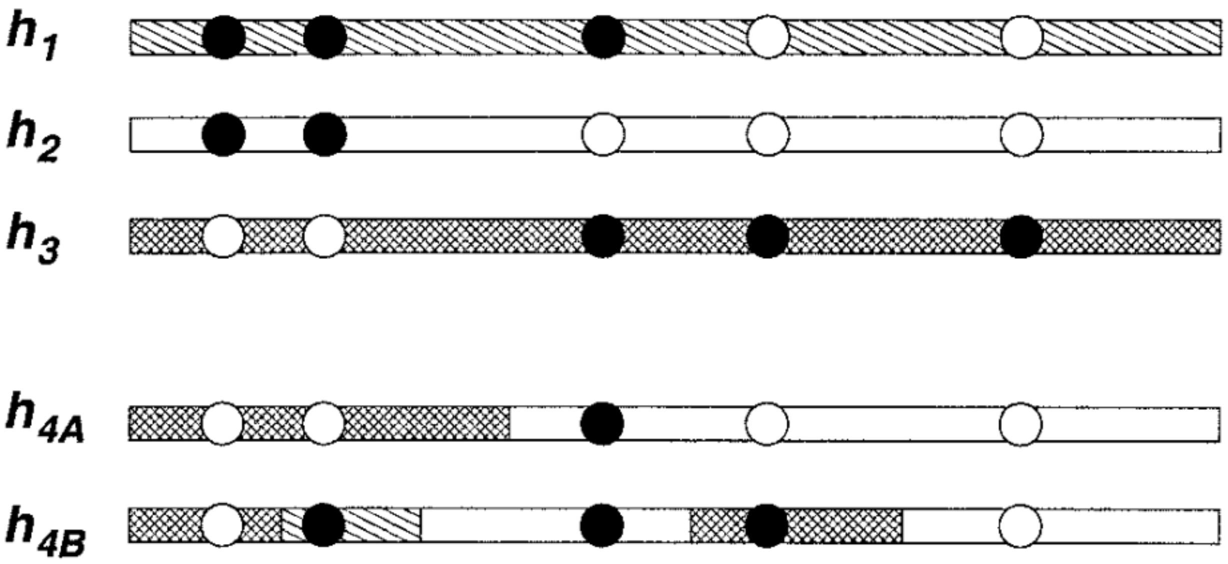
Phasing – from genotypes to haplotypes

Given a set of reference haplotypes H, we need to sample a mosaic of haplotypes that are consistent with the observed genoytpes G.

Extending to sequencing data D – including genotype likelihoods

$$\begin{aligned} P(D|H) &= \prod_i P(D_i|H) \\ P(D_i|H) &= \sum_{h_1} \sum_{h_2} \underset{\substack{\downarrow \\ \text{Genotype} \\ \text{likelihoods}}}{\boxed{P(D_i|G_i)}} P(G_i|H = \{h_1, h_2\}) P(H = \{h_1, h_2\})) \end{aligned}$$

