GWAS Experimental Design: genotypes

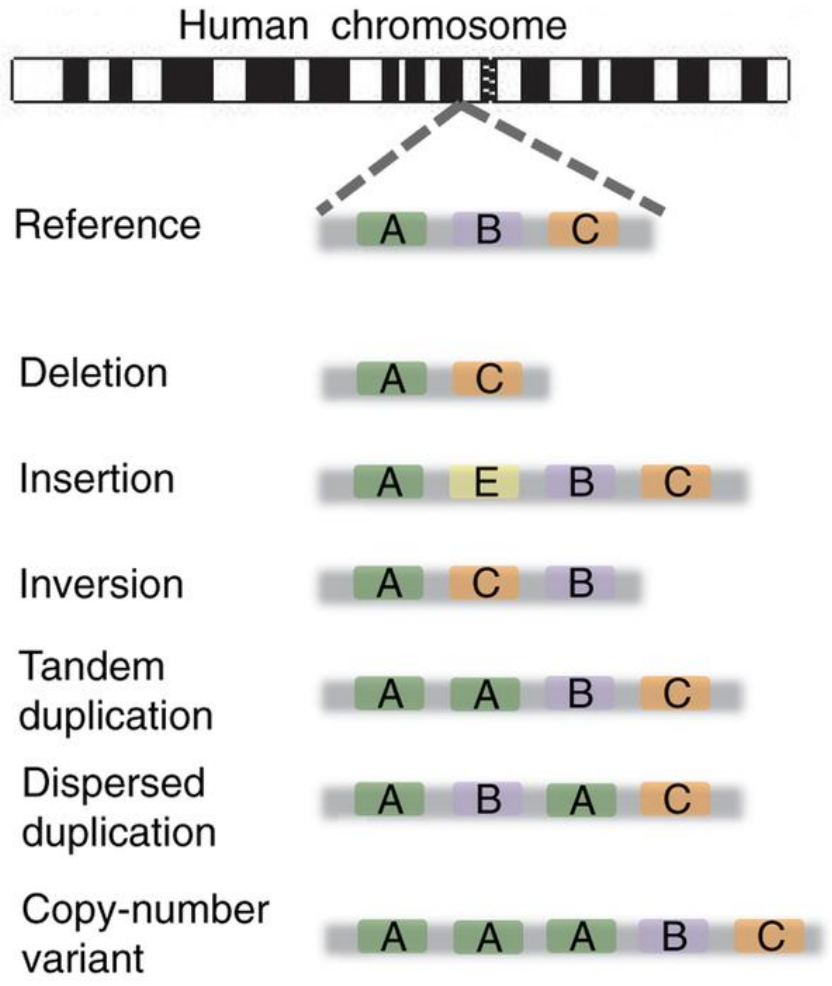
Outline of lecture

- Types of genetic data
  - SNP chips, whole genome sequence data

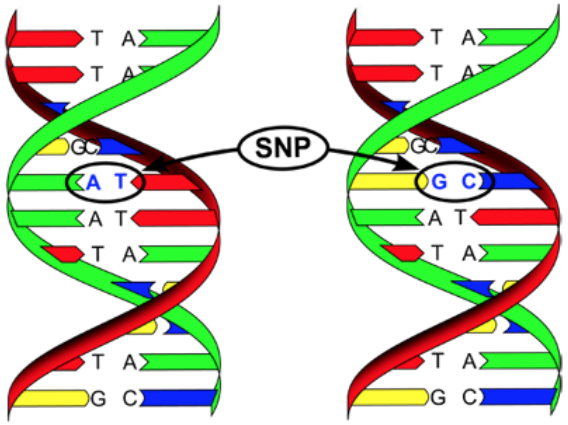
- Two types of ‘equilibrium’:
  - Hardy Weinberg Equilibrium
  - Linkage disequilibrium (LD)

Variation in DNA

- All people have 99.9% identical DNA
- We are interested in the 0.1% which is different *between* people
  - *e.g. How do these differences contribute to disease?*
- Different types of genetic variation
  - structural variants (deletions, inversions, insertions)
  - ...
  - SNP (single nucleotide polymorphism)



SNP = Single Nucleotide Polymorphism



- Most common type of variation in the genome
- Easily/reliably assayed (measured) at many places

What does ‘genetic data’ look like?

- Can assay ~1M SNPs per individual with ‘SNP chips’
- Data is typically counts of a ‘reference’ (A1) allele

genotype file:

	SNP1	SNP2	SNP3	SNP4
Bob	0	1	0	1
Fred	1	2	0	0
Jose	1	2	2	2
Andy	2	1	1	1

map file:

	chr	position	A1	A2
SNP1	1	52196307	A	T
SNP2	1	52462094	C	T
SNP3	1	52736008	A	G
SNP4	1	53010891	T	C

Whole Genome Sequencing

Genetic data

Either SNP chip or WGS data, once cleaned, is processed in similar manner.

In the practical tomorrow we will ‘clean’ the SNP chip genotypes, e.g.
--Missing genotypes
--Check for allele frequency differences
--Check for Hardy-Weinberg inconsistencies

We will spend rest of lecture on two important concepts
1. Hardy-Weinberg equilibrium
2. Linkage disequilibrium

Hardy-Weinberg Equilibrium

- HWE is the probabilistic relationship between allele and genotype frequencies at a single locus, i.e.

