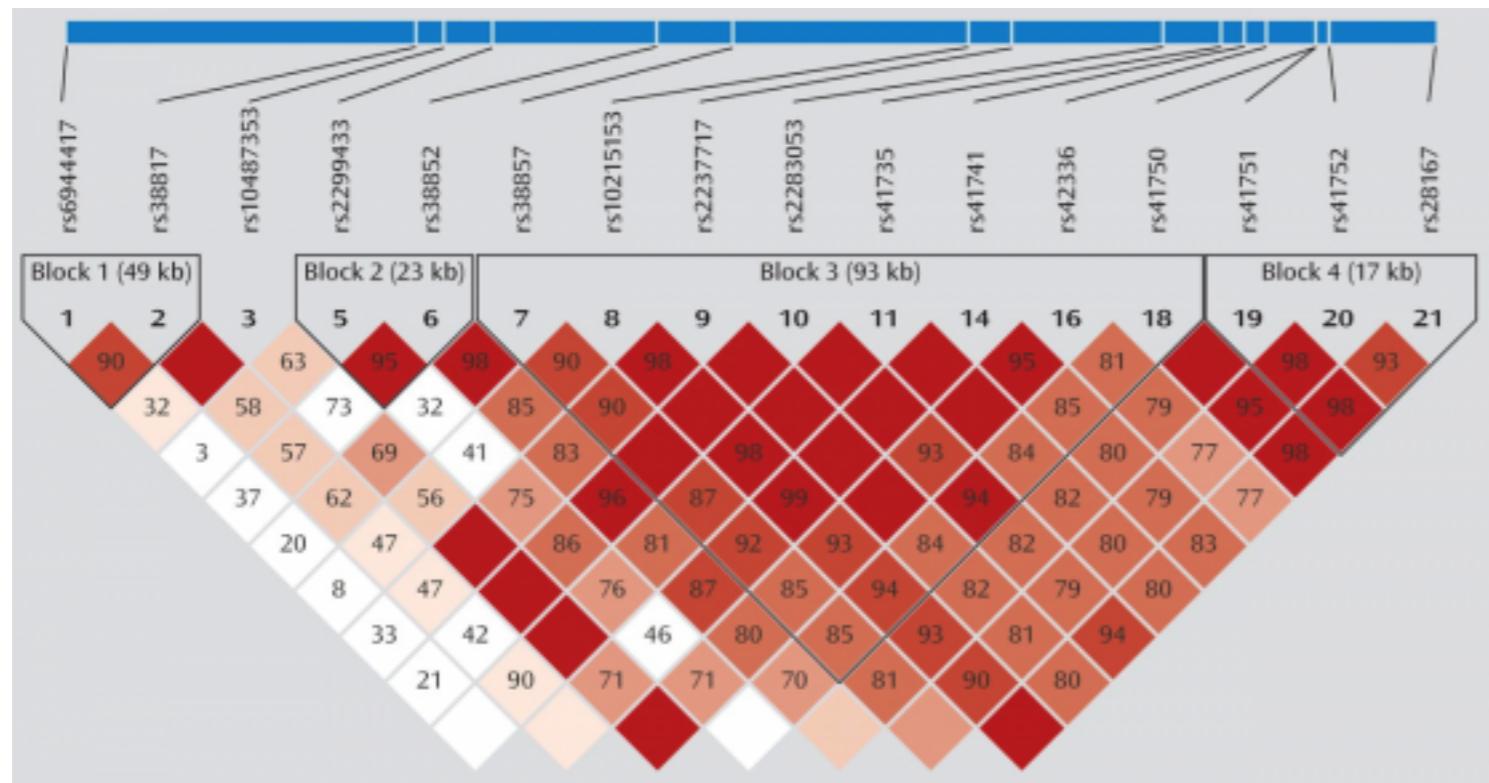
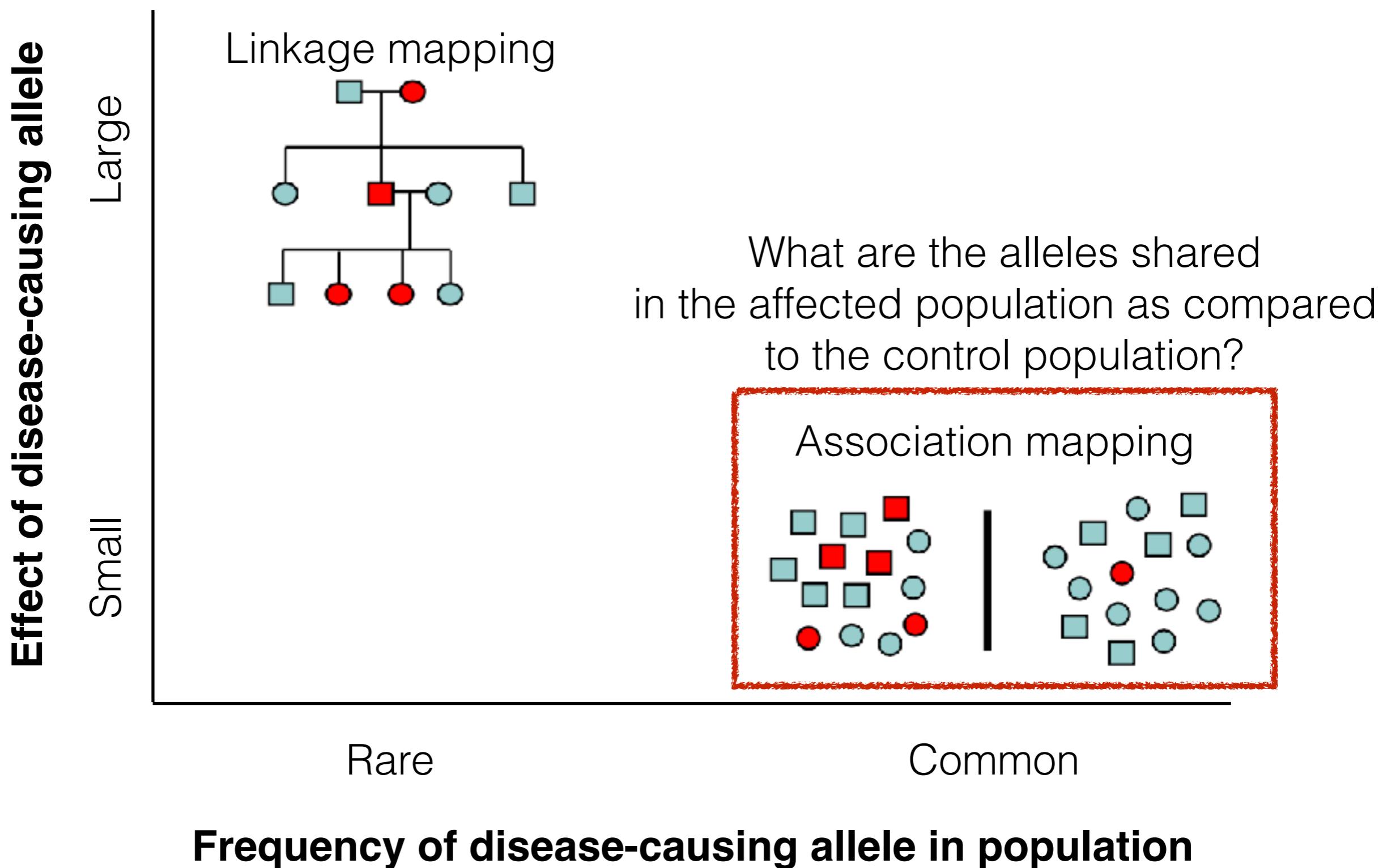


# Bio393: Genetic Analysis

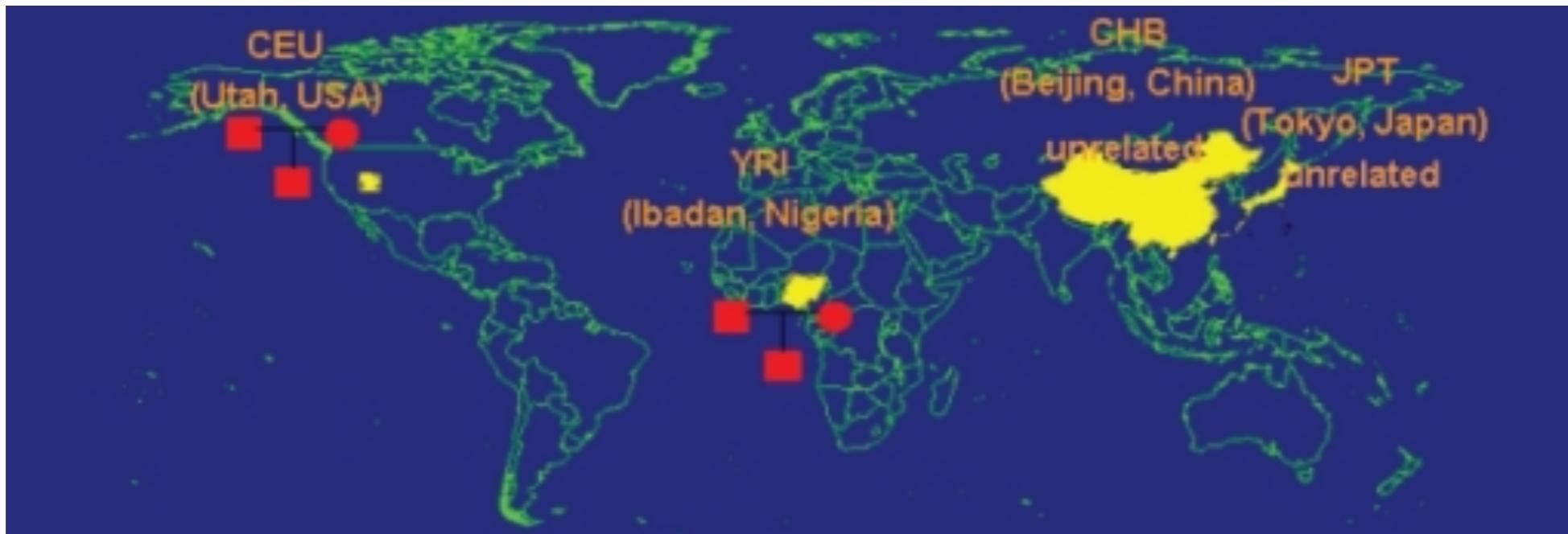
## Linkage disequilibrium, haplotypes, and GWAS



# Human gene mapping has two general flavors



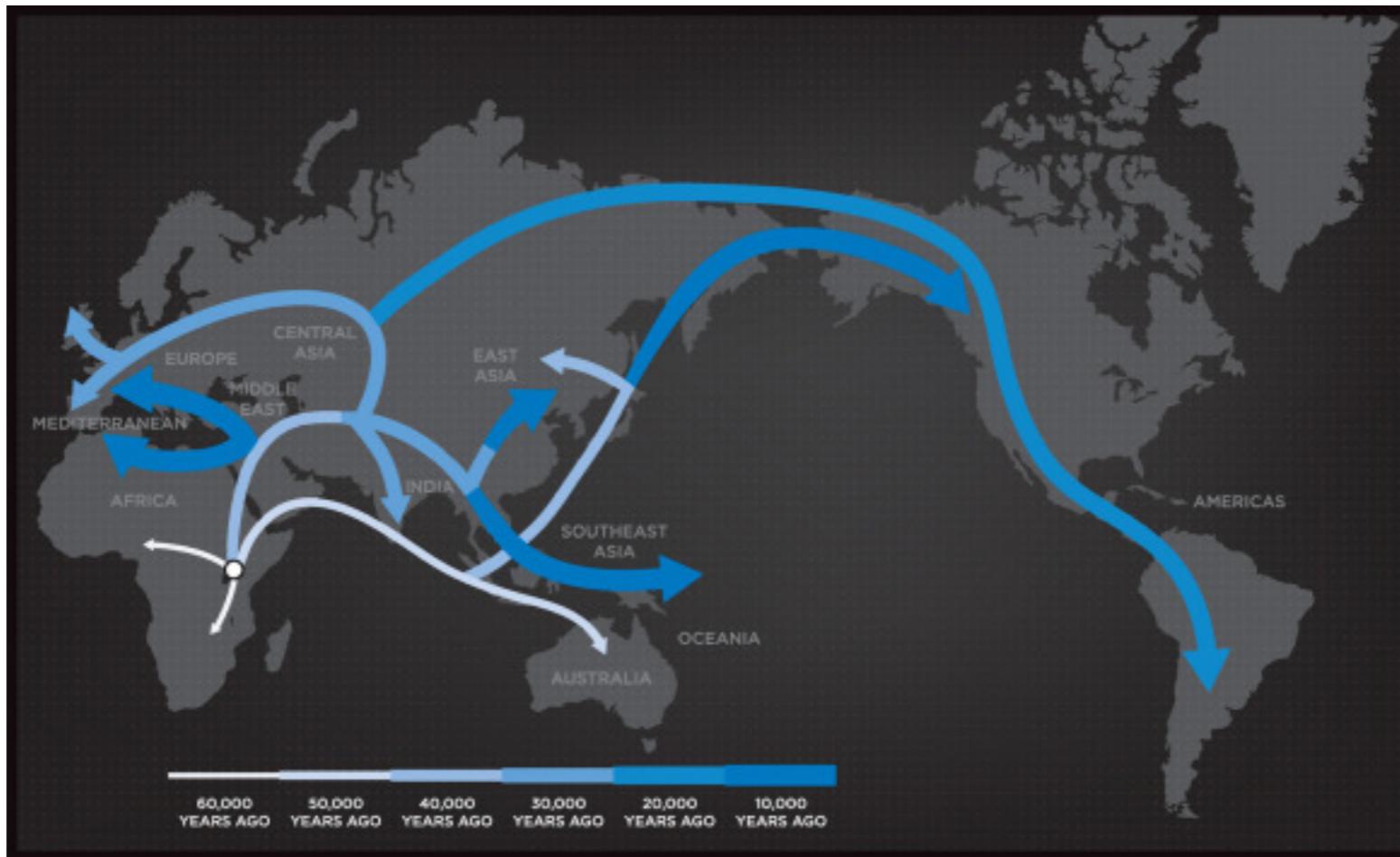
# Common variants facilitate genome-wide association (GWA) mapping



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

Do we have to test them all?

