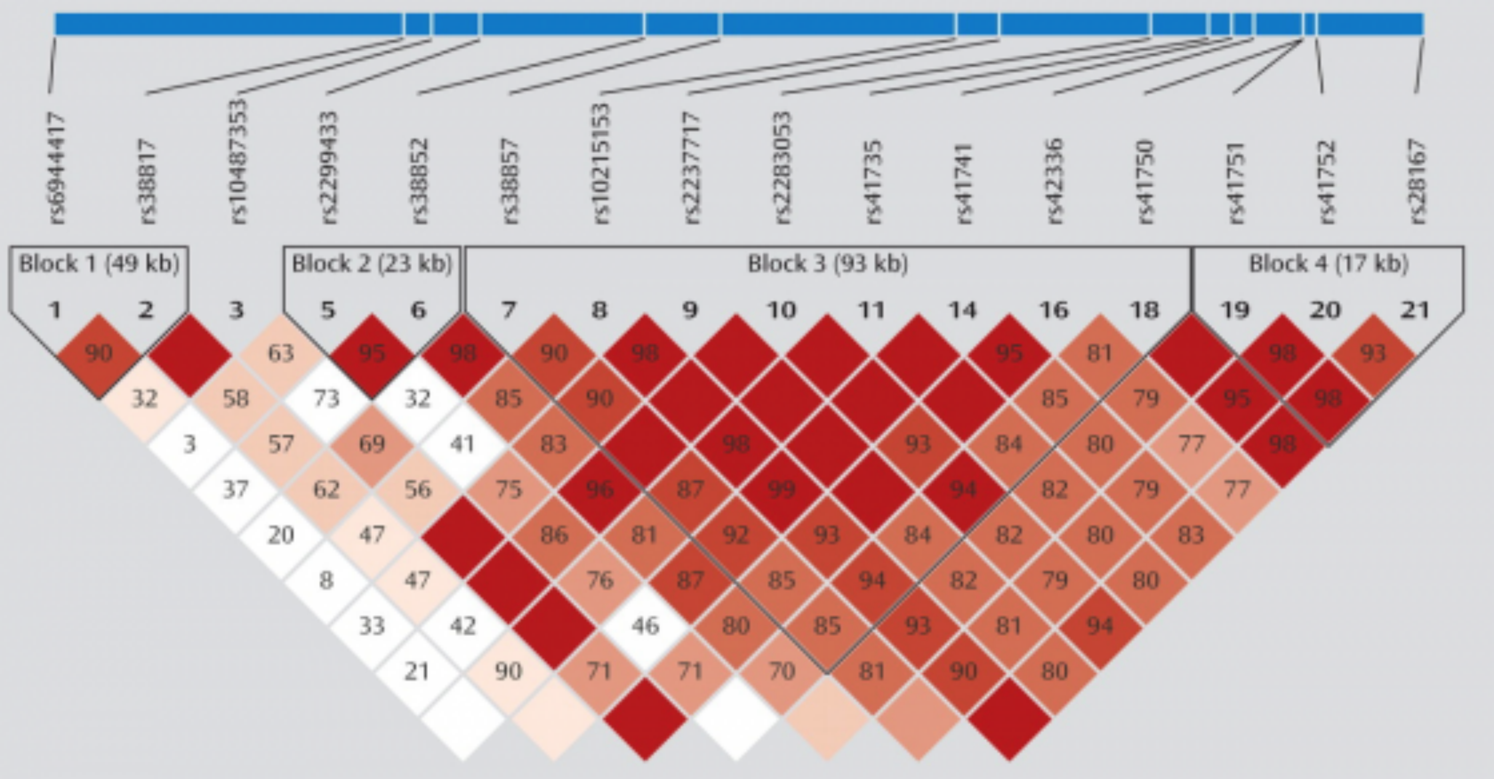
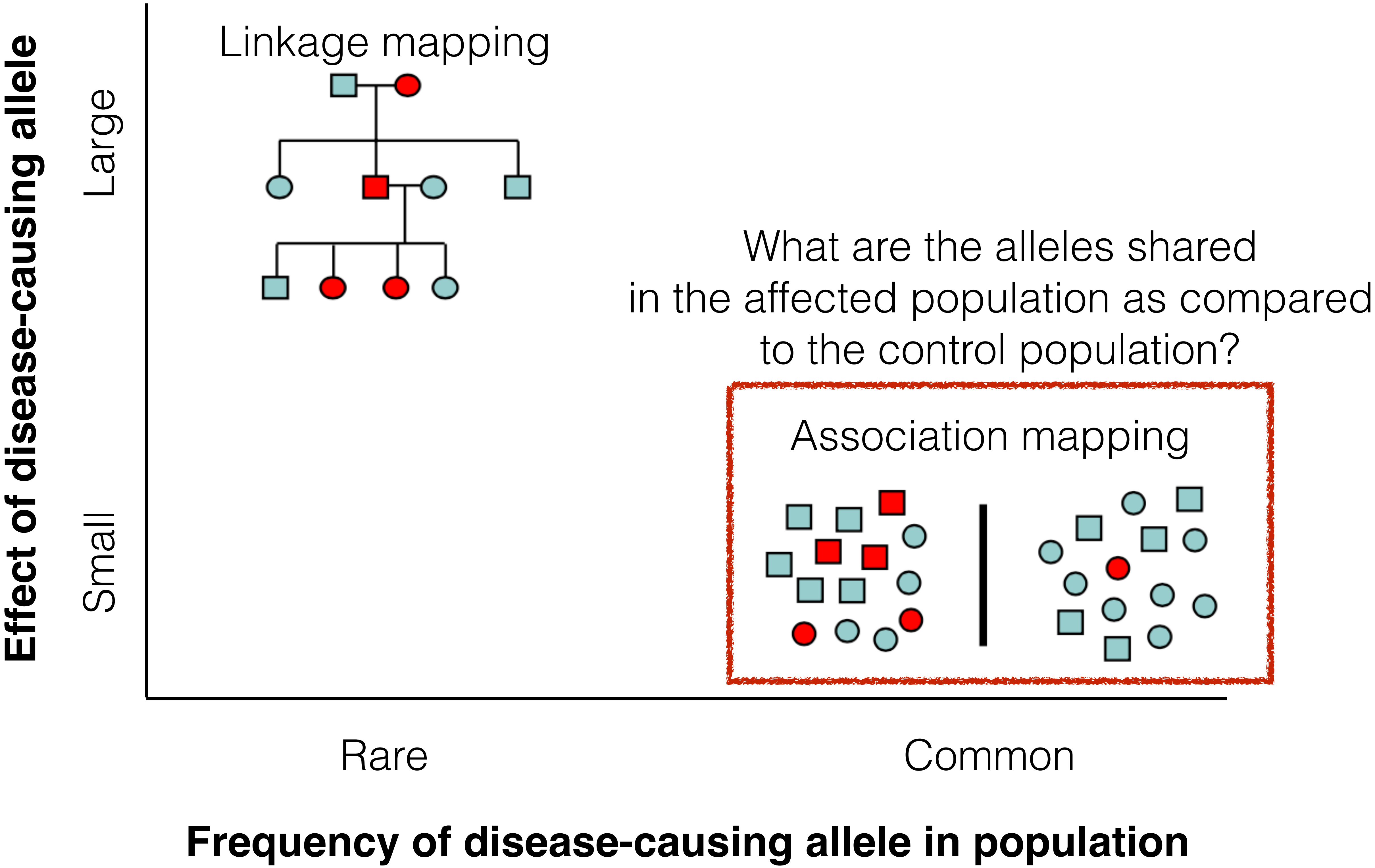
Linkage disequilibrium, haplotypes, and GWAS



Human gene mapping has two general flavors



Linkage mapping

What are the alleles shared
in the affected population as compared
to the control population?

