Imputation and haplotyping for genome-wide association analysis

Heather Cordell
Newcastle University

heather.cordell@ncl.ac.uk

Acknowledgement: thanks to Jonathan Marchini who originally created many of these slides

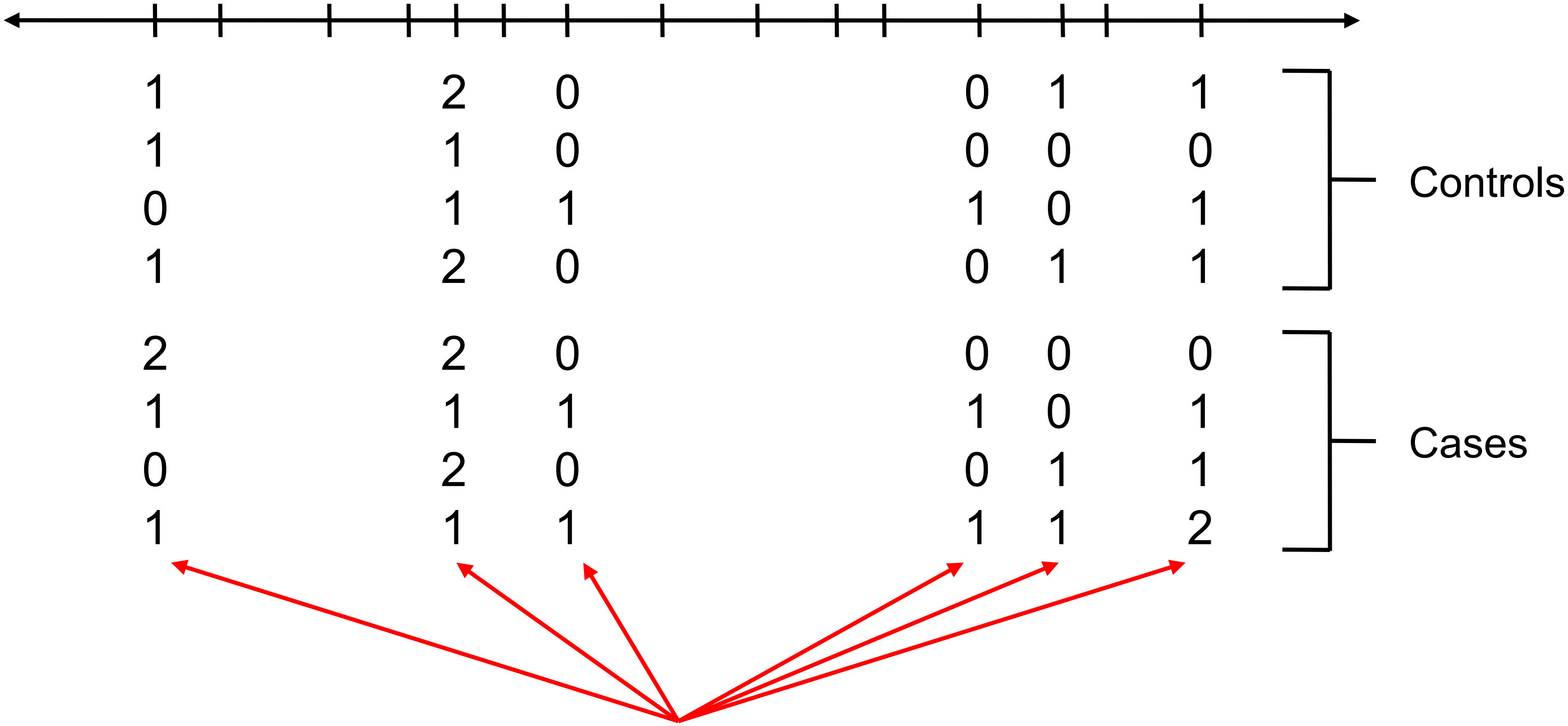
Outline

- The general concept of imputation
- Testing imputed SNPs for association
- Early examples of imputation in the WTCCC
- Using imputed data to empower meta-analysis

- The Strand Issue and Imputation QC metrics
- Properties of imputation - LD, populations, chips, allele frequency

- Pre-phasing and haplotype estimation
- Imputation panels (1000 Genomes, HRC, TOPMed)
- Imputation servers
  - Michigan Imputation Server
  - Sanger Imputation Server
  - TOPMed Imputation Server

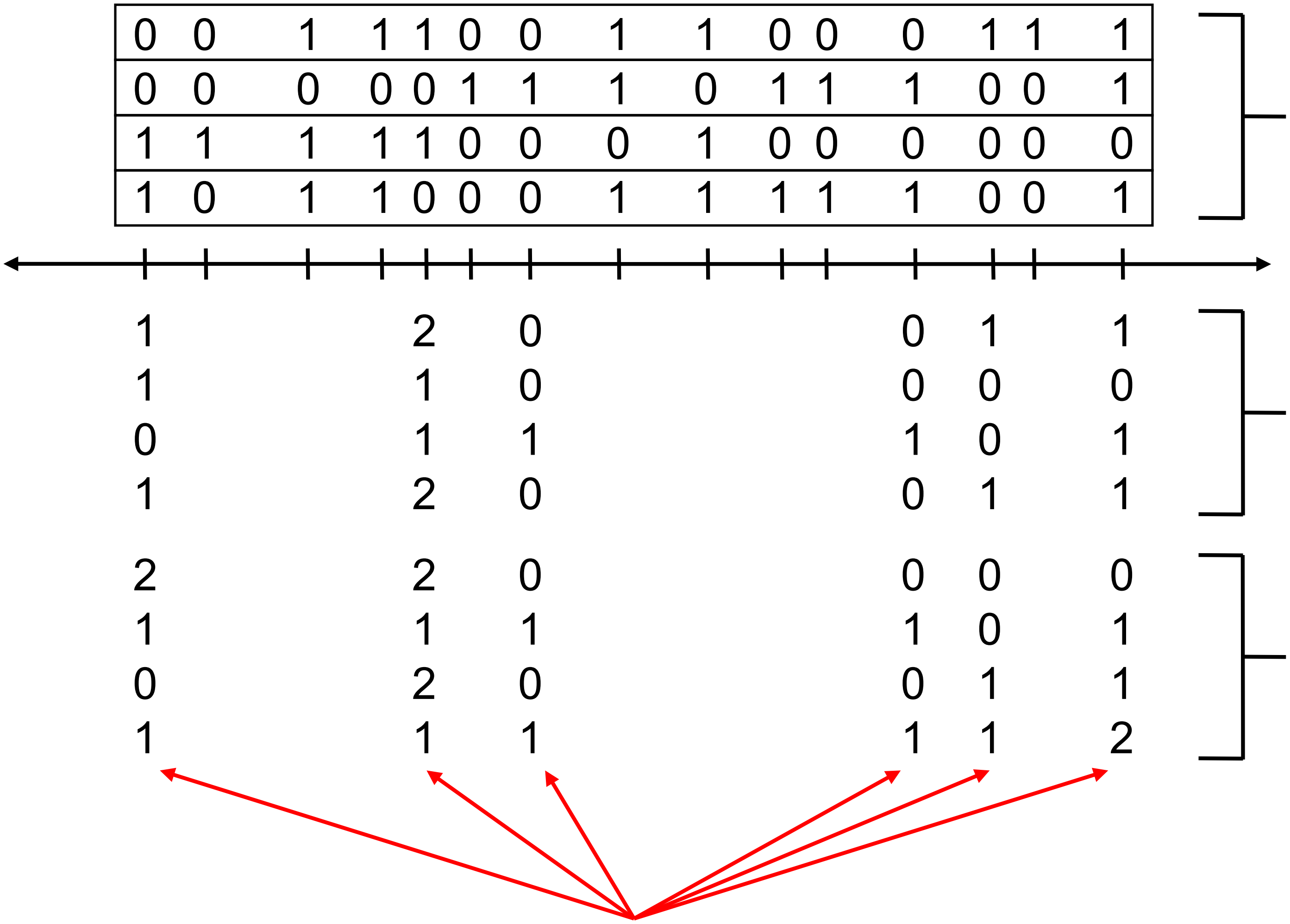
GWAS



1	2	0	0	1	1	Controls
1	1	0	0	0	0	
0	1	1	1	0	1	
1	2	0	0	1	1	
2	2	0	0	0	0	Cases
1	1	1	1	0	1	
0	2	0	0	1	1	
1	1	1	1	1	2	

SNPs genotyped in an association study. ~1M SNPs, 1000+ Cases/Controls

Imputation



Haploid Reference Panel

e.g.HapMap or 1000G or UK10K or HRC

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

Controls

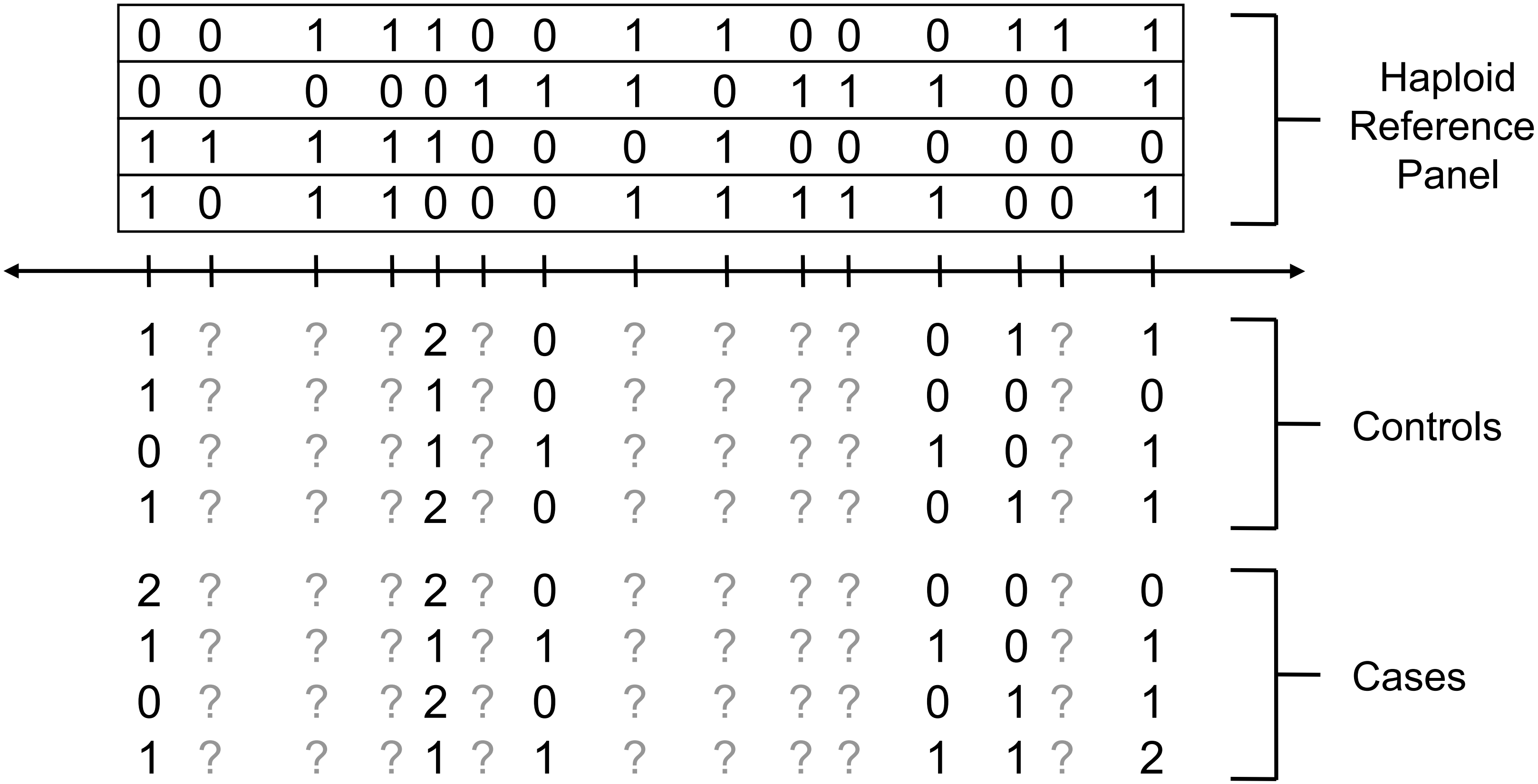
1				2		0					0	1		1
1				1		0					0	0		0
0				1		1					1	0		1
1				2		0					0	1		1