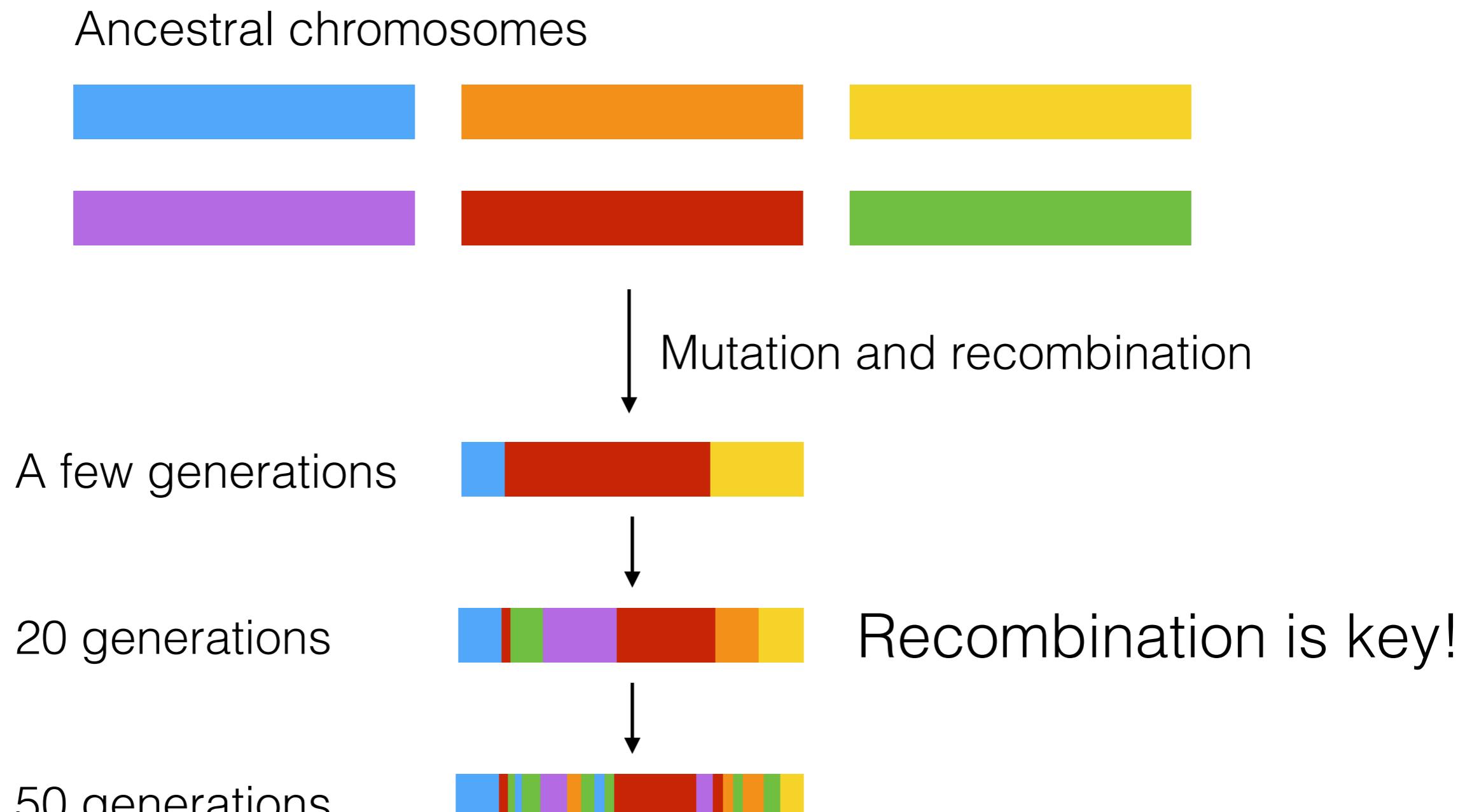
# Common variants facilitate genome-wide association (GWA) mapping



Our relatedness means that variants are correlated in populations

**Correlation between variants is called linkage disequilibrium (LD)**

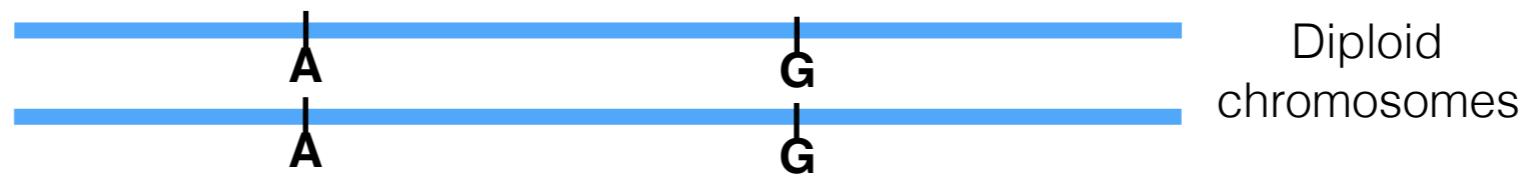
# Linkage disequilibrium (LD) is the non-random association of alleles at different loci



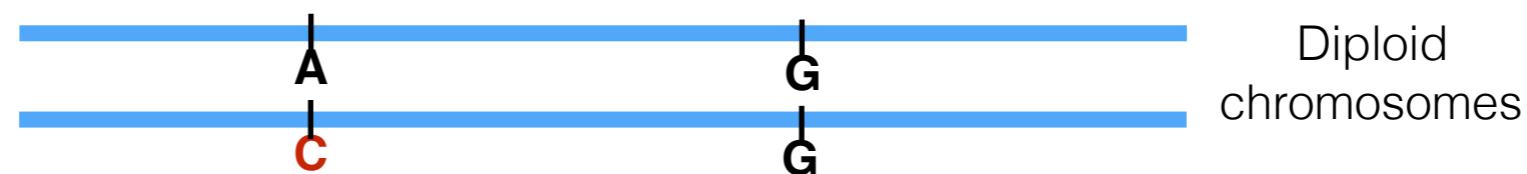
LD makes genotyping easier and cheaper

# Many alleles that exist today are from ancient mutation events

Before mutation



After mutation



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

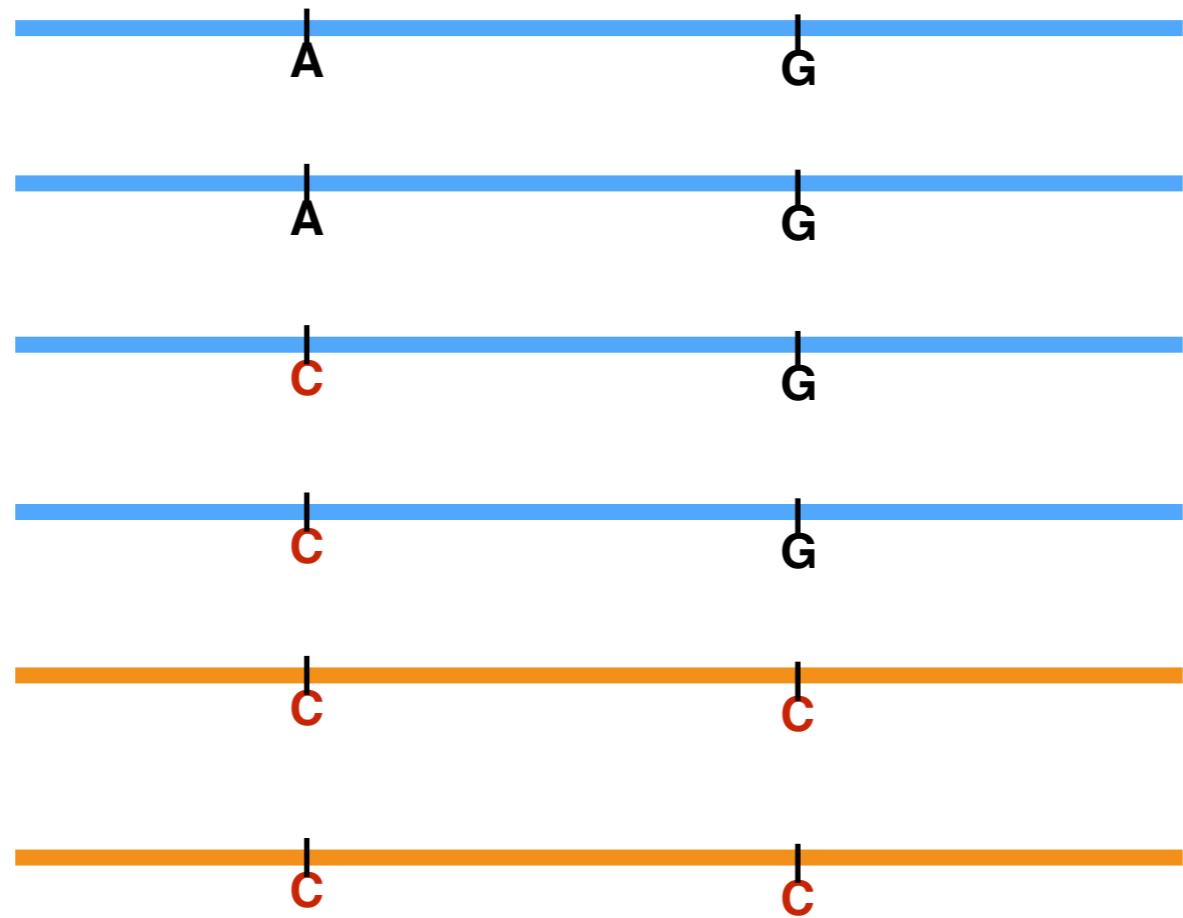
Before mutation



After mutation



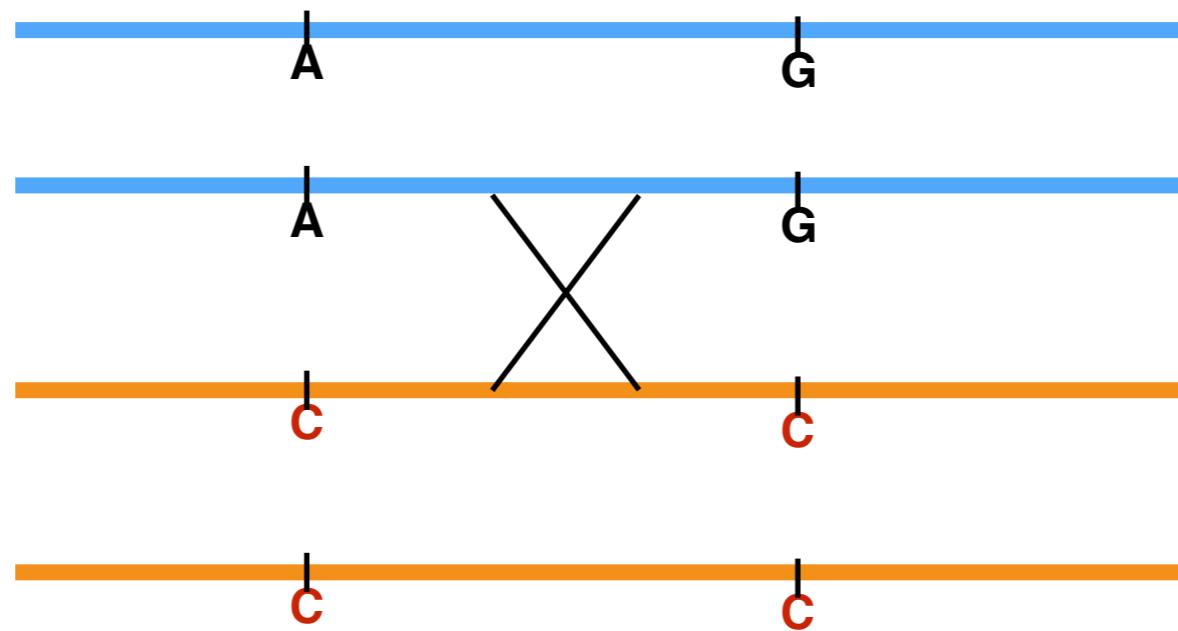
# Let's think about these chromosomes with different arrangements of alleles as haploid gametes



