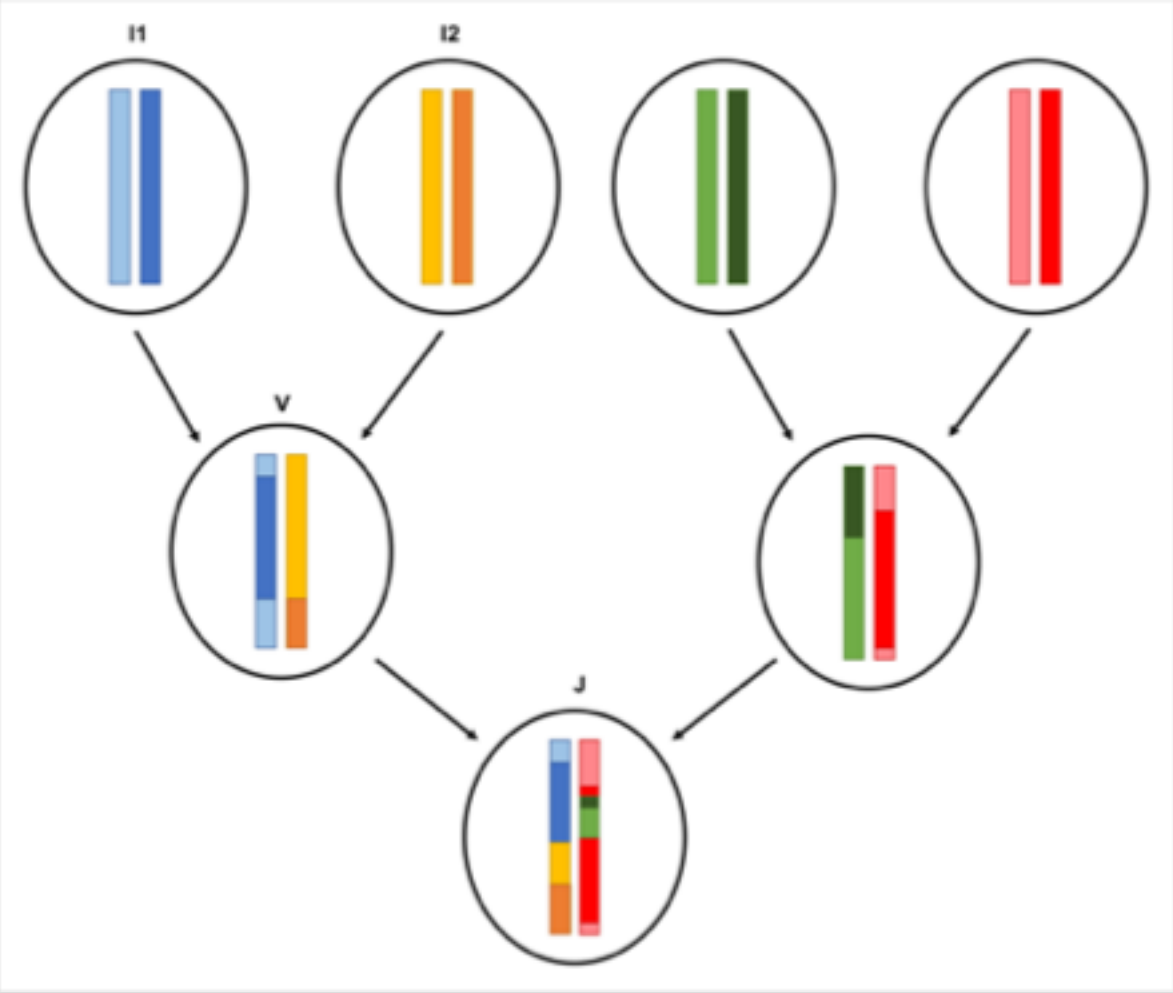
GWAS 10

Matti Pirinen
University of Helsinki
27.2.2019

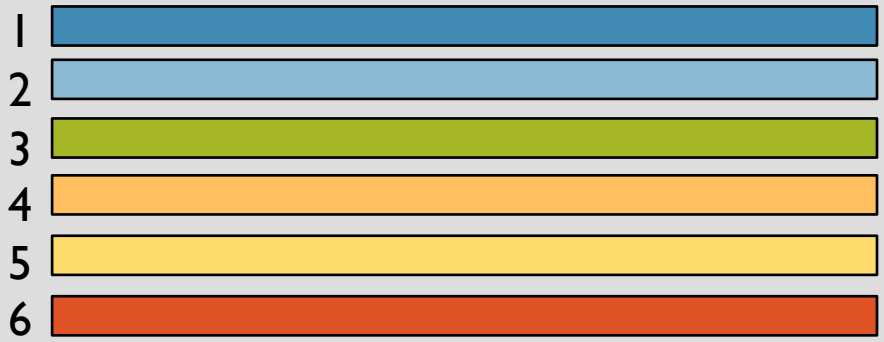
GENOMIC RECOMBINATION



Sini Kerminen; www.helsinki.fi/~mjxpirin/stamp/

- Offspring inherits genome in continuous segments of the parent's two genomes
- This is the source of LD in population
  - When two haplotypes share an allele at locus K, they are also more likely to share certain allele at nerby locus L than two random haplotypes from the population
    - sharing at locus K → common ancestor at locus K → a shared segment around locus K → increased probability of sharing also at nearby locus L
- How does this phenomenon show up in haplotypes sampled from a population?

PROBABILITY MODEL FOR HAPLOTYPE K+1 CONDITIONAL ON HAPLOTYPES 1…K

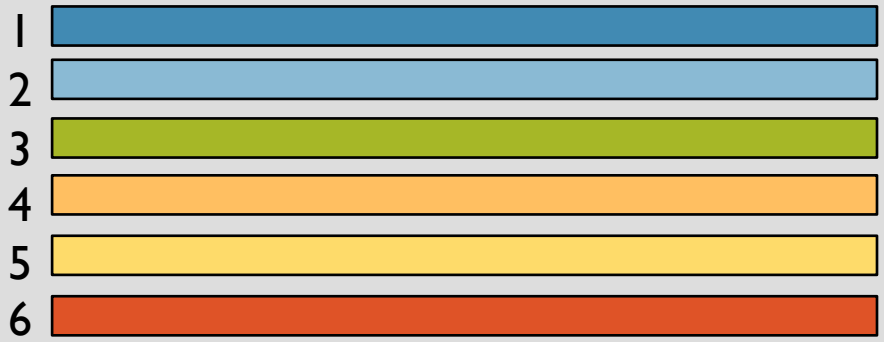


We have observed 6 haplotypes from the population.
How could a 7th one from the same population look like ?

Li & Stephens model (Genetics 2003)
build to capture following properties at any locus

1. The next haplotype is more likely to match a haplotype that has already been observed many times rather than one that has been observed less frequently.
2. The probability of seeing a novel haplotype not related to K haplotypes decreases as *K* increases.
3. If the next haplotype is not exactly the same as an existing (*i.e.*, previously seen) haplotype, it will tend to differ by a small number of mutations from an existing haplotype, rather than to be completely different from all existing haplotypes.