Consider an A/a bi-allelic locus:
and Alleles are A and a
Frequency of A is p
Frequency of a is q (thus p = 1 – q)

Three possible genotypes:
AA has expected frequency p^2
Aa has expected frequency 2pq
aa has expected frequency q^2

	A *p*	a *q*
A *(p)*	AA *(p^2)*	Aa *(pq)*
a *(q)*	aA *(qp)*	Aa *(q^2)*

Hardy-Weinberg Equilibrium

- HWE is the probabilistic relationship between allele and genotype frequencies at a single locus, i.e.

  BUT do our observed genotype frequencies match the expected frequencies?

  Test for HWE via Pearson's chi-squared test with 1df: $\chi^2 = \sum \frac{(O-E)^2}{E}$

Genotype	AA	Aa	aa	Total
Observed - number	233	385	129	747
Expected - frequencies	p^2	2pq	q^2	1
Expected - number	242.4	366.3	138.4	
$\chi^2 = 1.96$ with 1 df => P(X>1.96) = 0.162				

Hardy-Weinberg Equilibrium

- HWE is the probabilistic relationship between allele and genotype frequencies at a single locus
  - HWE makes many assumptions
- When is a locus *not* in HWE?

Hardy-Weinberg Equilibrium

- HWE is the probabilistic relationship between allele and genotype frequencies
  - HWE makes many assumptions
- When is a locus *not* in HWE?
  - Selection and/or demographic events
  - Unknown population structure in sample
  - Non-random mating
  - Genotyping errors (!)

Linkage Disequilibrium (LD)

- **Linkage disequilibrium (LD)** is the non-random association between genotypes at multiple sites in the genome.

Friend or foe?

Linkage Disequilibrium (LD)

- **Linkage disequilibrium (LD)** is the non-random association between alleles at multiple sites in the genome.

- GWAS exploit LD between common SNP and ‘causative mutations’
  - the SNP associations in GWAS are (usually) *indirect* associations between the genome and the trait of interest

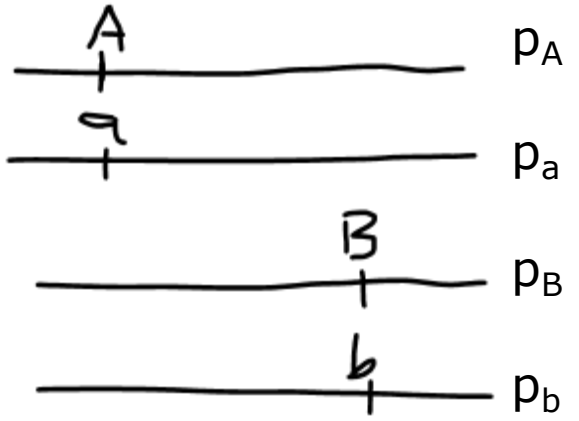
- LD is unhelpful for fine mapping or identifying ‘causative mutations’

Definitions of LD

Classical definition:

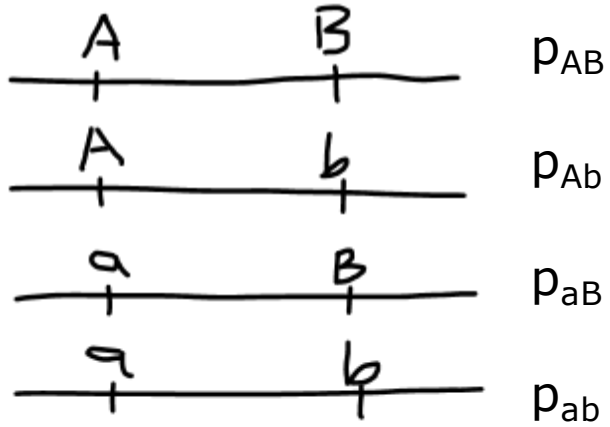
Imagine two bi-allelic loci

- locus 1: alleles are A, a with frequency p_A and p_a
- locus 2: alleles are B, b with frequency p_B and p_b



And there are 4 possible haplotypes:

- A_B with frequency p_{AB}
- A_b with frequency p_{Ab}
- a_B with frequency p_{aB}
- a_b with frequency p_{ab}



https://web.stanford.edu/group/pritchardlab/HGbook.html

Definitions of LD

Under linkage equilibrium…….

The alleles are <u>independent</u> :- therefore the frequency of the haplotype is determined by the frequency of the alleles, i.e.

$$p_{AB} = p_A \times p_B$$

$$p_{Ab} = p_A \times p_b$$

$$p_{aB} = p_a \times p_B$$

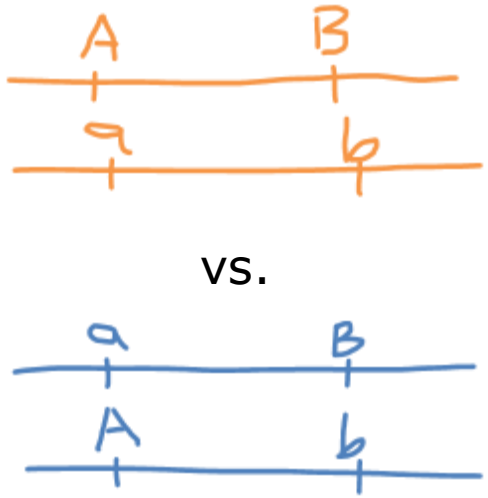
$$p_{ab} = p_a \times p_b$$

Definitions of LD

We can quantify the degree of independence between loci using 'D'

$$|D| = p_{AB}p_{ab} - p_{Ab}p_{aB}$$

If the two loci are independent, then D = 0



D measures if recombination has occurred. It is highly dependent on allele frequency & not suitable for comparing LD at different sites

