

# Hardy Weinberg Equilibrium Linkage Disequilibrium

Section 4

# Learning objectives:

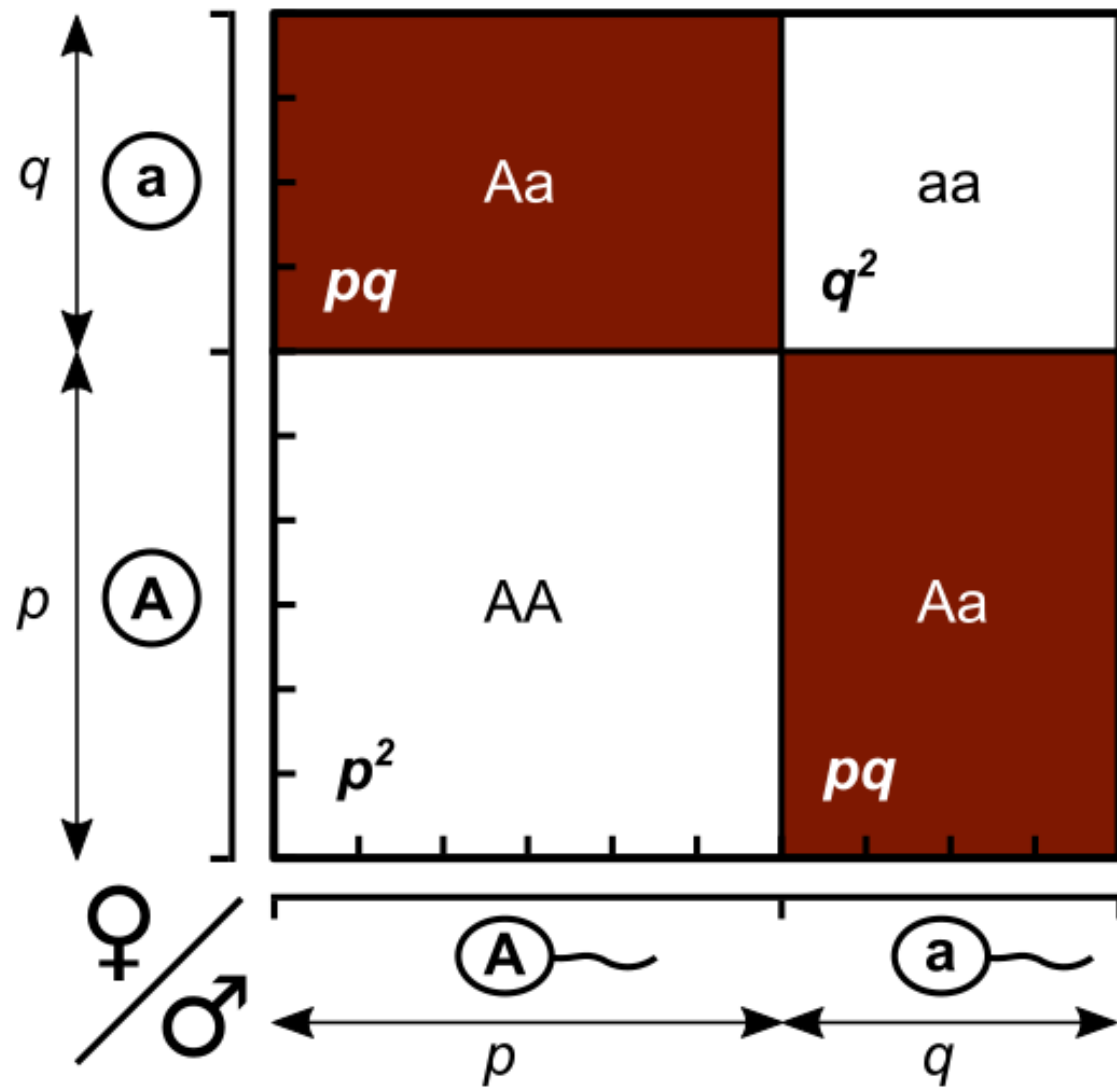
- Evaluate Hardy-Weinberg equilibrium and describe its importance for genetic epidemiology studies.
- Interpret linkage disequilibrium and describe how it helps and hinders genetic epidemiology studies.

