1

Genetic variation

UAB Universitat Autònoma de Barcelona

Marta Puig ORCID: 0000-0002-4439-6372

Nice collection of colorful butterflies from Freepik.com

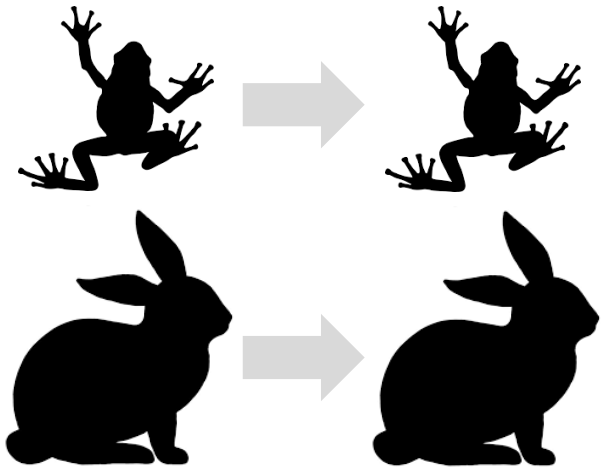
Genetic variation

Fixism *vs.* evolution

Fixism

Everything that exists, including all life forms, has not varied along time and everything is today exactly identical to how it was in the past and will be in the future

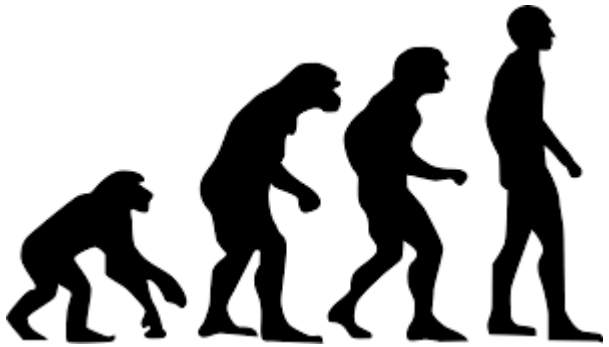
Variation = Inconvenience



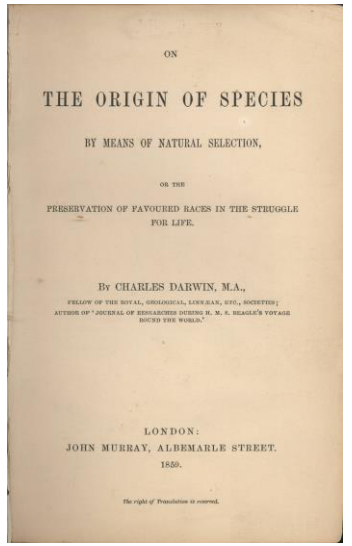
Evolution

Change in the heritable characteristics of populations over successive generations

Variation = Essential



Charles Darwin
1809 – 1882


ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

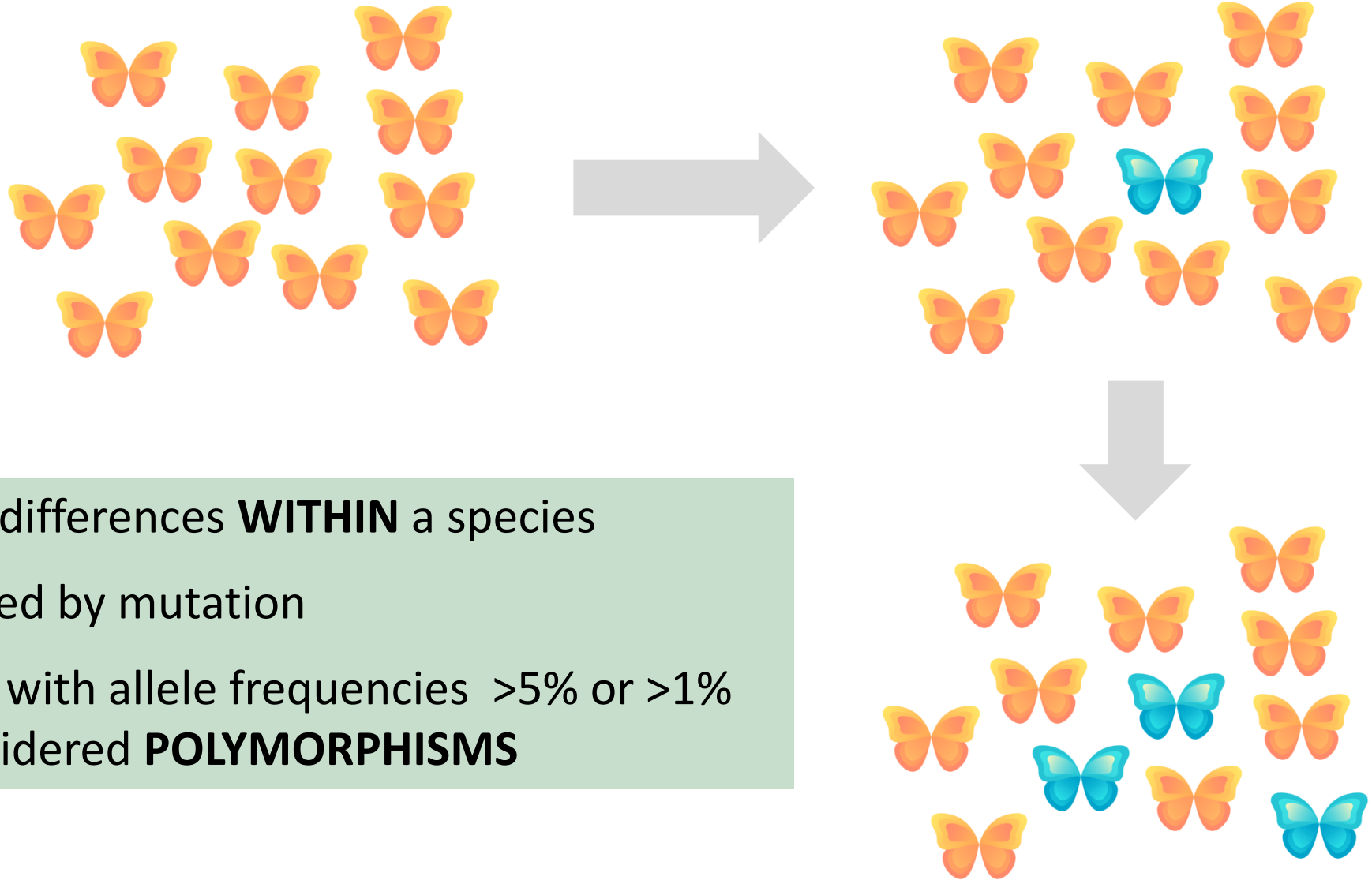
LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

1859

Evolution

Evolution does not take place in individuals, but in populations, which change over time

Polymorphism



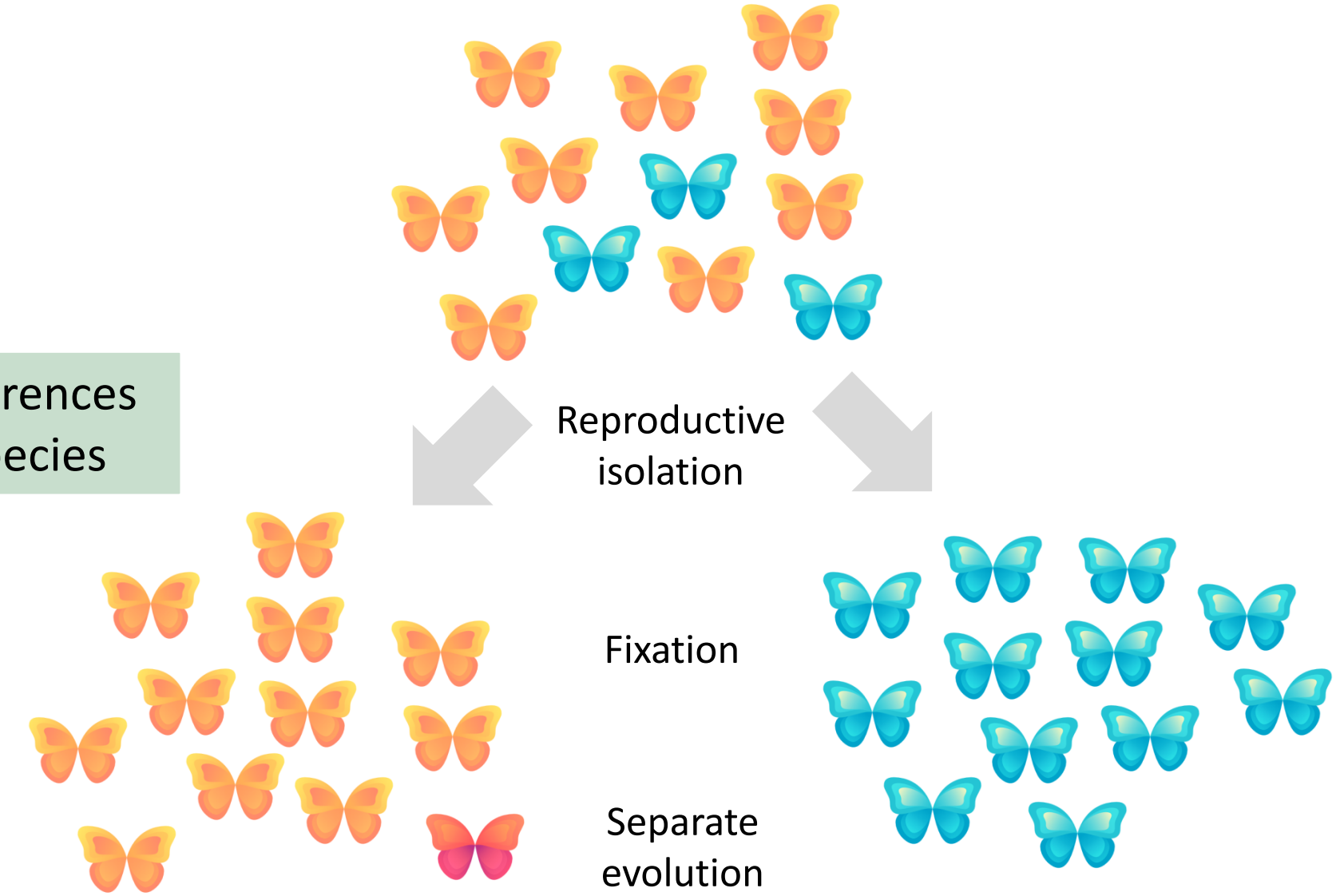
Genetic differences **WITHIN** a species

