

Data QC and Exploratory Data Analysis (Afternoon session)

Olivier Delaneau
University of Geneva
19/05/2017

Outline

- Haplotype estimation
- Genotype imputation
- Reference panels
- Imputation performance

Haplotype estimation

We observe genotypes for each individual

G 2 1 1 2 2 1 1 1 0 2 1 2 2 1 1 1 2 1 1 0

Haplotype estimation

We observe genotypes for each individual

G 2 1 1 2 2 1 1 1 0 2 1 2 2 1 1 1 2 1 1 0



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

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

We want to recover the underlying haplotypes

Complex problem:

N heterozygotes $\rightarrow 2^N$ haplotypes

N heterozygotes $\rightarrow 2^{N-1}$ pairs of haplotypes

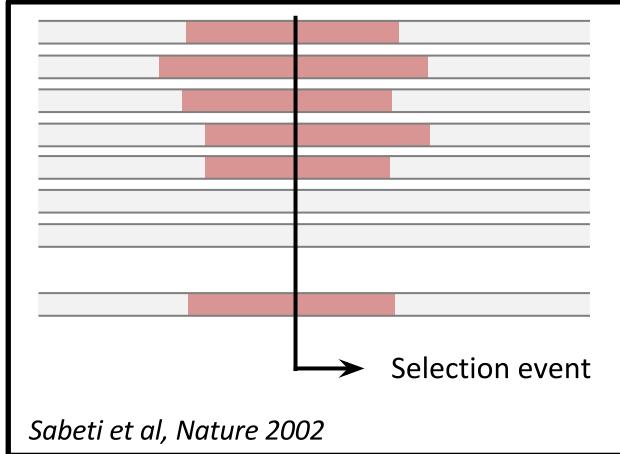
Haplotypes in disease genetics

Estimation of haplotypes from array-based and sequencing studies is an important problem for multiple reasons:

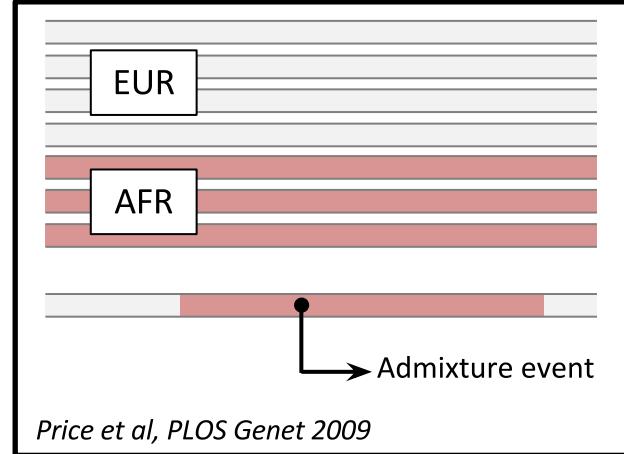
1. Construction of reference panels of haplotypes (HapMap, 1000 Genomes, UK10K, GoNL, etc...).
2. Imputation of Genome Wide Association Studies (i.e. GWAS).
3. Detection of novel associations in GWAS or fine-mapping of known effects e.g. compound heterozygotes.

Haplotypes in population genetics

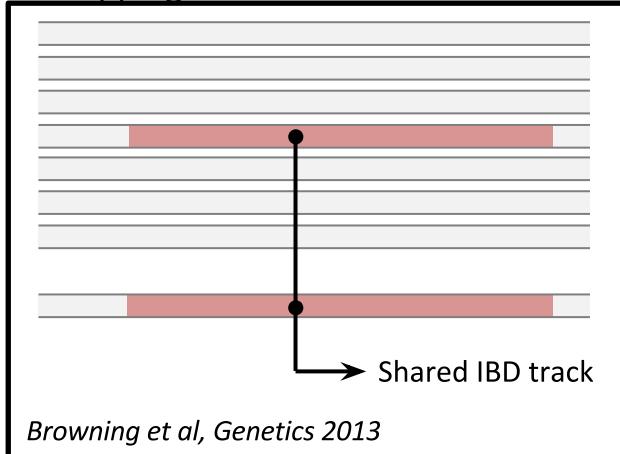
1. Mapping selection



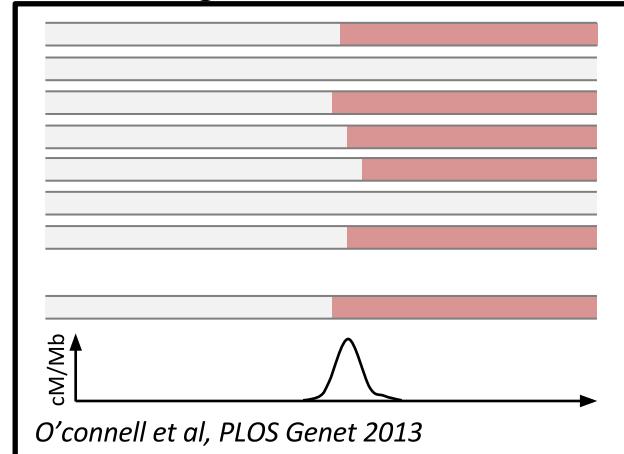
3. Mapping admixture



2. Mapping IBD



4. Estimating recombination rates



Basic idea of phasing

H

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

G

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

Haplotype estimation works iteratively.

Basic idea of phasing

H

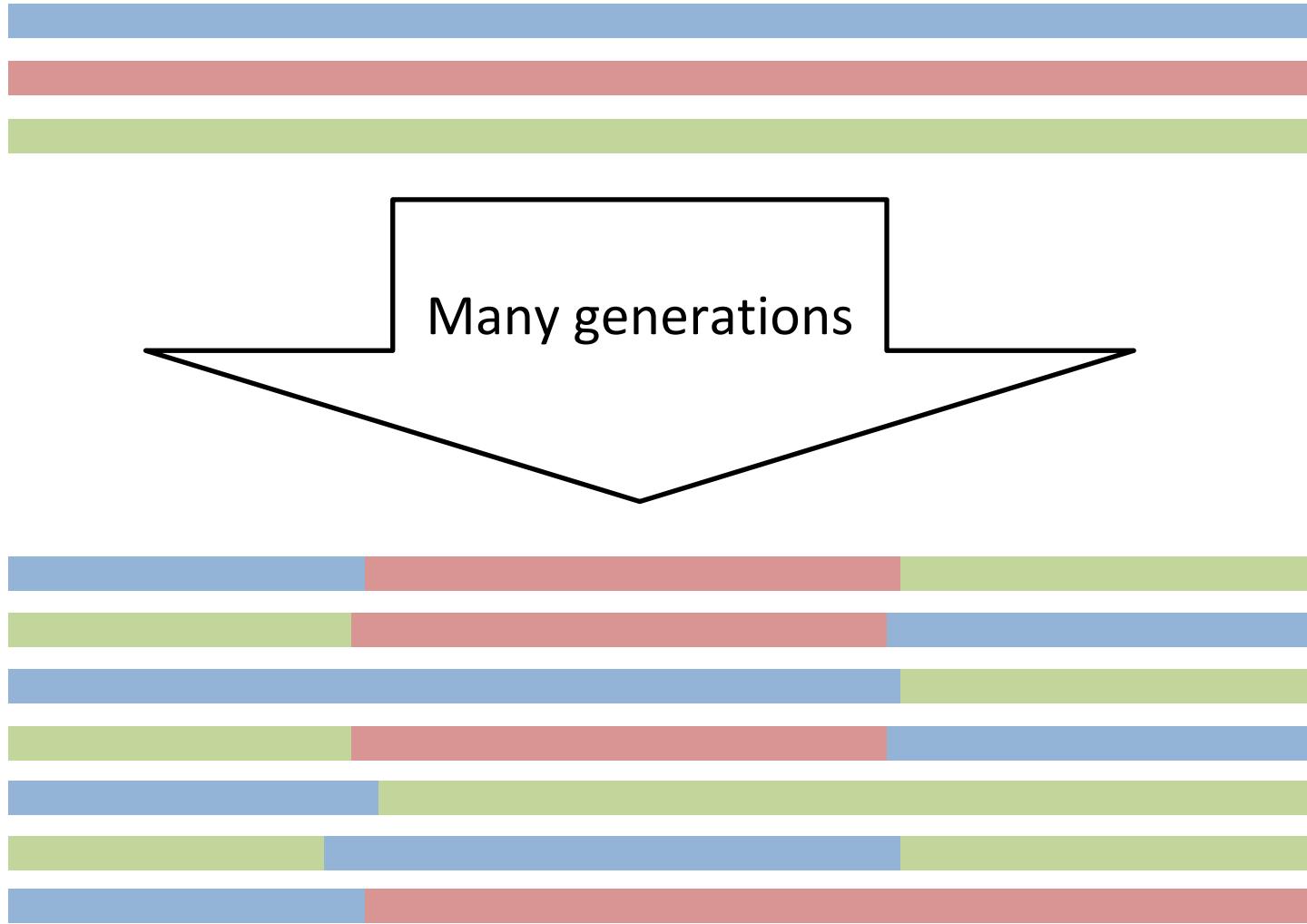
0	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	1	0		
0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	1	0	1	0	
1	0	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	
1	0	1	0	0	0	1	0	1	0	0	0	0	1	1	1	1	0	1	0	1	
0	1	0	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	
1	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	0	1	
0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0	
1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

G

2	1	1	2	2	1	1	1	0	2	1	2	2	1	1	1	2	1	1	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

It takes each individual's genotype vector (G) in turn and estimates the underlying haplotypes using the current haplotype estimates of other individuals (H).

Haplotype structure



Individuals share stretches of haplotypes inherited from common ancestors

Basic idea of phasing

	0	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	1	0
H	0	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	1	0
	1	0	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	0	0
	1	0	1	0	0	0	1	0	1	0	0	0	0	1	1	1	1	0	1	0
	0	1	0	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	0	0
	1	0	1	1	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0	1
G	1	0	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	0	0
	0	1	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	0	1	0
	2	1	1	2	2	1	1	1	0	2	1	2	2	1	1	1	2	1	1	0

It uses Hidden Markov Models (*Li and Stephens, 2005*) that models the underlying haplotypes as mosaics of haplotypes of other individuals.

Popular phasing methods

There are several popular methods for haplotype estimation from genotype data:

IMPUTE2 - https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

MACH - <http://www.sph.umich.edu/csg/abecasis/MACH/>

BEAGLE - <http://faculty.washington.edu/browning/beagle/beagle.html>

SHAPEIT2 - http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

EAGLE - <https://data.broadinstitute.org/alkesgroup/Eagle/>

SHAPEIT3 - http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

Measuring performance

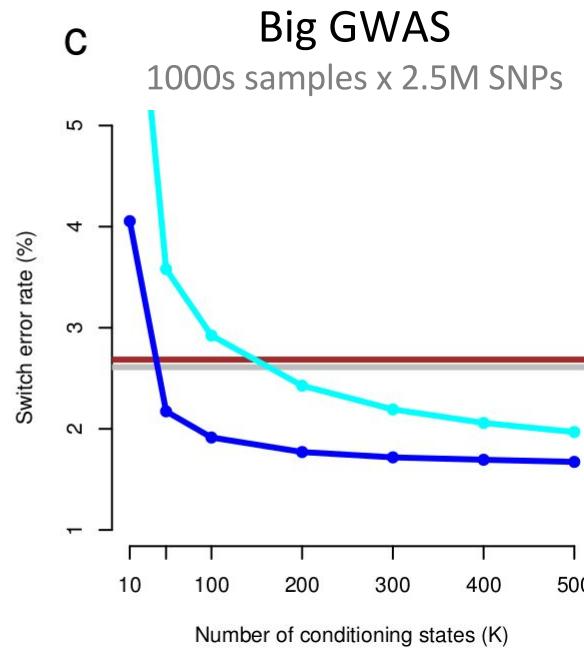
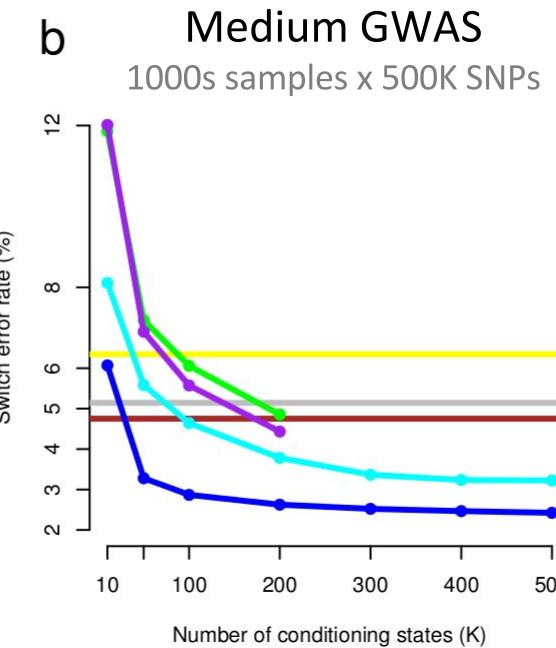
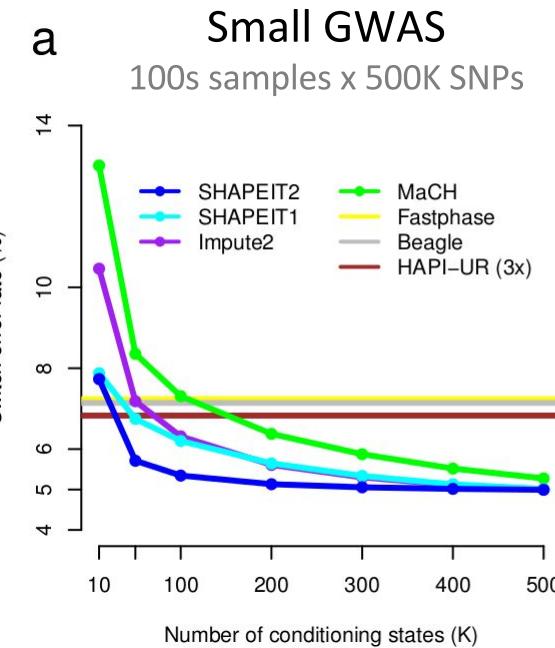
Truth