Cases

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

SNPs genotyped in an association study

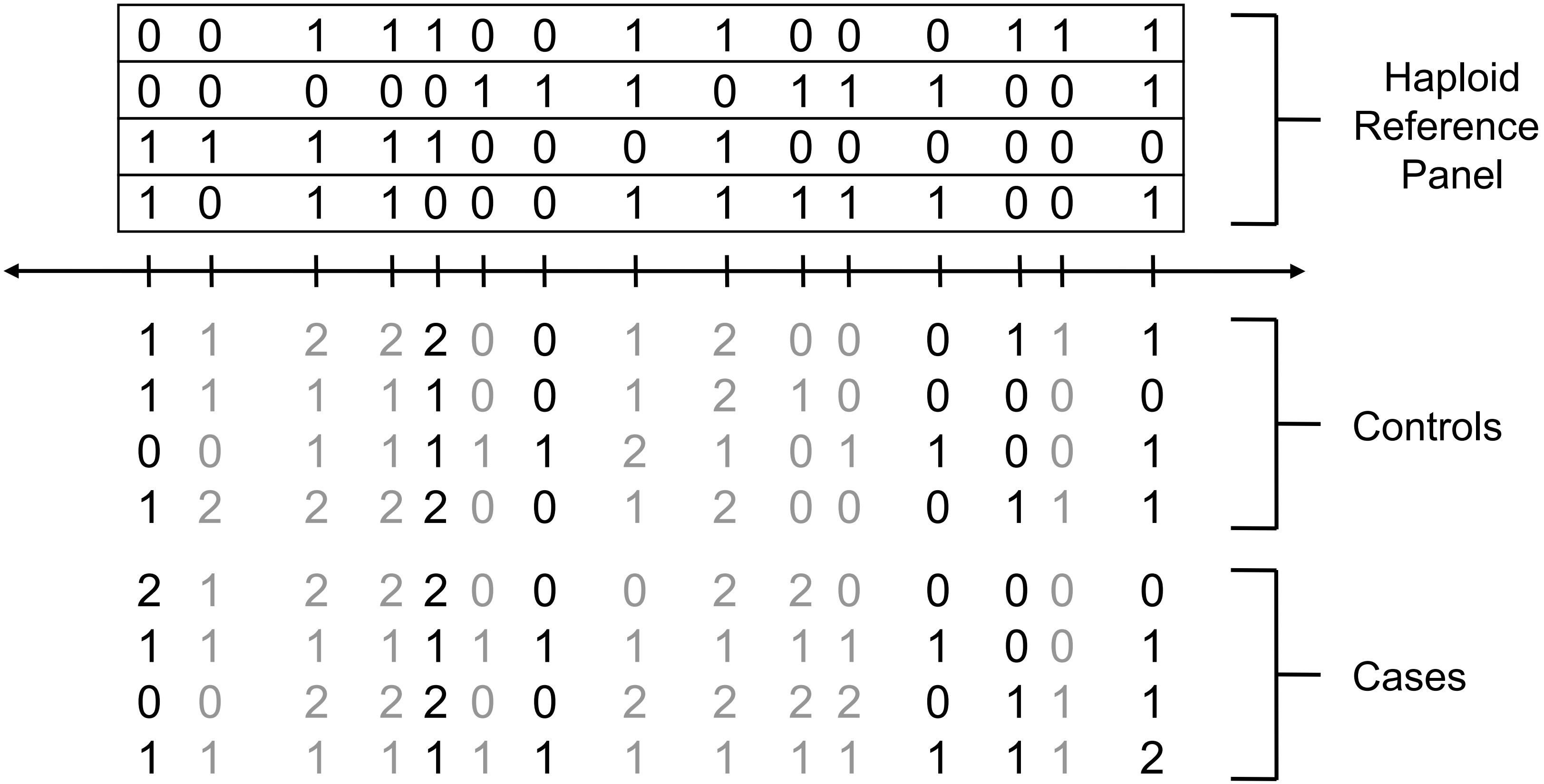
Imputation



0	0	1	1	1	0	0	1	1	0	0	0	1	1	1	Haploid Reference Panel
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1	
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1	
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1	Controls
1	?	?	?	1	?	0	?	?	?	?	0	0	?	0	
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1	
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1	
2	?	?	?	2	?	0	?	?	?	?	0	0	?	0	Cases
1	?	?	?	1	?	1	?	?	?	?	1	0	?	1	
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1	
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2	

Untyped SNPs are treated as missing data.

Imputation



Haploid Reference Panel	0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
	0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
	1	0	1	1	0	0	0	1	1	1	1	1	0	0	1
Controls	1	1	2	2	2	0	0	1	2	0	0	0	1	1	1
	1	1	1	1	1	0	0	1	2	1	0	0	0	0	0
	0	0	1	1	1	1	1	2	1	0	1	1	0	0	1
	1	2	2	2	2	0	0	1	2	0	0	0	1	1	1
Cases	2	1	2	2	2	0	0	0	2	2	0	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1
	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	1	1	2

The goal of imputation is to estimate the missing genotypes.

Imputation

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

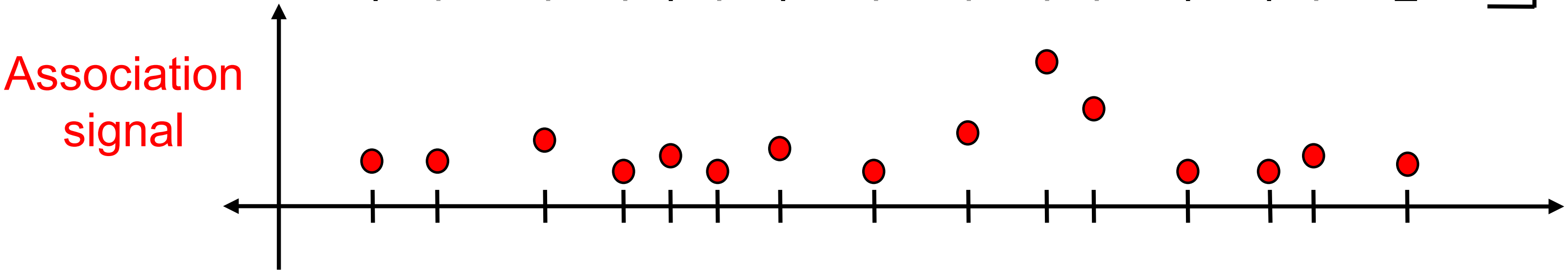
Haploid Reference Panel

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

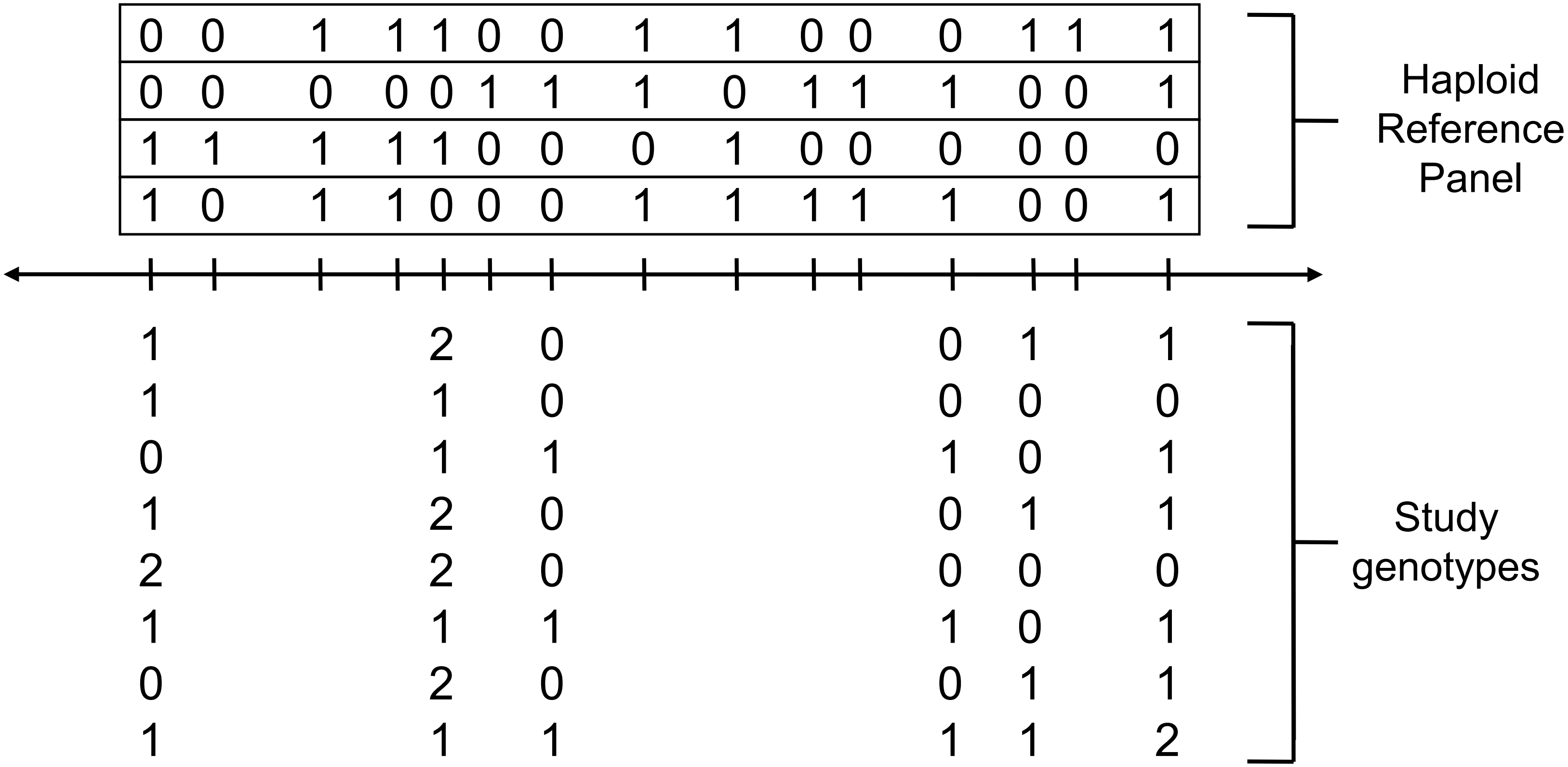
Controls

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

Cases



IMPUTE v1



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

Haploid Reference Panel

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

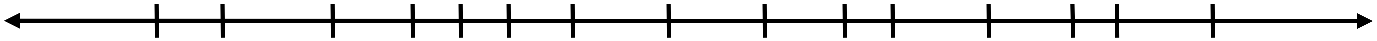
Study genotypes

Marchini et al. (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39:906-913

See also MaCH (Scott et al. 2007 Science 316:1341-5; Li et al. 2010 Genet Epidemiol. 34(8):816-34