Constructing new haplotypes: Li and Stephens copying model

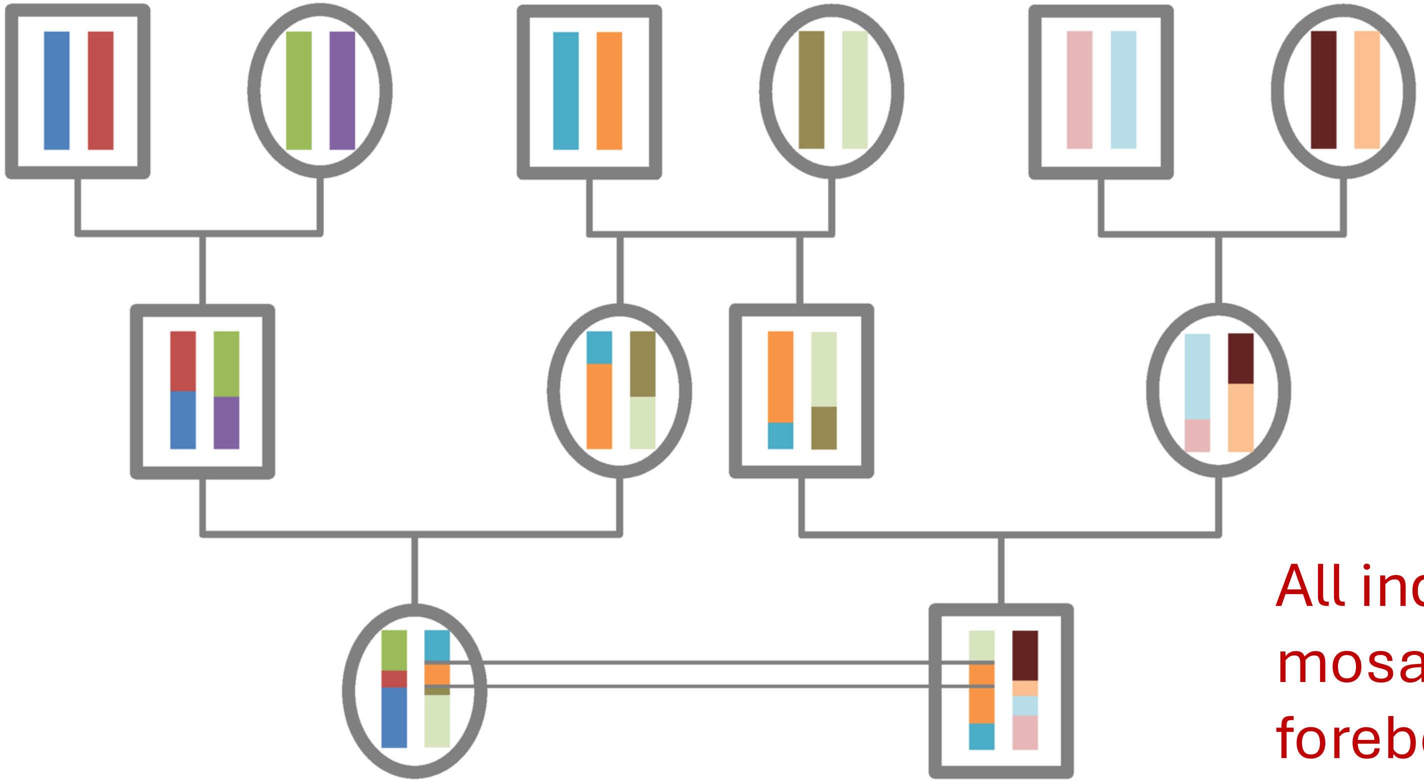


A new haplotype is made by copying from existing ones

- an imperfect mosaic of existing haplotypes

$P(h_k \mid h_1, \ldots, h_n, \Theta)$ is the probability of seeing a new haplotype h_k given haplotypes $h_1, \ldots, h_n$ and parameter Θ for the imperfect copying.

Why would this work? Some intuition



All individuals are made up by a mosaic of haplotypes from all their forebears.

