GWAS Experimental Design: genotypes

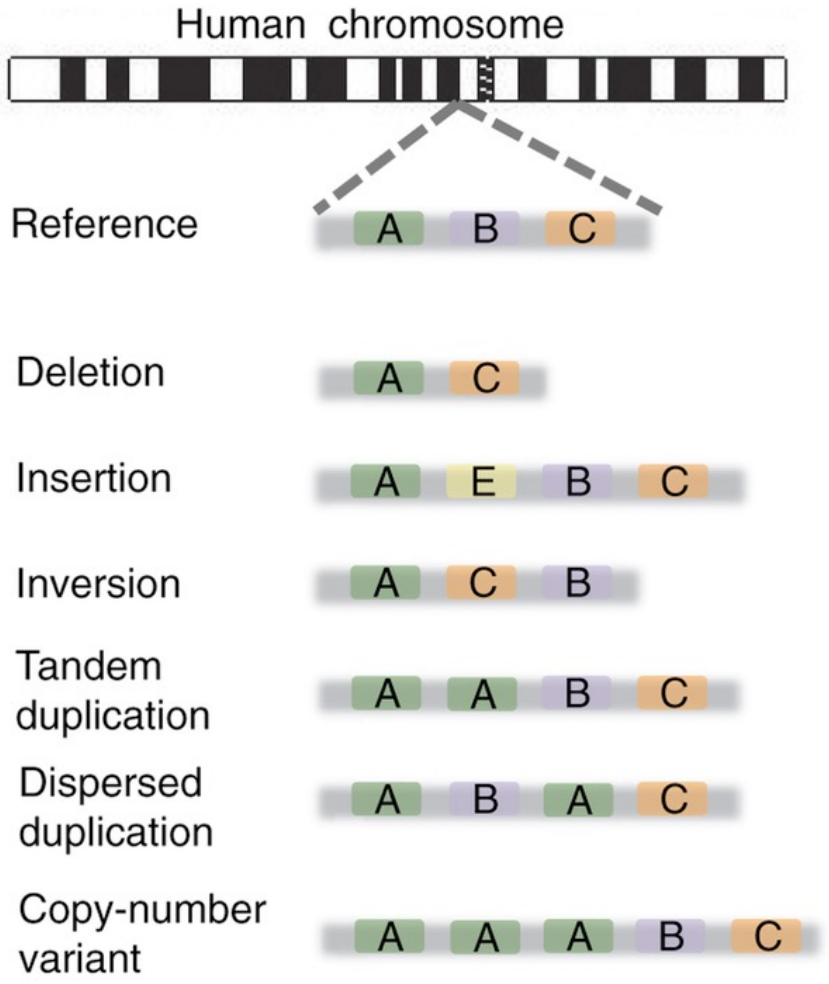
Outline of lecture

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

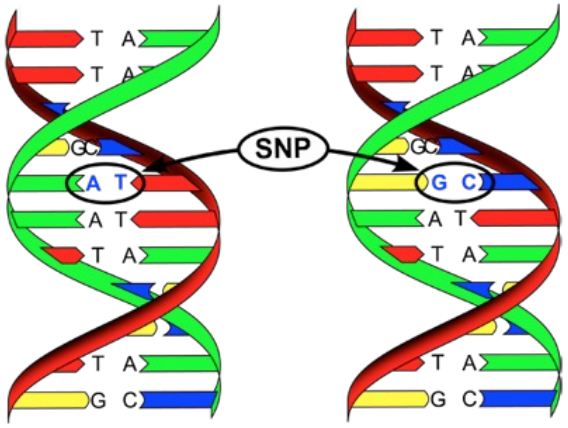
- Two types of ‘equilibrium’:
  - Hardy Weinberg Equilibrium
  - Linkage disequilibrium (LD)

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 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	ref	alt
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 this afternoon we will ‘clean’ the SNP chip genotypes
Missing genotypes
Check allele frequency
Check for Hardy-Weinberg inconsistencies

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

Hardy-Weinberg Equilibrium

- HWE is the probabilistic relationship between allele and genotype frequencies, i.e.

Consider a bi-allelic locus:

Alleles are A and a
freq(A) = p
freq(a) = 1 – p = q

Three possible genotypes:
AA, Aa, aa

Expected frequency of genotypes:
freq(AA) = p (allele1) x p (allele 2) = p^2
freq(Aa) = p (allele1) x q (allele 2) +
q (allele1) x p (allele 2) = 2pq
freq(aa) = q (allele1) x q (allele 2) = q^2

Does anybody recognize these types of probabilities?