# Introduction to genetics: single mating pair and offspring

	<b>A</b>	<b>a</b>
<b>A</b>	<b>AA</b>	<b>Aa</b>
<b>a</b>	<b>Aa</b>	<b>aa</b>

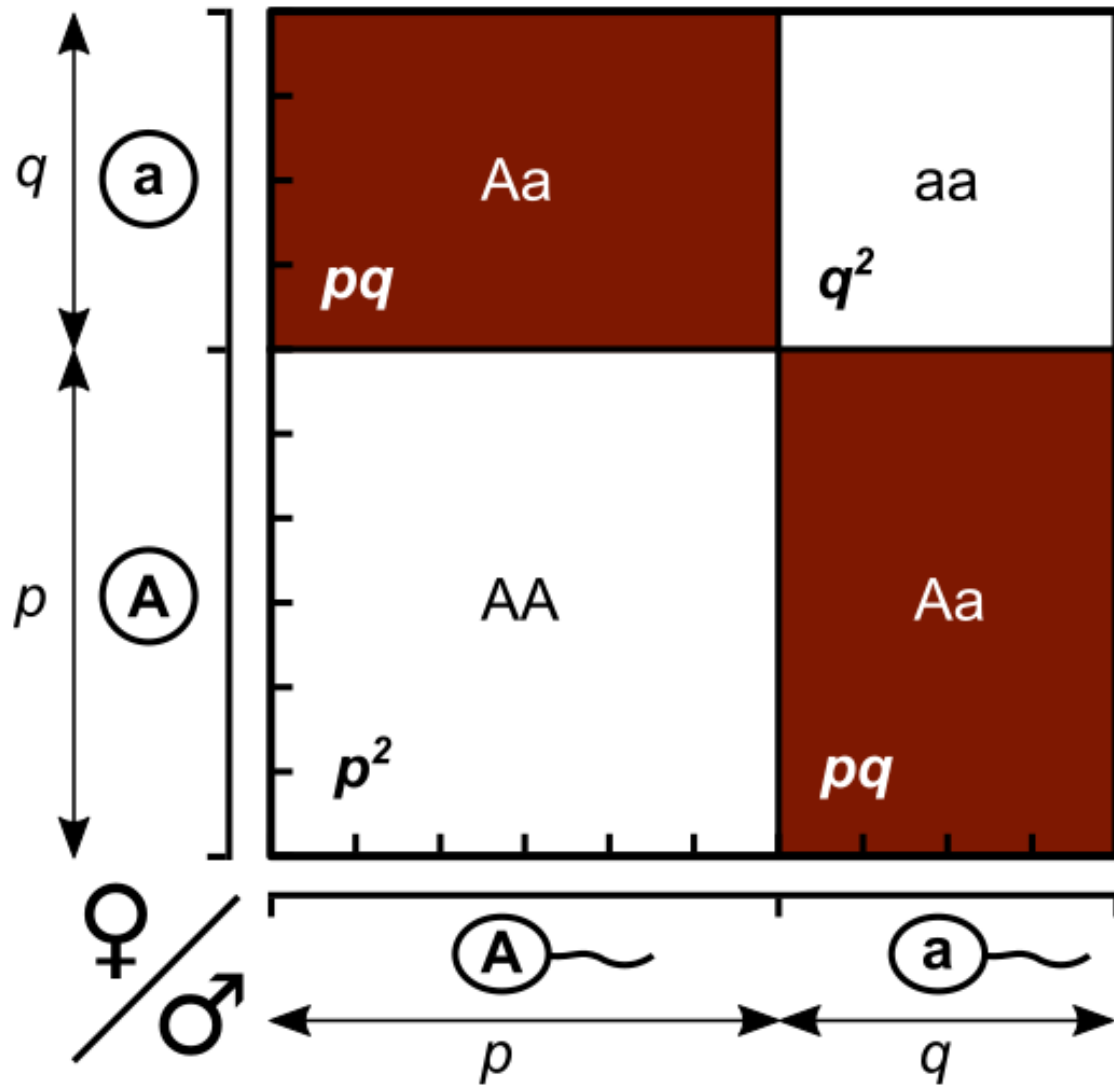
$$\frac{1}{4} (AA) + \frac{2}{4} (Aa) + \frac{1}{4} (aa)$$

# Population scale expected genotype combinations



Hardy Weinberg Equilibrium is a probabilistic relationship between ALLELE frequencies and GENOTYPE frequencies

# Population scale expected genotype combinations



**Based on random mating:**

Probability grab an “a” from the female is  $q$

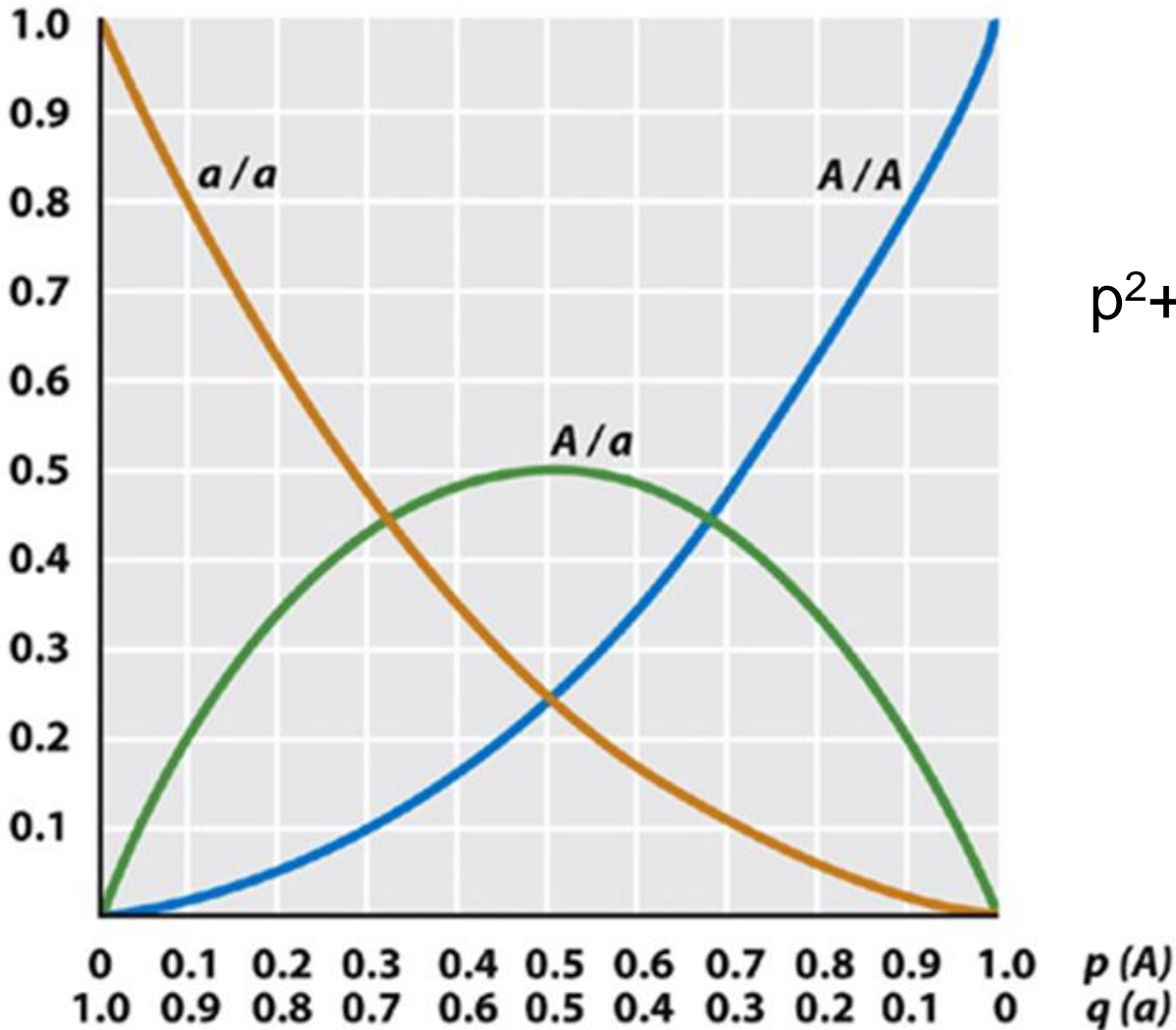
Probability grab an “a” from the male is  $q$