Computational complexity

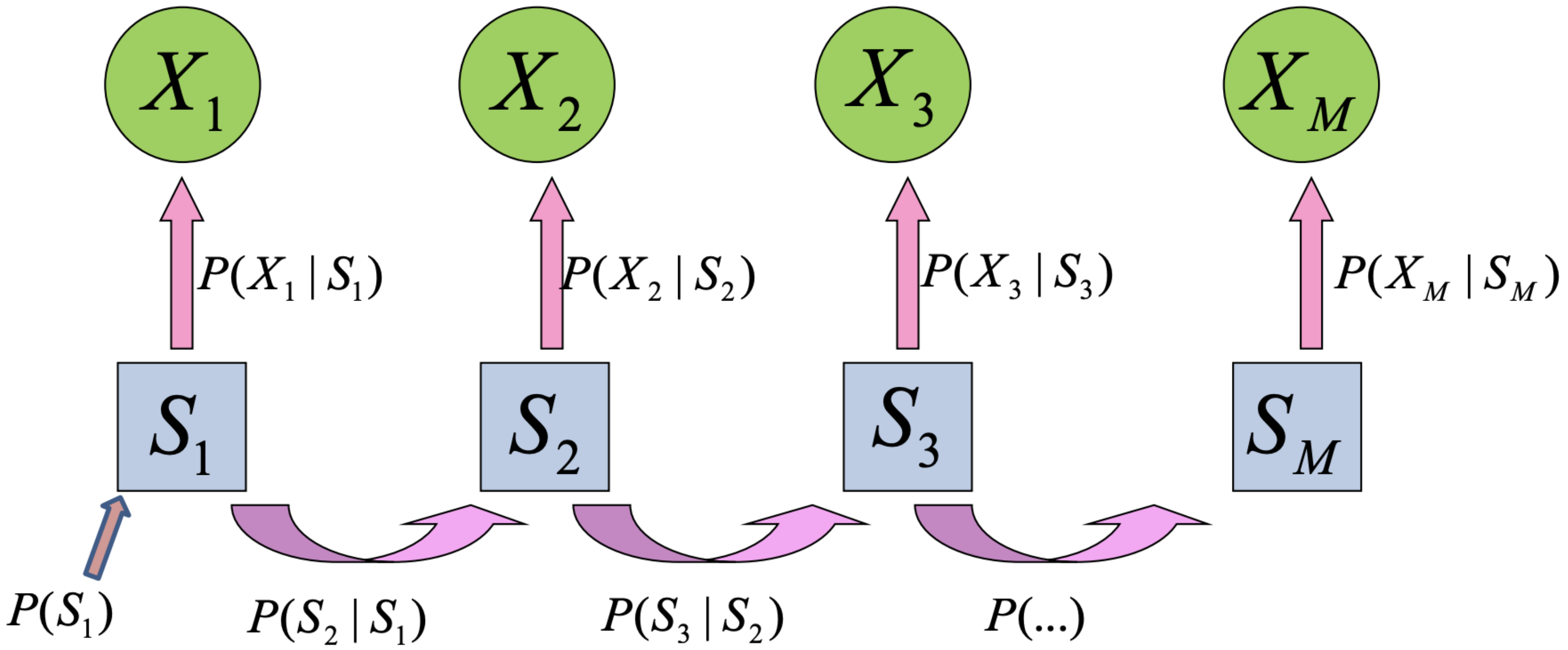
- Number of haplotypes
  - Explodes exponentially with number of markers

- Solutions
  - Markov model for process
  - Work in blocks, assume blocks are independent