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

# Recombination creates new arrangements of ancestral alleles

Before recombination



After recombination

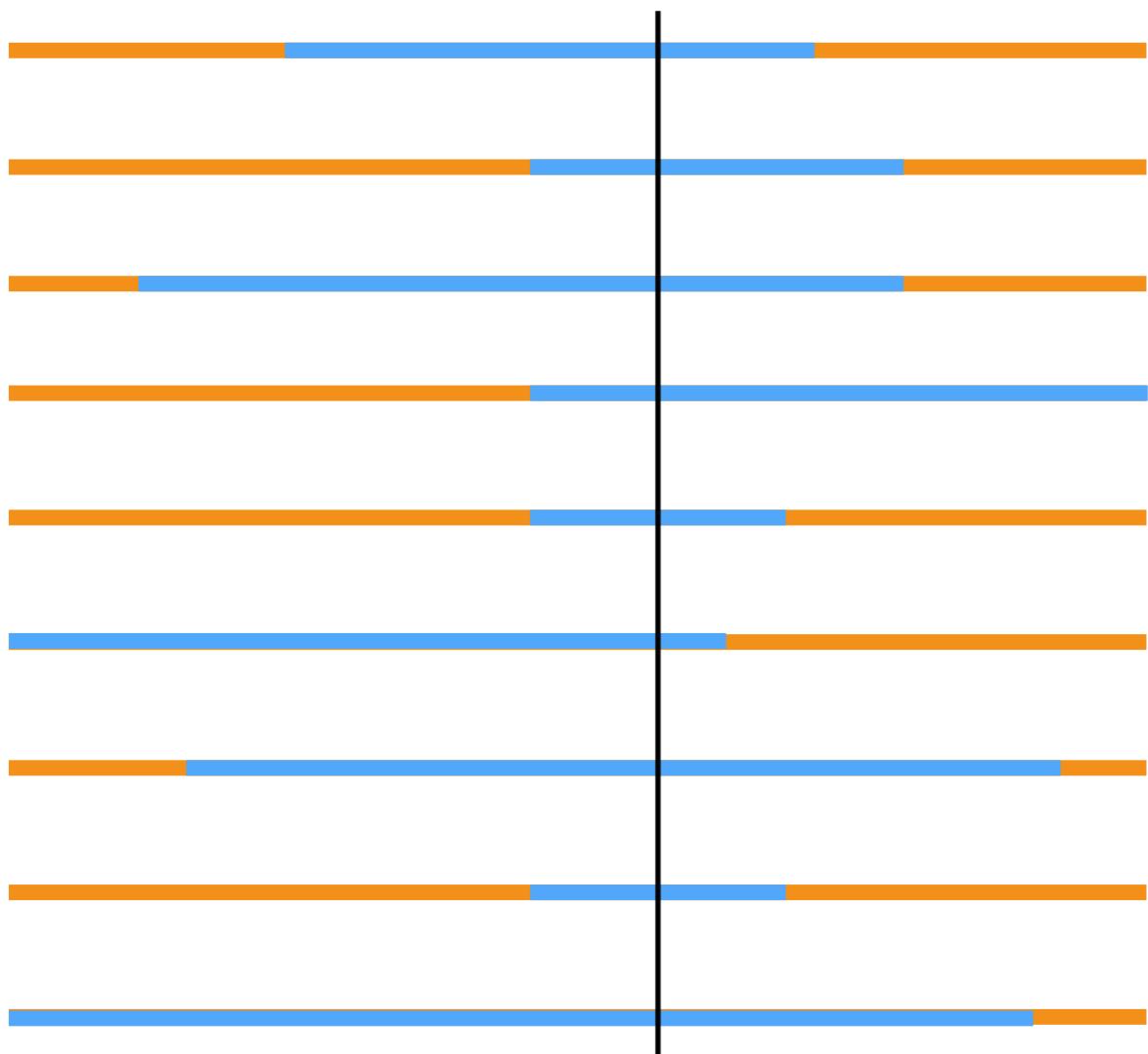


# Linkage disequilibrium is the non-random association of alleles at different loci

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  $p_A * p_B \neq p_{AB}$ , 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 rs2</u>		$p_{AB} = p_A * p_B$
		B	b	
<u>Locus rs1</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  $D_{rs1-rs2}$

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

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

Otherwise,  $D_{rs1-rs2} >$  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:

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

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

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

# How to calculate LD? An example

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

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

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

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

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

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:

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

$$p_a = 1 - p_A$$

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

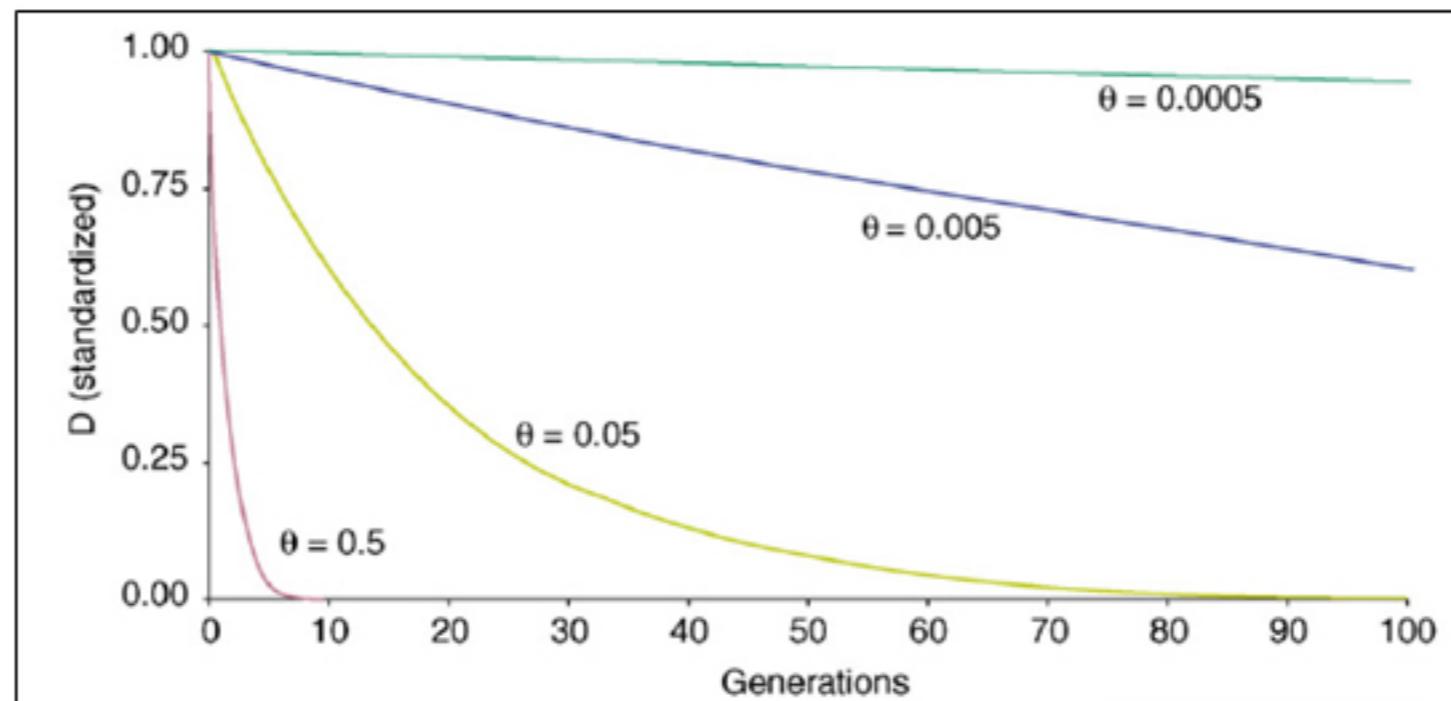
$$p_b = 1 - p_B$$

Allele	Number	Frequency
A	700	0.7
a	300	0.3
Haplotype	Number	Frequency
B	800	0.8
b	200	0.2
AB	600	0.6
Ab	100	0.1
aB	200	0.2
ab	100	0.1

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

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

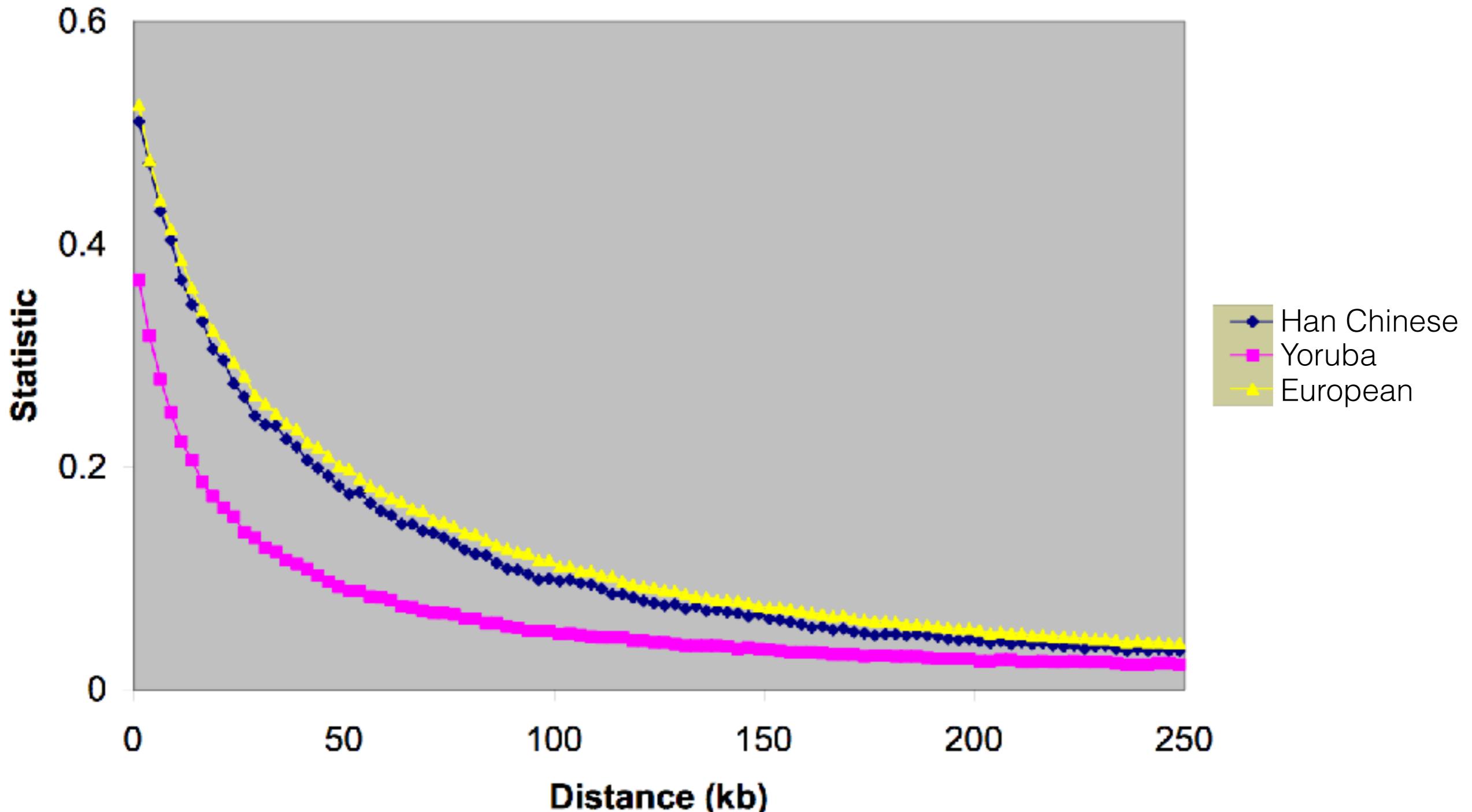
# Linkage disequilibrium decreases by distance and generation time



Mackay and Powell 2007

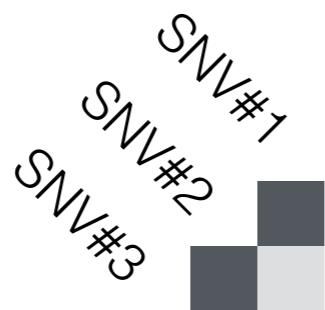
Recombination is key!