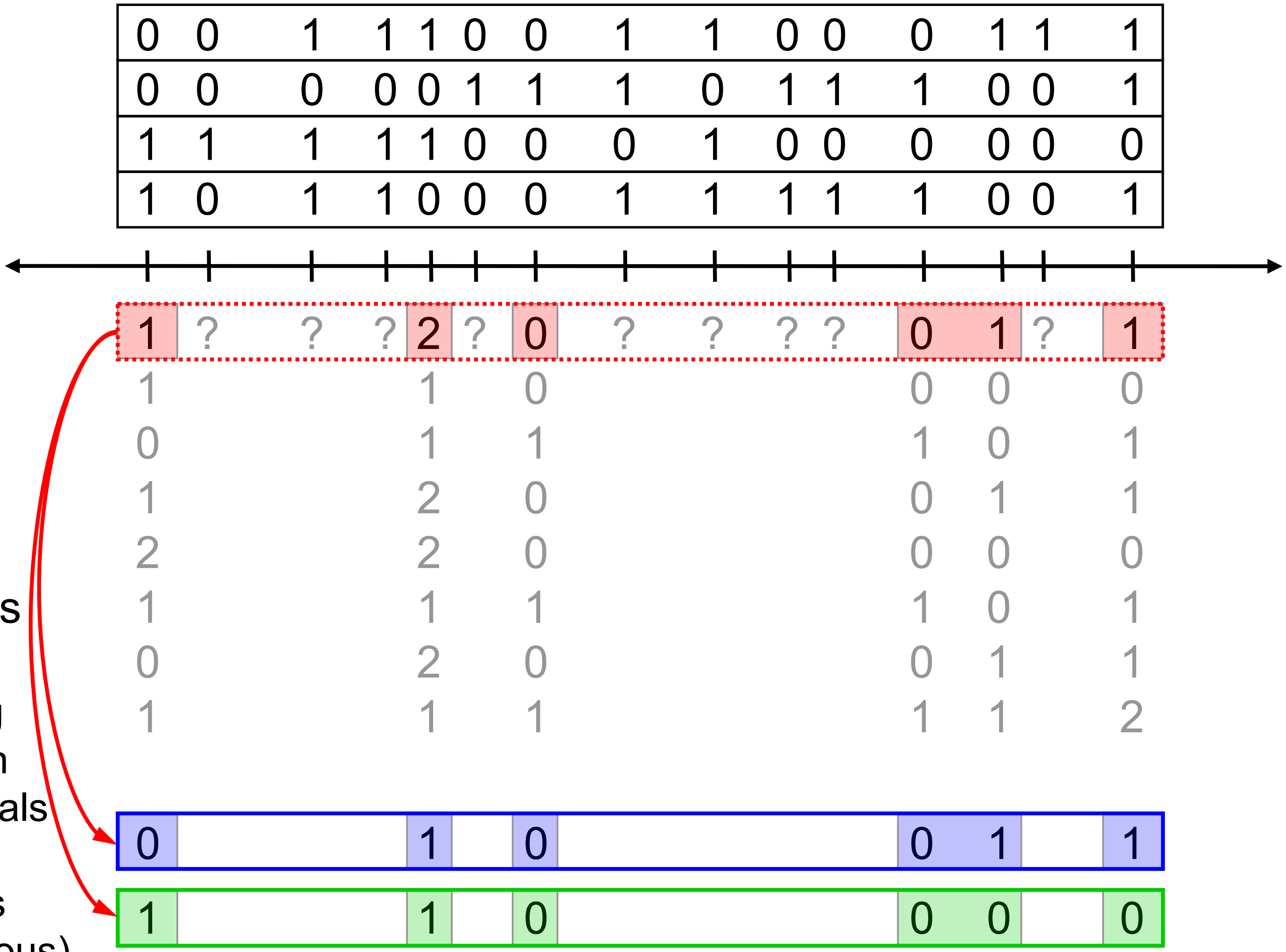
Basic idea

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



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

Basic idea



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

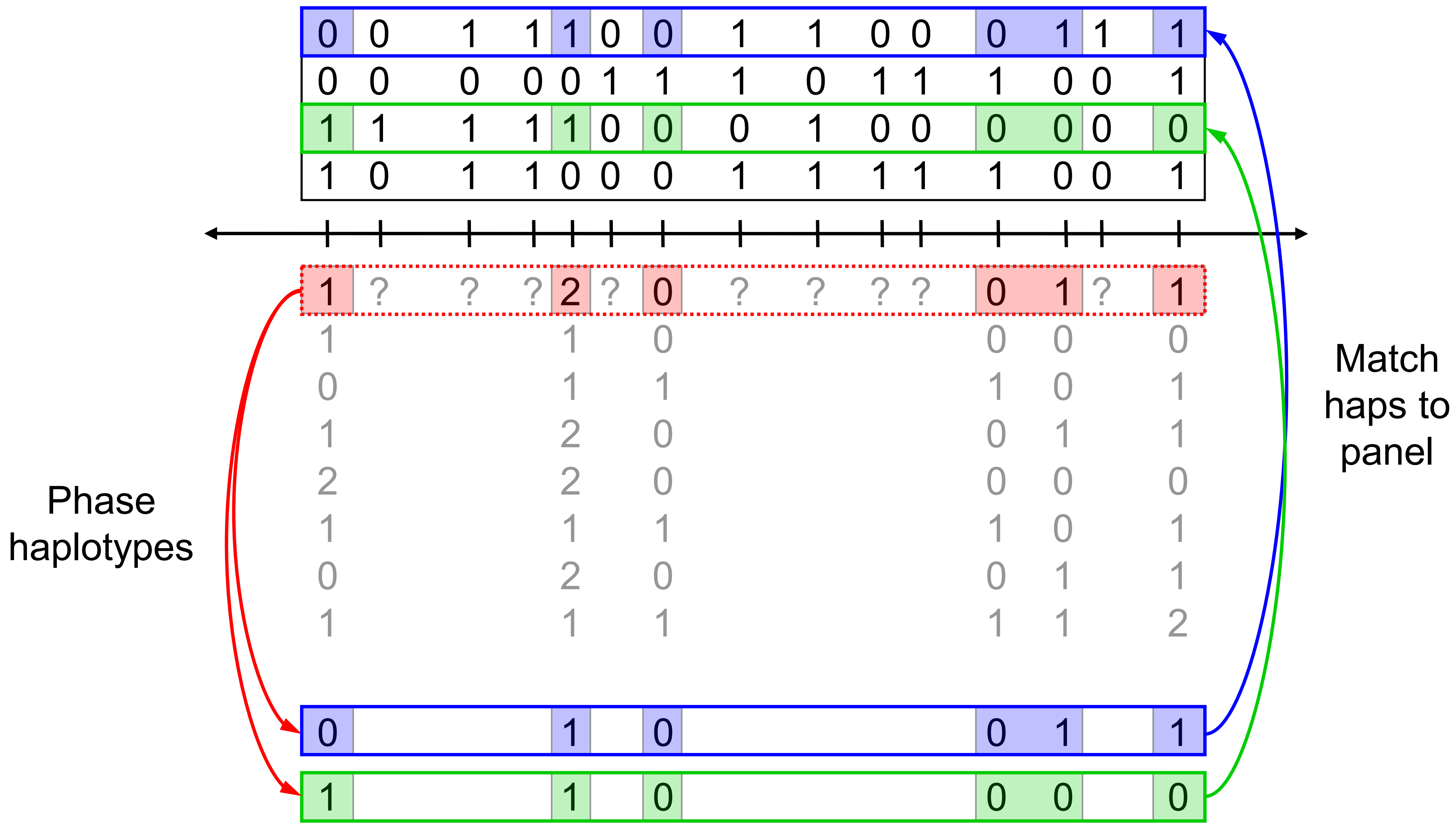
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1				1		0					0	0		0
0				1		1					1	0		1
1				2		0					0	1		1
2				2		0					0	0		0
1				1		1					1	0		1
0				2		0					0	1		1
1				1		1					1	1		2

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

Phase Haplotypes

(borrowing information from individuals whose haplotypes are unambiguous)

Basic idea

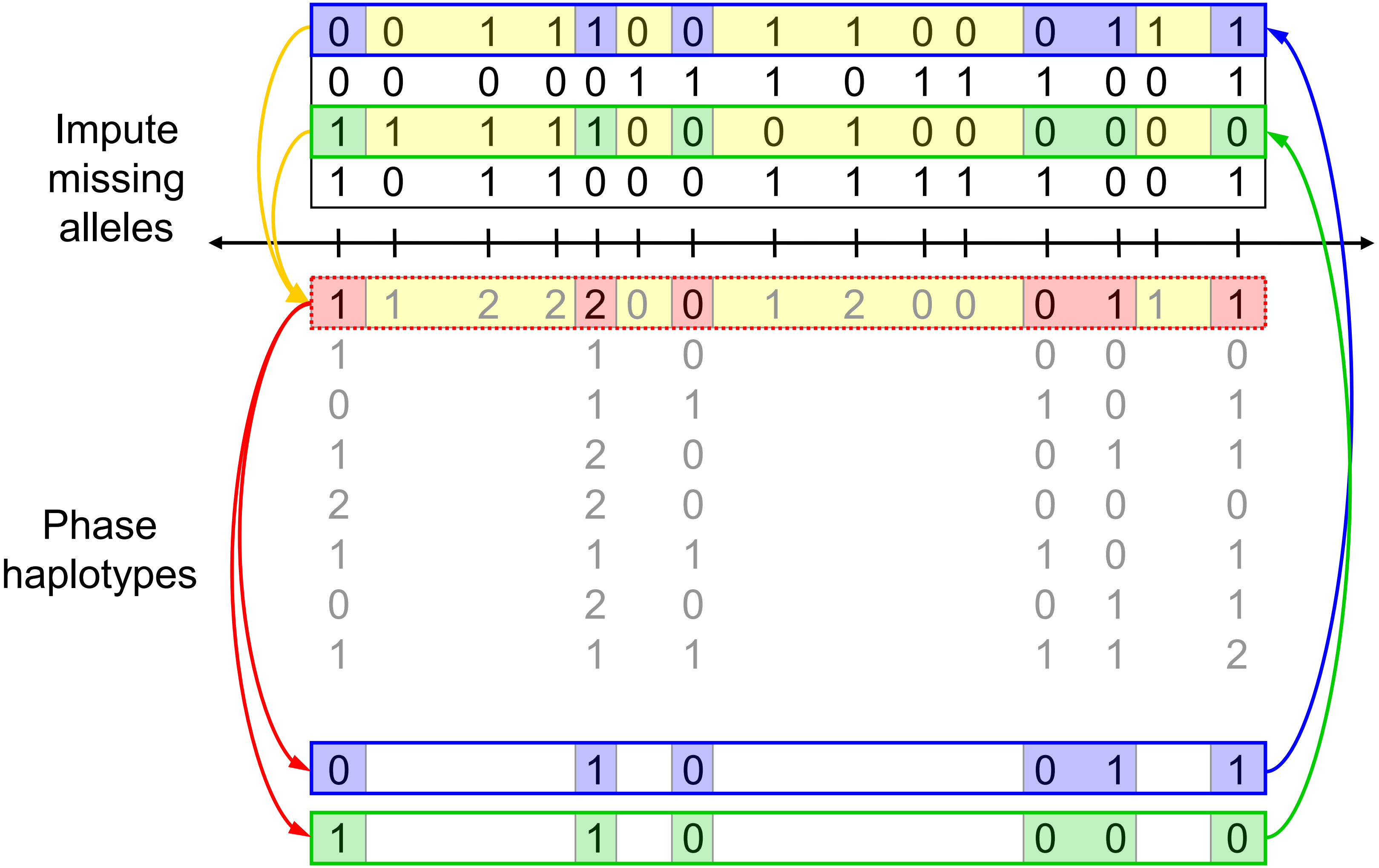


0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1				1		0					0	0		0
0				1		1					1	0		1
1				2		0					0	1		1
2				2		0					0	0		0
1				1		1					1	0		1
0				2		0					0	1		1
1				1		1					1	1		2
0		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