4. Due to recombination, the next haplotype will tend to look somewhat similar to existing haplotypes over contiguous genomic regions, the average physical length of these regions being larger in areas of the genome where the local rate of recombination is low.

PROBABILITY MODEL FOR HAPLOTYPE K+1 CONDITIONAL ON HAPLOTYPES 1…K

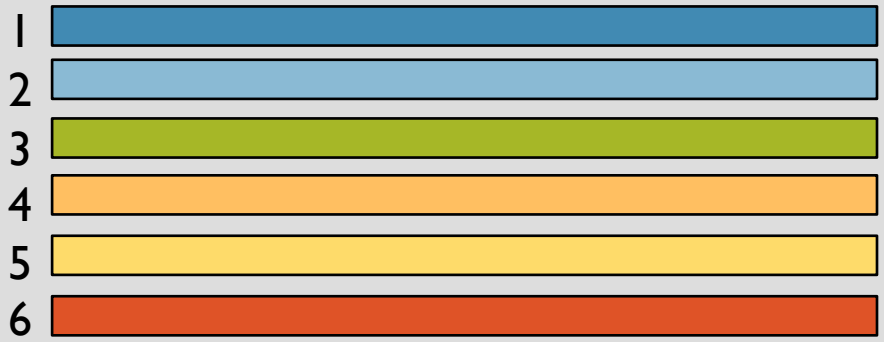


We have observed 6 haplotypes from the population.
How could a 7th one from the same population look like ?

- A recipe to generate a new haplotype sequentially locus-by-locus from left to right
  1) Choose one of the existing halpotypes uniformly at random and start following it from left end
  2) Repeat until reaching the right end:
     I. Choose a length of the segment that has not experienced any recombination events with other haplotypes in the population and follow the current haplotype for that distance to right
     II. Choose the next haplotype to follow uniformly at random among the K haplotypes
  3) Sample the allele of the haplotype by strongly preferring the allele of the reference haplotype being followed but allowing possibility of mutation

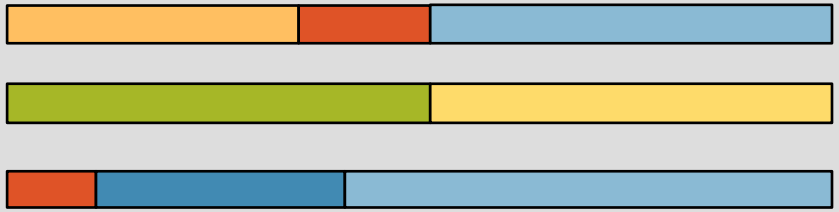
This is called **Li & Stephens model** (Genetics 2003)

PROBABILITY MODEL FOR HAPLOTYPE K+1 CONDITIONAL ON HAPLOTYPES 1…K



We have observed 6 haplotypes from the population.
How could a 7th one from the same population look like ?