Markov model for phasing and imputation

Observed data: genoytpes or reads

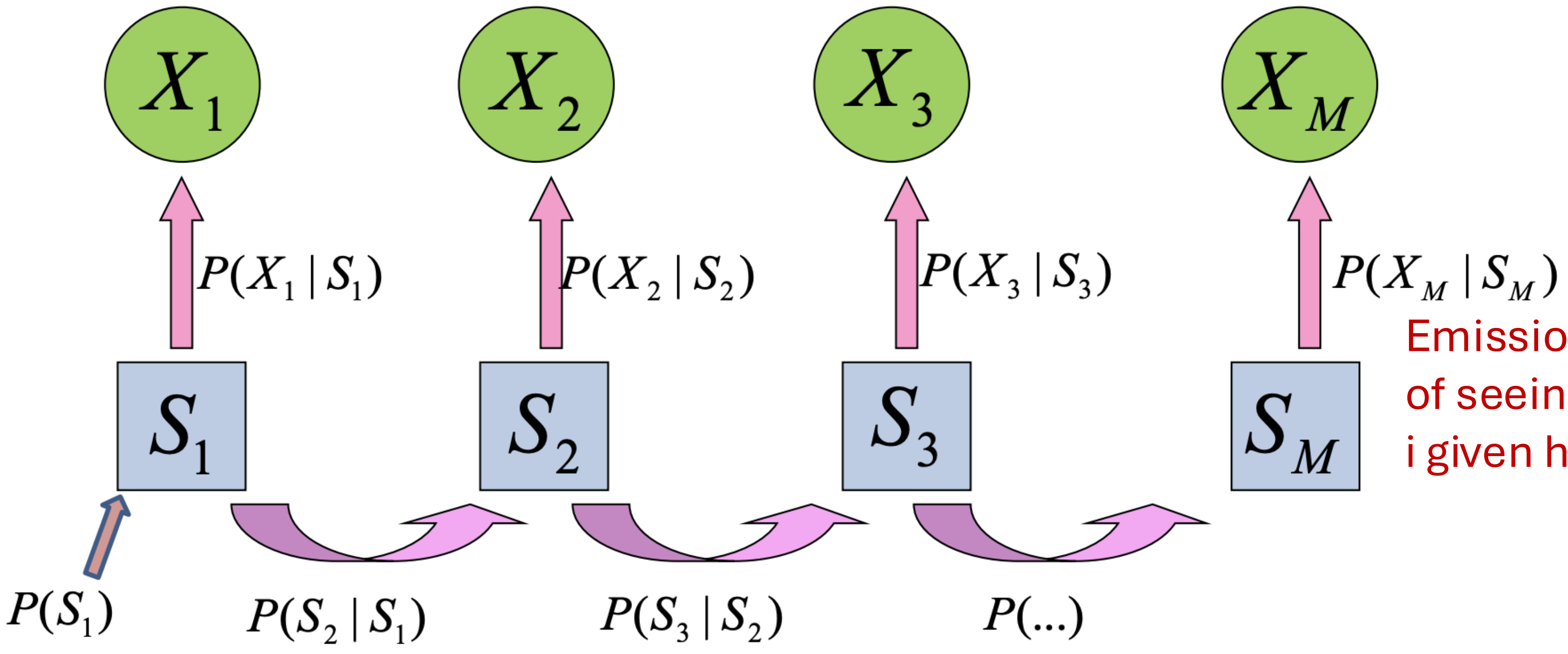
Hidden state: haplotype pair



Markov model for phasing and imputation

Observed data: genoytpes or reads

Hidden state: haplotype pair



Emission probability: Probability of seeing genotype or reads at site i given haplotype pair at site i

Transition probability: Probability of haplotype pair at site i+1 when we know haplotype pair at site i

Markov model details

- Transition probability
  - What would transitions depend on?
- Emission probability
  - How do we go from haplotype pair to genotypes or reads?

No reference panel?

- Can we still impute when we have no reference panel?
  - Yes! Here we also need to figure out the haplotypes on the fly, and the haplotype frequencies.

$$
\begin{aligned}
P(G|f) &= \prod_i P(G_i|f) \\
P(G_i|f) &= \sum_{h_1} \sum_{h_2} P(G_i|H = \{h_1, h_2\}) P(H = \{h_1, h_2\}|f)) \\
P(H = \{h_1, h_2\}|f) &= f_{h_1} f_{h_2}
\end{aligned}
$$

No reference panel?

- Can we still impute when we have no reference panel?
  - Yes! Here we also need to figure out the haplotypes on the fly, and the haplotype frequencies.

$$
\begin{aligned}
P(G|f) &= \prod_i P(G_i|f) \\
P(G_i|f) &= \sum_{h_1} \sum_{h_2} P(G_i|H = \{h_1, h_2\}) P(H = \{h_1, h_2\}|f)) \\
P(H = \{h_1, h_2\}|f) &= f_{h_1} f_{h_2}
\end{aligned}
$$

$$\hat{f} = argmax_f P(G|f)$$

Computational shortcuts

- Number of haplotypes – super high
    - Keep track of only haplotypes with “high” frequency.
    - Limit number of haplotypes being tracked.
    - Work on haplotype blocks at a time.