Match haps to panel

Phase haplotypes

Basic idea



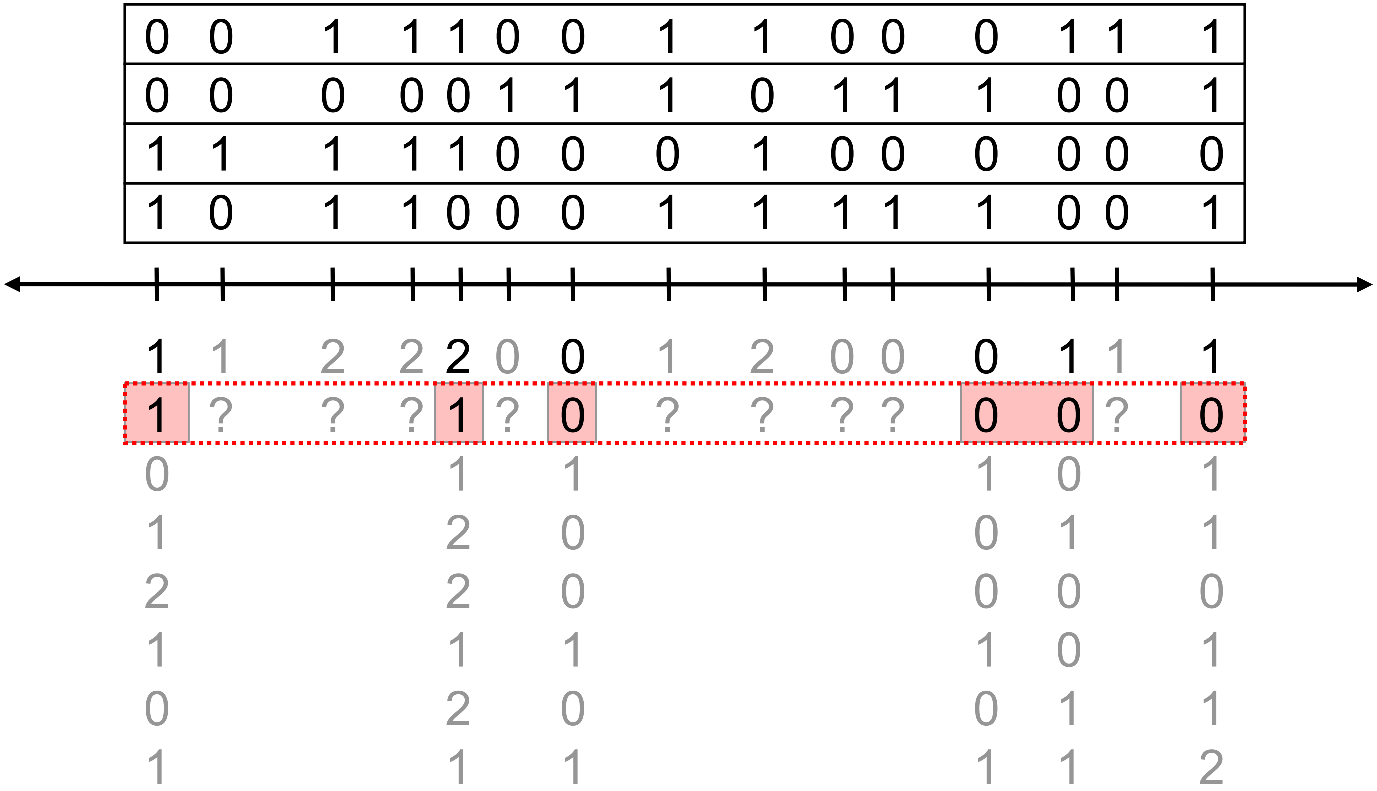
Impute missing alleles

Phase haplotypes

Match haps to panel

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

Basic idea

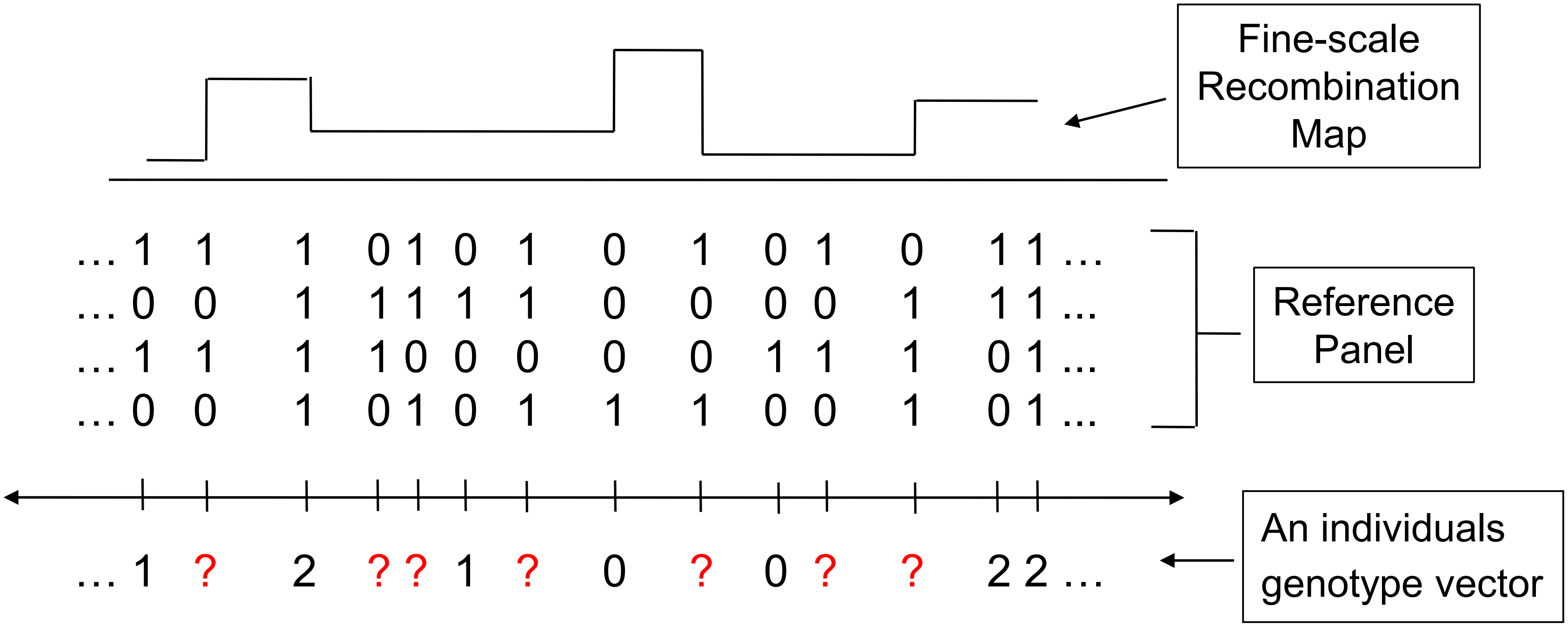


0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

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

Repeat independently for individuals 2,...,*N*

Li and Stephens model



Fine-scale Recombination Map

… 1 1 1 0 1 0 1 0 1 0 1 0 1 1 …

… 0 0 1 1 1 1 1 0 0 0 0 1 1 1 …

… 1 1 1 1 0 0 0 0 0 1 1 1 0 1 …

… 0 0 1 0 1 0 1 1 1 0 0 1 0 1 …

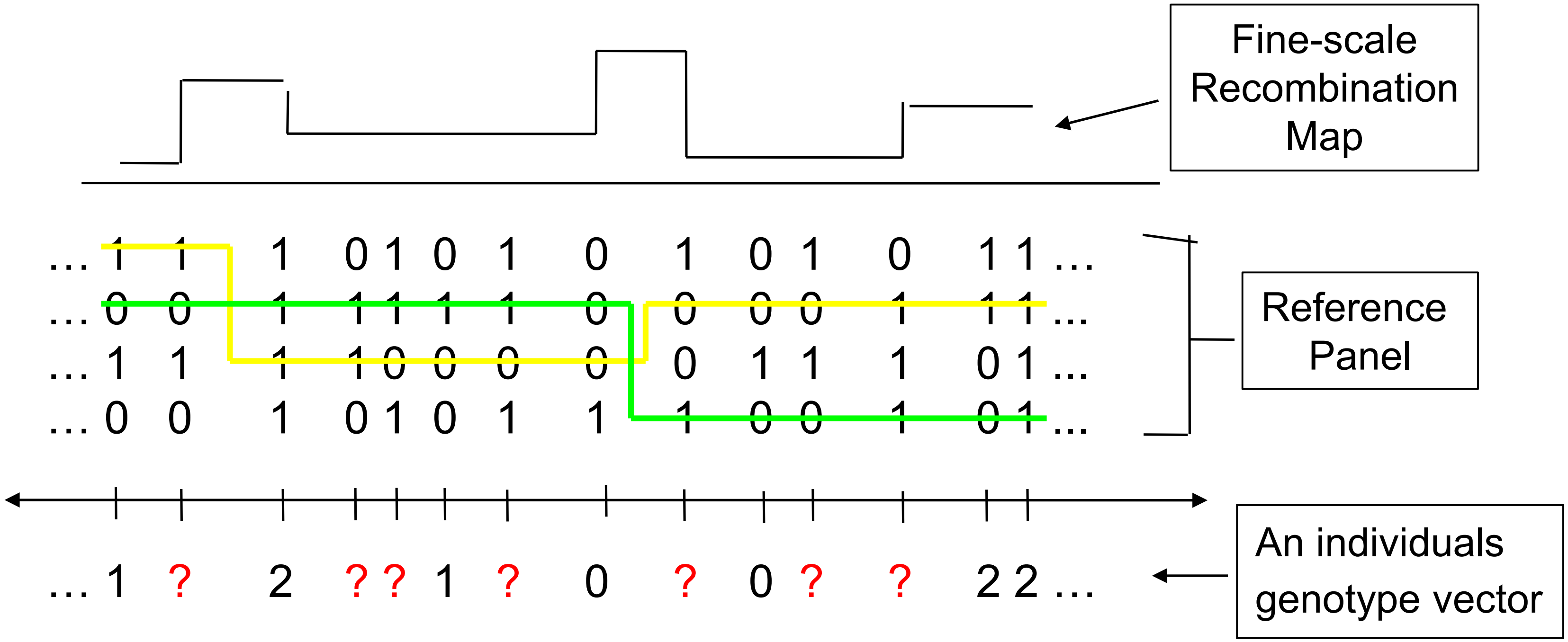
Reference Panel

… 1 ? 2 ? ? 1 ? 0 ? 0 ? ? 2 2 …

An individuals genotype vector

Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data Genetics. 2003 165(4):2213-33

IMPUTE v1



Fine-scale Recombination Map

… 1 1 1 0 1 0 1 0 1 0 1 0 1 1 …
… 0 0 1 1 1 1 1 0 0 0 0 1 1 1 …
… 1 1 1 1 0 0 0 0 0 1 1 1 0 1 …
… 0 0 1 0 1 0 1 1 1 0 0 1 0 1 …

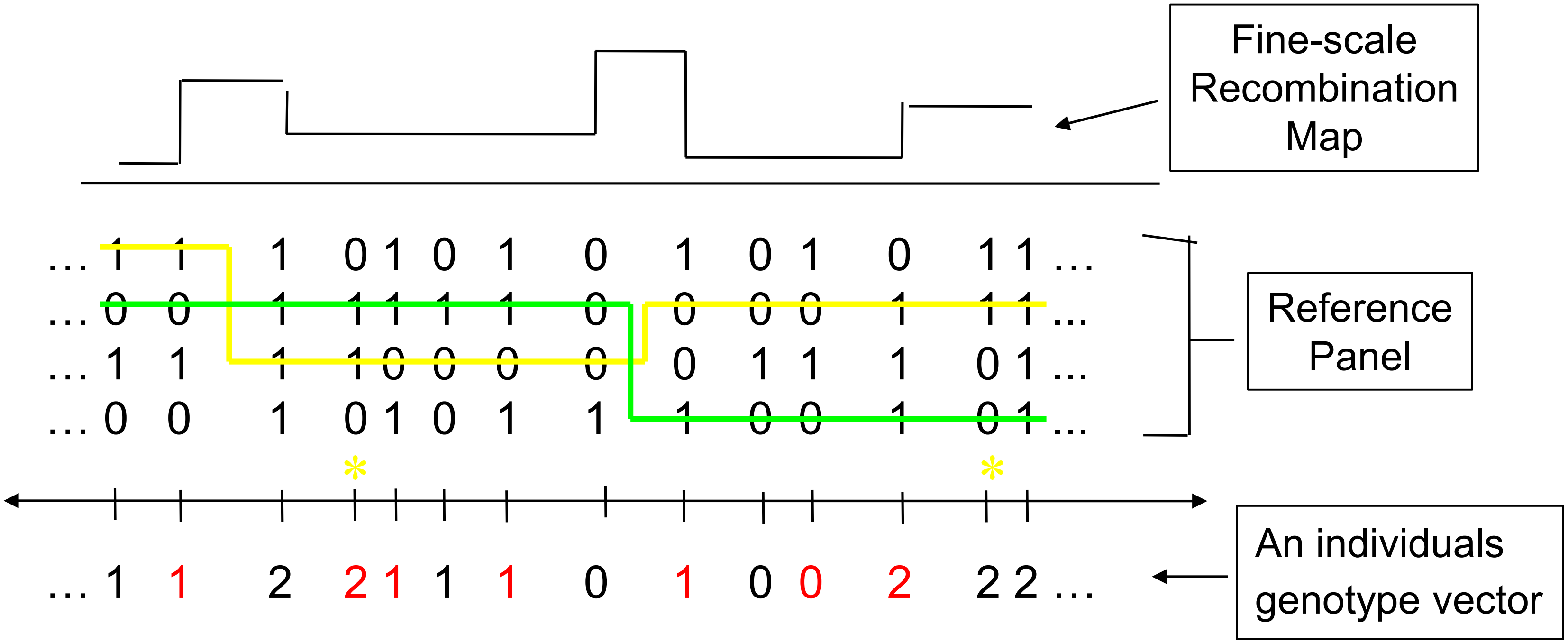
Reference Panel

… 1 ? 2 ? ? 1 ? 0 ? 0 ? ? 2 2 …

An individuals genotype vector

The model says that an individuals 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.

IMPUTE v1



Fine-scale Recombination Map

Reference Panel

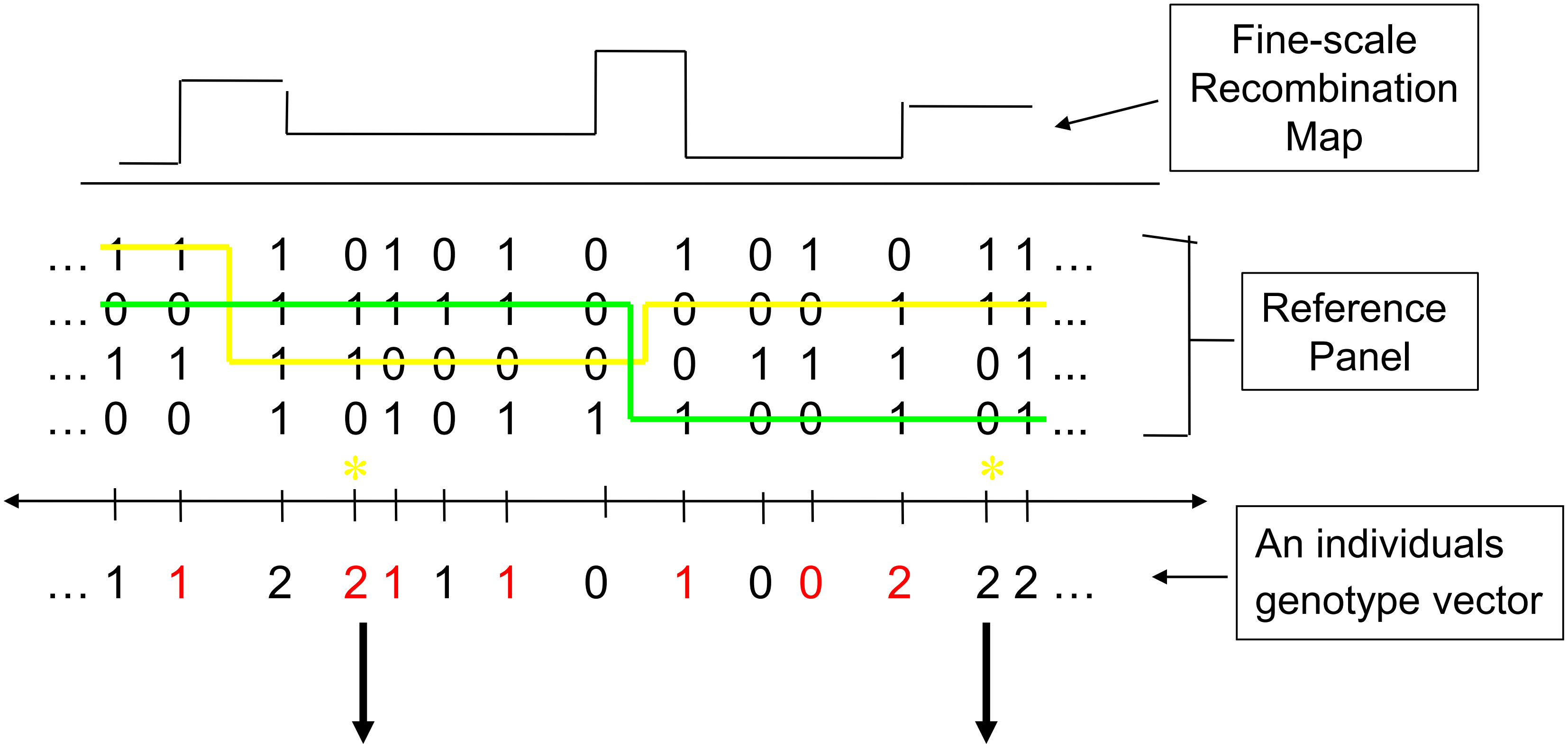
An individuals genotype vector

The model says that an individuals 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.

