Lluís Giménez Gabarró

1. Genetic Variation

When looking at any organism, we can see phenotypic variation (color, patrons…).

Fixism vs Evolution

Fixism: The world has always been as we see it now. The organisms don't evolve, they have always been there.
Here, the variation is inconvenient or complication. If you are trying to classify organisms, the fact that the same type of organism has 2 possible colors is confusing. A species that has different phenotypes is wrong.

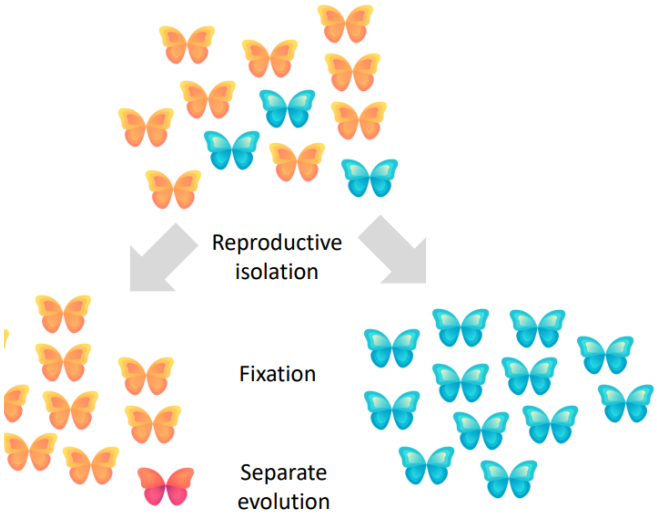
Evolution: Change in the heritable characteristics of populations over successive generations. Species change over time and this allows the population to change.
Thus, variation is essential.

Evolution does not take place in individuals but in populations, which change over time. Populations adapt to a new environment.

Polymorphism

Variation comes from mutations.
Polymorphism: Variation within species (they need to have a frequency >5-1%).

Once there are variants in the population, we can have an event of reproductive isolation (they can not reproduce with each other), where both groups will start to evolve independently.
Thus, they will accumulate a lot of changes (fixation of variations) and maybe they will become different species.



Concepts

Phenotype: Morphological, biochemical, physiological or behavioral attributes of an individual. In other words, any feature of an organism.
Examples: Color of the eyes, results of a blood test (high cholesterol or not)…

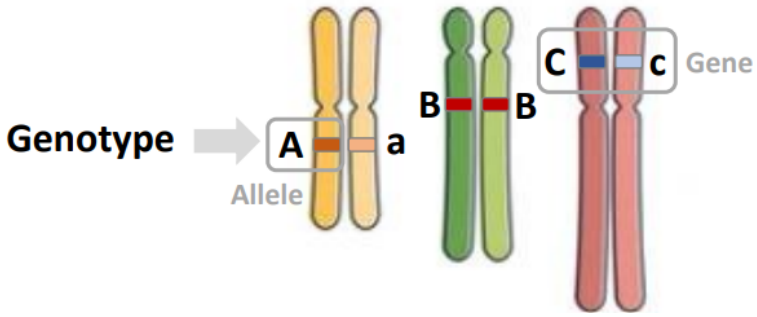
Genotype: The set of alleles that an organism has. It determines the phenotype.

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: Not an entire species, but a group of individuals of the same species living in a geographically restricted area so that any member can potentially mate with any other member.

Diploid organism has 2 sets of chromosomes. Each of the copies has been inherited from one of the parents.
Thus, we have 2 alleles of each gene. They can be equal or different.

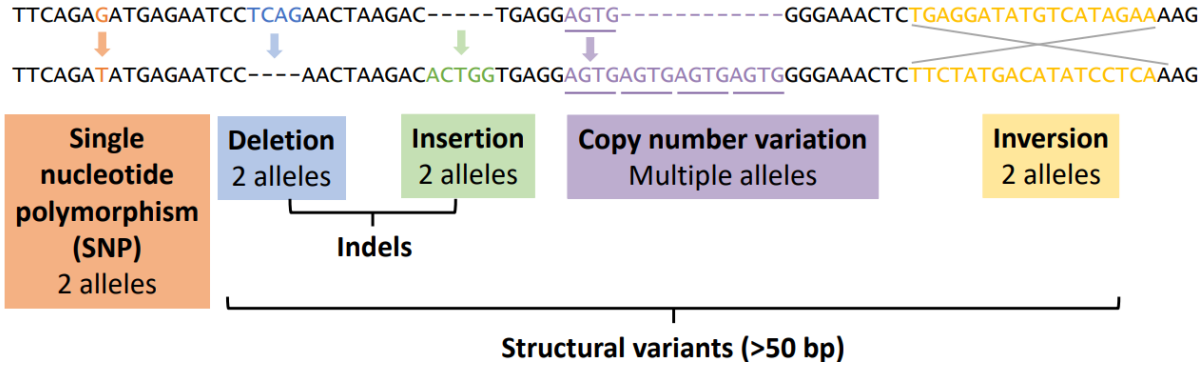


Types of variants

- Single nucleotide polymorphisms (**SNP**). Since we have 2 alleles, one allele can have the mutation and not the other or both can have it… So, in that position we have one DNA variant.
  We can have 3 alleles if there is another mutation in the same position (but this is not likely).

  Structural variants (>50 bp)
  - **Deletion or insert** (2 alleles). Indels because you do not know if there has been an addition or a deletion (you see the same thing). You can just look at different species and check if it has the sequence or not.
  - Copy number variations (**duplications**). Multiple alleles in the population, since you can have 2, 3, 4, 5, 6… duplications.
  - **Inversions**. 2 alleles (one orientation or the other)

  These structural variants can be huge. They can delete a whole gene (or more), duplicate a gene…



Lluís Giménez Gabarró

Microsatellites or Short Tandem Repeats (STRs)

It is a copy number variant of small sequences of 2-5 bp.
A species can have a microsatellite of 7, 8, 9, 10 or 11 copies.
You can easily amplify it by PCR since it is a small sequence.
You will be able to differentiate the different alleles because the PCR products will have different lengths and thus you can make an electroforesis.

They are used in paternity tests.

...TGAATGCGGAGTCTGCTTGAGATTCTCTCCCCTTCCTCTCCCTCTATCCCTCCT
CATGCTCTCTCTCATTATCTCATAAATAAATAAATAAATAAATAAATAAATAAATAA
ATAAATATCTTGAAACATTCGTTTGAAGGGTGGAGGTGATAGGCCCA...

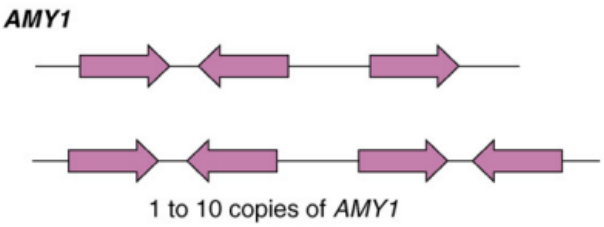
$(TAAA)_{7-11}$ → PCR product: 132-148 bp

TAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAATAAATAAA
TAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAA

5 alleles

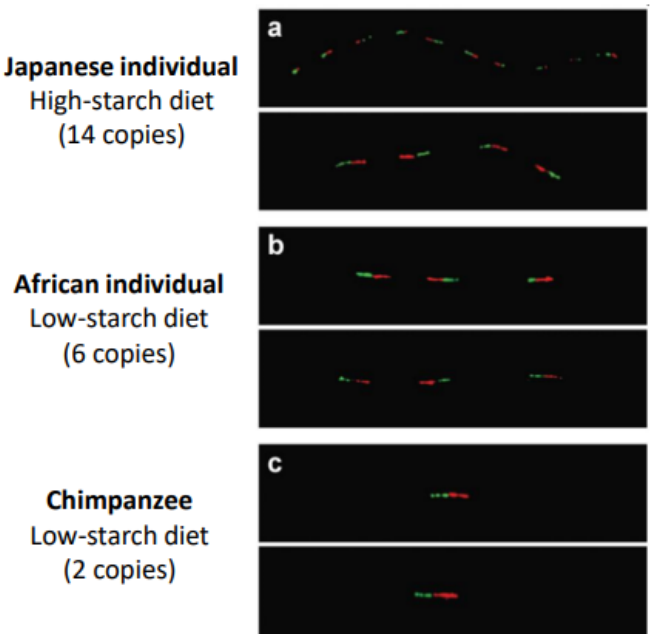
Amylase gene (AMY1)

This is another example of a copy number variant where a whole gene is duplicated (like the microsatellites but larger and include a gene).
Between 2-15 copies of the gene.



The amylase is an enzyme that can be found in the saliva and it is responsible for the digestion of the starch. If you eat more starch, you will have more copies of the amylase.

In the image we can see FISH of the 2 chromosomes of each individual. In green and red we have a copy of the amylase.

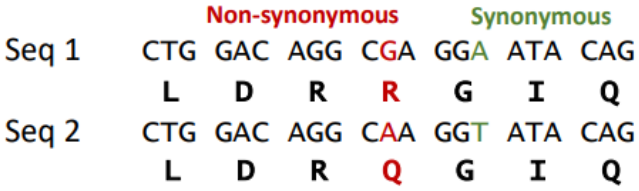


Japanese: 1 chromosome with 10 copies and the other with 4 copies.
African: 3+3
Chimp: 1+1

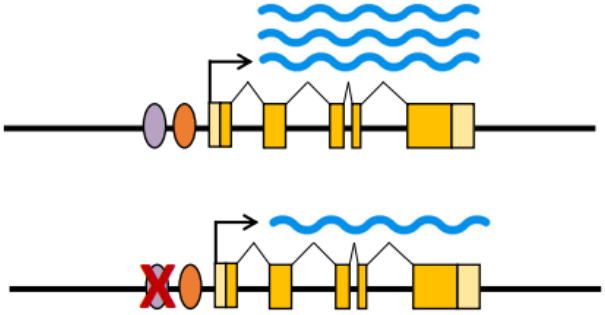
Do all variants have an effect on the phenotype?

Only polymorphisms in functional elements may have phenotypic effects.
Also, when changing an aa does not mean that the phenotype will change. Because maybe that region is not important for the phenotype or the properties of the new aa are similar.

Polymorphisms inside coding regions will NOT always have consequences.



Polymorphisms outside coding regions CAN have consequences. Maybe they affect the regulatory regions.



Synonymous variant: The aa is not modified.
Non-synonymous variant: aa is modified, but the properties can be similar.

Regulatory change

Lactase digests the lactose of the milk. Since we are mammals, we have this enzyme.
But when adults, most of the mammals lose the expression of the lactase.
In humans, we keep generating lactase (lactase persistence, a variant).
Thus, we have a mutation in the regulatory region. In fact, the variants that are associated with lactase persistence are located in introns of the gene located next to the lactase gene (MCM6).

If only one of the alleles has this mutation, we will be able to express the lactase (not at a high level). So, it is a dominant allele, because you produce lactase.

Lluís Giménez Gabarró

Allele and genotype frequencies

Allele number: Number of different alleles in a particular gene. It depends if it's a SNP, indel, duplication…

$$\text{Allele number} = k$$

Genotype frequency: Proportion of a given genotype among all individuals in a group. The genotype frequencies must add to 1.

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

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

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

If we want to compute the different types of genotypes:

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

Here there is an exception regarding the X-linked variants. Because out of 50 individuals we will not have 100 chromosomes X (man have 1 chromosome X)

Allele frequency: Proportion of a given allele among all the alleles in a group of individuals. The allele frequencies must add to 1.

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

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

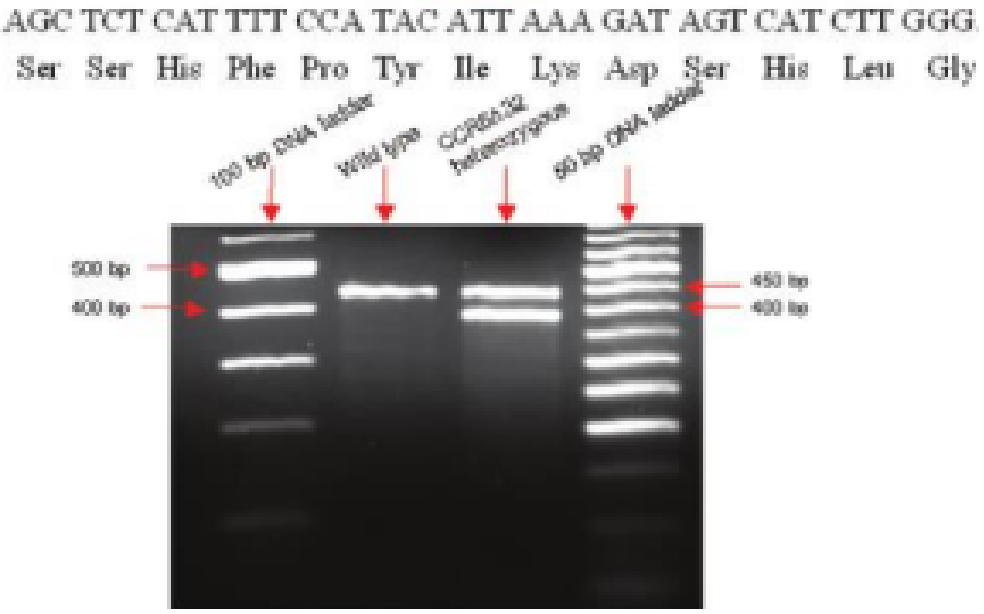
$$\text{Total number of alleles} = \text{Total number of individuals (N)} \times 2 = 2N$$

Lluís Giménez Gabarró

Allelic and genotype frequencies in CCR5 receptor gene

We have a 32bp deletion that disrupts a chemokine receptor gene (CCR5).
This receptor is used by HIV to infect human cells. Thus, individuals homozygous for deletion are strongly resistant to infection by HIV.

Since the reading-frame will be modified, the resulting protein will be completely different.
We can do a PCR and then an electrophoresis to check if an individual is homozygous or heterozygous for this allele.