Phasing and Imputation

To get the phase probability

$$p(h_1, h_2 | G_i, \hat{f}\,) = \frac{p(G_i|h_1,h_2)p(h_1,h_2|\hat{f}\,)}{p(G_i|\hat{f}\,)}$$

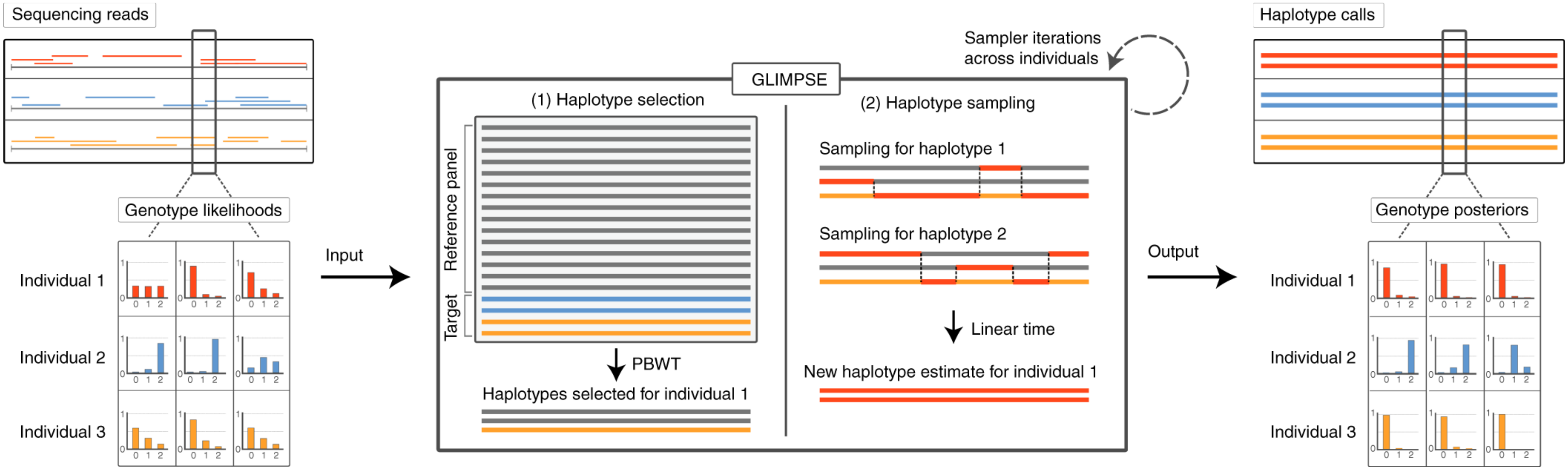
For imputation

$$p(G_{ij}|\hat{f}\,, G_{i,-j}) = \sum_{h_1}\sum_{h_2} p(G_{ij}|h_1, h_2)p(h_1, h_2|\hat{f}\,, G_{i-j})$$