Generated by mutation

Variants with allele frequencies >5% or >1% are considered **POLYMORPHISMS**

Divergence

Genetic differences **BETWEEN** species



Important concepts

Genotype/phenotype

Phenotype = morphological, biochemical, physiological or behavioral attributes of an individual

Genotype = set of alleles of an individual at one or many genes/locus/positions of the genome

Gene/allele

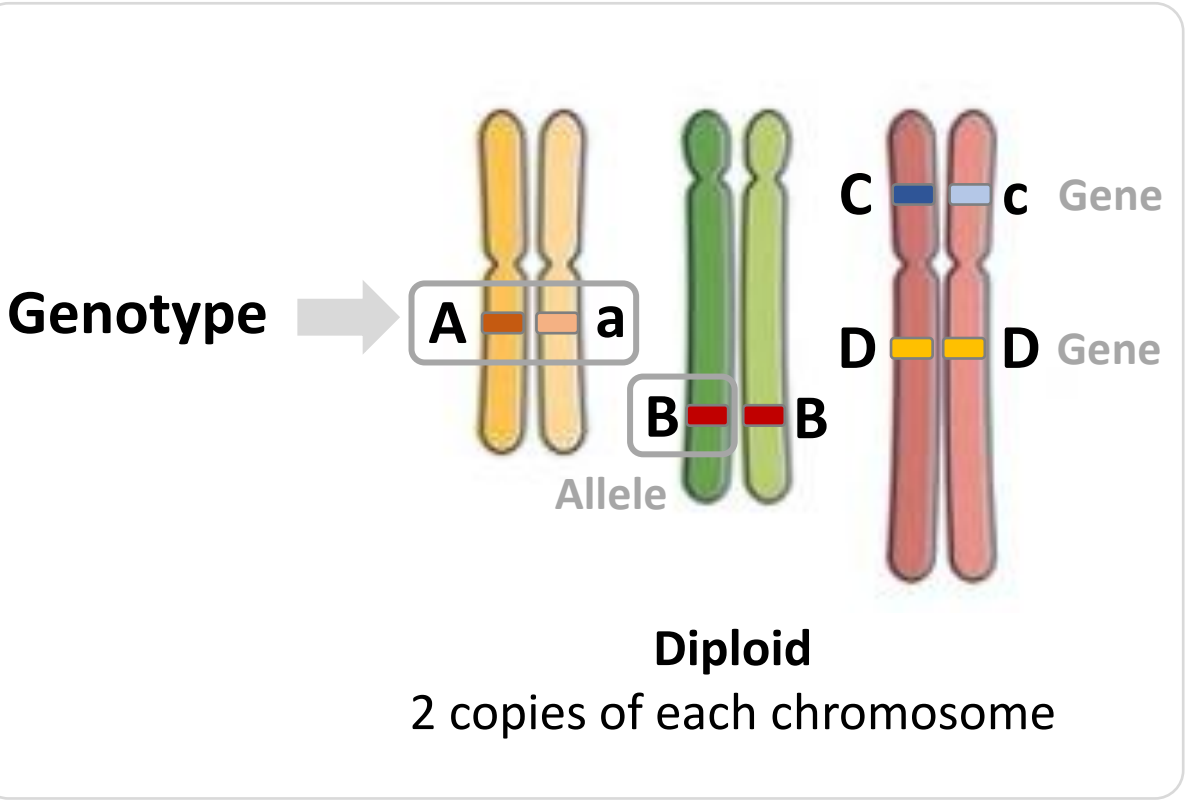
Gene = DNA sequence that codes for an RNA or protein

Allele = variant or alternative form of the DNA sequence at a given gene / copy of a gene in a diploid organism

Population

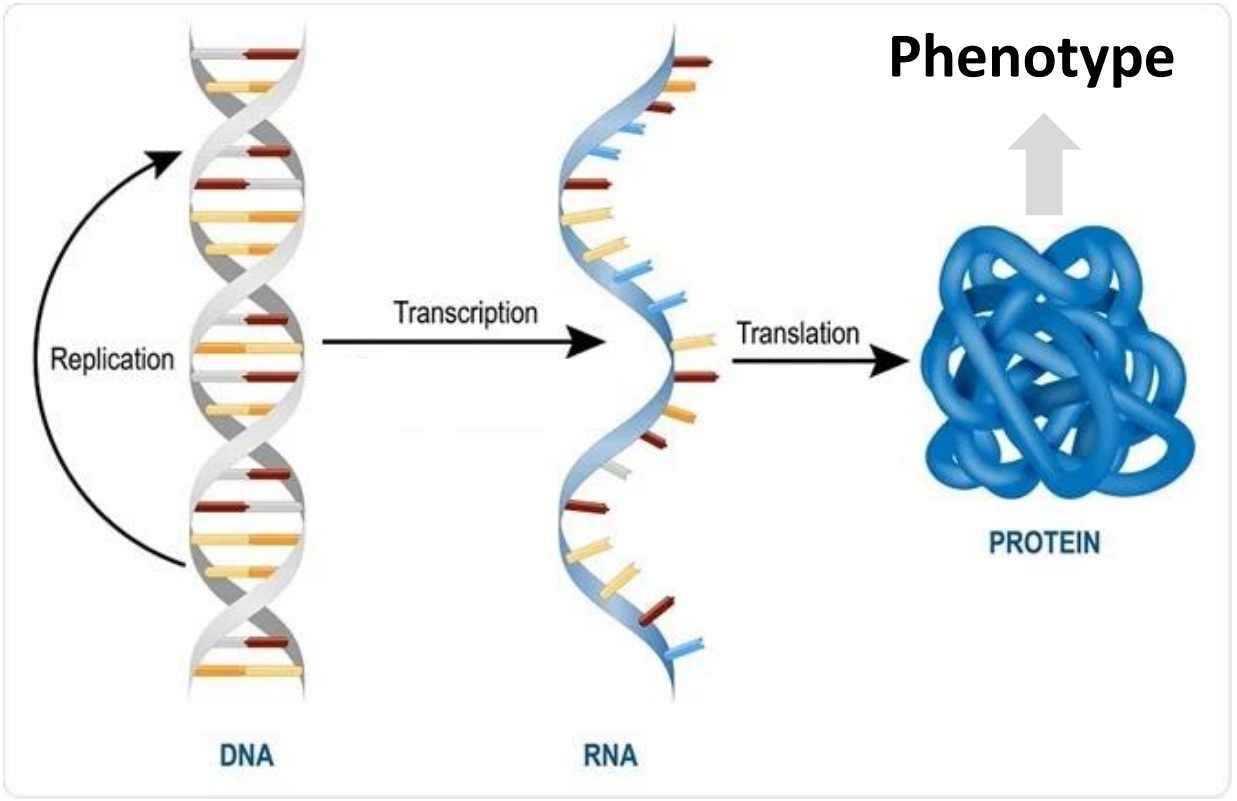
Population = not an entire species, but a group of individuals of same species living in a geographically restricted area so that any member can potentially mate with any other member

Important concepts

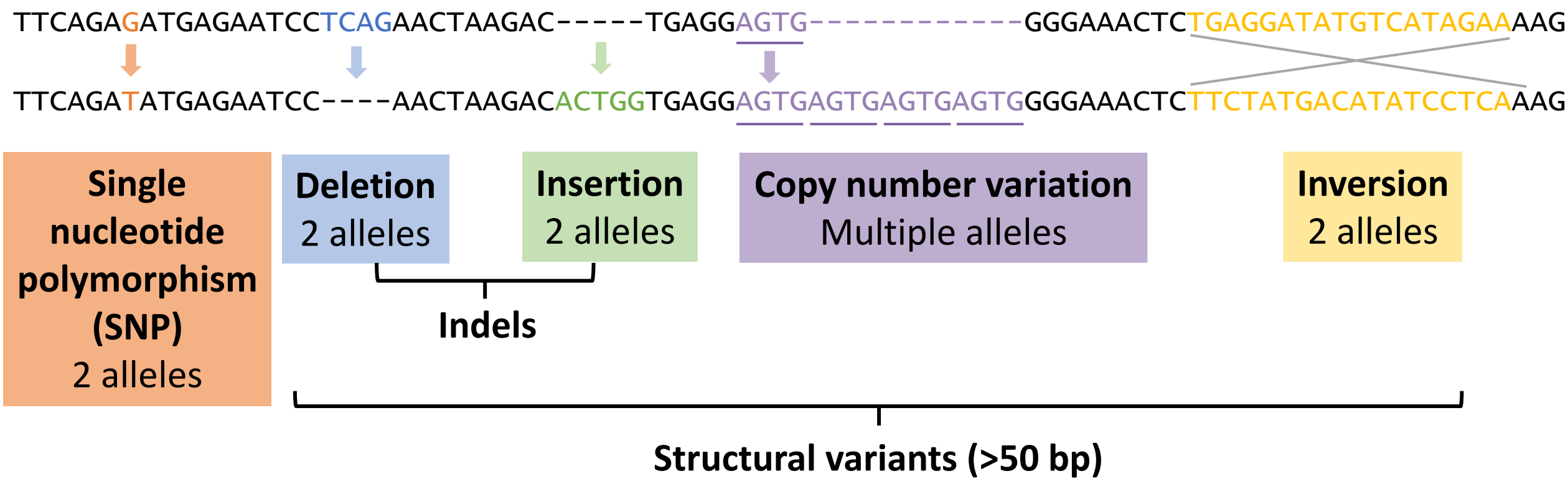


Diploid

2 copies of each chromosome



Types of variants



CNV: microsatellites or short tandem repeats

Microsatellites

Repeated sequences of 2-5 bp polymorphic in their length (multiallelic variants)
Also known as STRs (short tandem repeats)