Paths are sampled **probabilistically**

Generates a **probabilistic** assignment of the underlying genotype vector

IMPUTE v1


Fine-scale Recombination Map

… 1 1 1 0 1 0 1 0 1 0 1 0 1 1 …
… 0 0 1 1 1 1 1 0 0 0 0 1 1 1 …
… 1 1 1 1 0 0 0 0 0 1 1 1 0 1 …
… 0 0 1 0 1 0 1 1 1 0 0 1 0 1 …

Reference Panel

… 1 1 2 2 1 1 1 0 1 0 0 2 2 2 …

An individuals genotype vector

0	1	2
0.01	0.18	0.81

0	1	2
0.15	0.7	0.15

Produces estimates of genotype uncertainty at both untyped and typed genotypes.

Association testing allowing for genotype uncertainty

- There are several ways the imputed genotype probabilities can be used for subsequent association testing

1. Threshold the probability distribution to give "best guess" genotype calls
2. Use the expected allele counts as a "dosage"

AA	AB	BB
0.01	0.18	0.81

→

$$\boxed{0\text{x}0.01 + 1\text{x}0.18 + 2\text{x}0.81 = 1.8}$$

3. Average over the uncertainty
   - Can be done in both the Frequentist and Bayesian frameworks

- All these tests are implemented in the package SNPTEST

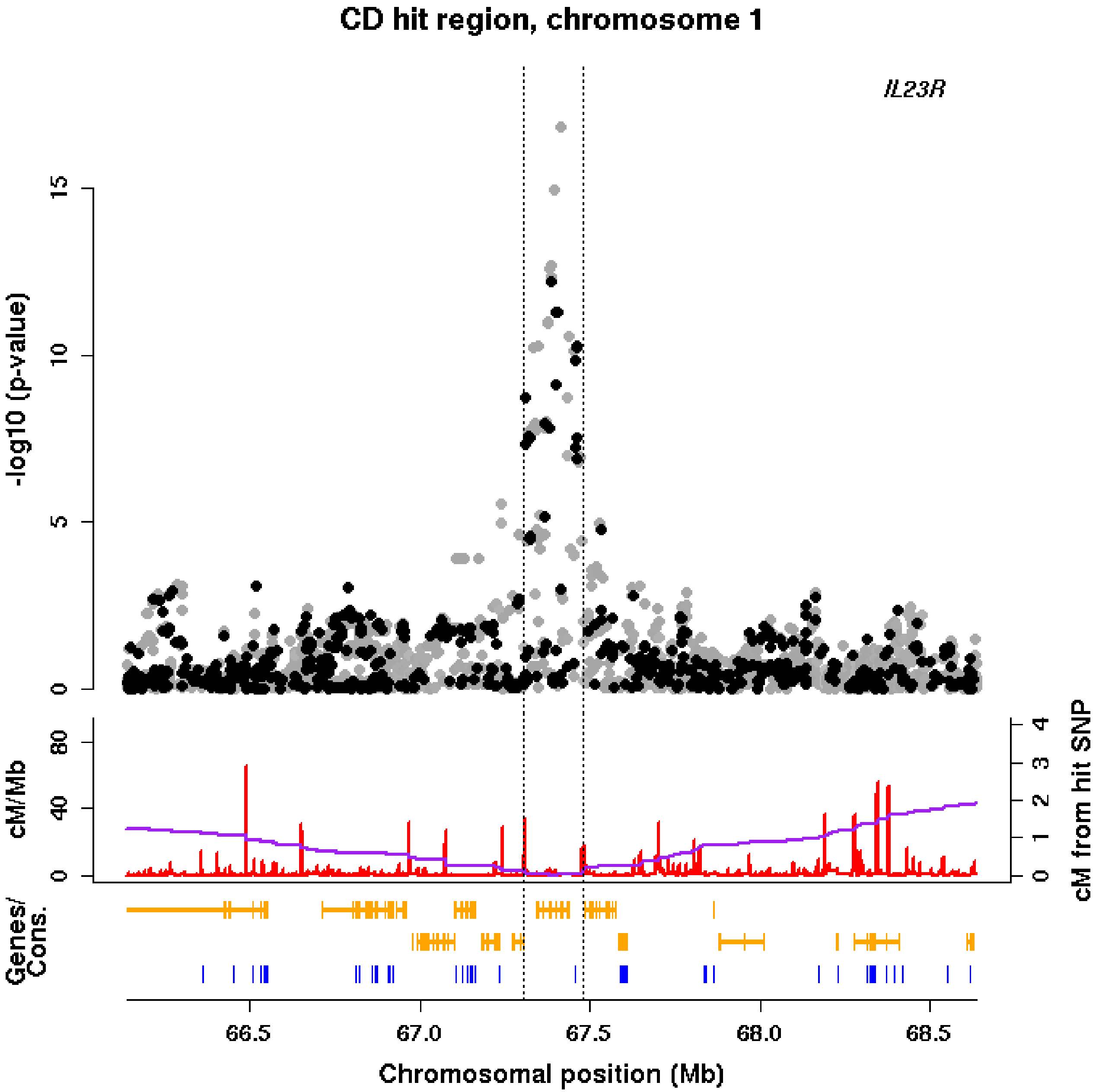


Image from The Wellcome Trust Case Control Consortium (2007) Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447;661-78. DOI: 10.1038/nature05911.

T1D hit region, chromosome 12

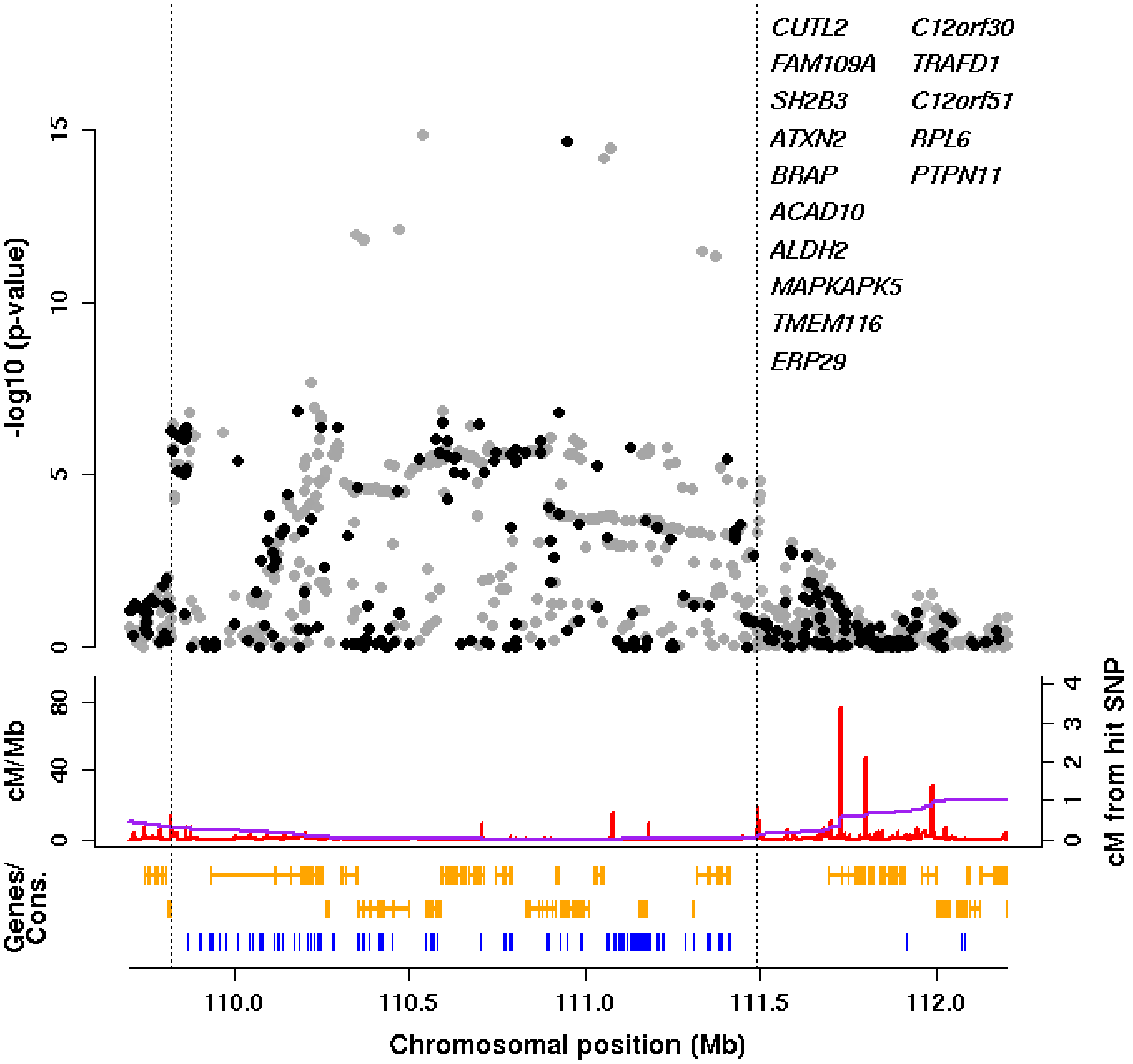
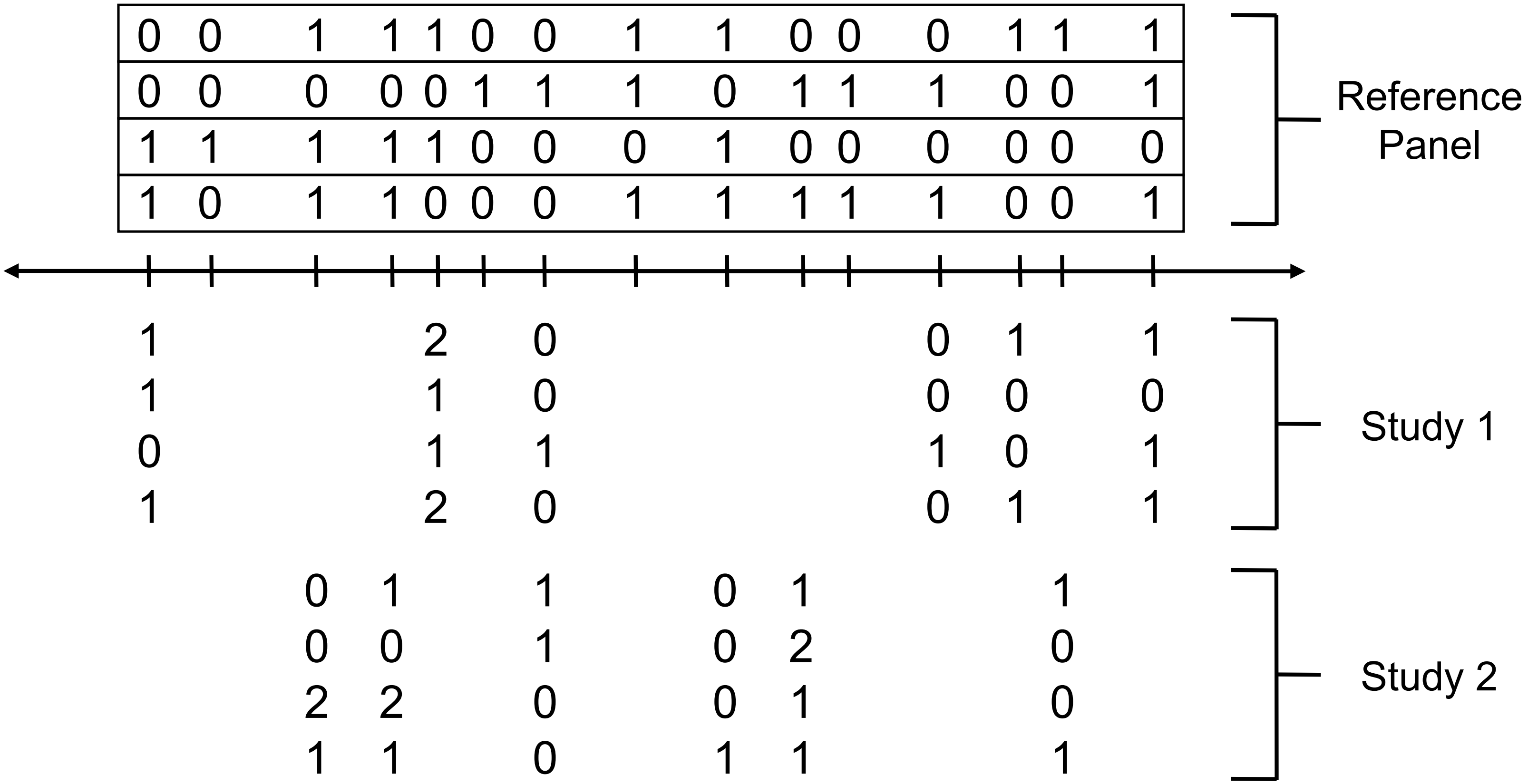


Image from The Wellcome Trust Case Control Consortium (2007) Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447;661-78. DOI: 10.1038/nature05911.