So, probability grab an “a” from the female and an “a” from the male is  $q * q$

# The Hardy-Weinberg principle

- Assume that...
  - Population is large (coin flip likelihoods)
  - Mating is random (selective genotype matches)
  - No immigration or emigration
  - Natural selection is not occurring (all genotypes have an equal chance of surviving and reproducing)
  - No mutations
- If these assumptions are true, we say that a population is not evolving (allele frequencies stay the same) and in **Hardy-Weinberg Equilibrium**

# The Hardy-Weinberg principle



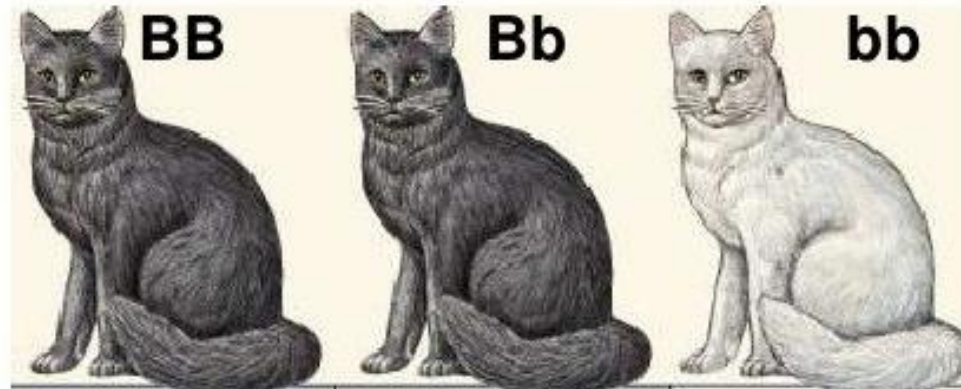
$$p+q=1 \text{ (allele frequencies)}$$

$$p^2+2qp+q^2=1 \text{ (genotype frequencies)}$$

# HWE example

- Assume 100 cats (200 alleles) with alleles B and b. B allele is dominant and results in black coloring. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?

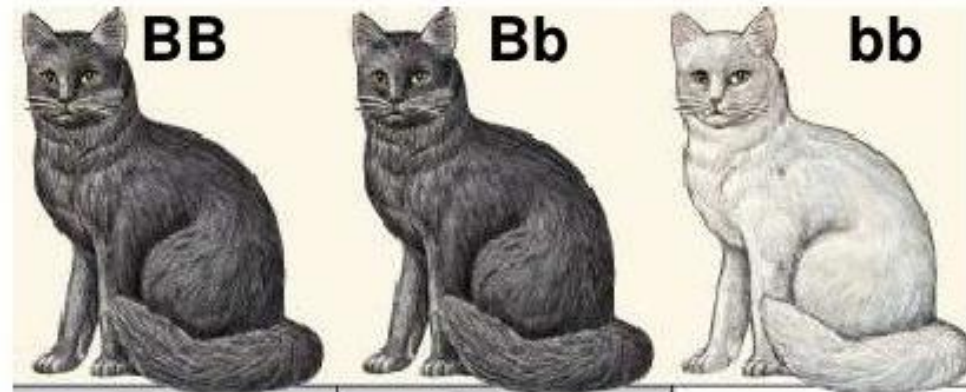
- $p+q=1$
- $p^2+2pq+q^2=1$





# HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2pq+q^2=1$
- $q^2=0.16$
- $q=0.4, p=0.6$
- $p^2=0.36$
- $2pq=2 \times 0.6 \times 0.4=0.48$
- 16 white cats and 84 black cats (28 Bb, 36 BB)



# Assessing deviation from HWE

$\chi^2$ -goodness-of-fit (GOF) tests with 1 degree of freedom

Sum of observed minus expected

- $O$  = observed counts,  $E$  = expected counts, sum across genotypes

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

Compare to chi-square distribution to determine whether the deviance is significant.

# Deviation from Hardy Weinberg?

Check chi-square distribution with 1-degree of freedom:

Degrees of Freedom	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086

# Reasons to defy Hardy Weinberg equilibrium

- True selective pressures
- Genotyping error! (most common reason)
- Undetected population substructure (differences in ancestry)
- Non-random procreation