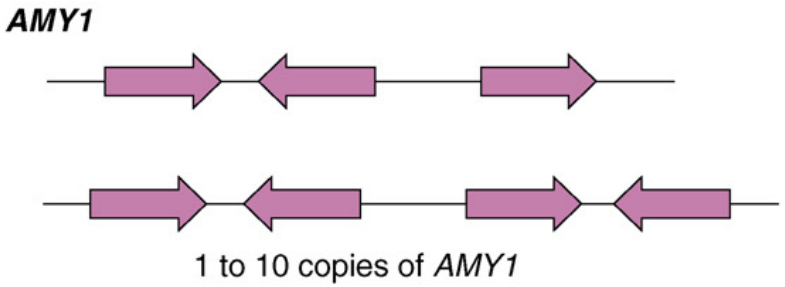
. . . TGAATGCGGAGTCTGCTTGAGATTCTCTC**CCCTTCCTCTCCCTCTATCCCTCCT**
CATGCTCTCTCTCATTATCTCATAAATAAATAAATAAATAAATAAATAAATAAATAA
ATAAA**TATCTTGAAACATTC**GTTTGAAGGGTGGAGGTGATA**GGCCCA . . .**

$(TAAA)_{7\text{-}11}$ → PCR product: 132-148 bp

TAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAA

5 alleles

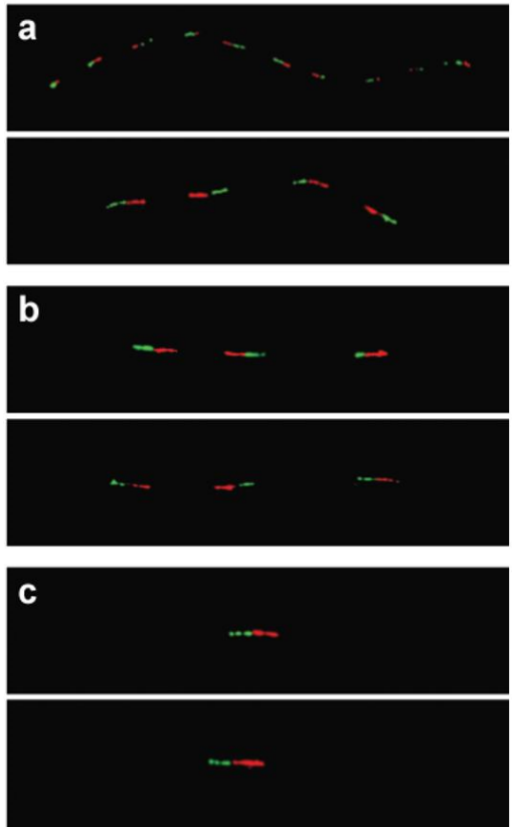
CNV: the amylase genes

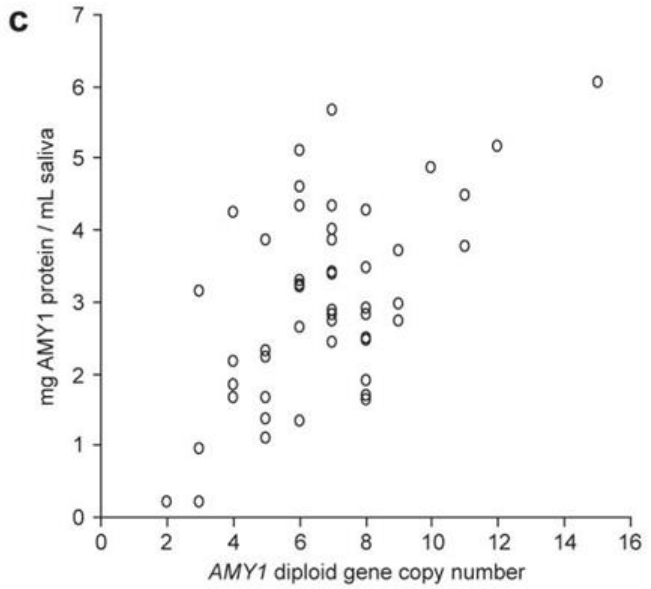


Japanese individual
High-starch diet
(14 copies)

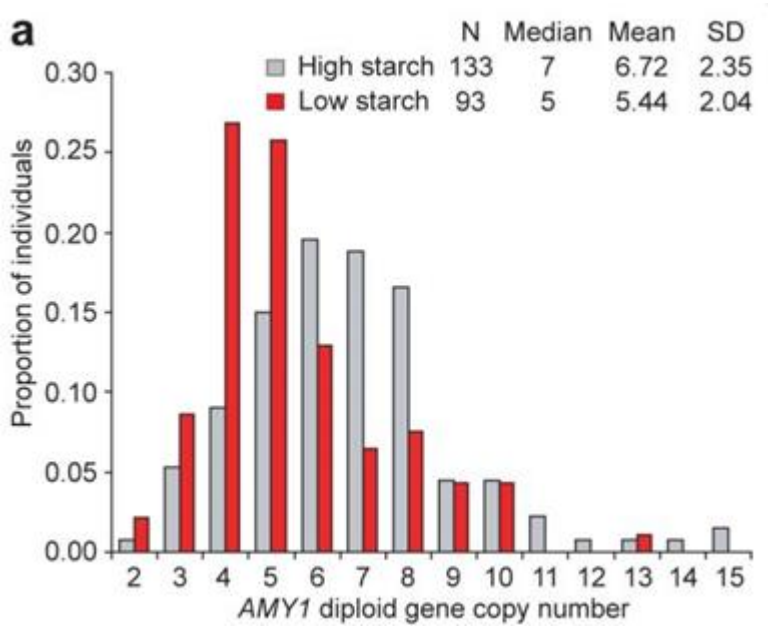
African individual
Low-starch diet
(6 copies)

Chimpanzee
Low-starch diet
(2 copies)





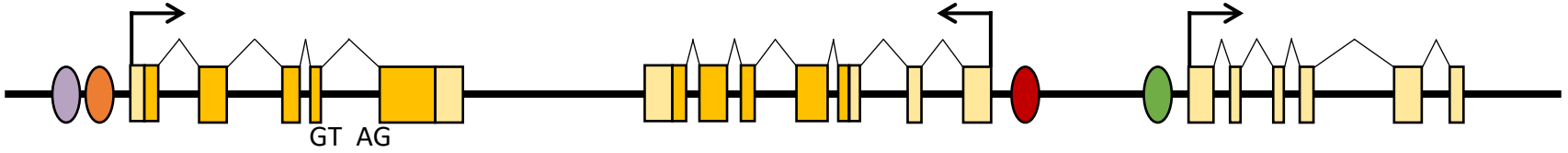
The amylase protein levels in saliva are proportional to the number of the *AMY1* gene copies



Individuals from populations with high-starch diets have on average more *AMY1* copies than those with traditionally low-starch diets.

Perry *et al*. (2007) *Nature Genetics*

Do all variants have effect on phenotype?

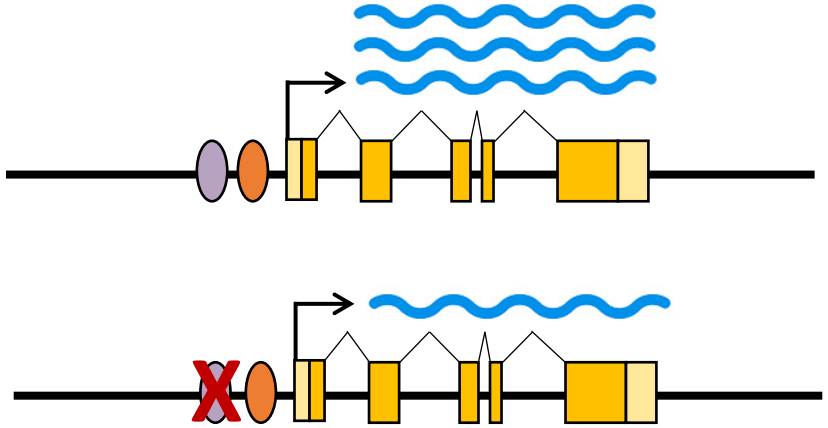


Only polymorphisms in **FUNCTIONAL ELEMENTS** may have phenotypical effects

Polymorphisms inside coding regions will NOT always have consequences

				Non-synonymous		Synonymous	
Seq 1	CTG	GAC	AGG	CGA	GGA	ATA	CAG
	L	**D**	**R**	**R**	**G**	**I**	**Q**
Seq 2	CTG	GAC	AGG	CAA	GGT	ATA	CAG
	L	**D**	**R**	**Q**	**G**	**I**	**Q**

Polymorphisms outside coding regions CAN have consequences



Amino acid change: white tigers

Panthera tigris

Pigmentation variant of the Bengal tiger (*Panthera tigris tigris*)

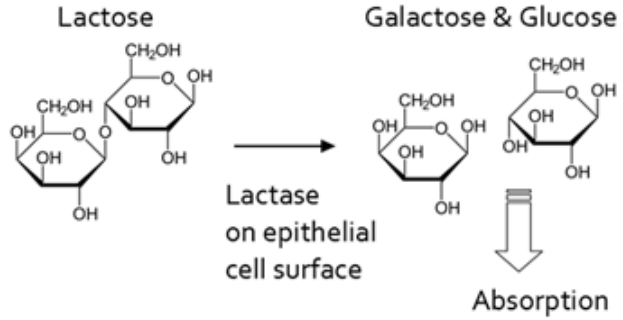
Recessive variant caused by alanine-to-valine substitution at amino acid residue 477 in the transporter protein SLC45A2 conserved in other species.