Meta-analysis



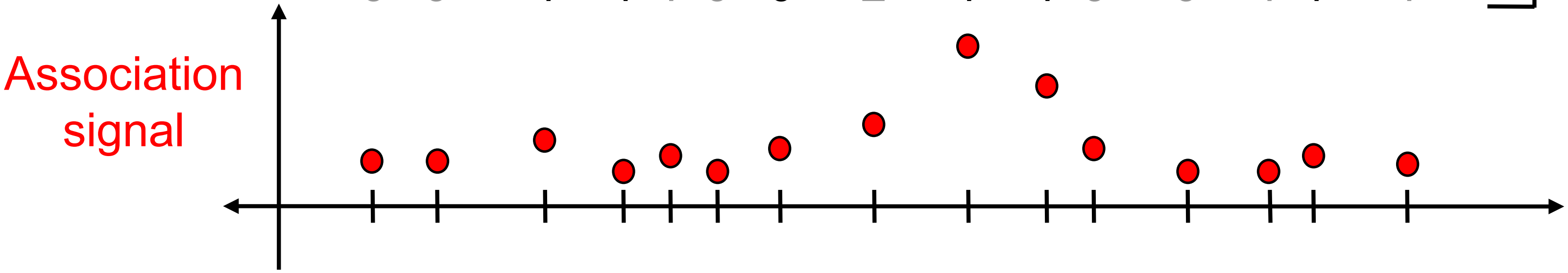
Reference Panel	0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
	0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
	1	0	1	1	0	0	0	1	1	1	1	1	0	0	1
Study 1	1				2		0					0	1		1
	1				1		0					0	0		0
	0				1		1					1	0		1
	1				2		0					0	1		1
Study 2			0	1			1		0	1				1	
			0	0			1		0	2				0	
			2	2			0		0	1				0	
			1	1			0		1	1				1	

Meta-analysis

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1	HapMap Phase II
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1	
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1	

1	1	2	2	2	0	0	1	1	2	0	0	1	1	1	Study 1
1	1	1	1	1	0	0	1	0	1	0	0	0	0	0	
0	0	1	1	1	1	1	2	1	1	1	1	0	0	1	
1	2	2	2	2	0	0	1	0	1	0	0	1	1	1	

0	0	0	1	1	1	1	2	0	1	1	1	1	1	2	Study 2
0	0	0	0	0	1	1	2	0	2	1	1	0	0	1	
1	1	2	2	1	0	0	1	0	1	1	1	0	0	1	
0	0	1	1	1	0	0	2	1	1	0	0	1	1	1	



Meta-analysis

- Beware of imputing cases and controls typed on different chips separately – this can induce spurious associations due to imputation artefacts

  - Need to first cut down the SNPs for each study to those in common between the case and control groups

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.

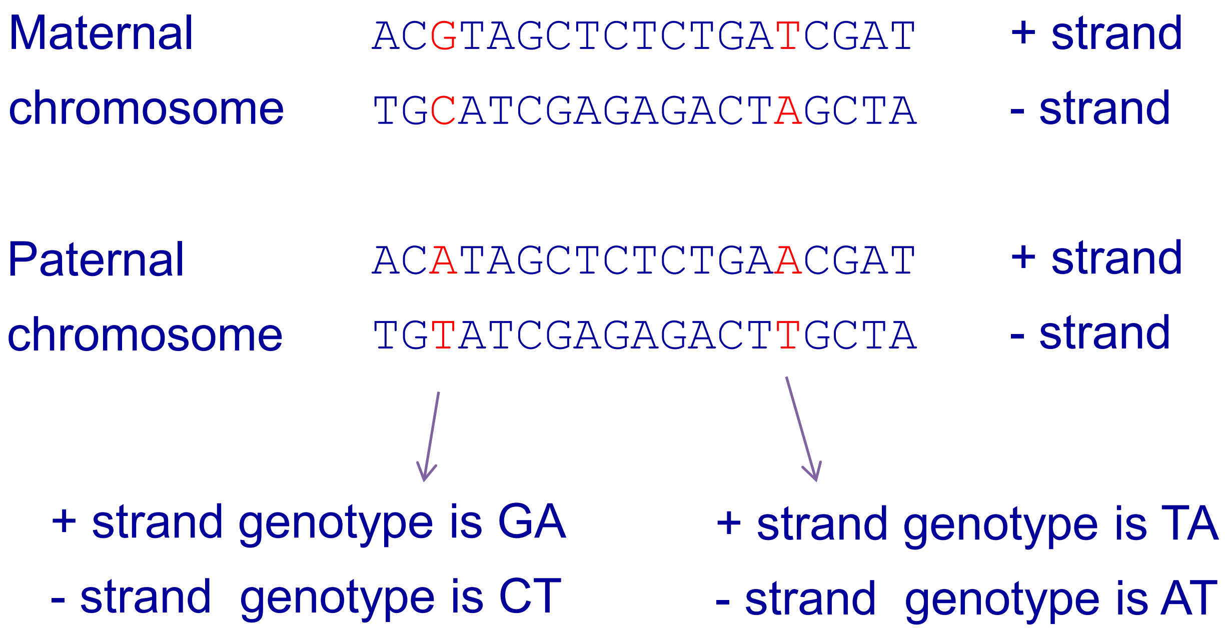
So it is very important that the genotypes of your GWAS are mapped to the same genome build as the reference panel you are using.

Currently, most commonly-used reference panels use build 37 (GRCh37).

- Some have now been updated to build 38
- Some imputation servers (e.g. Michigan and TOPMed) allow you to submit your genotypes in either build 37 or build 38.

The Strand Issue

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



Maternal chromosome

ACGTAGCTCTCTGATCGAT + strand

TGCATCGAGAGACTAGCTA - strand

Paternal chromosome

ACATAGCTCTCTGAACGAT + strand

TGTATCGAGAGACTTGCTA - strand

+ strand genotype is GA

- strand genotype is CT

+ strand genotype is TA

- strand genotype is AT

Haplotype reference panels usually 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 (strand of study sample genotypes **needs to match** the reference panel).

The strand info for many chips can be found at http://www.well.ox.ac.uk/~wrayner/strand/

Information metrics

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 estimated 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.

In many published studies (especially those that have used imputation for meta-analysis) SNPs with info score <0.3 (or 0.5, or even 0.8) are **excluded.**

Similar metrics are produced by other imputation packages (e.g. MaCH produces an Rsq measure)

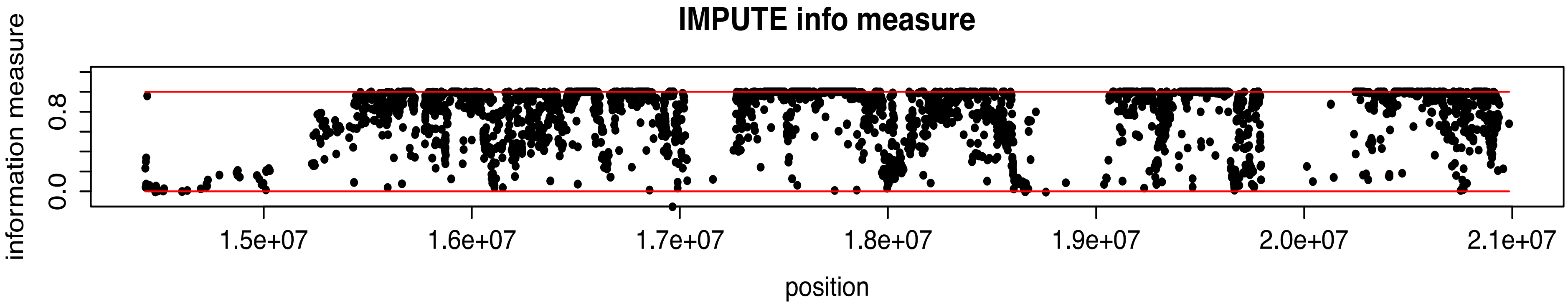
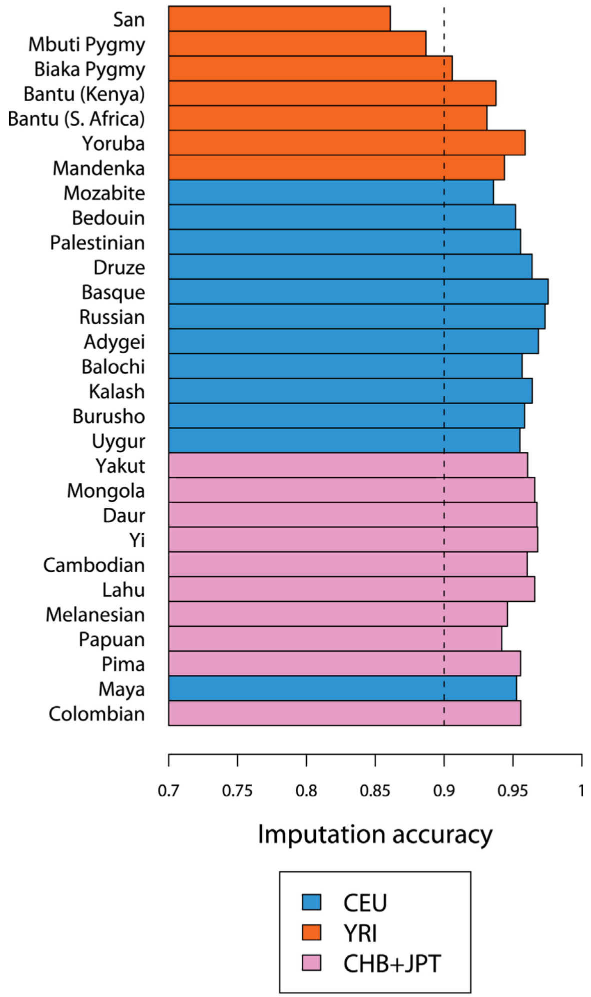


Image from J. Marchini and B. Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics doi:10.1038/nrg2796

Factors affecting accuracy : LD and ancestry

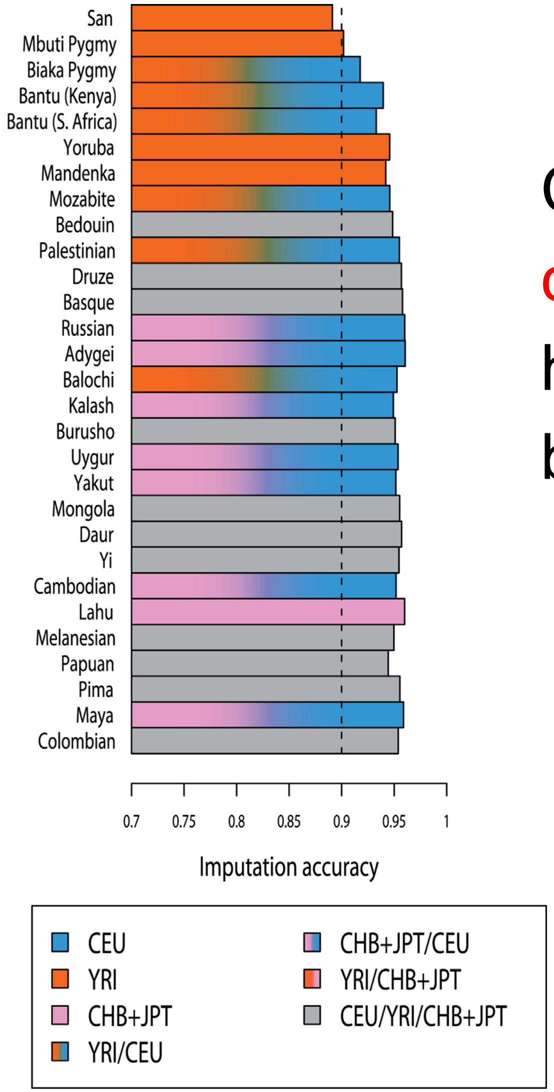


The figure shows the Maximal Imputation Accuracy Achieved by One of the Three HapMap Reference Panels, in Each of 29 Populations.