Hardy-Weinberg Equilibrium

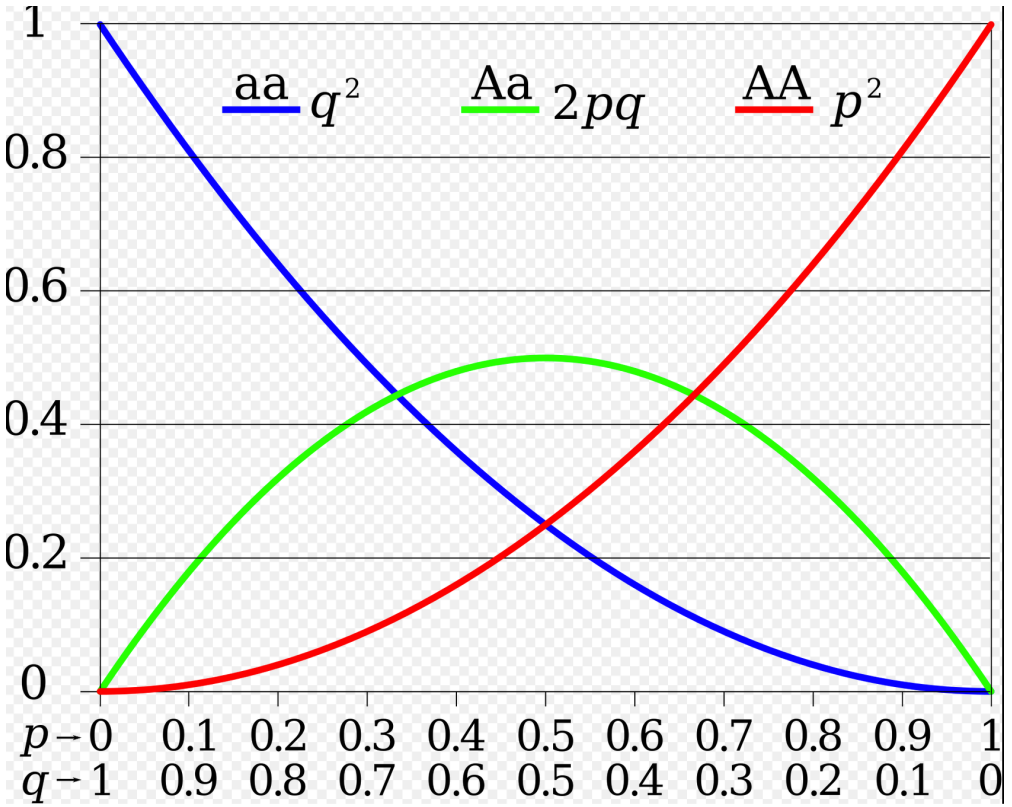
- HWE is the probabilistic relationship between allele and genotype frequencies, i.e.

Consider a bi-allelic locus:

Alleles are A and a
freq(A) = p
freq(a) = 1 – p = q

Three possible genotypes:
AA, Aa, aa

Expected frequency of genotypes:
freq(AA) = p (allele1) x p (allele 2) = p^2
freq(Aa) = p (allele1) x q (allele 2) +
q (allele1) x p (allele 2) = 2pq
freq(aa) = q (allele1) x q (allele 2) = q^2



Hardy-Weinberg principle wiki

Hardy-Weinberg Equilibrium

- HWE is the probabilistic relationship between allele and genotype frequencies, i.e.

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
- 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 genotypes 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:

Two markers A and B on the same chromosome
Alleles are:

- marker A: A1, A2
- marker B: B1, B2

Possible haplotypes are A1_B1, A1_B2, A2_B1, A2_B2

Definitions of LD

Linkage equilibrium…….

		Marker A		
		A1	A2	Frequency
Marker B	B1			0.5
	B2			0.5
	Frequency	0.5	0.5	

Definitions of LD

Linkage equilibrium…….

		Marker A		
		A1	A2	Frequency
Marker B	B1	0.25	0.25	0.5
	B2	0.25	0.25	0.5
	Frequency	0.5	0.5	

Definitions of LD

Linkage disequilibrium…….

		Marker A		
		A1	A2	Frequency
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Frequency	0.5	0.5	

Definitions of LD

Linkage disequilibrium…….

		Marker A		
		A1	A2	Frequency
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Frequency	0.5	0.5	

D = freq(A1_B1)*freq(A2_B2)-freq(A1_B2)*freq(A2_B1)

= 0.4 * 0.4 - 0.1 * 0.1

= 0.15

Definitions of LD

D = freq(A1_B1)*freq(A2_B2) - freq(A1_B2)*freq(A2_B1)