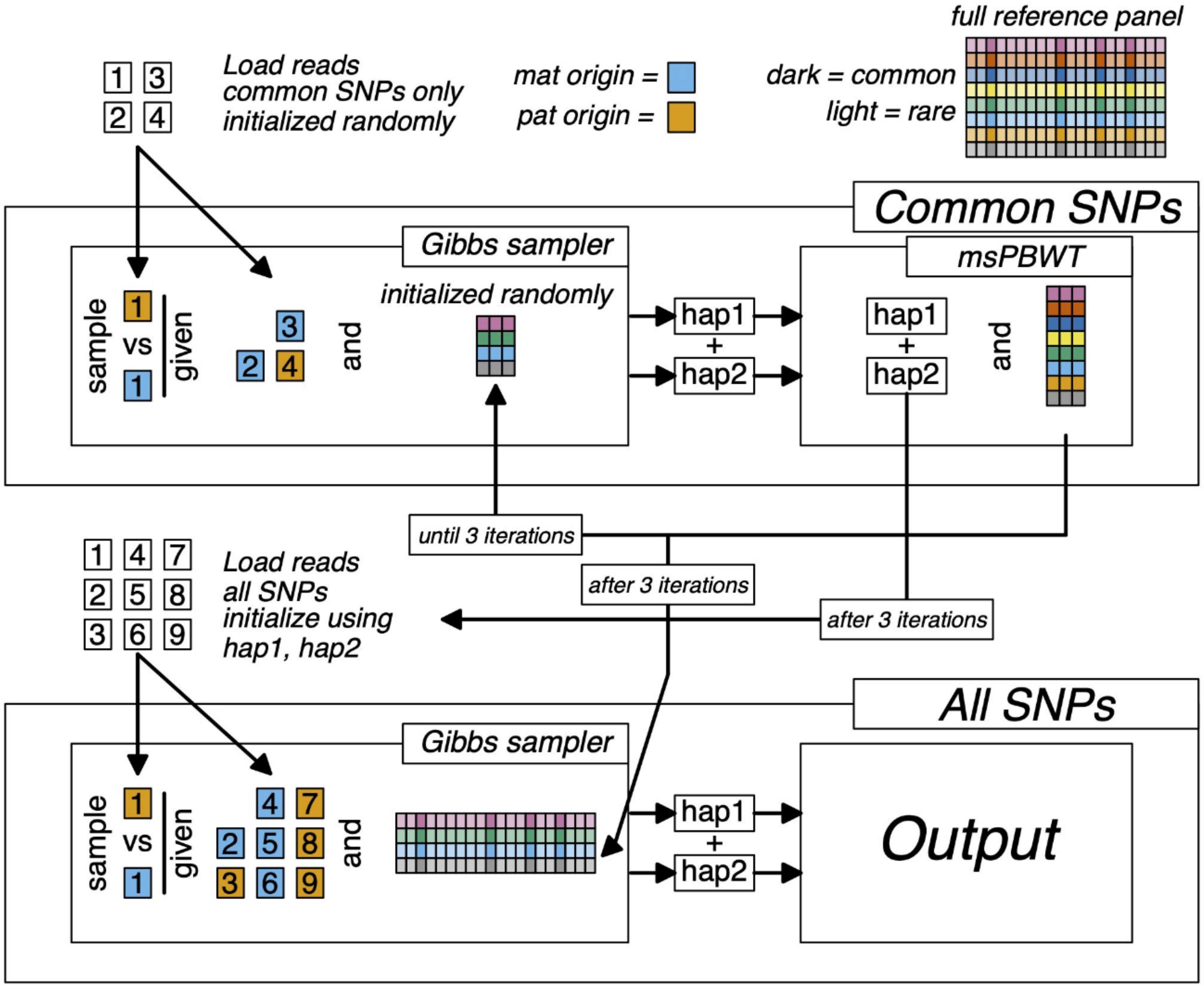
Imputation approaches

- Genotype calls + reference panels
  - Beagle 5
  - Eagle
  - Shapeit2 + impute2
- Low coverage sequencing + reference panel
  - QUILT2
  - GLIMPSE2
- Low coverage sequencing + no reference panel
  - Beagle 3 or 4
  - impute2

