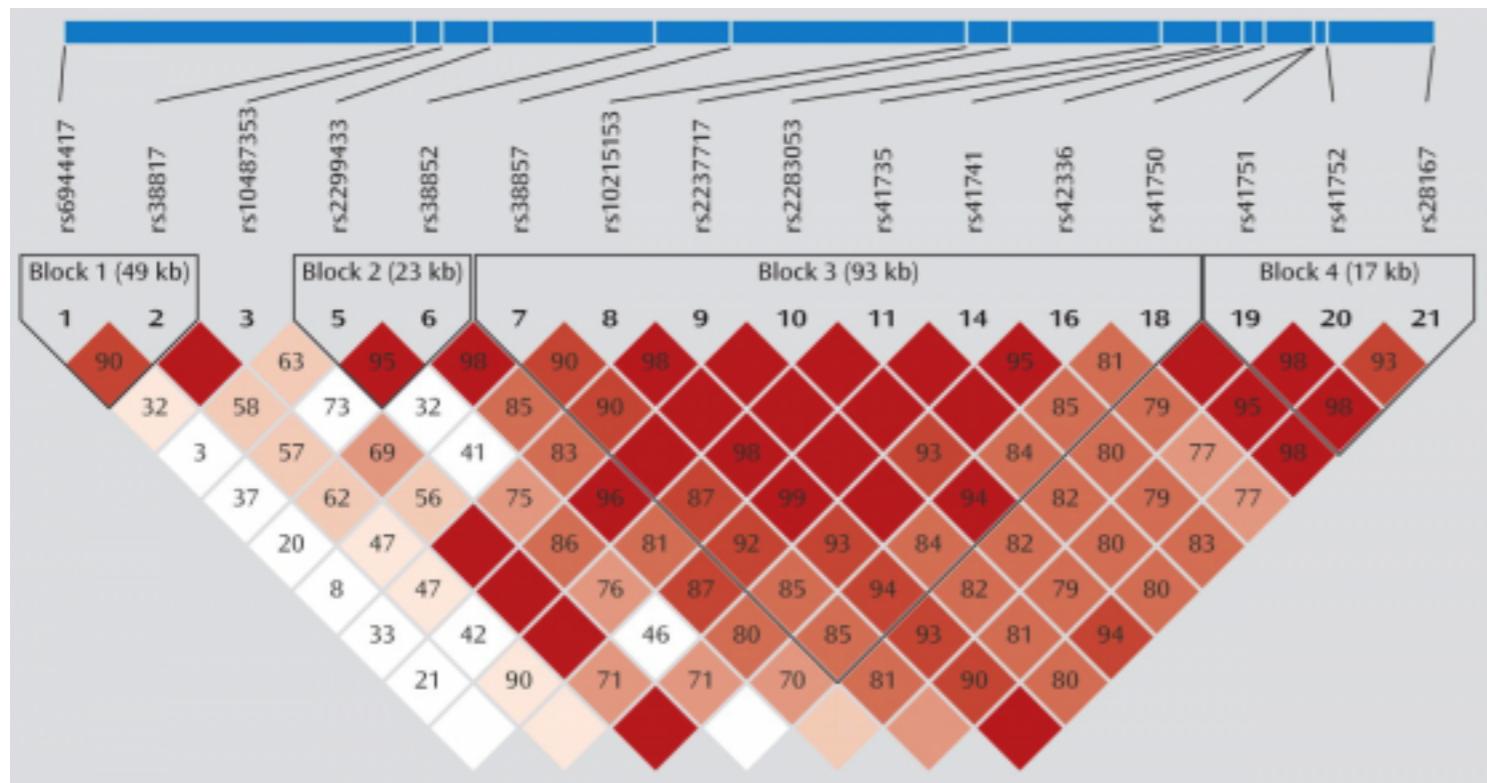
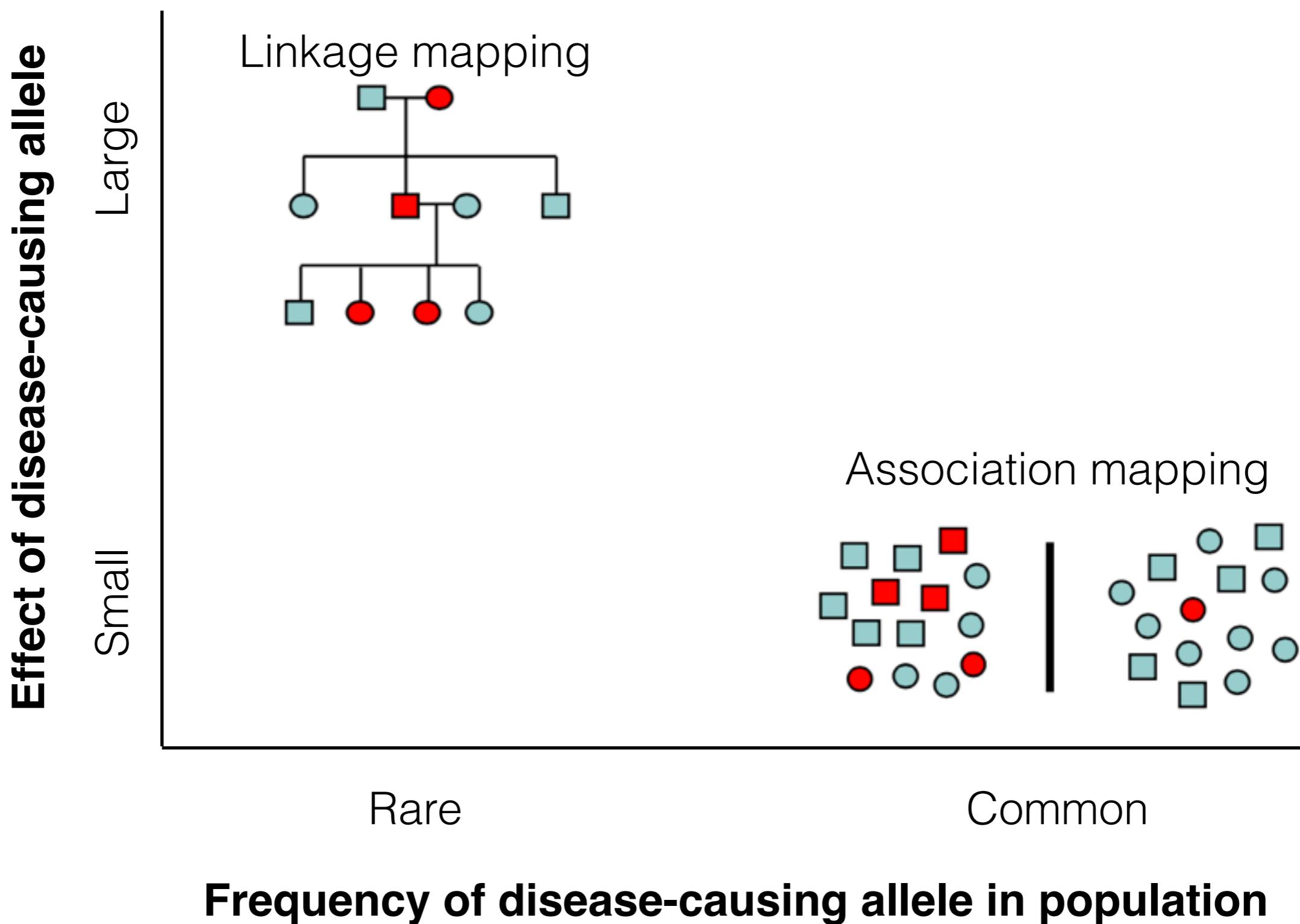


Bio393: Genetic Analysis

Linkage disequilibrium, haplotypes, and GWAS



Human gene mapping has two general flavors

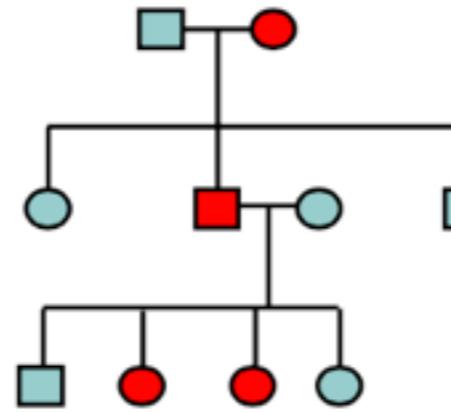


Human gene mapping has two general flavors

Effect of disease-causing allele

Large

Linkage mapping



Small

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

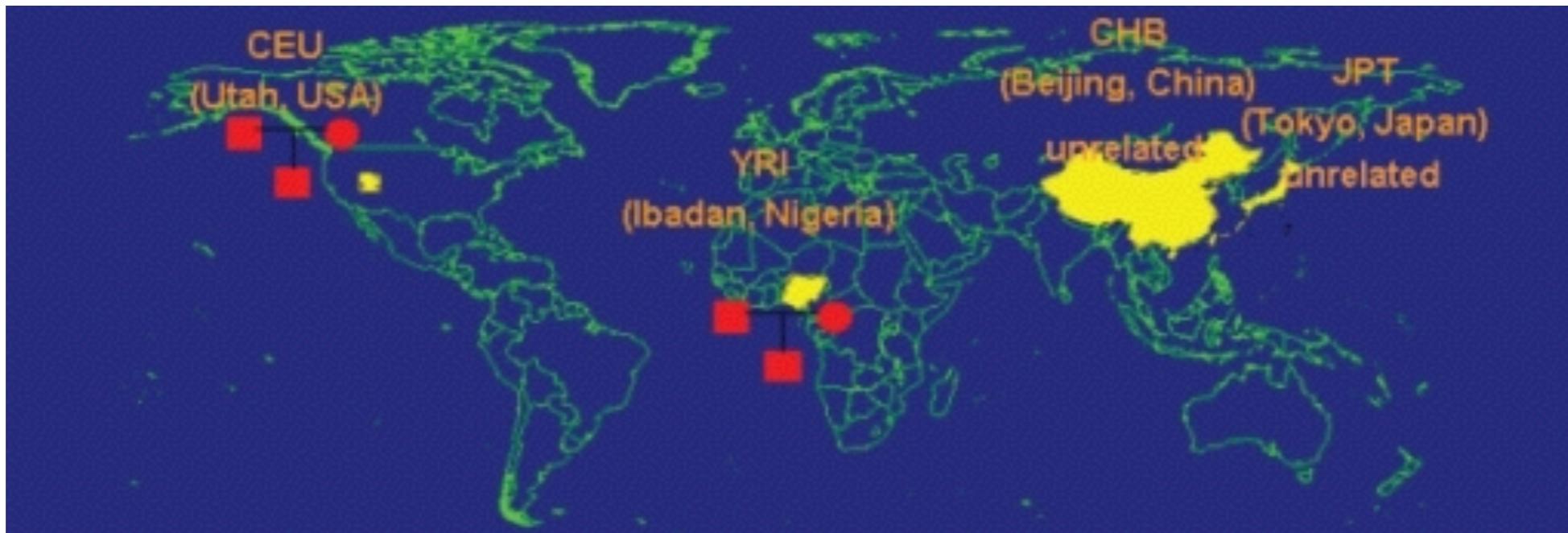
Association mapping

Rare

Common

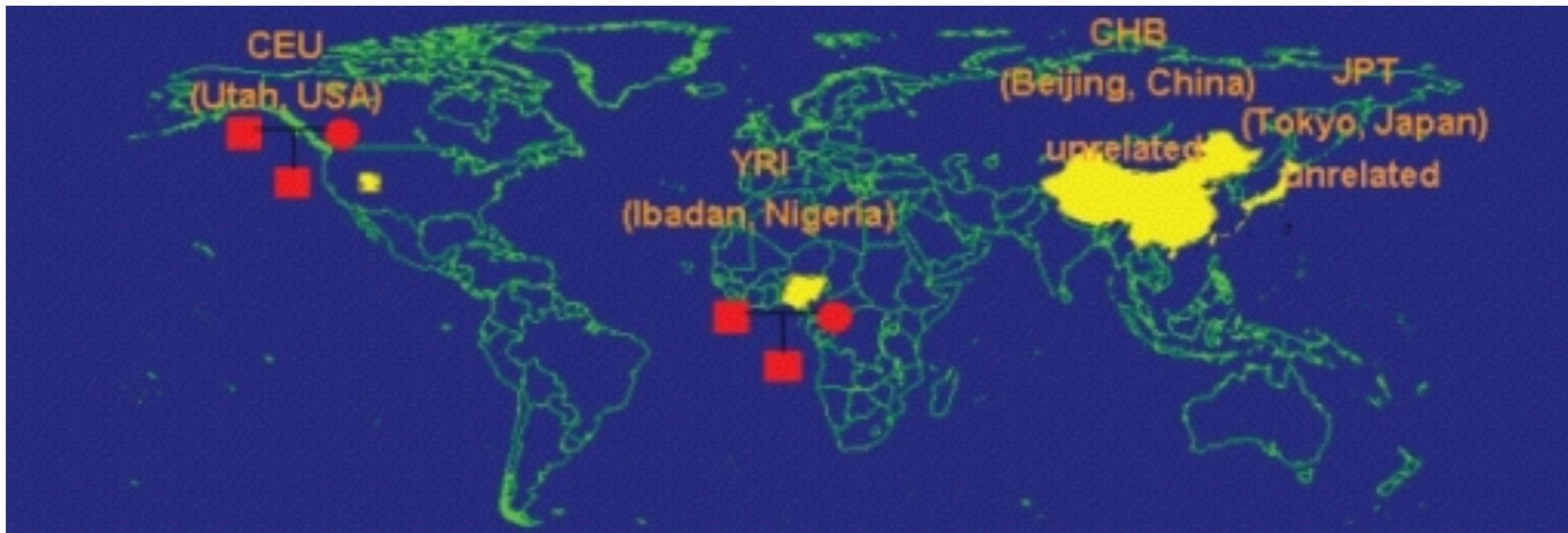
Frequency of disease-causing allele in population