Example. 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

The wild type allele is represented by a +.
We can compute the allele frequencies

ALLELE FREQUENCIES

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

We can compute the allele frequencies using 2 different formulas:

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

Evolution

Change of allele frequencies in a population over time

Hardy-Weinberg equilibrium

It 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.

There are some assumptions:

- Diploid organism
- Sexual reproduction. Meaning that the organism makes gametes (one from the father and one from the mother) that when binded they produce a new individual.
- Non-overlapping generations. For example an annual plant that grows in the summer, makes seeds in winter…
- Random mating
- Equal allele frequencies in both sexes
- Large population size
- No migration, mutation or selection

Principals:

- Genotype frequencies in a population with random mating are determined by allele frequencies. So, if I know the allele frequencies I can determine the genotype frequencies in a population.

$$P\ (AA) = p^2$$
$$Q\ (Aa) = 2pq$$
$$R\ (aa) = q^2$$

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

- Allele and genotype frequencies in a population in HWE do not change in the next generation (they all remain equal).

		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'**	**Q'**	**R'**

		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'**	**Q'**	**R'**

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

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

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

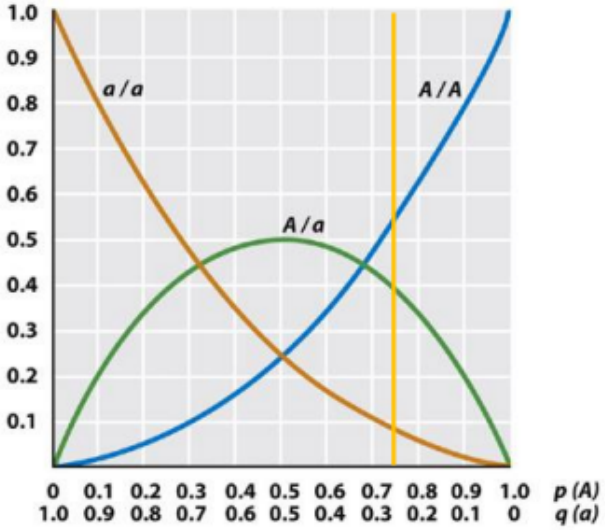
Same genotype frequencies in the next generation

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

Same allele frequencies in the next generation

Note that:

- P' = Progeny with a genotype AA
- Q' = Progeny with a genotype Aa
- R' = Progeny with a genotype aa
- p' = Frequency of allele A in the progeny
- q' = Frequency of allele a in the progeny



Here we can see that we will obtain the maximum number of heterozygotes when the probability of AA = aa = 0.25.
Also, when there is a large amount of homozygotes of one type, the other type will have a small amount and the rest will correspond to heterozygotes.

HWE in X-linked genes

If we want to know the number of alleles in a population that is linked to the X chromosome, we need to know the percentage of males and females.

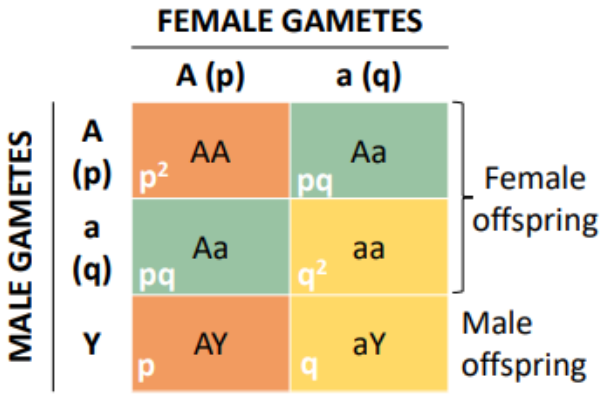
Genotype frequencies among females:

- AA = p^2
- Aa = 2pq
- aa = q^2

Genotype frequencies among males is equal to the allele frequencies:

- A = p
- a = q

We have 5 genotypes!



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

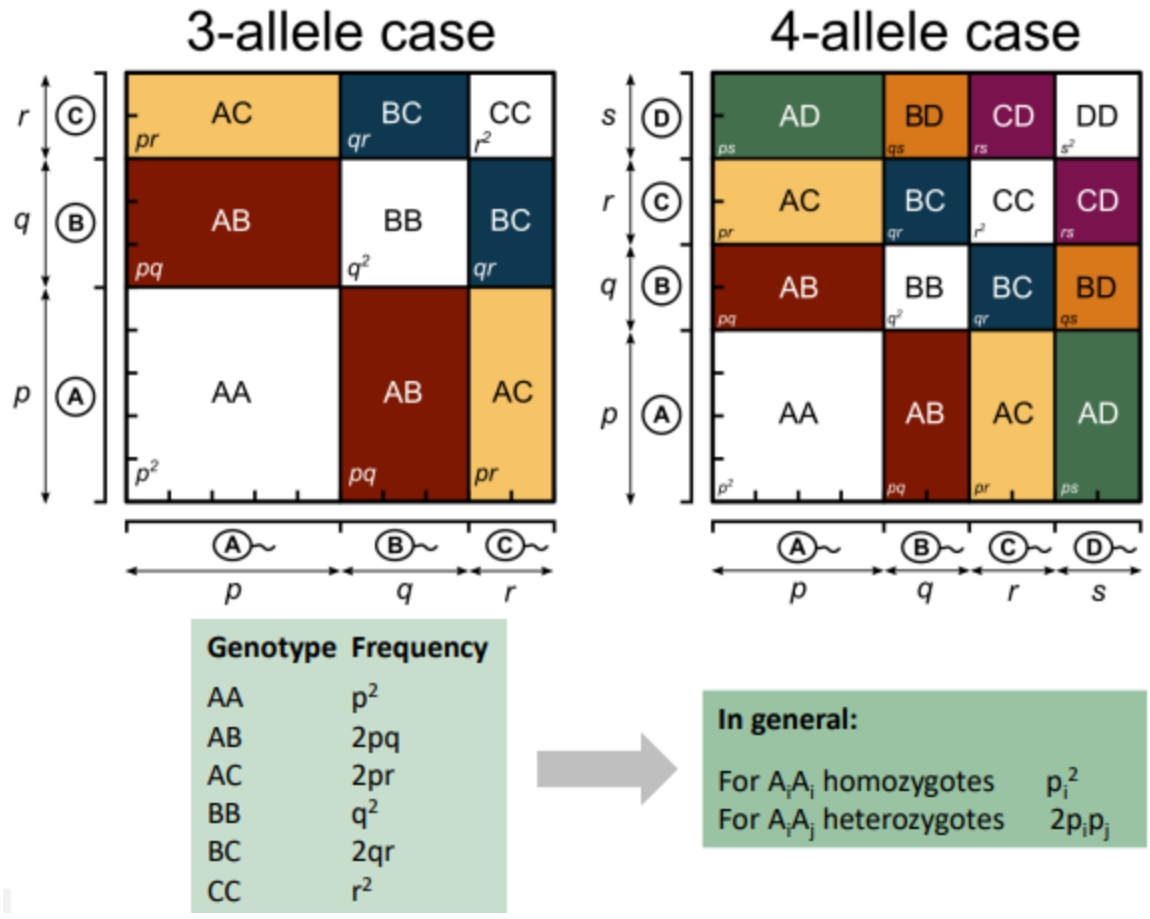
Generations needed to reach HWE

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.

So, it is really fast.

HWE with multiple alleles



All homozygotes will have the frequency of that allele elevated to 2.
All heterozygotes will have 2 times the frequency of the alleles involved.

How can we check if a population is in HWE or not?

We must test if the genotype frequencies adjust to the expected values. If they don't, one of the assumptions is not true.

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**

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

Estimating allele frequencies with dominance

The heterozygotes and dominant homozygotes will have the same phenotype.

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

$q = \sqrt{0.142} = 0.3768$ $\quad p = 1 - 0.3768 = 0.6232$

$2pq = 2 \cdot 0.3768 \cdot 0.6232 = 0.4697$

$p^2 = (0.6232)^2 = 0.3884$

PROPORTION OF HETEROZYGOTES WITHIN Rh+

$$\frac{0.4697}{0.4697 + 0.3884} = 0.547 = 54.7\%$$

Natural selection (NO HWE)

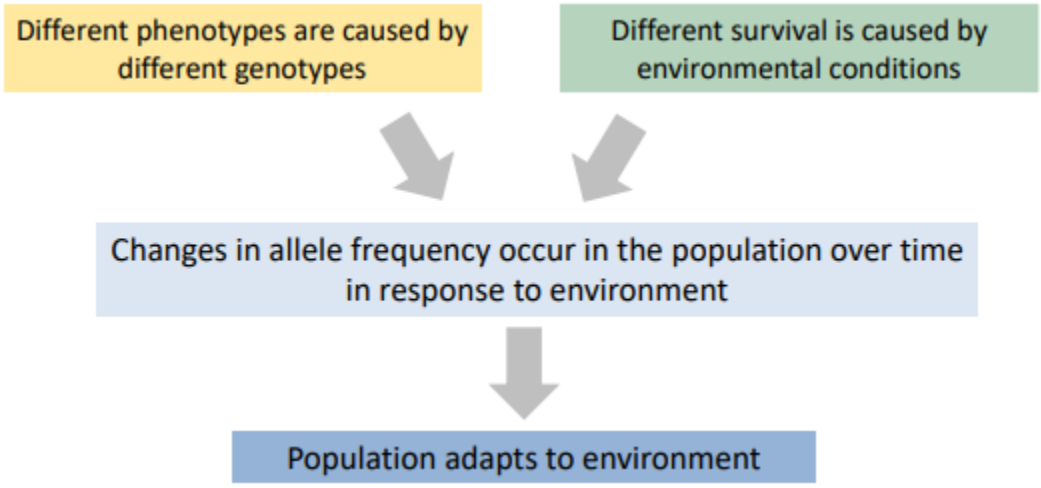
Process by which the genotypes that are superior in survival and reproduction will tend to leave more offspring than other genotypes, causing an increase in frequency of the favorable traits in the population over generations

Natural selection

- Requires existing heritable variation in a group
- Requires a causal relationship between genotype and number of offspring (the chance of surviving and having offspring)
- Depends on the environment. The alleles are good or bad depending on the context. One allele can be good in a certain scenario and bad in another one.
- Results in a greater adaptation of organisms to their environment over time

So, natural selection is going to cost a change in the allele frequencies in the populations. There are different evolutionary mechanisms that can make the allelic frequencies change, but only natural selection results in a greater adaptation of the organisms to the environment.

Selection can be reversible if the environment changes as long as none of the alleles have disappeared from the population.



Lluís Giménez Gabarró

Basic selection model

ASSUMPTIONS

Genetic system
- Single biallelic autosomal gene
- Diploidy

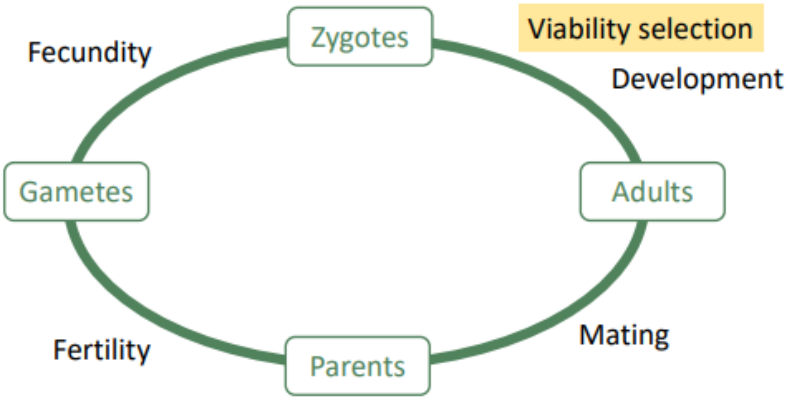
Selection
- Selection identical in both sexes
- Selection occurs through differences in viability
- Selection is constant through generations

Other factors
- Non-overlapping generations
- Random mating
- Large population size
- No mutation
- No migration

We have a cycle about what all individuals do.
From zygotes to adults we have a selection. Some genotypes will be more likely to become adults than others.

From adults to parents, we assume that all of them will be parents.

Not all individuals have the same number of gametes. But we assume that they all have the same probability and we also assume that all gametes have the same probability of becoming zygotes.



Fitness: Overall ability of an organism to survive and reproduce.

Absolute fitness (W): Measurement of the ability to survive of each genotype
Relative fitness (w): Ability of one genotype to survive to another genotype taken as reference.

Fitness depends on the environment

Example of mice

They have a polymorphism in MC1R gene.
Thus, they have 2 alleles, R and C

CC = Light
CR = Intermediate
RR = Dark

When they come from the continent to a new island, one of the variants is selected because of the camouflage. When they are on the continent, the RR genotype is selected but in the island the CC genotype is selected.

Fitness	$p_0 = 0.5$	Genotype		
		CC	CR	RR
Zygotes (N = 400)		100	200	100
Adults after selection (N = 160)		50	80	30
Absolute fitness (W)		50/100 = 0.5	80/200 = 0.4	30/100 = 0.3
Relative fitness (w) (to CC)		$w_{CC} = 0.5/0.5 = 1$	$w_{CR} = 0.4/0.5 = 0.8$	$w_{RR} = 0.3/0.5 = 0.6$
Relative fitness (w) (to CR)		$w_{CC} = 0.5/0.4 = 1.25$	$w_{CR} = 0.4/0.4 = 1$	$w_{RR} = 0.3/0.4 = 0.75$
Relative fitness (w) (to RR)		$w_{CC} = 0.5/0.3 = 1.67$	$w_{CR} = 0.4/0.3 = 1.33$	$w_{RR} = 0.3/0.3 = 1$

We can see the difference in number between the zygotes and adults after selection.
Compute the absolute fitness doing the ratio of the number of zygotes and adults.
Compute the relative fitness to CC, CR and RR dividing the absolute fitness of each genotype.