“African populations, whose levels of LD were generally quite similar, varied considerably in imputation accuracy, with the highest values occurring in the lower-LD Yoruba population and the lowest values occurring in the higher-LD Mbuti Pygmy and San populations. Instead of being highest for populations from the Americas and Oceania, who exhibit the highest LD levels, **imputation accuracy was highest in most analyses for European and East Asian populations that are closely related to populations from the reference panels.**”

Image from from Huang et al. (2009) *Genotype-Imputation Accuracy across Worldwide Human Populations*. 84, 235-250

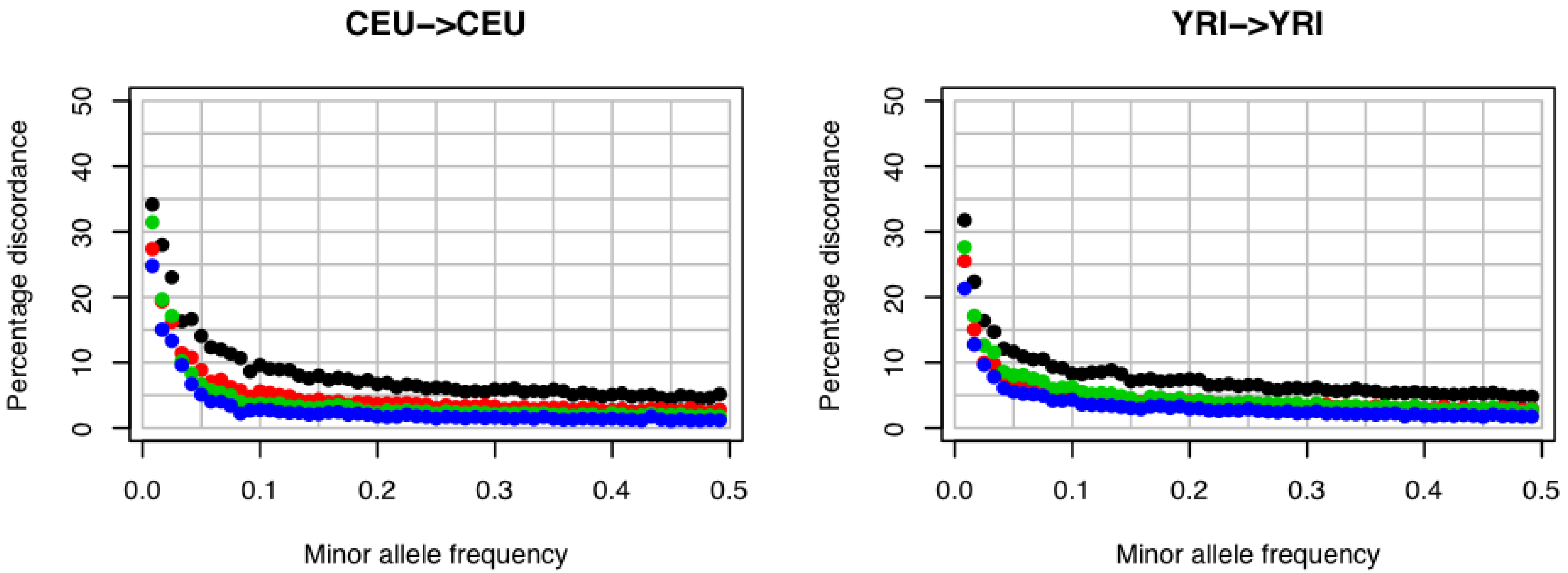
Factors affecting accuracy : LD and ancestry



Carrying out imputation based on combinations of reference sets of haplotypes can (in some cases) boost performance.

Image from from Huang et al. (2009) *Genotype-Imputation Accuracy across Worldwide Human Populations*. 84, 235-250

Factors affecting accuracy : Chip (density) and MAF



Genotyping chip : chips based on tag SNPs can help if used in the same population used to construct them. Random sets of markers work equally well across different populations.

Allele frequency : imputation of rare alleles is more difficult.

⬛ Affy 500k

🟥 Affy 6.0

🟩 Illm 670

🟦 Illm 1M

Image from J. Marchini and B. Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics

Factors affecting accuracy : Reference panel

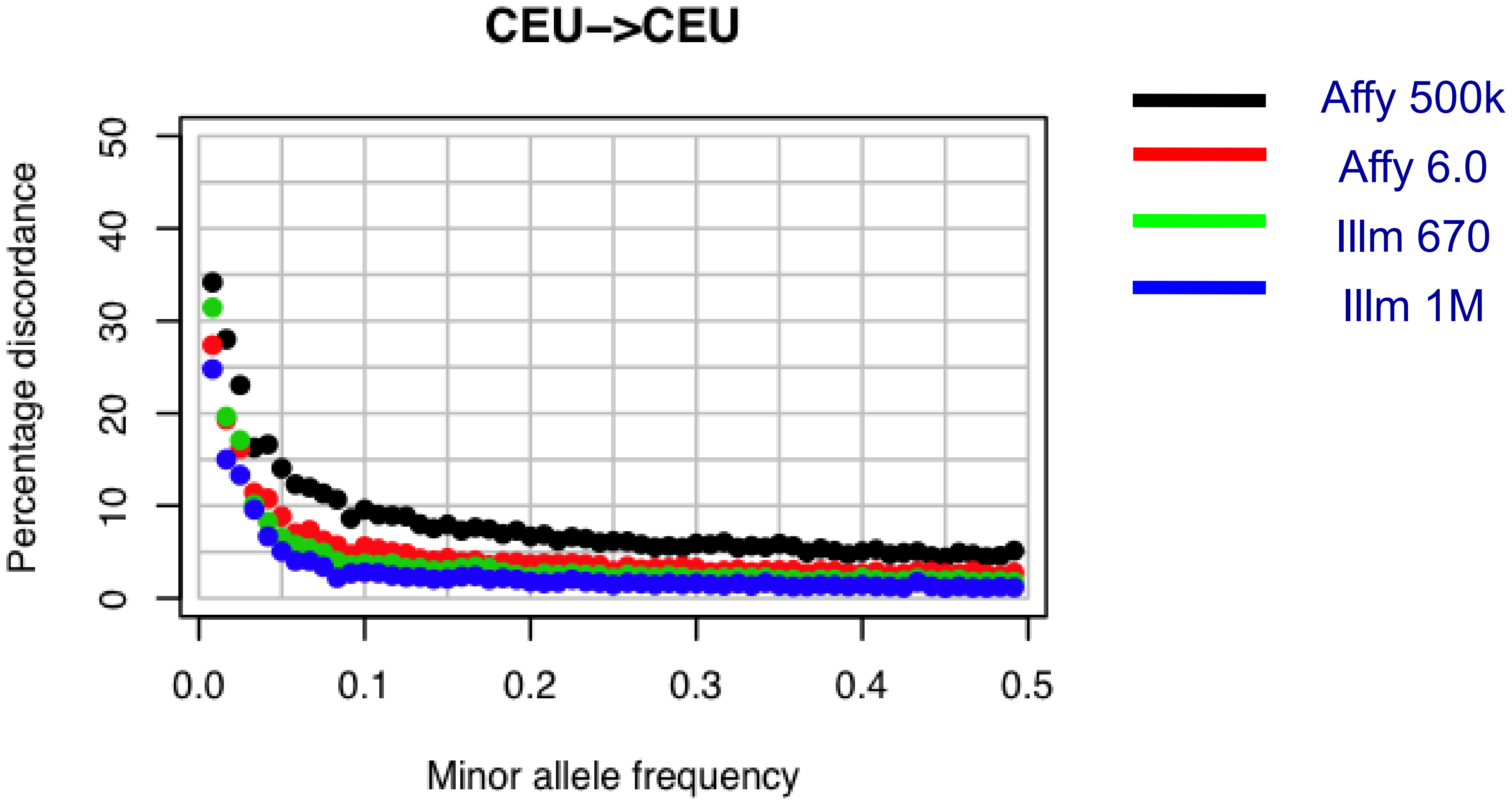
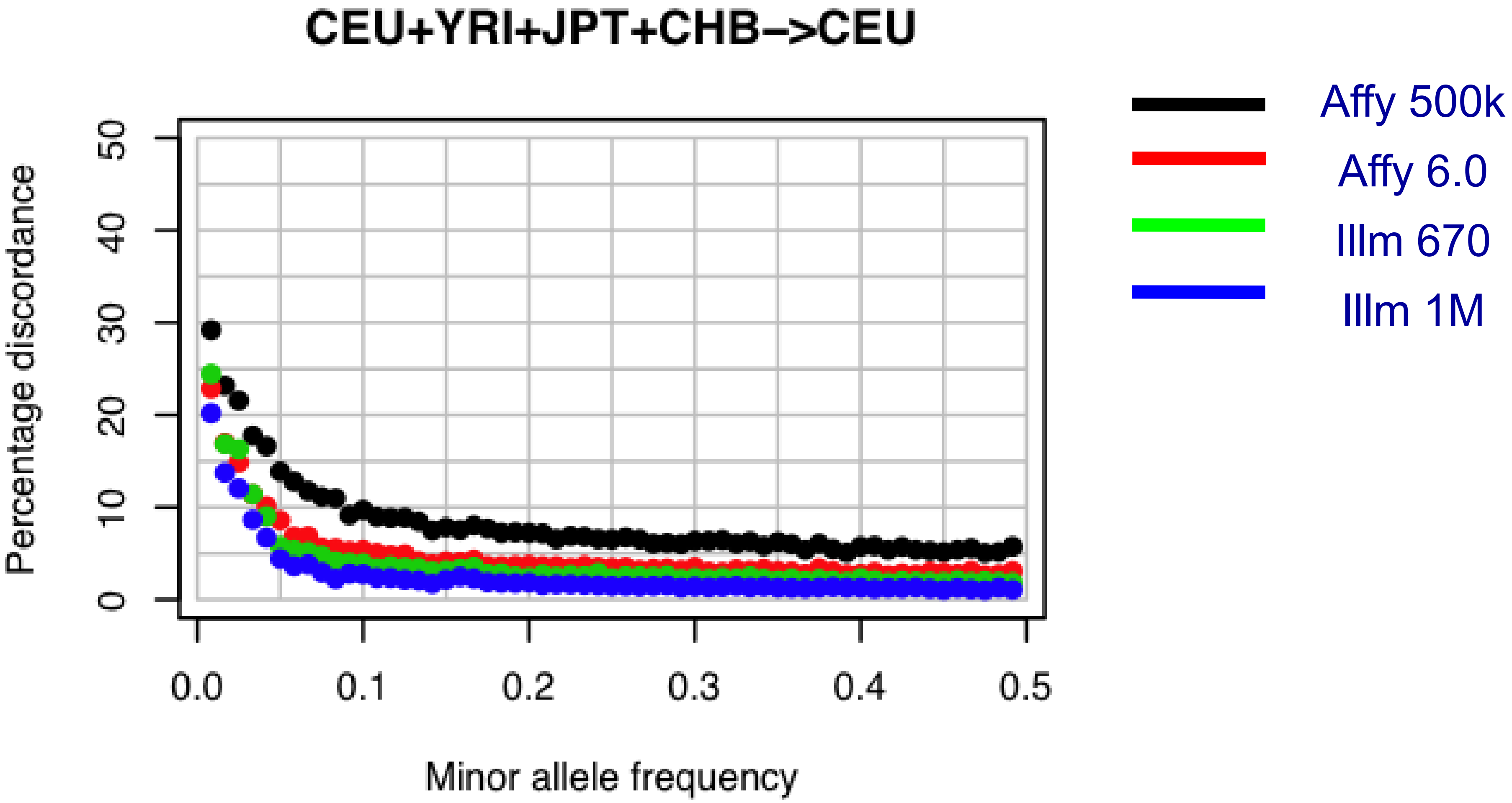


Image from J. Marchini and B. Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics

Factors affecting accuracy : Reference panel



A larger more diverse panel improves accuracy

Image from J. Marchini and B. Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics

Pre-phasing

Imputation is much faster if the GWAS samples are phased before imputation.

Phasing the GWAS samples takes about the same time as one imputation run.

Most imputation servers will do the pre-phasing for you.