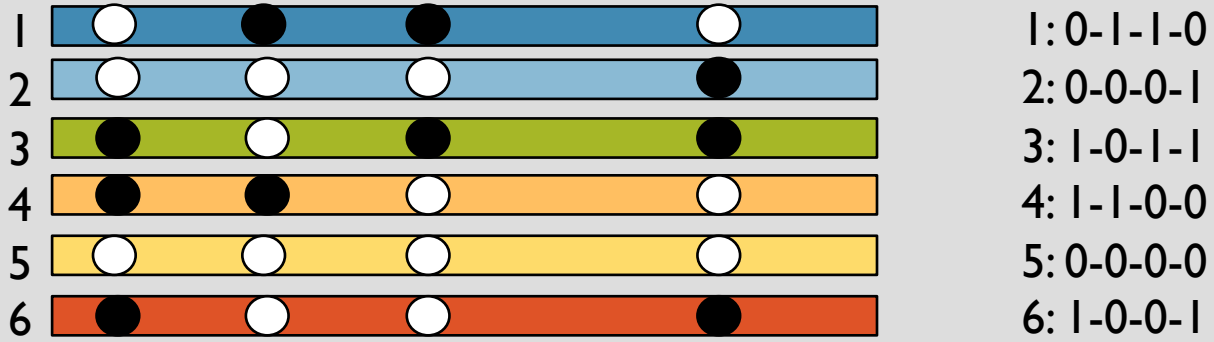
Here are 3 examples sampled from the L&S model



- A recipe to generate a new haplotype sequentially locus-by-locus from left to right
  1) Choose one of the existing halpotypes uniformly at random and start following it from left end
  2) Repeat until reaching the right end:
     I. Choose a length of the segment that has not had any recombination events with other haplotypes in the population and follow the current haplotype for that distance to right
     II. Choose the next haplotype to follow uniformly at random among the K haplotypes
  3) Sample the allele of the haplotype by preferring the allele of the reference haplotype being followed but allowing possibility of mutation

This is called **Li & Stephens model** (Genetics 2003)

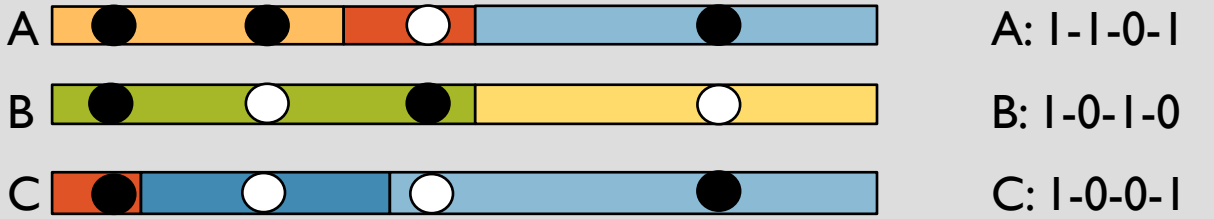
PROBABILITY MODEL FOR HAPLOTYPE K+1 CONDITIONAL ON HAPLOTYPES 1…K



1: 0-1-1-0
2: 0-0-0-1
3: 1-0-1-1
4: 1-1-0-0
5: 0-0-0-0
6: 1-0-0-1

What if we only observe alleles at certain SNPs (and not the sequence in between)?
Let's mark them by white (0) and black (1).

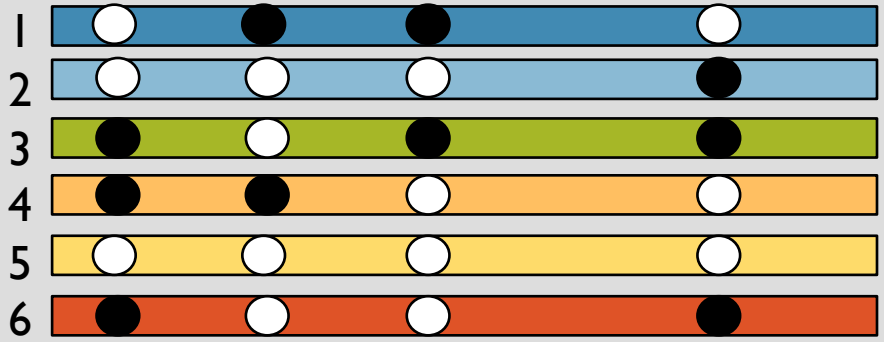
Here are 3 examples sampled from the L&S model