= 0.4 * 0.4 - 0.1 * 0.1

= 0.15

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

$r^2 = D^2/[freq(A1)*freq(A2)*freq(B1)*freq(B2)]$

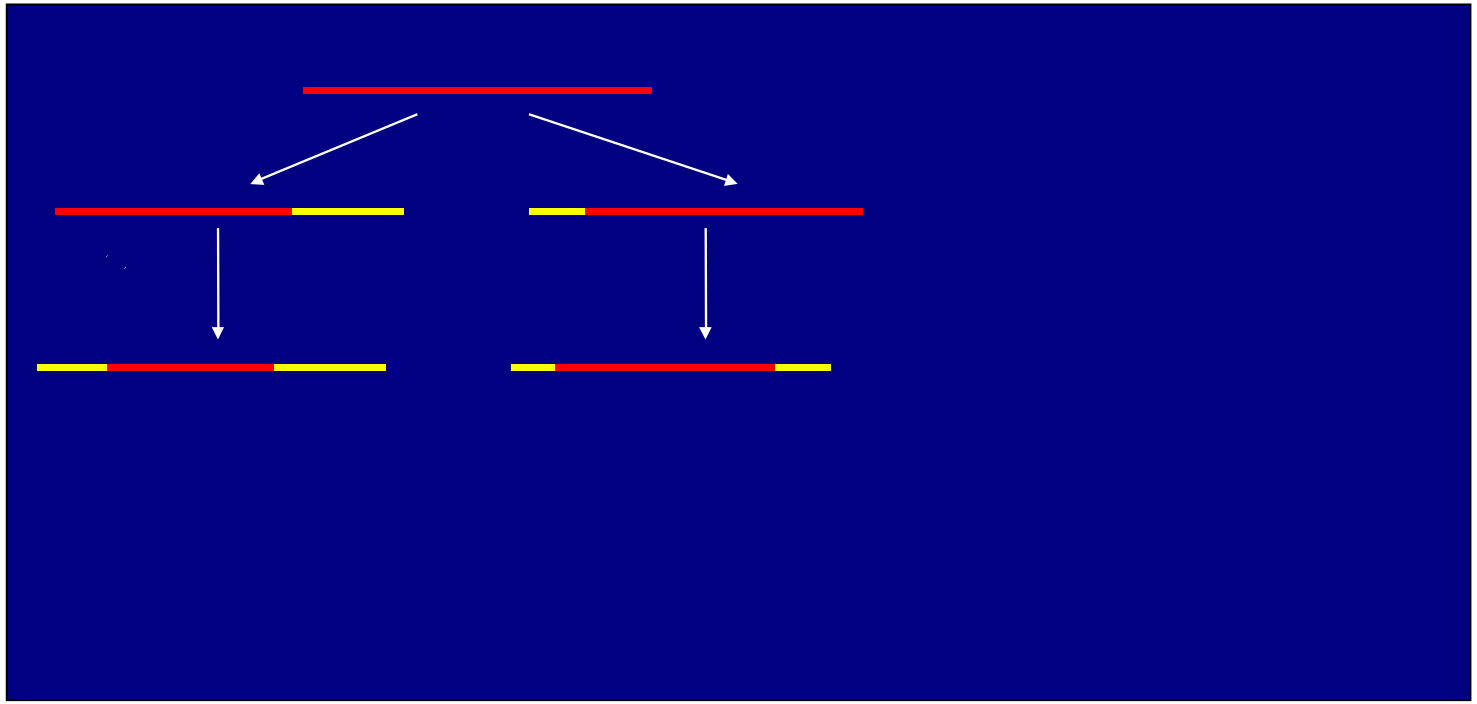
r^2 ranges from [0,1]

it is equivalent to the **squared correlation co-efficient** between alleles

- i.e. given the first allele, how well can we predict the second allele?

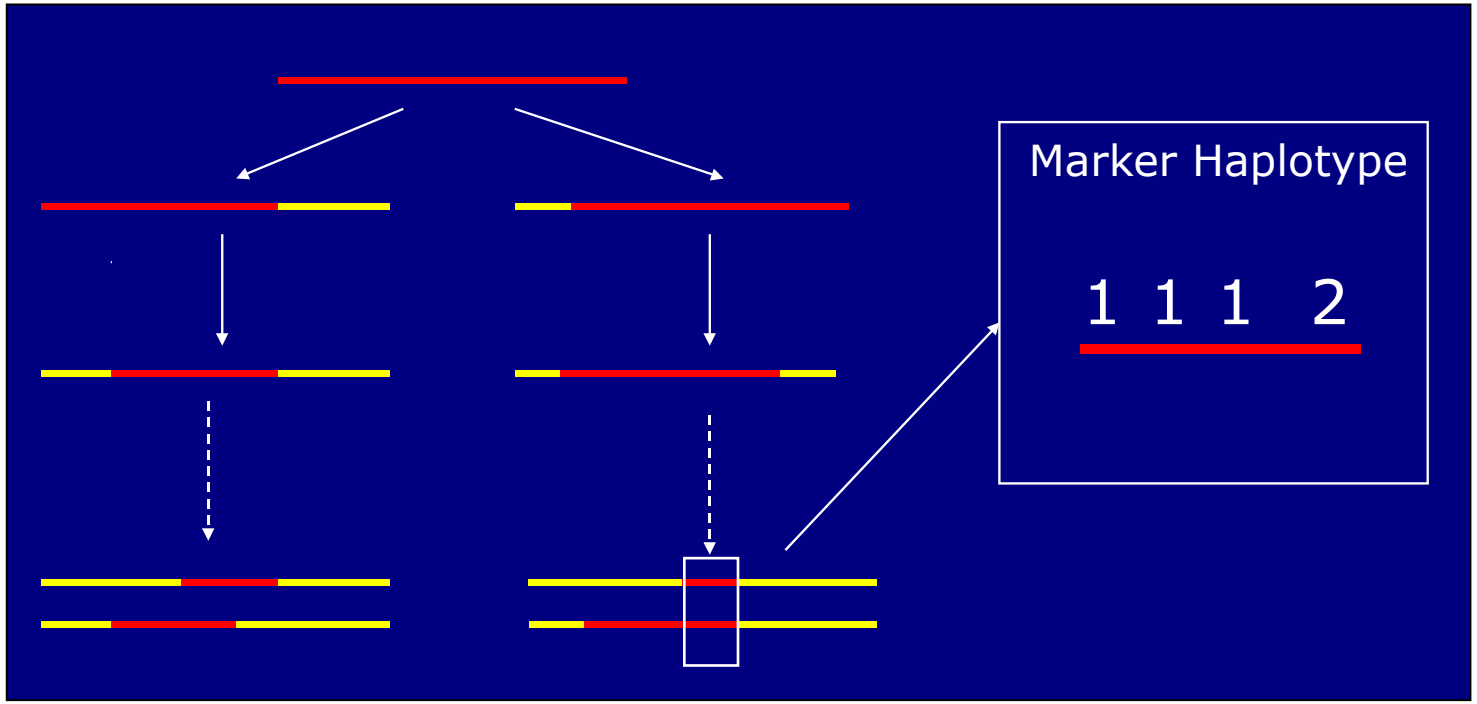
Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



Causes of LD

- A chunk of ancestral chromosome is conserved in the current population



Extent of LD is population dependent

Which within population LD is likely to be relatively high or low?