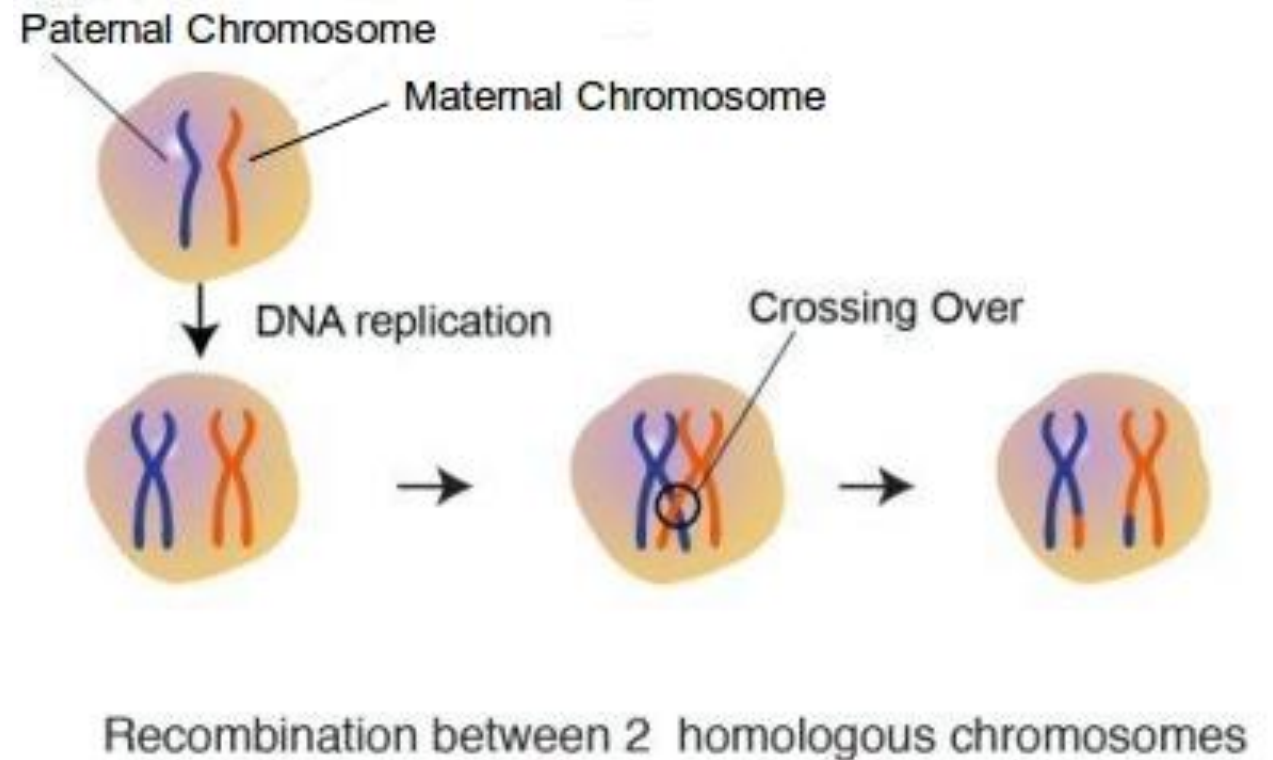
**Many statistical tests rely on SNPs being in Hardy Weinberg equilibrium, so we test this chi-square test on every SNP in a study.**

Zoom breakout Q1

Another important equilibrium...  
Linkage Disequilibrium

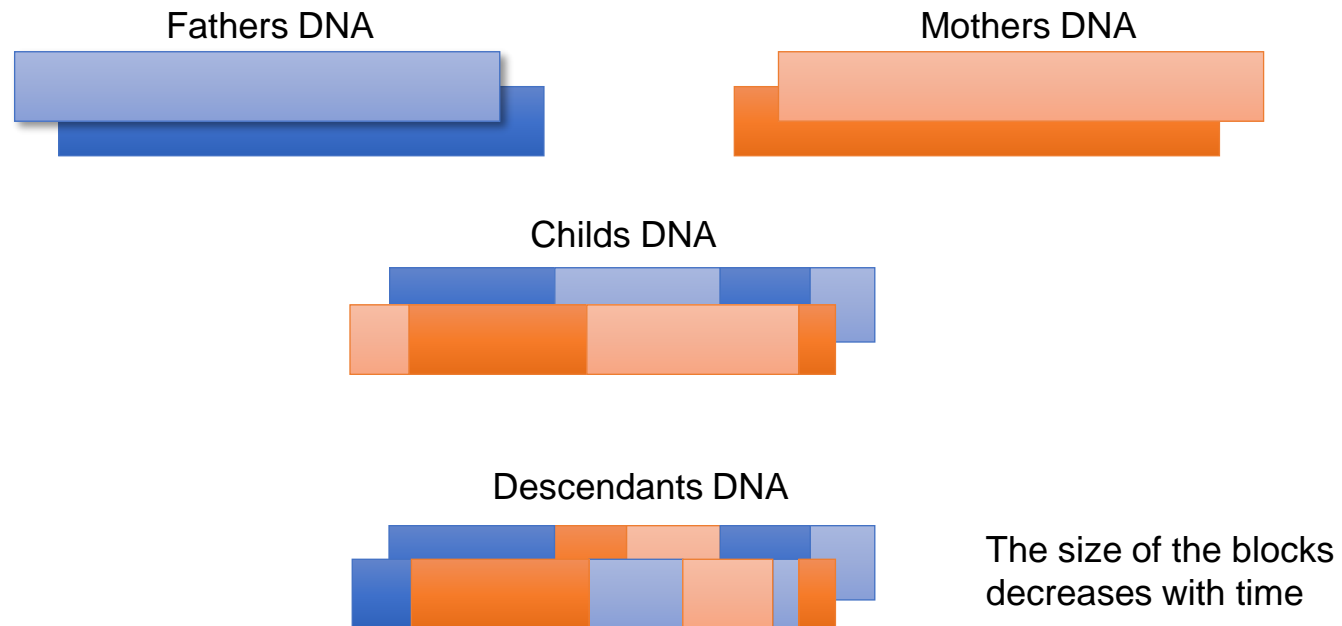
# Recombination

- Alleles on the same chromosome are inherited together unless *recombination (crossing over)* occurs
- The probability of recombination between two alleles increases with the distance between them