# Linkage disequilibrium varies among different populations



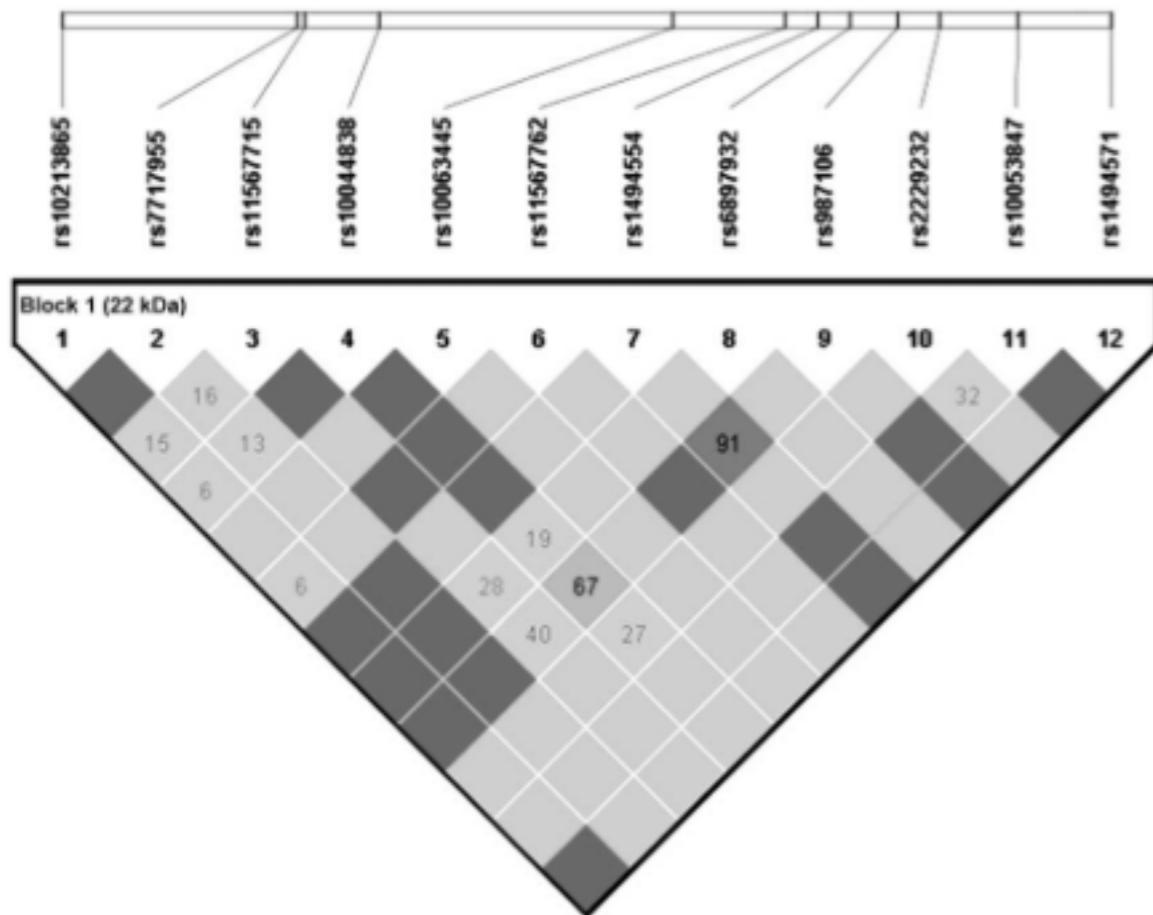
Recombination is key!

# Linkage disequilibrium is often shown as a triangle correlation plot

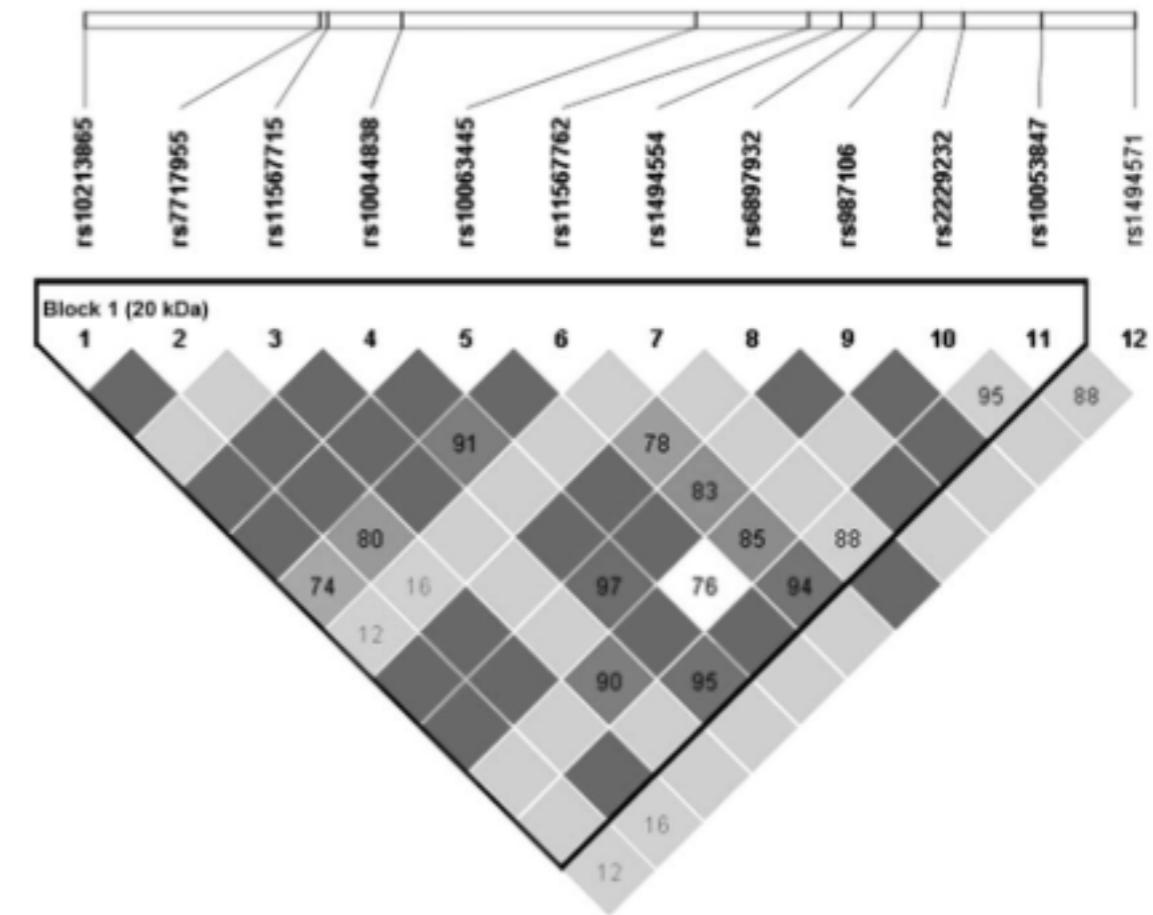


SNV#1 and SNV#2 have high LD  
SNV#2 and SNV#3 have high LD  
SNV#1 and SNV#3 have low LD

**A LD plot of Africans**



**B LD plot of Asians**

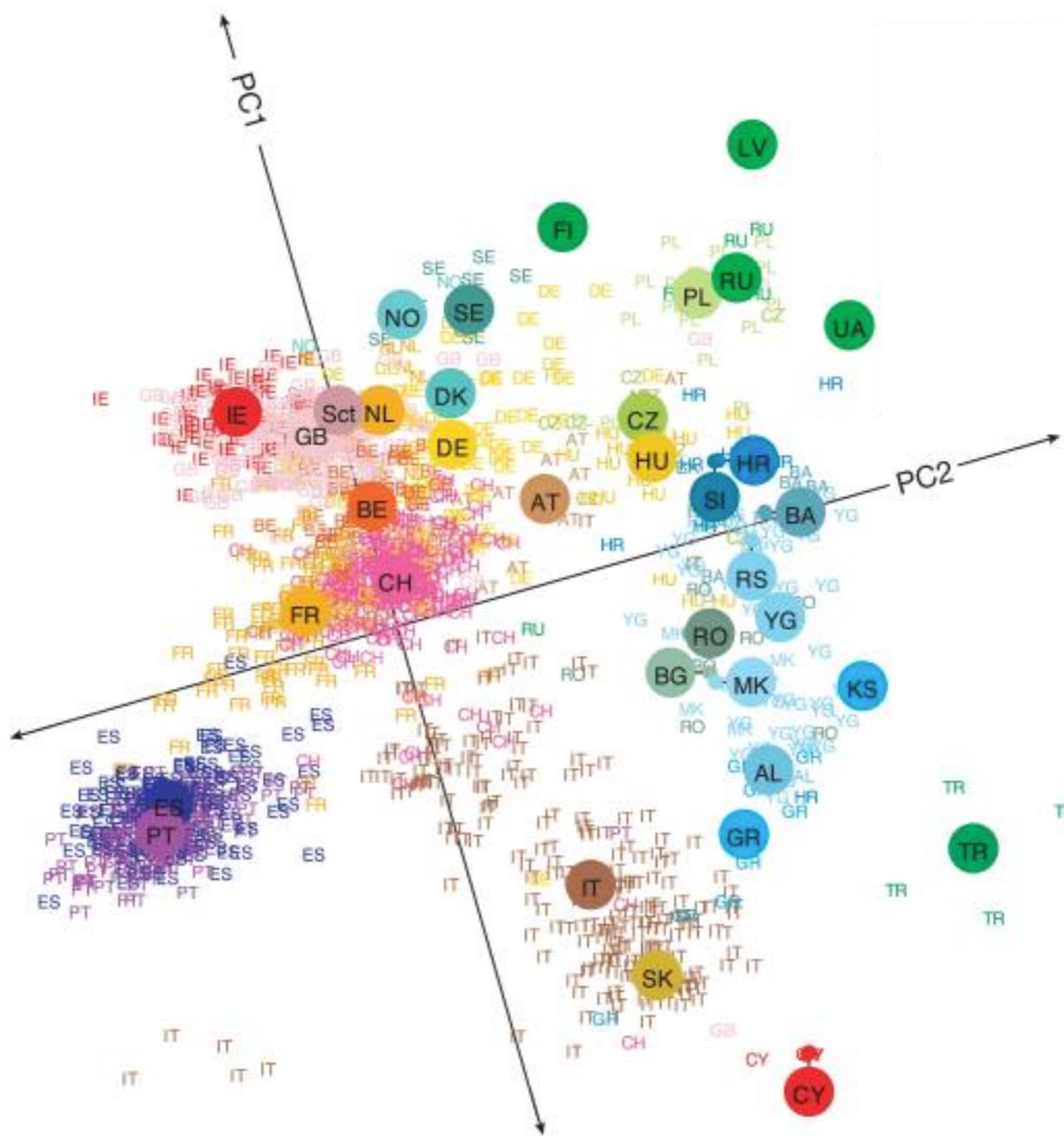


# **LD leads to population structure - alleles found together in populations**

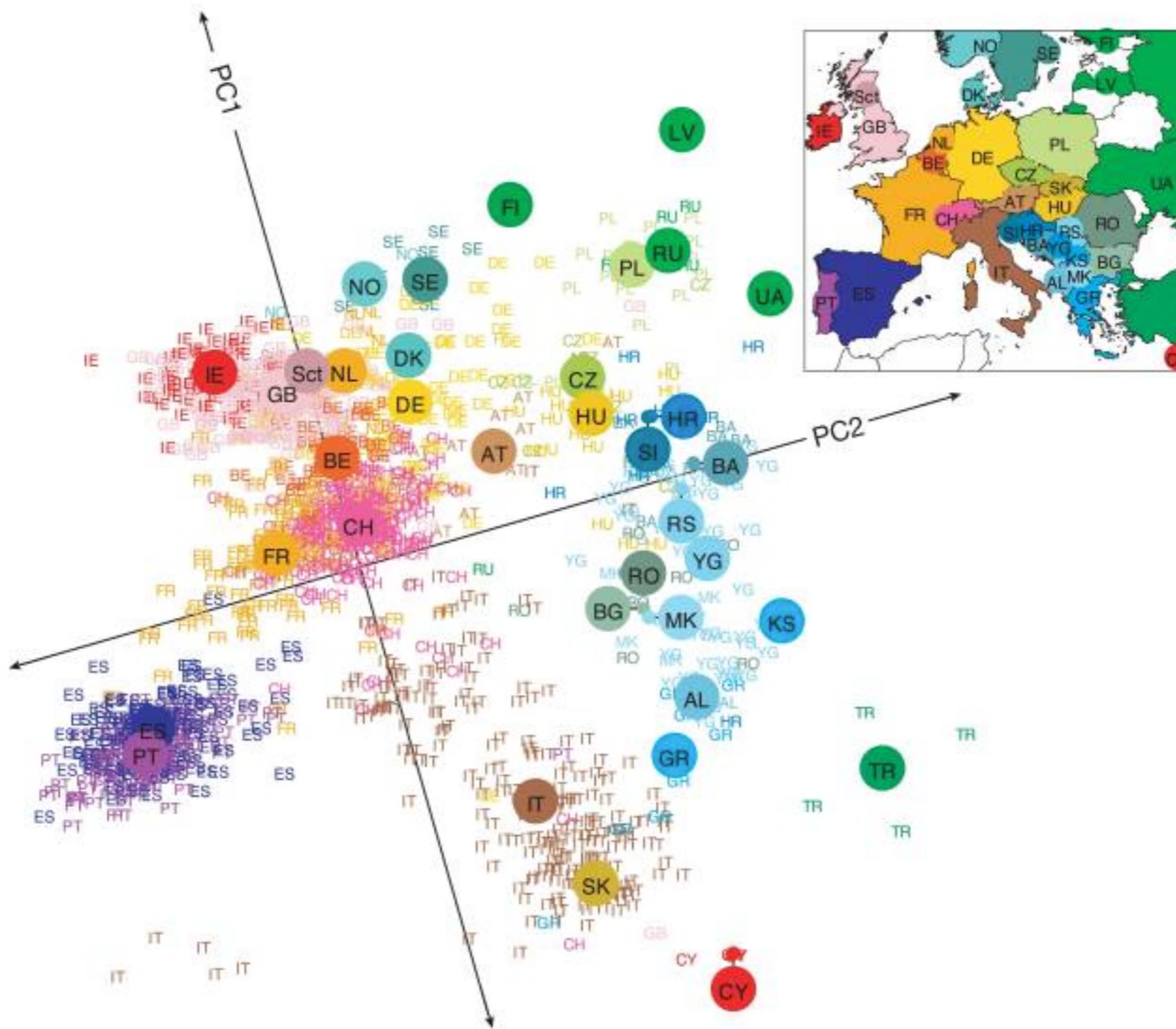


Relatedness of people caused by non-random mating  
is called population structure (or stratification)