A: 1-1-0-1
B: 1-0-1-0
C: 1-0-0-1

- L&S model produces haplotypes that are imperfect mosaic of the reference haplotypes
  - Allele at locus L does not need to match exactly to the allele at the reference haplotype being copied
    - See e.g. Hap C at SNP 2
    - This allows modeling also haplotypes that do not have close relative among the reference haplotypes
- Probability model for observed alleles at a new haplotype combines probabilities of
  - Recombination events quantified by the recombination rates at the proposed positions of recombination
  - Mutations from the reference haplotype
- A haplotype that needs many recombinations and/or many mutations compared to the reference haplotypes is improbable according to L&S model

INFERENCE FROM L&S MODEL



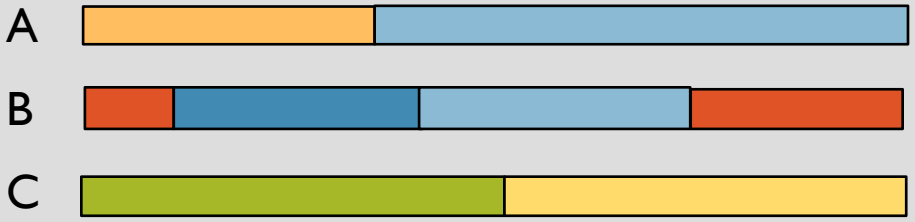
Suppose that we have observed a haplotype with alleles

With L&S model we can compute probabilities
for all possible candidate haplotypes, such as A,B, C

- Denote by H_l the reference haplotype 1,…,K that new haplotype K+1 is following at locus l
- Denote by $x_i(l)$ the allele that haplotype i has at locus l and by $X = \{x_i(l): i = 1,\ldots,K+1, l = 1,\ldots,L\}$ the observed allele data
- Given X, L&S model allows computing probabilities

$$P(H_l = h \mid X) \text{ for all } h=1,\ldots K \text{ and } l = 1,\ldots,L$$

- There is an efficient algorithm (Hidden Markov Model) to compute these probabilities
- As a result, we have, for each locus, a probability that the new haplotype follows each of the reference haplotypes