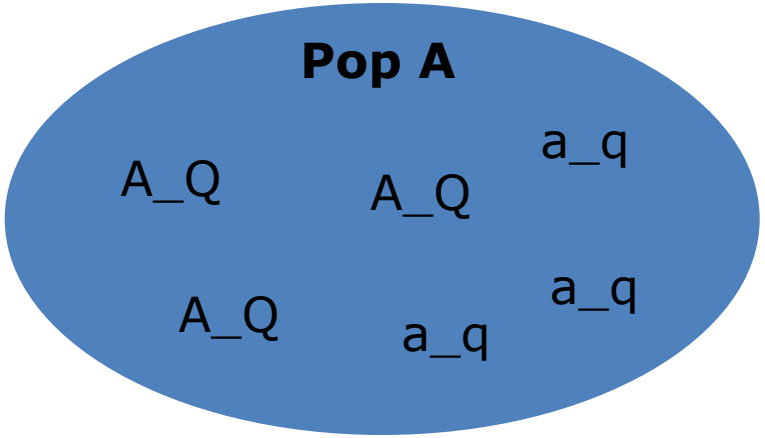
Commercial wheat vs. wild relative

African ancestry vs. European ancestry

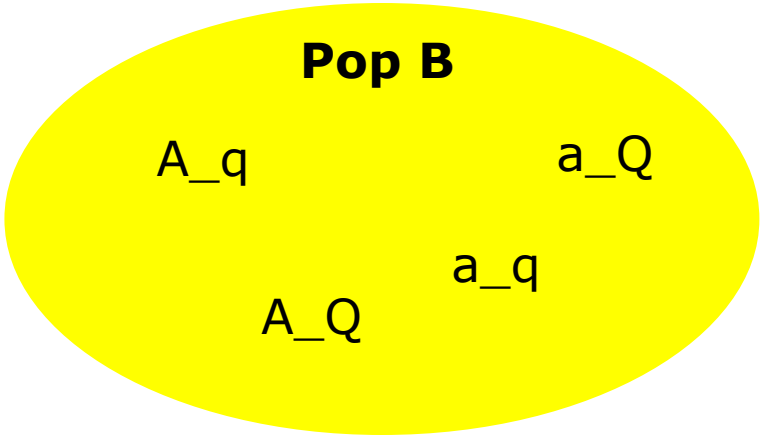
European ancestry vs. Finnish ancestry

LD is population dependent

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

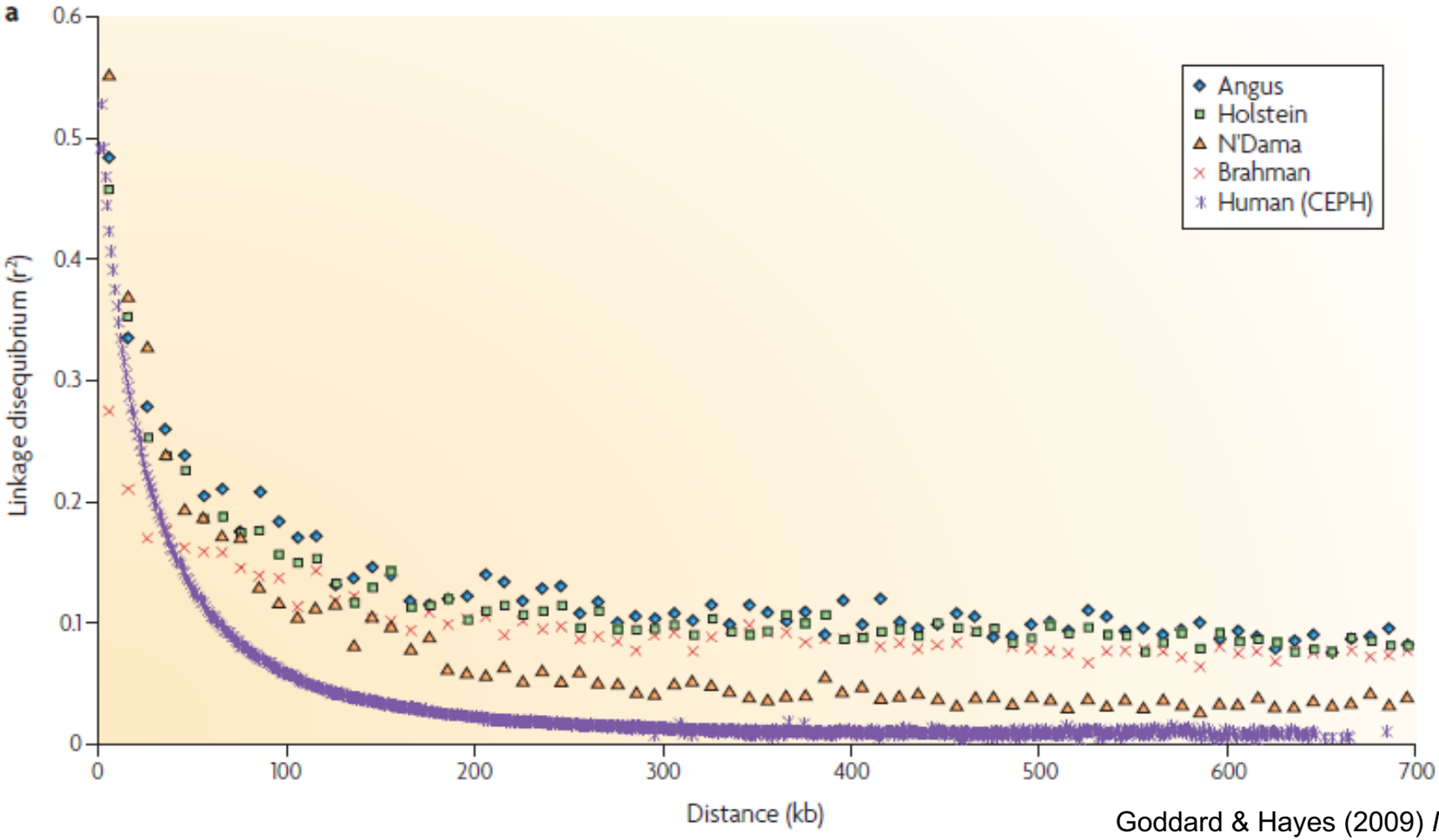