Freq(C) = p_0 = 0.3
Freq(R) = q_0 = 0.7

	Genotype			
	AA	Aa	aa	Total
Absolute fitness (W)	0.5	0.4	0.3	-
Zygotes (before selection)	$p^2 = 0.09$	$2pq = 0.42$	$q^2 = 0.49$	1
Adults (after selection)	$p^2 w_{AA} = 0.045$	$2pq w_{Aa} = 0.168$	$q^2 w_{aa} = 0.147$	$0.36 = \bar{w}$
Adults (normalized)	$\frac{p^2 w_{AA}}{\bar{w}} = 0.1250$	$\frac{2pq w_{Aa}}{\bar{w}} = 0.4667$	$\frac{q^2 w_{AA}}{\bar{w}} = 0.4083$	1

$\bar{w}$ = average fitness

$$p' = P + \frac{H}{2} = 0.36$$

$$q' = Q + \frac{H}{2} = 0.64$$

Absolute fitness values ten times smaller

	Genotype			
	AA	Aa	aa	Total
Absolute fitness (W)	0.05	0.04	0.03	-
Zygotes (before selection)	$p^2 = 0.09$	$2pq = 0.42$	$q^2 = 0.49$	1
Adults (after selection)	$p^2 w_{AA} = 0.0045$	$2pq w_{Aa} = 0.0168$	$q^2 w_{aa} = 0.0147$	$0.036 = \bar{w}$
Adults (normalized)	$\frac{p^2 w_{AA}}{\bar{w}} = 0.1250$	$\frac{2pq w_{Aa}}{\bar{w}} = 0.4667$	$\frac{q^2 w_{AA}}{\bar{w}} = 0.4083$	1

$$p' = P + \frac{H}{2} = 0.36$$

$$q' = Q + \frac{H}{2} = 0.64$$

Same values are obtained

We normalize so that the probabilities add to 1.

This will be the correct computation, because in the other case we were using the absolute fitness (they don't matter, it only tells us if the genotype adapts better than another):

	Genotype		
	CC	CR	RR
Relative fitness (w)	$w_{CC} = 0.5/0.5 = 1$	$w_{CR} = 0.4/0.5 = 0.8$	$w_{RR} = 0.3/0.5 = 0.6$

GENERATION 0

Freq(C) = $p_0 = 0.3$
Freq(R) = $q_0 = 0.7$

	Genotype			
	AA	Aa	aa	Total
Relative fitness (w)	1	0.8	0.6	
Zygotes (before selection)	$p^2 = 0.09$	$2pq = 0.42$	$q^2 = 0.49$	1
Adults (after selection)	$p^2w_{AA} = 0.09$	$2pqw_{Aa} = 0.336$	$q^2w_{aa} = 0.294$	$0.72 = \bar{w}$
Adults (normalized)	$\frac{p^2w_{AA}}{\bar{w}} = 0.1250$	$\frac{2pqw_{Aa}}{\bar{w}} = 0.4667$	$\frac{q^2w_{AA}}{\bar{w}} = 0.4083$	1

$p_0 = P + \frac{H}{2} = 0.36$

$q_0 = Q + \frac{H}{2} = 0.64$

GENERATION 1

	Genotype			
	AA	Aa	aa	Total
Zygotes (before selection)	$p^2 = 0.1296$	$2pq = 0.4608$	$q^2 = 0.4096$	1
Adults (after selection)	$p^2w_{AA} = 0.1296$	$2pqw_{Aa} = 0.3686$	$q^2w_{aa} = 0.2458$	$0.744 = \bar{w}$
Adults (normalized)	$\frac{p^2w_{AA}}{\bar{w}} = 0.1742$	$\frac{2pqw_{Aa}}{\bar{w}} = 0.4954$	$\frac{q^2w_{AA}}{\bar{w}} = 0.3304$	1

$p_1 = P + \frac{H}{2} = 0.42$

$q_1 = Q + \frac{H}{2} = 0.58$

	Genotype			
	AA	Aa	aa	Total
Relative fitness (w)	$w_{AA} = \frac{W_{AA}}{W_{AA}}$	$w_{Aa} = \frac{W_{Aa}}{W_{AA}}$	$w_{aa} = \frac{W_{aa}}{W_{AA}}$	Highest fitness value used to normalize
Zygotes (before selection)	p^2	$2pq$	q^2	1
Adults (after selection)	p^2w_{AA}	$2pqw_{Aa}$	q^2w_{aa}	$p^2w_{AA} + 2pqw_{Aa} + q^2w_{aa} = \bar{w}$
Adults (normalized) = p'	$\frac{p^2w_{AA}}{\bar{w}}$	$\frac{2pqw_{Aa}}{\bar{w}}$	$\frac{q^2w_{aa}}{\bar{w}}$	1

$\bar{w}$ = average fitness

$$p' = P + \frac{H}{2} = \frac{p^2w_{AA}}{\bar{w}} + \frac{pqw_{Aa}}{\bar{w}} = \frac{p^2w_{AA} + pqw_{Aa}}{\bar{w}} = \frac{p(pw_{AA} + qw_{Aa})}{\bar{w}}$$

$$q' = Q + \frac{H}{2} = \frac{q^2w_{aa}}{\bar{w}} + \frac{pqw_{Aa}}{\bar{w}} = \frac{q^2w_{aa} + pqw_{Aa}}{\bar{w}} = \frac{q(qw_{aa} + pw_{Aa})}{\bar{w}}$$

We can also compute the **fitness of an allele**

Allele A

$\bar{w}_A = p\,w_{AA} + q\,w_{Aa}$

Allele a

$\bar{w}_a = p\,w_{Aa} + q\,w_{aa}$

$$p' = \frac{p(pw_{AA} + qw_{Aa})}{\bar{w}} = p\frac{\bar{w}_A}{\bar{w}}$$

If we keep computing the next generations, we will see that allele RR will keep decreasing.

The average fitness will increase each generation, because the population is adapting. It will reach 1, meaning that it can not adapt any more.

Equilibrium

When we reach an equilibrium, the allele frequency (p) does not change.
Thus, the p in generation 10 will be equal to p in generation 29990

When will we reach equilibrium?

- When p = 0
- When Wa = W

We have 3 scenarios:

$$\Delta p = p' - p = p\frac{\bar{w}_A}{\bar{w}} - p = p\frac{(\bar{w}_A - \bar{w})}{\bar{w}}$$

$\bar{w}_A > \bar{w}$	Individuals with alleles A have an average fitness **HIGHER** than the average fitness of the population	$\Delta p > 0$	Frequency of allele A will **INCREASE**
$\bar{w}_A < \bar{w}$	Individuals with alleles A have an average fitness **LOWER** than the average fitness of the population	$\Delta p < 0$	Frequency of allele A will **DECREASE**
$\bar{w}_A = \bar{w}$	Individuals with alleles A have **THE SAME** fitness than the population	$\Delta p = 0$	No frequency change

Directional selection: Selection will cause the fixation of advantageous alleles and the loss of deleterious alleles. This is what we have been talking about during the class.

This directional selection can be seen in 2 different ways:

- **Purifying selection:** A new allele has a negative effects and is selectively removed from the population
- **Positive selection:** A new allele has advantageous effects and is selectively fixed.

Balanced selection: Maintenance of both alleles in the population that occurs when the heterozygote genotype has the higher fitness.

Lactase Persistence

In this case, it will not affect the fitness of the population because it is not relevant to survive. It was relevant at a certain moment, but not when being an adult.

In the case of adaptation to high altitude, there are many ways in which we can adapt to that environment (Andes and Tibetan).

Lluís Giménez Gabarró

Selection coefficient (s)

Reduction in fitness of a given genotype as compared to another.

EXTREME VALUES

w = 1 → s = 0 → No selection

w = 0 → s = 1 → Lethal

$w_{AA} = 1$

$w_{aa} = 1 - s$

$w_{Aa} = 1 - ?$

In heterozygotes it depends on the relationship between both alleles, as we are going to see.

Degree of dominance (h)

Parameter that modulates the selection coefficient in heterozygotes depending on the dominance relationship of the alleles.

$w_{Aa} = 1 - hs$

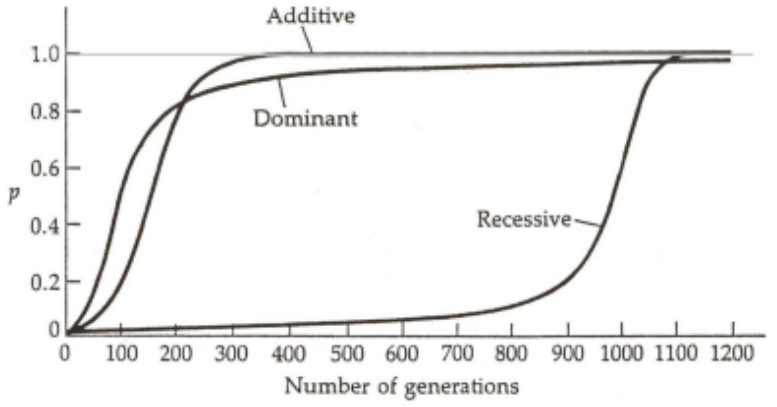
Favorable allele = A

Dominance	h	Fitness			Fitness in heterozygotes
		AA	AB	BB	
Dominant	**0**	**1**	**1**	**1 – s**	Same as AA
Recessive	**1**	**1**	**1 – s**	**1 – s**	Same as BB
Additive	**1/2**	**1**	**1 – s/2**	**1 – s**	Intermediate

Changes in frequency of a favored allele

Since selection acts on phenotypes, the rate of change in allele frequency (the time needed for fixation or loss of an allele) will depend on how phenotypes are related to genotypes. Alleles can be invisible to selection

s = 0.05



p = 1 Fixation

p = 0 Loss

Favored allele	Slowest rate of change	Reason
Dominant	When allele is common	Recessive alleles hidden in heterozygotes
Recessive	When allele is rare	No homozygotes with high fitness

So, in the last diagram we are looking at the rate in which an allele is lost or fixed.

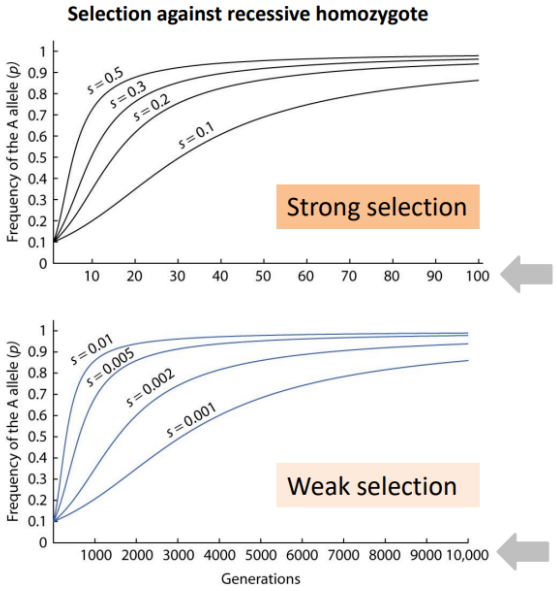
- In the case of a **dominant** allele, its frequency will increase very fast but at a certain moment it will stop increasing as fast as it was. This is because the recessive one can “hide” in the heterozygotes.
- In the case of a **recessive** allele, its frequency will increase really slowly. Because there are a small number of alleles “a” and the only genotype that is selected is “aa”. But once the frequency of the allele “a” increases, then it will start growing exponentially and reach p= 1. Because in this case, dominant alleles can not hide anywhere.
- In the case of an **additive** allele, it will reach p=1 really fast. Because the genotype AA adapts better than Aa and this one adapts better than aa. Thus, the recessive allele can not hide and the allele A is always selected. Here the selection is very effective.

The strength of natural selection

The rate of change in allele frequencies depends on the strength of natural selection. So, depending on the value of the selection coefficient, we will obtain p=1 faster or slower.

Because if the “s” for an allele is big, then its fitness will be really small and the allele will be purified.

Thus, it depends on how relevant the allele is.



Overdominance or heterozygote superiority

The heterozygote has a fitness of 1 and the homozygotes have a lower fitness (different selection coefficients).

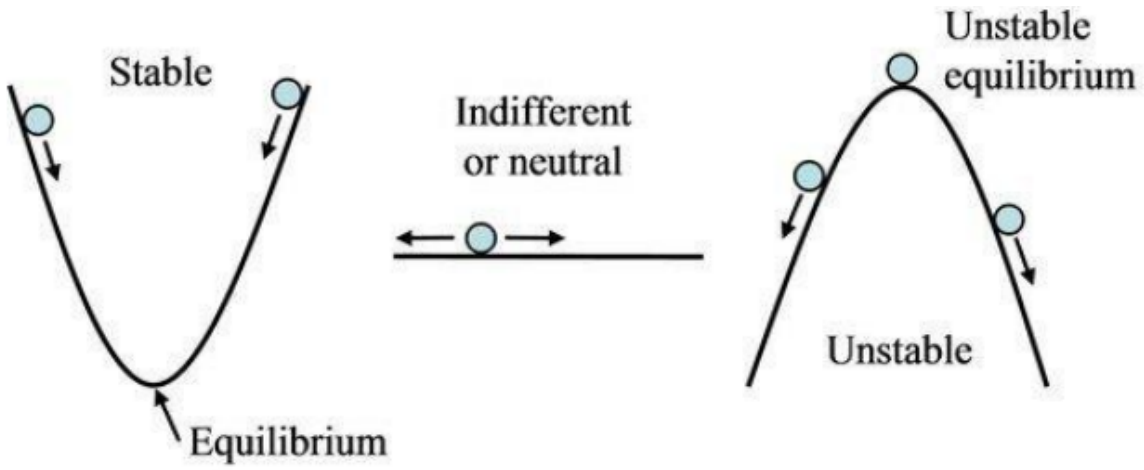
None of the two alleles can be fixed in the population (because the heterozygote has both alleles), but we can reach an equilibrium in which allele frequencies do not change across generations.

Equilibrium

$$\Delta p = 0 \quad \Rightarrow \quad \hat{p} = \frac{s_2}{s_1 + s_2}$$

If we know the selection coefficients, we will know the frequency of the alleles.