Definitions of LD

$$r^2 = D^2/[p_A p_a p_B p_b]$$

r^2 ranges from [0,1]

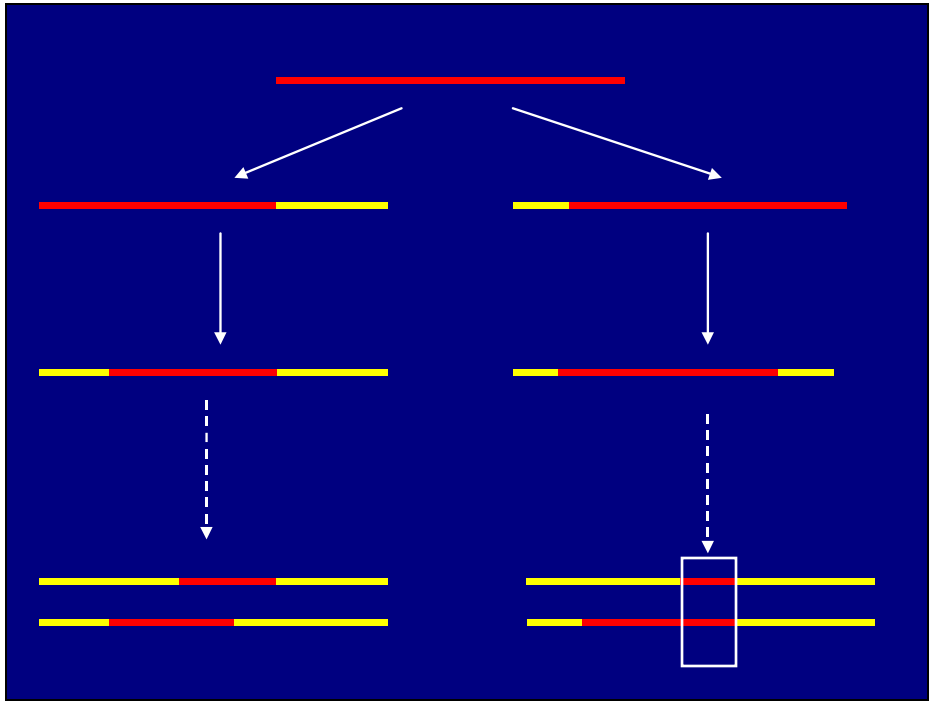
$r^2 = 0$ means the two loci are independent

$r^2 = 1$ means the two loci are 'in perfect' LD. Only occurs if two of the possible four haplotypes are observed, i.e. only A_B / a_b or only A_b / a_B

r^2 is equivalent to the **squared correlation co-efficient** between alleles

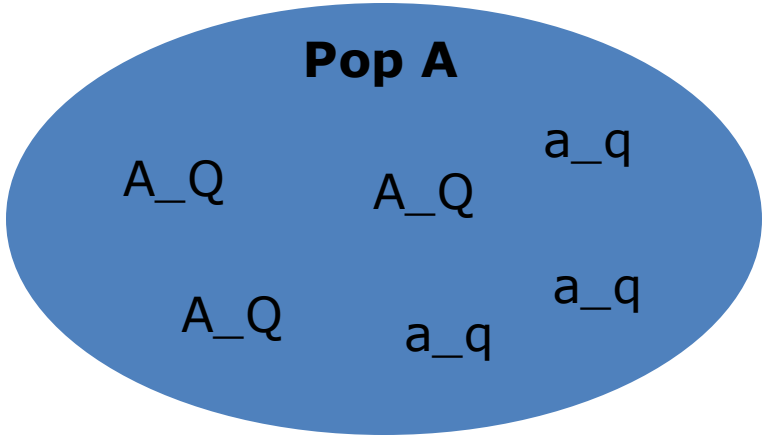
Recombination breaks up LD

- A chunk of ancestral chromosome is conserved in the 2nd population, but not the 1st population

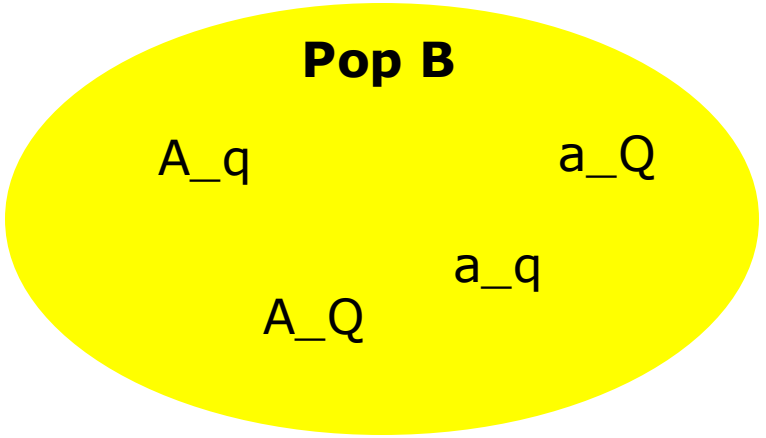


LD is population dependent

The association between a marker and a 'causative mutation' may be population dependent