Common polymorphisms facilitate genome-wide association (GWA) mapping



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

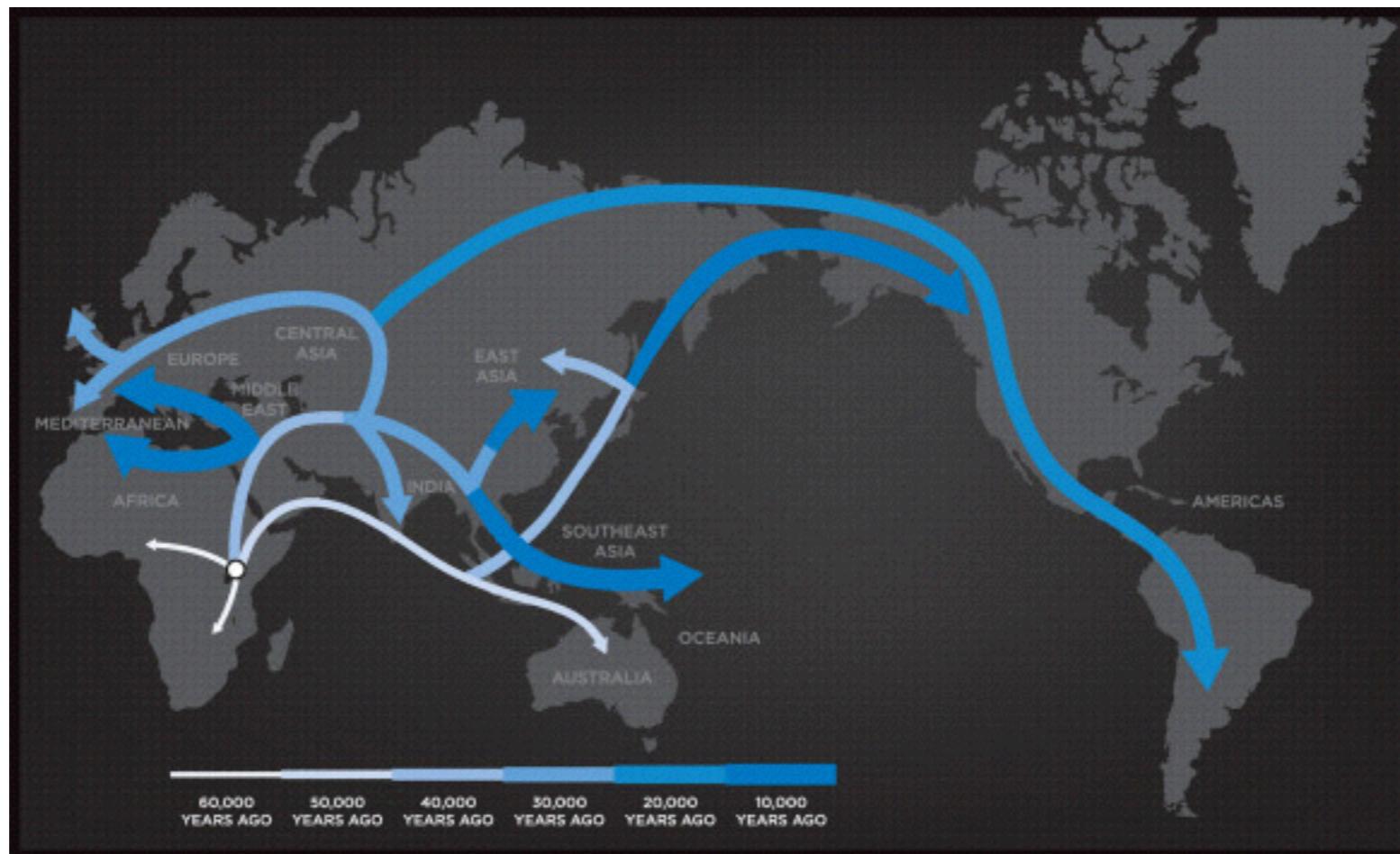
Common polymorphisms facilitate genome-wide association (GWA) mapping



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

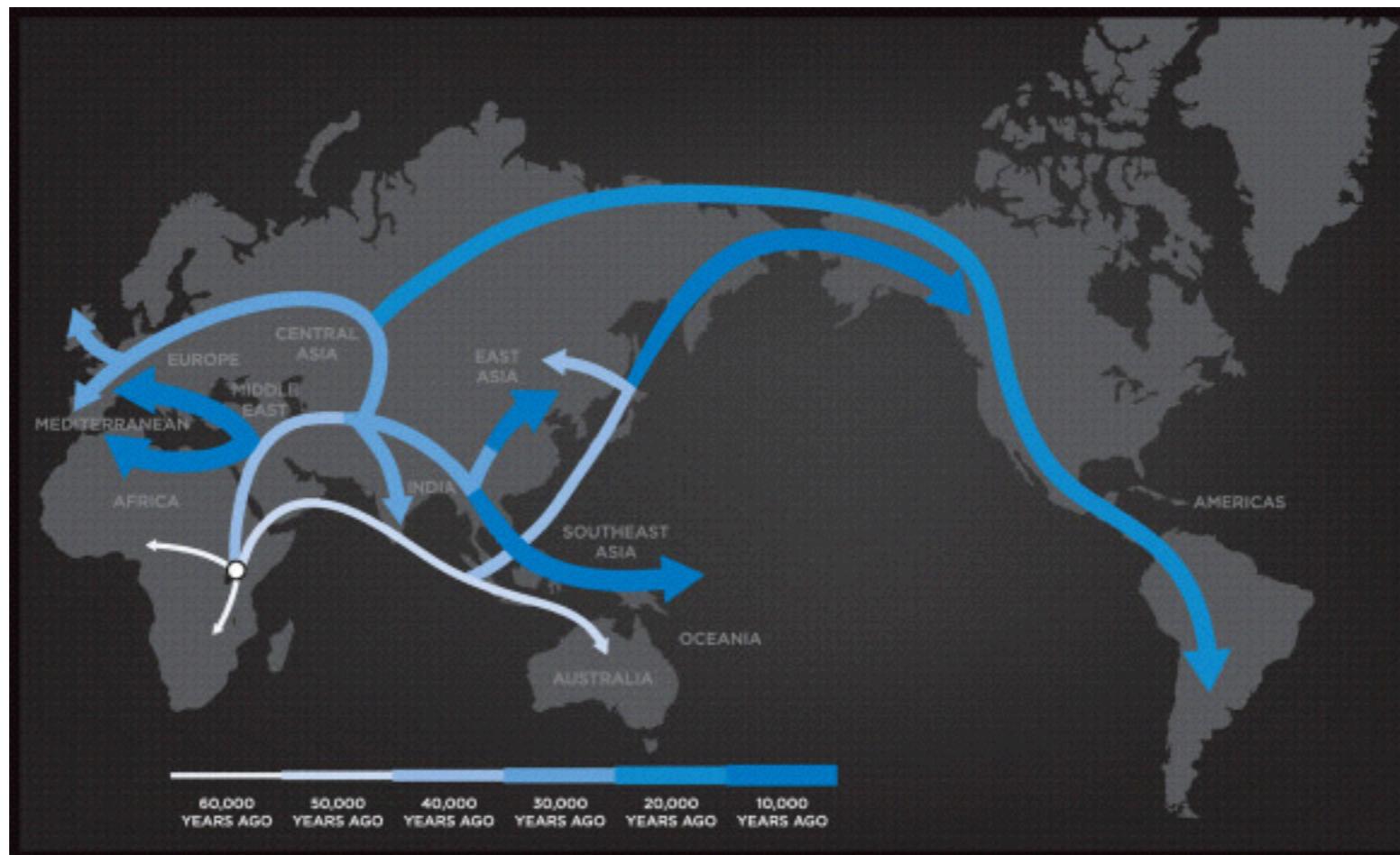
Do we have to test them all?

Common polymorphisms facilitate genome-wide association (GWA) mapping



Our relatedness means that variants
are correlated in populations

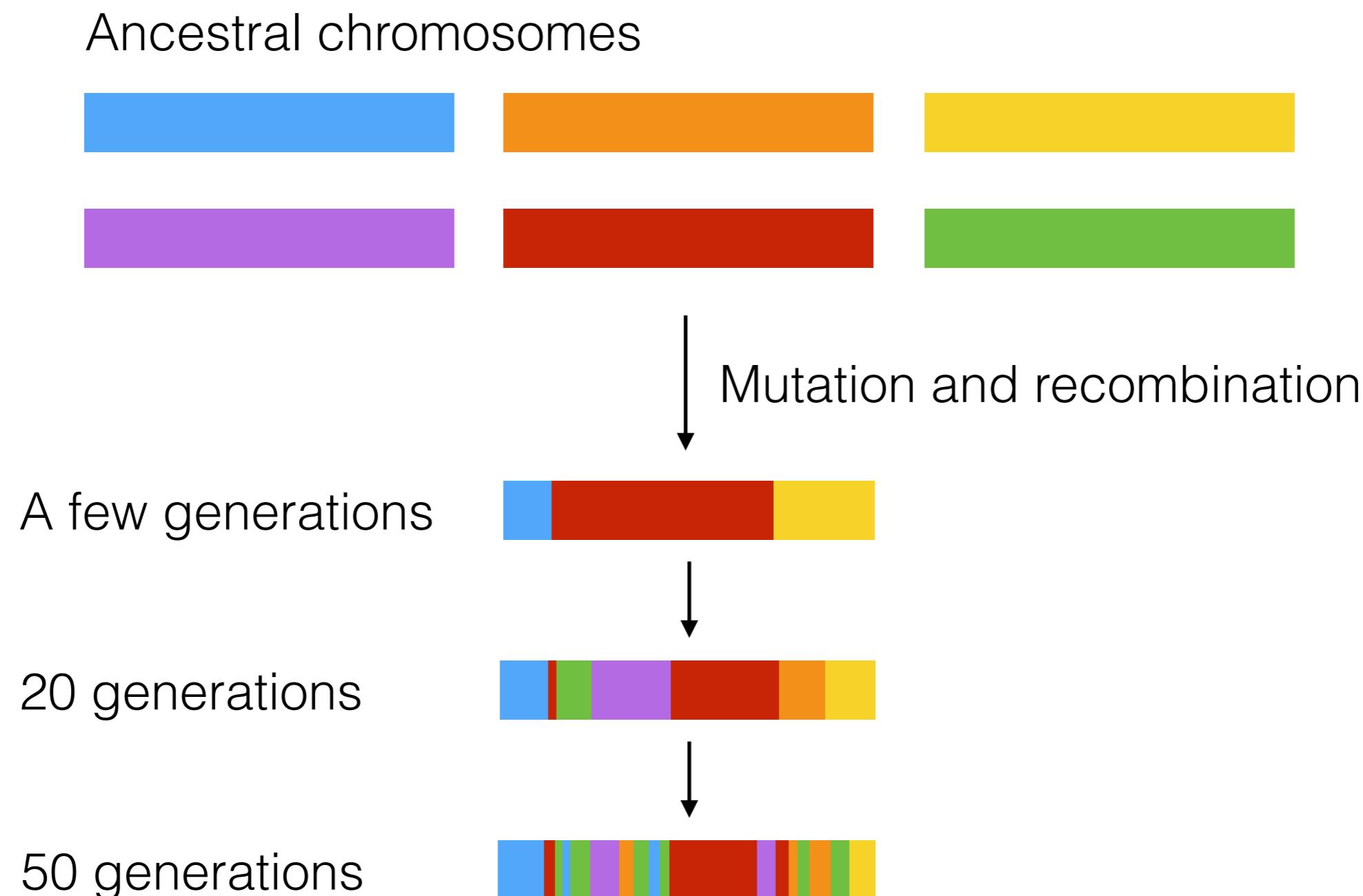
Common polymorphisms facilitate genome-wide association (GWA) mapping



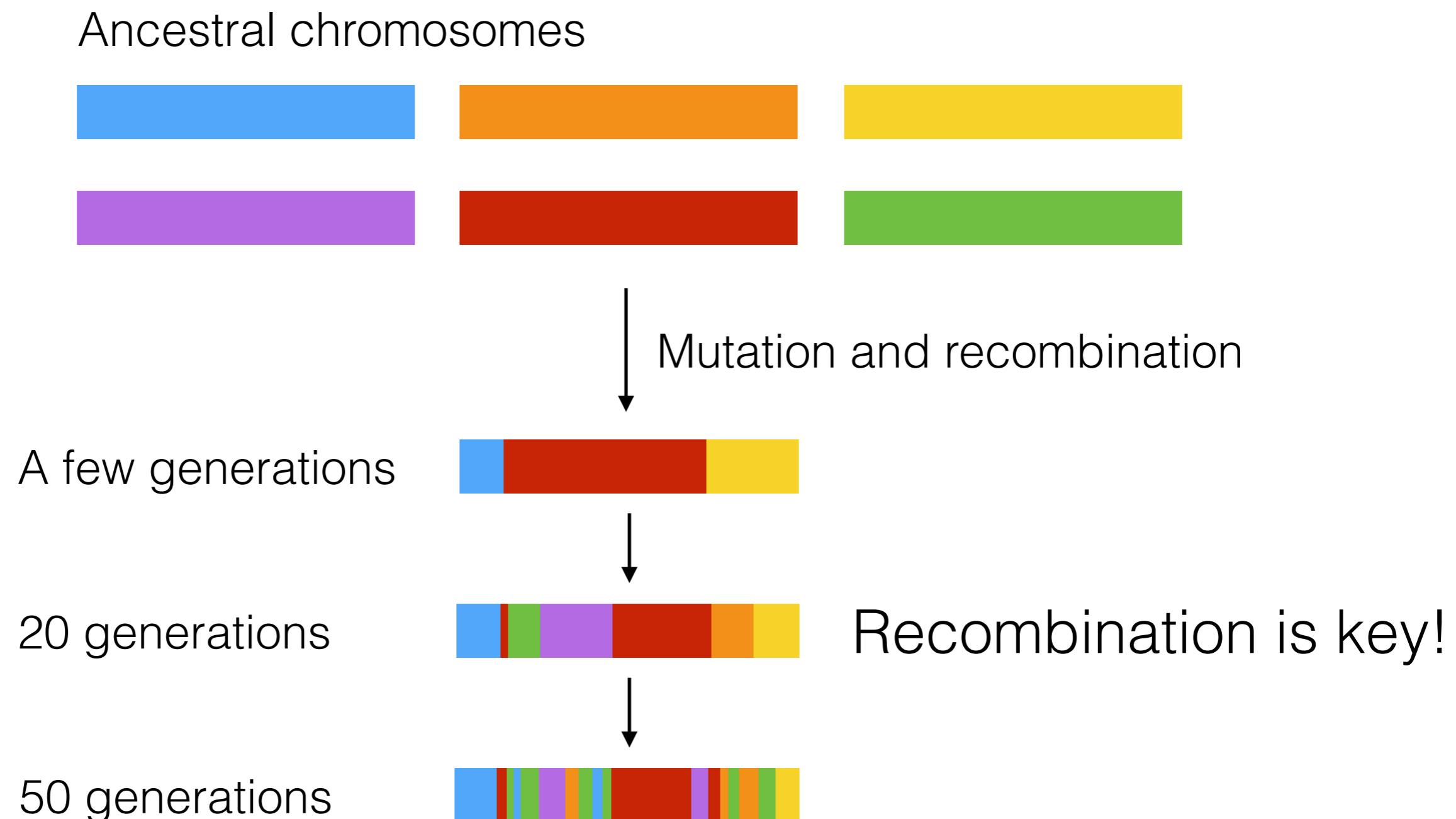
Our relatedness means that variants are correlated in populations

Correlation between variants is called linkage disequilibrium (LD)