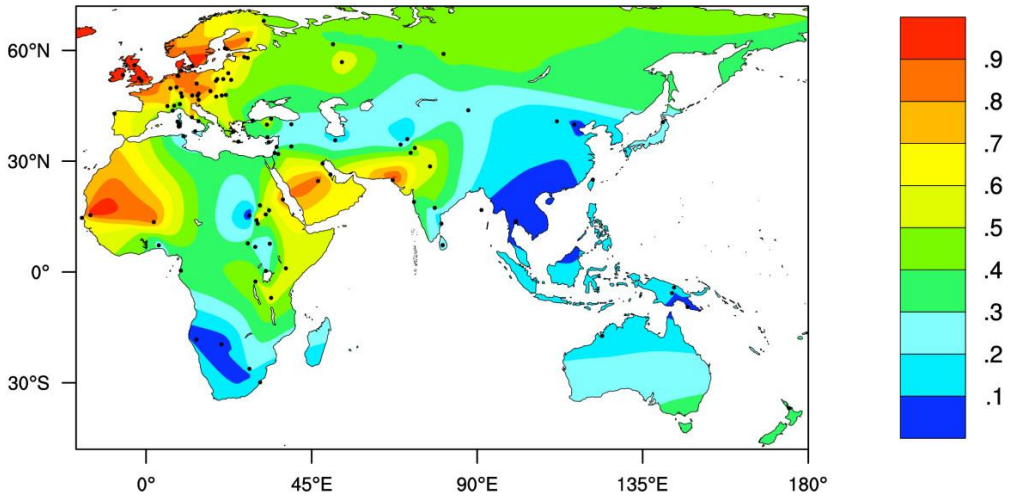
The *SLC45A2* missense mutation primarily inhibits the synthesis of red/yellow pheomelanin, with no or only minor effect on black eumelanin.

Xu *et al.* (2013) *Current Biology*

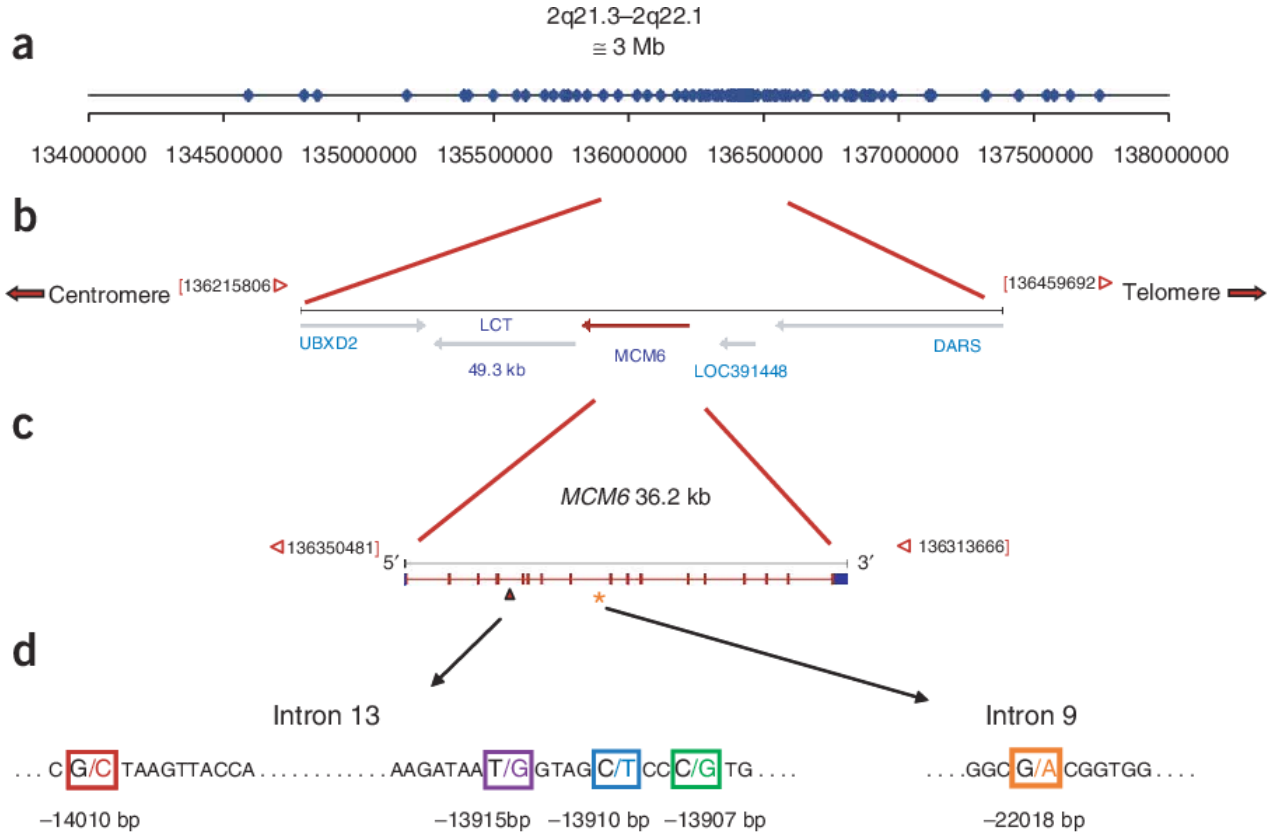
Regulatory change: lactase persistence

In most mammals ability to digest milk disappears with age and is related to the production of the lactase enzyme (*LCT* gene)





Lactase production in adults shows large variability in human populations and seems related to pastoralism



SNPs associated to lactase persistence are located in introns of flanking gene *MCM6*

Itan *et al*. (2010) *BMC Evolutionary Biology*
Tishkoff *et al*. (2007) *Nature Genetics*

Allele and genotype frequencies

Allele number = number of different alleles in a particular gene

Genotype frequency = proportion of a given genotype among all individuals in a group

Allele frequency = proportion of a given allele among all the alleles in a group of individuals

Allele and genotype numbers in diploid organisms

Allele number = k

$$\text{Genotype number} = \frac{k\,(k+1)}{2}$$

Allele number	Genotype number
1	1
2	3
3	6
4	10

2 alleles
AA
AB
BB

3 alleles
AA
BB
CC
AB
AC
BC

4 alleles
AA
BB
CC
DD
AB
AC
AD
BC
BD
CD

EXCEPTION X-linked variants

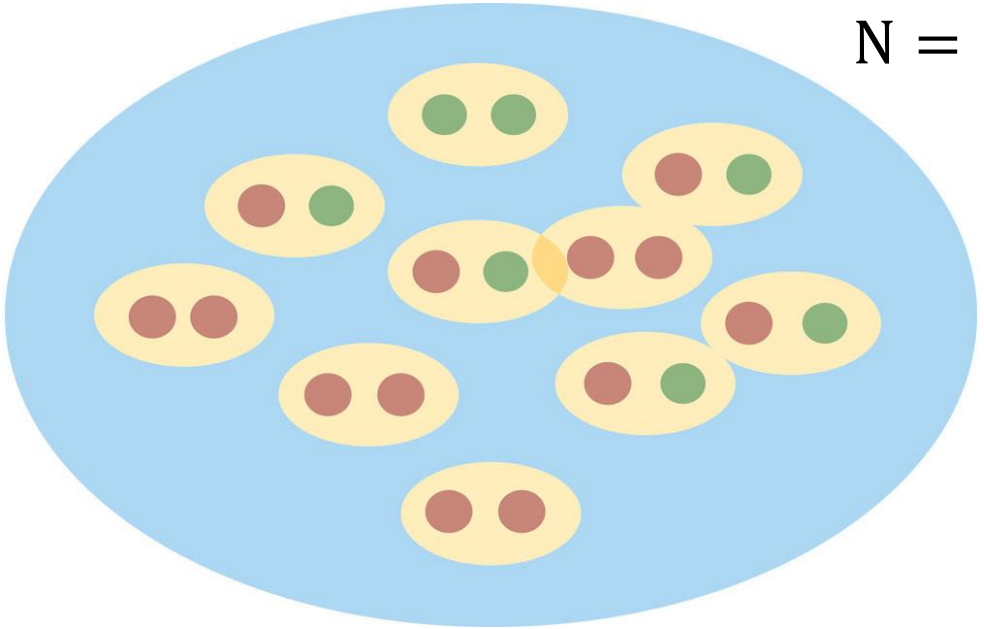
Allele frequencies in diploid organisms

2 alleles: *A* and *a*

$$\text{Freq}(A) = \text{p} = \frac{\text{number of } A \text{ alleles}}{\text{total number of alleles}}$$

$$\text{Freq}(a) = \text{q} = \frac{\text{number of } a \text{ alleles}}{\text{total number of alleles}}$$