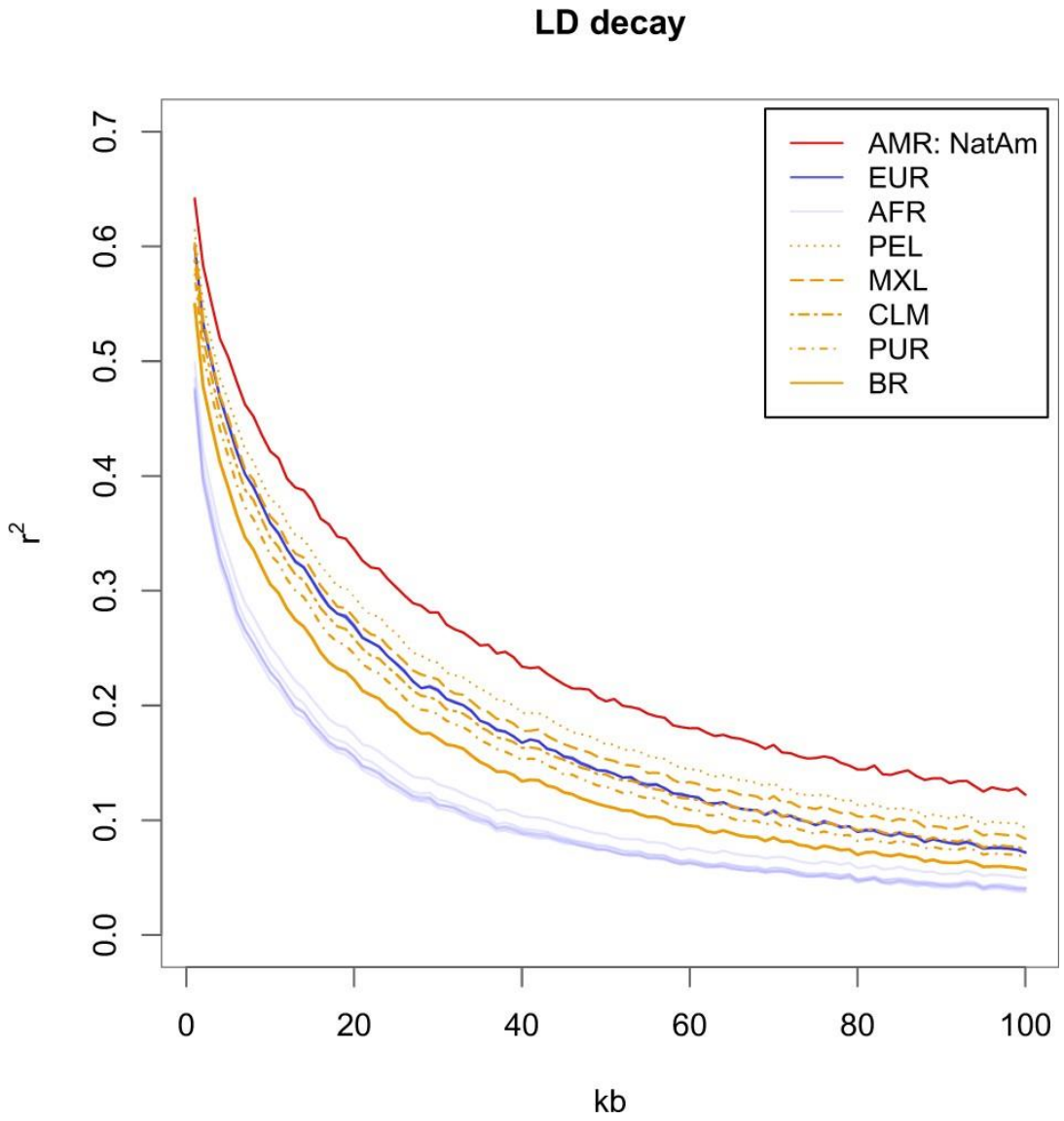


'A' in LD with (unobserved) mutation 'Q'



'A' in uncorrelated with mutation 'Q'