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

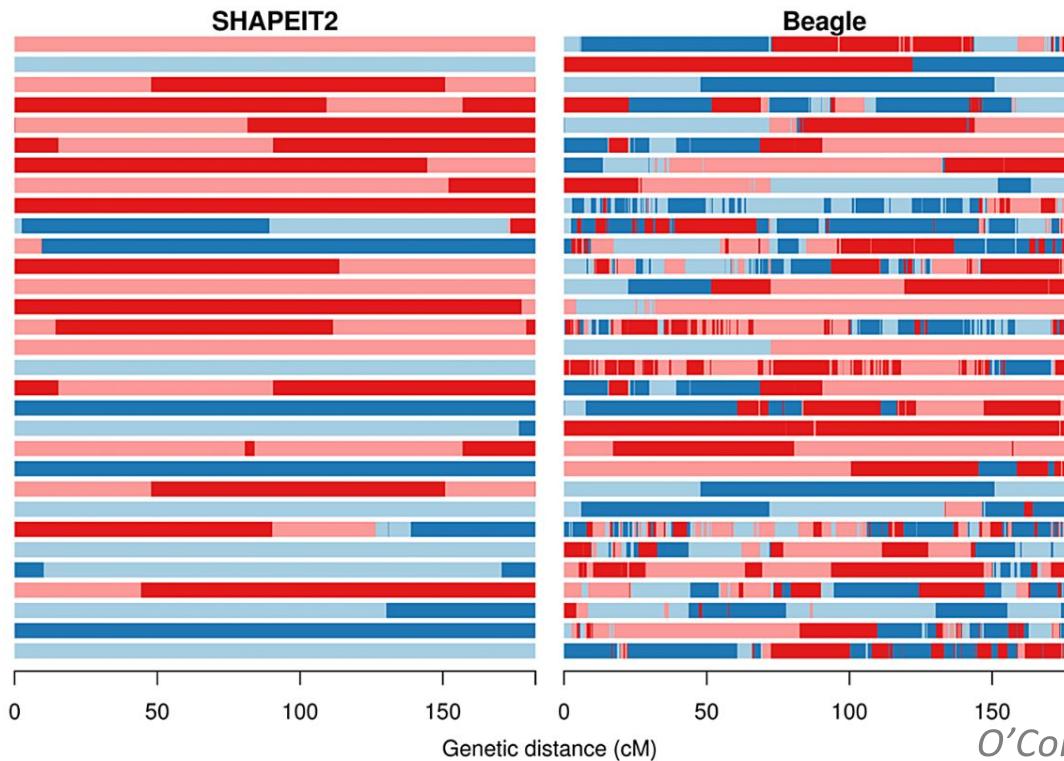
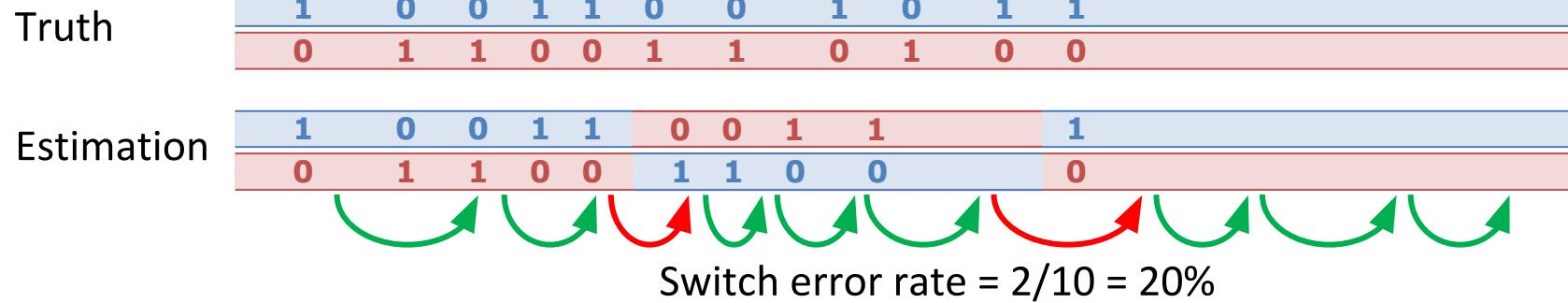
Estimation

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

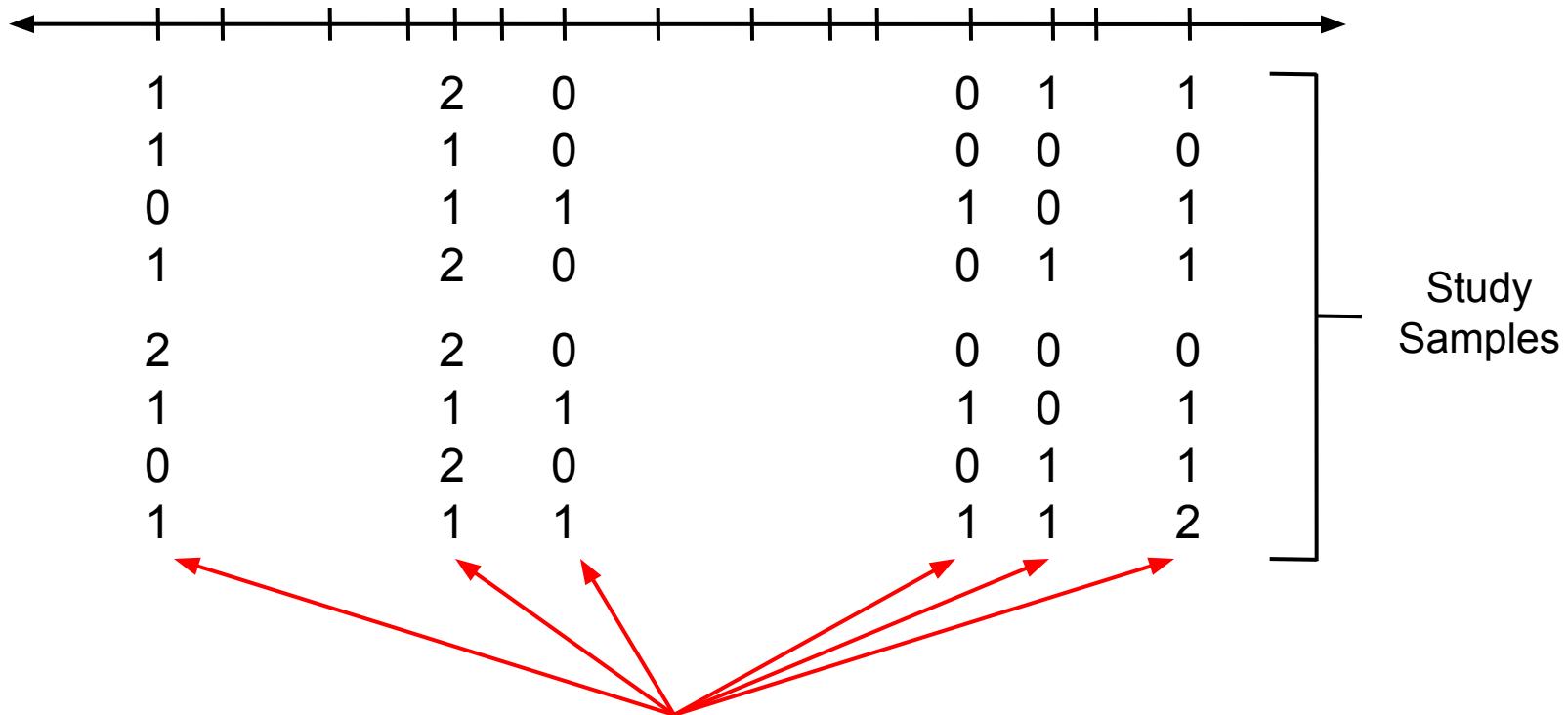
Switch error rate = 2/10 = 20%



Measuring performance

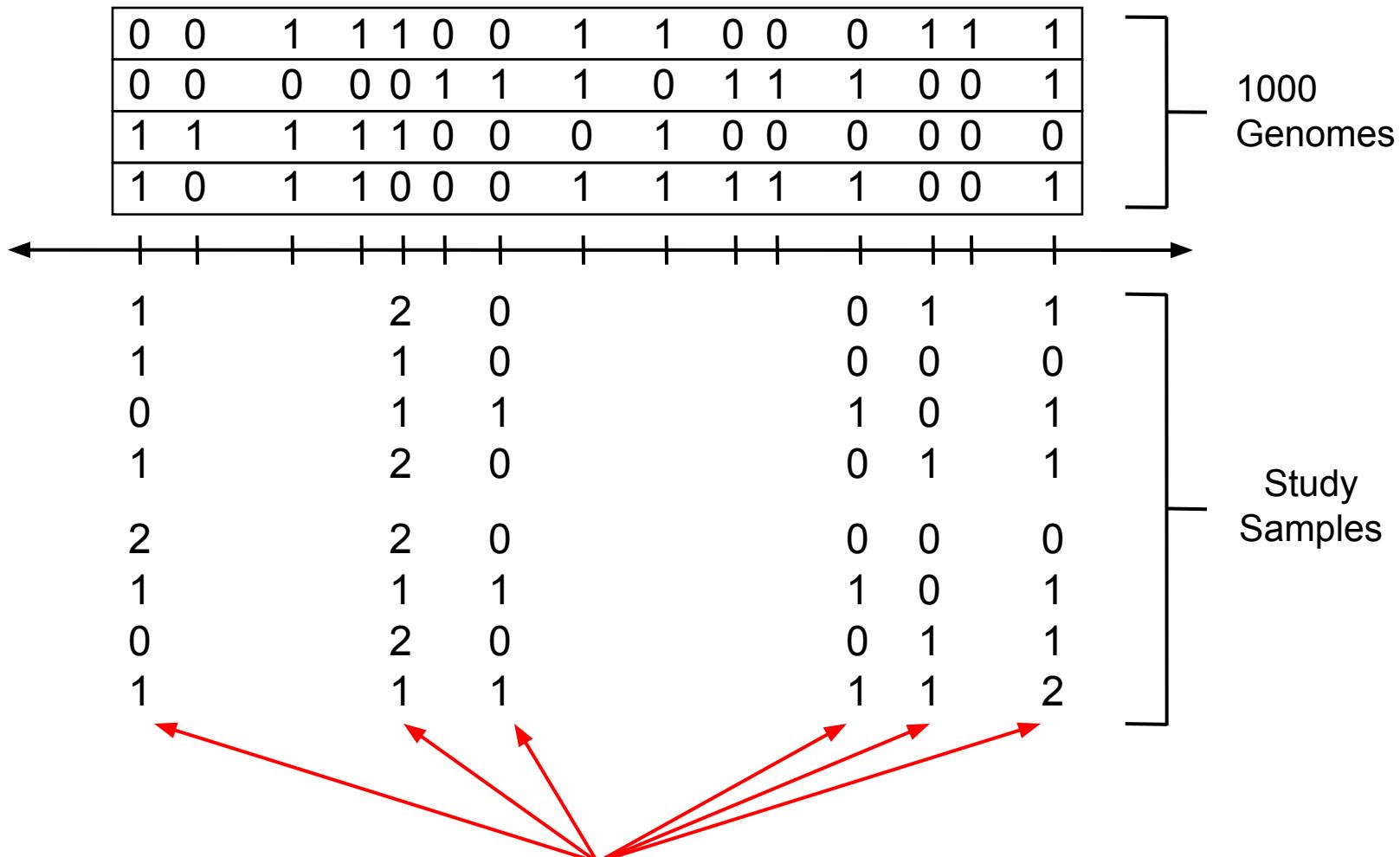


Limitation of genotyping data



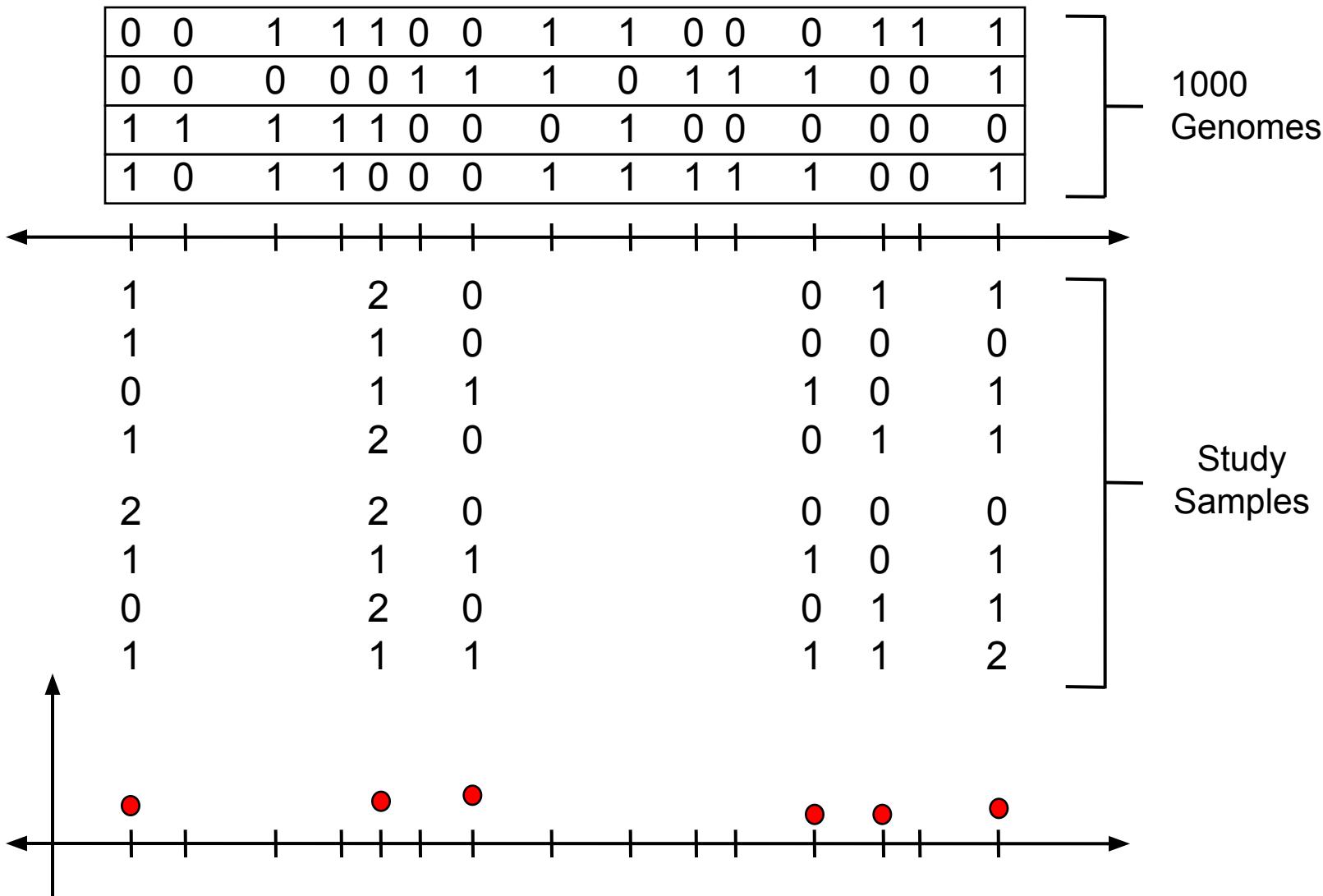
SNPs genotyped in a genomic study: 1 to 5M variants, 100s to 1000s of samples