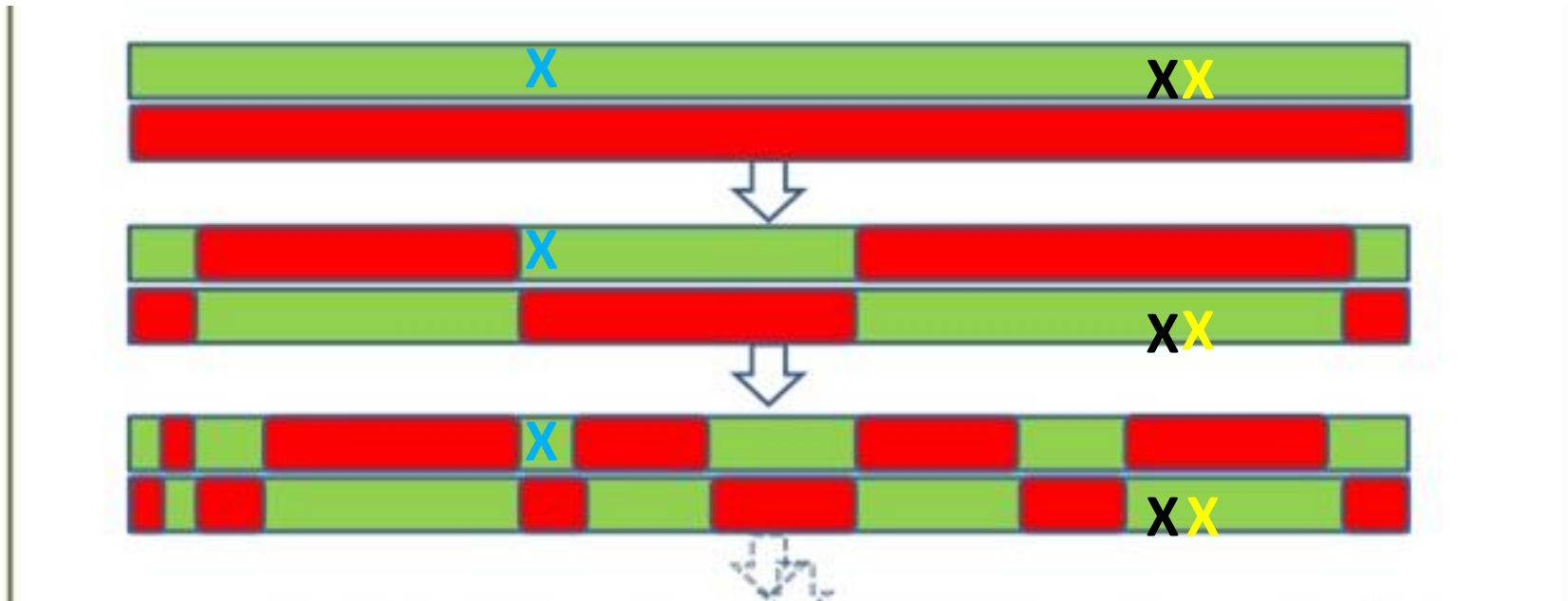
# DNA inheritance breaks DNA into blocks

We inherit “blocks” of the genome from our parents (and not independent base-pairs)





# Recombination and linkage disequilibrium



How well can you tell me if the **X** is present if your chromosome also has the **X**?  
How about if your chromosome has the **XX**?

# Linkage Disequilibrium

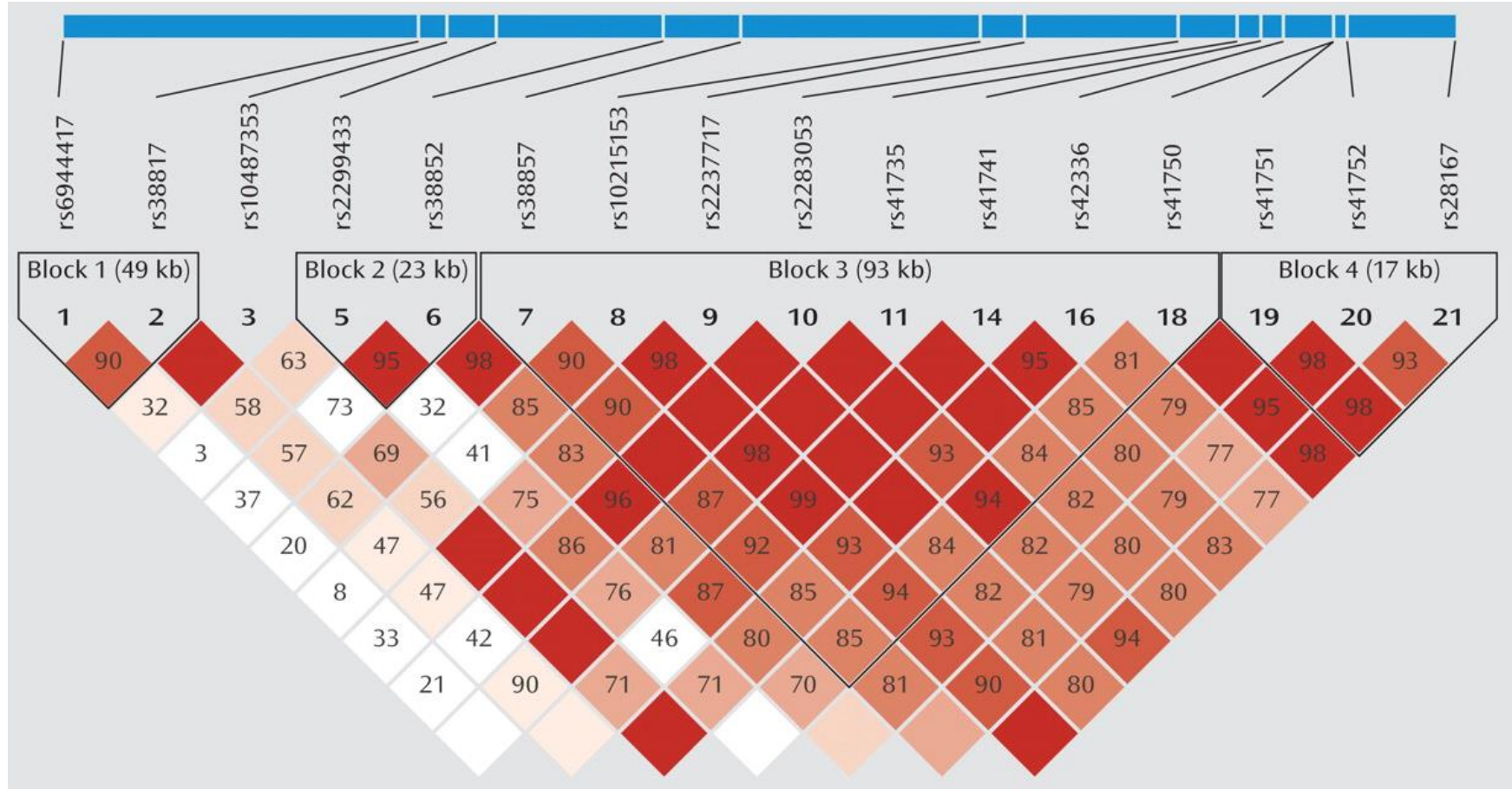
- Non-random association of alleles at different loci within a population
  - Alleles are found together more or less often than you would expect by chance if all alleles randomly and independently segregated during meiosis.
- Scores range from 0 (not in LD) to 1 (in complete LD).
- Two ways to measure LD:  $D'$  and  $r^2$