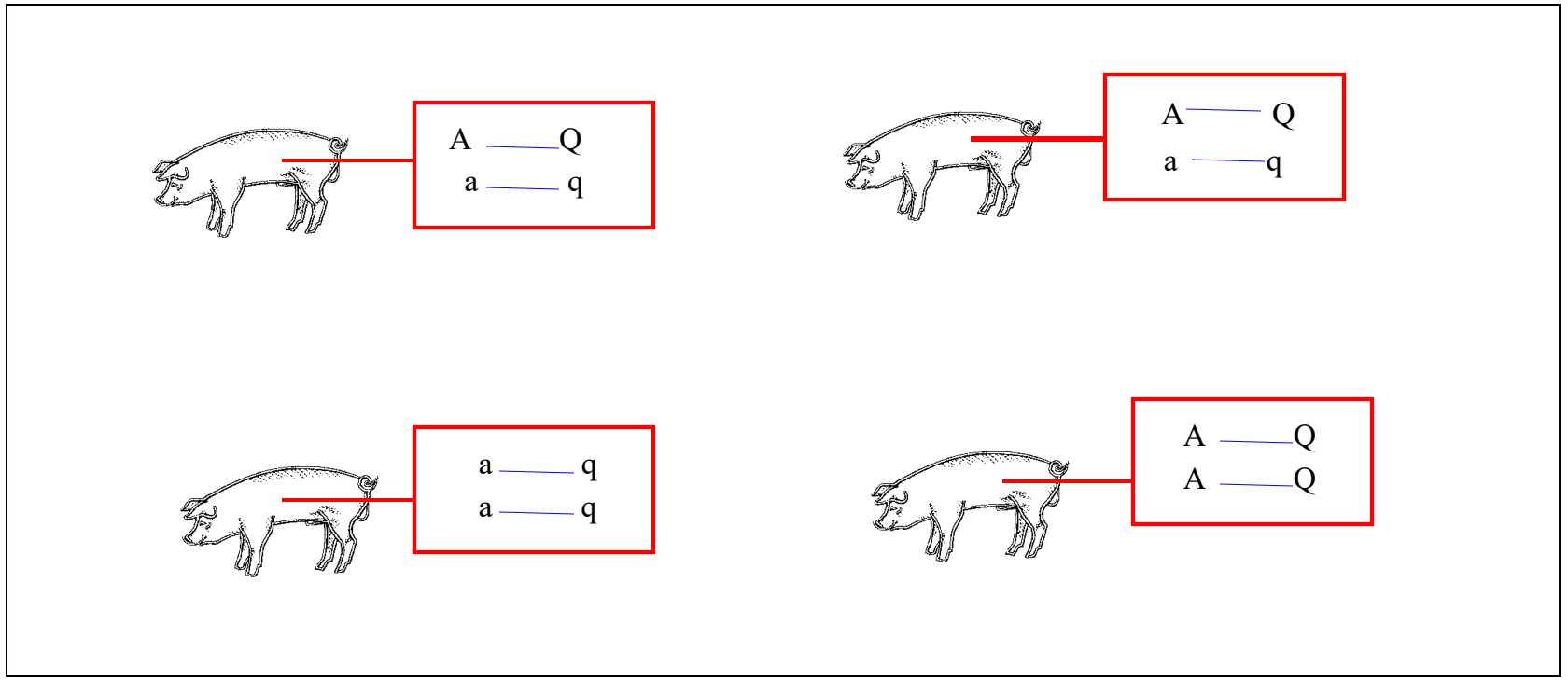
LD decays relative to genomic distance and is also population dependent



G3 Genes|Genomes|Genetics, Volume 12, Issue 7, July 2022, jkac111, https://doi.org/10.1093/g3journal/jkac111

Why do we care about LD?

1. we can use genetic markers as proxies to detect associations between genomic regions & a trait



Why do we care about LD?

2. We can use LD to fill in missing genotypes via imputation

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

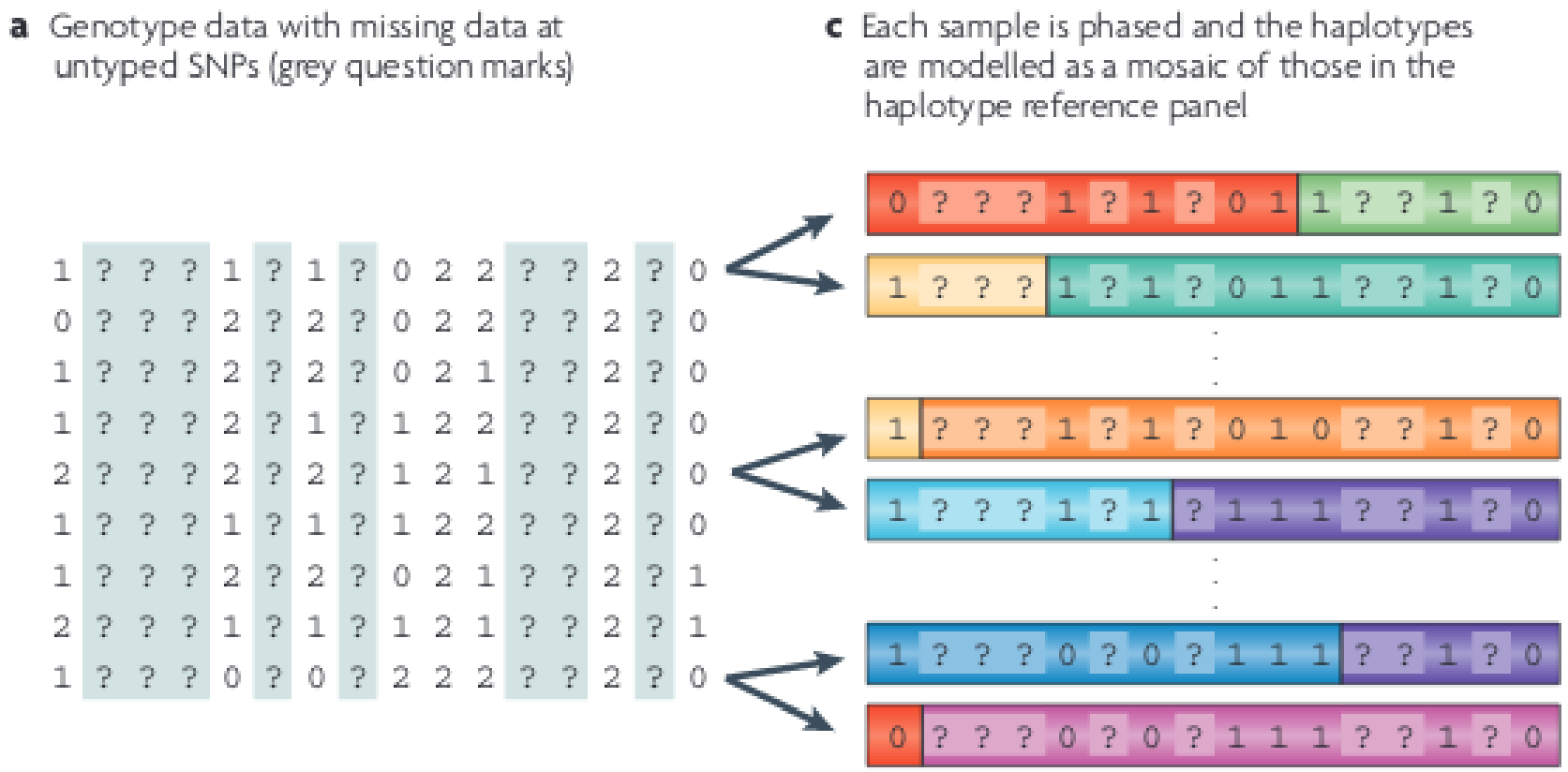
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