# **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



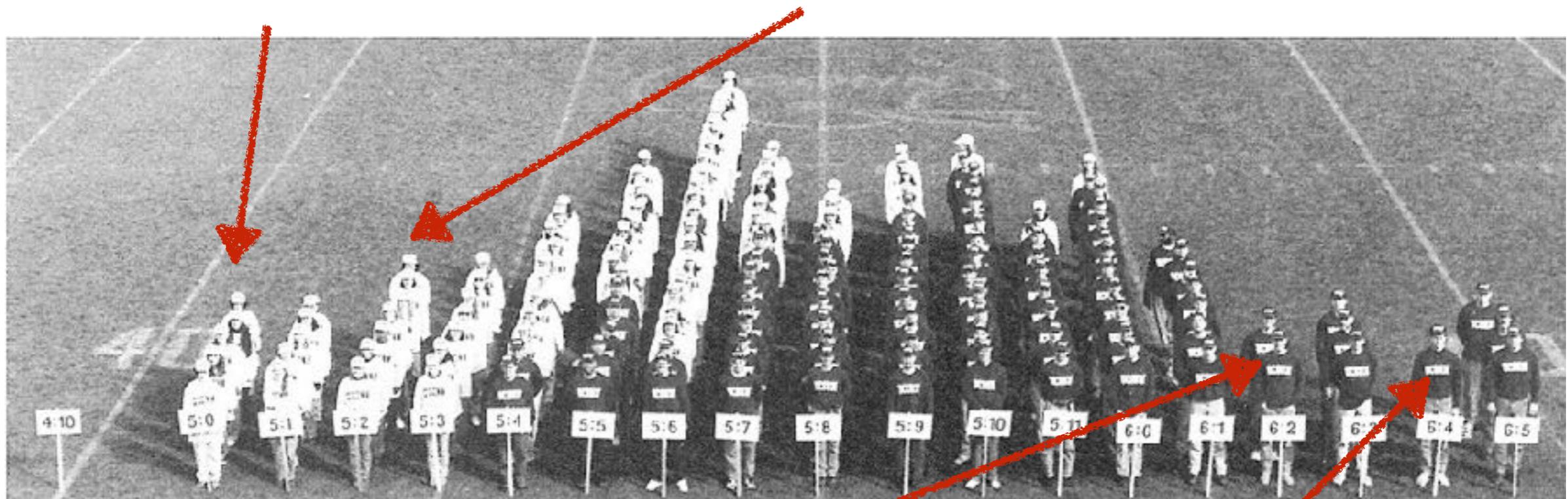
vs.



How many people need to be genotyped?

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

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



University of Connecticut, 1997

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

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

# 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 SNV every 10 kb

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

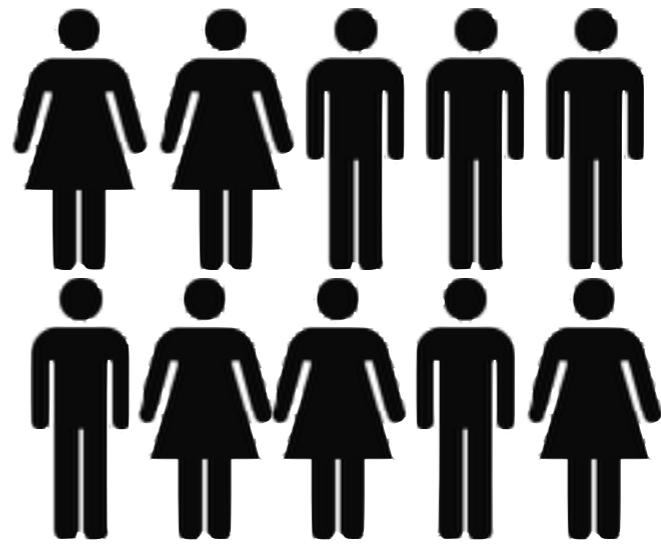
Why do 4.3M SNV tests on current arrays?

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

Case-control study design



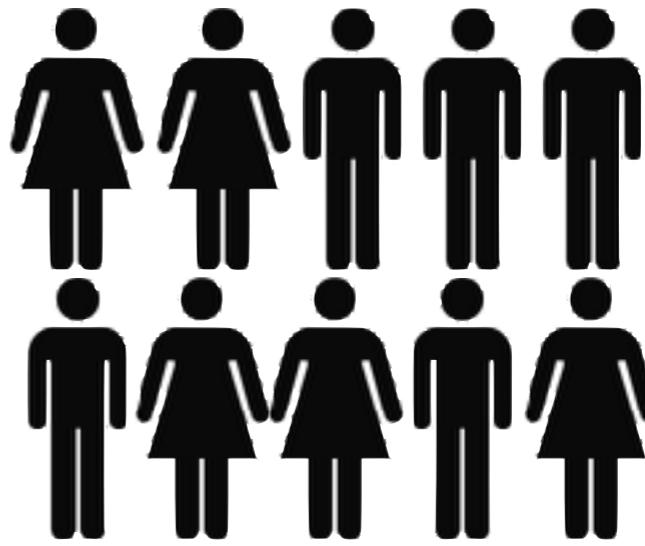
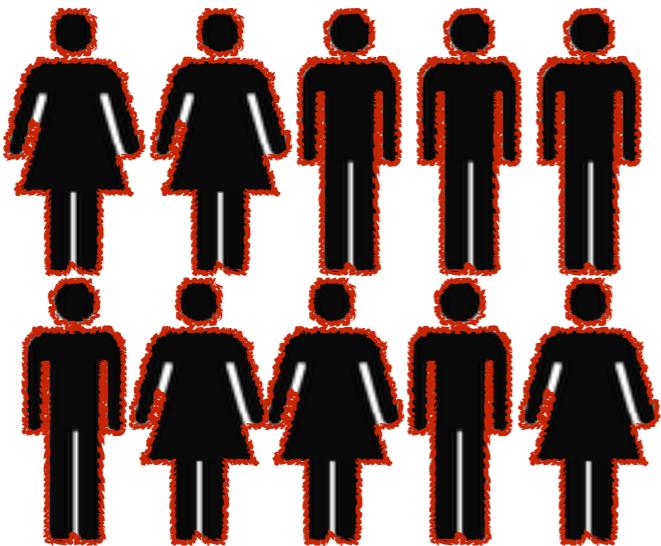
Cases  
(People with trait)



Controls  
(People without trait)

What alleles do the cases share that the controls lack?