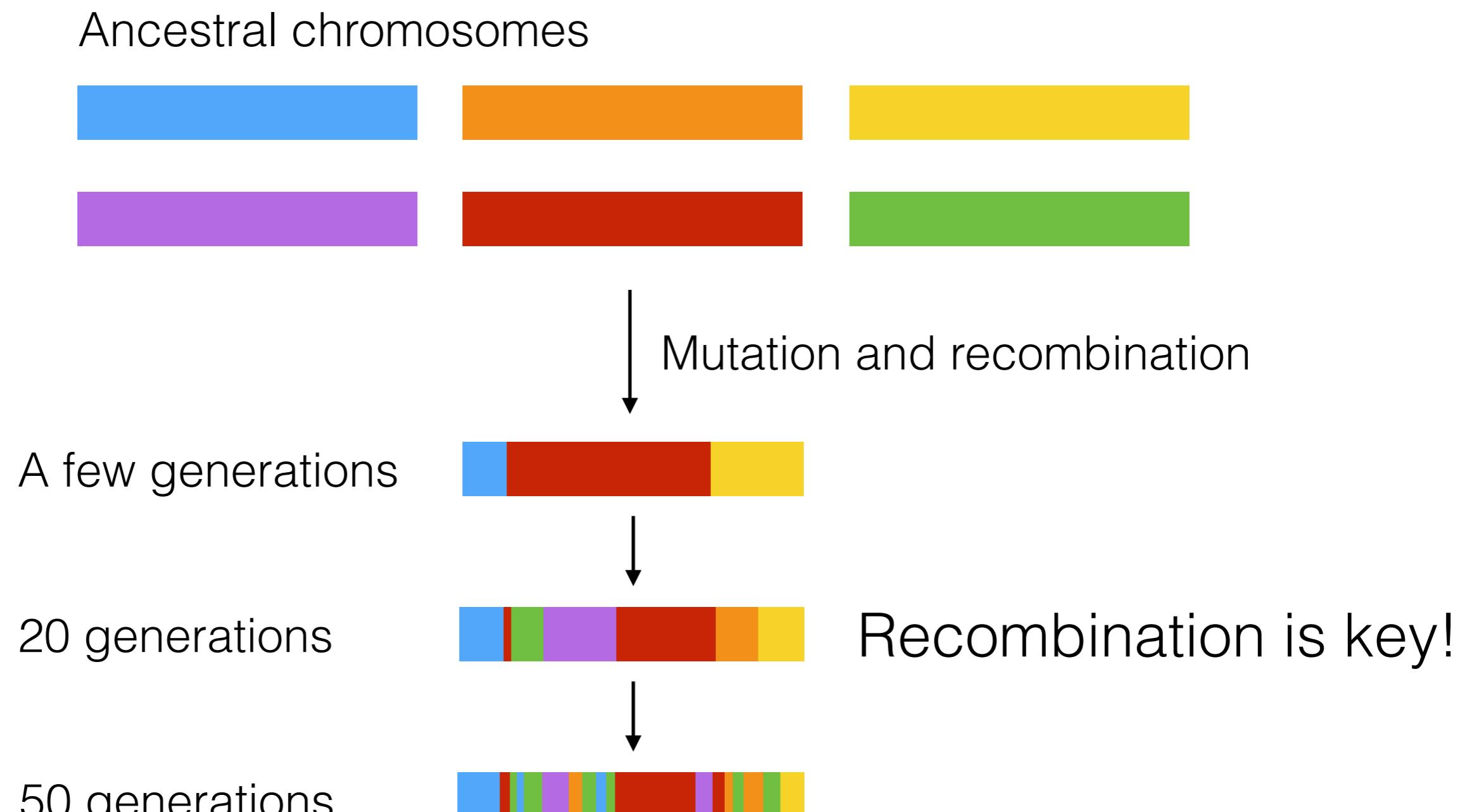
Linkage disequilibrium (LD) is the non-random association of alleles at different loci



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

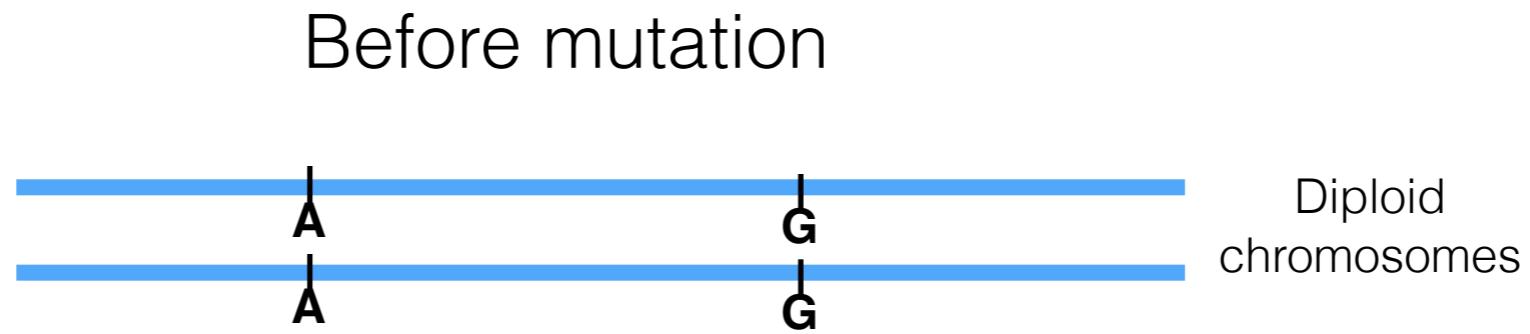


Linkage disequilibrium (LD) is the non-random association of alleles at different loci



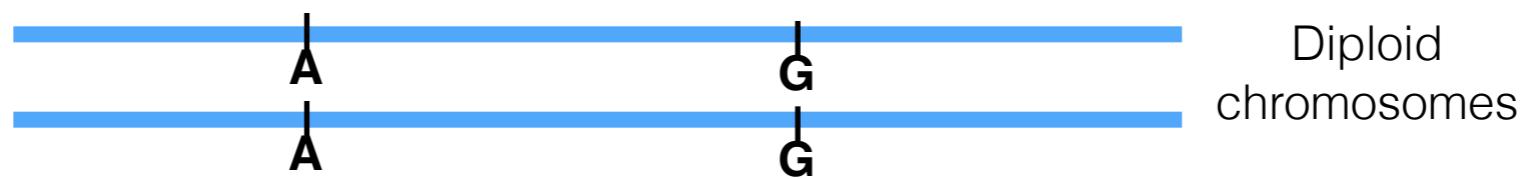
LD makes genotyping easier and cheaper

Many alleles that exist today are from ancient mutation events

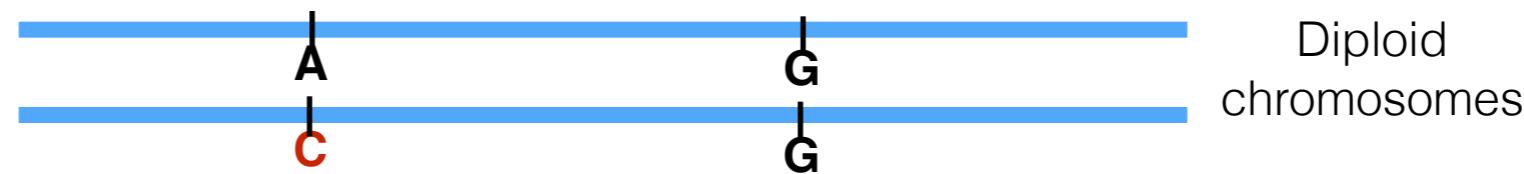


Many alleles that exist today are from ancient mutation events

Before mutation



After mutation



**That allele spreads throughout the population,
then another mutation occurs**

Before mutation



**That allele spreads throughout the population,
then another mutation occurs**

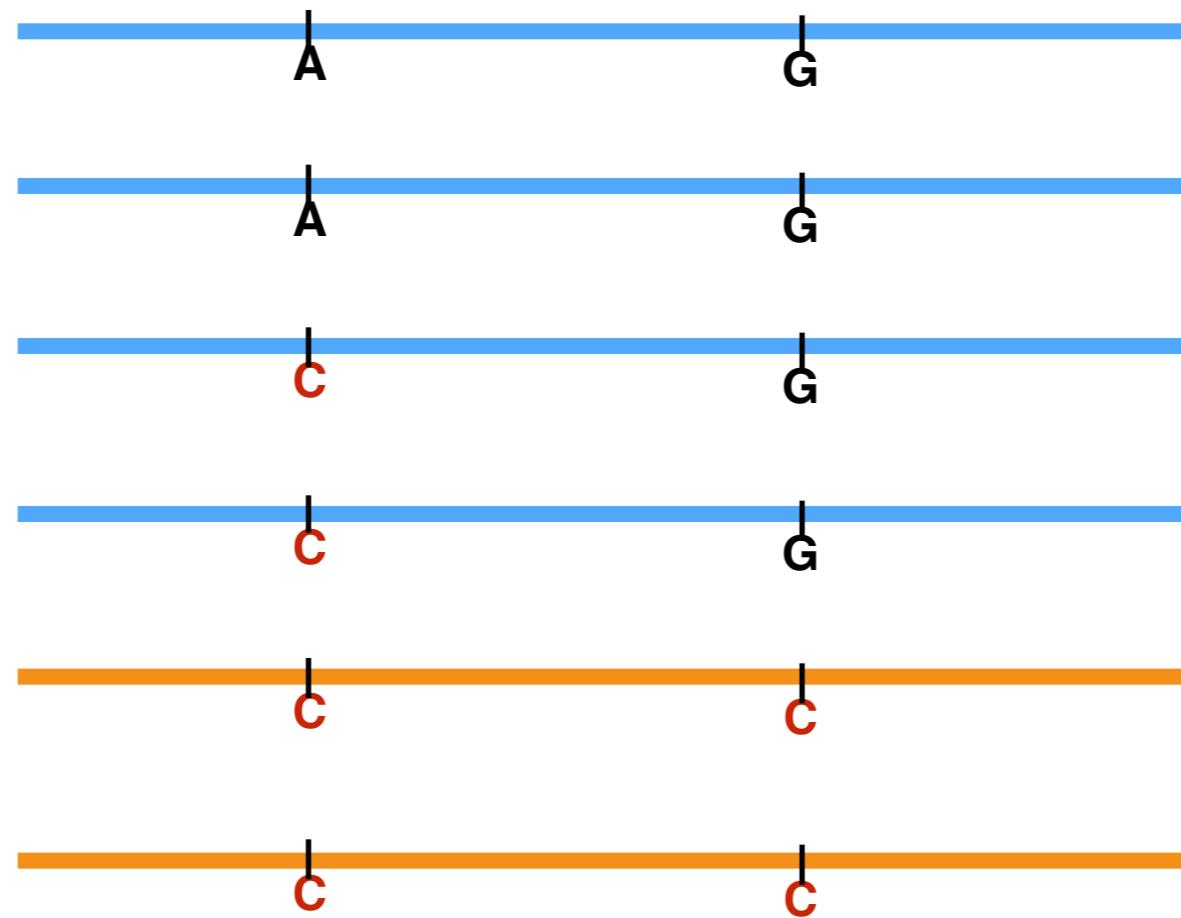
Before mutation



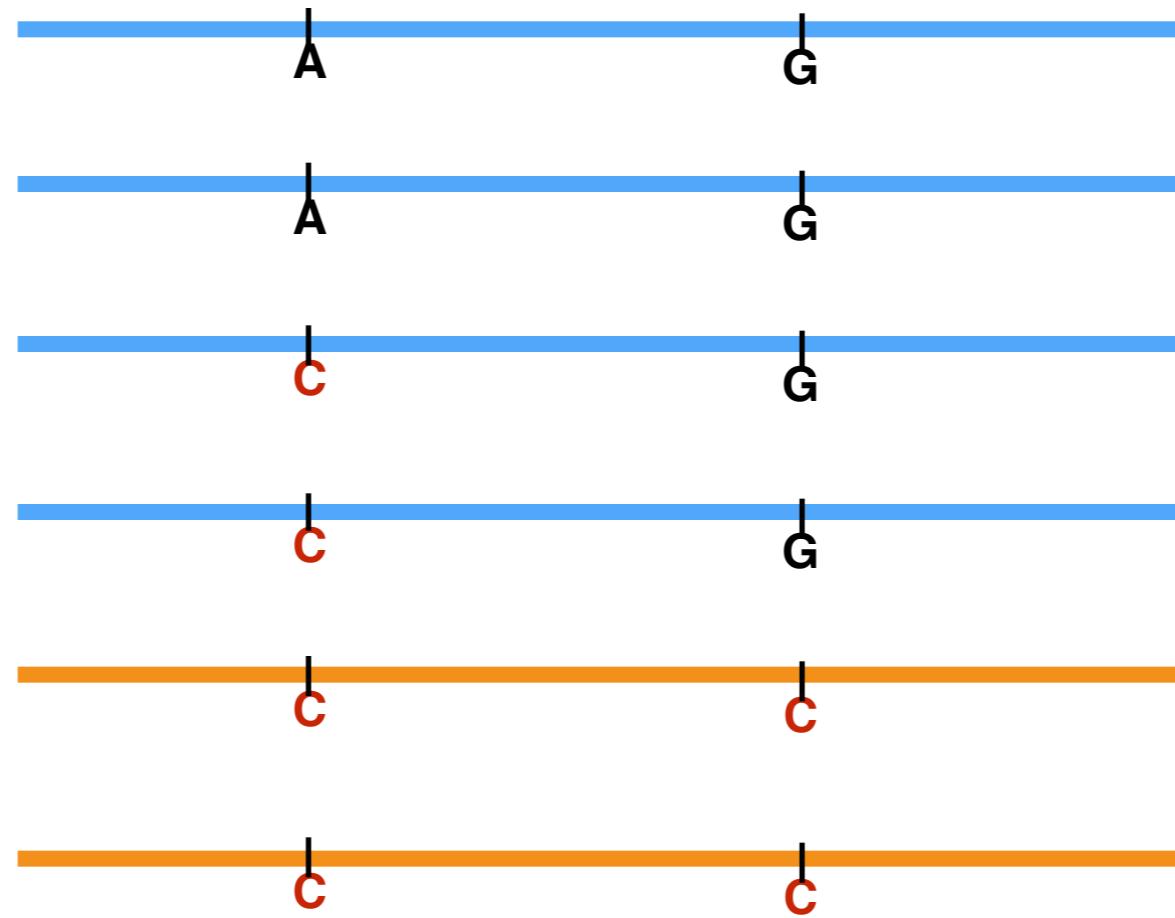
After mutation



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



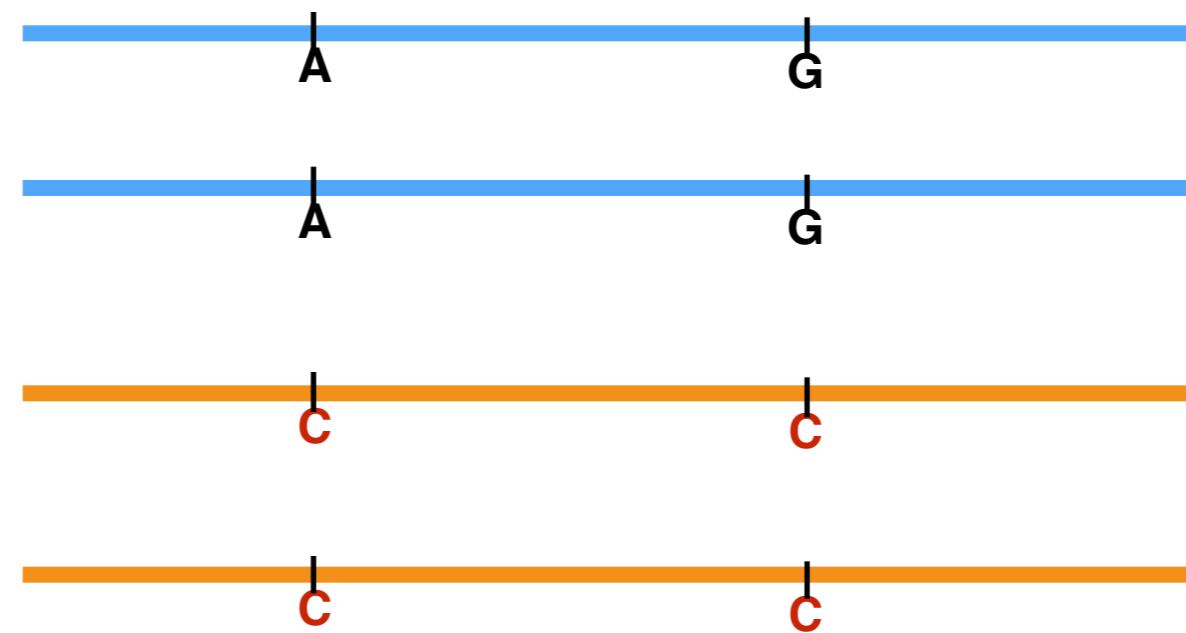
Let's think about these chromosomes with different arrangements of alleles as haploid gametes