d 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

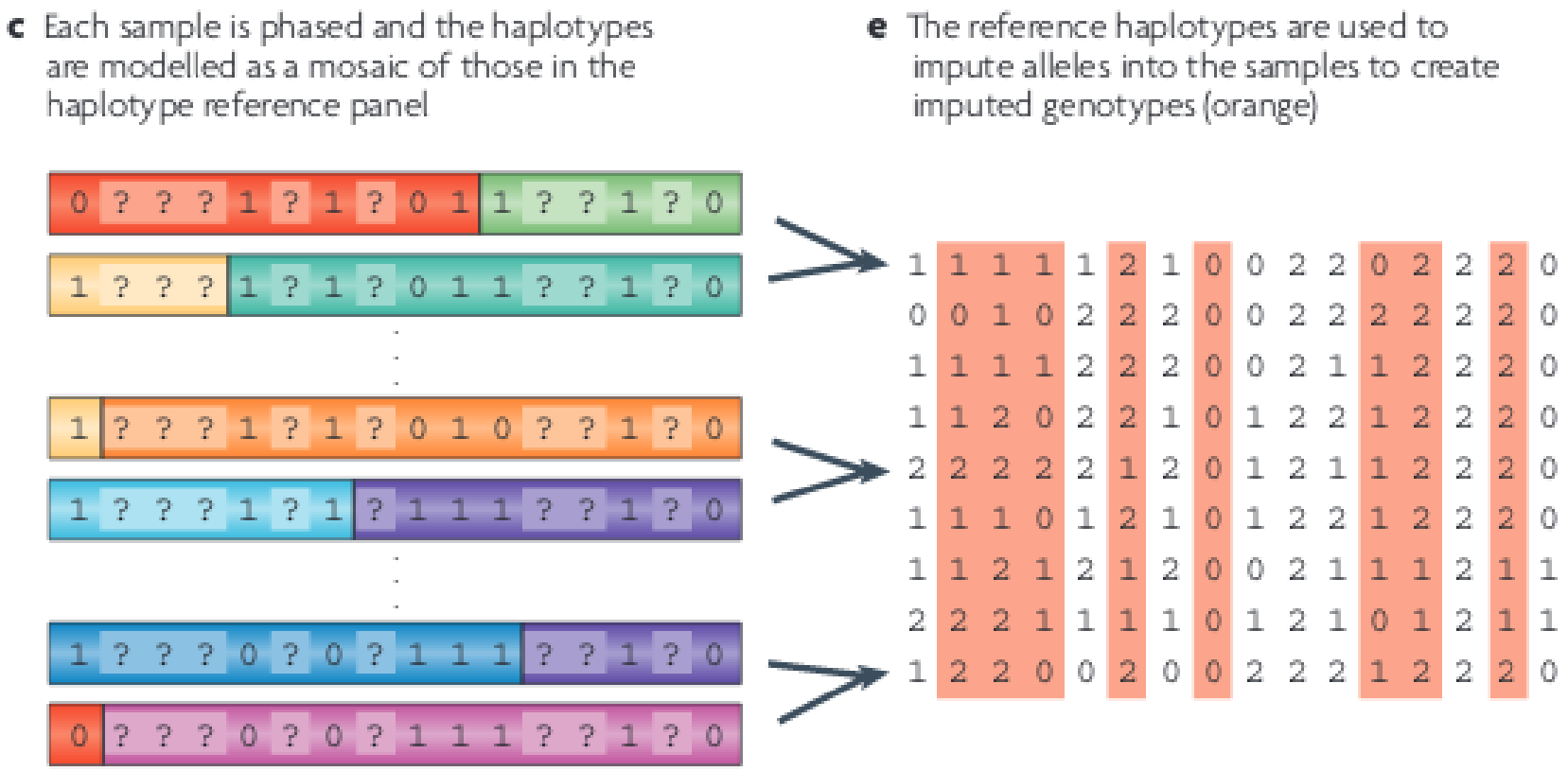
Why do we care about LD?