marker ‘A’ in linkage with causative mutation ‘Q’



marker ‘A’ in linkage equilibrium with ‘Q’

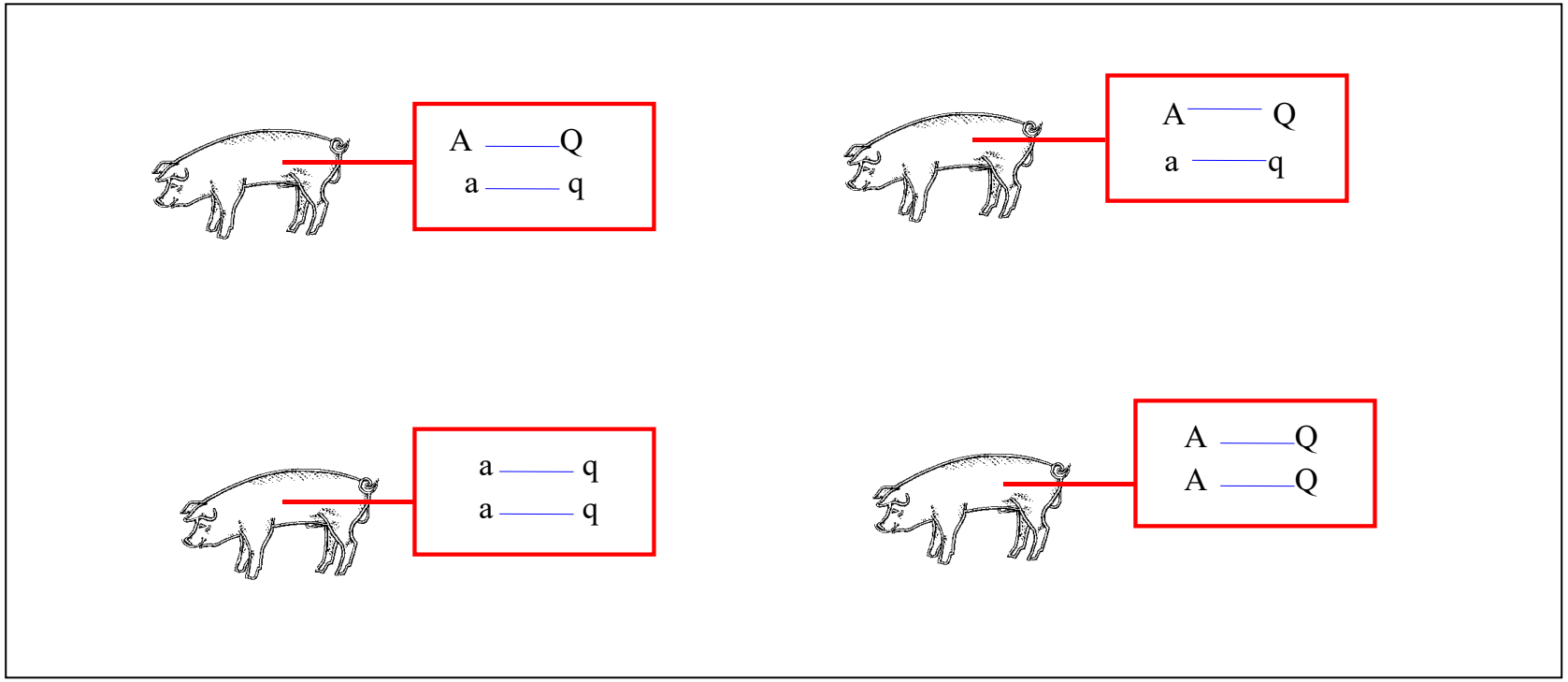
Extent of LD is population dependent



Goddard & Hayes (2009) *Nature Reviews Genetics*

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)