Association mapping

Effect of disease-causing allele

Large

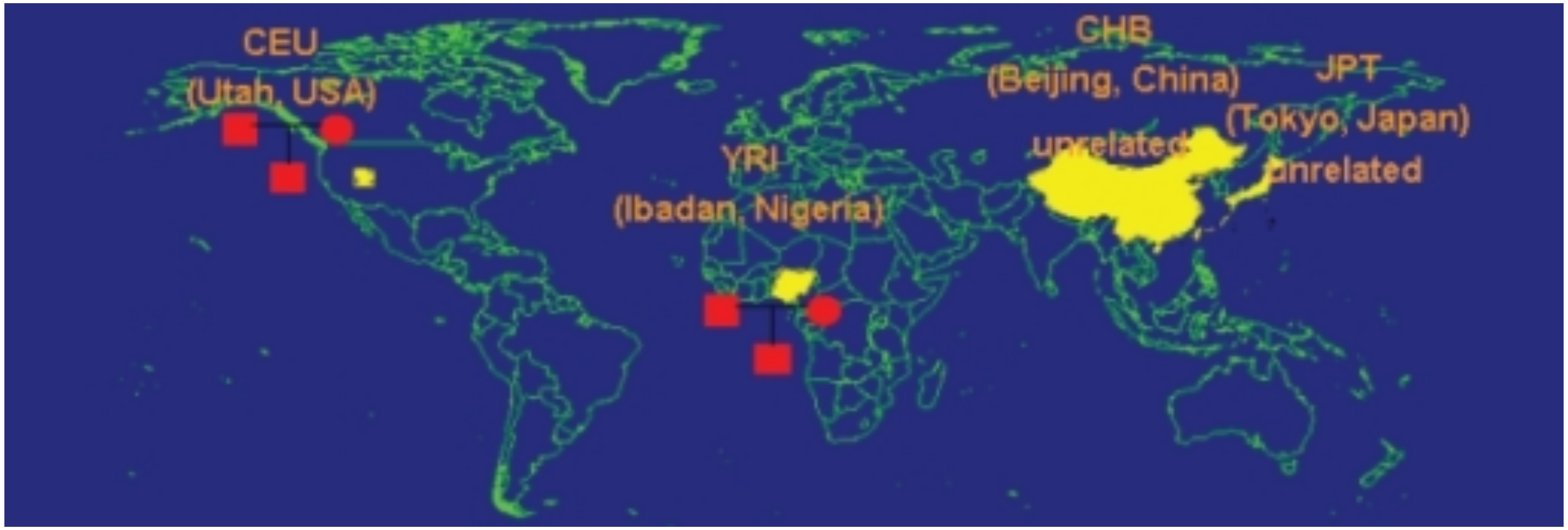
Small

Rare

Common

Frequency of disease-causing allele in population

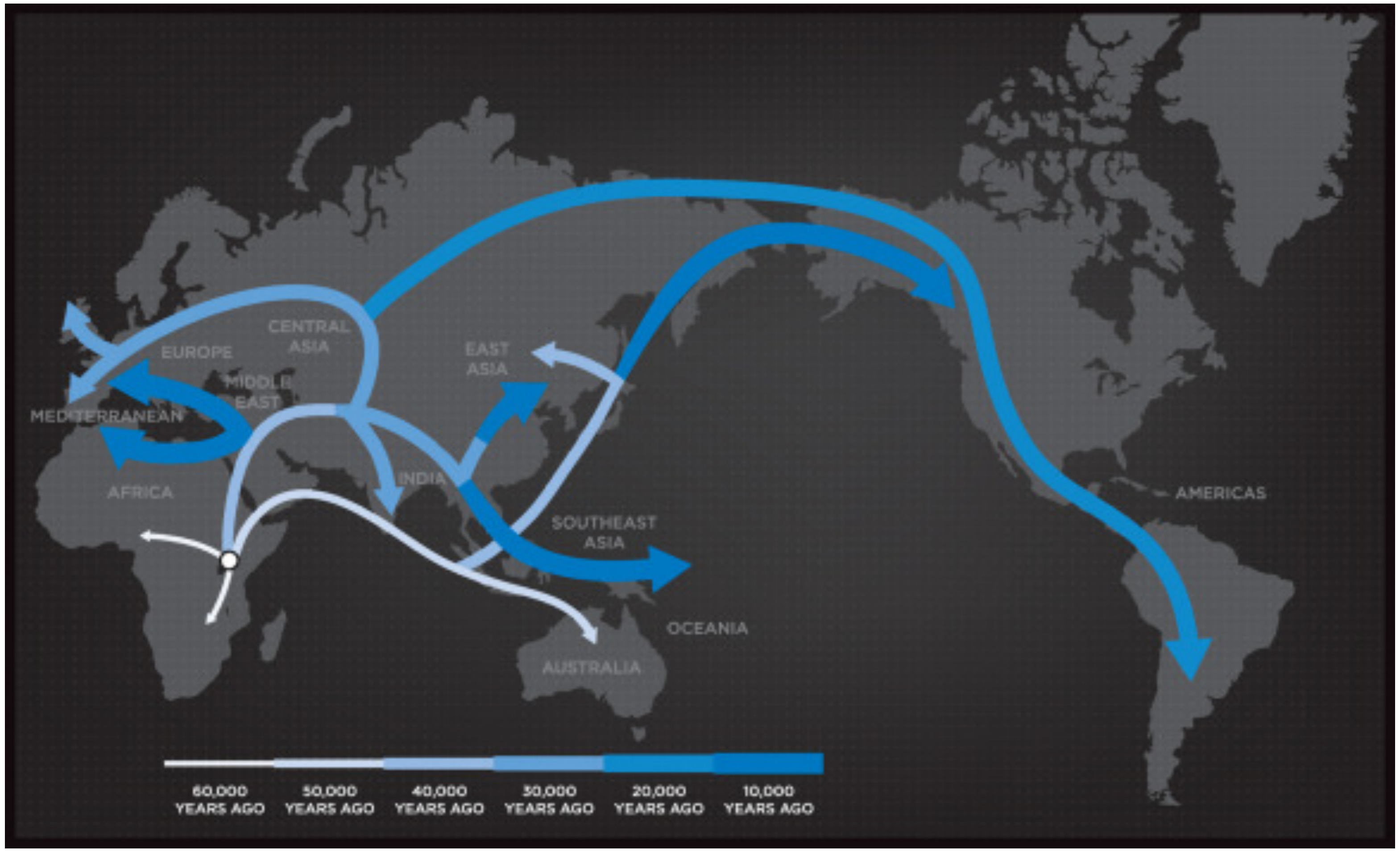
Common variants facilitate genome-wide association (GWA) mapping



The Human Haplotype Map (HapMap) identified
10 million common variants

Do we have to test them all?