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

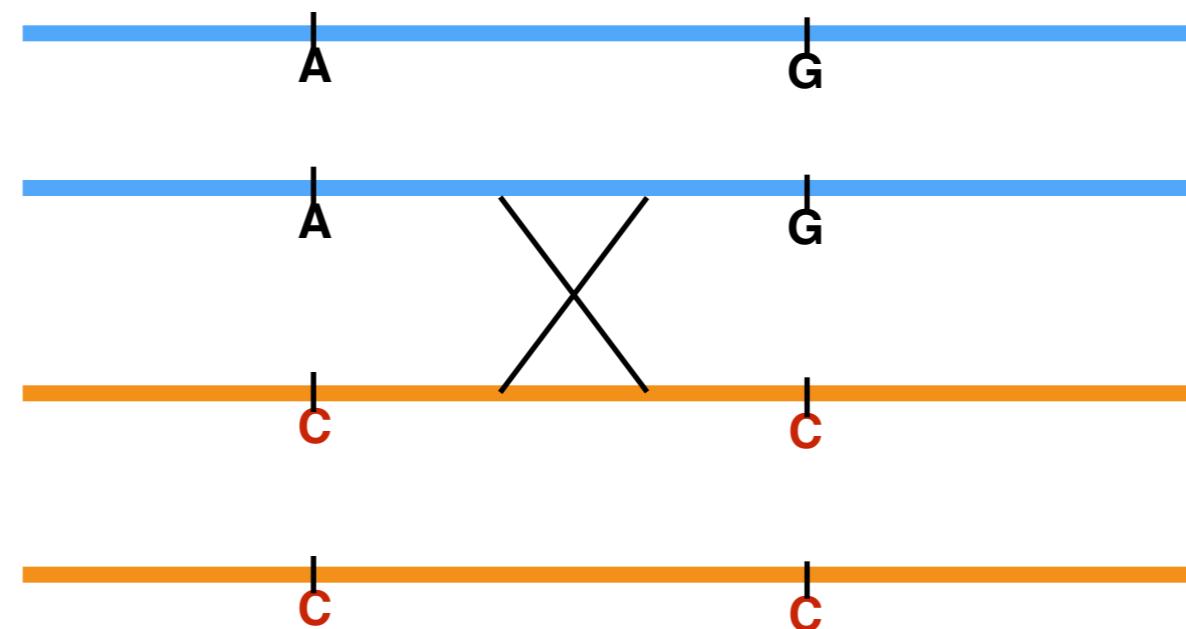
Recombination creates new arrangements of ancestral alleles

Before recombination



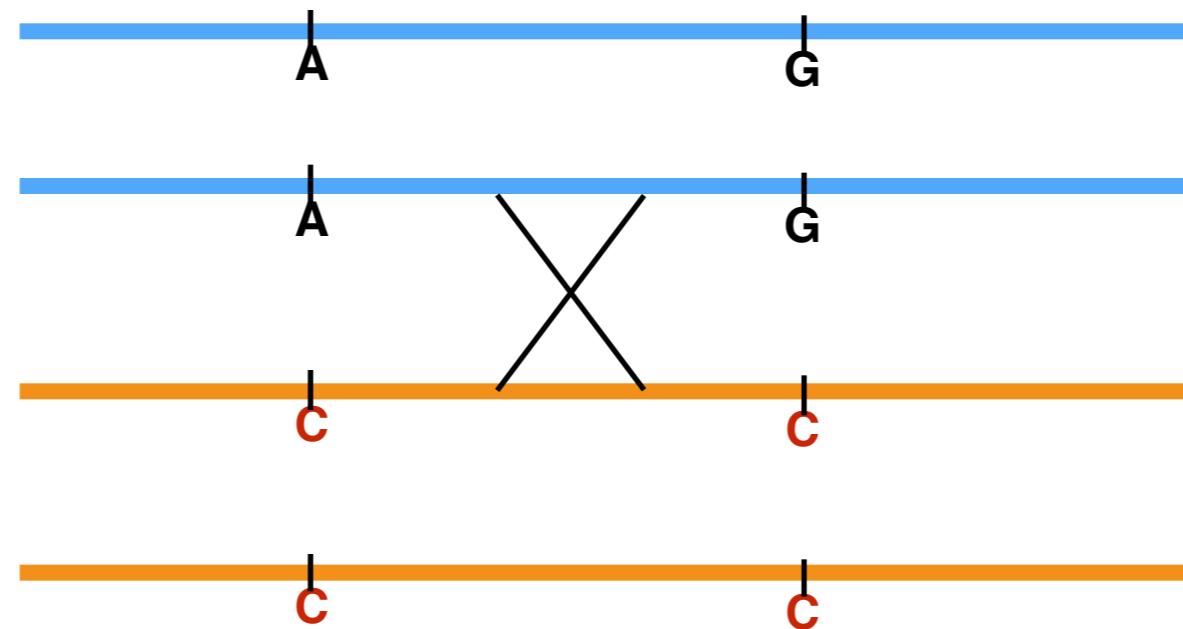
Recombination creates new arrangements of ancestral alleles

Before recombination



Recombination creates new arrangements of ancestral alleles

Before recombination



After recombination

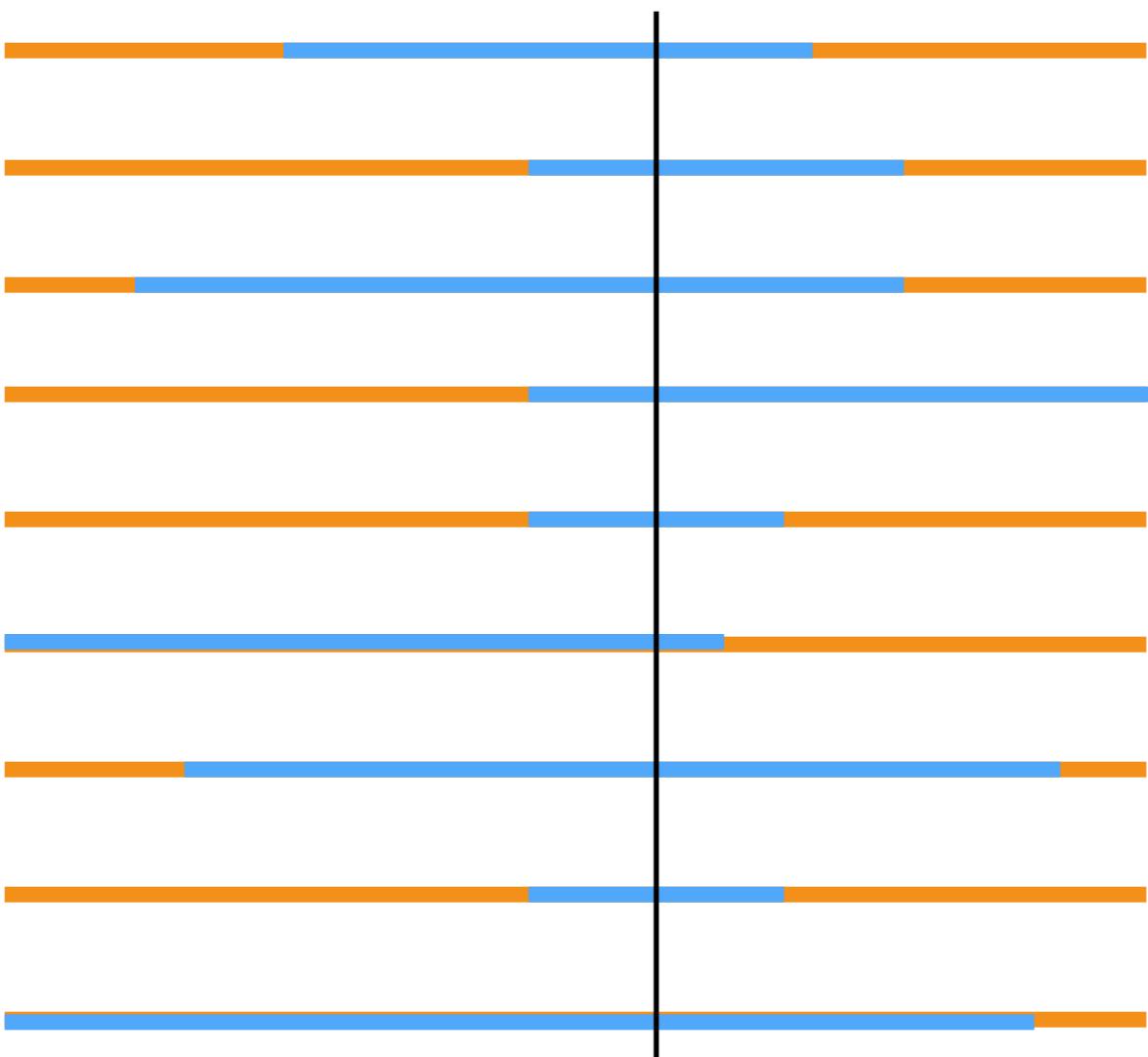


Linkage disequilibrium is the non-random association of alleles at different loci

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 (A and B) 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

$$p_b = 1 - p_B$$

Haplotype frequencies in a population

Let's say we have two linked loci (A and B) 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

Haplotype frequencies in a population

Let's say we have two linked loci (A and B) 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 (A and B) 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

Haplotype frequencies in a population

Let's say we have two linked loci (A and B) 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

$$p_A * p_B$$

Haplotype frequencies in a population

Let's say we have two linked loci (A and B) 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 $p_A * p_B \neq p_{AB}$, 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

		<u>Locus B</u>		$p_{AB} = p_A * p_B$
		B	b	
<u>Locus A</u>	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_{AB}

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

How to calculate LD?

The Disequilibrium coefficient D_{AB}

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

When in equilibrium, $D_{AB} = 0$

Otherwise, $D_{AB} >$ or < 0

How to calculate LD?

The Disequilibrium coefficient D_{AB}

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

When in equilibrium, $D_{AB} = 0$

Otherwise, $D_{AB} >$ or < 0

The sign is arbitrary. Set A, B 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:

$$r^2 = (D_{AB})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{AB} = p_{AB} - p_A * p_B$

How to calculate LD?

The correlation is the preferred term:

$$r^2 = (D_{AB})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{AB} = p_{AB} - p_A * p_B$

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

How to calculate LD? An example

$$r^2 = (D_{AB})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{AB} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers A and B with two forms A1 or A2 and B1 or B2?

We genotype 1000 people to get:

Haplotype	Number
A1B1	600
A1B2	100
A2B1	200
A2B2	100

How to calculate LD? An example

$$r^2 = (D_{AB})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{AB} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers A and B with two forms A1 or A2 and B1 or B2?

We genotype 1000 people to get:

Haplotype	Number
A1B1	600
A1B2	100
A2B1	200
A2B2	100

Convert to numbers of individuals into haplotype frequencies:

Haplotype	Number	Frequency
A1B1	600	0.6
A1B2	100	0.1
A2B1	200	0.2
A2B2	100	0.1

How to calculate LD? An example

$$r^2 = (D_{AB})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{AB} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers A and B with two forms A1 or A2 and B1 or B2?

Frequencies of haplotypes:

Haplotype	Number	Frequency
A1B1	600	0.6
A1B2	100	0.1
A2B1	200	0.2
A2B2	100	0.1

Convert to frequencies of alleles:

$$p_{A1} = p(A1B1) + p(A1B2)$$

$$p_{A2} = 1 - p_{A1}$$

$$p_{B1} = p(A1B1) + p(A2B1)$$

$$p_{B2} = 1 - p_{B1}$$

Allele	Number	Frequency
A1	700	0.7
A2	300	0.3
B1	800	0.8
B2	200	0.2

How to calculate LD? An example

$$r^2 = (D_{AB})^2 / (p_A * (1 - p_A) * p_B * (1 - p_B))$$

Remember $D_{AB} = p_{AB} - p_A * p_B$

What is the disequilibrium between two markers A and B with two forms A1 or A2 and B1 or B2?

Convert to frequencies of alleles:

$$p_{A1} = p(A1B1) + p(A1B2)$$

$$p_{A2} = 1 - p_{A1}$$

$$p_{B1} = p(A1B1) + p(A2B1)$$

$$p_{B2} = 1 - p_{B1}$$

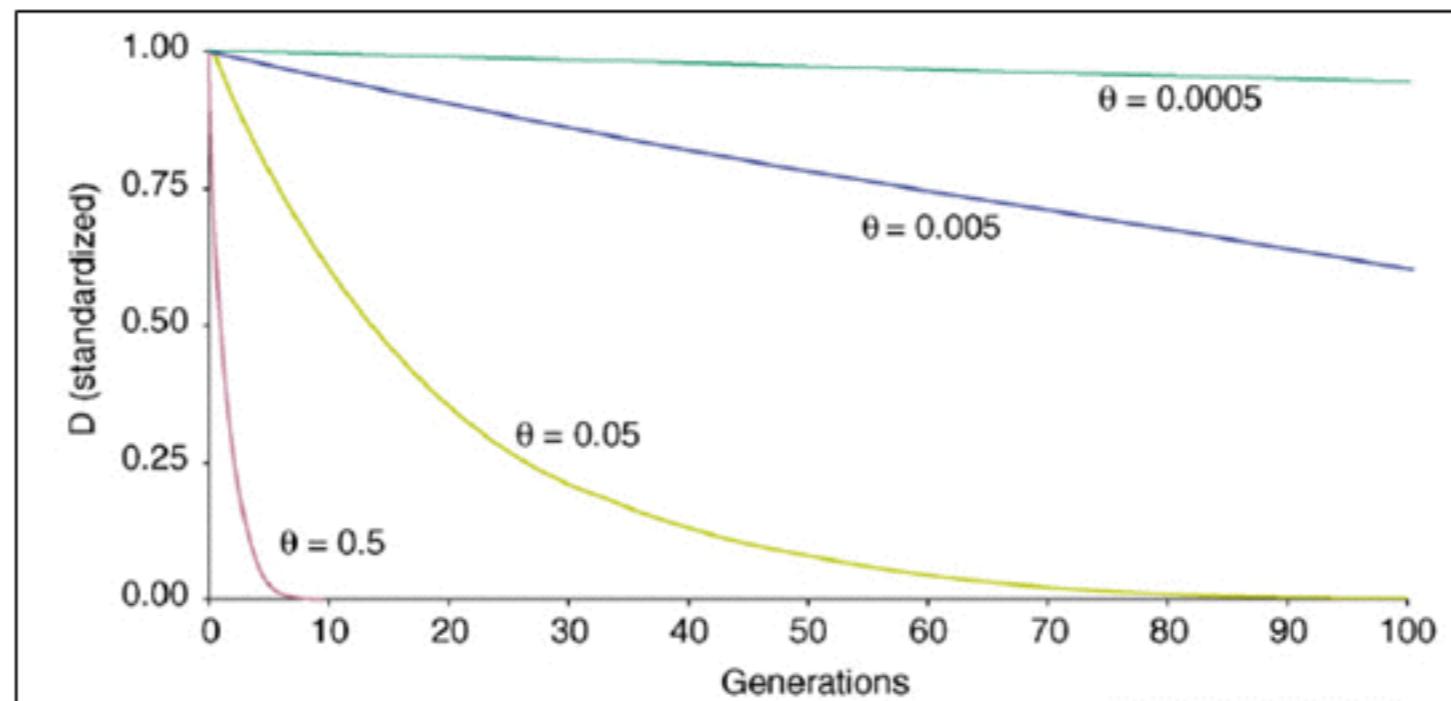
Allele	Number	Frequency
A1	700	0.7
A2	300	0.3
B1	800	0.8
B2	200	0.2