Total number of alleles =
Total number of individuals (N) x 2 = 2N



$$\text{N} = 10$$

$$\text{Freq}(A) = \text{p} = \frac{7}{10 \cdot 2} = \frac{7}{20} = 0.35$$

$$\text{Freq}(a) = \text{q} = \frac{13}{10 \cdot 2} = \frac{13}{20} = 0.65$$

All allele frequencies must add 1

Genotype frequencies

With 2 alleles there are 3 possible genotypes:
AA, Aa, aa

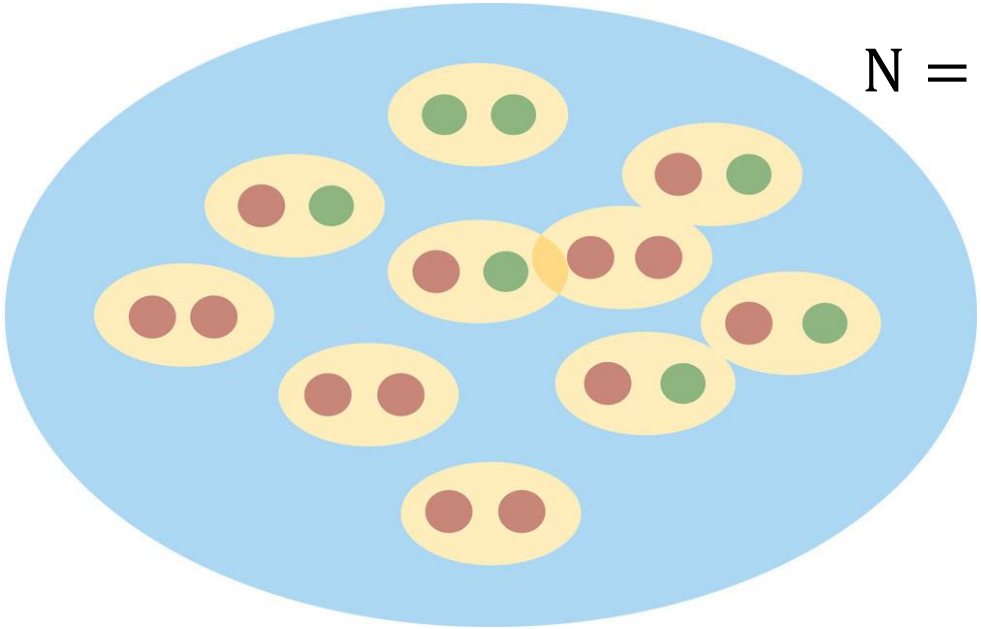
$$\text{Freq}(AA) = \text{P} = \frac{\text{number of } AA \text{ individuals}}{\text{total number of individuals}}$$

$$\text{Freq}(Aa) = \text{H} = \frac{\text{number of } Aa \text{ individuals}}{\text{total number of individuals}}$$

$$\text{Freq}(aa) = \text{Q} = \frac{\text{number of } aa \text{ individuals}}{\text{total number of individuals}}$$

IMPORTANT

All genotype frequencies must add 1



$\text{N} = 10$

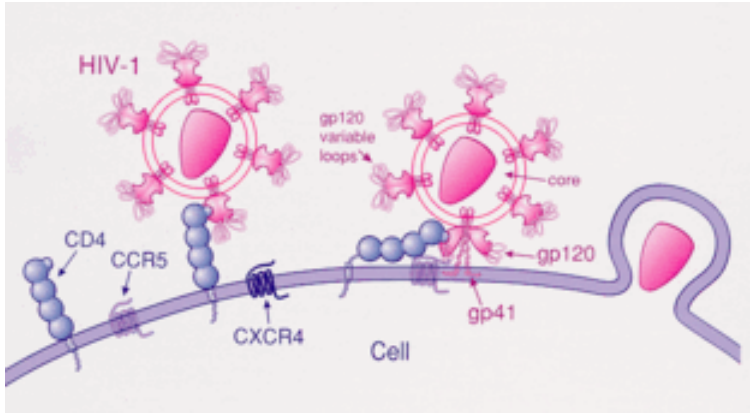
$$\text{Freq}(AA) = \text{P} = \frac{1}{10} = 0.1$$

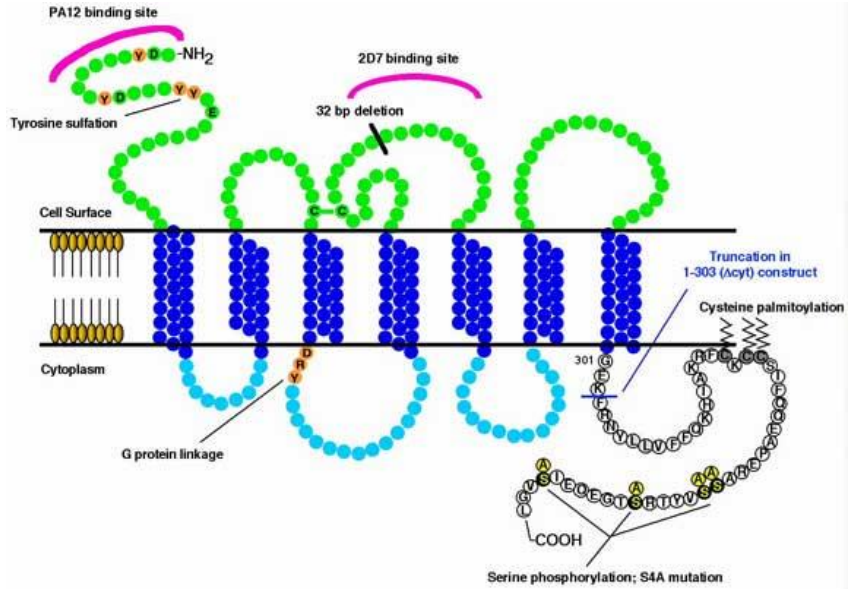
$$\text{Freq}(Aa) = \text{H} = \frac{5}{10} = 0.5$$

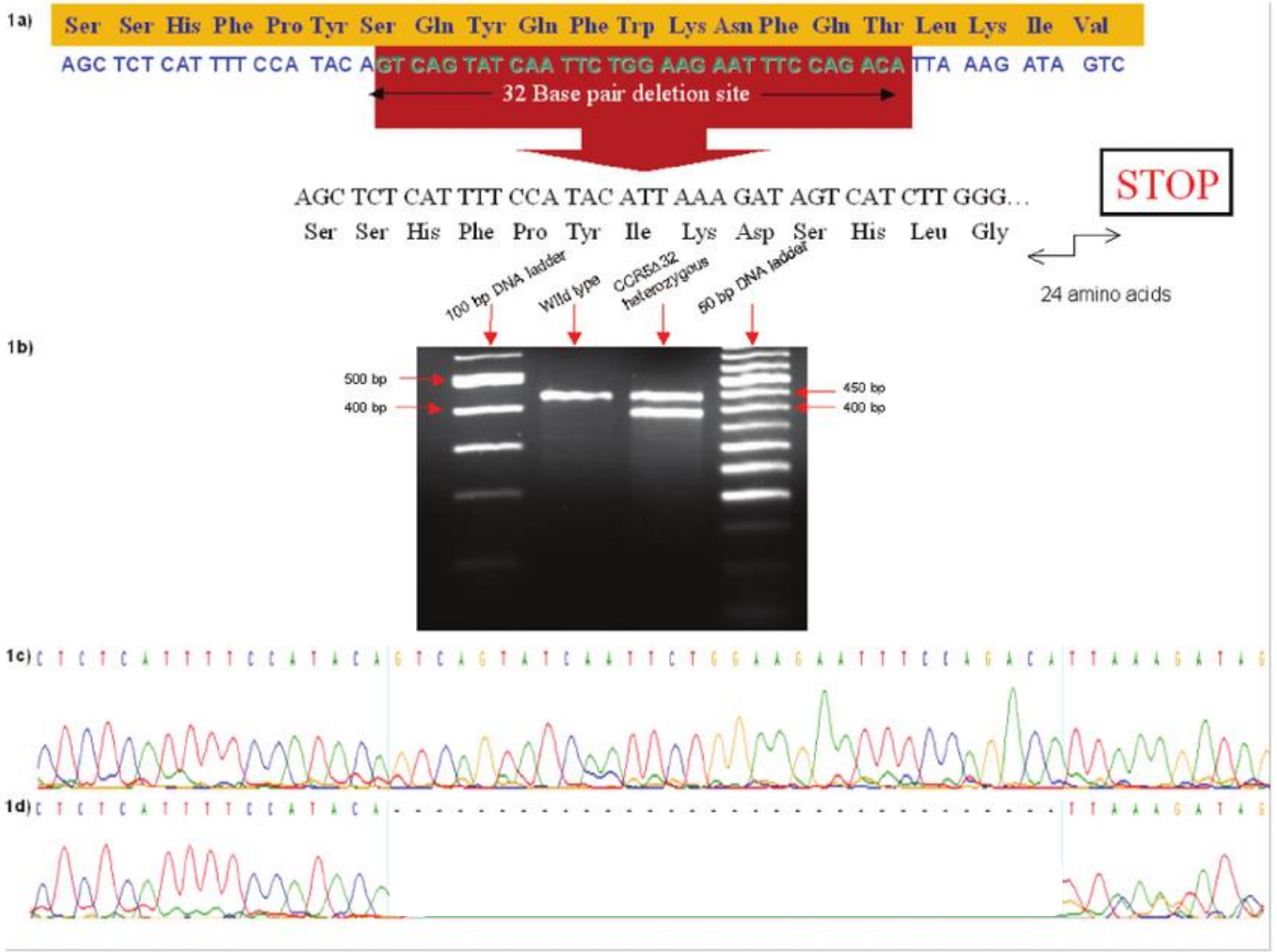
$$\text{Freq}(aa) = \text{Q} = \frac{4}{10} = 0.4$$

Allelic and genotype frequencies in *CCR5* receptor gene

32-bp indel that disrupts human chemokine receptor gene *CCR5*
Individuals homozygous for deletion are strongly resistant to infection by HIV-1







Bhatnagar *et al.* (2009) *Balkan Journal of Medical Genetics*

Allele and genotype frequencies in *CCR5* receptor gene

GENOTYPE DATA FROM A POPULATION IN PARIS

Genotype	Number of individuals	Genotype frequencies	Number of + alleles	Number of Δ32 alleles
+/+	224	P = 224/294 = 0.7619	224 · 2 = 448	0
Δ32/+	64	H = 64/294 = 0.2177	64	64
Δ32/Δ32	6	Q = 6/294 = 0.0204	0	6 · 2 = 12
Total	294	1	512	76

Total alleles = 588

ALLELE FREQUENCIES

$$p = \frac{512}{588} = 0.8707 \qquad q = \frac{76}{588} = 0.1293$$

p + q = 1

P + H + Q = 1

Allele and genotype frequencies – Counting method

Genotype	Number of individuals	Genotype frequencies	Number of + alleles	Number of Δ32 alleles
A_1/A_1	N_1	$P = N_1/N$	$2N_1$	0
A_1/A_2	N_2	$H = N_2/N$	N_2	N_2
A_2/A_2	N_3	$Q = N_3/N$	0	$2N_3$
Total	N	1	$2N_1 + N_2$	$2N_3 + N_2$