\*  $r^2$  depends on allele frequencies.  $D'$  is more difficult to compare across different allele frequencies.  $r^2$  is used more often.

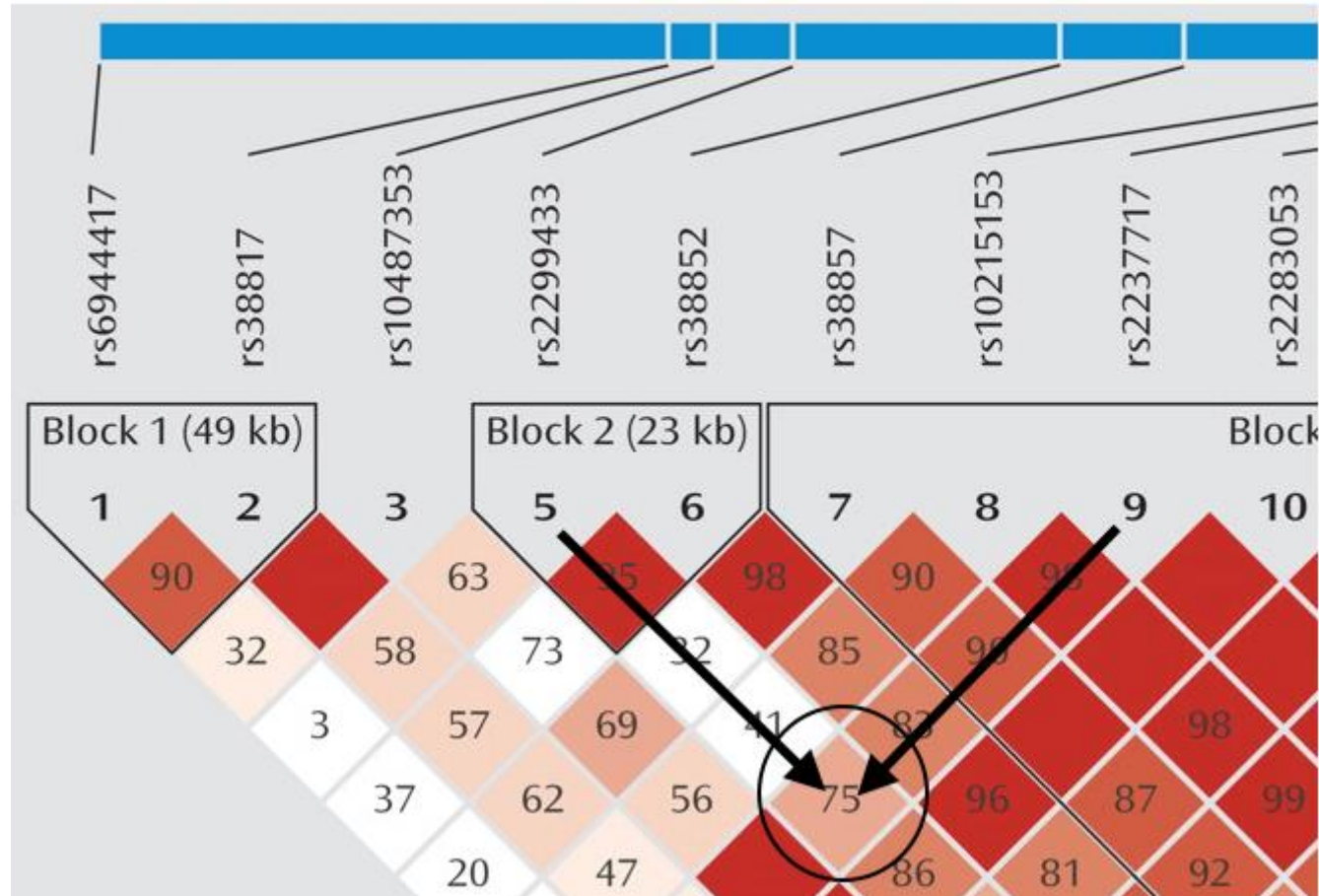
# Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
- Natural selection
  - Haplotypes that carry favorable mutations increase in frequency