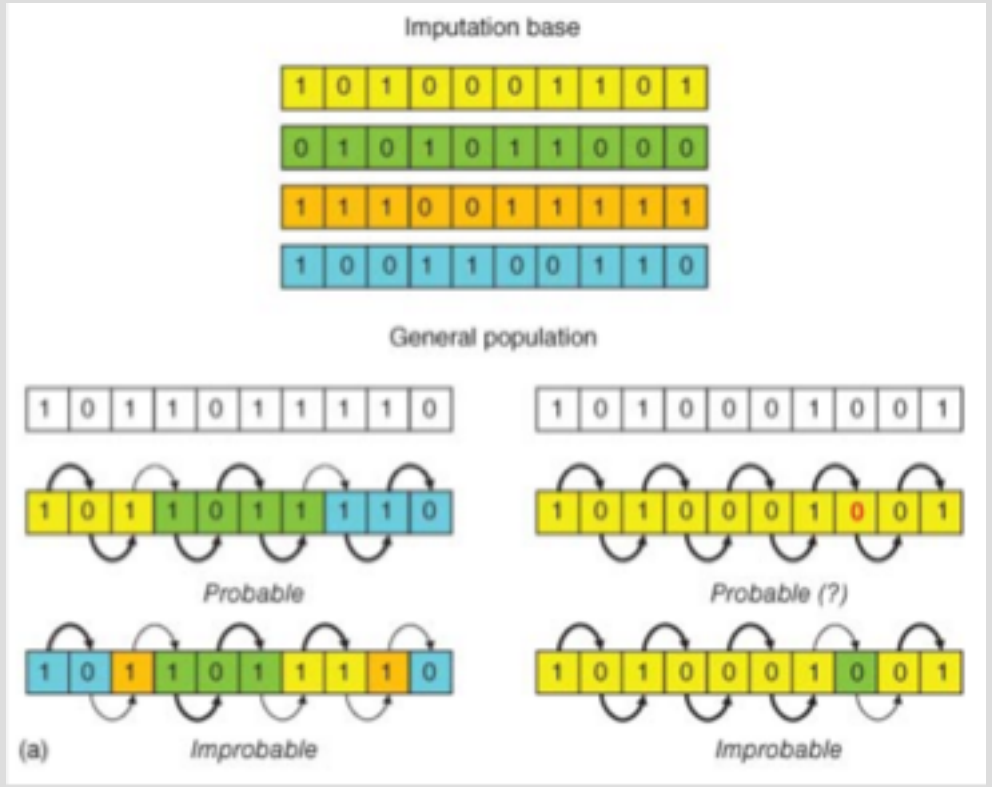


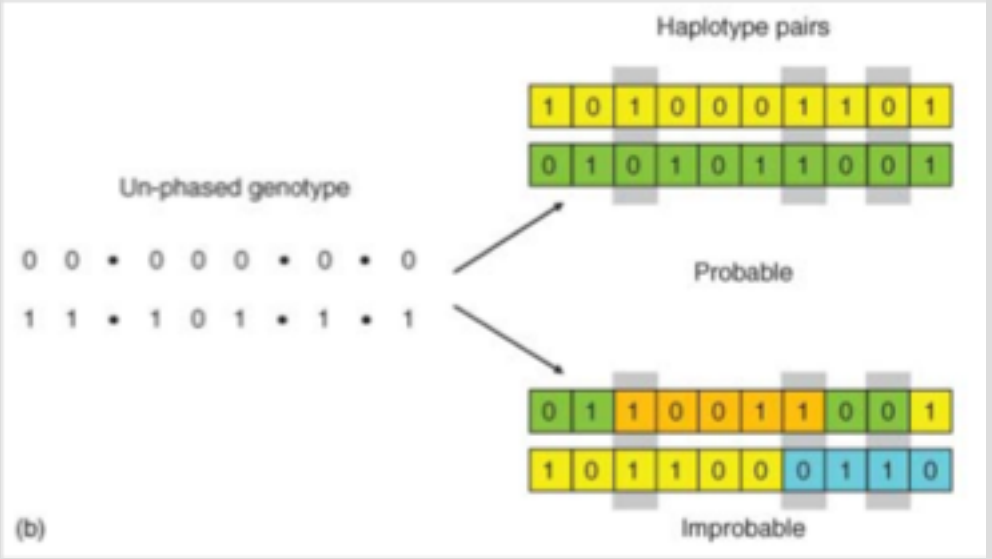
A: Needs no mutations, 1 recombination, highly probable

B: Needs no mutations, 3 recombinations, moderately probable

C: Needs 3 mutations, 1 recombination very improbable

GENOTYPE IMPUTATION