2. We can use LD to fill in missing genotypes via imputation



Why do we care about LD?

2. We can use LD to fill in missing genotypes via imputation



Why do we care about LD?

2. We can use LD to fill in missing genotypes via imputation

Imputation is used to:

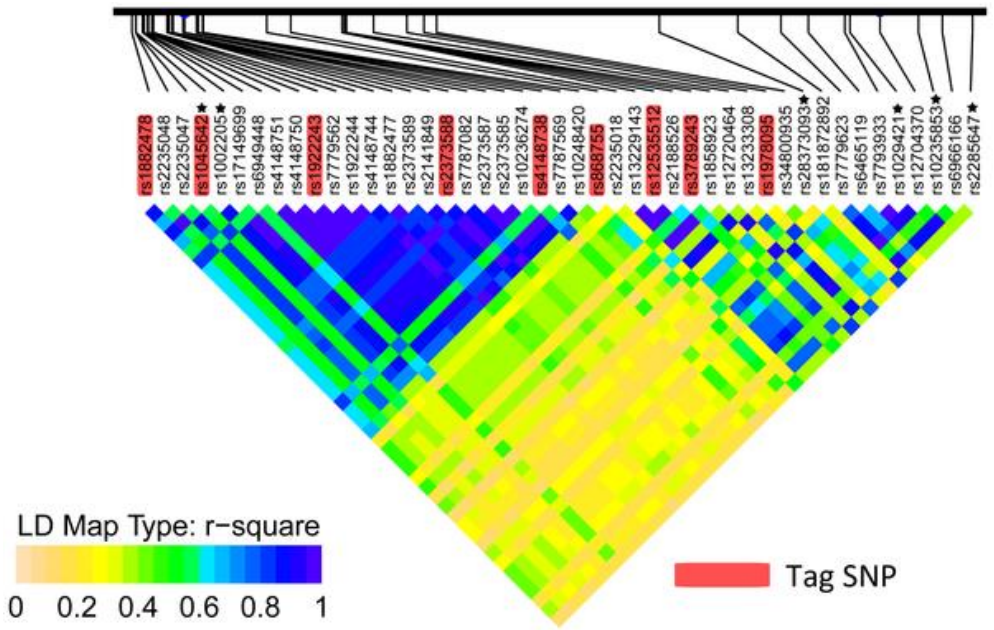
- fill in missing data, i.e. SNP removed during QC or poorly genotyped in some samples
- completely impute (unobserved) SNP in genotyped individuals from the reference panel