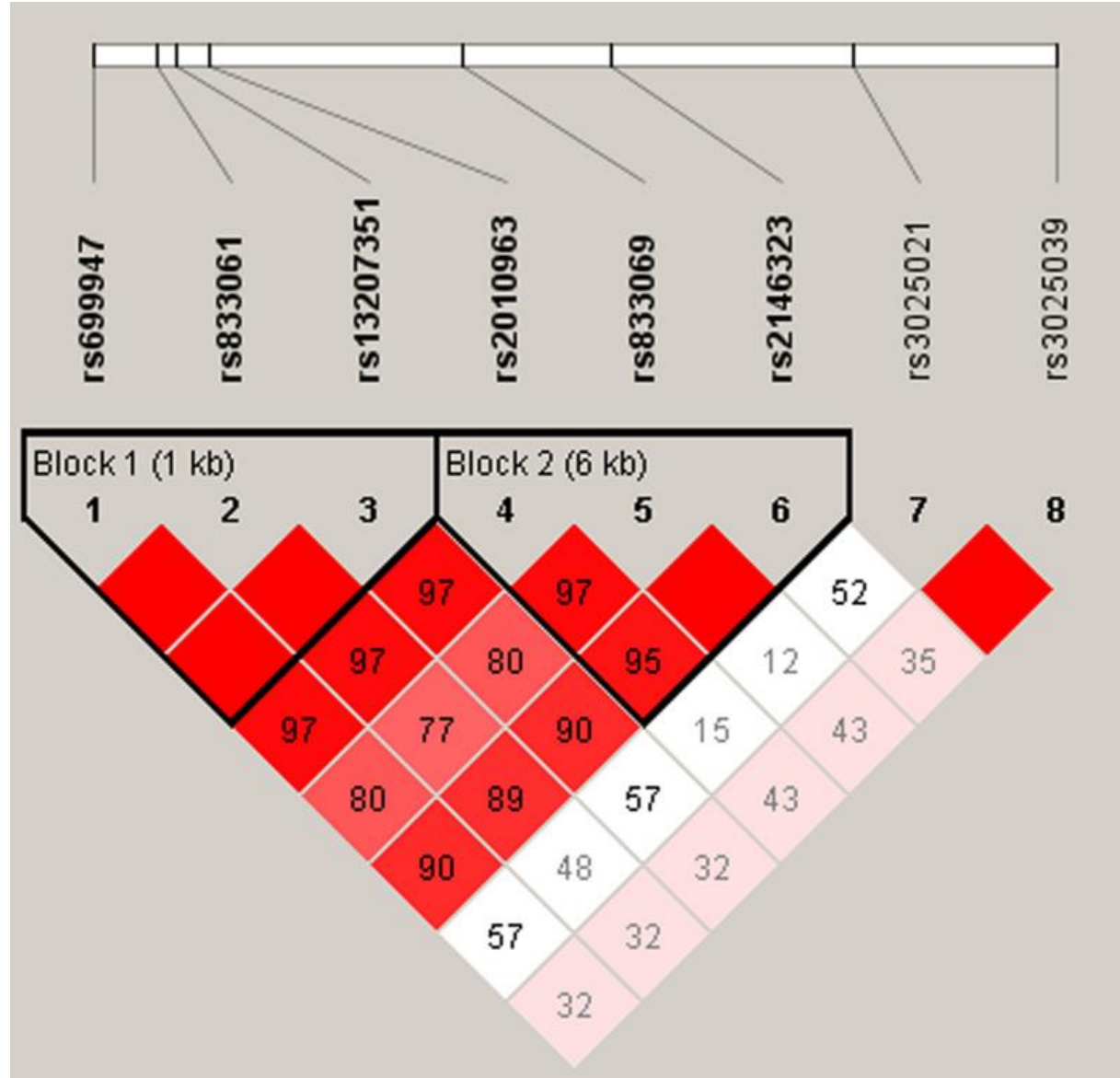
# Reading an LD map



# Reading an LD map



# What is the correlation between rs13207351 and rs3025021?





## Welcome to LDlink

LDlink is a suite of web-based applications designed to easily and efficiently interrogate linkage disequilibrium in population groups. Each included application is specialized for querying and displaying unique aspects of linkage disequilibrium.

### LDassoc

Interactively visualize association p-value results and linkage disequilibrium patterns for a genomic region of interest.

### LDhap

Calculate population specific haplotype frequencies of all haplotypes observed for a list of query variants.

### LDmatrix

Create an interactive heatmap matrix of pairwise linkage disequilibrium statistics.

### LDpair

Investigate correlated alleles for a pair of variants in high LD.

### LDpop

Investigate allele frequencies and linkage disequilibrium patterns across 1000G populations.

### LDproxy

Interactively explore proxy and putatively functional variants for a query variant.

### LDtrait

Search if a list of variants (or variants in LD with those variants) have previously been associated with a trait or disease.

### SNPchip

Find commercial genotyping platforms for variants.

### SNPclip

Prune a list of variants by linkage disequilibrium.

Zoom breakout Q2 and Q3



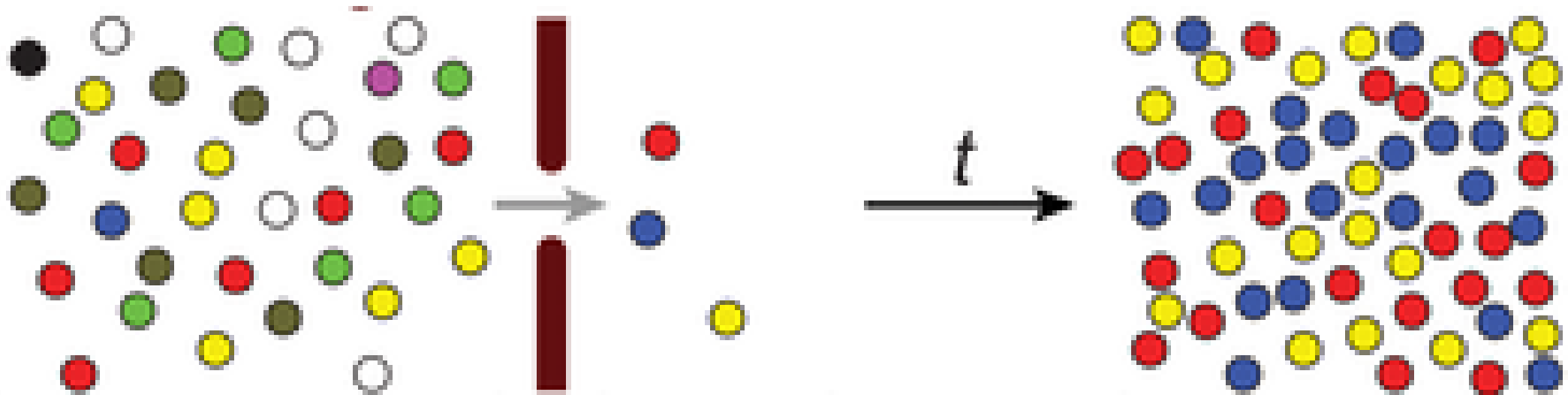
# Benefit of LD in genetic studies?