Types of equilibrium



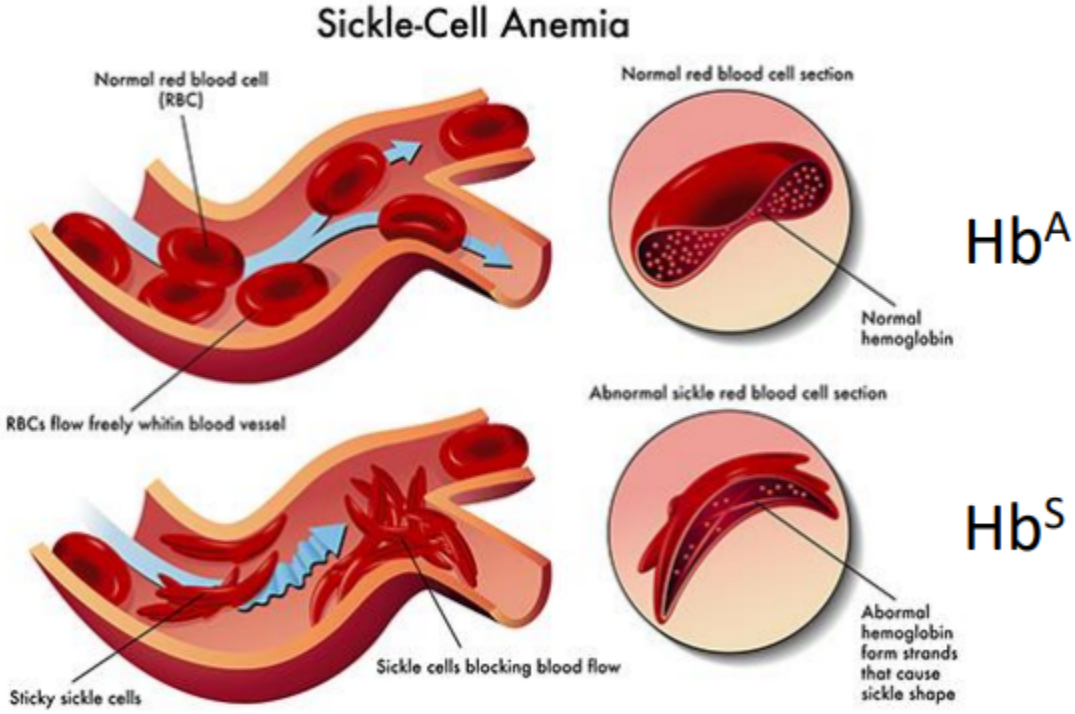
If there is an heterozygote superiority, it will reach an stable equilibrium, where the allele frequencies converge to an equilibrium value irrespective of initial frequencies.

It will have a maximum of the average fitness equal to 0.667.
It would be 1 if everyone was an heterozygote.

Sickle-cell anemia

Disease caused by an allele (a single aa change) in the hemoglobin gene:

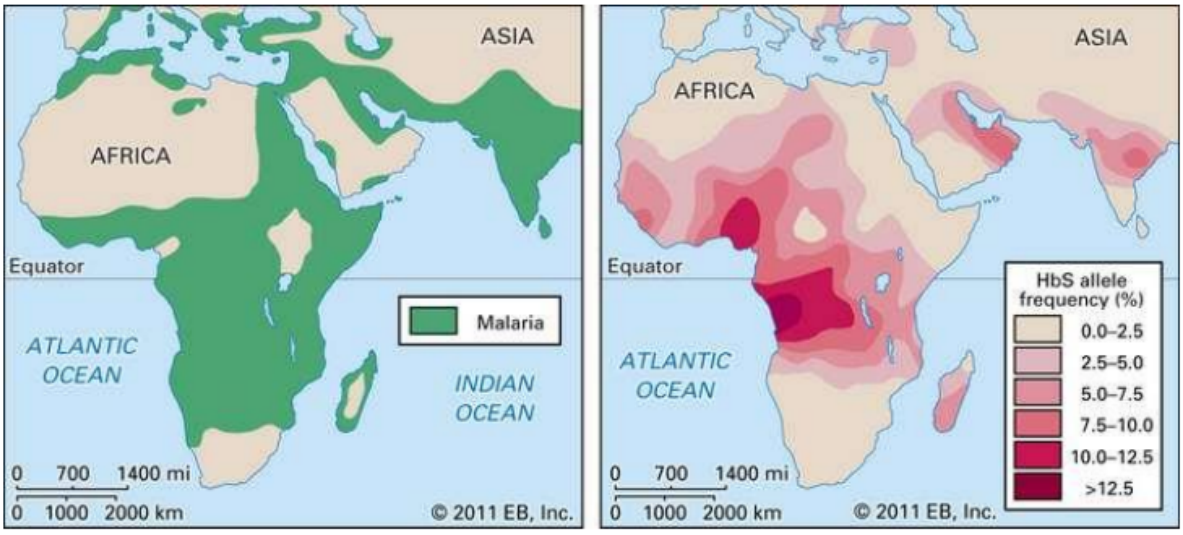
- A is the wild type allele
- S is the disease allele



The red blood cells change their shape and block the blood flow.
These cells tend to have a shorter life and the body can not replace them fast enough.

Persistence of this allele has been explained by the fact that heterozygous persons are resistant to malaria.
Because when red sickle cells are invaded by the malarial parasite, they adhere to blood vessel walls, become deoxygenated and then they are destroyed (also the parasite).
In regions where malaria is endemic, the heterozygote is the fittest genotype because it is resistant to malaria and does not suffer the disease.

The homozygotes suffer malaria, the other homozygotes suffer anemia and the heterozygotes suffer a small anemia.



Warfarin resistance in rats

Warfarin is an anticoagulant used as a drug (for humans) and venom for rats.
Warfarin inhibits the enzyme VKORC1, which reduces vitamin k (essential to activate blood coagulation factors).

This enzyme has 2 alleles:

- One that is affected by the warfarin
- The other is not affected so much. Thus, warfarin is not able to anticoagulate your blood (because you will have vitamin k). But, you will need a lot of vitamin k because this allele is not as effective as the other in normal conditions.

So, we have 2 escenarios. Here we can see the fitness of the genotypes when there is warfarin in the environment and when there is no warfarin.

Condition	SS	RS	RR
WITHOUT WARFARIN	1	0.77	0.46
WITH WARFARIN	0.68	1	0.37

In normal conditions:

- Genotype SS is the fittests because it does not require too much vitamin k.
- Genotype RS has a lower fitness because it needs more vitamin k
- Genotype RR has the lowest fitness because it needs a lot of vitamin k

In warfarin conditions:

- Genotype RS is the fittest because it's resistant to warfarin
- Genotype SS has a lowest fitness because it is not resistant to warfarin
- Genotype RR has the lowest fitness because it needs a lot of vitamin k

Underdominance or heterozygote inferiority

When the heterozygote has the lowest fitness.

Polymorphism can be maintained under certain equilibrium allele frequencies, but any deviation from these frequencies will lead to fixation of one of the alleles, so it is an extremely rare phenomenon.

So, this is a case of unstable equilibrium.

General categories of relative fitness values (summary)

	Genotype fitness		
Category	w_{AA}	w_{Aa}	w_{aa}
Selection against recessive phenotype	**1**	**1**	**1 – s**
Selection against dominant phenotype	**1 – s**	**1 – s**	**1**
Intermediate dominance ($0 \leq h \leq 1$)	**1**	**1 – hs**	**1 – s**
Heterozygote advantage	**$1 - s_1$**	**1**	**$1 - s_2$**

Mutation facts

It's a permanent change in an organism's DNA (from nucleotide substitutions to large structural variants). It is the source of all genetic variation.

It's the result of unrepaired damage in DNA and errors during DNA replication or repair

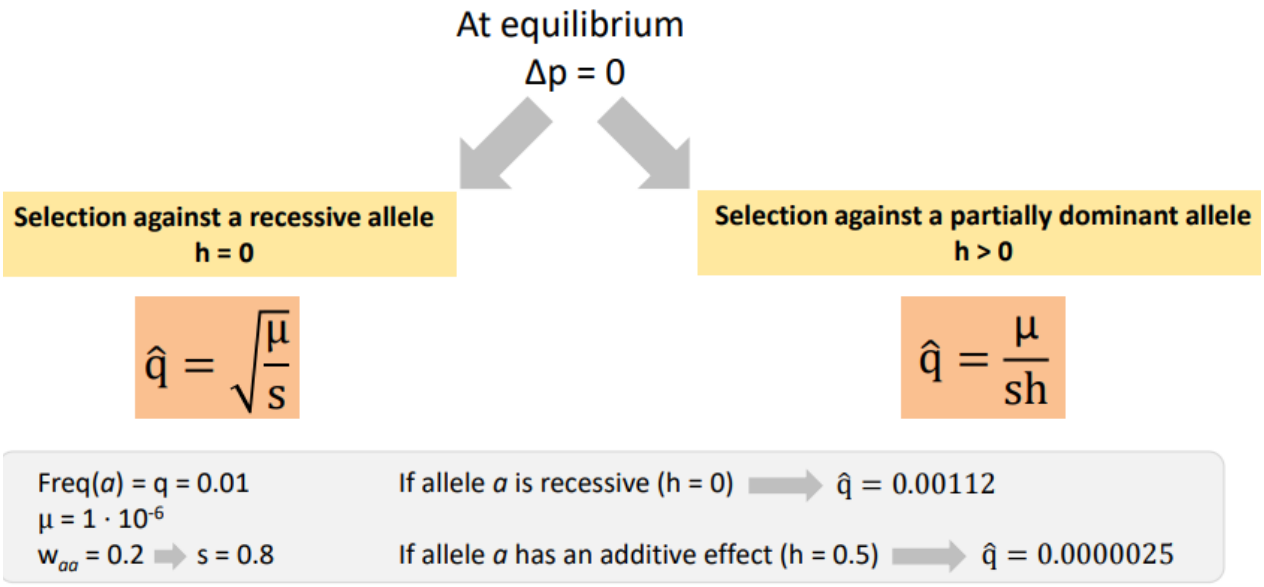
Mutation-selection balance

Mutation will generate non-functional alleles with negative effects on fitness.

We will have 2 forces acting on this allele:

- Mutation that is creating this non-functional allele (a)
- Purifying selection that eliminates the non-functional allele.

We will reach an equilibrium where we will have a constant frequency.



mu is the mutation rate.

If the bad allele is recessive, we will have a certain frequency of that allele.
If it is additive, we will have a smaller frequency (because in this case selection acts directly in the genotype). They can not hide…

This is why genetic diseases exist!

Lluís Giménez Gabarró

Genetic Drift

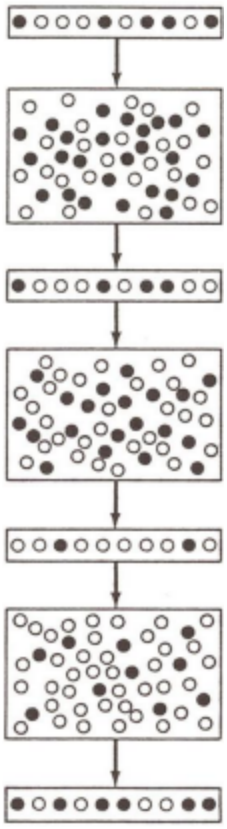
The allele frequencies can change by chance (genetic drift). Similar to a boat that has no capitan and thus it drifts.

Assumptions

- We will use a small population. This is not assumed in the HWE

When we were talking about natural selection, we made emphasis on the genotypes. But in this case, we will forget about the genotypes and we will think of our population as a group of alleles.

Example: We have a population with 10 individuals. 5 of them are white and 5 black.
After mating, we select 10 gametes at random. It could happen that we select 6 white individuals and 4 black individuals.



Thus, the proportion of alleles has changed and we do another round of mating. If we do the same experiment again and again, the white individuals will replace all black individuals because of the random variation of the allele frequency.

This is due to the sampling error (selection of the gametes that are going to represent the alleles in the next population).

If there is no drift, we will always select 5 white individuals and 5 black individuals.

Genetic drift is a stochastic process, we can not tell what is going to happen. We can just tell the probability of an event happening.

Natural selection is a deterministic process (if we put the same conditions, we will obtain the same result).

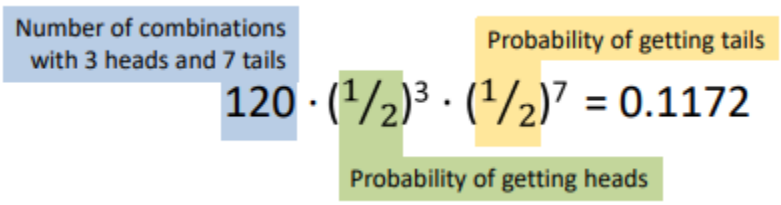
How can we calculate the probability of the next results?
We will use the binomial distribution when:

- There are 2 possible outcomes of a trial
- The probability of each outcome remains the same across all trials
- All trials are independent of each other (the result I obtain the first time does not affect the result of the other experiments)

The probability of getting exactly *k* successes with *p* probability in *n* trials is:

$$P = \frac{n!}{k!\,(n-k)!}\,p^k\,(1-p)^{n-k}$$

How to compute the probability of getting 3 heads and 7 tails if I flip a coin 10 times?



To compute the number of combinations I will do the factorial. If I want to consider the order, I don't have to compute all the combinations.

We can compute the probability of all possible combinations and we will see that it is not likely that we obtain 5 white and 5 black individuals. In fact, 75% of the time we will obtain a different allele frequency.