Total number of alleles = 2N

ALLELE FREQUENCIES

$$p = \frac{2N_1 + N_2}{2N} = P + \frac{1}{2}H \qquad q = \frac{2N_3 + N_2}{2N} = Q + \frac{1}{2}H$$

From individual counts with each genotype

From genotype frequencies

Example problem

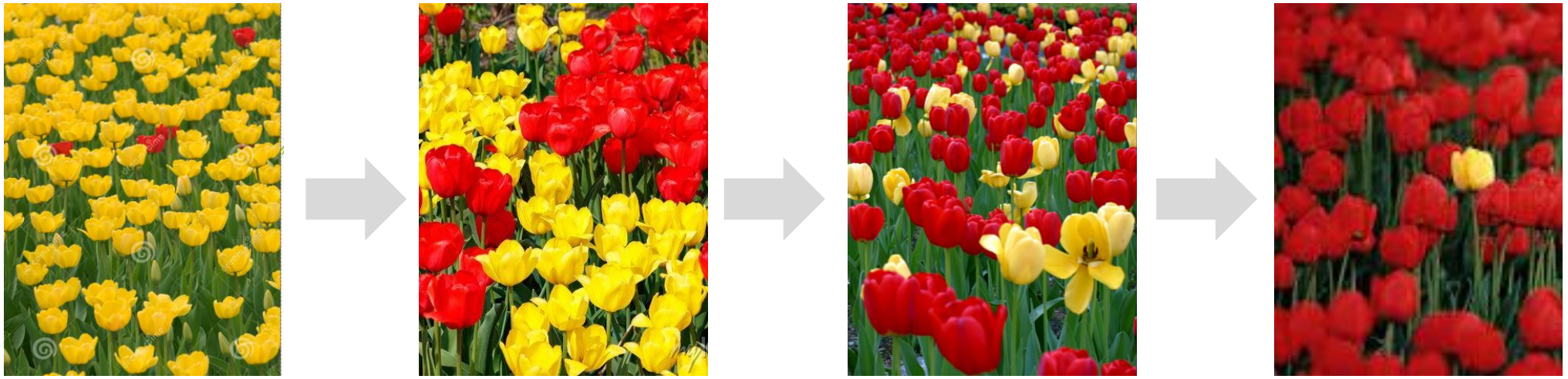
In a plant population, flower color is determined by a single gene with two codominant alleles. There are 170 plants with pink flowers, 340 plants with purple flowers, and 21 plants with flowers combining pink and purple.

Calculate allele and genotype frequencies in this population.



Evolution definition

Evolution = Change of allele frequencies in a population over time



Hardy-Weinberg equilibrium

ASSUMPTIONS

- Diploid organism
- Sexual reproduction
- Non-overlapping generations
- Random mating
- Equal allele frequencies in both sexes
- Large population size
- No migration
- No mutation
- No selection

Hardy–Weinberg provides a **null model**, a prediction based on a simplified or idealized situation, where no biological processes are acting and genotype frequencies are the result of random combination.

Hardy-Weinberg equilibrium

PRINCIPLES

1. Genotype frequencies in a population with random mating are determined by allele frequencies
2. Allele and genotype frequencies in a population in Hardy-Weinberg equilibrium do not change in the next generation

G. H. Hardy
(1877 – 1947)

FIGURE 1.—G. H. Hardy.

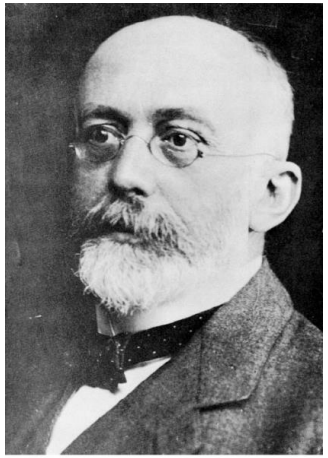
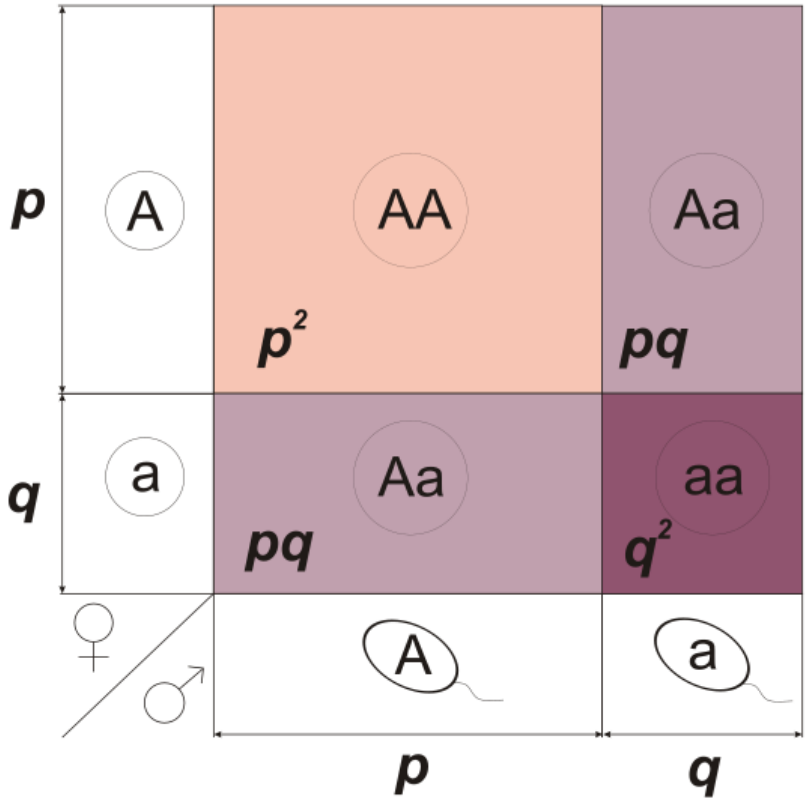


FIGURE 1.—Wilhelm Weinberg (from STERN 1962).

W. Weinberg
(1862 – 1937)

Hardy-Weinberg principles

1. **Genotype frequencies in a population with random mating are determined by allele frequencies**



GENOTYPE FREQUENCIES

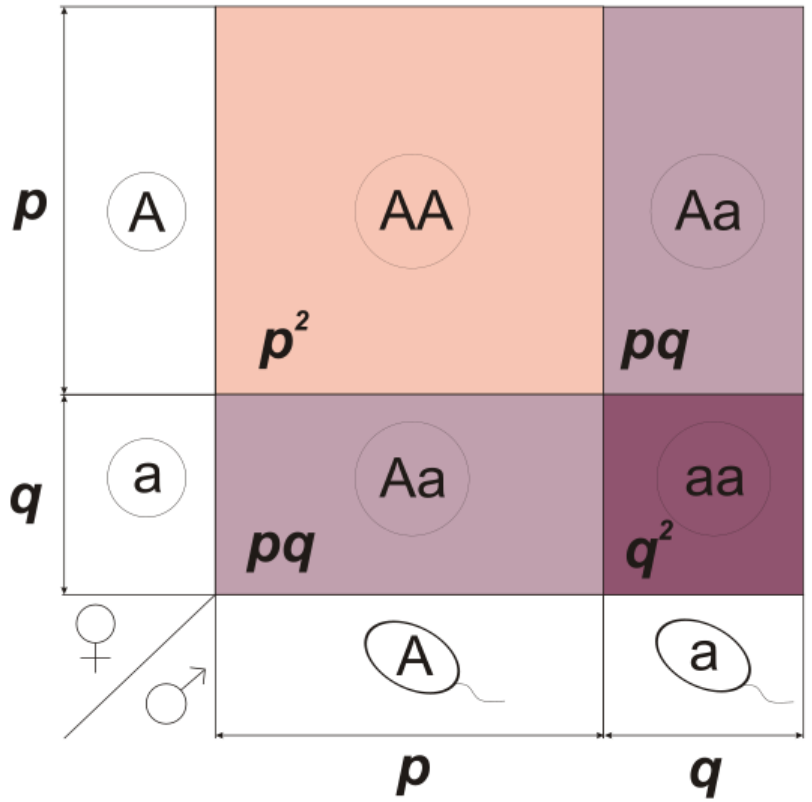
P (AA) = p^2
H (Aa) = 2pq
Q (aa) = q^2

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

Image from https://en.wikipedia.org/wiki/Hardy-Weinberg_principle

Hardy-Weinberg principles

1. **Genotype frequencies in a population with random mating are determined by allele frequencies**



$p = 0.6$
$q = 0.4$

$p + q = 0.6 + 0.4 = 1$

GENOTYPE FREQUENCIES

$P(AA) = 0.6 \cdot 0.6 = 0.36 = p^2$
$H(Aa) = 0.4 \cdot 0.6 + 0.6 \cdot 0.4 = 0.24 + 0.24 = 0.48$
$= pq + pq = 2pq$
$Q(aa) = 0.4 \cdot 0.4 = 0.16 = q^2$

$p^2 + 2pq + q^2 = 1$
$0.36 + 0.48 + 0.16 = 1$

Hardy-Weinberg principles

2. Allele and genotype frequencies in a population in Hardy-Weinberg equilibrium do not change in the next generation

PARENTAL MATING FREQUENCIES

			Fathers		
			AA	Aa	aa
			P	H	Q
Mothers	AA	P	P^2	PH	PQ
	Aa	H	PH	H^2	HQ
	aa	Q	PQ	HQ	Q^2