Haplotype	Number	Frequency
A1B1	600	0.6
A1B2	100	0.1
A2B1	200	0.2
A2B2	100	0.1

$$D_{AB} = 0.6 - 0.7 * 0.8 = 0.04$$

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

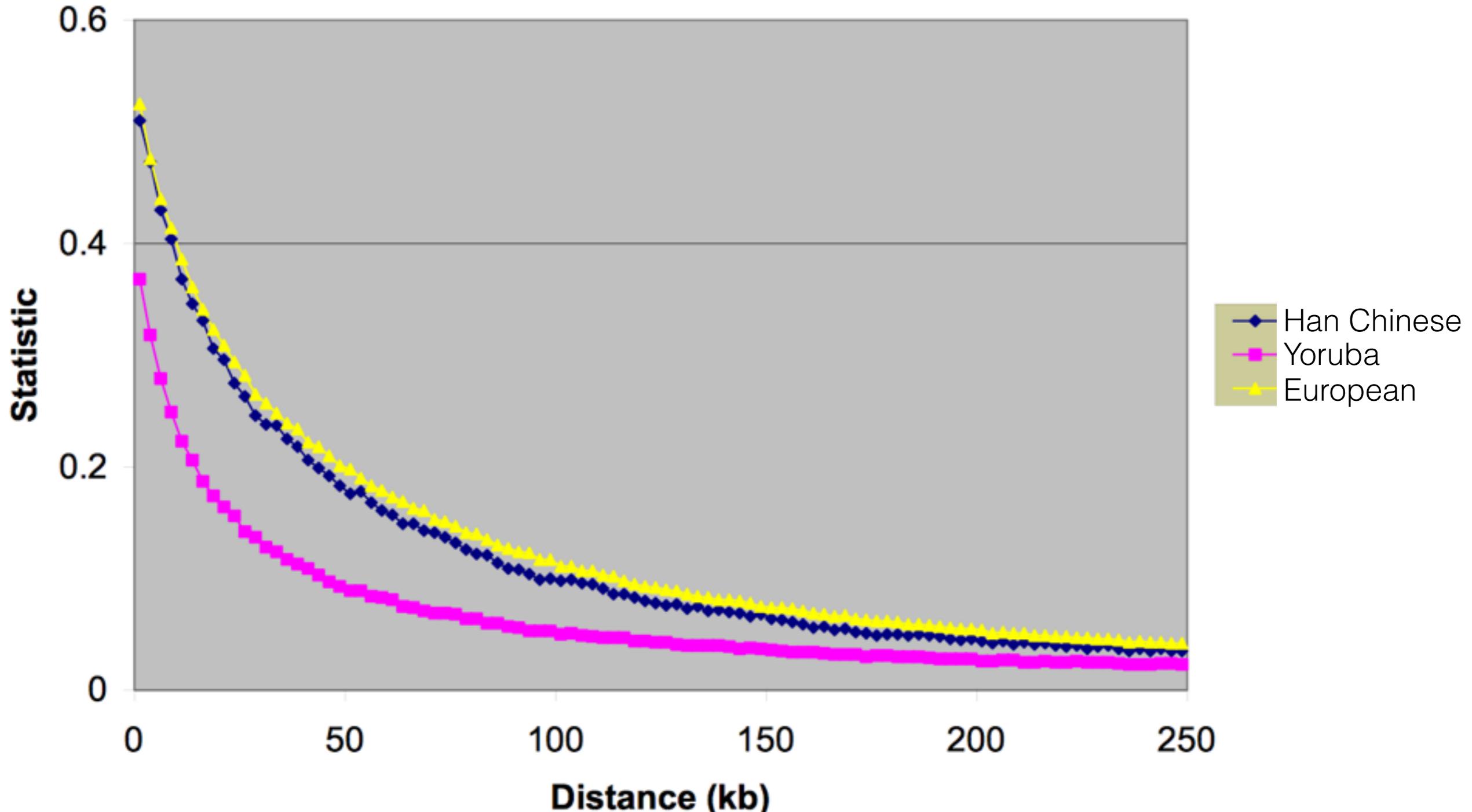
Linkage disequilibrium decreases by distance and generation time



Mackay and Powell 2007

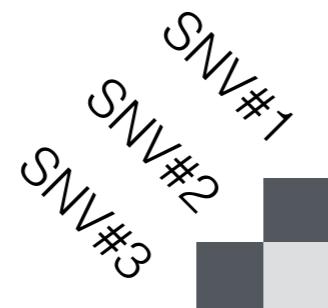
Recombination is key!

Linkage disequilibrium varies among different populations



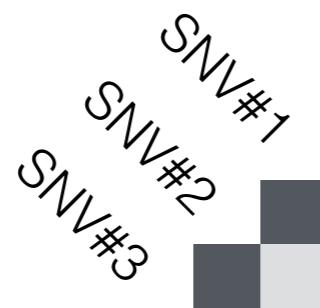
Recombination is key!

Linkage disequilibrium is often shown as a triangle correlation plot



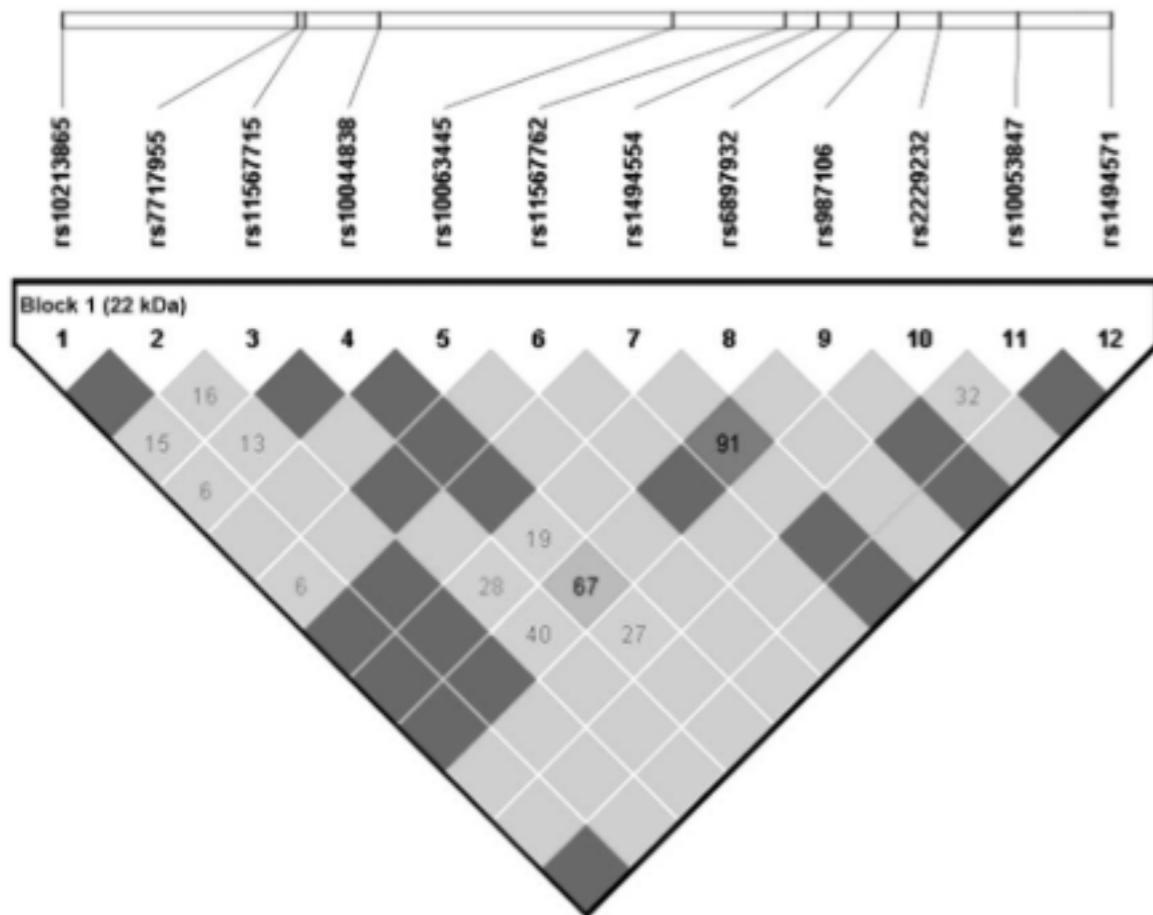
SNV#1 and SNV#2 have high LD
SNV#2 and SNV#3 have high LD
SNV#1 and SNV#3 have low LD

Linkage disequilibrium is often shown as a triangle correlation plot

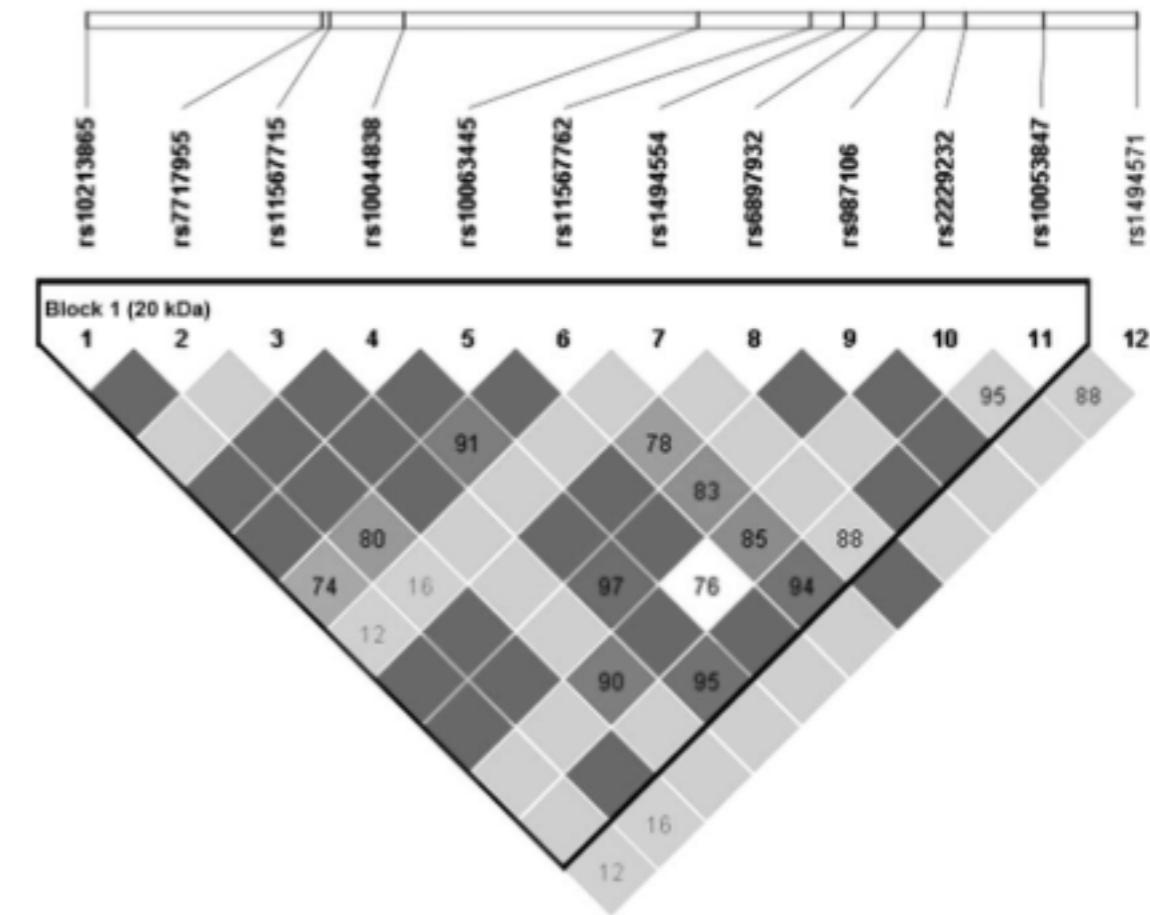


SNV#1 and SNV#2 have high LD
SNV#2 and SNV#3 have high LD
SNV#1 and SNV#3 have low LD

A LD plot of Africans



B LD plot of Asians



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

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

Small haplotype blocks (little LD) = fine mapping



vs.



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

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

Small haplotype blocks (little LD) = fine mapping

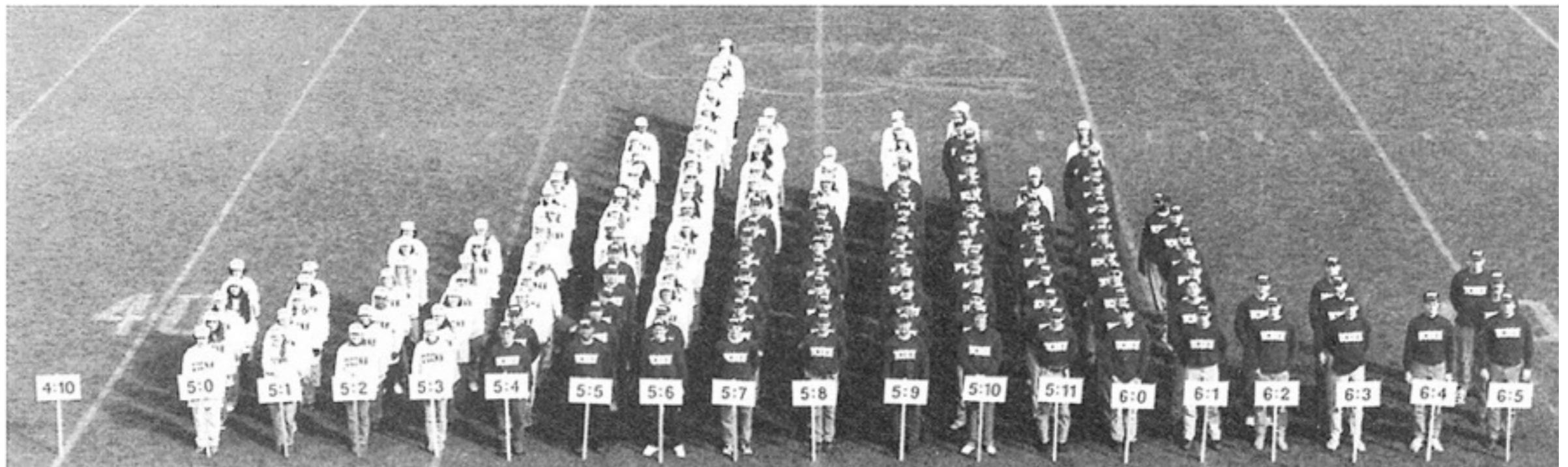


vs.



How many people need to be genotyped?