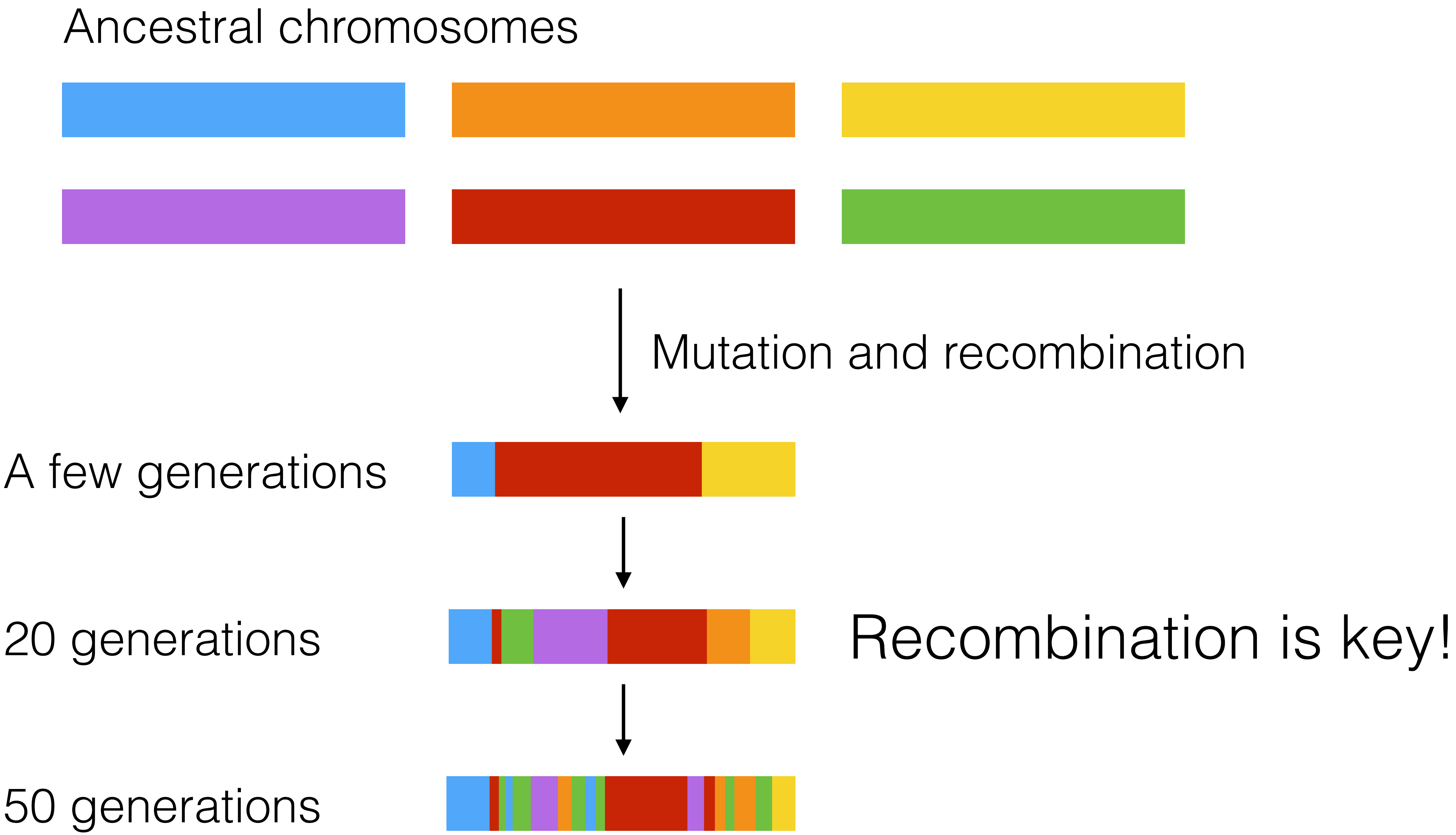
Common variants facilitate genome-wide association (GWA) mapping



Our relatedness means that variants are correlated in populations

Correlation between variants is called linkage disequilibrium (LD)

Linkage disequilibrium (LD) is the non-random association of alleles at different loci



Ancestral chromosomes

Mutation and recombination

A few generations

20 generations

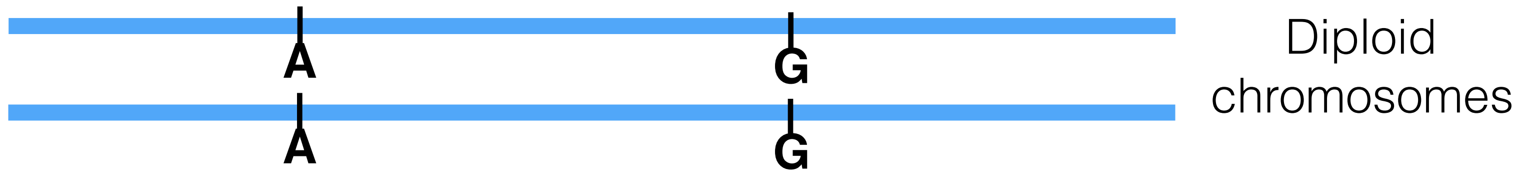
Recombination is key!