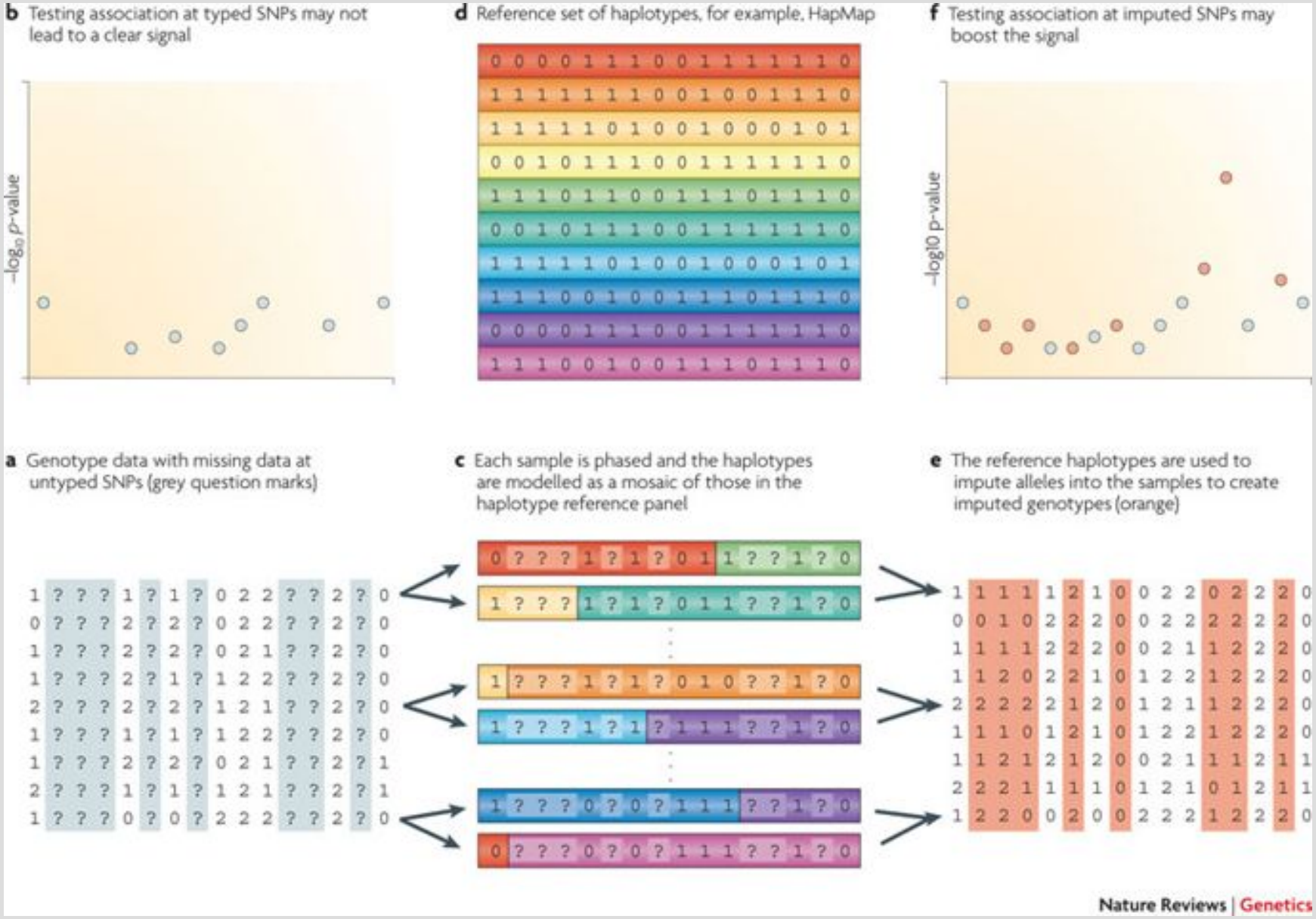
Genotype imputation with IMPUTE. The method is illustrated for 10 linked SNPs with alleles encoded by 0 and 1 (indicating, e.g. the presence or absence of a reference allele), using an imputation reference panel of 4 haplotypes.