		Offspring genotype frequencies		
Mating	Frequency	AA	Aa	aa
AA x AA	P^2	1		
AA x Aa	2PH	1/2	1/2	
AA x aa	2PQ		1	
Aa x Aa	H^2	1/4	1/2	1/4
Aa x aa	2HQ		1/2	1/2
aa x aa	Q^2			1
Totals next generation		**P'**	**H'**	**Q'**

Hardy-Weinberg principles

2. Allele and genotype frequencies in a population in Hardy-Weinberg equilibrium do not change in the next generation

		Offspring genotype frequencies		
Mating	**Frequency**	**AA**	**Aa**	**aa**
AA x AA	P^2	P^2		
AA x Aa	2PH	PH	PH	
AA x aa	2PQ		2PQ	
Aa x Aa	H^2	$H^2/4$	$H^2/2$	$H^2/4$
Aa x aa	2HQ		HQ	HQ
aa x aa	Q^2			Q^2
Totals next generation		**P'**	**H'**	**Q'**

$$P' = P^2 + \frac{2PH}{2} + \frac{H^2}{4} = \left(P + \frac{H}{2}\right)^2 = p^2$$

$$H' = \frac{2PH}{2} + 2PQ + \frac{H^2}{2} + \frac{2HQ}{2} = 2\left(P + \frac{H}{2}\right)\left(Q + \frac{H}{2}\right) = 2pq$$

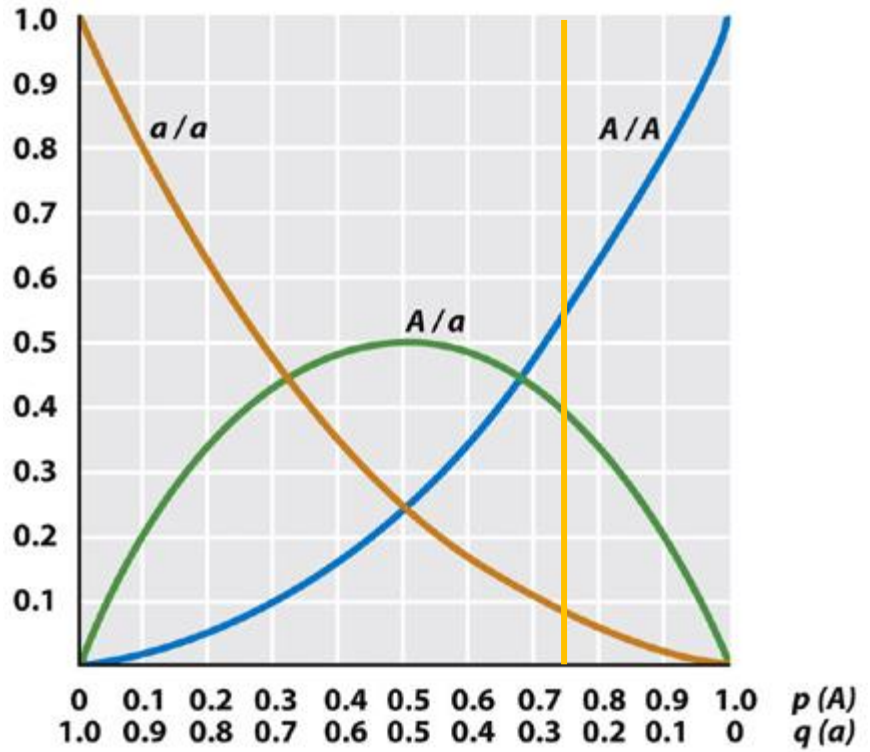
$$Q' = \frac{H^2}{4} + \frac{2HQ}{2} + Q^2 = \left(Q + \frac{H}{2}\right)^2 = q^2$$

Same genotype frequencies in the next generation

$$p' = P' + \frac{1}{2}H' = p^2 + \frac{1}{2}2pq = p^2 + pq = p(p+q) = p$$

Same allele frequencies in the next generation

Hardy-Weinberg equilibrium expected genotype frequencies

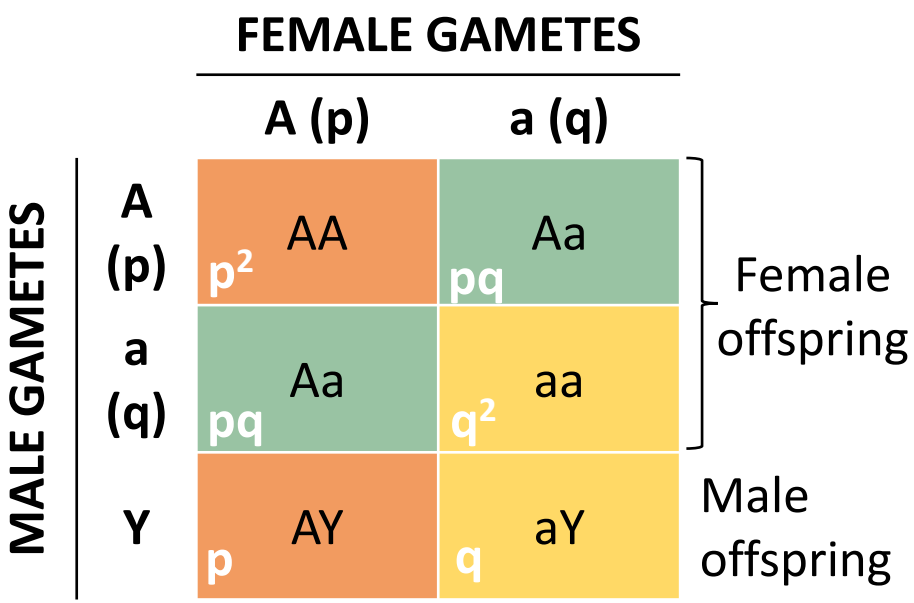


p = 0.75	p^2 = 0.5625	2pq = 0.375	q^2 = 0.0625
q = 0.25	**56.25% AA**	**37.5% Aa**	**6.25% aa**

Griffiths, Wessler, Lewontin & Carroll (2008) *Introduction to Genetics Analysis*

HW equilibrium in X-linked genes

- Males only have one X chromosome
- Genotype frequencies among females:
  $AA = p^2$
  $Aa = 2pq$
  $aa = q^2$
- Genotype frequencies in males:
  $A = p$
  $a = q$



If *a* is a recessive allele there will be many more males exhibiting the trait than females because the frequency of affected females (q^2) will be much smaller than the frequency of affected males (q)

Generations needed to reach HW equilibrium

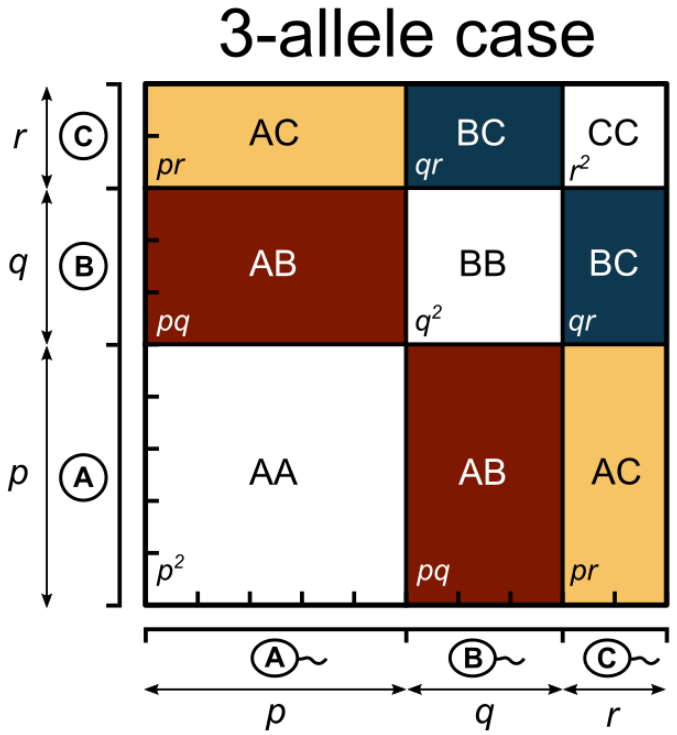
If allele frequencies are identical in males and females

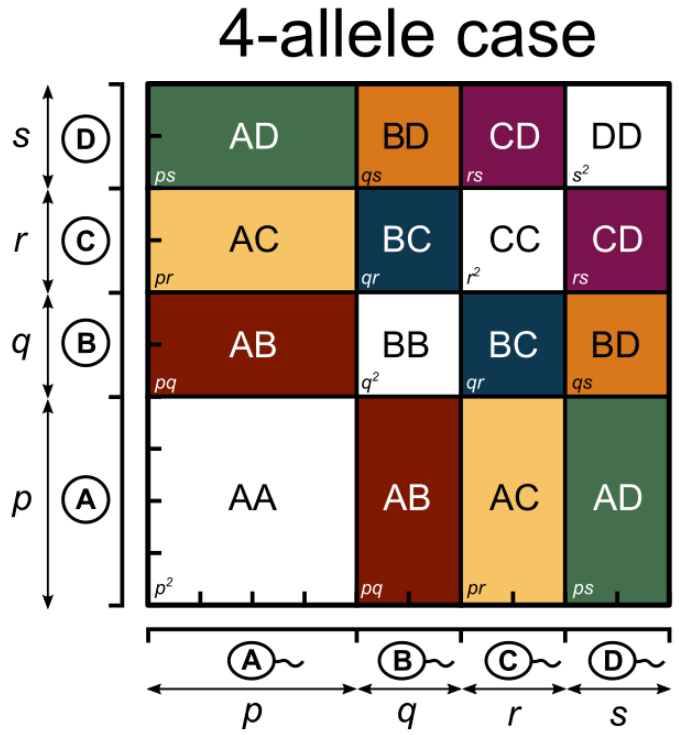
After one round of random mating, we obtain HWE allele and genotype frequencies

If allele frequencies are NOT identical in males and females

After the first round of random mating, same allele frequencies in both sexes
After the second round of random mating, HWE will be established