50 generations

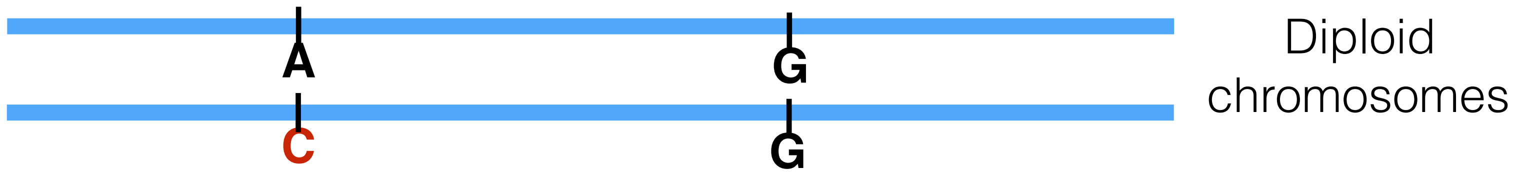
LD makes genotyping easier and cheaper

Many alleles that exist today are from ancient mutation events

Before mutation

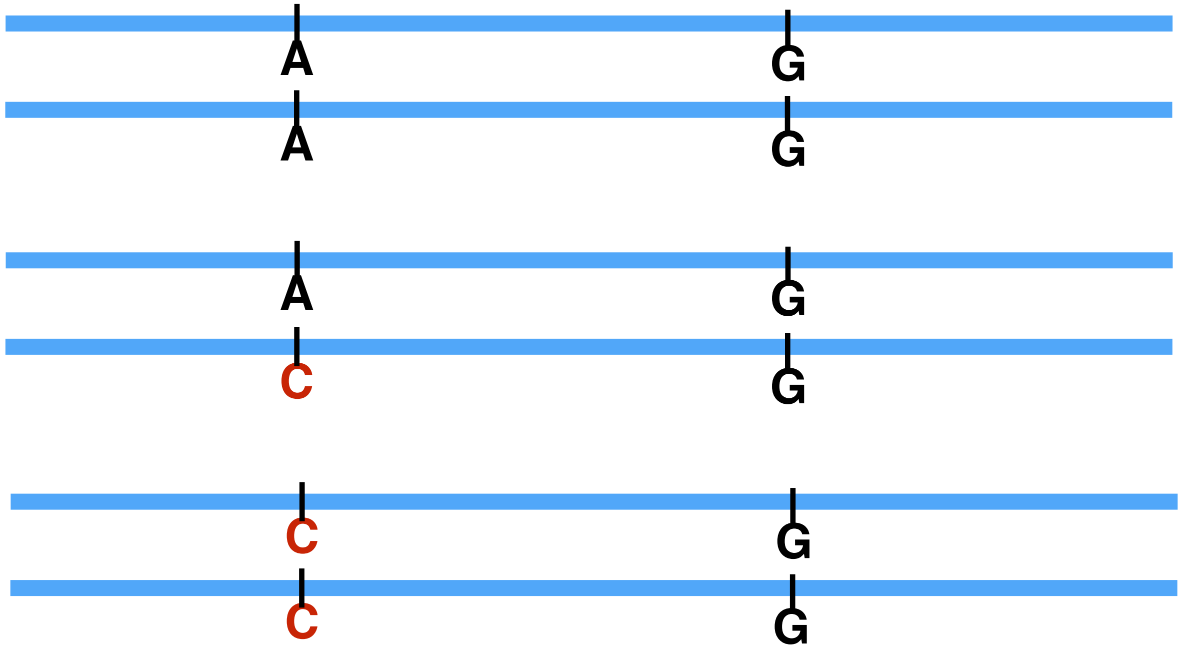


After mutation

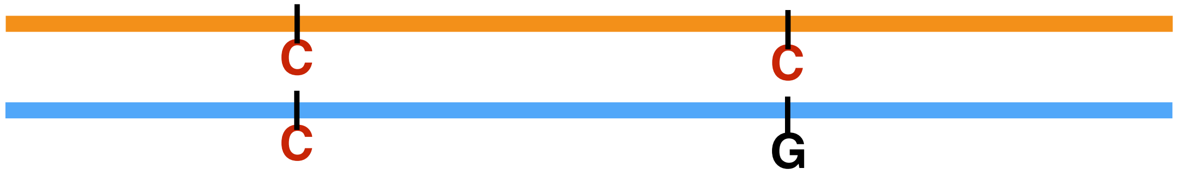


That allele spreads throughout the population, then another mutation occurs

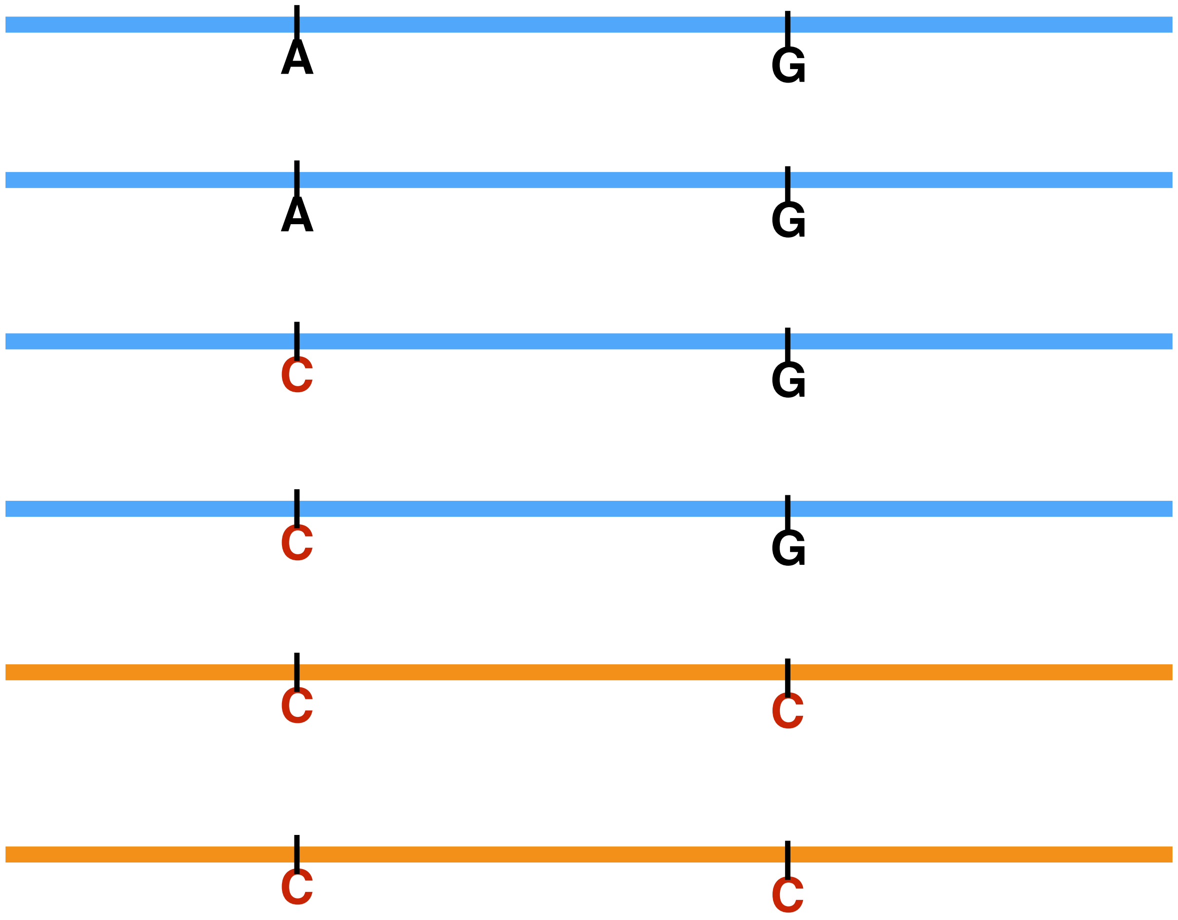
Before mutation



After mutation



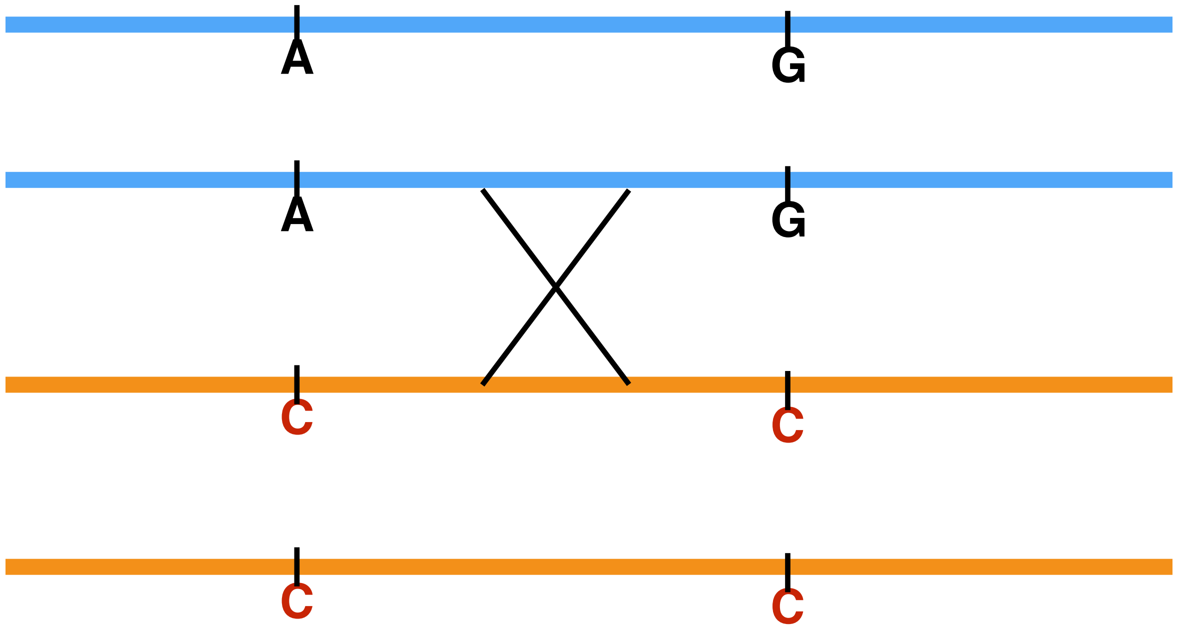
Let's think about these chromosomes with different arrangements of alleles as haploid gametes