Possible results	Combinations	Probability
10 white	1	0.0009765
9 white + 1 black	10	0.009765
8 white + 2 black	45	0.043945
7 white + 3 black	120	0.1172
6 white + 4 black	210	0.2051
5 white + 5 black	252	0.2461
4 white + 6 black	210	0.2051
3 white + 7 black	120	0.1172
2 white + 8 black	45	0.043945
1 white + 9 black	10	0.009765
10 black	1	0.0009765

The binomial formula adapted to the population genetics, is the following:

$$P = \frac{2N!}{k!\,(2N-k)!}\, p^k\, q^{2N-k}$$

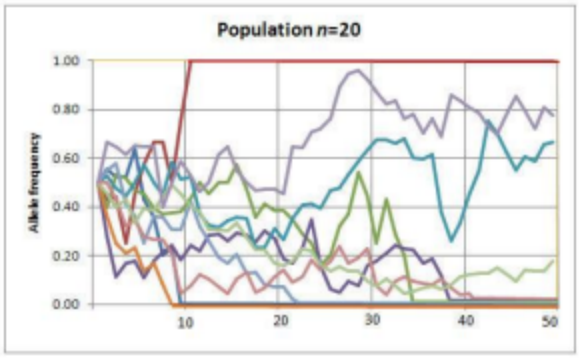
How will the size of the population affect this process?
We can make the allele distribution and compute confidence intervals of the possible values of the allele frequency "p". Consider that p = 0.5 and SD = 0.2

- If the population has 10 alleles, the CI = [0.184 - 0.816]
- If the population has 100 alleles, the CI = [0.4 - 0.6]
- If the population has 1000 alleles, the CI = [0.468 - 0.532]

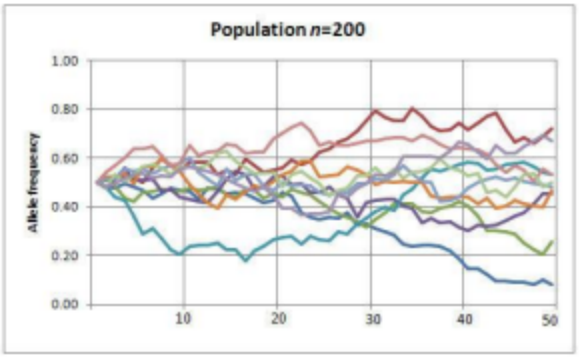
Thus, if the population size is small, the genetic drift will be more relevant.

Allele frequencies will change by chance in populations of all sizes, but the amount of change due to sampling error decreases as population size increases.

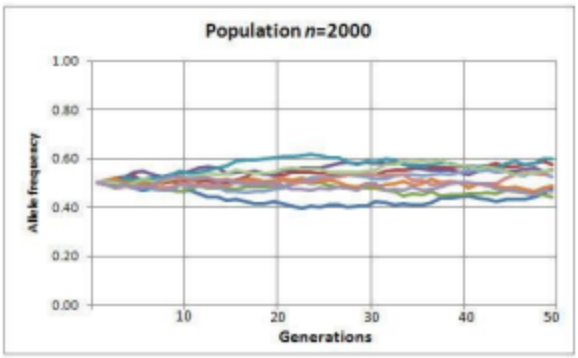
When we represent what is going to happen, to the allele frequencies, for a number of generations, due to genetic drift, we will use this type of graph.
The different colors represent different populations that evolve.



2N = 40



2N = 400



2N = 4000

When the allele frequency reaches a value of 1, it means that one of the alleles has been fixed and the other lost (it will not change any more). This normally happens in small populations, since genetic drift is more relevant.

Lluís Giménez Gabarró

Wright-Fisher model for genetic drift

Model to predict what is going to happen due to genetic drift.

The Wright-Fisher model describes the sampling of alleles (A and a) in a population with no selection, no mutation, no migration, non-overlapping generation times and random mating in a constant size population.

We have to imagine that we have a lot of different populations that are evolving by drift. Meaning that, out of N individuals of one generation, we obtain an infinite pool of gametes and I select the 2N alleles randomly. So, we are making the population evolve by drift.

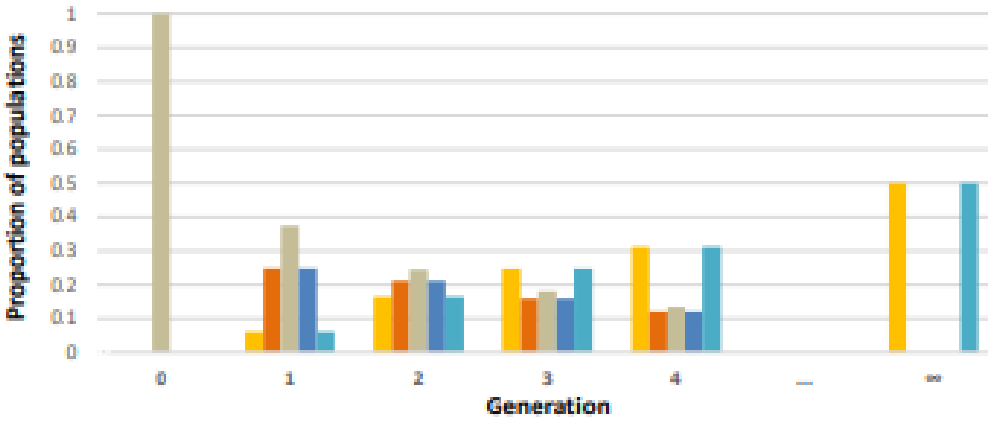
Imagine we have the smallest population (2 individuals) that is only affected by genetic drift. At the beginning, all of our populations have a frequency of 0.5 and all the possible combinations for the next generation are:

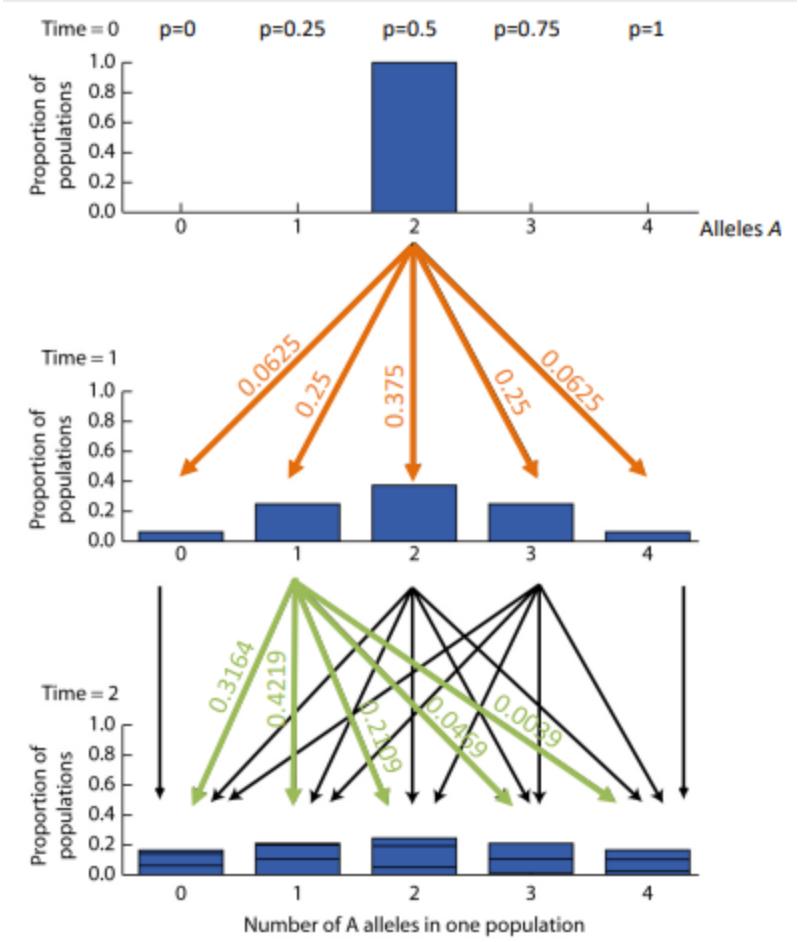
N = 2 individuals 2N = 4 alleles $p_0 = q_0 = 0.5$

A alleles	a alleles	Probability
0	4	$1 \cdot (0.5)^0 \cdot (0.5)^4 = 0.0625$
1	3	$\frac{4!}{1!\,3!} \cdot (0.5)^1 \cdot (0.5)^3 = 0.25$
2	2	$\frac{4!}{2!\,2!} \cdot (0.5)^2 \cdot (0.5)^2 = 0.375$
3	1	$\frac{4!}{3!\,1!} \cdot (0.5)^3 \cdot (0.5)^1 = 0.25$
4	0	$1 \cdot (0.5)^4 \cdot (0.5)^0 = 0.0625$

We compute this using the binomial distribution. Thus, the probability of losing A in the next generation is 6.25%.

In the second generation, we can compute the new probabilities using the binomial distribution. Note that the populations that have fixed one of the alleles, will not change in the next generation. Thus, the allele frequencies tend to be fixed or lost.



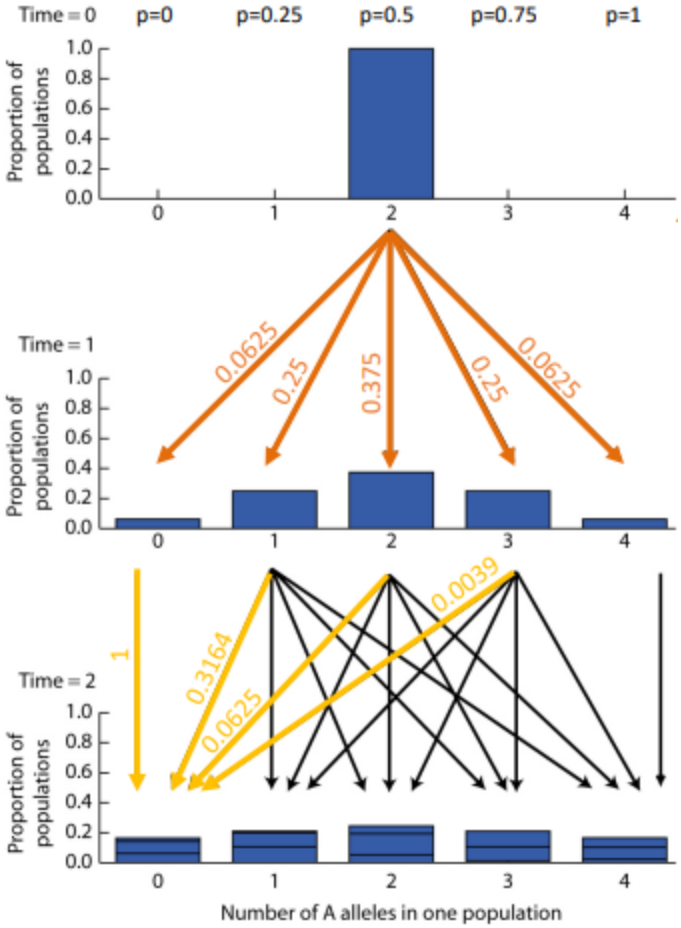


Probability transition matrix for a population of size 2N = 4

Number of alleles *A* in generation t

Number of alleles *A* in generation t+1	0	1	2	3	4
0	1	0.3164	0.0625	0.0039	0
1	0	0.4219	0.25	0.0469	0
2	0	0.2109	0.375	0.2109	0
3	0	0.0469	0.25	0.4219	0
4	0	0.0039	0.0625	0.3164	1

What is the probability of having 0 alleles A in generation 2. We will have to look at all scenarios.



Probability transition matrix for a population of size 2N = 4

Number of alleles *A* in generation t

Number of alleles *A* in generation t+1	0	1	2	3	4
0	1	0.3164	0.0625	0.0039	0
1	0	0.4219	0.25	0.0469	0
2	0	0.2109	0.375	0.2109	0
3	0	0.0469	0.25	0.4219	0
4	0	0.0039	0.0625	0.3164	1

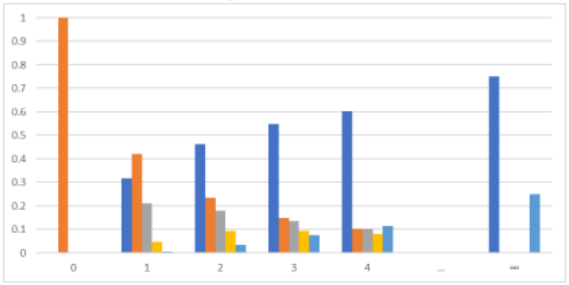
$$P(0 \text{ alleles A in generation } 2) = (0.0625 \cdot 1) + (0.25 \cdot 0.3164) + (0.375 \cdot 0.0625) + (0.25 \cdot 0.0039) = 0.166$$

An increasing number of populations accumulate at states of 0 and 4 alleles A, eventually reaching fixation or loss for all populations.
Mean frequency does not change with time
Mean heterozygosity decreases with time, because the alleles tend to be fixed.
Variance increases with time.

Number of alleles A \ Generation	0	1	2	3	4	...	∞
0 (p=0)	0	0.0625	0.1660	0.2490	0.3117	...	0.5
1 (p=0.25)	0	0.25	0.2109	0.1604	0.1205	...	0
2 (p=0.5)	1	0.375	0.2461	0.1813	0.1356	...	0
3 (p=0.25)	0	0.25	0.2109	0.1604	0.1205	...	0
4 (p=1)	0	0.0625	0.1660	0.2490	0.3117	...	0.5
$\bar{p}$	0.5	0.5	0.5	0.5	0.5	...	0.5
$\bar{H}$	0.5	0.375	0.2812	0.2109	0.1582	...	0
Var(p)	0	0.0625	0.1094	0.1445	0.1709	...	0.25

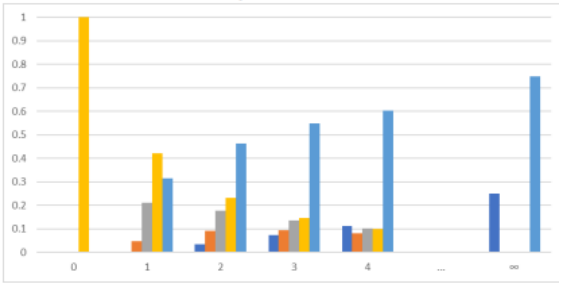
What happens if we start with an allele frequency of 0.25 and 0.75?
In this case, we will fix 25% and 75% of the time for each respective allele.
So, the probability of fixation of a neutral allele is equal to its initial frequency in the population.