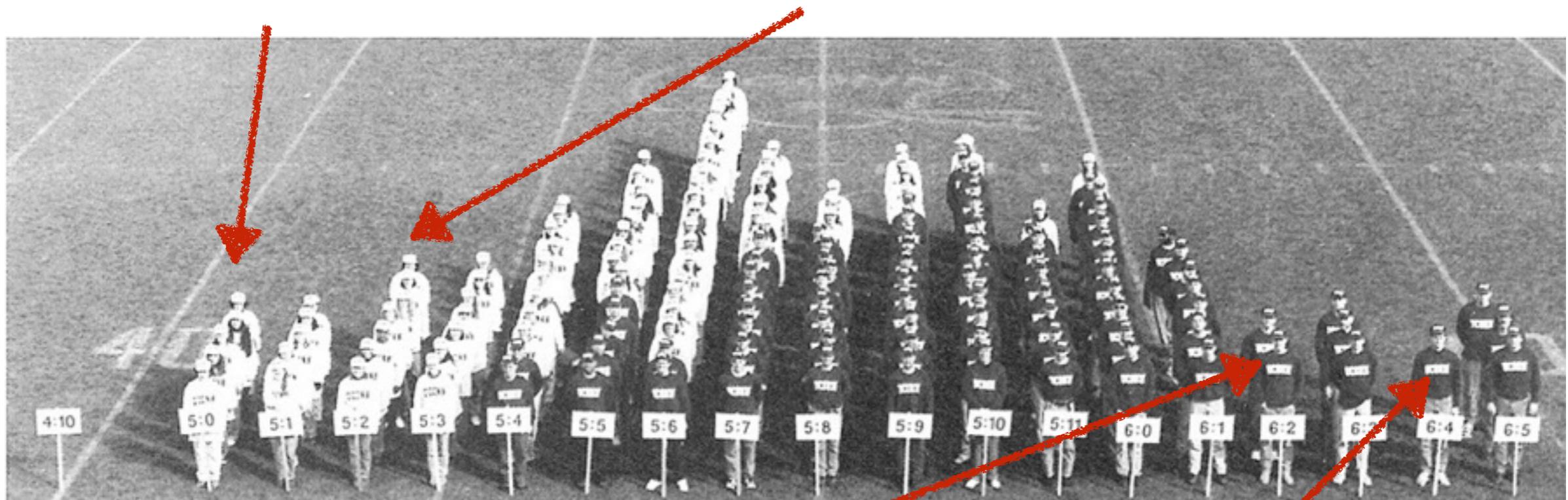
Genome-wide association studies measure correlation between “tag-SNP” and disease-causing allele



University of Connecticut, 1997

Genome-wide association studies measure correlation between “tag-SNP” and disease-causing allele

CAGCGATAGGCTTAATGTT	CAGCGATAGGCTTAATGTT
AGCCC GTTT <ins>T</ins> ATGACCAACG	AGCCC GTTT <ins>T</ins> ATGACCAACG
GGGTTCACAGTGAGCTGTGT	GGGTTCACAGTGAGCTGTGT

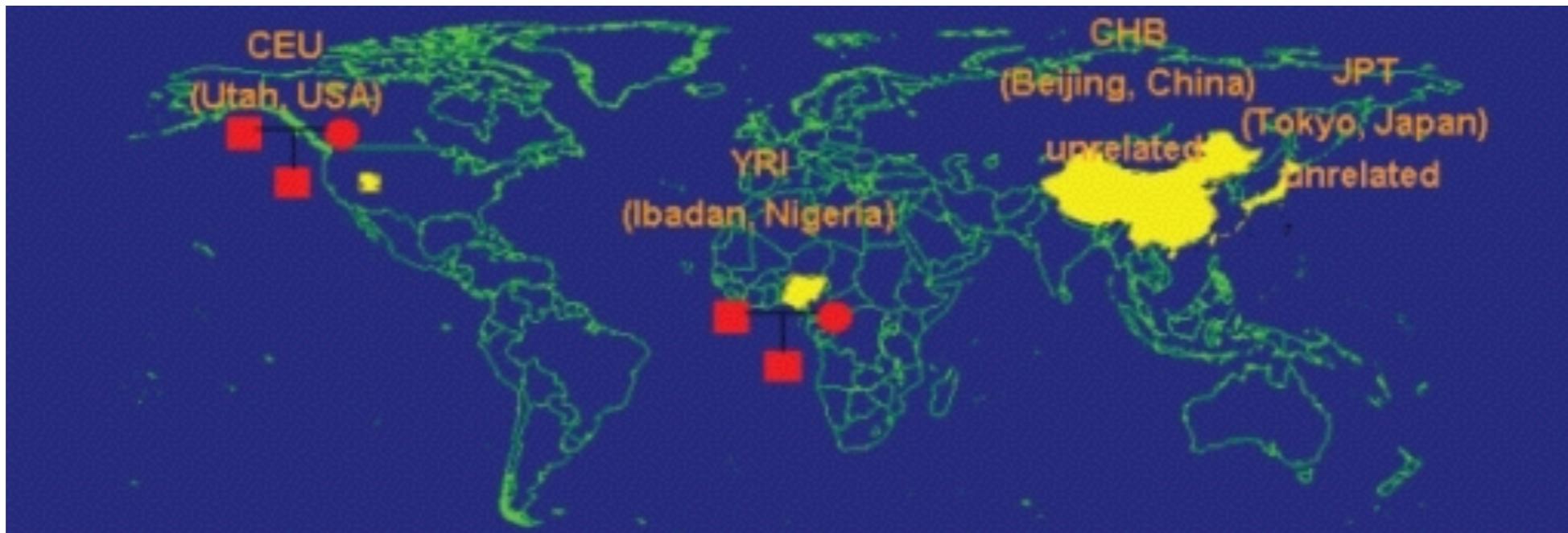


University of Connecticut, 1997

CAGCGATAGGCTTAATGTT
AGCCC GTTT <ins>G</ins> ATGACCAACG
GGGTTCACAGTGAGCTGTGT

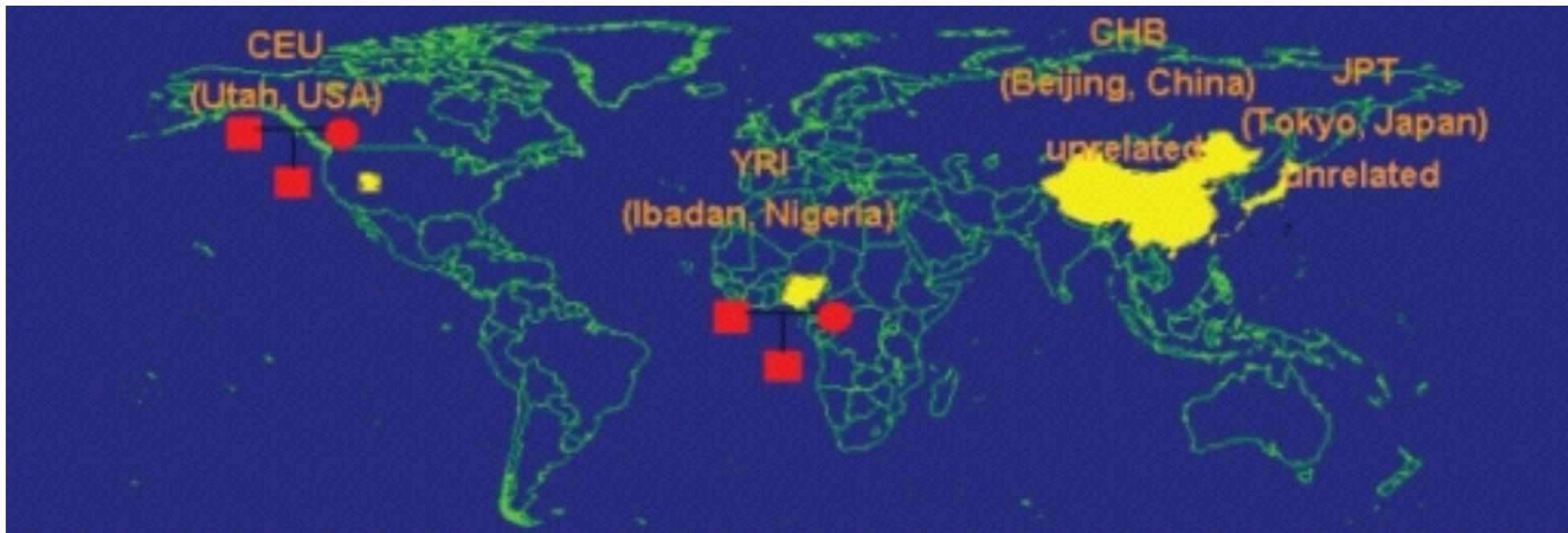
CAGCGATAGGCTTAATGTT
AGCCC GTTT <ins>G</ins> ATGACCAACG
GGGTTCACAGTGAGCTGTGT

Common polymorphisms facilitate genome-wide association (GWA) mapping



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

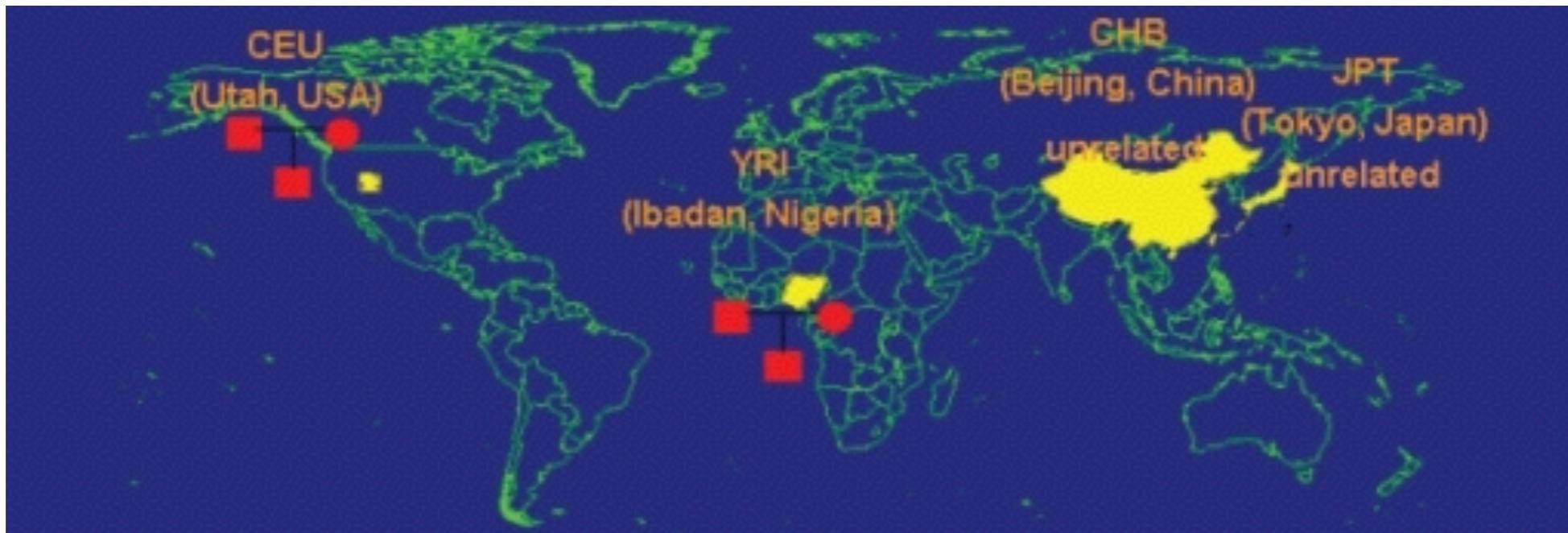
Common polymorphisms facilitate genome-wide association (GWA) mapping



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

LD blocks in humans are 20-100 kb

Common polymorphisms facilitate genome-wide association (GWA) mapping

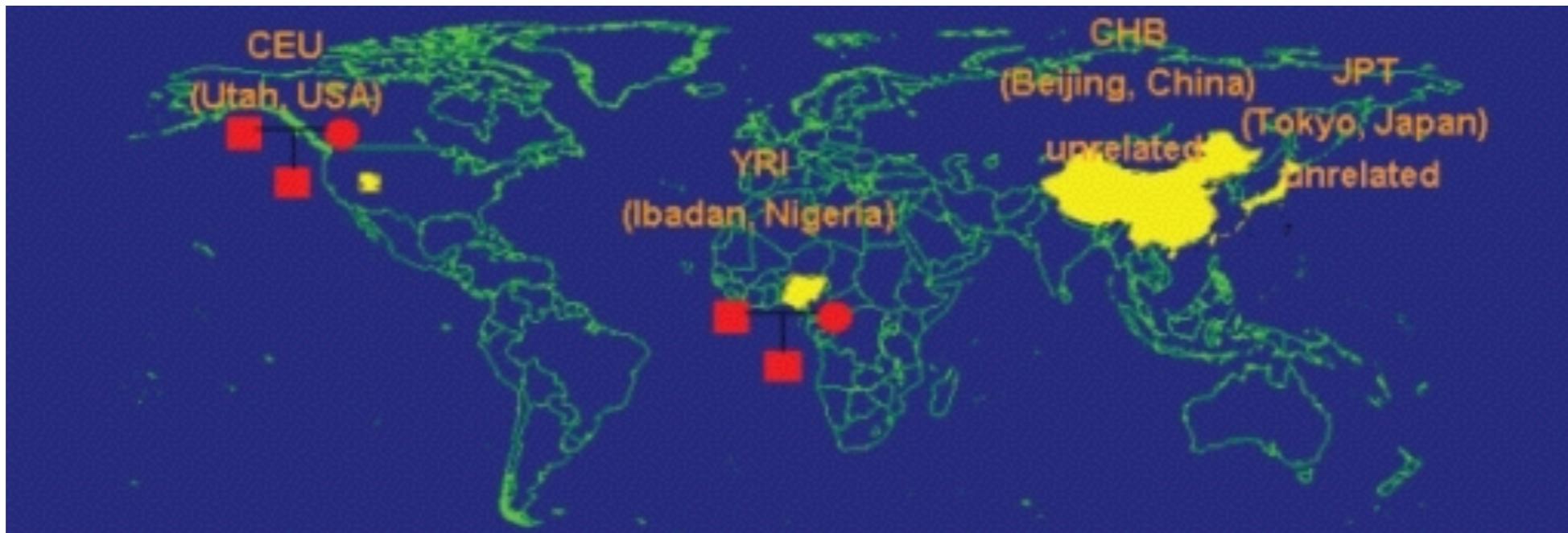


The Human Haplotype Map (HapMap) identified
10 million common polymorphisms

LD blocks in humans are 20-100 kb

500,000 common variants gives us a SNP every 10 kb

Common polymorphisms facilitate genome-wide association (GWA) mapping



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

LD blocks in humans are 20-100 kb

500,000 common variants gives us a SNP every 10 kb

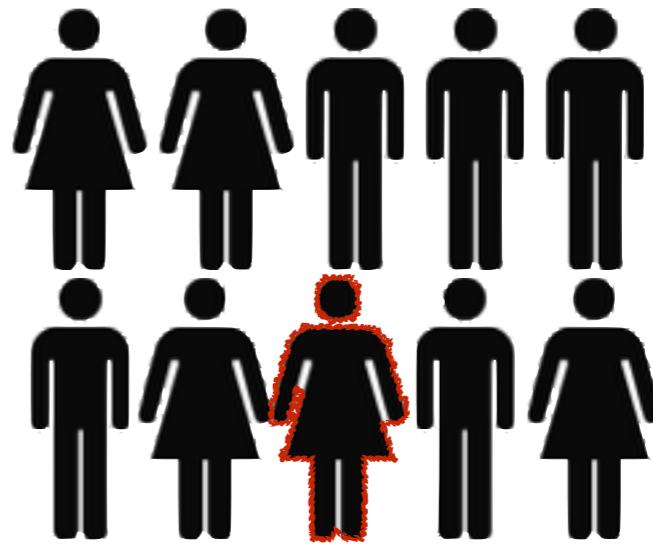
2-10 SNPs mark each LD block for the statistical test

The set up of a genome-wide association (GWA) mapping

Case-control study design



Cases
(People with trait)



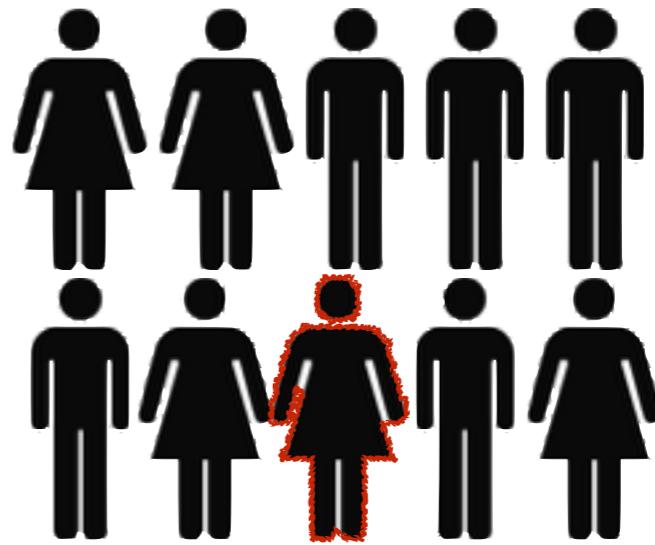
Controls
(People without trait)