(a) Every population haplotype is assumed to be a mosaic of haplotypes from the reference according to L&S model. High transition probabilities are indicated by bold arrows; thin arrows indicate lower transition probabilities. Possible mutations are highlighted in red.

(b) The probability of an unphased genotype with missing data is evaluated by considering all possible pairs of mosaic haplotypes that would be compatible with the observed data. The most probable pair determines the most probable genotypes at untyped SNPs
(i.e. 0–1, 1–1 and 0–0 rather than 1–1, 0–1 and 0–1 in the present example).

Krawczak 2015 eLS
https://doi.org/10.1002/9780470015902.a0022399

GENOTYPE IMPUTATION

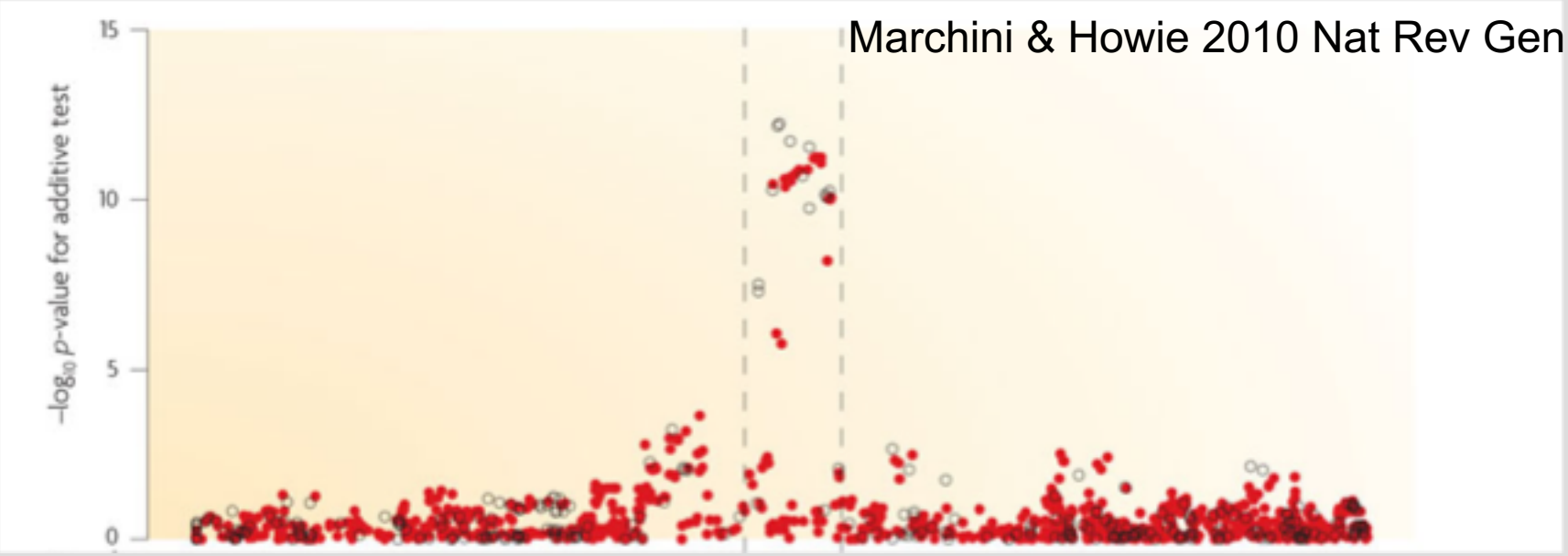


Howie & Marchini (2010) *Nature Reviews Genetics* **11:** 499–511

The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (**a**). Testing for association at just these SNPs may not lead to a significant association (**b**). Imputation attempts to predict these missing genotypes. Algorithms essentially involve phasing each individual in the study at the typed SNPs(**c**). These haplotypes are compared to the dense haplotypes in the reference panel (**d**). The phased study haplotypes have been coloured according to which reference haplotypes they match. The haplotypes of a given individual are modelled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (**e**). Testing these imputed SNPs can lead to detailed view of associated regions (**f**).

USES OF IMPUTATION

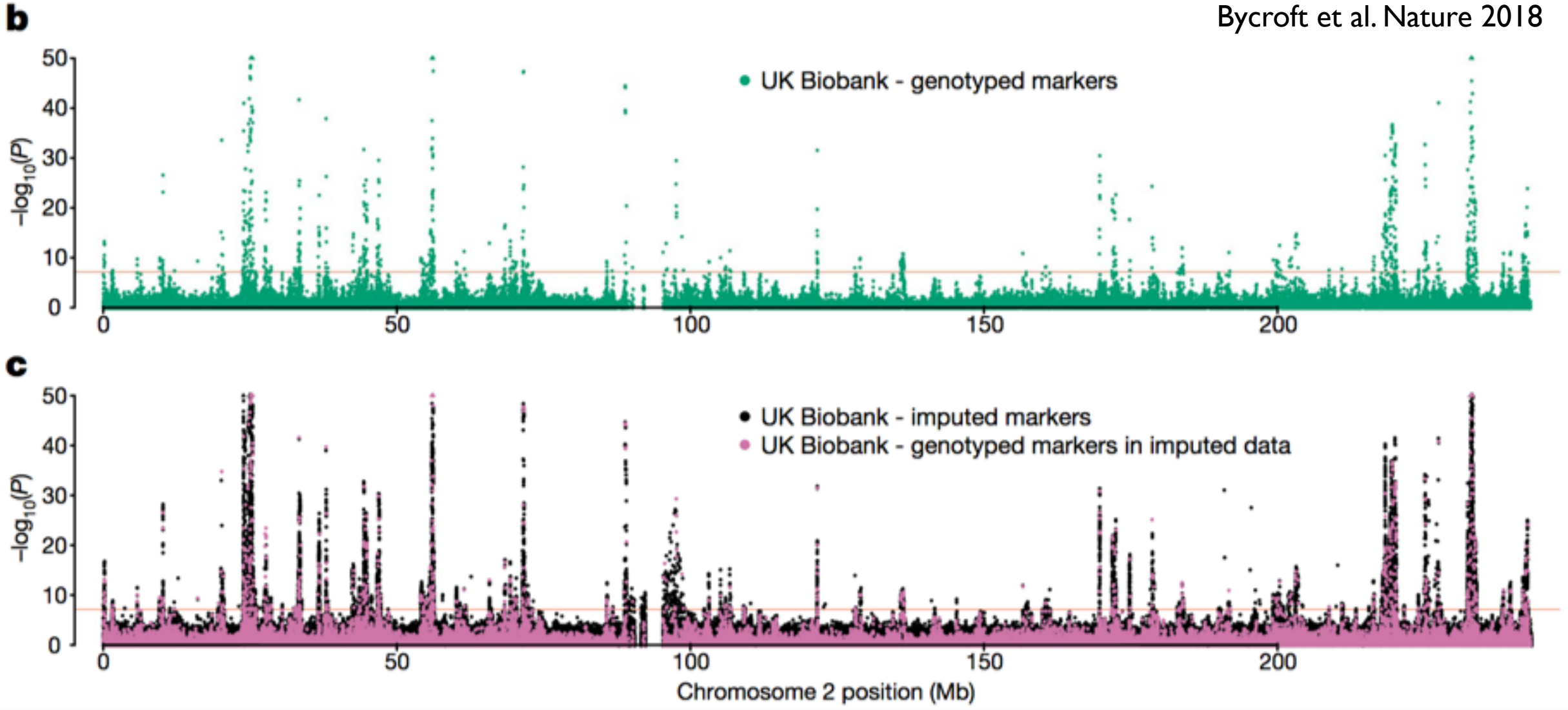
- Boosting power as more variation is captured
- Fine-mapping to reveal best candidates
- Meta-analysis to combine data on the union of SNPs across cohorts