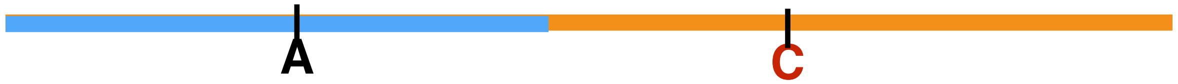
Mutations arose in particular genetic backgrounds,
so not every allelic combination is present

Recombination creates new arrangements of ancestral alleles

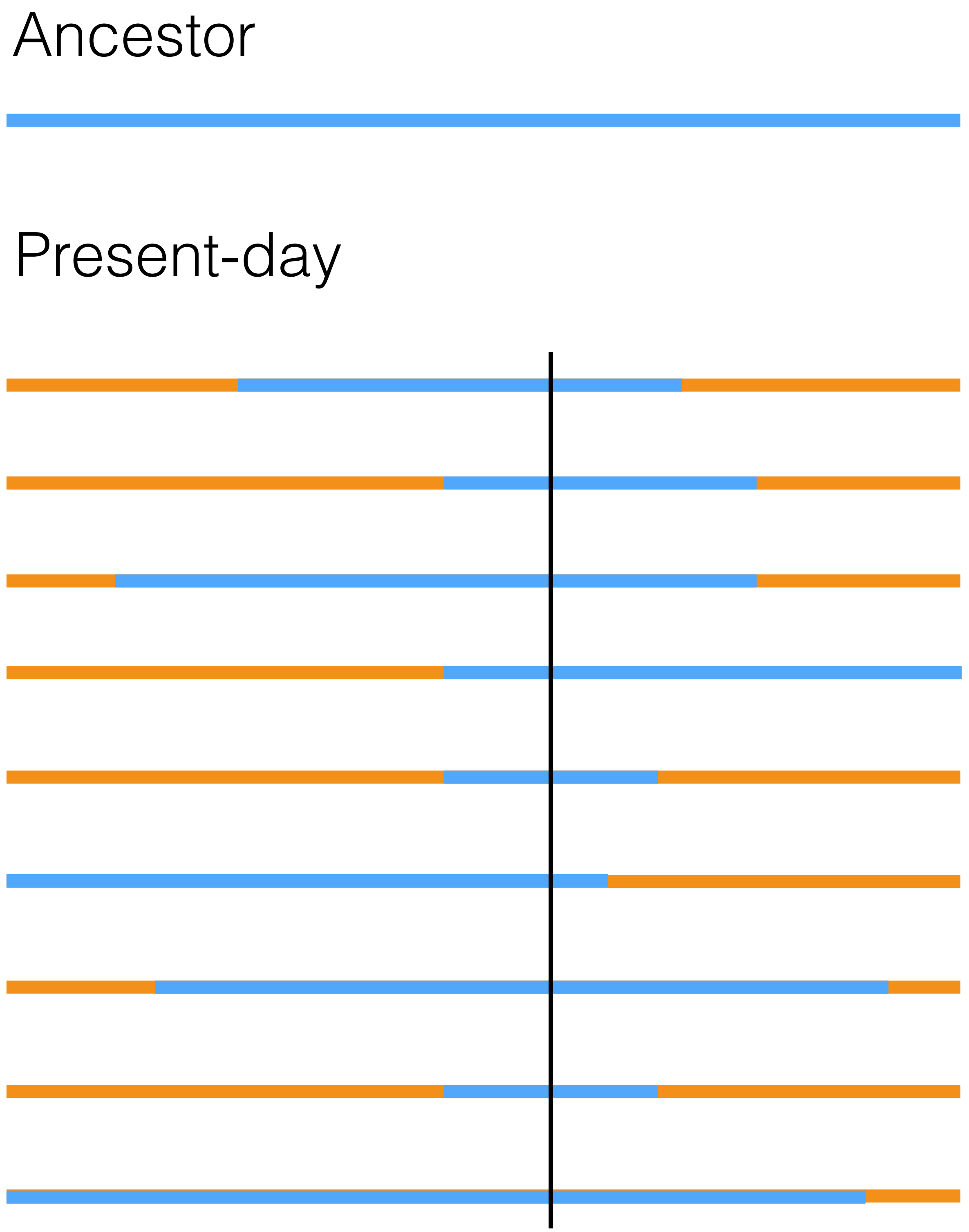
Before recombination



After recombination



Linkage disequilibrium is the non-random association of alleles at different loci



Ancestor

Present-day

Chromosomes are mosaics

Degree of mosaicism depends on:

- Recombination rate
- Mutation rate
- Population size
- Natural selection

Combinations of linked alleles close together reflect ancestral haplotypes

Haplotype frequencies in a population

Let’s say we have two linked loci (rs1 and rs2) that each have two alleles (A or a and B or b)

Four combinations exist:

A	**B**
A	**b**
a	**B**
a	**b**

p_A = frequency of A in the population or proportion of gametes with A

$p_a = 1 - p_A$

p_B = frequency of B in the population or proportion of gametes with B

$p_b = 1 - p_B$

p_{AB} = frequency of A and B occurring together in the same gamete
or frequency of the AB haplotype

These numbers come from genotyping populations

Haplotype frequencies in a population

Let’s say we have two linked loci (rs1 and rs2) that each have two alleles (A or a and B or b)

p_A = frequency of A in the population or proportion of gametes with A

$p_a = 1 - p_A$

p_B = frequency of B in the population or proportion of gametes with B

$p_b = 1 - p_B$

p_{AB} = frequency of A and B occurring together in the same gamete
or frequency of the AB haplotype

At equilibrium, the probability of A and B occurring together is the just probability that A and B independently occur in the same gamete

$$p_A * p_B$$

If $p_A * p_B \mathrel{!}= p_{AB}$, then non-random association or disequilibrium is observed

Haplotype frequencies in a population

p_A = frequency of A in the population or proportion of gametes with A

$p_a = 1 - p_A$

p_B = frequency of B in the population or proportion of gametes with B

$p_b = 1 - p_B$

p_{AB} = frequency of A and B occurring together in the same gamete
or frequency of the AB haplotype

$$p_{AB} = p_A * p_B$$

Locus rs2

		B	b	
Locus rs1	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
		p_B	p_b	

How to calculate LD?

The Disequilibrium coefficient $D_{rs1\text{-}rs2}$

$D_{rs1\text{-}rs2} = p_{AB} - p_A * p_B$

When in equilibrium, $D_{rs1\text{-}rs2} = 0$

Otherwise, $D_{rs1\text{-}rs2} >$ or < 0

The sign is arbitrary. Set rs1, rs2 to the common alleles.

Range depends on allele frequencies, so comparisons between different pairs of markers are difficult.

How to calculate LD?