HW equilibrium with multiple alleles





Genotype	Frequency
AA	p^2
AB	$2pq$
AC	$2pr$
BB	q^2
BC	$2qr$
CC	r^2

In general

For A_iA_i homozygotes p_i^2

For A_iA_j heterozygotes $2p_ip_j$

Applications of Hardy-Weinberg equilibrium

- Null model to analyze the effect of different factors on the genetic composition of a population
- Test if genotype frequencies adjust to expected values > If they do not, one of the assumptions is not true
- Estimation of allele frequencies in case of dominance
- Test alternative models of inheritance
- Forensic DNA profiling

Adjustment of genotype frequencies to Hardy-Weinberg

Genotype	Observed	Expected	$\chi^2 = \frac{(O - E)^2}{E}$
MM	298	$p^2 \cdot N = 294.3$	0.0465
MN	489	$2pq \cdot N = 496.4$	0.1103
NN	213	$q^2 \cdot N = 209.3$	0.0654
Total	N = 1000	1000	0.222

ALLELE FREQUENCIES

$$p = \frac{298 \cdot 2 + 489}{2000} = 0.5425 \qquad q = \frac{213 \cdot 2 + 489}{2000} = 0.4575$$

χ^2 TEST

df = 3 – 1 – 1 = 1

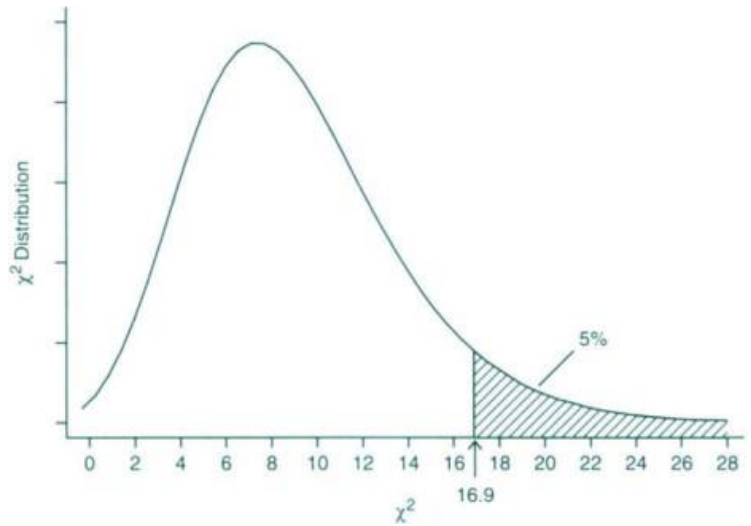
$\chi^2_{0.05,1} = 3.81$

0.222 < 3.81

H_0 = Equal

H_1 = Different

χ^2 test



Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	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
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209

The χ^2 (chi-square) distribution for df = 9 with an $\alpha = 0.05$ and its corresponding chi-square value of 16.9.

The α probability is shown as the shaded area under the curve to the right of a critical chi-square, in this case, representing a 5% probability that a value drawn randomly from the distribution will exceed a critical chi-square of 16.9

ONLINE TOOLS

https://www.socscistatistics.com/tests/chisquare2/default2.aspx
https://www.mathsisfun.com/data/chi-square-calculator.html

Example problem

In a plant population, flower color is determined by a single gene with two codominant alleles. There are 170 plants with pink flowers, 340 plants with purple flowers, and 21 plants with flowers combining pink and purple.

Determine if this population is in Hardy-Weinberg equilibrium with a statistic test.



Applications of Hardy-Weinberg equilibrium

- Null model to analyze the effect of different factors on the genetic composition of a population
- Test if genotype frequencies adjust to expected values > If they do not, one of the assumptions is not true
- Estimation of allele frequencies in case of dominance
- Test alternative models of inheritance
- Forensic DNA profiling

Estimating allele frequencies with dominance

Genotype	Phenotype	Expected frequencies	Observed frequencies
DD	Rh+	$p^2 + 2pq$	0.858
Dd			
dd	Rh-	q^2	0.142
Total	N	1	1

ALLELE FREQUENCIES

$$\text{Freq(d)} = q = \sqrt{0.142} = 0.3768$$

$$\text{Freq(D)} = p = 1 - 0.3768 = 0.6232$$

GENOTYPE FREQUENCIES

$$\text{Freq(Dd)} = 2pq = 2 \cdot 0.3768 \cdot 0.6232 = 0.4697$$

$$\text{Freq(DD)} = p^2 = (0.6232)^2 = 0.3884$$

PROPORTION OF HETEROZYGOTES WITHIN Rh+

$$\frac{\text{Het}}{\text{Rh}+} = \frac{2pq}{p^2 + 2pq} = \frac{0.4697}{0.4697 + 0.3884} = 0.547 = 54.7\%$$

Hartl (2000) *A primer of population genetics*

Applications of Hardy-Weinberg equilibrium

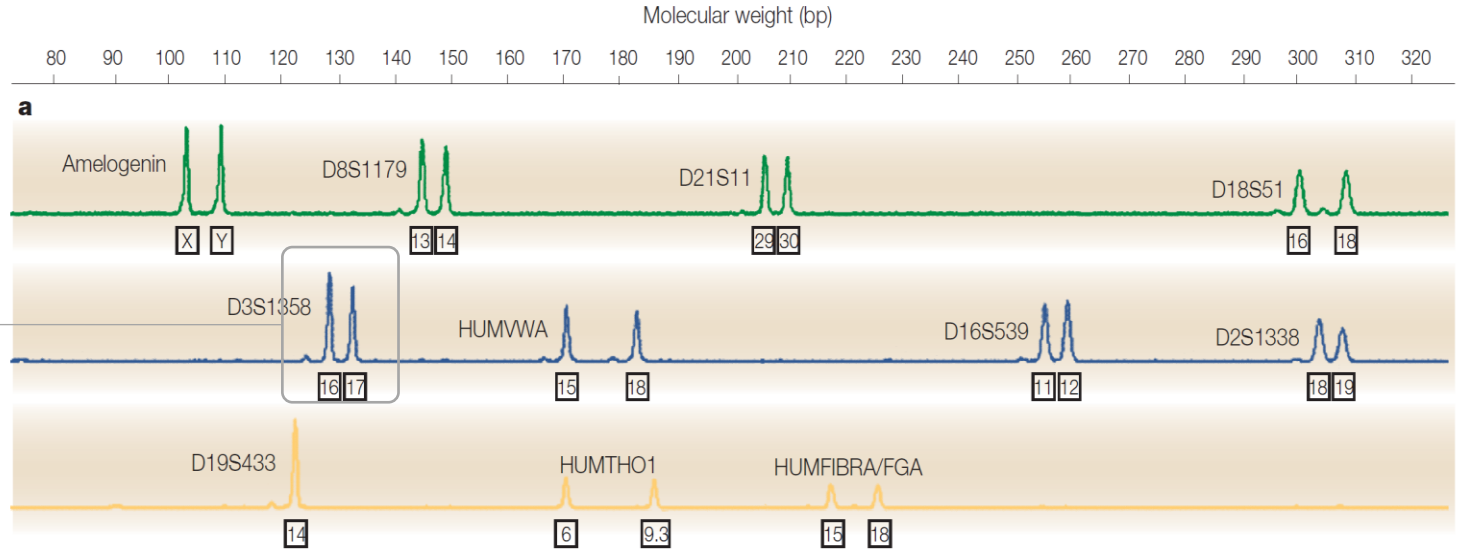
- Null model to analyze the effect of different factors on the genetic composition of a population
- Test if genotype frequencies adjust to expected values > If they do not, one of the assumptions is not true
- Estimation of allele frequencies in case of dominance
- Test alternative models of inheritance
- Forensic DNA profiling

Forensic DNA profiling

Microsatellite amplification and analysis

Allele 1 AGTGAGTGAGTGAGTG

Allele 2 AGTGAGTG



D3S1358		FGA		D5S818		D7S820		D8S1179		vWA		D13S317		D18S51		D21S11	
Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq
13	0.0025	18	0.0306	9	0.0308	6	0.0025	>9	0.0179	13	0.0051	8	0.0995	>11	0.0127	24.2	0.0051
14	0.1404	19	0.0561	10	0.0487	7	0.0172	9	0.0102	14	0.1020	9	0.0765	11	0.0128	27	0.0459
15	0.2463	20	0.1454	11	0.4103	8	0.1626	10	0.102	15	0.1122	10	0.051	12	0.1276	28	0.1658
16	0.2315	20.2	0.0026	12	0.3538	9	0.1478	11	0.0587	16	0.2015	11	0.3189	13	0.1224	29	0.1811
17	0.2118	21	0.1735	13	0.1462	10	0.2906	12	0.1454	17	0.2628	12	0.3087	14	0.1735	30	0.2321
18	0.1626	22	0.1888	14	0.0077	11	0.202	13	0.3393	18	0.2219	13	0.1097	15	0.1276	30.2	0.0383
19	0.0049	22.2	0.0102	15	0.0026	12	0.1404	14	0.2015	19	0.0842	14	0.0357	16	0.1071	31	0.0714
		23	0.1582			13	0.0296	15	0.1097	20	0.0102			17	0.1556	31.2	0.0995
		24	0.1378			14	0.0074	16	0.0128					18	0.0918	32	0.0153
		25	0.0689					17	0.0026					19	0.0357	32.2	0.1122
		26	0.0179											20	0.0255	33.2	0.0306
		27	0.0102											21	0.0051	35.2	0.0026
														22	0.0026		

Microsatellite	Genotype	Frequency
D3S1358	17, 18	0.0689
FGA	24, 25	0.0190
D5S818	12, 13	0.1035
D7S820	11, 12	0.0567
D8S1179	13, 14	0.1367
vWA	17, 17	0.0691
D13S317	9, 12	0.0472
D18S51	18, 18	0.0084
D21S11	29, 30	0.0841
Amelogenin	X, Y	0.5

P = 0.000000000001213

1 in 862,379,847,814 people will have this 10-loci profile