The set up of a genome-wide association (GWA) mapping

Case-control study design



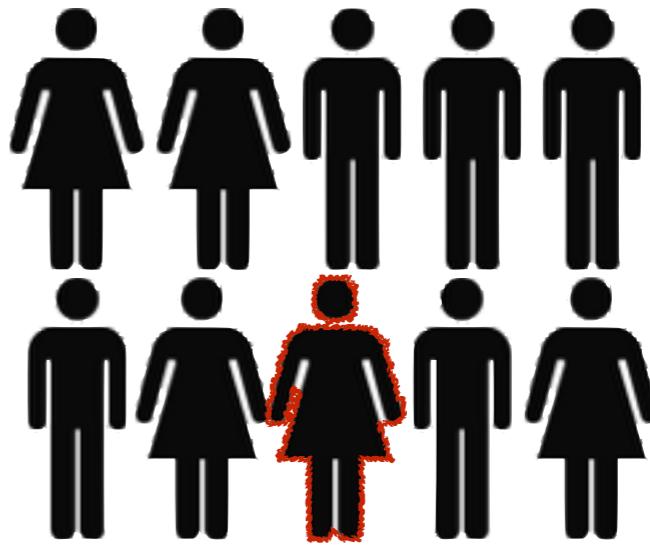
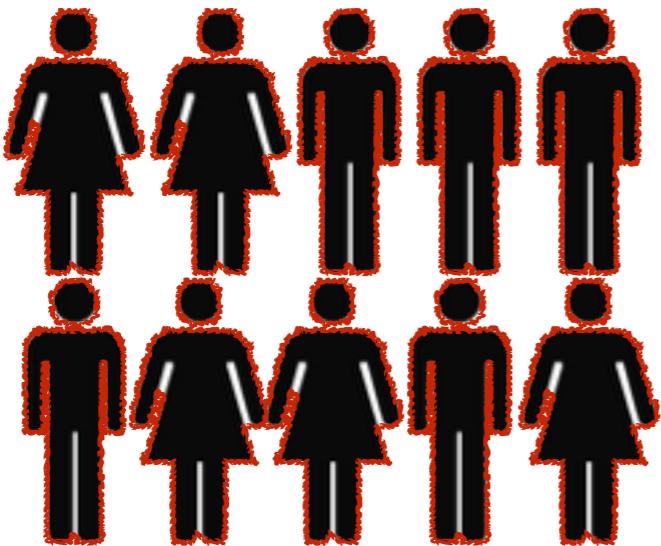
Cases
(People with trait)



Controls
(People without trait)

What alleles do the cases share that the controls lack?

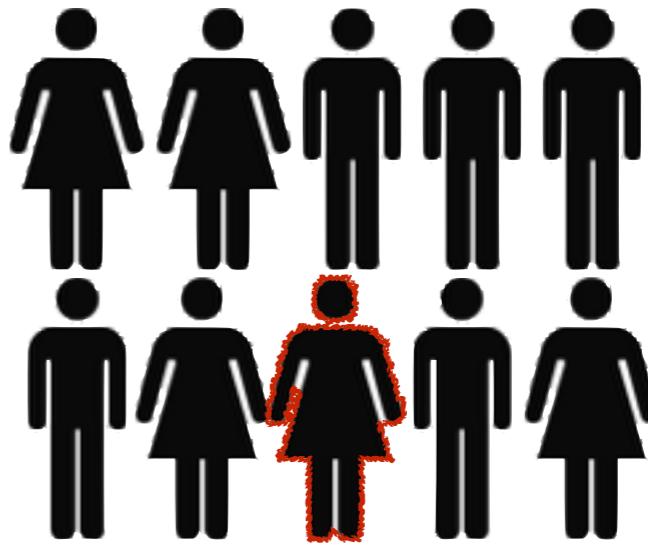
Collect genotype and phenotype data for lots of people



Genotype: SNP arrays (500k) or sequencing

Phenotype: Measure quantitative values

Collect genotype and phenotype data for lots of people

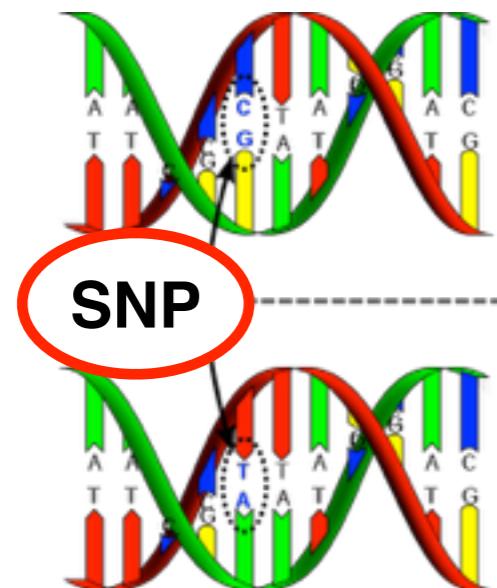


Genotype: SNP arrays (500k) or sequencing

Phenotype: Measure quantitative values

\$250 million spent since 2006 on GWAS

Measure correlation between genetic variation and phenotypic variation in cases and compare to controls

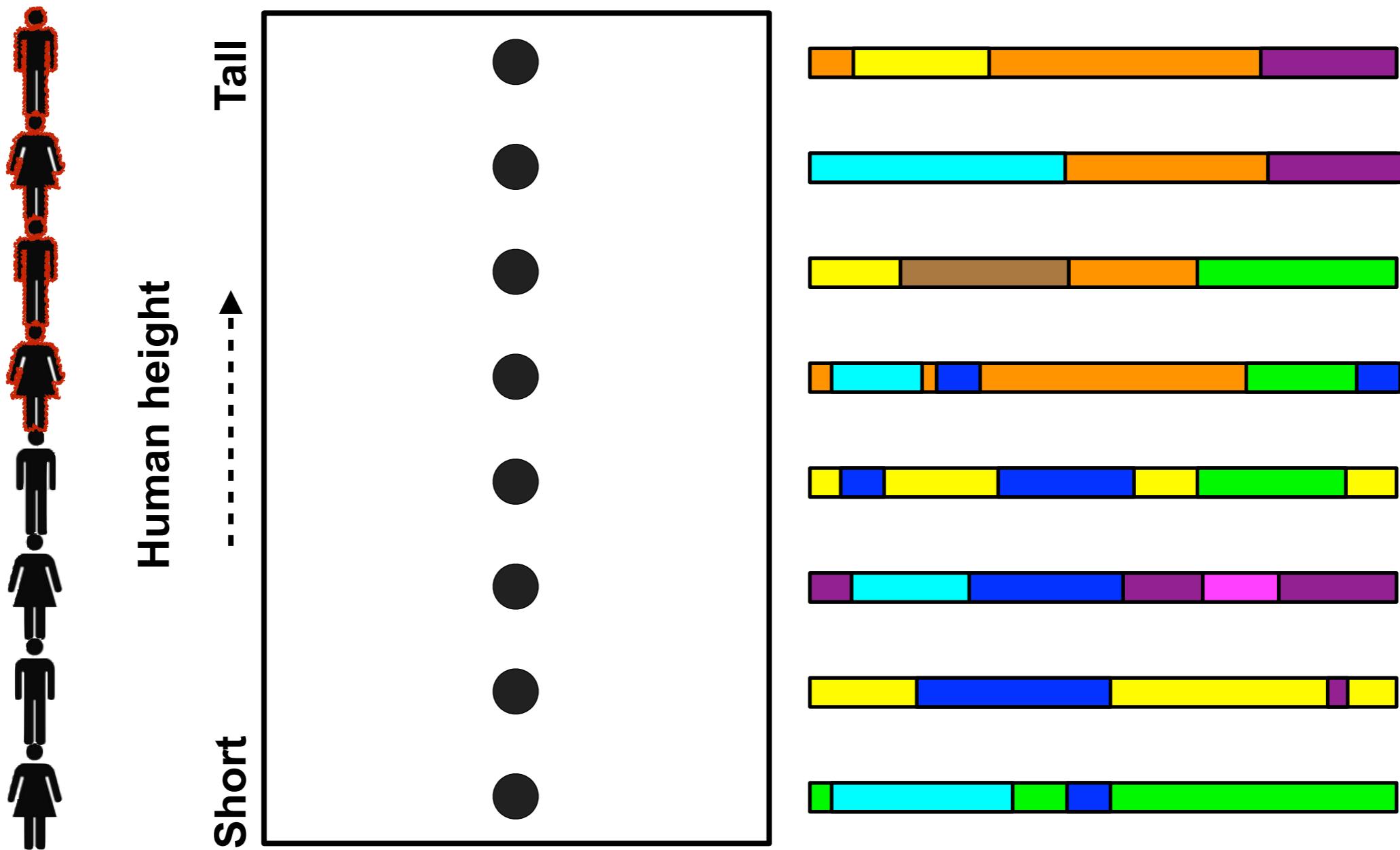


Genetic variation

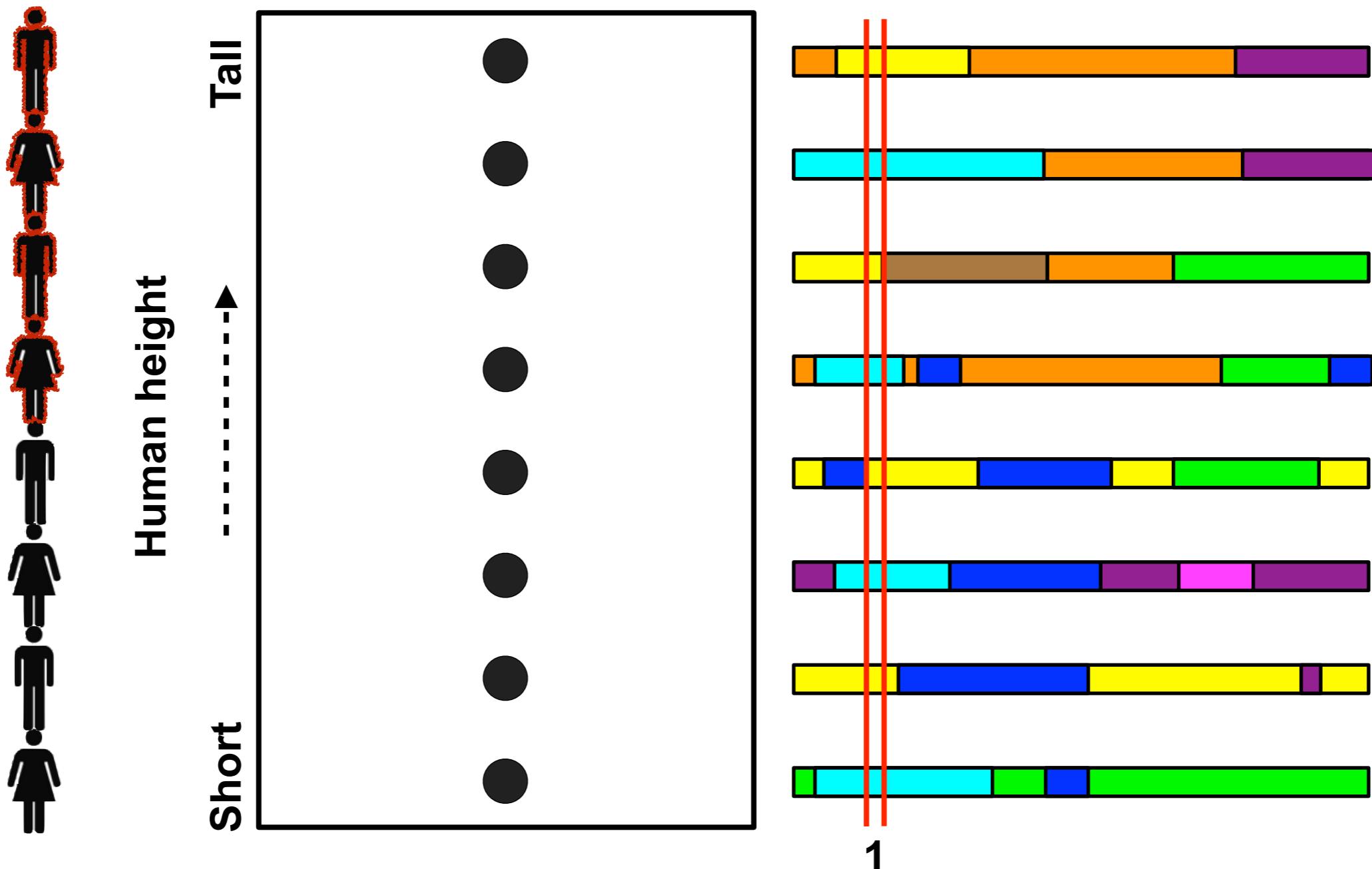


Phenotypic variation

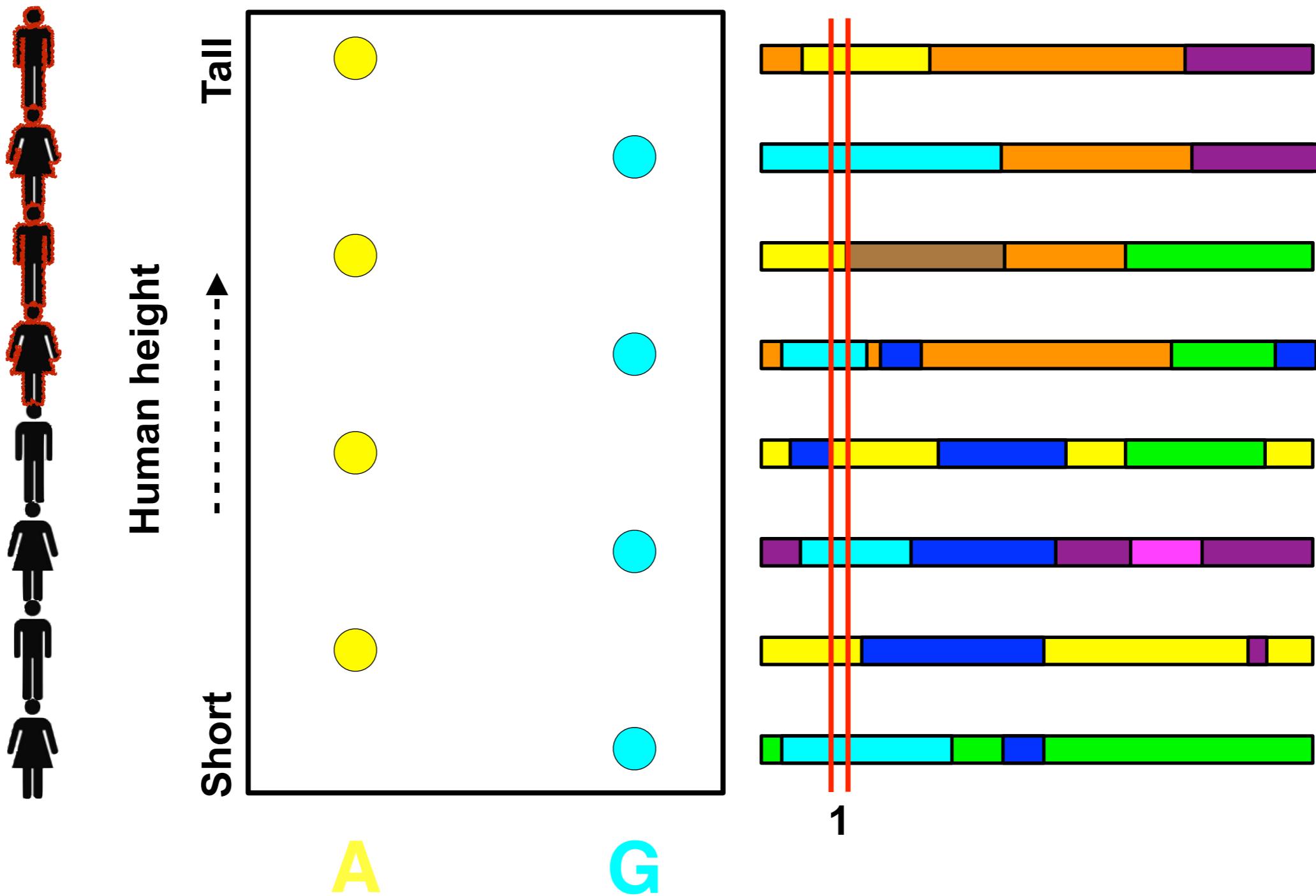
Association mapping: Correlating genotype with phenotype