Haplotype estimation for phasing/pre-phasing

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

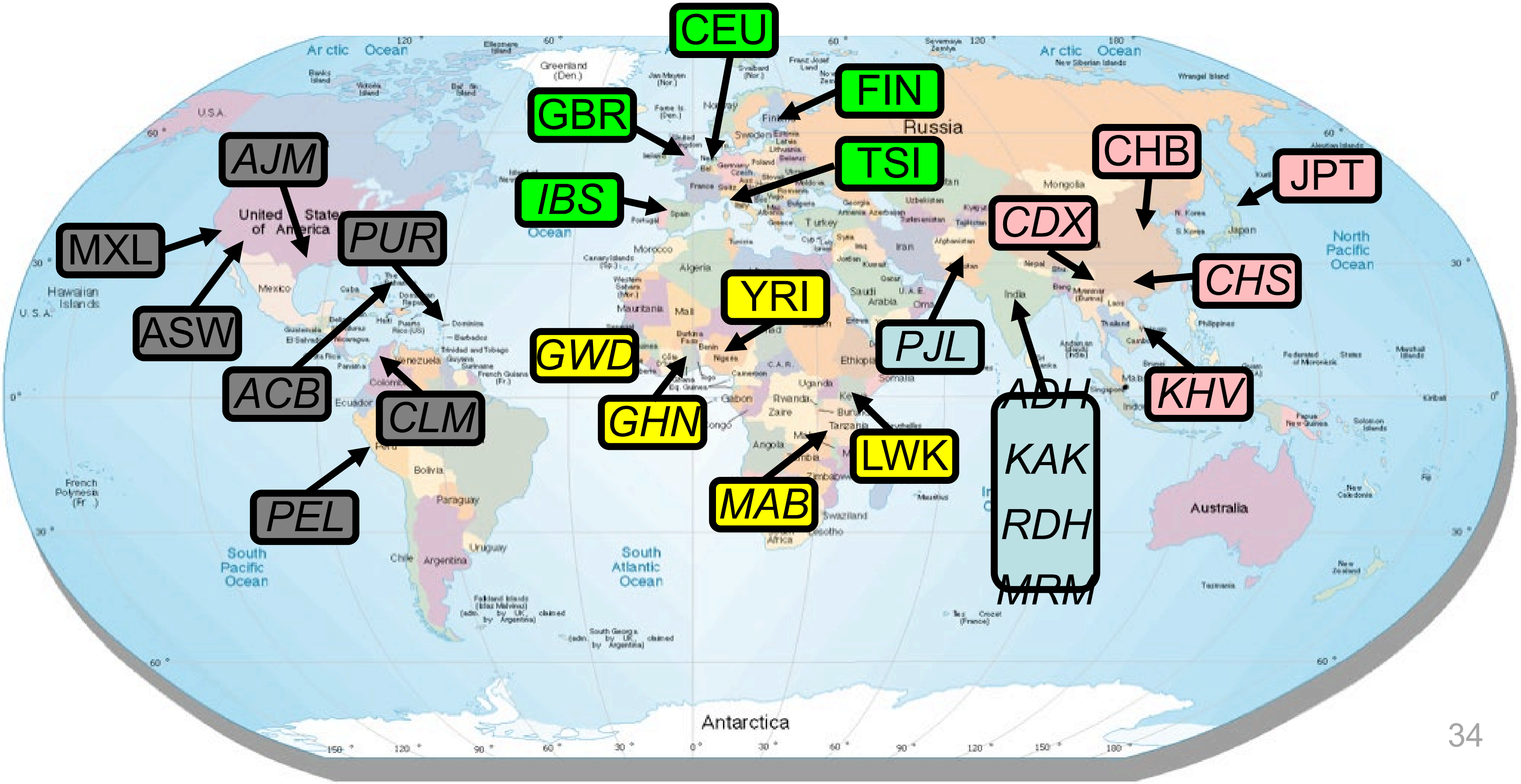
SHAPEIT2 - http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

EAGLE - https://www.hsph.harvard.edu/alkes-price/software/

1000 Genomes Project

The 1000 Genomes Project dataset constructed a reference panel ~5000 haplotypes of SNPs, indels and structural variants from 26 populations.

This reference panel has been widely used for imputation and is freely available at http://www.1000genomes.org/



The Haplotype Reference Consortium (HRC)

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

McCarthy et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 10.1038/ng.3643

Downstream imputation accuracy

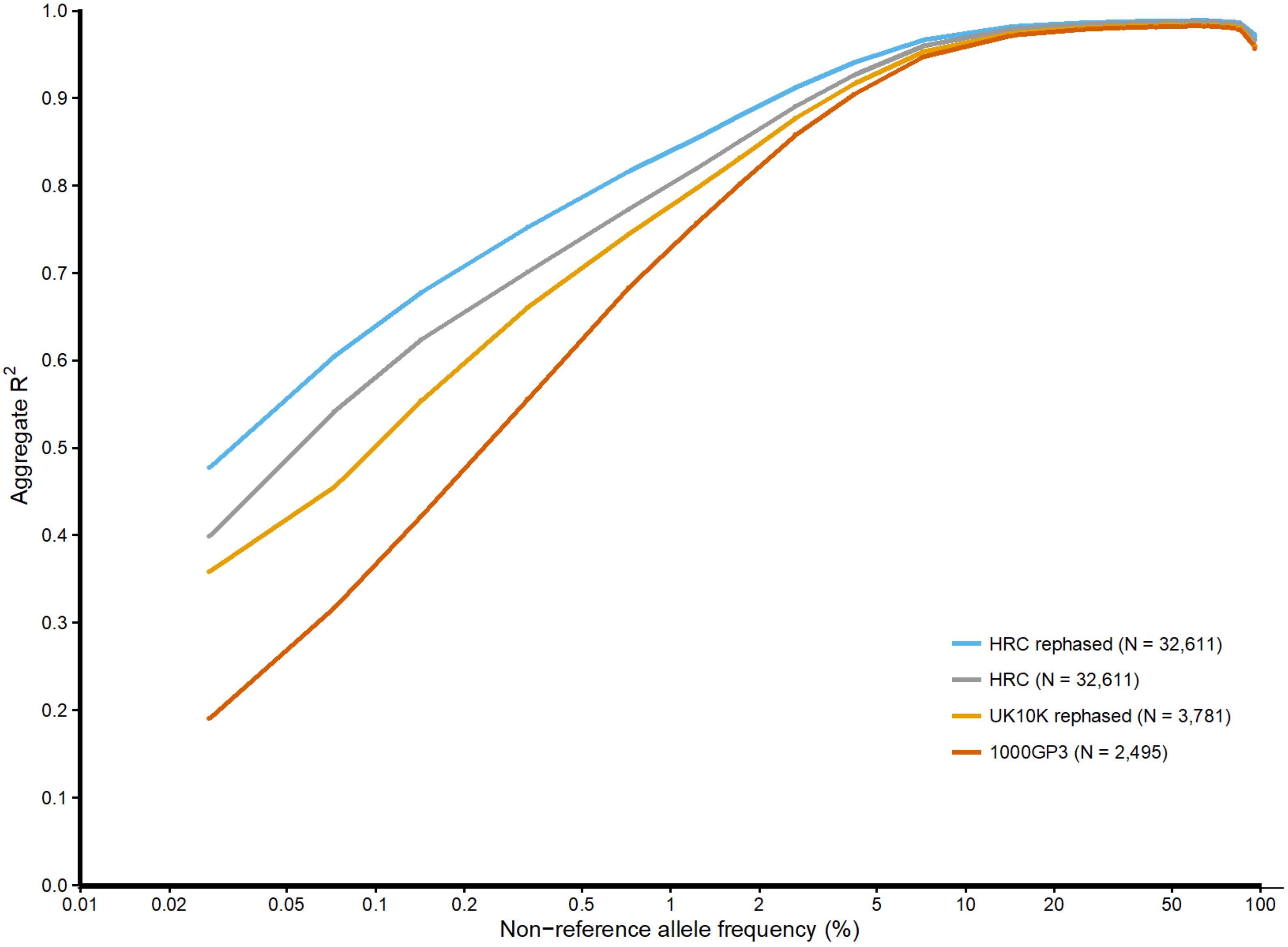


Image from McCarthy et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 10.1038/ng.3643

Using the HRC for imputation

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

- The HRC data is NOT publicly available, as the HapMap and 1000GP haplotypes are, due to consent issues.

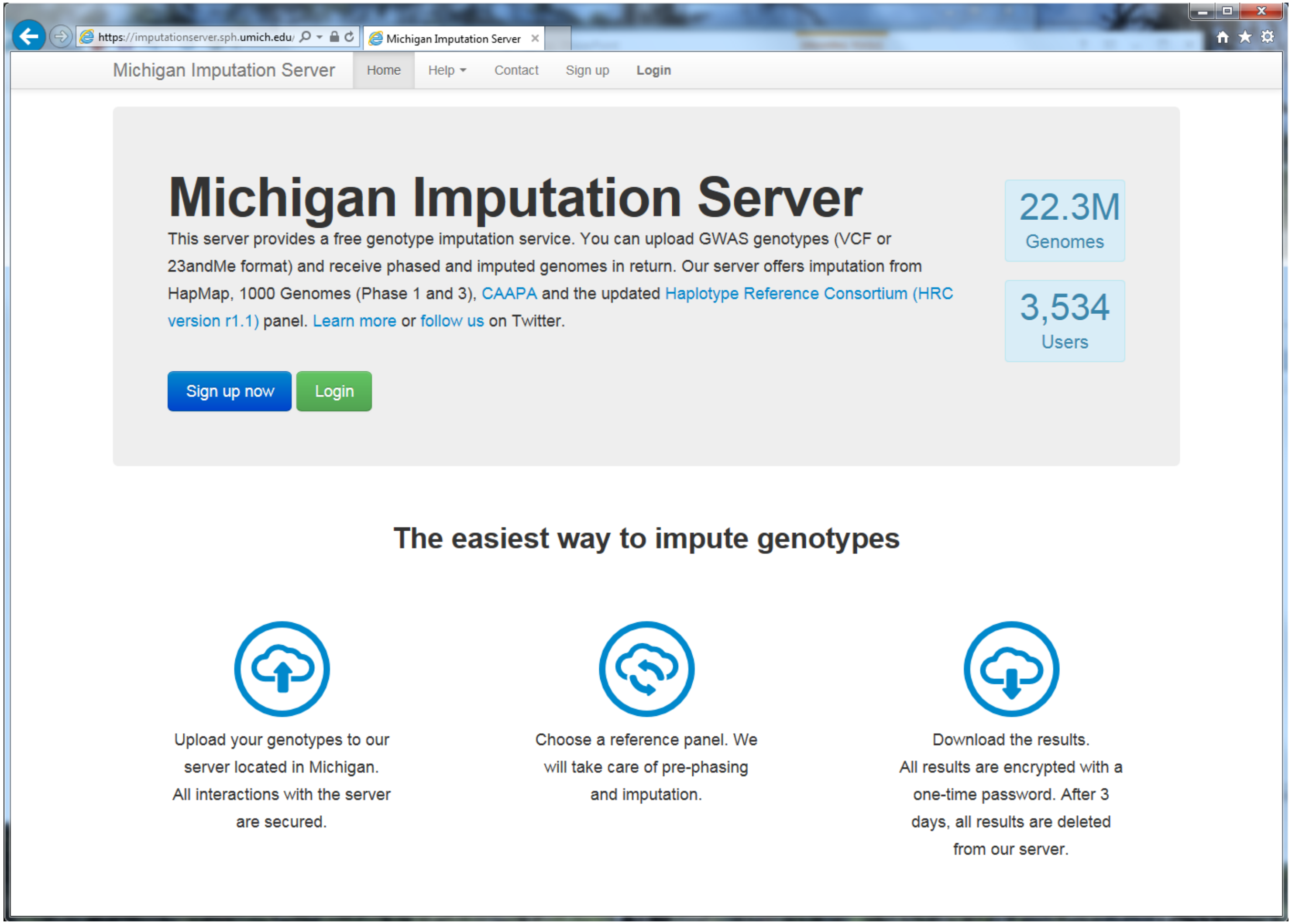
- A subset of HRC haplotypes has been made available for the sole purpose of imputation.

- Currently 2 imputation servers exist that allow users to upload genotypes from their GWAS samples, and have the phasing and imputation (based on either HRC or 1000G) carried out remotely and efficiently:

https://imputation.sanger.ac.uk/
https://imputationserver.sph.umich.edu

Michigan imputation server

https://imputationserver.sph.umich.edu


https://imputationserver.sph.umich.edu/
Michigan Imputation Server
Michigan Imputation Server
Home
Help
Contact
Sign up
Login

Michigan Imputation Server

This server provides a free genotype imputation service. You can upload GWAS genotypes (VCF or 23andMe format) and receive phased and imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and 3), CAAPA and the updated Haplotype Reference Consortium (HRC version r1.1) panel. Learn more or follow us on Twitter.

22.3M Genomes

3,534 Users

The easiest way to impute genotypes

Upload your genotypes to our server located in Michigan. All interactions with the server are secured.

Choose a reference panel. We will take care of pre-phasing and imputation.

Download the results. All results are encrypted with a one-time password. After 3 days, all results are deleted from our server.

Michigan imputation server

1. Prepare your data
2. Register (if necessary) and Login
3. Upload your data
4. Start the Imputation
5. Download Results
6. Carry out post-imputation QC to retain only reliable genotypes

Michigan imputation server

1. Prepare your data:

- Perform standard GWAS QC (**and carry out a preliminary GWAS** to identify any problem SNPs)
- Check/update your data (BP positions, strand, alleles etc.) to match that of your chosen reference panel http://www.well.ox.ac.uk/~wrayner/tools/
- Convert to sorted VCF files (one per chromosome) compressed by bgzip (*.vcf.gz)
  - E. g. using plink2/VCFtools/VcfCooker and VCFtools and tabix (including bgzip)

Michigan imputation server

5. Download Results:

- Potentially **huge** files (!) – use wget
- Then need to unzip the files using the password provided by email

6. Carry out **post-imputation checks**
e.g. filtering by info (Rsq) score, MAF, genotype probabilities or call rates etc. etc.

TOPMed imputation server

- https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home
- Diverse reference panel including information from 97,256 deeply sequenced human genomes
- Samples derived from Trans-Omics for Precision Medicine (TOPMed) program
  - To elucidate the genetic architecture and disease biology of heart, lung, blood, and sleep disorders
  - Via whole genome sequencing
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7875770/

TOPMed imputation server

- Very similar to Michigan imputation server
- Allows you to filter results to only download those variants passing a specified Rsq (e.g. 0.3)

- Beware of using Will Raynor's tool as described in "Usage with the TOPMed reference panel"
  - Uses list of sites in ALL.TOPMed_freeze5_hg38_dbSNP.vcf.gz to update your genotypes to match the TOPMed panel (which is in Build 38)
  - This list does not use SNP names (rs IDs), instead it has SNP IDs like TOPMed_freeze5?chr1:76,766
    - So most of your will get excluded, if they are not already in Build 38!

- Better use Will Raynor's tool as described in "Usage with HRC reference panel", which updates to Build 37.