$p_0 = 0.25$



	0	1	2	3	4	...	∞
0 (p=0)	0	0.3164	0.4633	0.5484	0.6038	...	0.75
1 (p=0.25)	1	0.4219	0.2329	0.1471	0.1003	...	0
2 (p=0.5)	0	0.2109	0.1780	0.1353	0.1017	...	0
3 (p=0.75)	0	0.0469	0.0923	0.0944	0.0805	...	0
4 (p=1)	0	0.0039	0.0336	0.0748	0.1137	...	**0.25**

$p_0 = 0.75$



	0	1	2	3	4	...	∞
0 (p=0)	0	0.0039	0.0336	0.0748	0.1137	...	0.25
1 (p=0.25)	0	0.0469	0.0923	0.0944	0.0805	...	0
2 (p=0.5)	0	0.2109	0.1780	0.1353	0.1017	...	0
3 (p=0.75)	1	0.4219	0.2329	0.1471	0.1003	...	0
4 (p=1)	0	0.3164	0.4633	0.5484	0.6038	...	**0.75**

All alleles started as a single allele in the population. Meaning that all alleles are created by mutations and this mutation only happens in one chromosome. So, for every new allele created in a population, its initial frequency is p = 1/2N. Thus, it is more likely to be lost than fixed.

Genetic drift causes a reduction in heterozygosity

Heterozygosity = observed proportion of heterozygotes in a population.

If we have a high variance, we will have a lower proportion of heterozygotes. If we have a low variance, we will have a higher proportion of heterozygotes.
One of the consequences of evolving by drift is that we will increase variance and thus we will reduce the proportion of heterozygotes in the population.

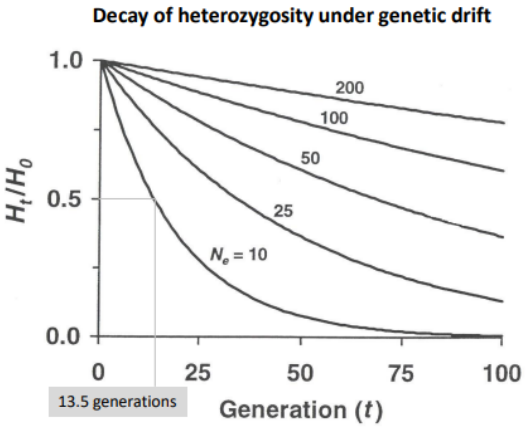
Heterozygosity declines by a factor of 1 - (1/2N) every generation due to drift.

$$H_t = H_0\left(1-\frac{1}{2N}\right)^t \quad = \quad \frac{H_t}{H_0} = \left(1-\frac{1}{2N}\right)^t$$

H0 is the initial proportion of heterozygotes.
When one allele is fixed, we will have H=0

So, with this formula we can compute the proportion of heterozygotes that you are going to have in a given generation depending on the size of the population.

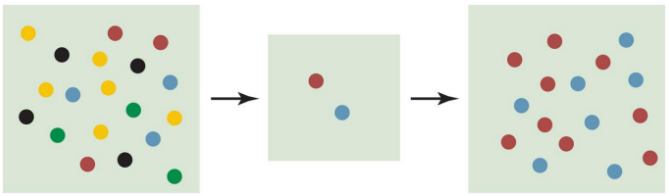
The smaller the population, the faster we will lose heterozygosity.
In a large population you also lose heterozygosity but slower.



Reductions in population size

Genetic drift acts more quickly to reduce genetic variation in small populations
The resulting population can have (when there is a reduction):

- Reduced variation and reduced ability to adapt to new selection pressures. The remaining alleles do not represent the variation that we had before.

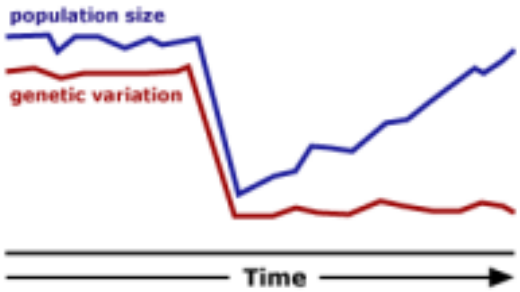


  Although we can recover the size of the population, we can not recover the variation that we had. All the new individuals will come from a small variance population (thus they will be “copies”).

- A non-random sample of the genes in the original population

This reduction in size can happen in 2 different ways.
Bottleneck: Occurs when a population's size is reduced for at least one generation

- Consequences: Long term reduction of genetic variation



Example of species that are in danger of extinction. We can increase the number of individuals, but not the genetic variation. They will all be copies of each other.

Founder effect: Occurs when a new colony is started by a few members of the original population.

- Consequences: Substantial loss in genetic diversity and rapid divergence between source and founder populations

Example of the founder effect: Huntington's disease in Venezuela

Huntington's disease is inherited as an autosomal dominant trait and causes degeneration of nerve cells in the brain ultimately leading to death.

In Venezuela this is a very common disease.

It is caused by a protein that has a series of repeated aa.

- If this repeated region is short (10 to 26 repetitions), then the protein is normal.
- If this repeated region is medium (27 to 39 repetitions), then the individual is a carrier.
- If this repeated region is large (more than 40 repetitions), then the individual has the disease. The protein accumulates in the brain (prionic) and causes the disease. This is why it is a dominant trait, since if one allele is affected then you will have the disease.

How is it possible that a disease that is so bad for the people is still present in Venezuela?
It is due to 3 reasons:

- **Founder effect:** One person of the people that established in this region was a carrier. This woman had many children. So, the founder effect changed the frequency of this allele (it is rare in the rest of the world).
- **Mutation:** This gene has a high mutation rate. It is very easy to have multiple repetitions by error.
- **Weak selection:** This disease does not affect people until after they have reproduced. So, for natural selection, this is not a bad gene and thus it does not purify it.

Effective population size (Ne)

We said that the population size is the most important factor in genetic drift. But this is not true, the important thing is the size of the population that actively participates in the reproductive process.
Size of an idealized population that would have the same effect of random sampling on allele frequencies as that of the actual population.

Census population (Nc)

Total number of individuals in a population, we don't care if they actively participate in the reproductive process or not.

Ideal population (Nc = Ne)

- There are equal number of males and females, all of whom are able to reproduce
- All individuals are equally likely to produce offspring, and the number of offspring that each produces varies no more than expected by chance
- Mating is random
- The number of breeding individuals is constant from one generation to the next.

Ne is usually much smaller than Nc. Factors that contribute to this difference:

- Different number of males and females
- Fluctuations in population size
- Variation in the number of offspring among individuals. Not because some individuals are better but just by chance.
- Bottlenecks
- Overlapping generations

When there is a different number of males and females

This happens when 1 male controls a group of females, for example. The other males are not participating.
Ne in a population unequal sex ratio for autosomic genes:

$$N_e = \frac{4N_m N_f}{N_m + N_f}$$

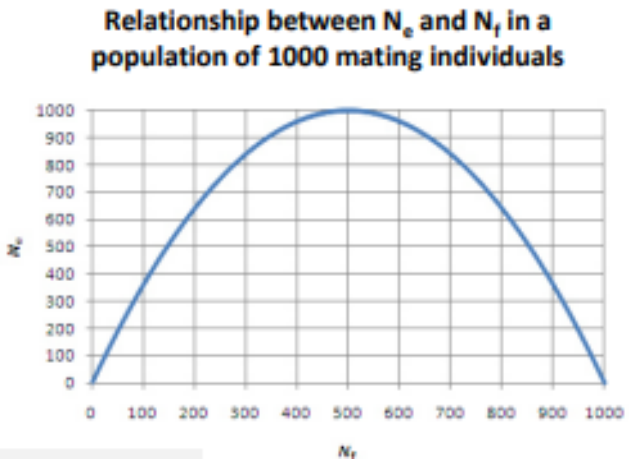
Autosomic genes

Nm = number of males
Nf = number of females

What is the effective population size of a honey bee hive?

$N \approx 100.000$
$N_f = 1$
$N_m >>> 1$

$N_e = 4$



When there are fluctuations in population size
Populations can show regular cycles of increase and decrease spanning a number of years (prey and predator relationship).

Small population numbers will cause an increased chance of fixation or loss of alleles by genetic drift.
We can estimate the effect of fluctuations in populations on the overall effective size using the harmonic mean, which gives more weight to small values.

The moments where the population size is small are the important ones. They lose variance, which can not be recovered.

$$\frac{1}{N_e} = \frac{1}{t}\left(\frac{1}{N_0} + \frac{1}{N_1} + \cdots + \frac{1}{N_t}\right)$$

What is the effective population size of a population with 100 individuals that was reduced to 10 for 1 generation but has recovered its original census in the following generation?

$N_e = 25$

Low effective population sizes (Ne) in humans

Table 1 | **Effective population size (N_e) estimates from DNA sequence diversities**

Species	N_e	Genes used	Refs
Species with direct mutation rate estimates			
Humans	10,400	50 nuclear sequences	145
Drosophila melanogaster (African populations)	1,150,000	252 nuclear genes	108
Caenorhabditis elegans (self-fertilizing hermaphrodite)	80,000	6 nuclear genes	41
Escherichia coli	25,000,000	410 genes	146
Species with indirect mutation rate estimates			
Bonobo	12,300	50 nuclear sequences	145
Chimpanzee	21,300	50 nuclear sequences	145
Gorilla	25,200	50 nuclear sequences	145
Gray whale	34,410	9 nuclear gene introns	147
Caenorhabditis remanei (separate sexes)	1,600,000	6 nuclear genes	43
Plasmodium falciparum	210,000 –300,000	204 nuclear genes	148

For data from genes, synonymous site diversity for nuclear genes was used as the basis for the calculation, unless otherwise stated.

It's strange that the Ne of humans is 10.000 since the total population is more than 7.000 milions. This is due to the fact that we are taking into consideration the historical world population and in the beginning there was a really small population size. Only in the last few years the population has increased a lot.

Important ideas

- **Allele frequencies change randomly due to sampling error**
- **The direction of the change is unpredictable** (allele frequencies will randomly increase and decrease over time)
- **Cumulative behavior** (each generation allele frequency will tend to deviate more and more from initial frequency and probability of fixation increases with time)
- **The amount of change due to sampling error decreases as the population size increases** (smaller populations will be more affected by genetic drift than larger populations)
- Given enough time and in the absence of factors that maintain both alleles, **one allele will drift to fixation and the other will drift to extinction**
- **The probability of fixation of an allele is equal to its initial frequency**
- **Heterozygosity will decrease over time in a finite population** (it will eventually become 0 when an allele is fixed)
- **Effective population size (N_e)** will determine the effect of genetic drift in a population instead of census size

Mutation

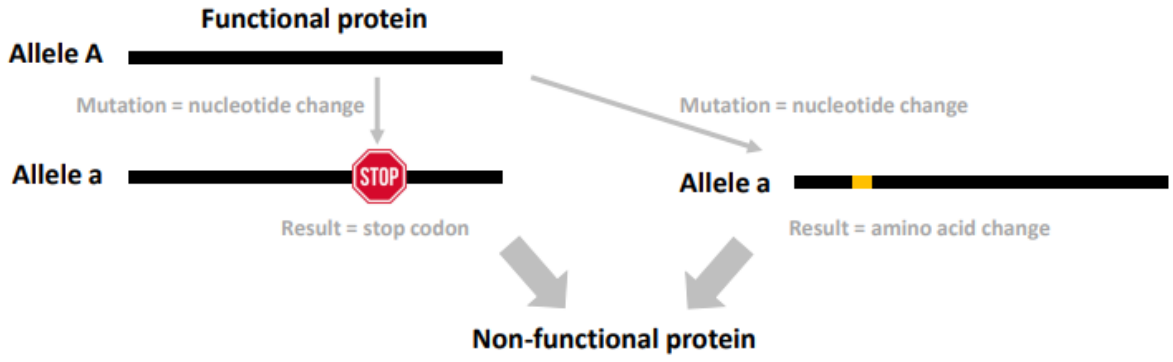
In this case we will not assume a No mutation (assumption required in the HWE)

- A mutation is any permanent change in the organism's DNA (from nucleotide substitution to large structural variants) and is the result of unrepaired damage in DNA and errors during DNA replication or repair.

- Mutation is the source of all genetic variation. Mutation introduces new alleles in populations.

- Mutations in the germinal line are transmitted to offspring but somatic mutations are not.

- At phenotypical level, mutation can be considered recurrent. At the molecular level most mutations are unique. Meaning that many mutations can produce the same phenotype (example of a non-functional protein).



Irreversible mutation

In each generation, a proportion of the allele A will be transformed into an allele a.

$$A \xrightarrow{\mu} a$$

So, in the next generation we will have the following proportion of allele A

$$p_{t+1} = p_t\,(1-\mu)$$

If we keep computing the frequency of allele A in the other generations, we will obtain the following formula:

$$p_{t+2} = p_{t+1}\,(1-\mu)$$

$$p_{t+2} = p_t\,(1-\mu)\,(1-\mu)$$

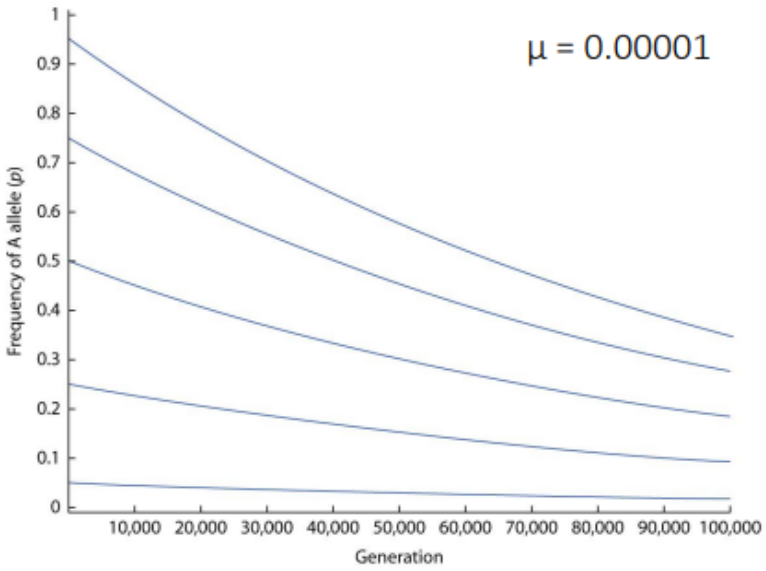
$$p_{t+2} = p_t\,(1-\mu)^2$$

$$p_{t+3} = p_t\,(1-\mu)^3$$

$$\vdots$$

$$\mathbf{p_t = p_0\,(1-\mu)^t}$$

The mutation rate is really small. Mutation is extremely slow compared to genetic drift, as we can see in the figure.