Imputed SNPs can be used in GWAS like genotyped SNPs

- increases the power to detect associations

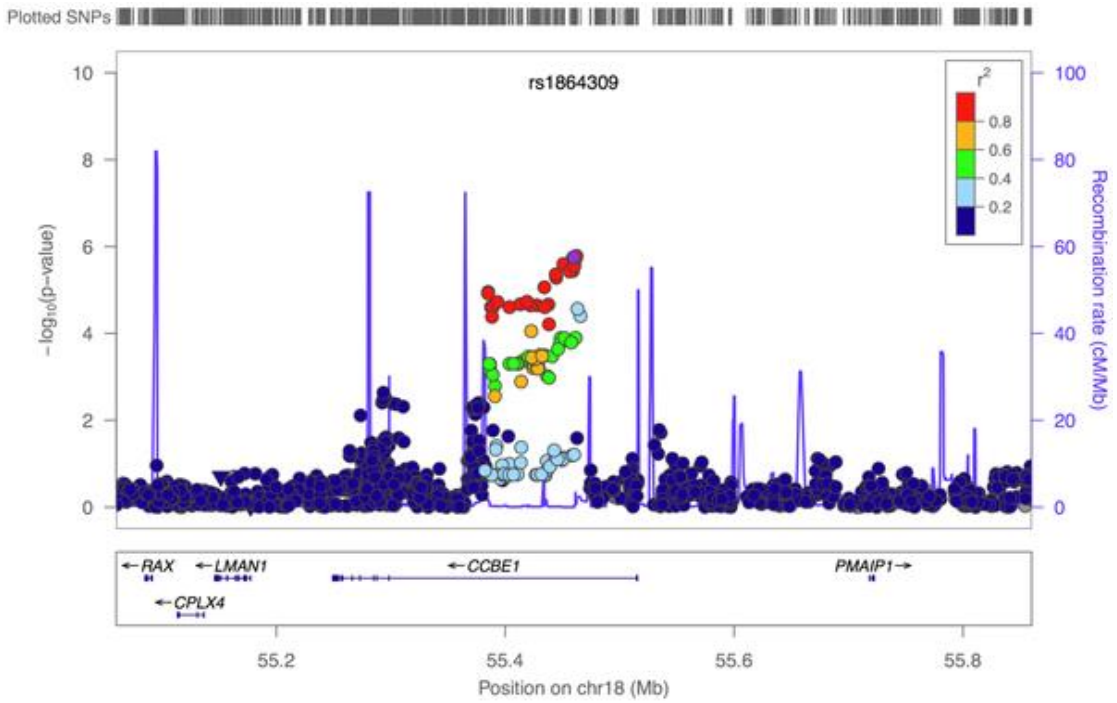
Representing LD in GWAS

- Pair-wise LD plot



Shou et al. (2012) *PLOS ONE*

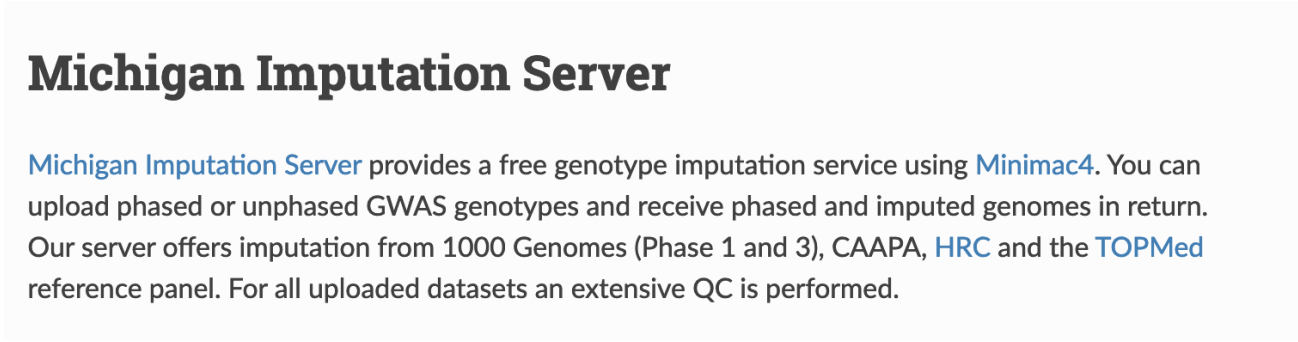
- Recombination graphs

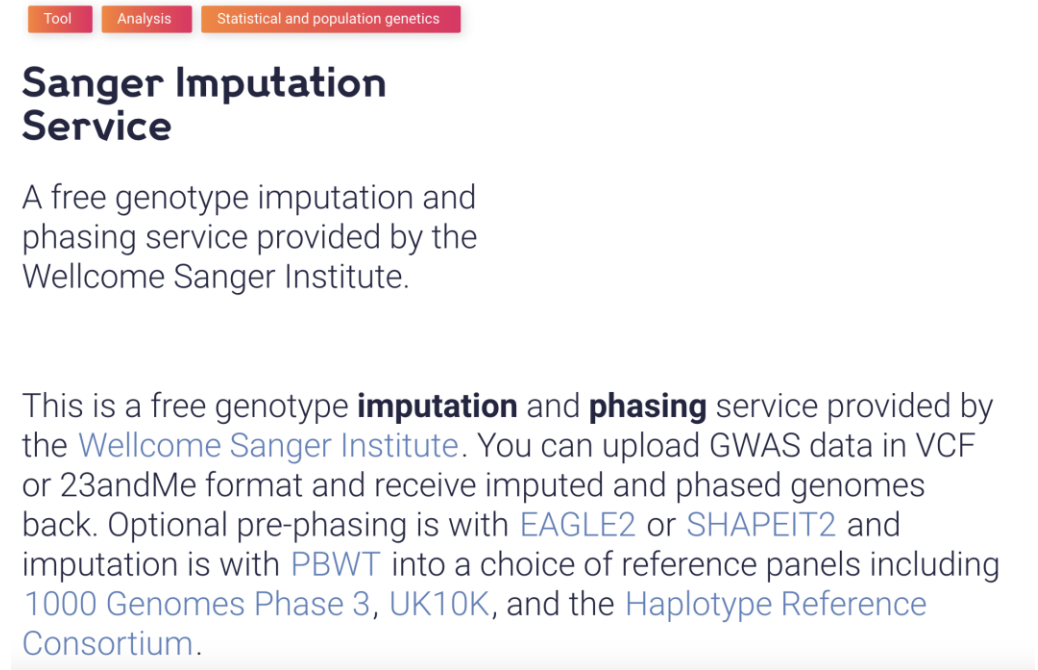


Fledel-Alon et al. (2011) *PLOS ONE*

Imputation

- SNP-chip data is typically imputed to full sequence. *Why?*
- Imputation requires a relevant reference dataset & phased genotypes
- In human genetics, can be done for *free* using online imputation servers
  - e.g. Michigan or Sanger imputation servers





Summary

GWAS use (a minimum of) ~1M carefully selected bi-allelic SNP from SNP-chips

- Increasingly this is imputed to full sequence and/or GWS data is available

Two important ‘equilibriums’

Within a locus: **Hardy-Weinberg equilibrium** test tells us about non-random genotype frequencies at a locus

Between loci: **Linkage disequilibrium** tells us of non-random association between two loci

HWE is typically used in GWAS context to detect genotyping errors

LD is useful/essential for GWAS & imputation

it tells us about population history

but is annoying for fine mapping ‘causal’ mutations

Practical Session

Choose Part 1 or 2

Part 1: LD between loci

Part 2: A simple GRM