# Collect genotype and phenotype data for lots of people

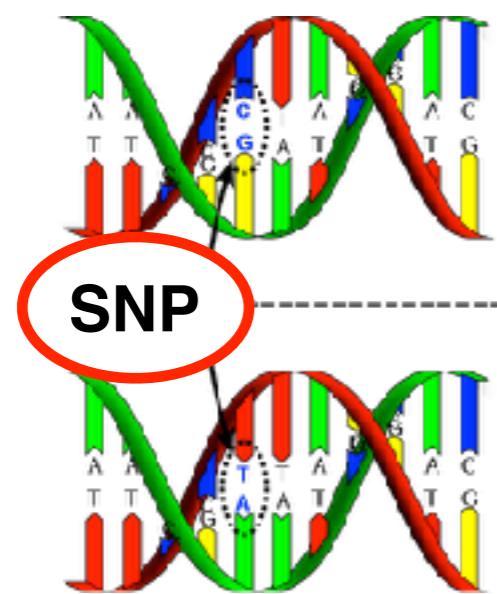


Genotype: SNV arrays (>500k) or sequencing

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

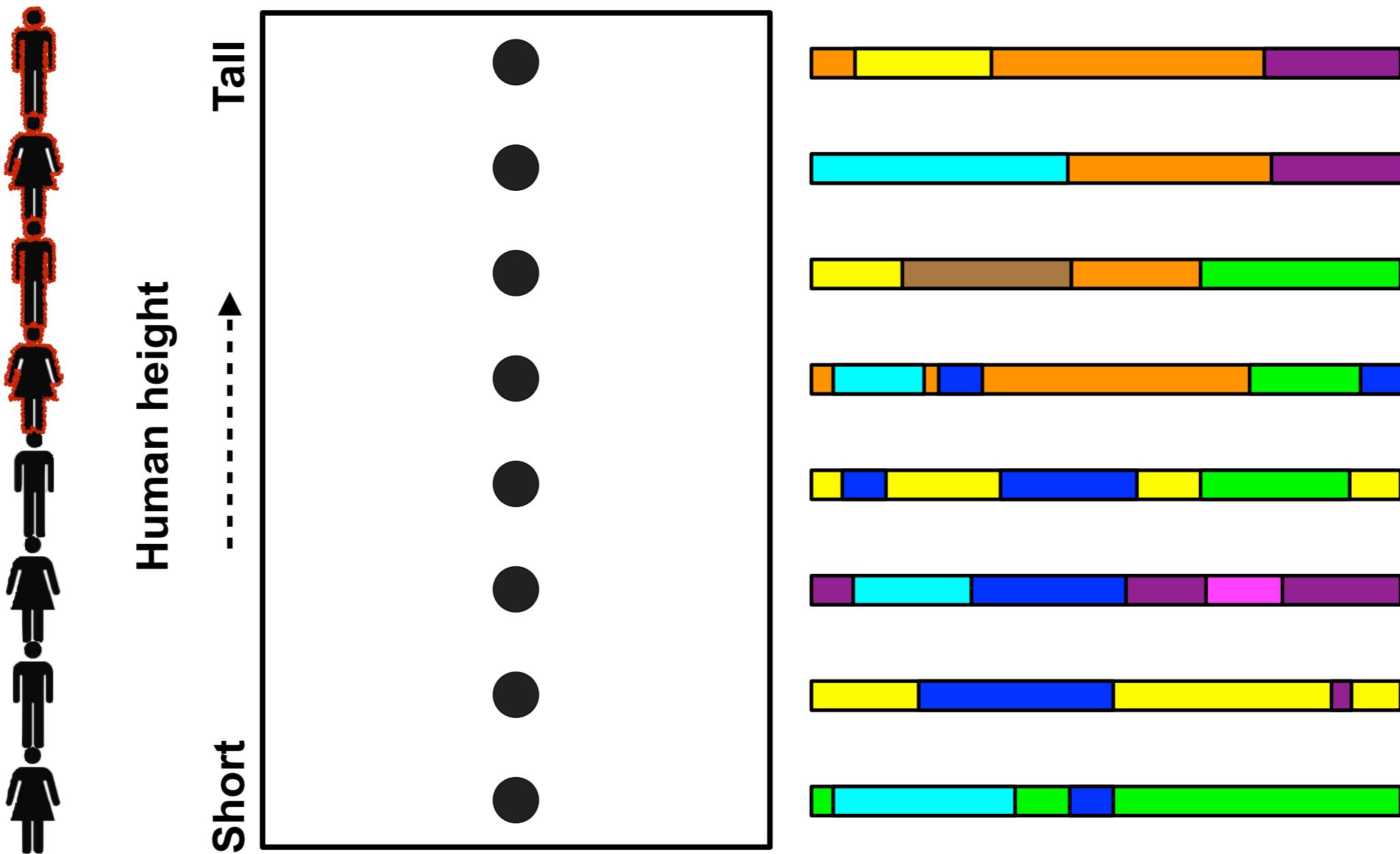


**Genetic variation**

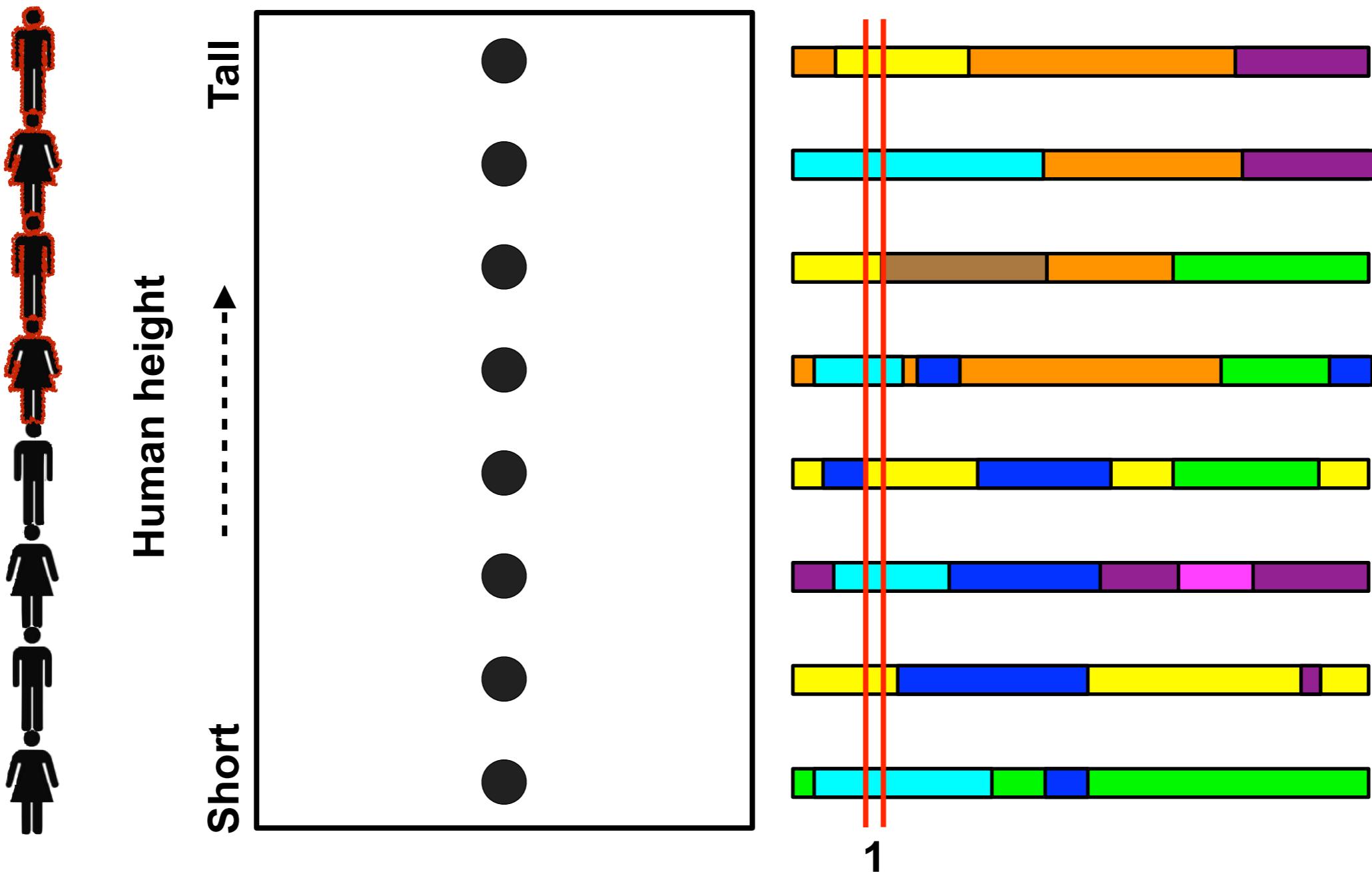


**Phenotypic variation**

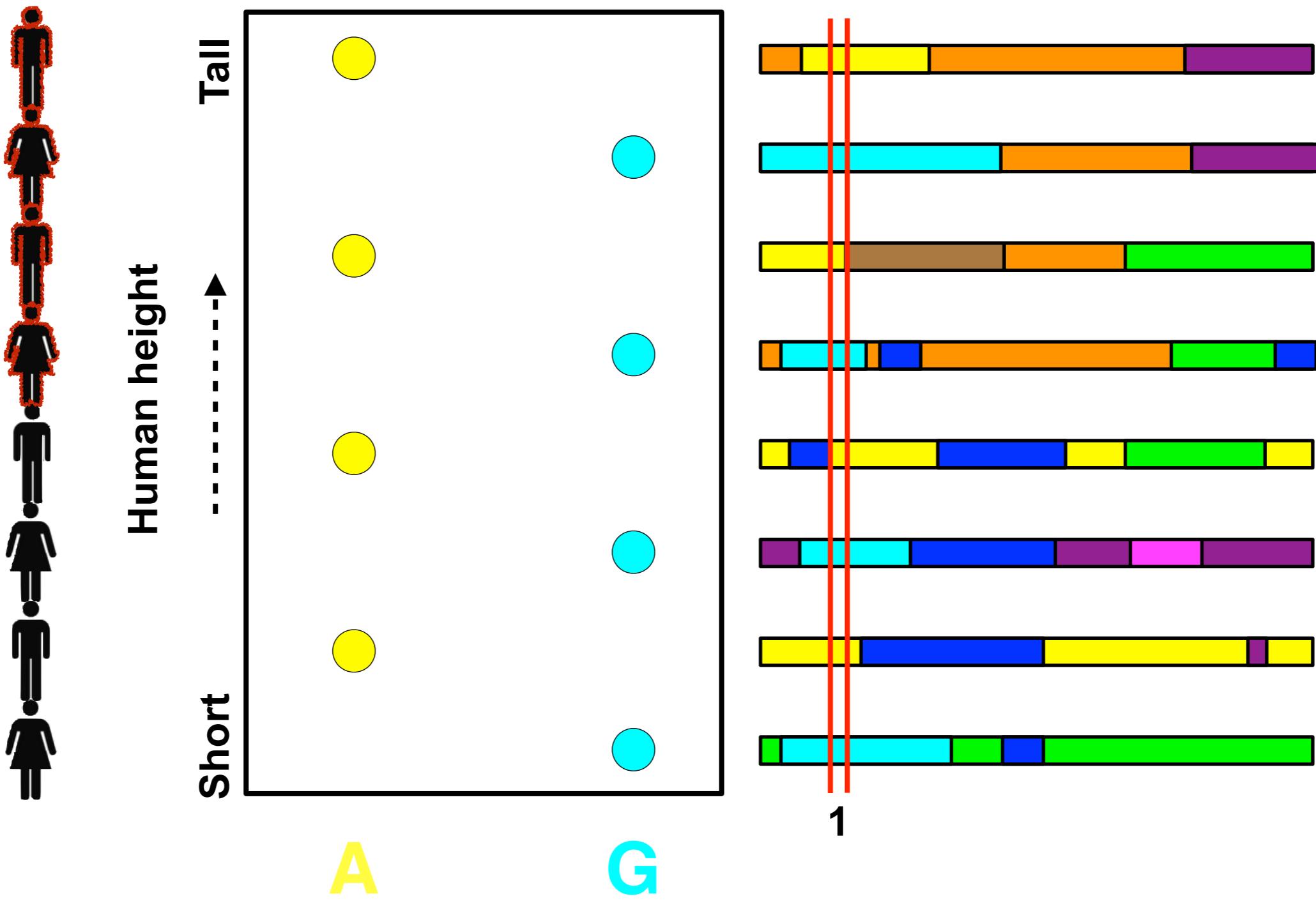
# Association mapping: Correlating genotype with phenotype



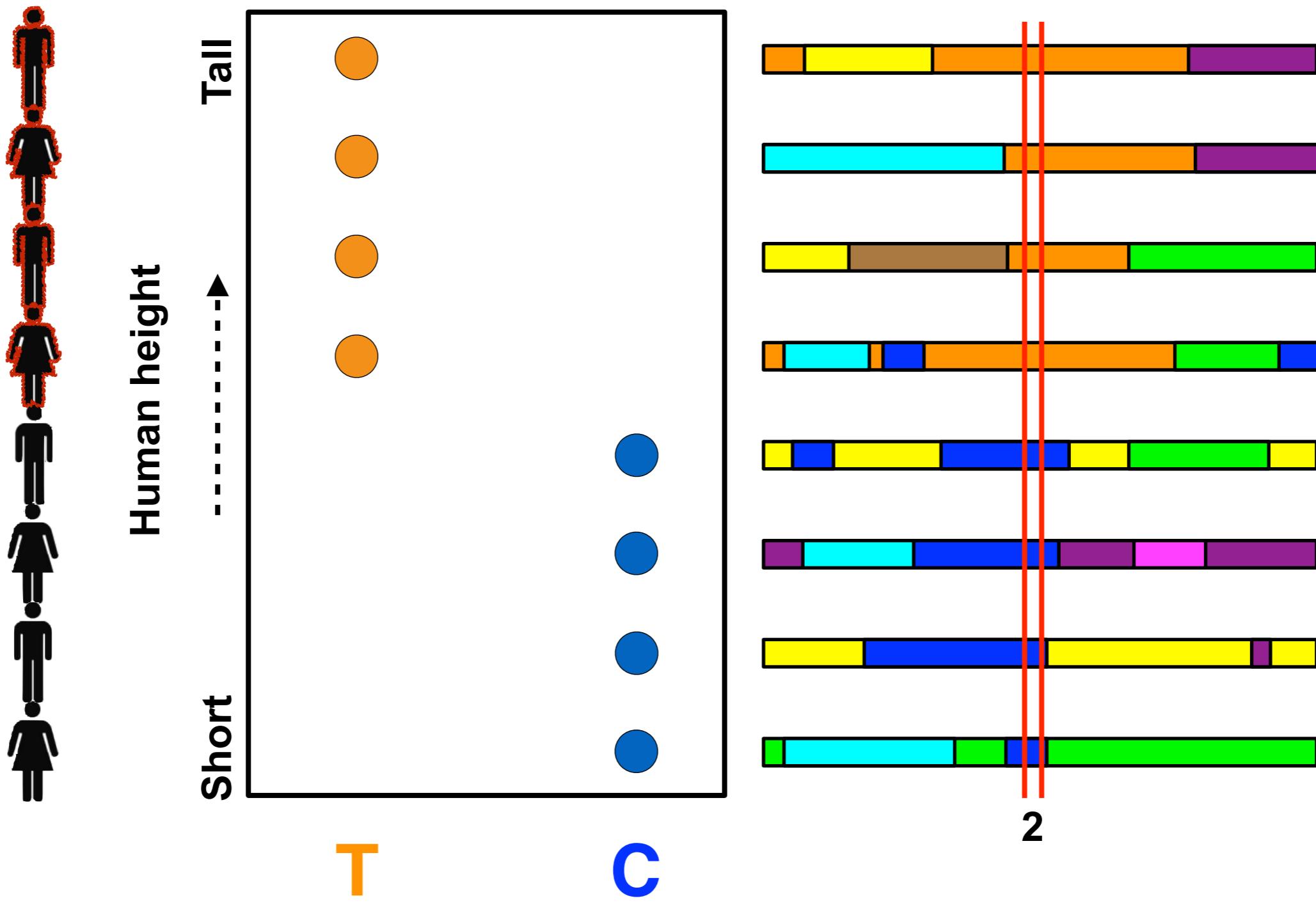
# Association mapping: Correlating genotype with phenotype