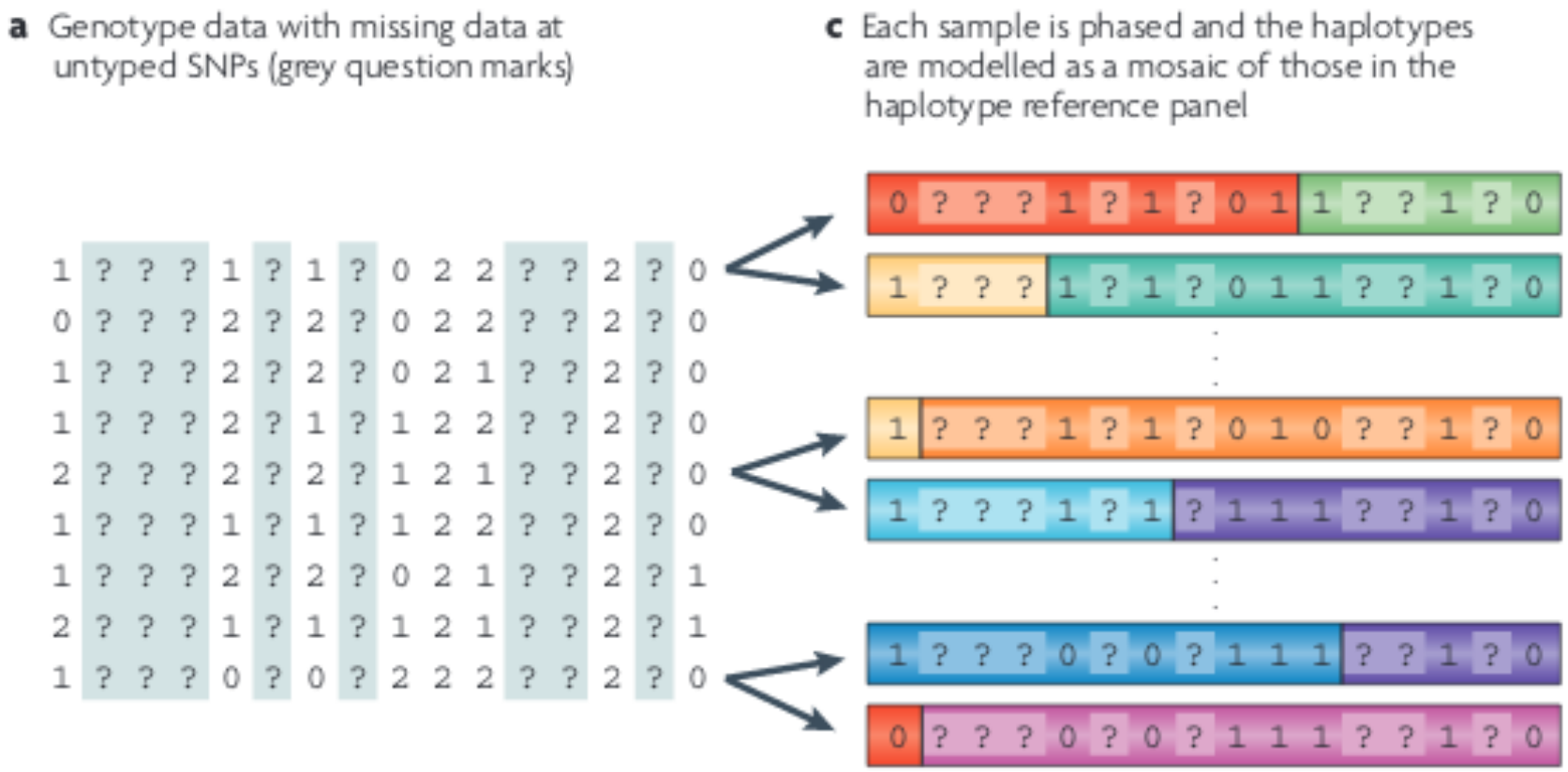
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

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

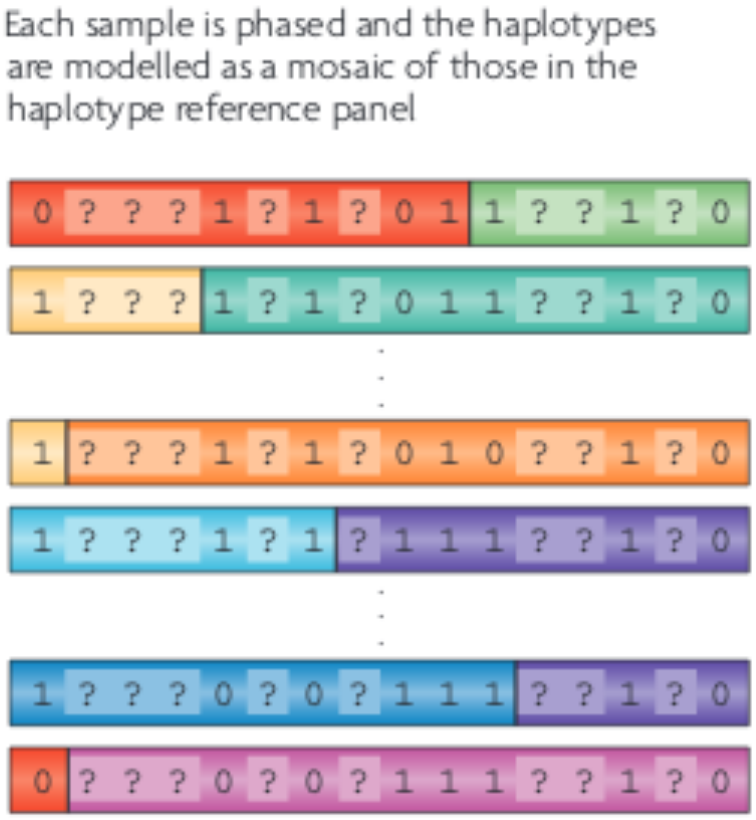
Why do we care about LD?