When $t \to \infty$ $\quad p_t \to 0$ $\quad q_t \to 1$

Lluís Giménez Gabarró

Reversible mutation

It can happen that a also transforms into A. Thus, there is an equilibrium where the allele frequencies will not change.

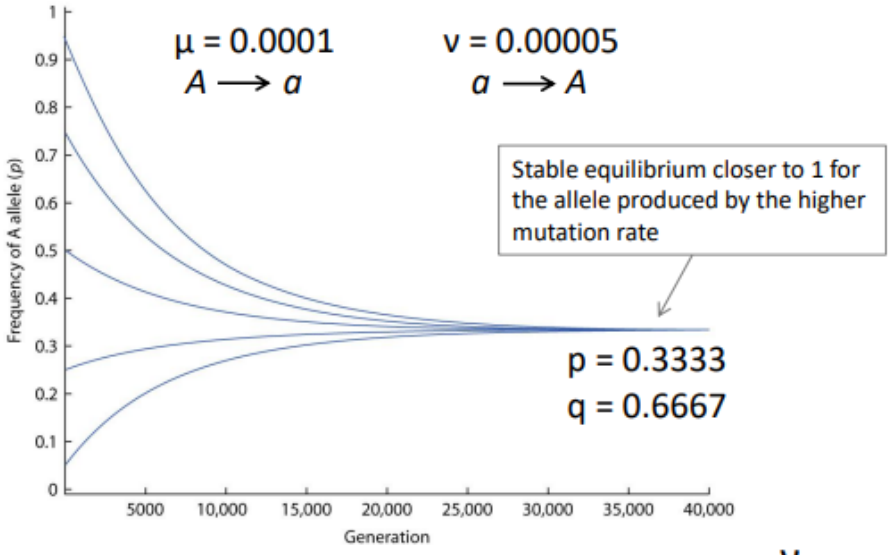
$$A \underset{\nu}{\overset{\mu}{\rightleftharpoons}} a$$

The formula to compute the frequency of allele A is:

$$p_t = \frac{\nu}{\mu+\nu} + \left(p_0 - \frac{\nu}{\mu+\nu}\right)(1-\mu-\nu)^t$$

In equilibrium

$$\hat{p} = \frac{\nu}{\mu+\nu}$$



When $t \to \infty$ $\quad (1-\mu-\nu)^t \to 0$ $\quad p_t \to \frac{\nu}{\mu+\nu}$

The rates of mutation from wild type to a novel allele (forward mutations) are nearly a factor of 10 more common than mutations from a novel allele to a wild type (reverse mutations)

This asymmetry occurs because there are more ways mutation can cause a normal allele to malfunction than there are ways to exactly restore that function once it is disrupted.

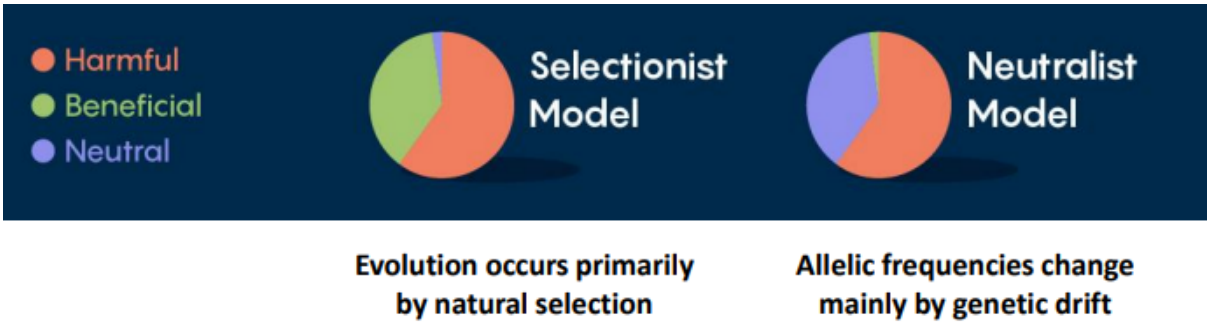
Lluís Giménez Gabarró

Natural theory of molecular evolution (Kimura)

Out of all the variants that we have, how many of them have evolved by selection and how many of them have evolved by genetic drift (2 most important in evolution).

Selectionists vs neutralists models

For many years, biologists have argued about whether genome evolution depends more on natural selection or on genetic drift.
Selectionist and neutralist models of evolution predict different proportions of beneficial and neutral mutations.



Infinite alleles model

Kimura proposed this model to solve the question. We have already seen that:

- Mutations are recurrent at phenotypic level but not at DNA level.
- The mutation rate is small (highly unlikely to make the same mutation in the same place twice).
- Every time a mutation occurs, we are creating a new allele. So, the number of alleles that we can create is really high.

In a finite population, mutation creates alleles and genetic drift eliminates alleles.

So, when we create a new allele, 3 things can happen:

- It can be detrimental and thus, a purifying selection will eliminate it.
- It can be beneficial and thus, positive selection will fix it.
- It can be neutral and thus, it will evolve by genetic drift (it will be lost or fixed at random)

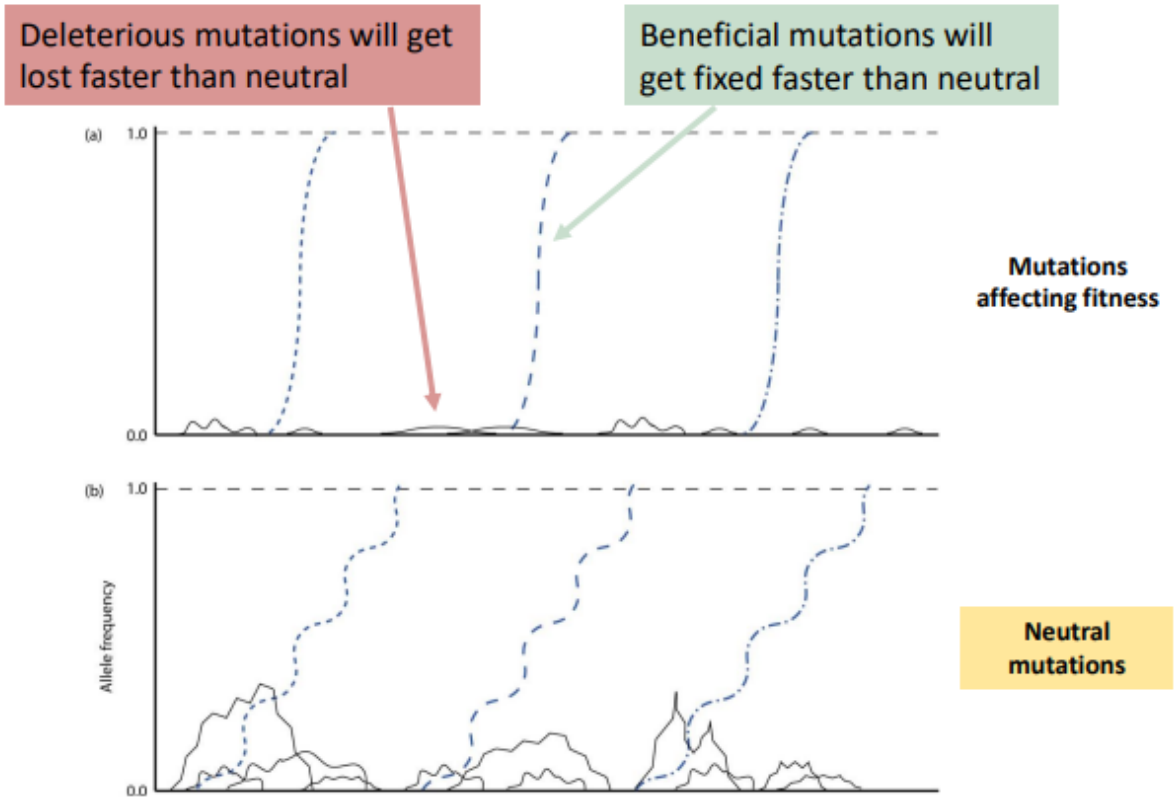
We must take into consideration that most non-neutral mutations are deleterious.

Neutral theory of molecular evolution

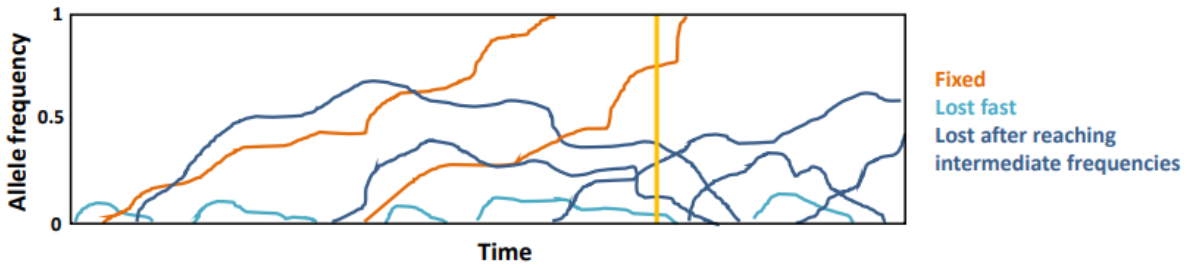
We have this population where mutation is introducing new alleles. If the allele is not beneficial it will be lost really fast and if it is beneficial it will be selected really fast (natural selection is fast).

But these variants are not polymorphic for a long time. They are lost or selected fast.

The neutral mutation, on the other hand, will evolve by drift and they will be polymorphic for a longer period of time.



So, the “Neutral Theory” says that most of the polymorphic variants that we see in the population are because they are neutral.

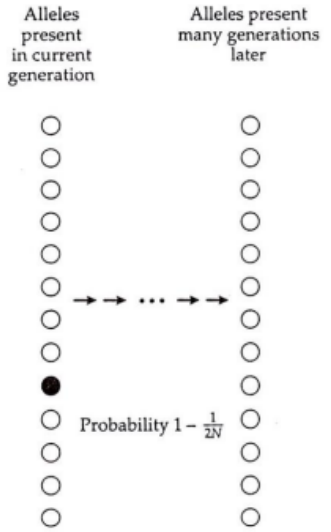


Most of the variation within and between species is due to random genetic drift of alleles that are selectively neutral.

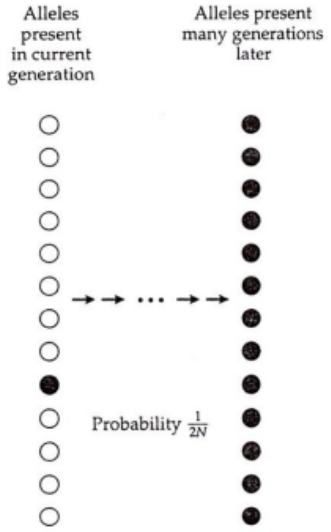
Lluís Giménez Gabarró

Fixation or loss of an allele in a finite population

The probability of fixing a variant is equal to the initial proportion of that variant (when it's neutral).



New allele is lost



New allele is fixed

$$P_{fix} = p_0$$

The probability of getting lost is 1 - 1/(2N) and the probability of being fixed is 1/(2N). So, it's more likely to be lost.

Neutral evolution rate

This neutral theory of molecular evolution allows us to predict the evolution rate (how fast the sequences are going to change).

Mutation rate (μ): Rate at which changes are incorporated in a nucleotide sequence during DNA replication and reparation processes. Rate at which a new allele appears.
Substitution rate (K): Rate at which new mutations are fixed in the population

To know the substitution rate, we will need to know how many substitutions happen per generation and multiply by the probability of fixation. If we develop the formulas, we can see that the mutation rate is equal to the substitution rate.

NEUTRAL MUTATIONS

$$\frac{\text{substitutions}}{\text{generation}} = \frac{\text{mutations}}{\text{generation}} \cdot P(\text{fixation})$$

$$K = 2N\mu \cdot \frac{1}{2N} = \mu$$

$$K = \mu$$

EXAMPLE

$$\mu = \frac{1}{1000} = 1 \cdot 10^{-4} \text{ mut/allele/generation}$$

$$N = 10000 \text{ individuals} \qquad 2N = 20000 \text{ alleles}$$

$$P(\text{fix}) = \frac{1}{20000}$$

$$2N\mu = \frac{1}{1000} \cdot 20000 = 20 \text{ mut/gener}$$

$$\lambda = 20 \cdot \frac{1}{20000} = \frac{1}{1000} \text{ subst/gen}$$

The substitution rate is equal to the mutation rate
Neutral divergence among species depends only on divergence time and mutation rate

Lluís Giménez Gabarró

Average time to fixation or loss

Average time that an allele takes to reach fixation or loss depends on its initial frequency when under the influence of genetic drift alone.