Limitation of genotyping data

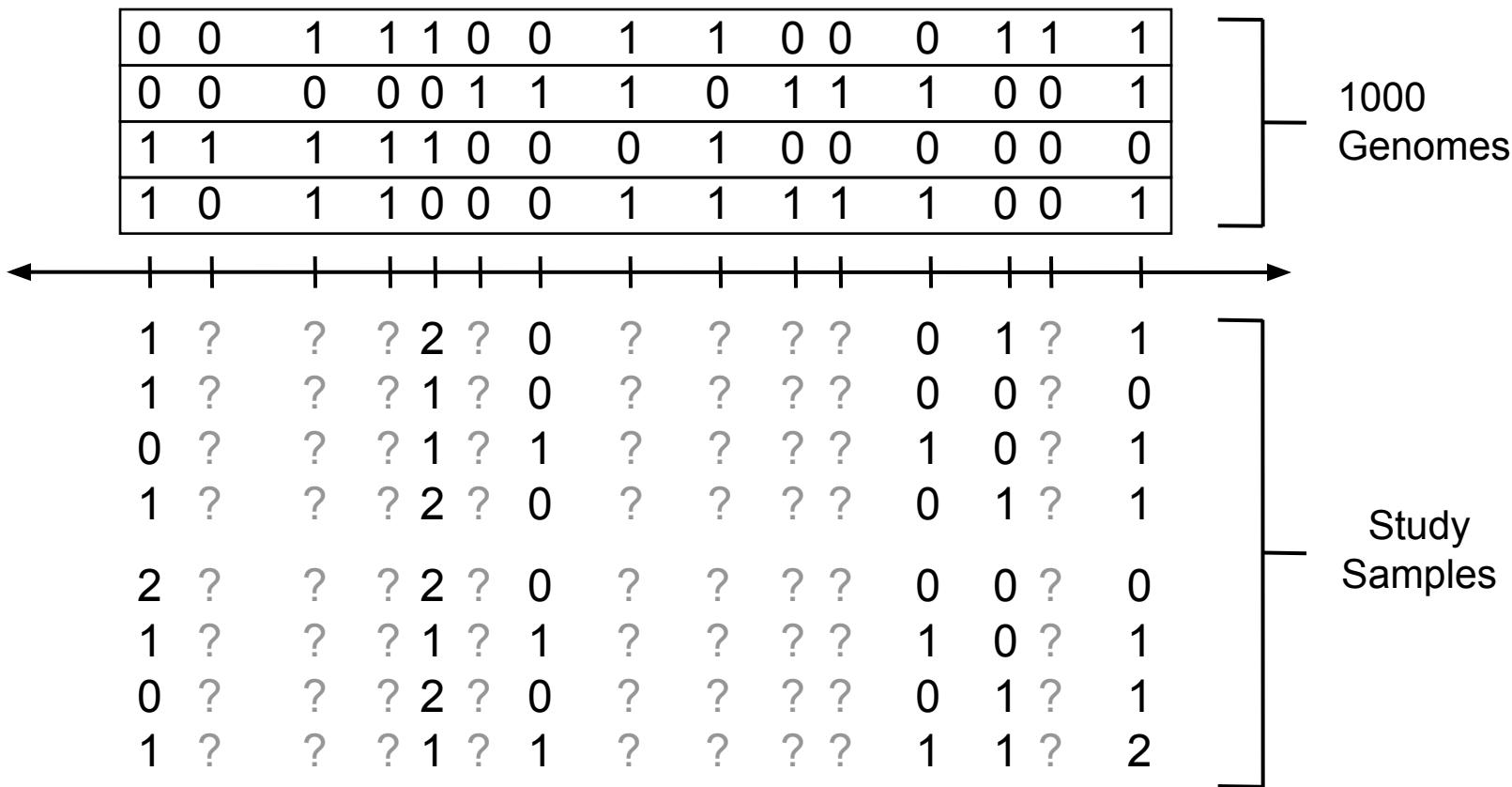


This is far less than what we've got in sequencing based studies

Limitation of genotyping data



Solution: Imputation



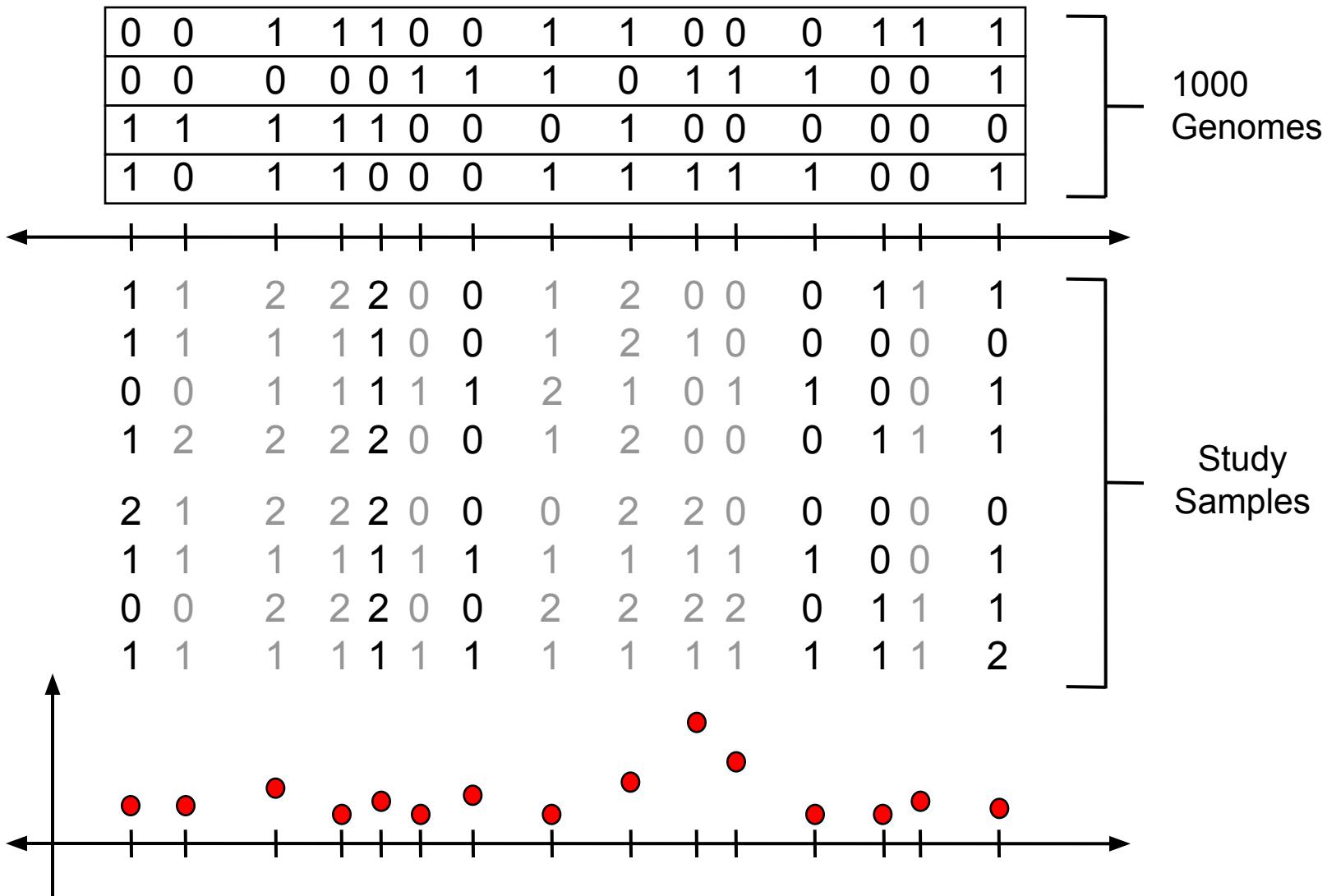
Un-typed variants are treated as missing data.

Solution: Imputation

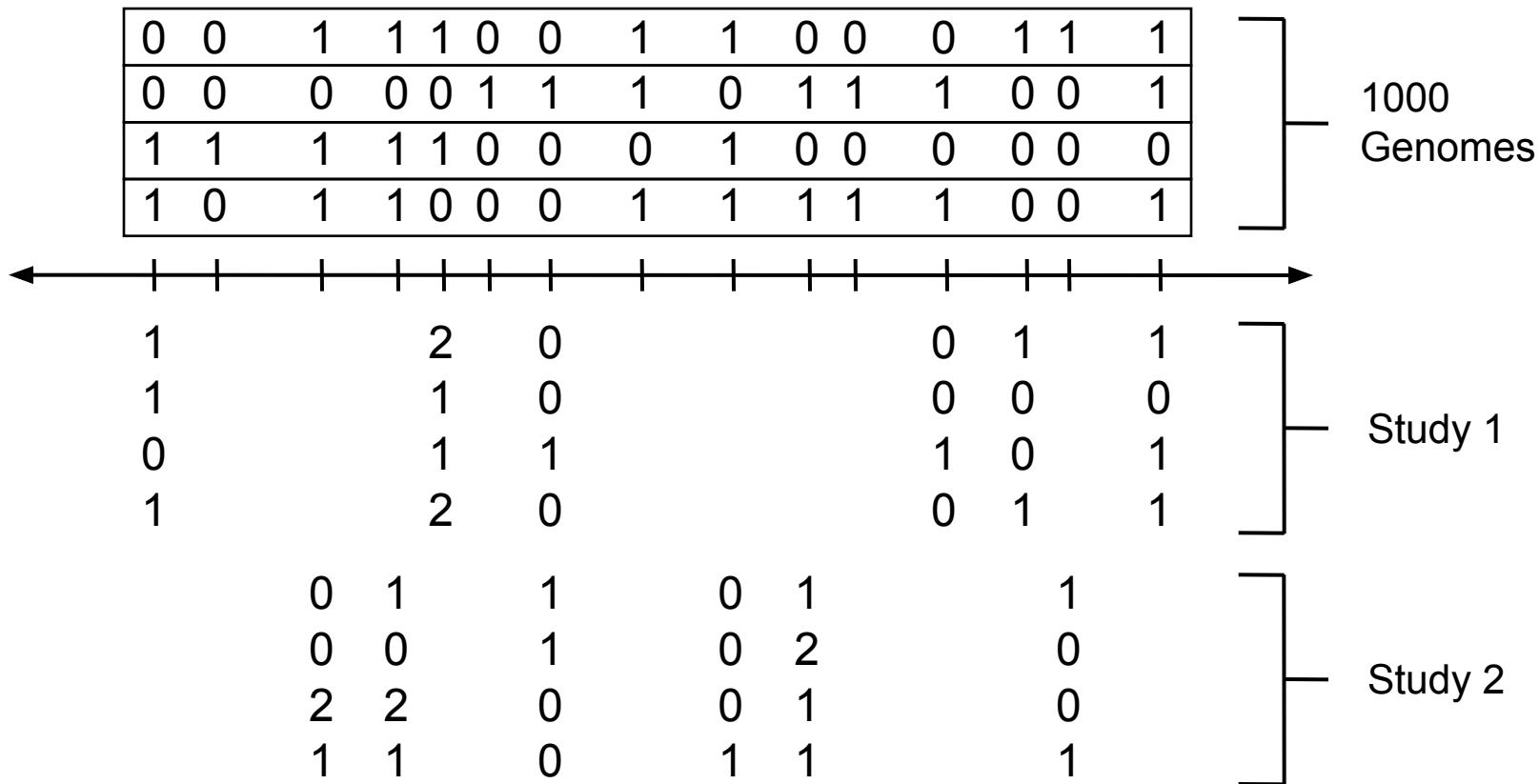
0	0	1	1	1	0	0	1	1	0	0	0	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0
1	1	1	1	1	0	0	0	1	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0
1	1	2	2	2	0	0	1	2	0	0	0	1	1
1	1	1	1	1	0	0	1	2	1	0	0	0	0
0	0	1	1	1	1	1	2	1	0	1	1	0	0
1	2	2	2	2	0	0	1	2	0	0	0	1	1
2	1	2	2	2	0	0	0	2	2	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	0	0
0	0	2	2	2	0	0	2	2	2	2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