GLIMPSE 2

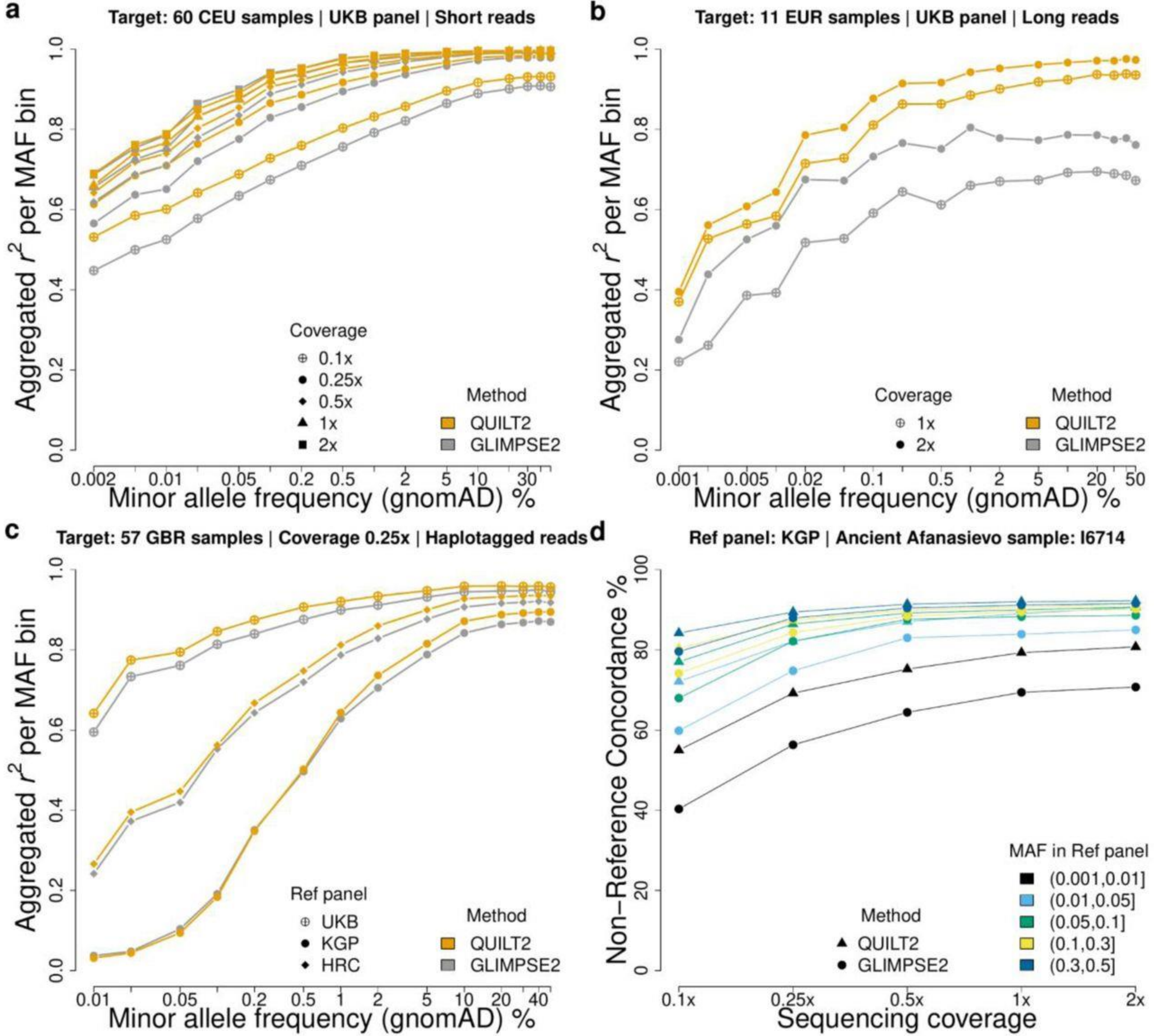

Sequencing reads
Genotype likelihoods
Individual 1
Individual 2
Individual 3
0 1 2
Input
GLIMPSE
(1) Haplotype selection
Reference panel
Target
PBWT
Haplotypes selected for individual 1
(2) Haplotype sampling
Sampling for haplotype 1
Sampling for haplotype 2
Linear time
New haplotype estimate for individual 1
Sampler iterations across individuals
Output
Haplotype calls
Genotype posteriors
Individual 1
Individual 2
Individual 3

QUILT2



full reference panel
1 3
2 4
Load reads
common SNPs only
initialized randomly
mat origin =
pat origin =
dark = common
light = rare
Common SNPs
Gibbs sampler
msPBWT
sample
1
vs
1
given
3
2
4
and
initialized randomly
hap1
+
hap2
hap1
+
hap2
and
until 3 iterations
after 3 iterations
after 3 iterations
1 4 7
2 5 8
3 6 9
Load reads
all SNPs
initialize using
hap1, hap2
All SNPs
Gibbs sampler
sample
1
vs
1
given
4 7
2 5 8
3 6 9
and
hap1
+
hap2
Output

Performance comparison



Conclusions

- Imputation – great tool for saving money and increasing power
- Different methods perform differently on different types of data
- Works best for well represented variants (common variants)