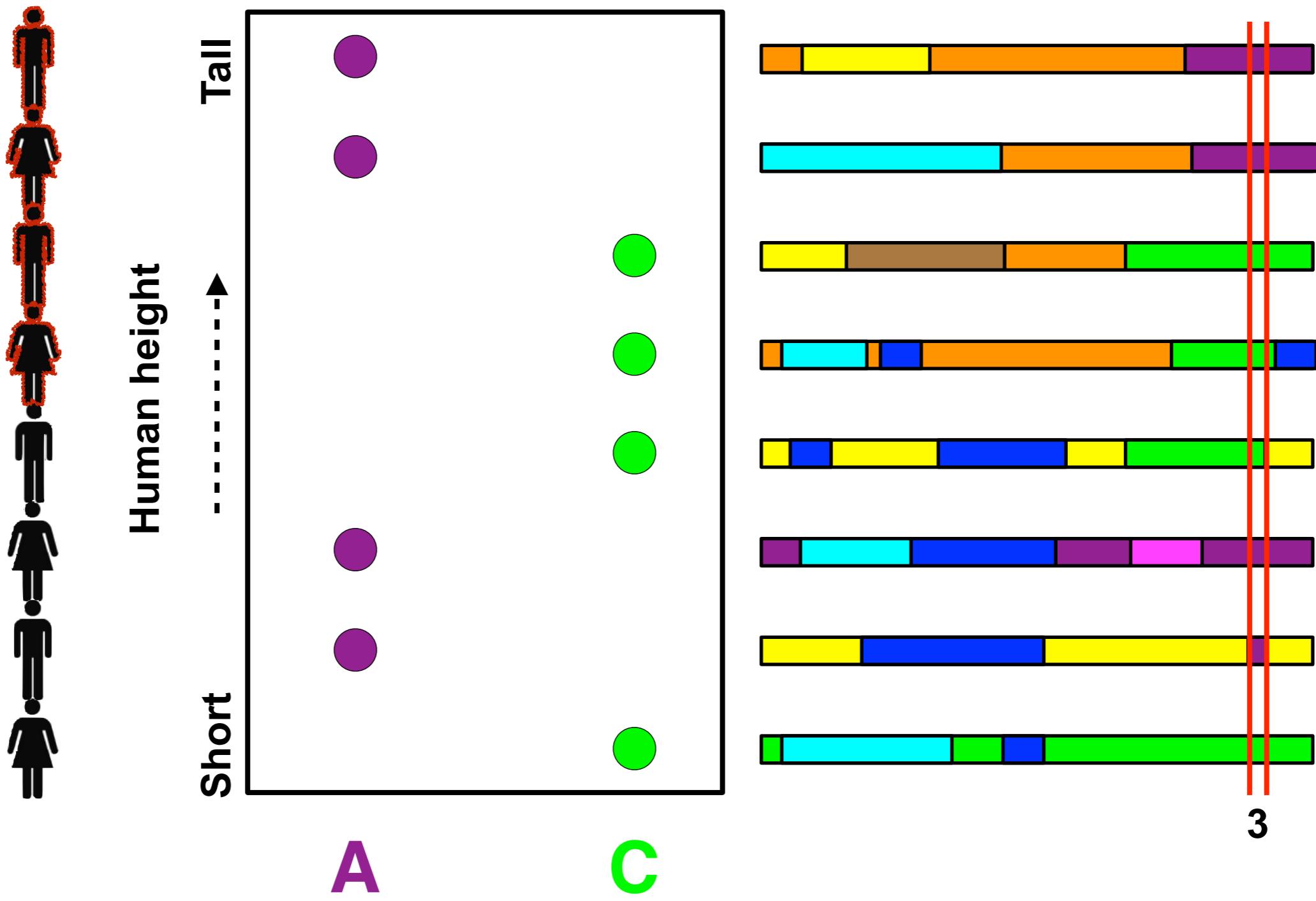
# Association mapping: Correlating genotype with phenotype



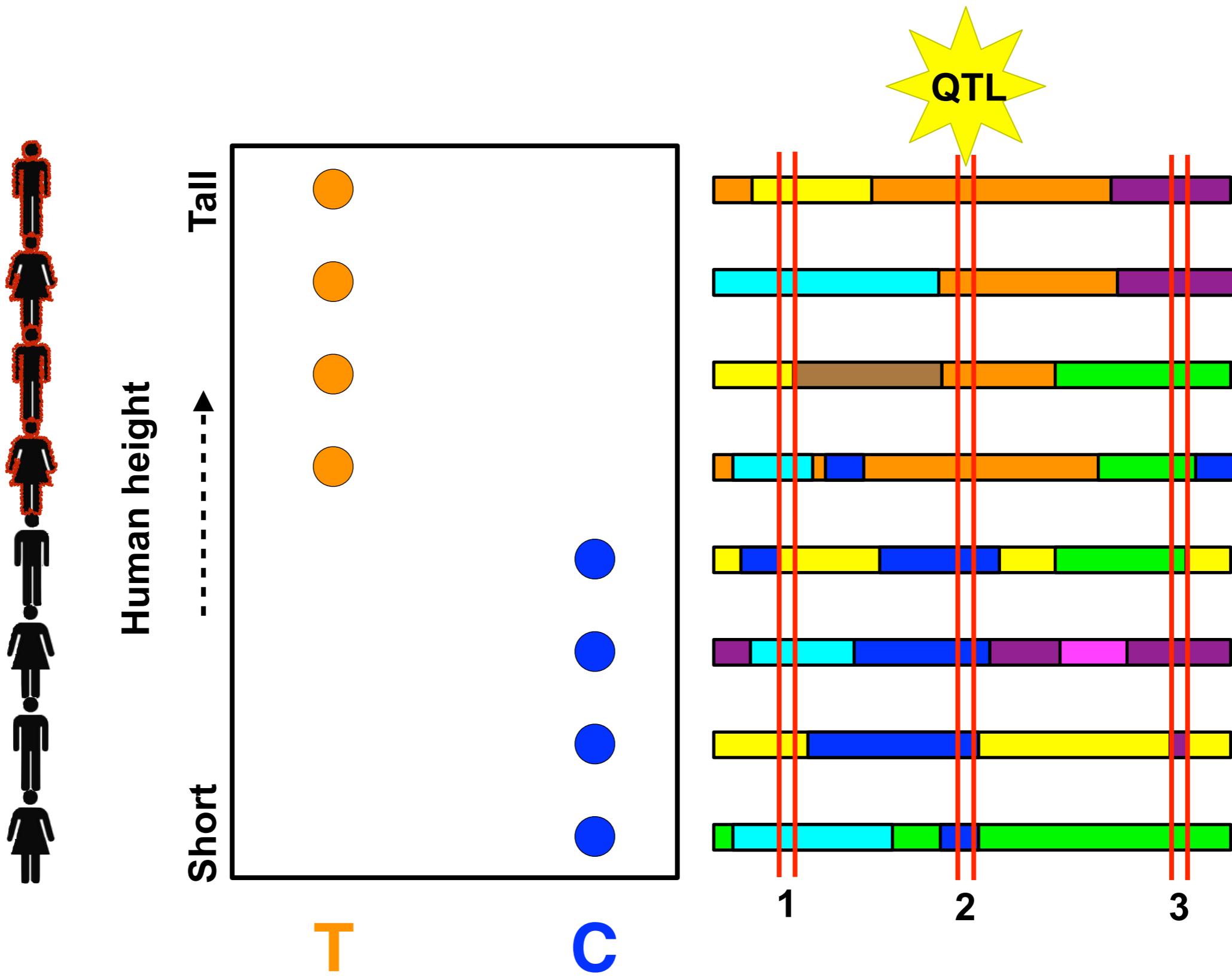
# Association mapping: Correlating genotype with phenotype



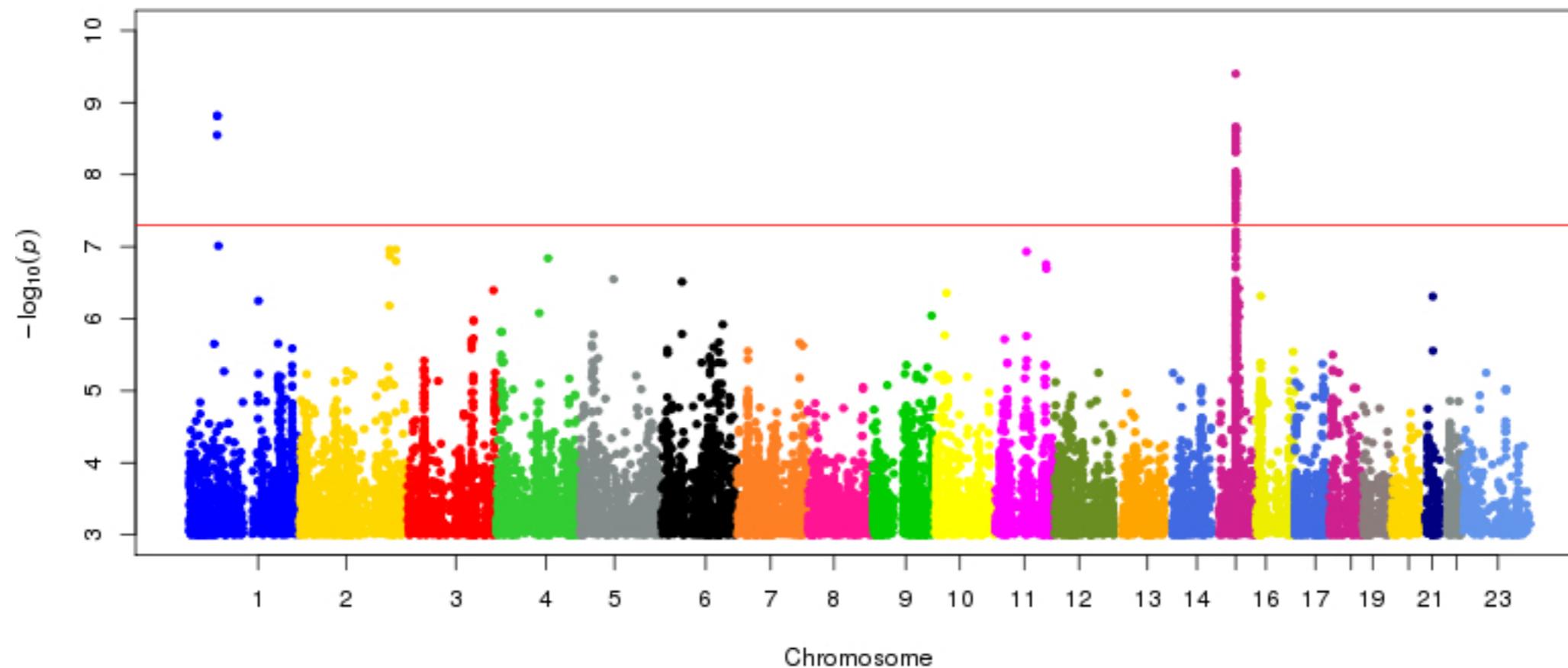
# Association mapping: Correlating genotype with phenotype



# Association mapping: Correlating genotype with phenotype

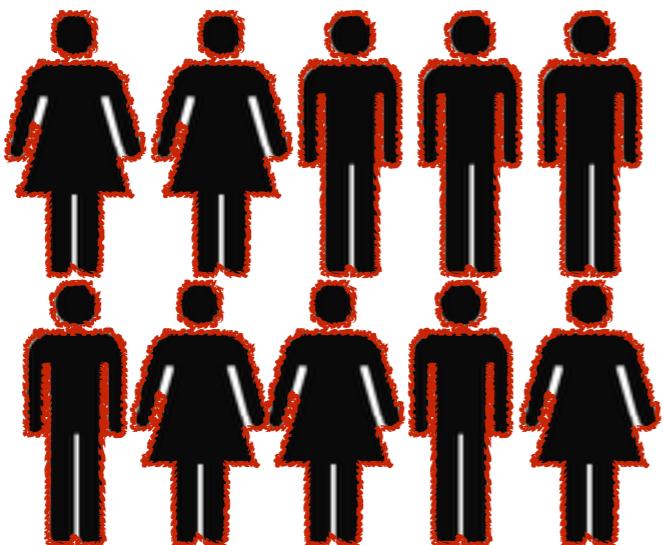


# An example Manhattan plot of GWA mapping results

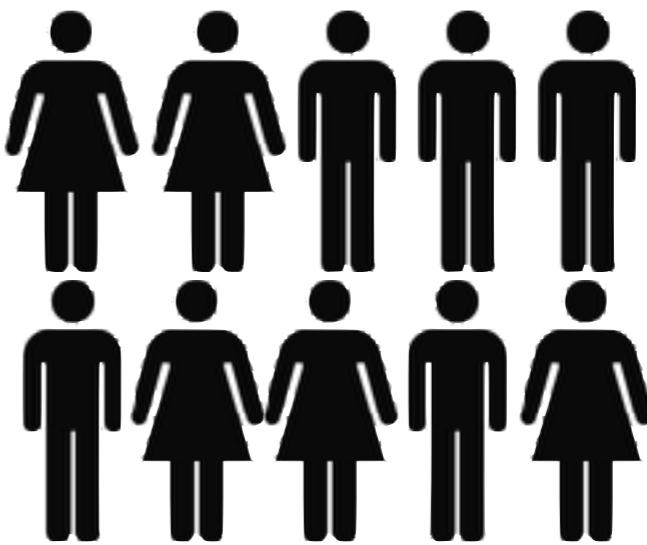


Styrkarsdottir *et al.* Nature 2014

# GWAS calculation



4000 Cases



6000 Controls

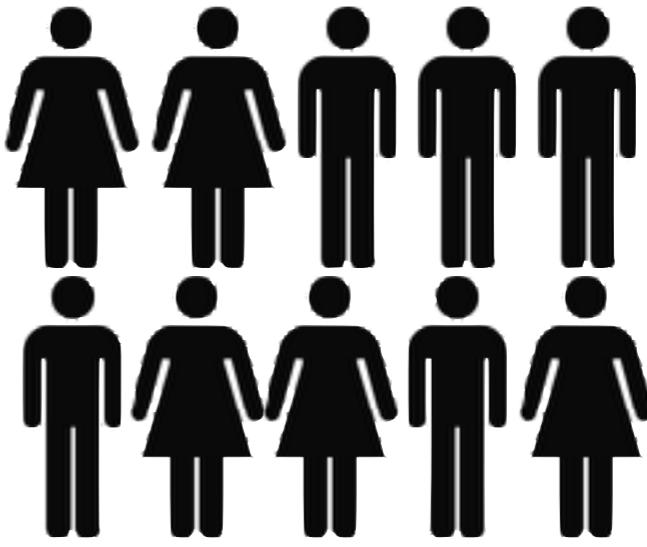
SNV1  
(G or A) 4000 of 8000 (50%)