The goal of imputation is to estimate the missing genotypes at these variants

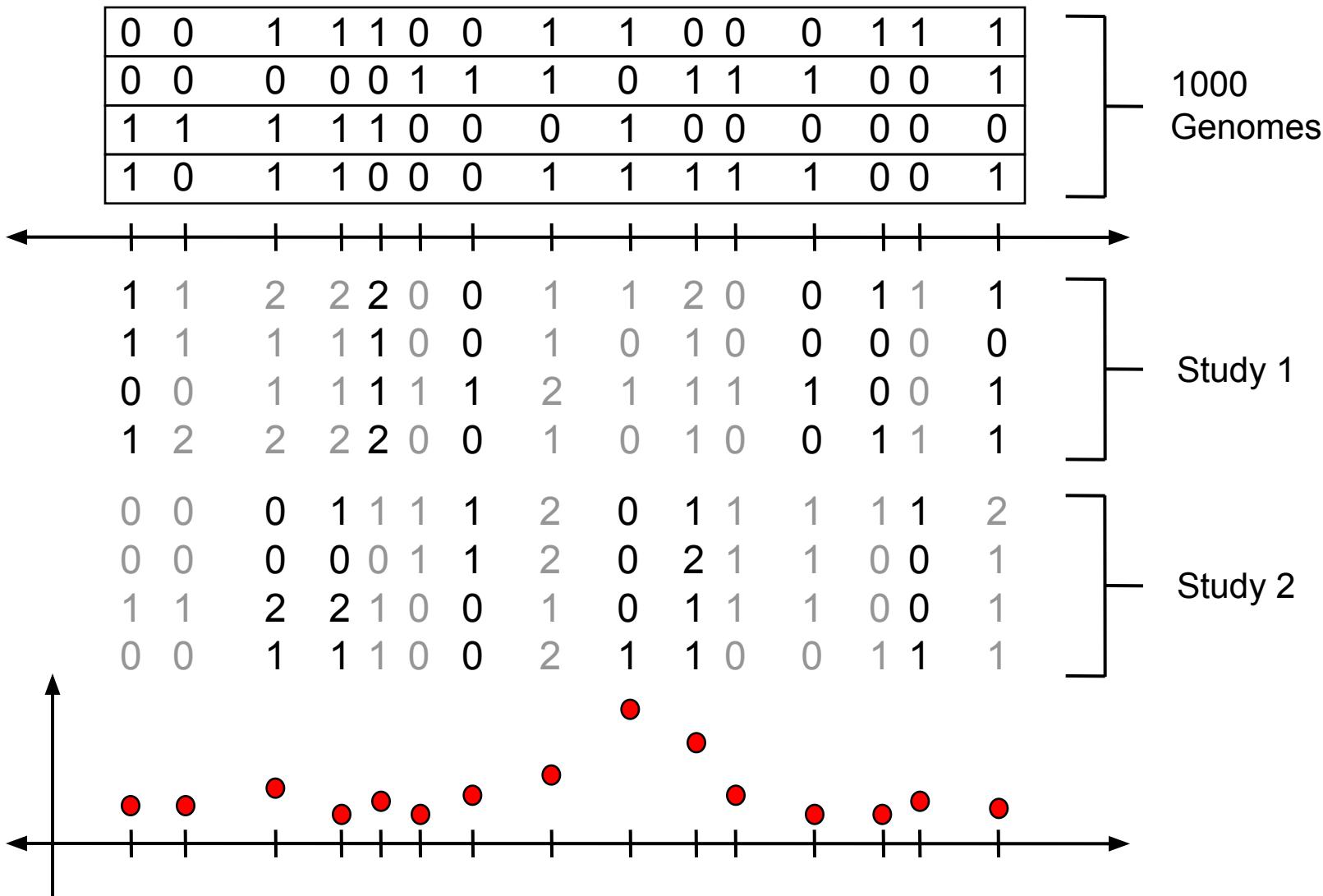
Solution: Imputation



Combining genotype data sets



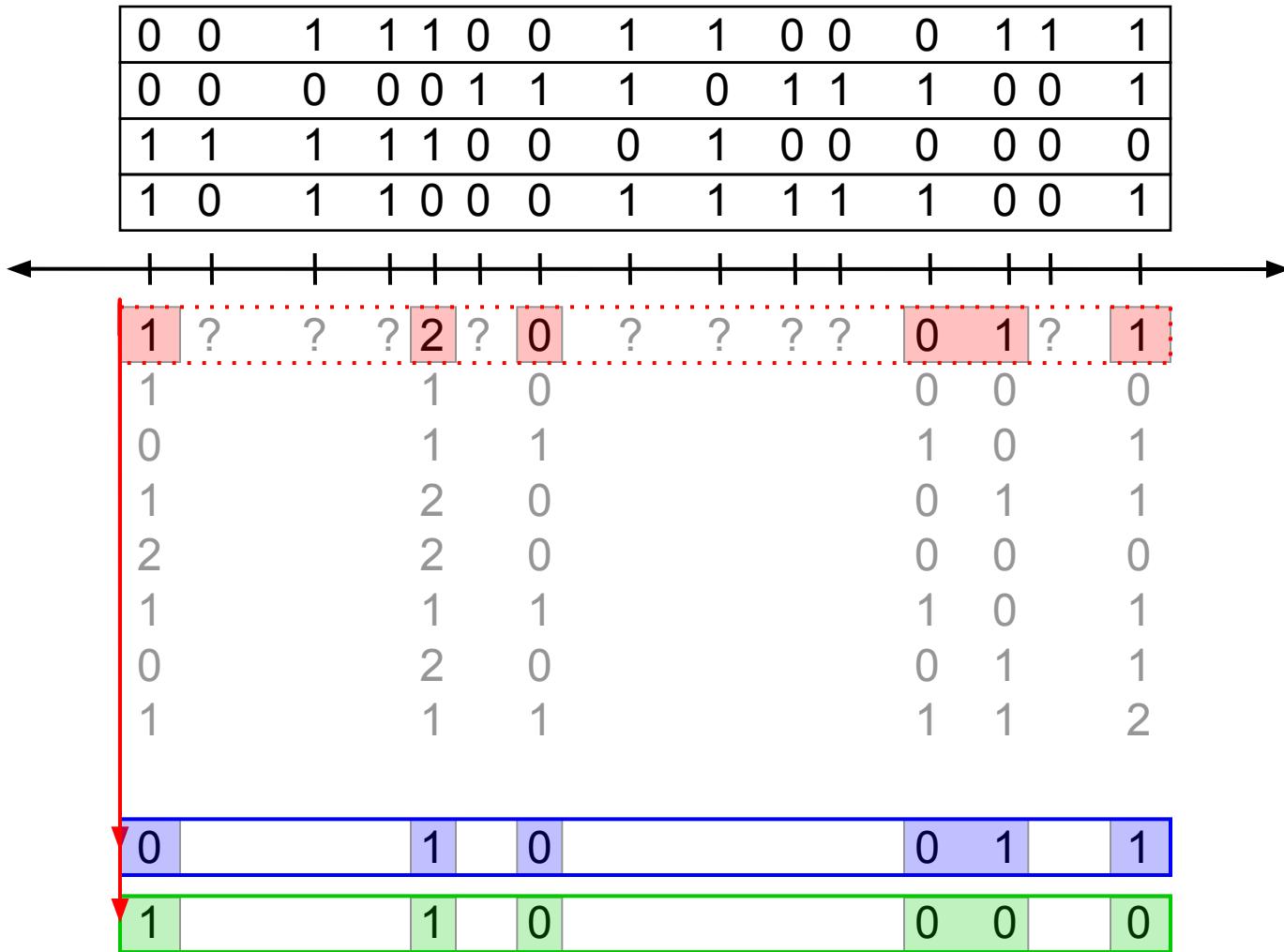
Combining genotype data sets



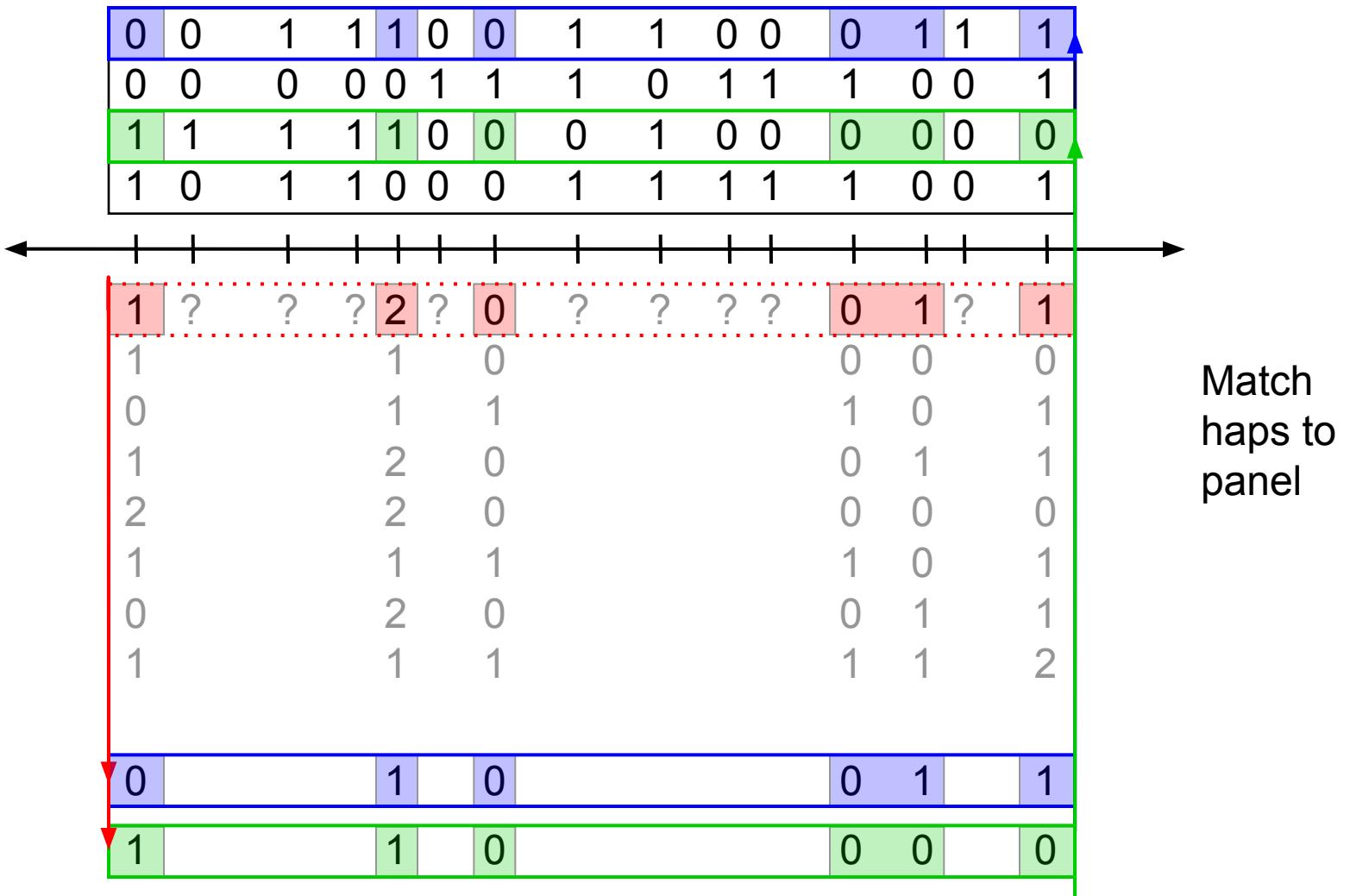
How does it work?

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

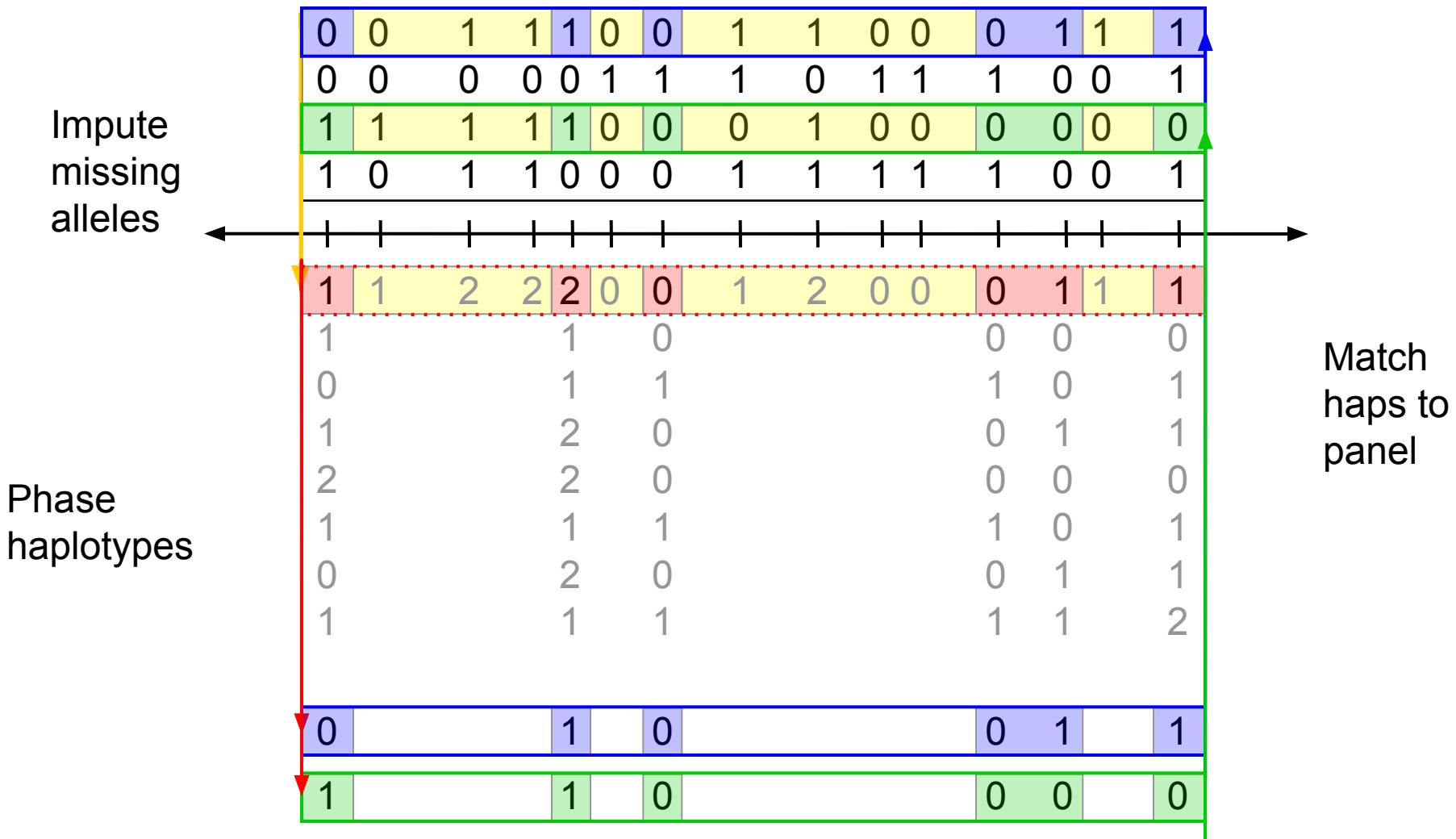
How does it work?