The correlation is the preferred term:

$$r^2 = (D_{rs1\text{-}rs2})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{rs1\text{-}rs2} = p_{AB} - p_A * p_B$

Ranges between 0 and 1
with 0 being equilibrium and 1 being perfect linkage

How to calculate LD? An example

$r^2 = (D_{rs1\text{-}rs2})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$ Remember $D_{rs1\text{-}rs2} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers rs1 and rs2 with two forms A or a and B or b?

We genotype 500 people to get:

Haplotype	Number
AB	600
Ab	100
aB	200
ab	100

Convert to numbers of alleles into haplotype frequencies:

Haplotype	Number	Frequency
AB	600	0.6
Ab	100	0.1
aB	200	0.2
ab	100	0.1

How to calculate LD? An example

$r^2 = (D_{rs1\text{-}rs2})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$ Remember $D_{rs1\text{-}rs2} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers rs1 and rs2 with two forms A or a and B or b?

Frequencies of haplotypes:

Haplotype	Number	Frequency
AB	600	0.6
Ab	100	0.1
aB	200	0.2
ab	100	0.1

Convert to frequencies of alleles:

$p_A = p(AB) + p(Ab)$
$p_a = 1 - p_A$
$p_B = p(AB) + p(aB)$
$p_b = 1 - p_B$

Allele	Number	Frequency
A	700	0.7
a	300	0.3
B	800	0.8
b	200	0.2

How to calculate LD? An example

$r^2 = (D_{rs1\text{-}rs2})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$ Remember $D_{rs1\text{-}rs2} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers rs1 and rs2 with two forms A or a and B or b?

Convert to frequencies of alleles:

Allele	Number	Frequency
A	700	0.7
a	300	0.3
B	800	0.8
b	200	0.2

Haplotype	Number	Frequency
AB	600	0.6
Ab	100	0.1
aB	200	0.2
ab	100	0.1

$p_A = p(AB) + p(Ab)$

$p_a = 1 - p_A$

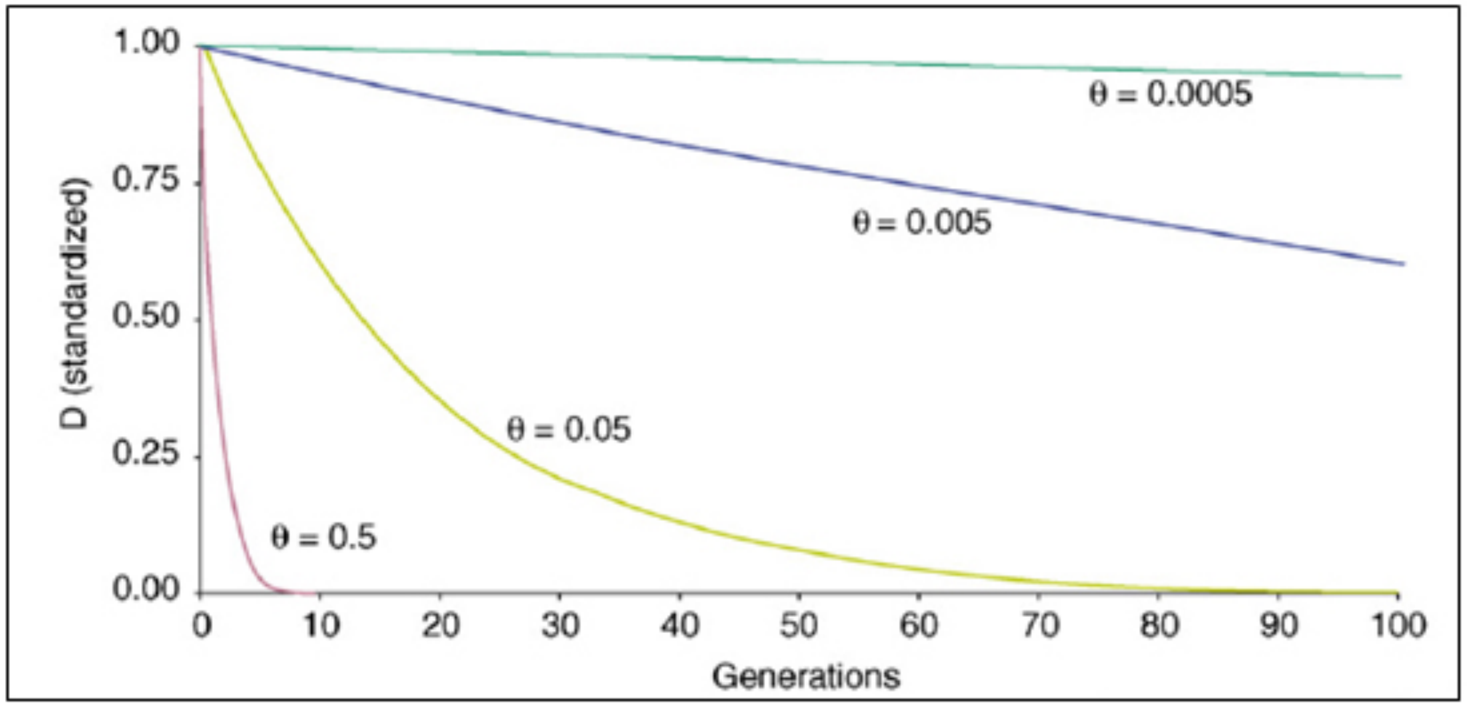
$p_B = p(AB) + p(aB)$

$p_b = 1 - p_B$

$$D_{rs1\text{-}rs2} = 0.6 - 0.7 * 0.8 = 0.04$$

$$r^2 = 0.04^2 / (0.7 * 0.3 * 0.8 * 0.2) = 0.048$$

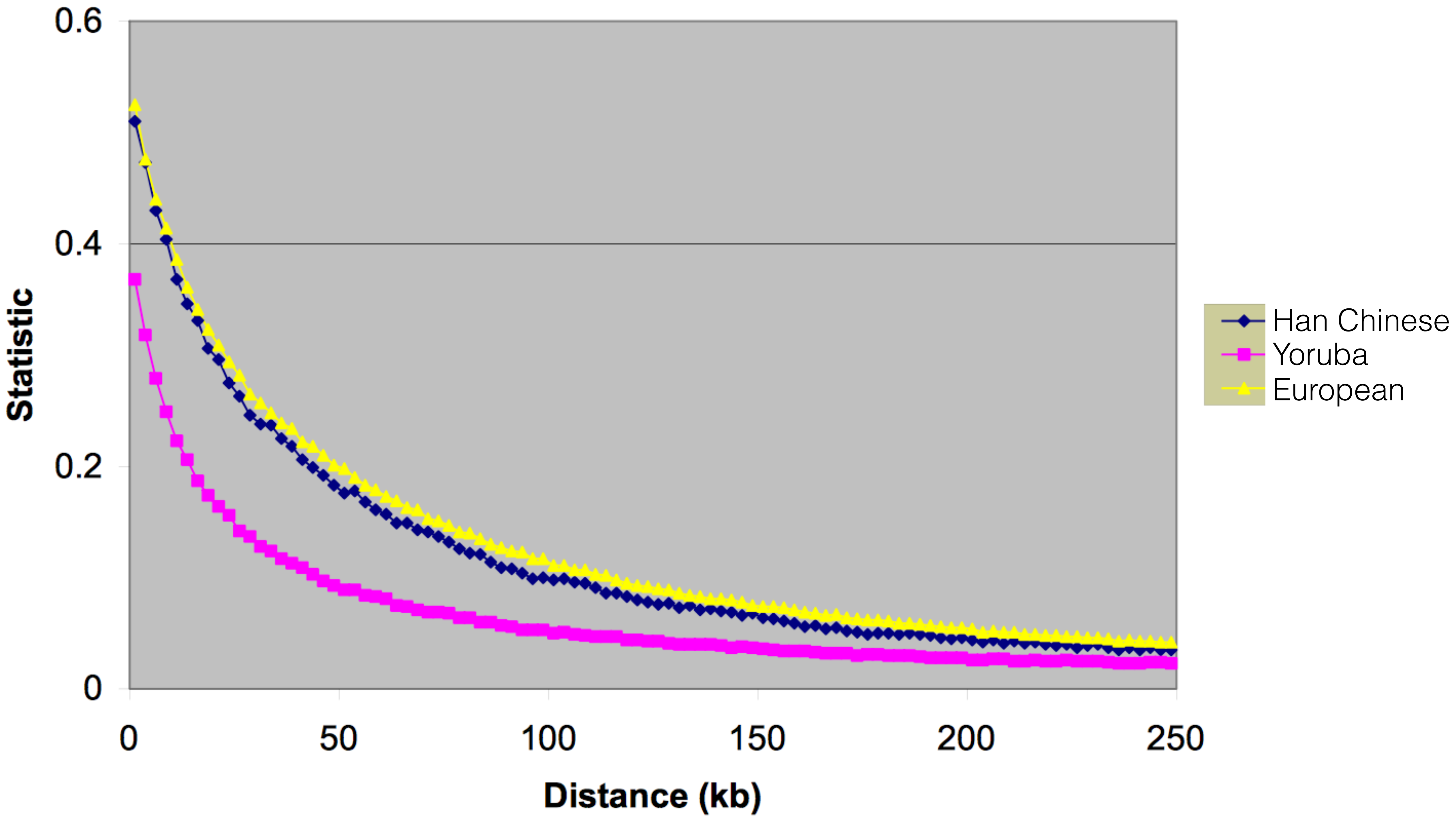
Linkage disequilibrium decreases by distance and generation time



Mackay and Powell 2007

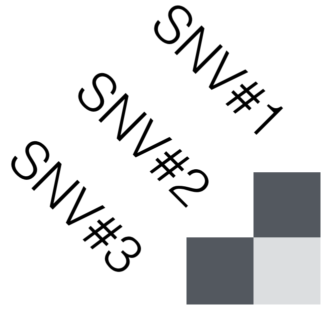
Recombination is key!

Linkage disequilibrium varies among different populations

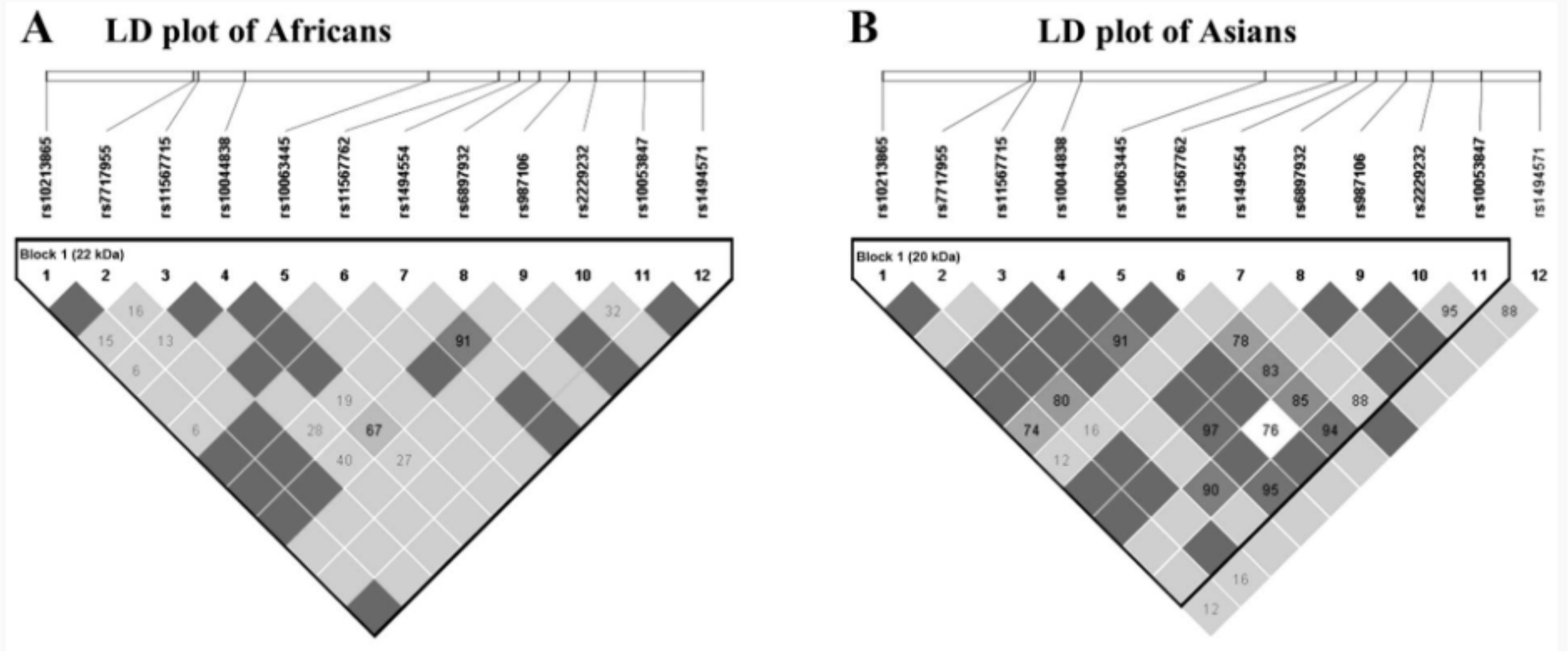


Recombination is key!

Linkage disequilibrium is often shown as a triangle correlation plot



SNV#1 and SNV#2 have high LD
SNV#2 and SNV#3 have high LD
SNV#1 and SNV#3 have low LD