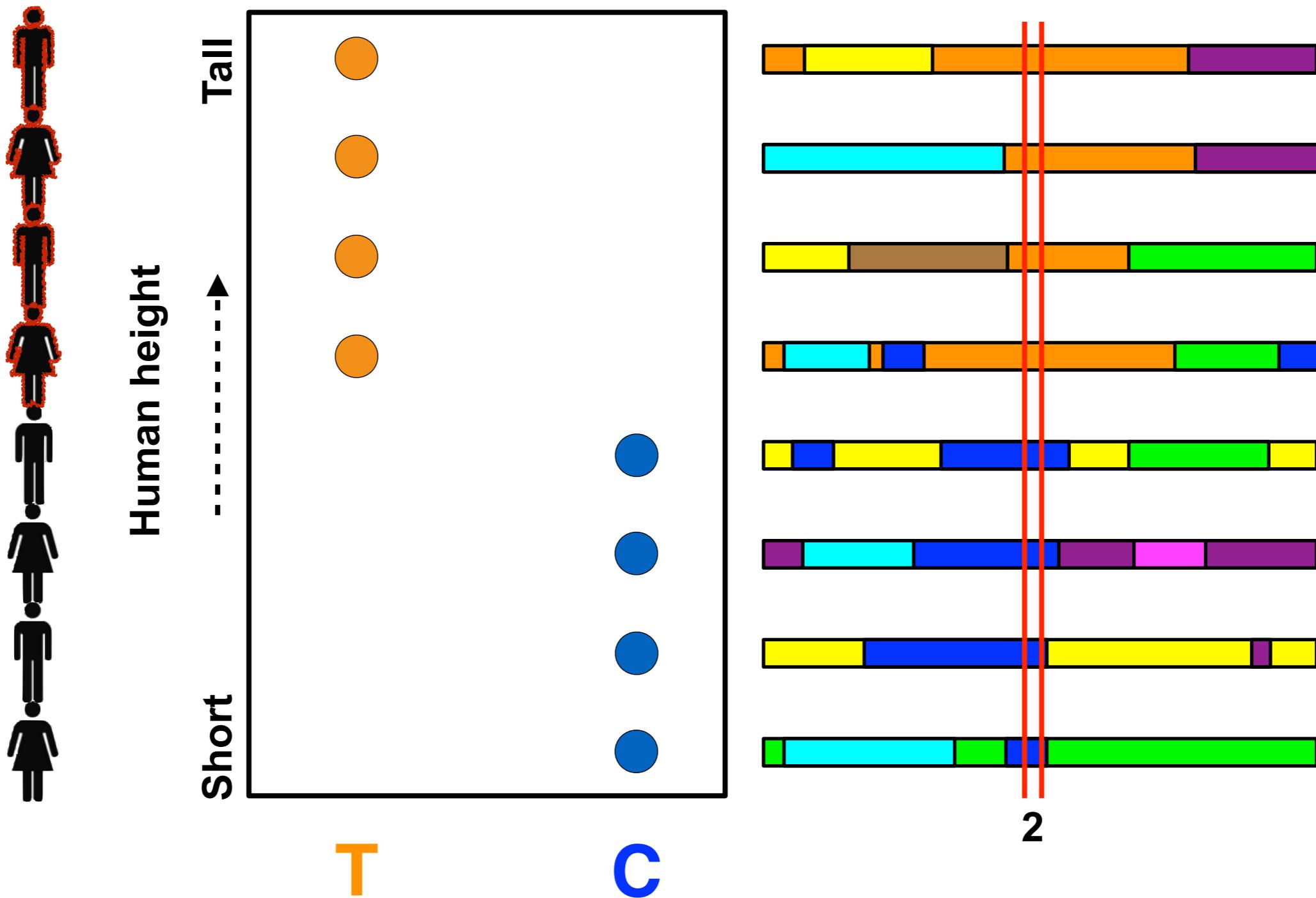
Association mapping: Correlating genotype with phenotype



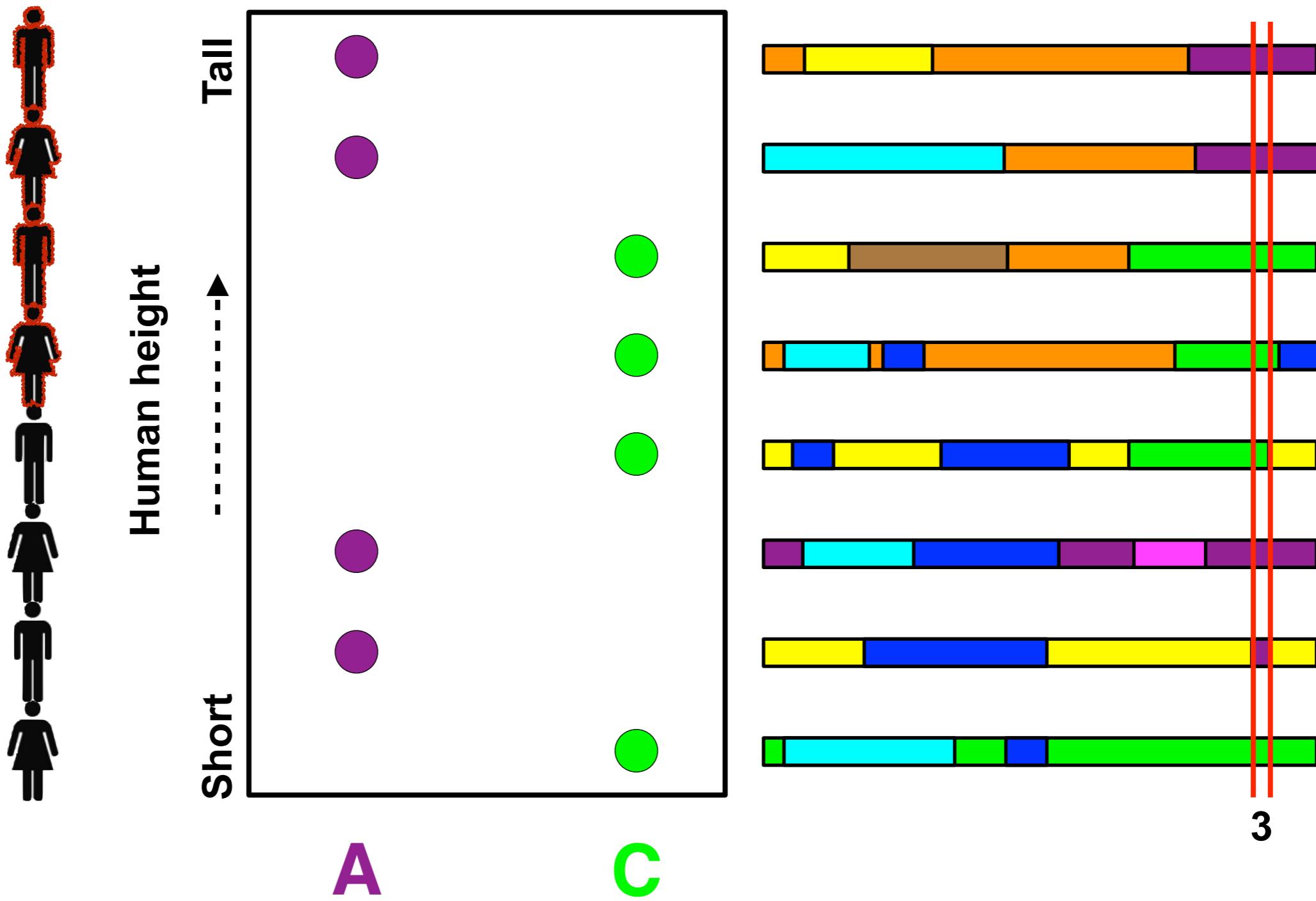
Association mapping: Correlating genotype with phenotype



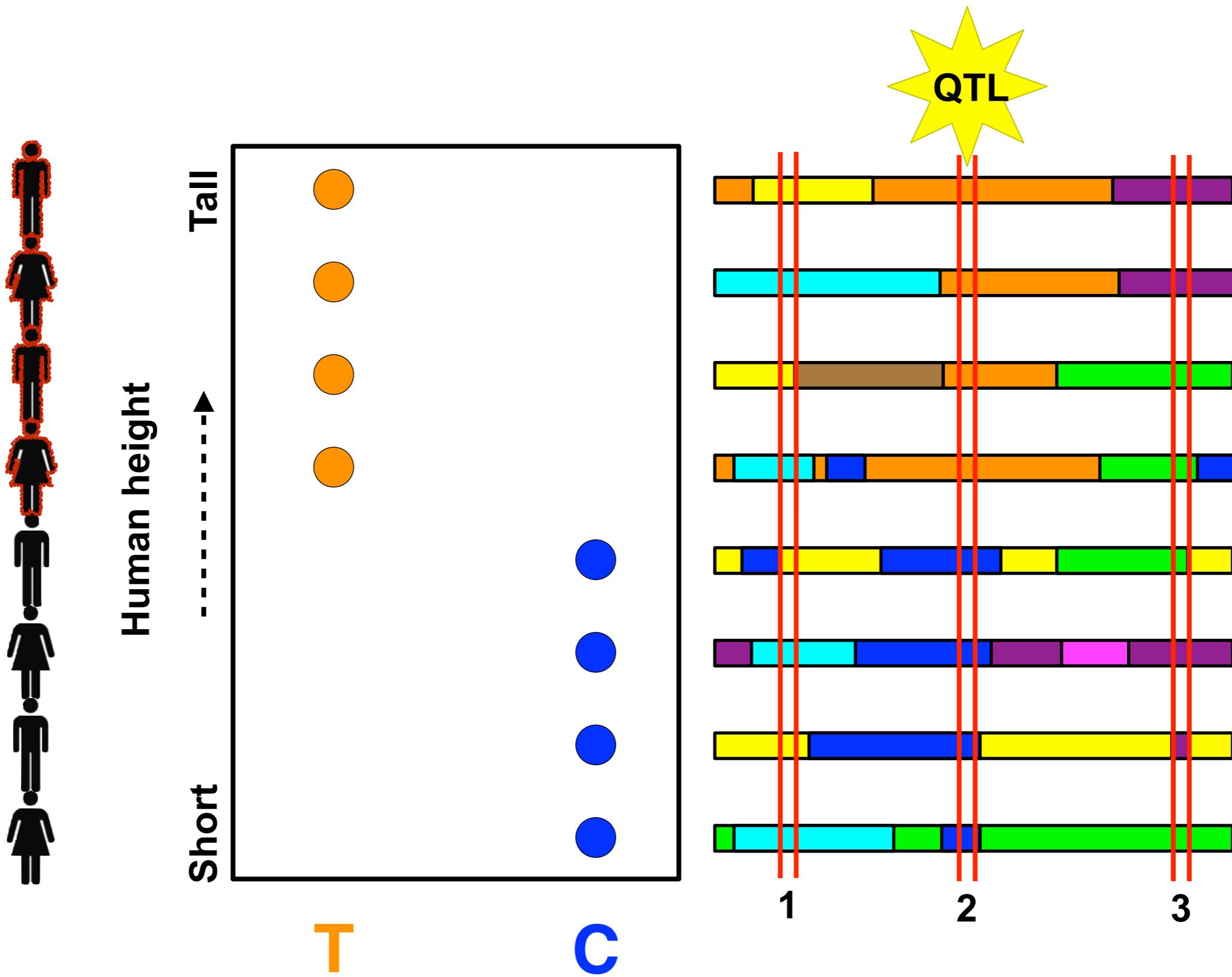
Association mapping: Correlating genotype with phenotype



Association mapping: Correlating genotype with phenotype

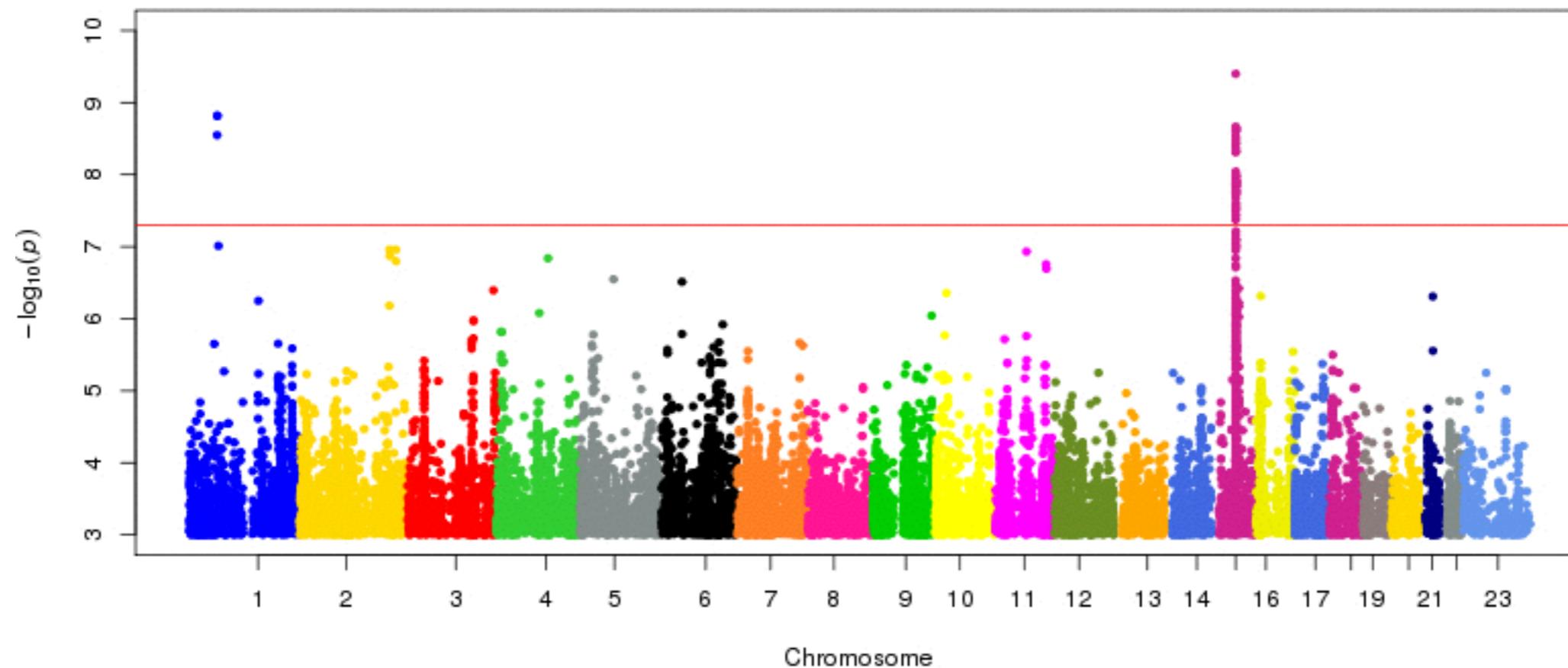


Association mapping: Correlating genotype with phenotype



QTL=Quantitative Trait Locus

An example Manhattan plot of GWA mapping results



Styrkarsdottir *et al.* Nature 2014