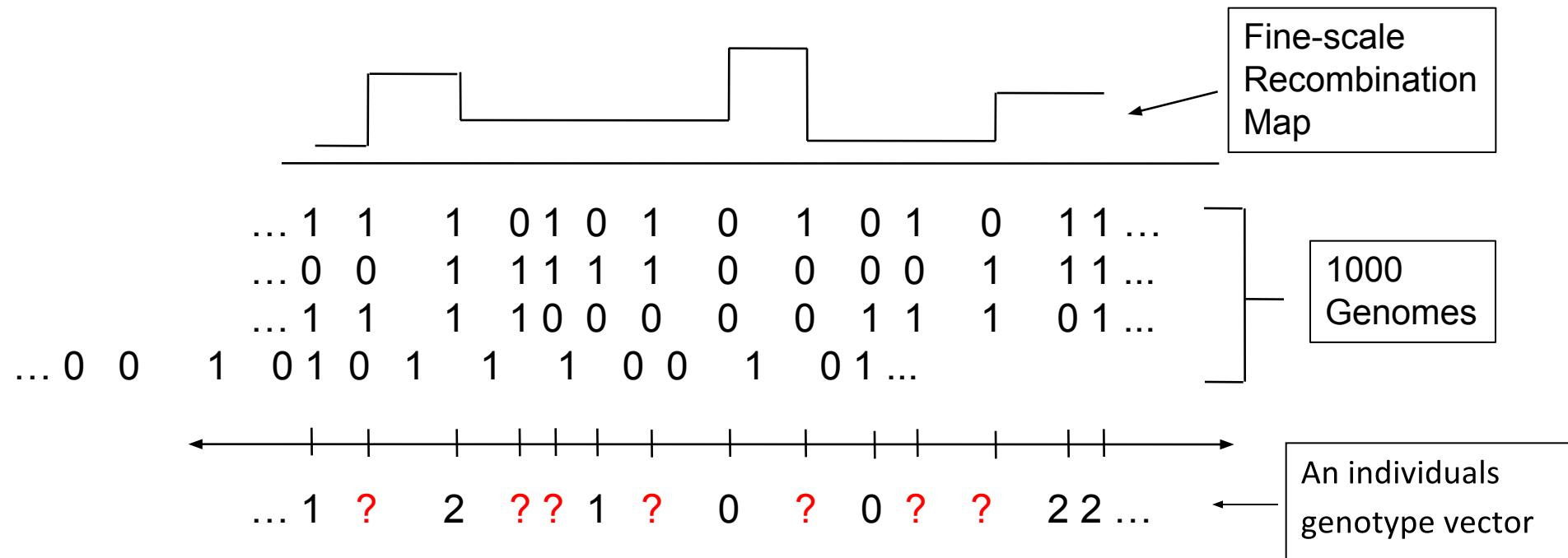
How does it work?



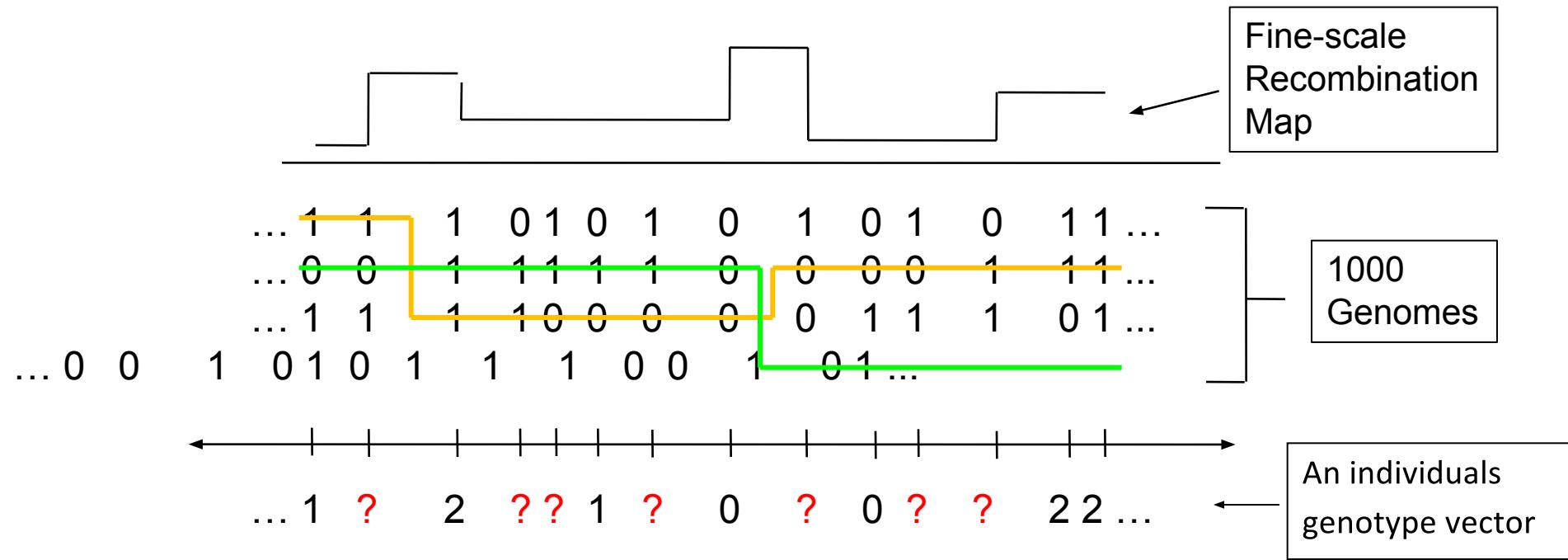
How does it work?



How does it work?

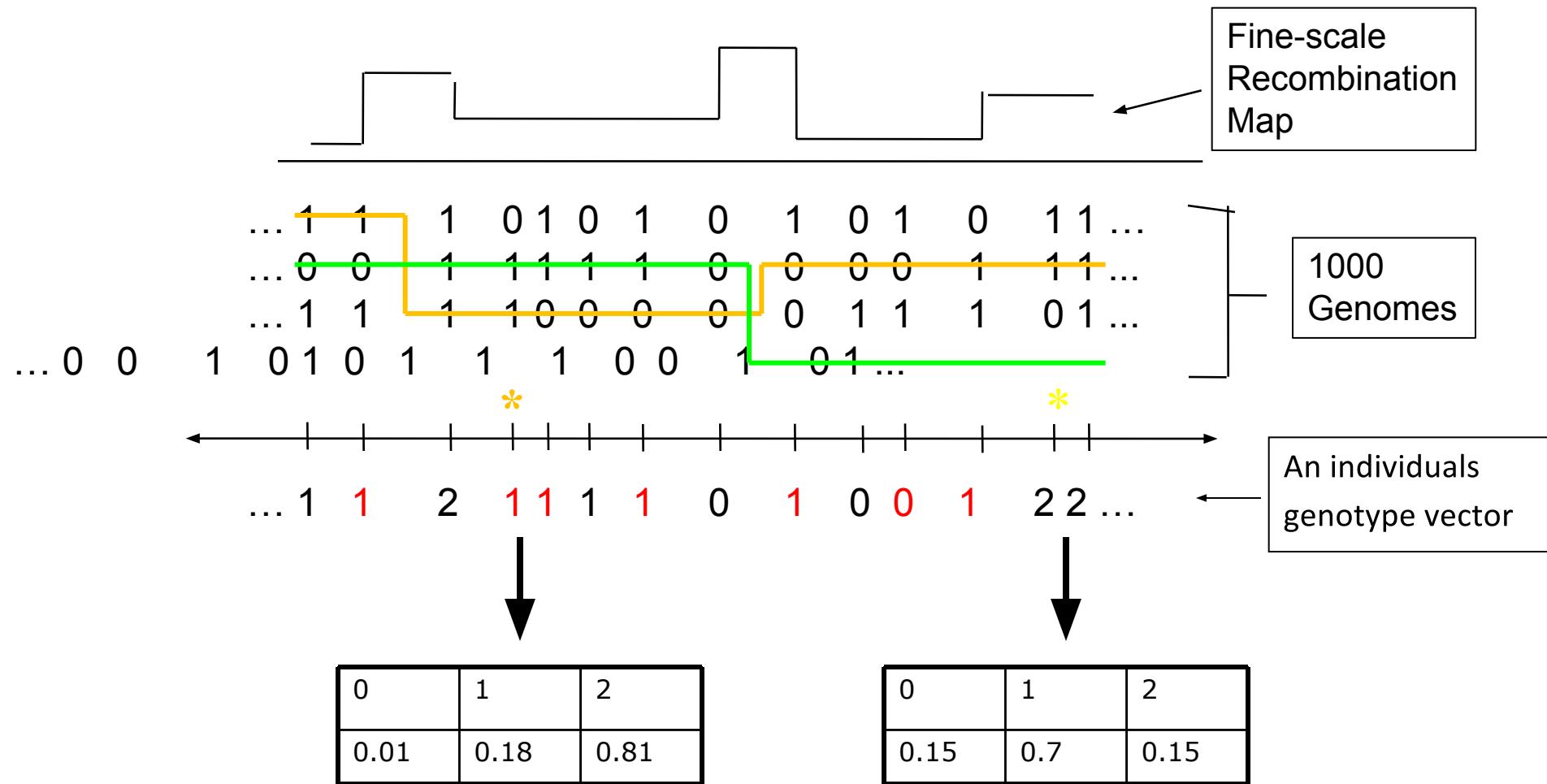


How does it work?



The model says that an individual's genotype is constructed by copying alleles along two paths through the space of haplotypes. The switch rates of the paths are controlled by the recombination map. Mutation events are also allowed.

How does it work?



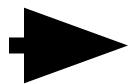
We produce estimates of genotype uncertainty at both untyped and typed genotypes

Accounting for genotype uncertainty

There are several ways the imputed genotype probabilities can be used.

1. Threshold the probability distribution to give genotype calls
2. Use the expected allele counts

AA	AB	BB
0.01	0.18	0.81



$$0 \times 0.01 + 1 \times 0.18 + 2 \times 0.81 = 1.8$$

3. Average over uncertainty
 - Can be done in both the Frequentist and Bayesian frameworks
 - Very few implementations (SNPtest)

CD hit region, chromosome 1

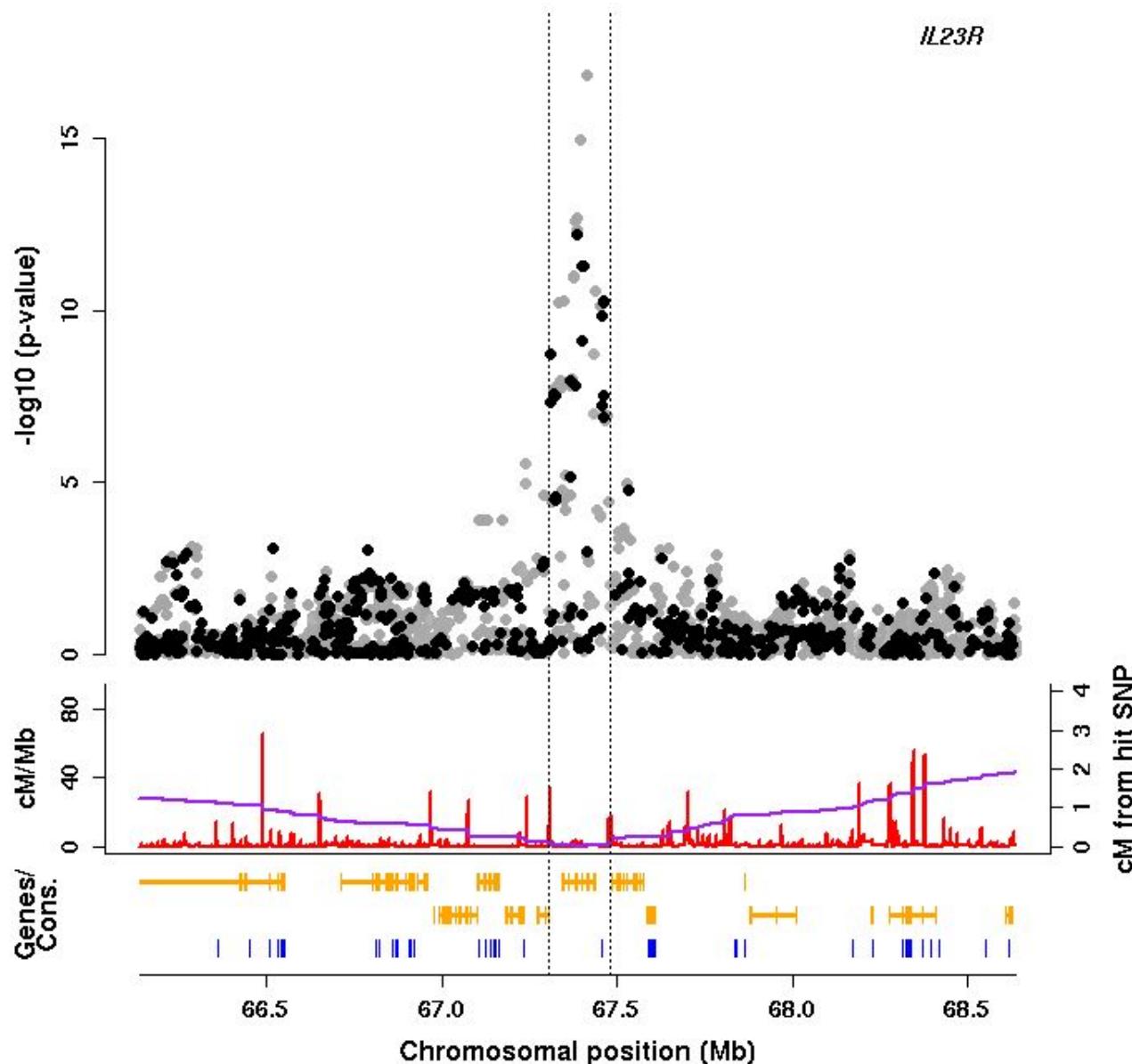


Image from The Wellcome Trust Case Control Consortium (2007) Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447;661-78. DOI: 10.1038/nature05911.

Genome build

The Human Reference Sequence is updated periodically, each version is referred to a ‘genome build’.

Positions of SNPs can change between builds.

Almost all imputation programs align SNPs between the reference panels and the GWAS datasets using the position of SNPs.

So it is very important that the genotypes of your GWAS are mapped to the same genome build as the reference panel you are using.

Currently, all the most commonly used reference panels use build 37/38.

Strand issue

Genotypes from SNP chips are called relative to either the + or – strand of the human reference genome.