5000 of 12000 (42%)

	Cases	Controls
G	4000	5000
A	4000	7000

Observed                      Expected

# GWAS calculation



SNV1  
(G or A) 4000 of 8000 (50%)

5000 of 12000 (42%)

	Cases	Controls
G	4000	5000
A	4000	7000

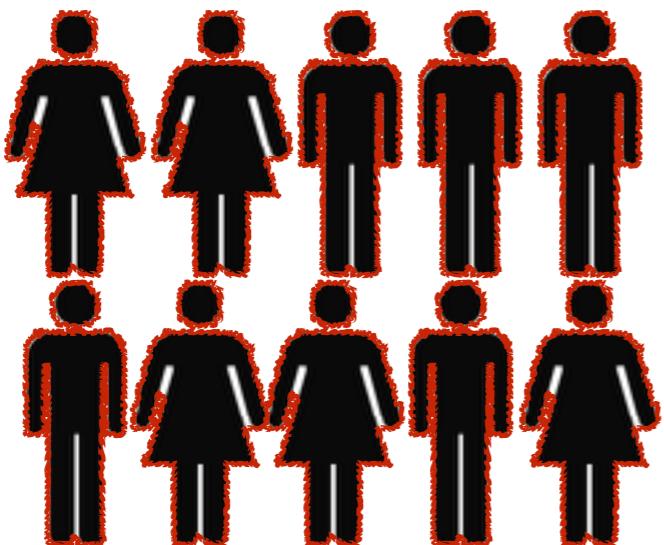
Observed

Expected

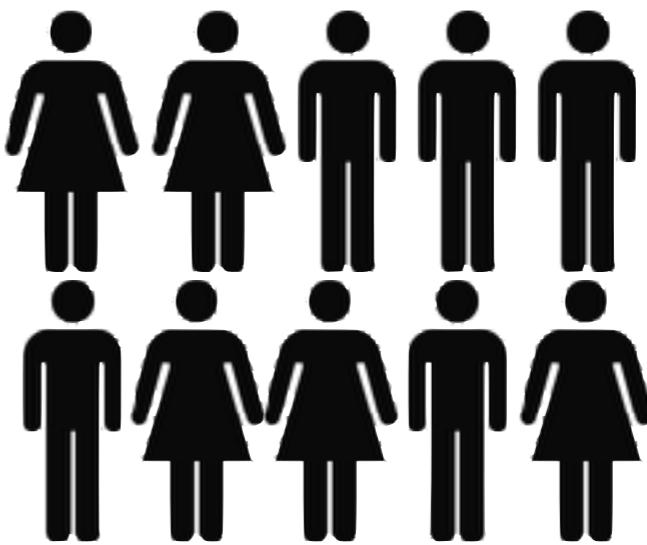
Pearson's chi-squared test  
with one degree of freedom

67.0038 or p-value of 2.71e-16

# GWAS calculation



4000 Cases



6000 Controls

SNV1  
(G or A) 4000 of 8000 (50%)

5000 of 12000 (42%)

SNV2  
(T or C) 3200 of 8000 (40%)

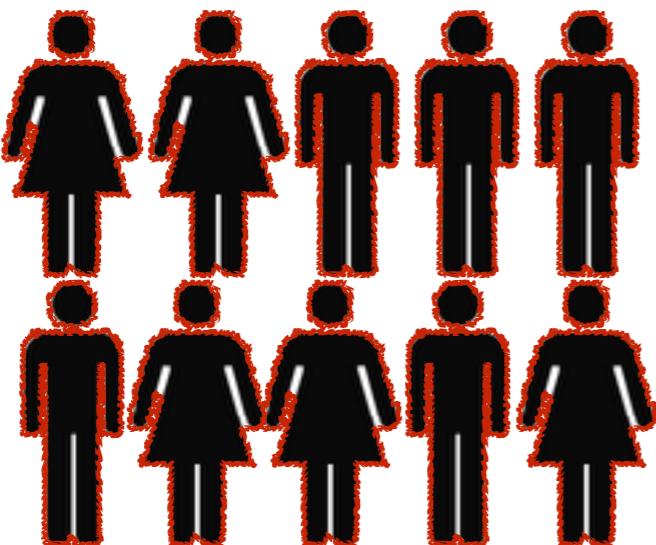
4600 of 12000 (38%)

	Cases	Controls
T	3200	4600
C	4800	7400

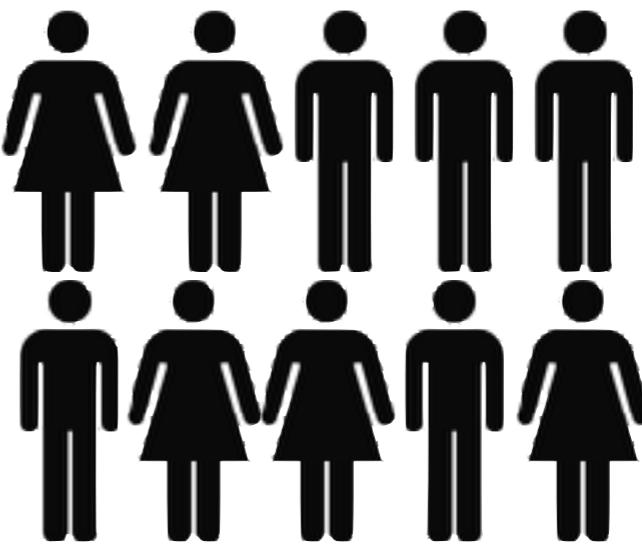
Observed

Expected

# GWAS calculation



4000 Cases



6000 Controls

SNV1  
(G or A) 4000 of 8000 (50%)

5000 of 12000 (42%)

SNV2  
(T or C) 3200 of 8000 (40%)

4600 of 12000 (38%)

	Cases	Controls
T	3200	4600
C	4800	7400

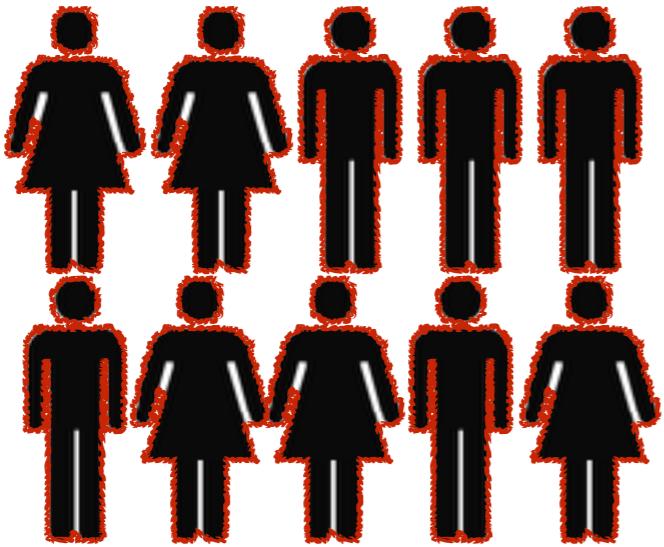
Observed

Expected

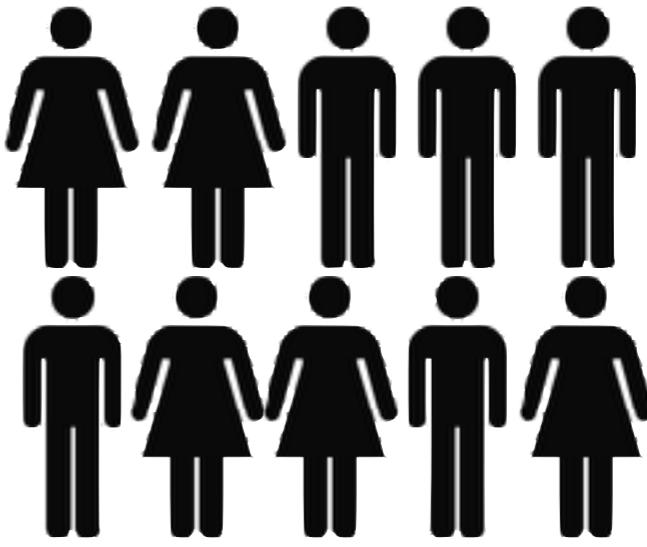
Pearson's chi-squared test  
with one degree of freedom

2.7327 or p-value of 0.09831

# GWAS calculation



4000 Cases



6000 Controls

SNV1  
(G or A) 4000 of 8000 (50%)

5000 of 12000 (42%)

SNV2  
(T or C) 3200 of 8000 (40%)

4600 of 12000 (38%)

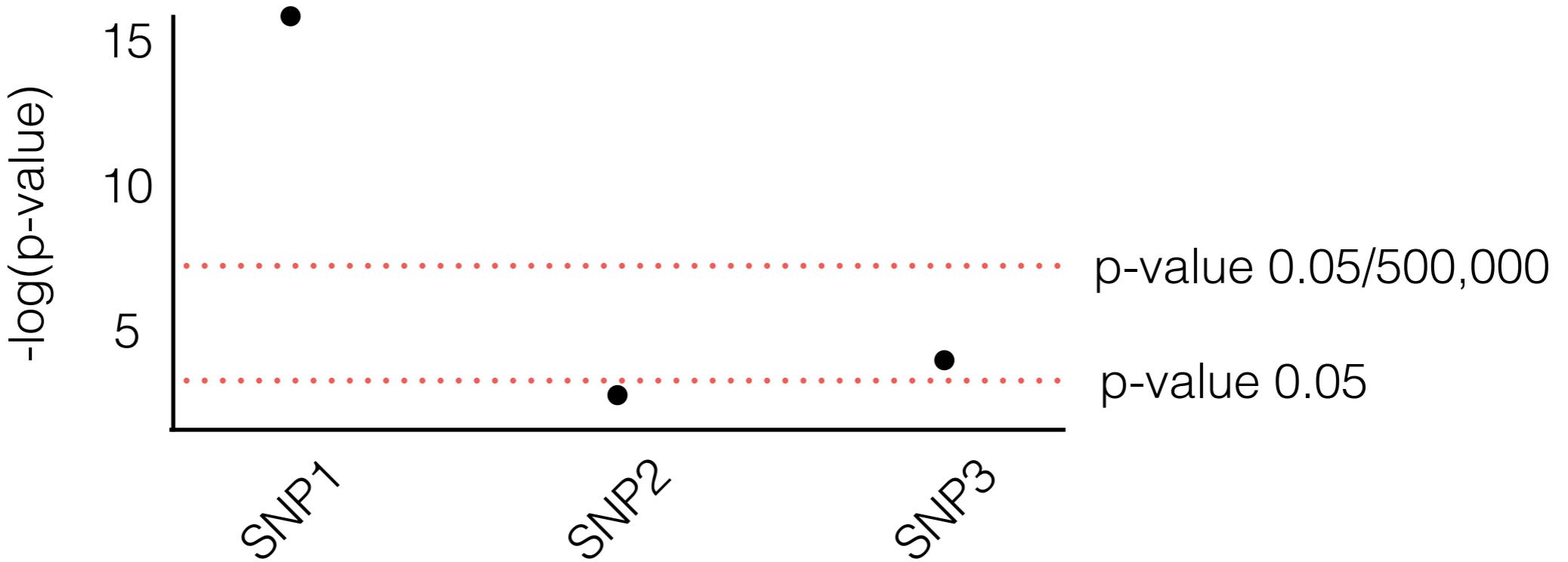
SNV3  
(C or A) 3600 of 8000 (45%)

5000 of 12000 (40%)

	Cases	Controls
T	3600	5000
C	4400	7000

10.7443 or p-value of 0.001046

# GWAS results



500,000 SNVs across the whole genome



500,000 tests with a p-value of 0.05 means  
that we would reject the null hypothesis  
for 25,000 SNVs by chance

Bonferroni correction  $0.05 / 500,000$  or  $1e-7$

# Three possibilities for the results of any GWA mapping

1. Marker is the *functional variant*
2. Marker is in *linkage disequilibrium* with functional variant
3. Marker is associated because of *population relatedness*  
*(population structure)*

# GWA mapping within groups and replication



GWA mapping works best within a related population

The mapping *might* be replicated in different populations