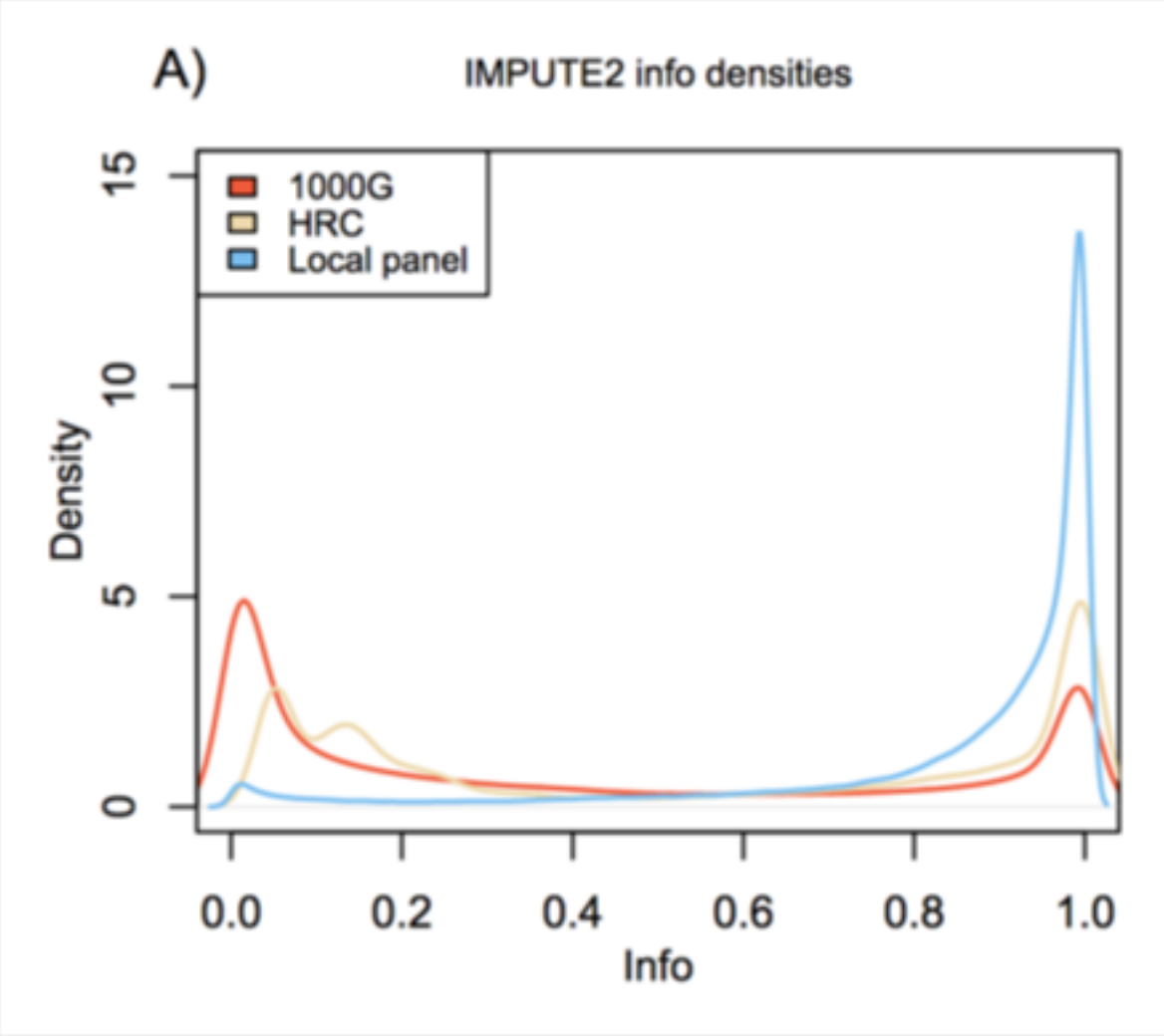
Marchini & Howie 2010 Nat Rev Gen

HEIGHT GWAS IN UK BIOBANK CHR 2

Bycroft et al. Nature 2018



INFO DISTRIBUTION



Imputation quality of imputing Finnish data (chromosome 17) using 3 reference panels:

1000G: 1000 Genomes panel (N = 1,092);

HRC: Haplotype reference Consortium panel (N = 32,488);

Local panel: Combined Finnish low-pass whole genome sequence and high-pass whole exome sequence panel (N = 6,873)

Surakka et al. 2016 bioRxiv