Maternal chromosome	ACGTAGCTCTCTGATCGAT TGCAATCGAGAGAGACTAGCTA	+ strand - strand
---------------------	---	----------------------

Paternal chromosome	ACATAGCTCTCTGAACGAT TGTATCGAGAGAGACTTGCTA	+ strand - strand
---------------------	--	----------------------

+ strand GA
- strand CT

+ strand TA
- strand
AT

Haplotype reference panels have alleles aligned to the + strand.
Genotype chips can have a mixture of genotypes from + and - strand
This needs to be fixed prior to imputation.
The strand info about all chips can be found here

<http://www.well.ox.ac.uk/~wrayner/strand/>

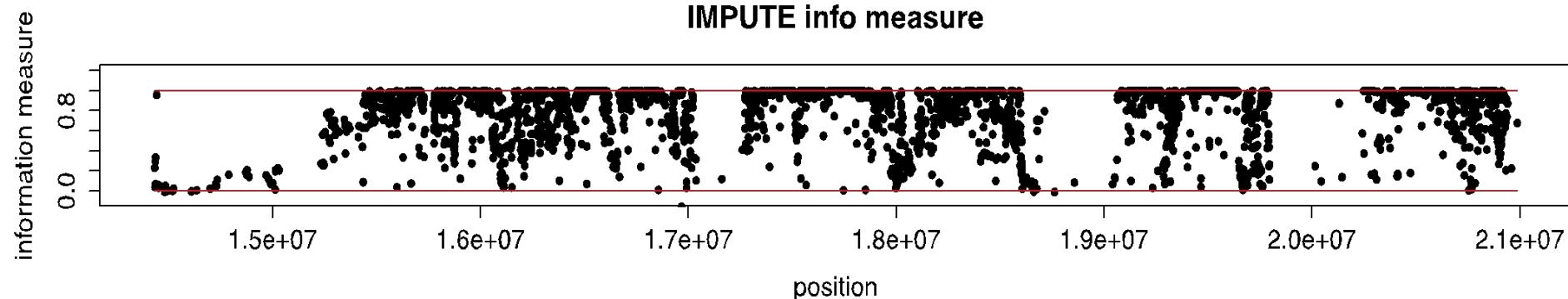
Information metric

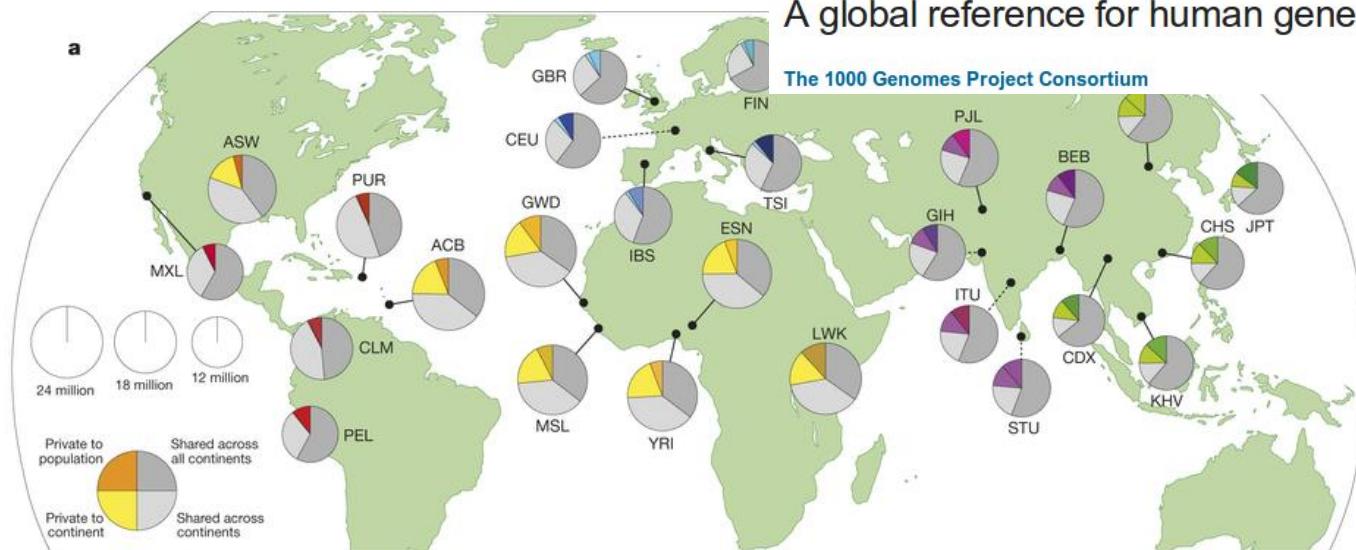
Once imputation has been carried out, it is a good idea to try and measure how well imputed the genotypes are at each SNP. IMPUTE produces an information measure for each SNP in the range [0,1]:

- 1 means there is no uncertainty at all in any of the imputed genotypes.
- 0 means there is complete uncertainty for all of the genotypes.

Approximately, an information measure of α means the imputed genotypes contain as much information as αN genotypes that are completely certain (where N is the total number of genotypes).

In recent published studies (especially those that have used imputation for meta-analysis), variants with $\text{info} < 0.4$ have been excluded from the study.





Population	Number of samples
African	661
Americas	347
Est Asian	504
European	503
South Asian	489
Total	2,504

Type of variants	Number of variants
Bi-allelic SNPs	81,102,777
Bi-allelic Indels	3,196,364
Multi-allelic SNPs	274,425
Multi-allelic Indels	169,601
Structural variants	58,713
Total	84,801,880

Why did we phase 1000 Genomes?

1. To get fully resolved haplotypes, obviously,

ATGC**A**TGAGCT**G**ACTGAGG**A**CTGGACTAG**C**GATCAG
ATG**C**TTCGAGCT**T**ACTGAGG**C**CTGGACTAG**G**GATCAG

ATGC**A**TGAGCT**G**ACTGAGG**A**CTGGACTAG**C**GATCAG
ATGC**A**TGAGCT**T**ACTGAGG**A**CTGGACTAG**G**GATCAG

Why did we phase 1000 Genomes?

1. To get fully resolved haplotypes, obviously,
2. But also to refine the genotype calls.

ATGC A TCG	G C GATCAG
CGAGCT G A	
ATGC A TCGAGCT G ACTGAGG A CTGGACTAG C GATCAG	
ATG C TCGAGCT T ACTGAGG C CTGGACTAG G GATCAG	
ATG C TCG CTTACTGA GC C CTGGAC	
C TTCGAGC	G G GATCAG

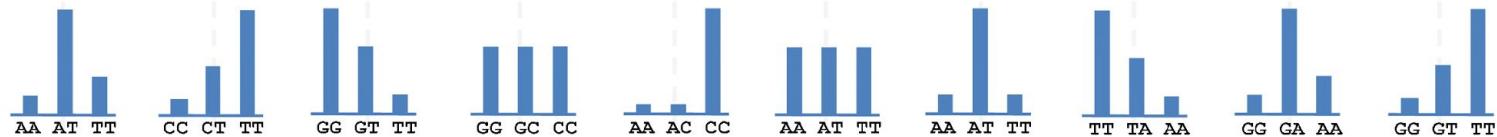
GC A TCGAG	CTGAGG A C	CTAG C GAT
	GG A CTGGA	
ATGC A TCGAGCT G ACTGAGG A CTGGACTAG C GATCAG		
ATGC A TCGAGCT T ACTGAGG A CTGGACTAG G GATCAG		
ATGC A TCG	AGG A CTGG	
CGAGCT T A		G G GATCAG

How does it work?

H

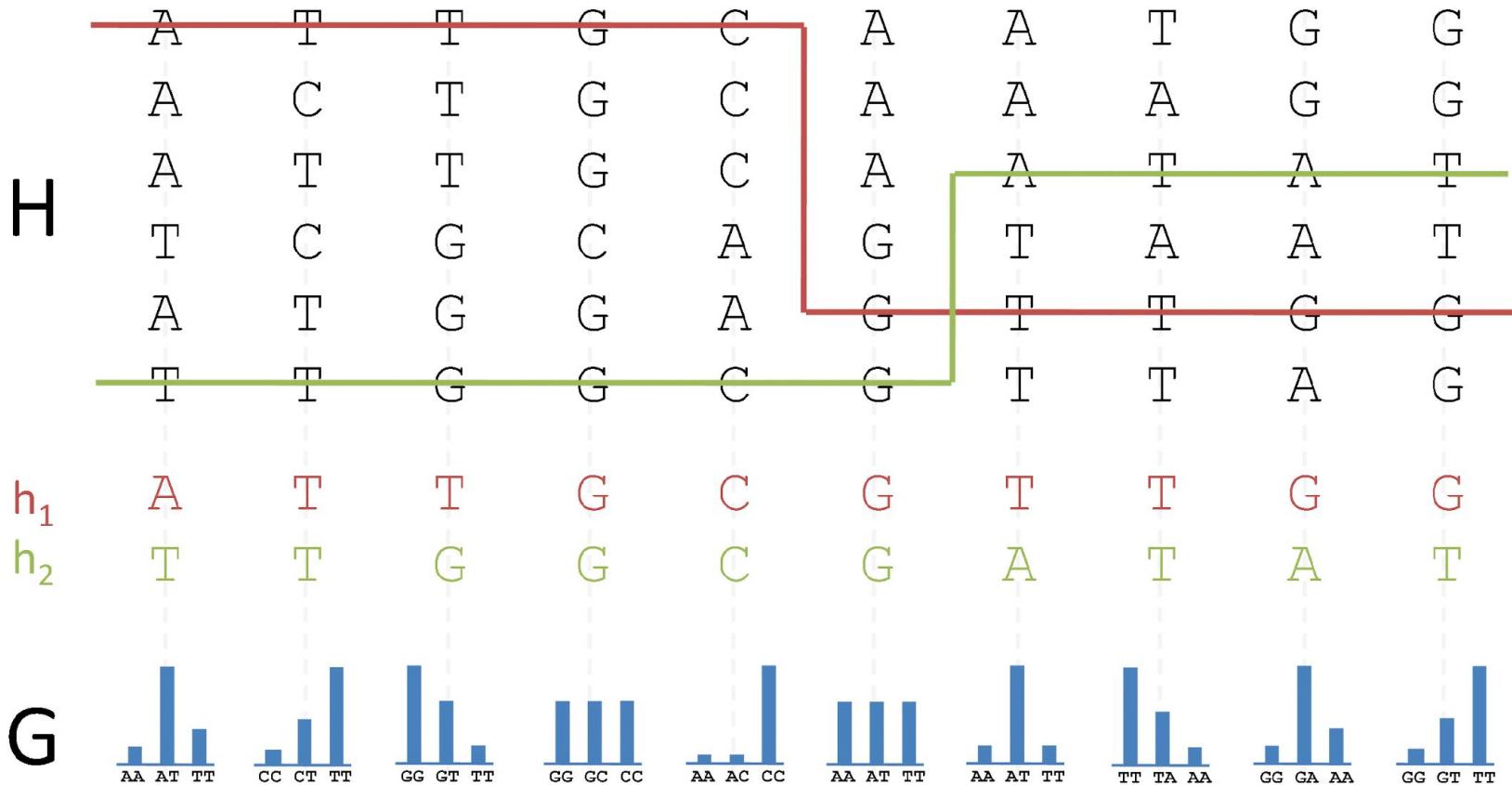
A	T	T	G	C	A	A	T	G	G
A	C	T	G	C	A	A	A	G	G
A	T	T	G	C	A	A	T	A	T
T	C	G	C	A	G	T	A	A	T
A	T	G	G	A	G	T	T	G	G
T	T	G	G	C	G	T	T	A	G

G



Uncertainty in genotypes is modeled with genotype likelihoods (GLs)

How does it work?



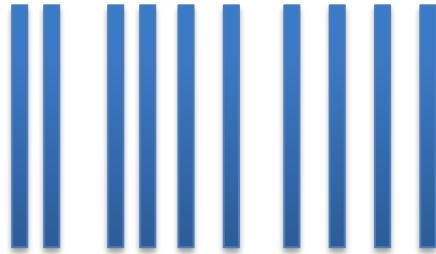
Estimating haplotypes produces genotype calls.

Measuring imputation performance

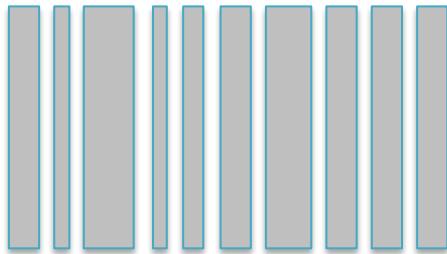
Complete Genomics
Genotypes

Measuring imputation performance

Genotypes on Illumina 1M



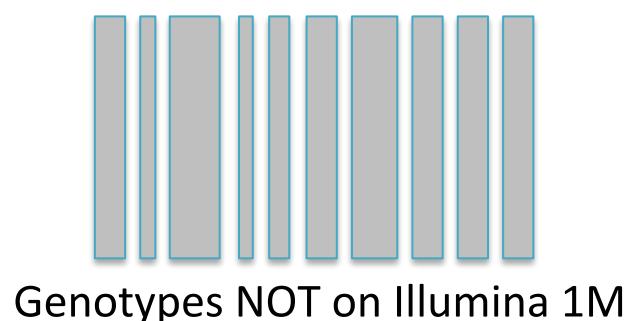
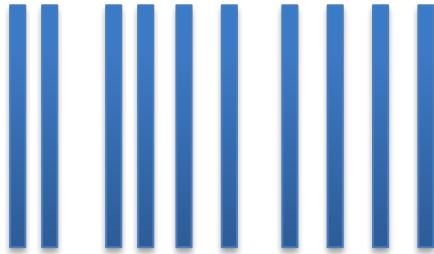
Genotypes NOT on Illumina 1M