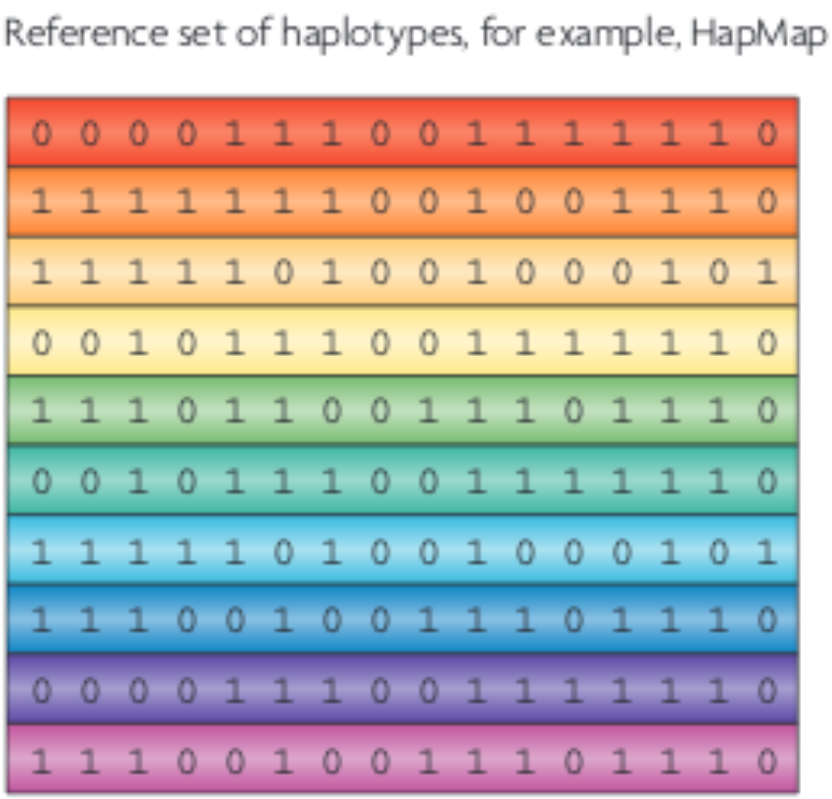
2. We can use LD to fill in missing genotypes via imputation



Why do we care about LD?

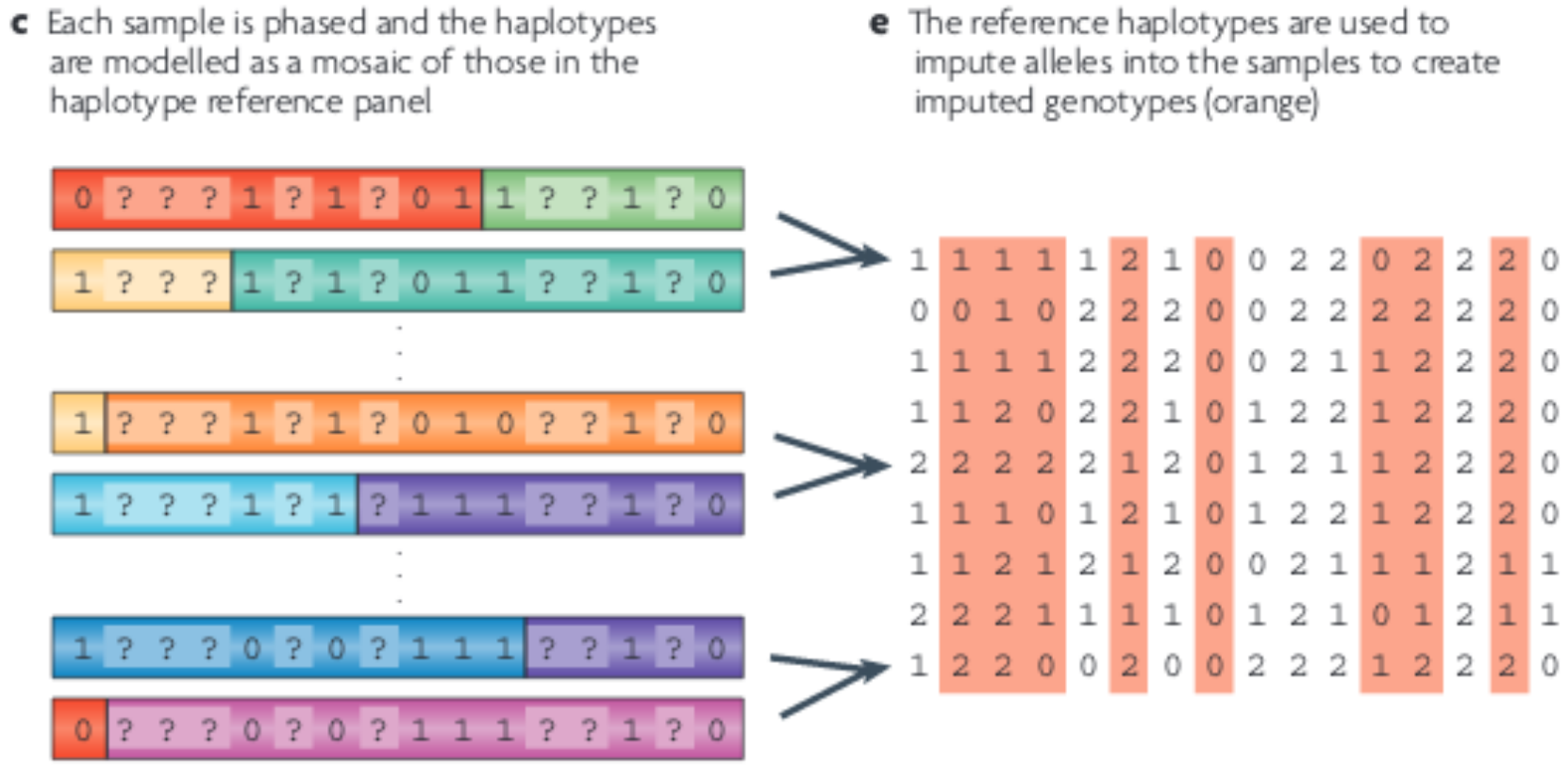
2. We can use LD to fill in missing genotypes via imputation





Why do we care about LD?