- We don't need to look at every location! Because of LD patterns, knowing an allele at one position can tell us about alleles at other positions. We can save \$ and time by only looking at a subset.

# Considerations of LD in genetic studies?

- We might not be able to tell if the SNP we are looking at actually **CAUSES** the phenotype, or if it is just in tight LD with something that does.
- Different populations have different LD patterns.

# Founding populations reset LD



# How does LD influence our study power?

- If a SNP C and causal SNP G are in LD with  $r^2$ , then a study with N cases and controls which measures C (but not G) will have the same power to detect an association between C and disease as a study with  $r^2 N$  cases and controls that directly measured G.
- $r^2 N$  is the “effective sample size”
  - If the  $r^2$  between your measured SNP C and causal SNP G is 0.5 you need to double your sample size to obtain the same power as if you had measured (genotyped) G directly.

# Summary

- Hardy Weinberg disequilibrium tests can indicate underlying population structure or selective pressure.
- Linkage disequilibrium is the nonrandom assortment of alleles in a genomic region.

# Calculation of LD

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other.  
(This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB} = p_A p_B$	Ab $p_{Ab} = p_A p_b$	$p_A$
	aB $p_{aB} = p_a p_B$	ab $p_{ab} = p_a p_b$	$p_a$
	$p_B$	$p_b$	1

# Calculation of LD

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other  
(This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB} = p_A p_B$	Ab $p_{Ab} = p_A p_b$	$p_A$
	aB $p_{aB} = p_a p_B$	ab $p_{ab} = p_a p_b$	$p_a$
	$p_B$	$p_b$	1

What do we actually see?

We can infer LD as the deviation of observed haplotype frequency from its corresponding allele frequencies if SNP1 and SNP2 are independent of each other

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_A p_B + D$	Ab $p_A p_b - D$	$p_A$
	aB $p_a p_B - D$	ab $p_a p_b + D$	$p_a$
	$p_B$	$p_b$	1

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

# Two measures of LD: $D'$ and $r^2$

$$D' = \frac{D}{D_{max}},$$

$$D_{max} = \begin{cases} \max\{-p_A p_B, -(1-p_A)(1-p_B)\}, & \text{when } D < 0 \\ \min\{p_A(1-p_B), (1-p_A)p_B\}, & \text{when } D > 0 \end{cases}$$

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$



# LD calculation exercise

SNPs rs6025 and rs4524 are both associated with venous thromboembolism (blood clot in a vein). The number of alleles for each SNP based on 503 individuals are displayed in the table below. Based on these numbers, calculate

- Frequencies of the four alleles (rs6025-C, rs6025-T, rs4524-G, rs4524-A)**
- Frequencies for the four haplotypes (C-G, C-A, T-G and T-A)**
- D' and  $r^2$  between the two SNPs.**

Distribution of alleles for rs6025 and rs4524 across 503 individuals.

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

Instead of  $D$ , we often express LD in terms of  $D'$  (normalized  $D$ ) or  $r^2$  (correlation coefficient)

$$D' = \frac{D}{D_{max}}$$

$$D_{max} = \begin{cases} \max\{-p_A p_B, -(1-p_A)(1-p_B)\}, & \text{when } D < 0 \\ \min\{p_A(1-p_B), (1-p_A)p_B\}, & \text{when } D > 0 \end{cases}$$

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

# LD calculation exercise

a) Frequencies of the four alleles (rs6025-C, rs6025-T, rs4524-G, rs4524-A)

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

rs6025/ rs4524	G	A		
C	255	739	C=994	pC=0.988
T	0	12	T=12	pT=0.012
	G=255	A=751	1006	1
	pG=0.253	pA=0.747	1	

# LD calculation exercise

b) Frequencies for the four haplotypes (C-G, C-A, T-G and T-A)

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

rs6025/rs4524	G	A
C	$p_{CG} = 255/1006 = 0.253$	$p_{CA} = 739/1006 = 0.735$
T	$p_{TG} = 0$	$p_{TA} = 12/1006 = 0.0119$

# LD calculation exercise

**c)  $D'$  and  $r^2$  between the two SNPs.**

$$\begin{aligned} D &= p_{CG} \cdot p_{TA} - p_{CA} \cdot p_{TG} \\ &= 0.253 \cdot 0.0119 - 0.735 \cdot 0 \\ &= 0.0030 \end{aligned}$$

$$\begin{aligned} D' &= D / D_{\max} \\ &= 0.003 / \min\{0.253 \cdot (1 - 0.988), (1 - 0.253) \cdot 0.998\} \\ &= 0.003 / \min\{0.003, 0.746\} \\ &= 0.003 / 0.003 \\ &= 1 \end{aligned}$$

$$\begin{aligned} r^2 &= D^2 / (p_{rs6025-C} \cdot p_{rs6025-T} \cdot p_{rs4524-G} \cdot p_{rs4524-A}) \\ &= 0.003^2 / (0.988 \cdot 0.012 \cdot 0.253 \cdot 0.747) \\ &= 9.217 \times 10^{-6} / 0.0022 \\ &= 0.0041 \end{aligned}$$