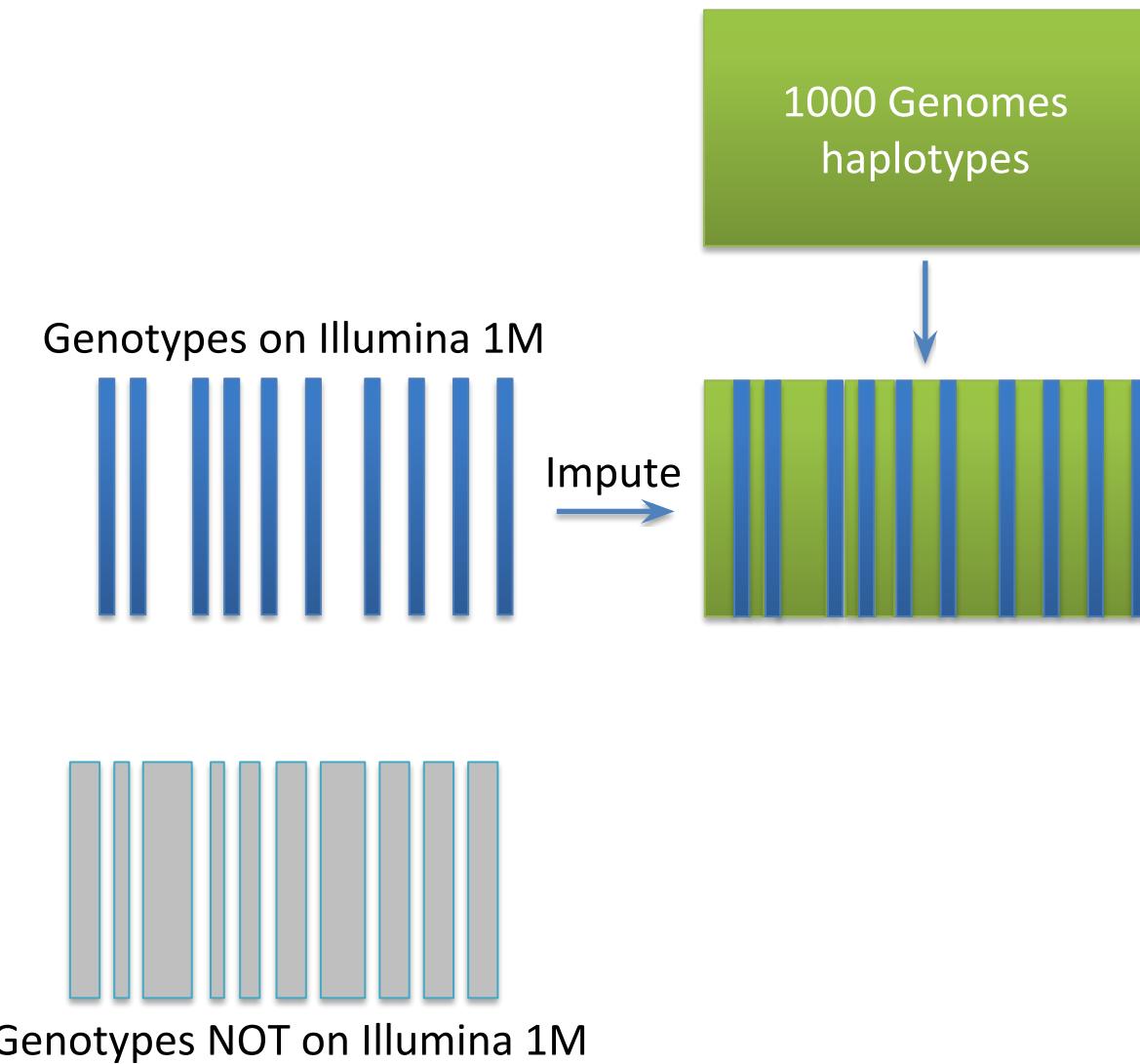
Measuring imputation performance

1000 Genomes
haplotypes

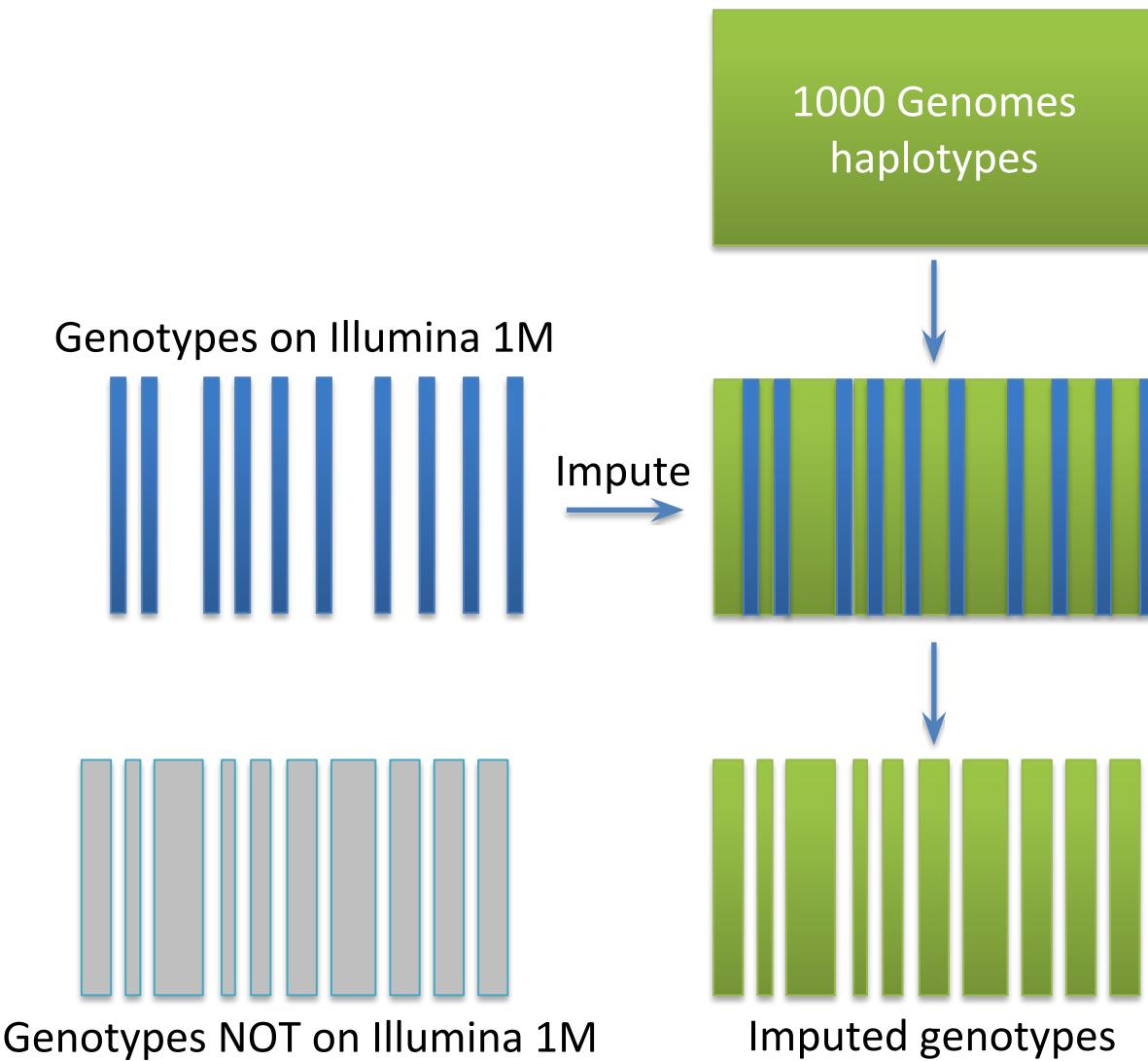
Genotypes on Illumina 1M



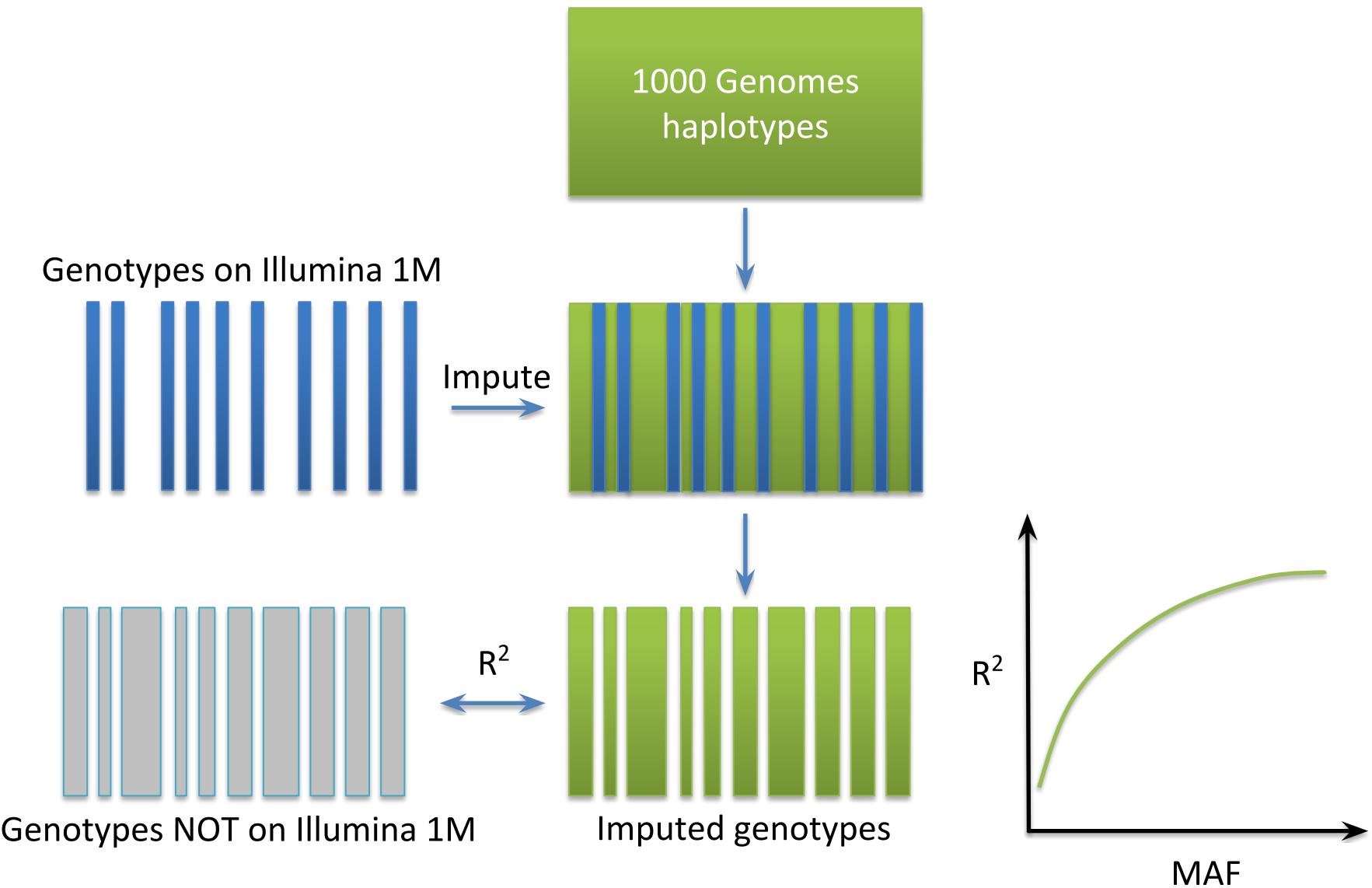
Measuring imputation performance



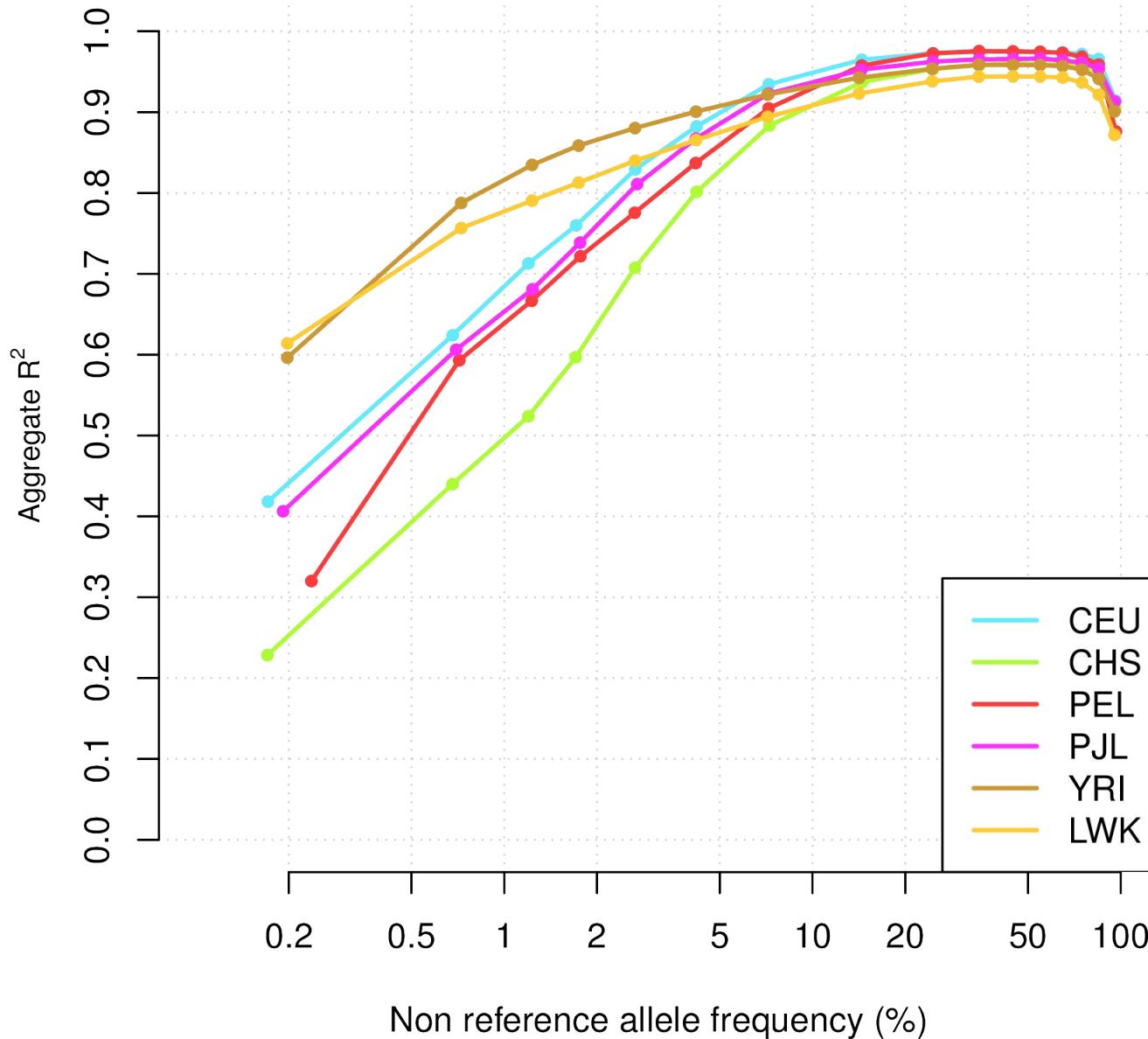
Measuring imputation performance



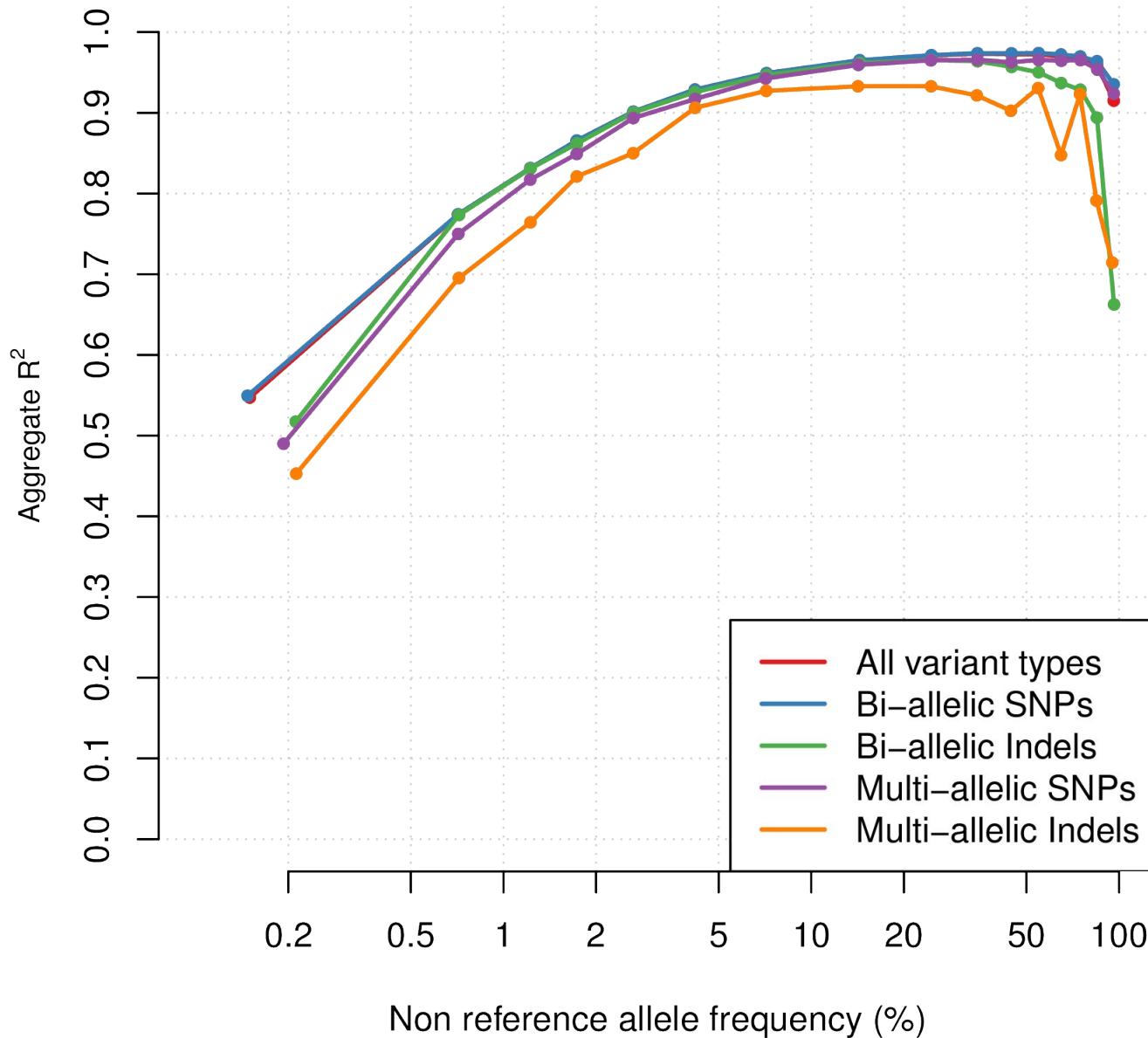
Measuring imputation performance



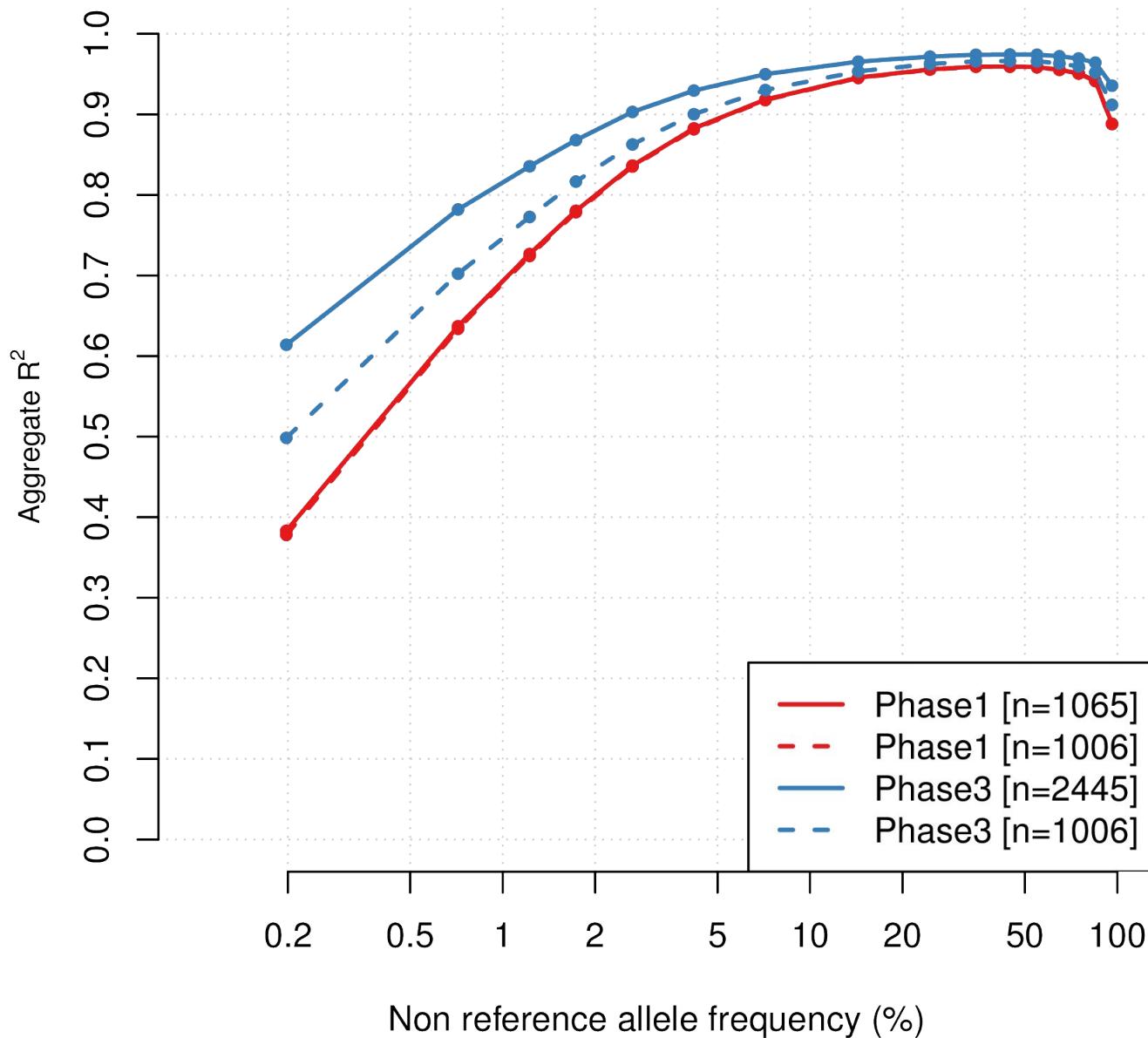
Stratified by populations



Stratified by variant types



Phase1 versus Phase3



Haplotype reference panels

Reference	Year	# haplotypes	# populations	# variants
HAPMAP 2	2006	420	3	2,139,483
HAPMAP 3	2009	2,368	11	1,440,616
1000GP Pilot	2010	358	3	14,894,361
1000GP 2010	2010	1,258	14	22,242,654
1000GP Interim	2011	2,186	14	38,558,931
1000GP Phase Reference	2012	2,186	14	38,219,282
	Year	# haplotypes	# populations	# variants
1000GP Final	2014	~5,000	25	~80,000,000
UK10K	2013/14	~8,000	UK	~50,000,000
GoT2D	2013/14	~5,600	European	~27,000,000

Q: Which reference panel to use for imputation into GWAS?
A: 2013/14, 1,000 NL, 20,000,000

The Haplotype Reference Consortium

<http://www.haplotype-reference-consortium.org/>