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



Why do we care about LD?

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

Imputation is used to:

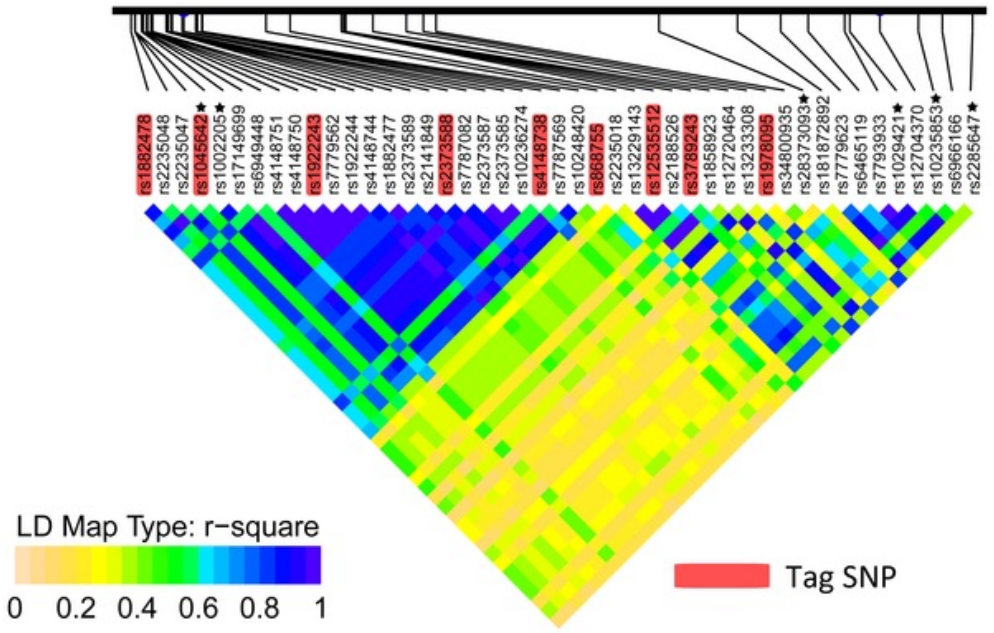
- fill in missing data, i.e. SNP removed during QC or poorly genotyped in some samples
- completely impute genotyped not genotyped but in the reference panel

Imputed SNPs can be used in GWAS like genotyped SNPs

- increases the power to detect associations

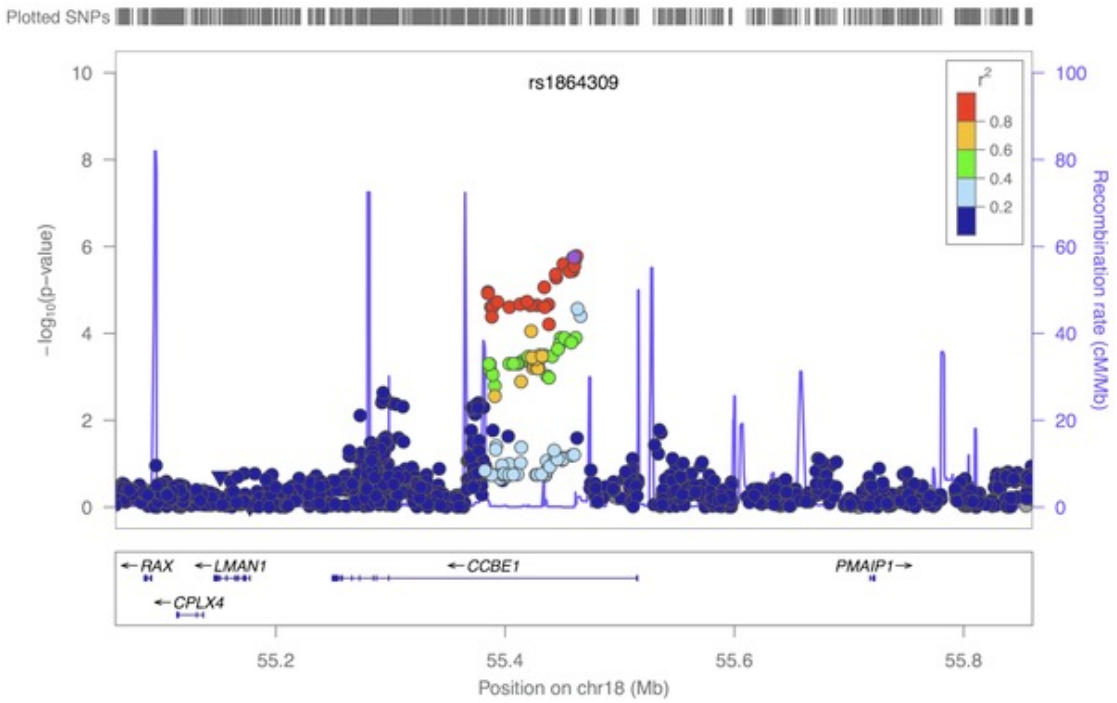
Representing LD in the GWAS

- Pair-wise LD plot



Shou et al. (2012) *PLOS ONE*

- Recombination graphs



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

Summary

- GWAS typically use ~1M carefully selected bi-allelic SNP from SNP-chips

- 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 also tells us about population history
  - but is annoying for identifying fine mapping