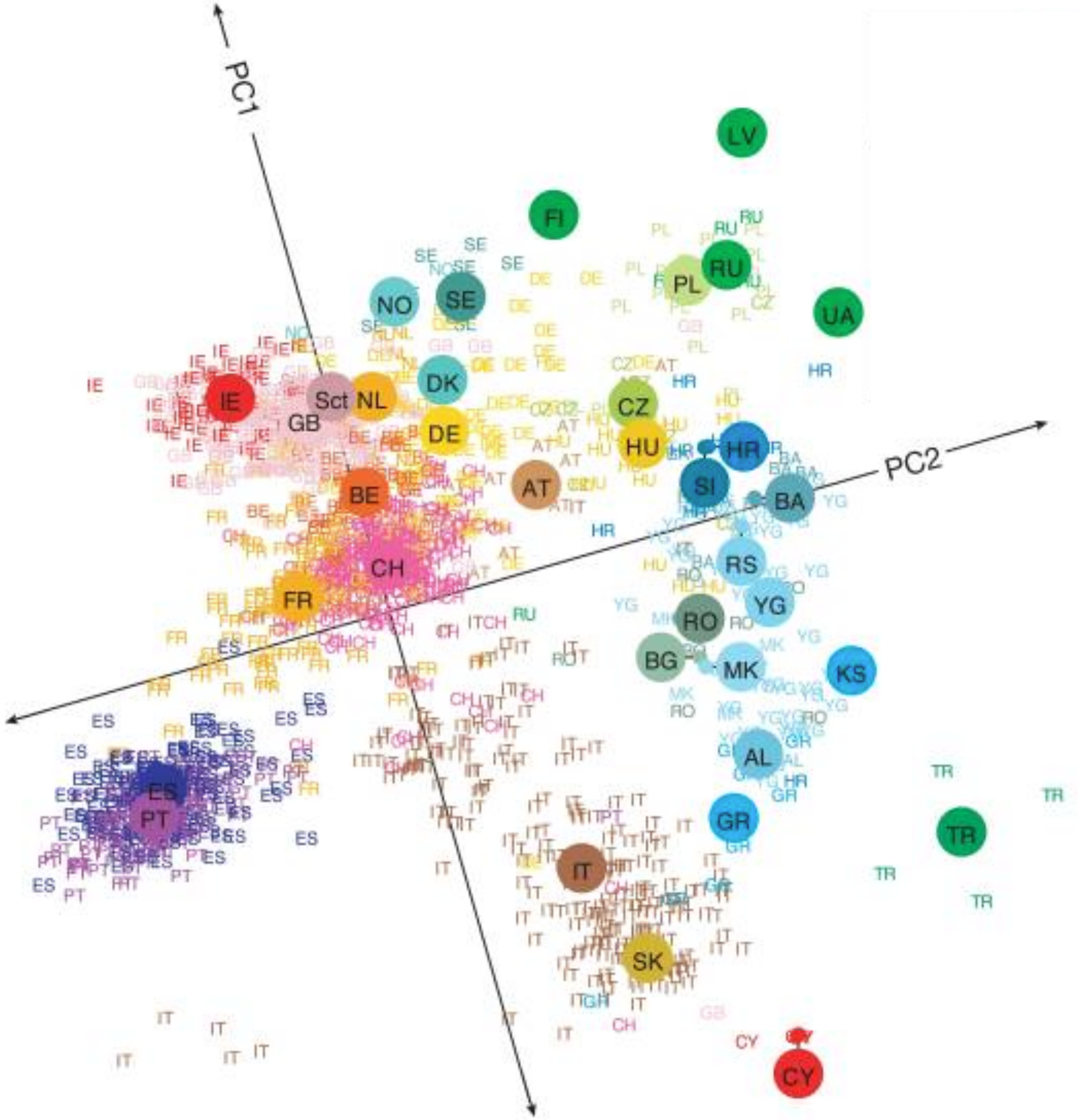


Kim *et al.* Molecular Medicine Reports 2013

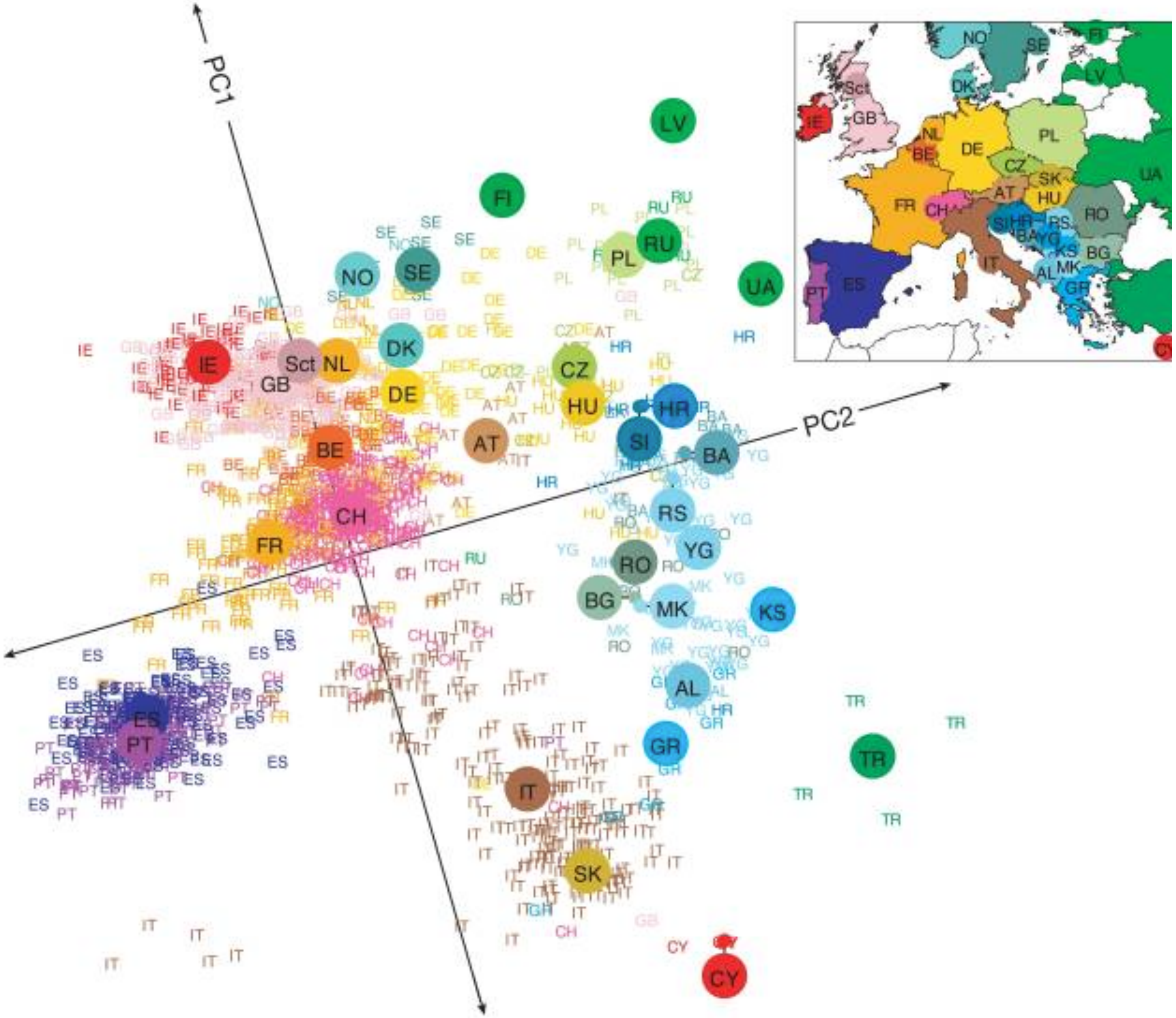
LD leads to population structure - alleles found together in populations

Relatedness of people caused by non-random mating is called population structure (or stratification)

LD leads to population structure - alleles found together in populations



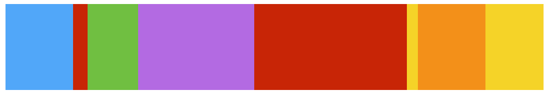
LD leads to population structure - alleles found together in populations



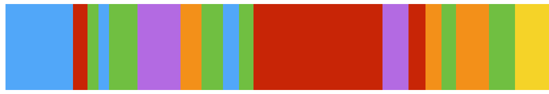
Correlation between marker and disease-causing allele drastically affects how well mappings will work

Big haplotype blocks (long-range LD) = coarse mapping

Small haplotype blocks (little LD) = fine mapping



vs.



How many people need to be genotyped?