- The HRC data **will NOT be publically available**, as HapMap and 1000GP haplotypes are, due to consent issues. A subset of HRC haplotypes will be available via EGA for the sole purpose of imputation.
- Currently 2 imputation servers exist that allow users to upload genotypes from their GWAS samples, and have imputation carried out remotely and efficiently

<https://imputation.sanger.ac.uk/>

<https://imputationserver.sph.umich.edu>

- A phasing server for high coverage sequenced samples is also available

<https://phasingserver.stats.ox.ac.uk/>

The Haplotype Reference Consortium

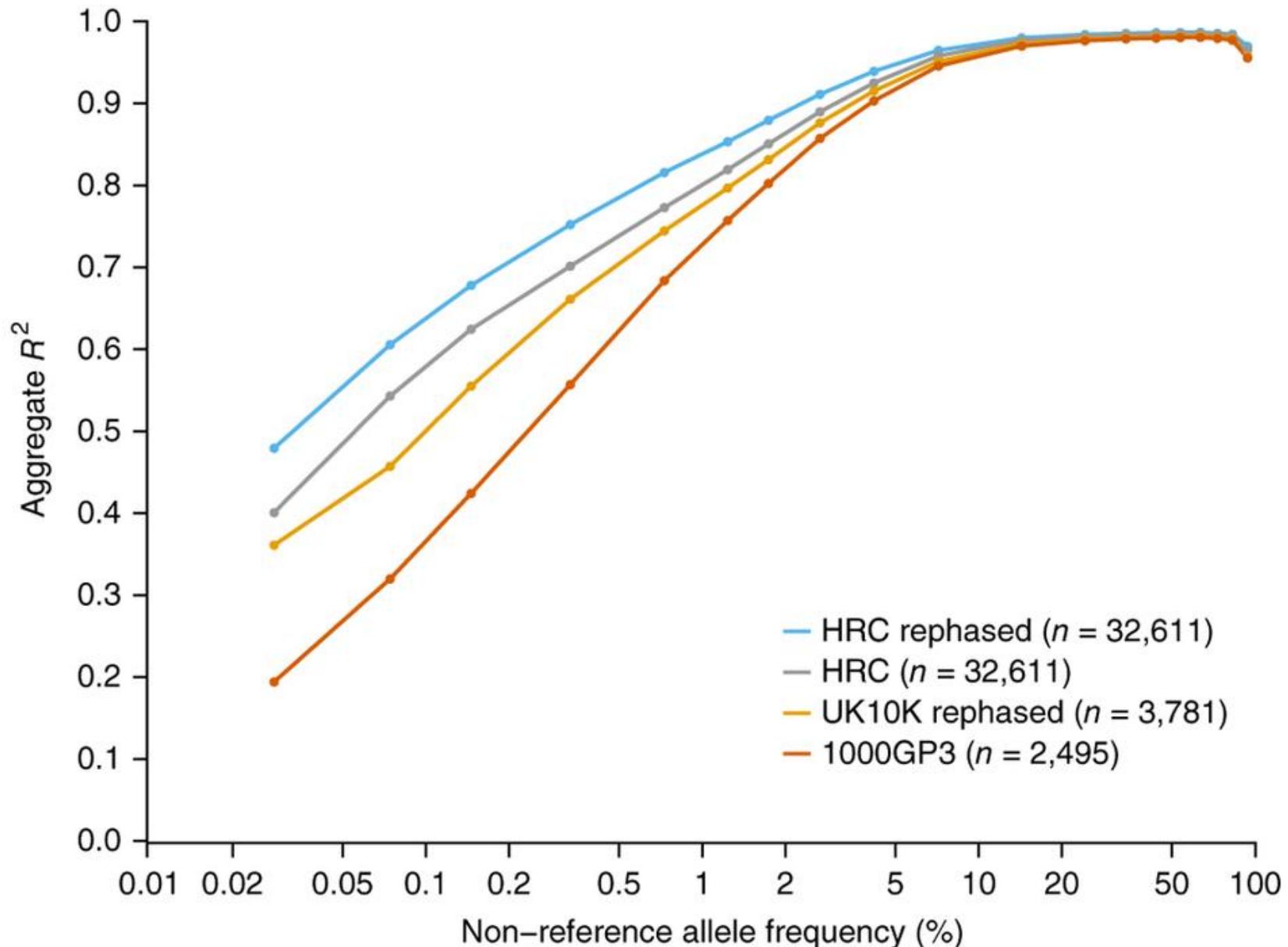
Dataset	Samples	Coverage
IBD	4514	2-4x
UK10K	3781	6.5x
Sardinia	3514	4x
GoT2D	2874	4x + Exome
1000GP Phase 3	2535	4x + Exome
BRIDGES	2489	6-8x
AMD	2099	4x
Finland	1941	4-6x
MCTFR	1339	10x
HUNT	1024	4x
GECCO	954	4-6x
Project MinE	943	45x
GPC	767	30x
GoNL	748	12x
inCHIANTI	680	7x
Orkney	399	4x
Neptune	253	4x
FVG	250	4-10x
MANOLIS	249	4x
Val Borbera	225	6x
	32,488	

Goal : create a European haplotype map of over 50,000+ haplotypes by combining together many low-coverage sequencing studies.

Release 1

64,976 haplotypes
39,235,157 SNPs
estimated MAC ≥ 5

The Haplotype Reference Consortium



Acknowledgments

The afternoon practical