Average time to fixation

$$T_{fix} = -4N \frac{(1-p)\ln(1-p)}{p}$$

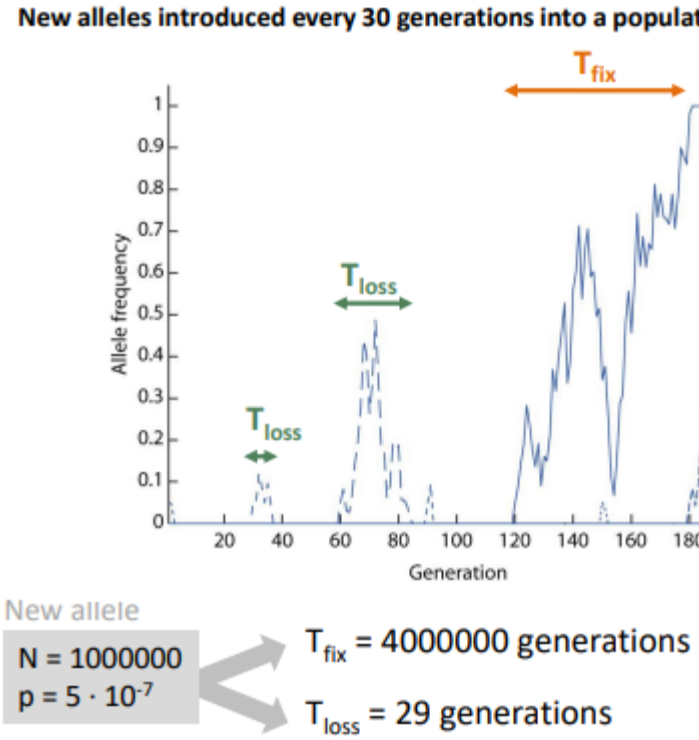
NEW ALLELE $p = \frac{1}{2N}$ → $T_{fix} \approx 4N$

Average time to loss

$$T_{loss} = -4N \frac{p \ln(p)}{1-p}$$

NEW ALLELE $p = \frac{1}{2N}$ → $T_{loss} \approx 2\ln(2N)$

When a single allele appears, the time to fix it is equal to 4 times the total population on **average** (4 hundred generations, for example).



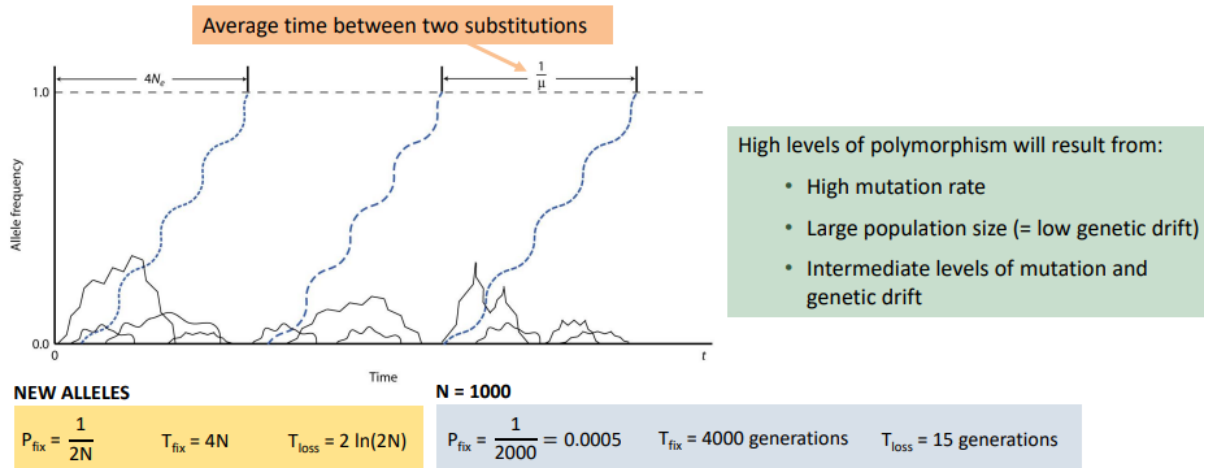
So, it is more likely to lose the allele.

Lluís Giménez Gabarró

Polymorphism under neutral theory

Polymorphism results from the transient dynamics of allele frequencies before they reach fixation or loss. While new alleles are segregating there is polymorphism in the population.

Very few mutations fix, but those segregate for much longer time than the mutations that end in loss.



The variants that are going to be fixed are going to remain more time than the ones that are going to be lost. So, these are the ones we are going to see as polymorphisms (neutral and fixated).

The average time between 2 substitutions is equal to 1/mu.

We are going to have a high level of polymorphisms when:

- There is a high mutation rate (a lot of new alleles will emerge)
- There is a large population size (thus the genetic drift is not relevant and thus, it will not fix a lot of polymorphisms, resulting in a larger amount of polymorphisms)
- Intermediate levels of mutation and genetic drift (because there will be an equilibrium of losing and fixing)

Lluís Giménez Gabarró

Migration and population structure

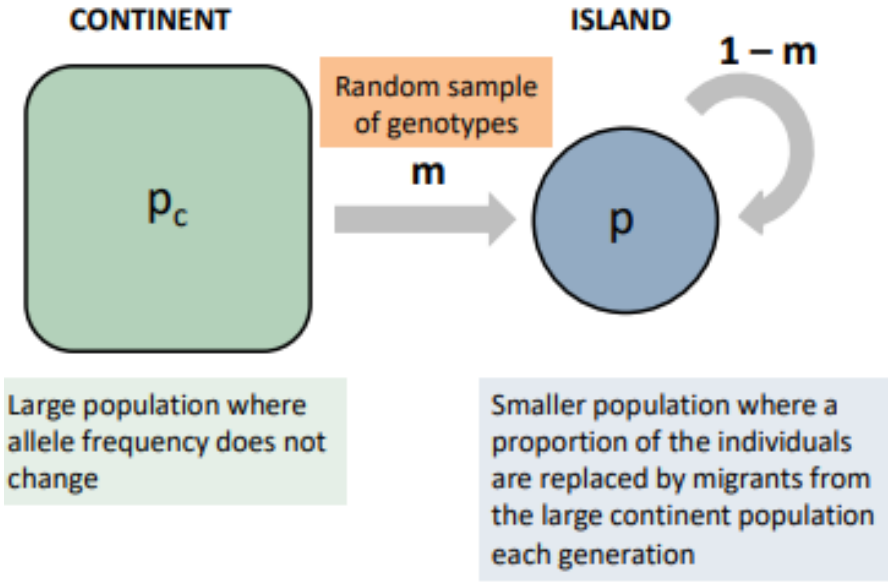
We are going to allow migration, which is also going to make changes in allele frequencies. So, now we know the 4 methods that change allele frequencies: Natural selection (the only one that makes the population better adapted), genetic drift, mutation and migration.

Two different populations of the same species can have different allele frequencies and even different alleles. This is due to the fact that genetic drift acts randomly in both, selection favors different alleles according to the environment and mutations will also produce variants.

Both populations must be isolated.

If there is movement between both populations, we will have migration.
Migration causes gene flow or transfer of genetic material from one population to another. It also limits the genetic divergence that can occur among subpopulations (if there is migration, the species can not diverge so much).

Example. Continent-island model: Imagine that we have a population in a continent that has a constant frequency. We have a migration rate (percentage of the alleles of the population of the island that come from the continent) of individuals from the continent to the island. Thus, if m = 0.1, 10% of the alleles of the island come from the continent.



We are going to suppose that the migration rate is constant. So, the allele frequency of the island is going to change in each generation.

How can we know the allele frequency of the island at any time?
In generation 1 I am going to have:

- The allele frequency of the generation before from the island. p0 · (1-m)
- Allele frequency that comes from the continent. pc · m

$$p_1 = p_0(1 - m) + p_C\, m$$

$$p_2 = p_1(1 - m) + p_C\, m$$

$$p_2 = (p_0(1 - m) + p_C\, m)(1 - m) + p_C\, m$$
$$p_2 = p_0(1 - m)^2 + p_C\,(1 - (1 - m)^2)$$
$$p_2 = p_C + (p_0 - p_C)\,(1 - m)^2$$

$$p_3 = p_C + (p_0 - p_C)\,(1 - m)^3$$
$$\vdots$$

$$p_t = p_C + (p_0 - p_C)(1 - m)^t$$

pc = frequency in the continent
pt = frequency in the island
p0 = Initial frequency in the island

Imagine that we have 1000 alleles with p = 0.5 in the island and, each generation, 100 alleles come from the continent (migration rate is equal to 0.1).
In the continent, p = 0.8.

We can calculate the new frequencies.

	m = 0.1		Island $p_0 = 0.5$	Continent $p = 0.8$		
Generation	**Alleles from island**	**Alleles from continent**	**Island**	**Continent**	**p**	**q**
0	1000	0	500 A 500 a	0	$p_0 = 0.5$	$q_0 = 0.5$
1	900	100	$p_0 \cdot 900 = 450$ A $q_0 \cdot 900 = 450$ a	p·100 = 80 A q·100 = 20 a	$p_1 = \frac{450 + 80}{1000} = 0.53$	$q_1 = \frac{450 + 20}{1000} = 0.47$
2	900	100	$p_1 \cdot 900 = 477$ A $q_1 \cdot 900 = 423$ a	80 A + 20 a	$p_2 = \frac{477 + 80}{1000} = 0.557$	$q_2 = \frac{423 + 20}{1000} = 0.443$
3	900	100	$p_2 \cdot 900 = 501$ A $q_2 \cdot 900 = 399$ a	80 A + 20 a	$p_3 = \frac{501 + 80}{1000} = 0.581$	$q_3 = \frac{399 + 20}{1000} = 0.419$

10% of the alleles of the island are going to come from the continent each generation.

We can compute the variation of the frequency of an allele using:

$$\Delta p = p_1 - p_0 = -m\,(p_0 - p_C)$$

Since “m” is always positive, depending on the value of (p0-pc) we will increase or decrease the frequency of that allele.

If $p_0 > p_C$ then $p_0 - p_C > 0$ and allele frequency in the island will DECREASE

If $p_0 < p_C$ then $p_0 - p_C < 0$ and allele frequency in the island will INCREASE

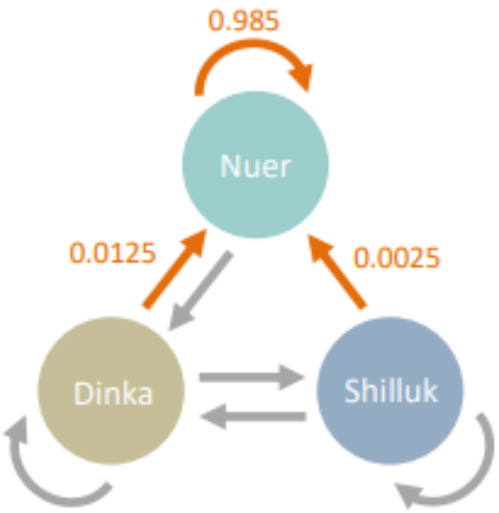
We will reach an equilibrium until there is the same frequency in both populations. Since the frequency of the continent does not change, we will reach an equilibrium when pt = pc.

Equilibrium is reached more slowly when there is less migration.

General Model

We can apply this model to any possible situation.

Example: We have 3 populations with the following mutation rates.



Here we have all the mutation rates between all the populations

Migration rates (m_{ij})	Donor population (i)		
Recipient population (j)	NUER ($p_0 = 0.575$)	DINKA ($p_0 = 0.567$)	SHILLUK ($p_0 = 0.505$)
NUER	0.9850	0.0125	0.0025
DINKA	0.0138	0.9775	0.0087
SHILLUK	0.0000	0.0098	0.9902

What is going to happen in the next generation?

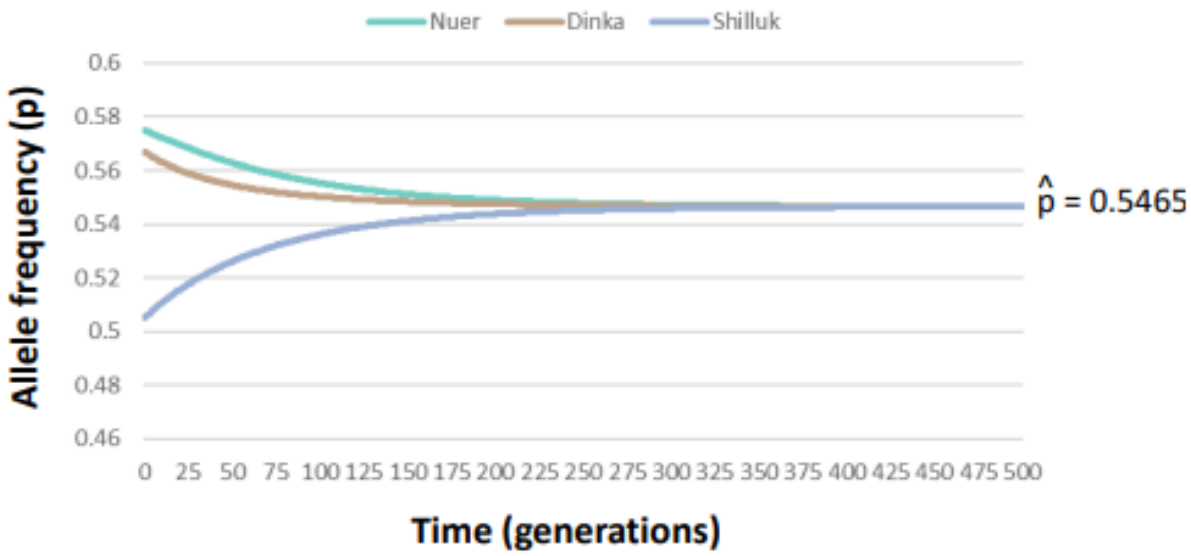
NUER

$$p_1 = \underbrace{0.9850}_{1-m} \cdot \overbrace{0.575}^{\text{NUER } p_0} + \underbrace{0.0125}_{m_{D\to N}} \cdot \overbrace{0.567}^{\text{DINKA } p_0} + \underbrace{0.0025}_{m_{S\to N}} \cdot \overbrace{0.505}^{\text{SHILL } p_0} = 0.5747$$

$$p_2 = 0.9850 \cdot \overbrace{0.5747}^{\text{NUER } p_1} + 0.0125 \cdot \overbrace{0.5666}^{\text{DINKA } p_1} + 0.0025 \cdot \overbrace{0.5057}^{\text{SHILL } p_1} = 0.5744$$

DINKA	SHILLUK
$p_1 = 0.5666$	$p_1 = 0.5057$

After a long number of generations, we will reach an equilibrium



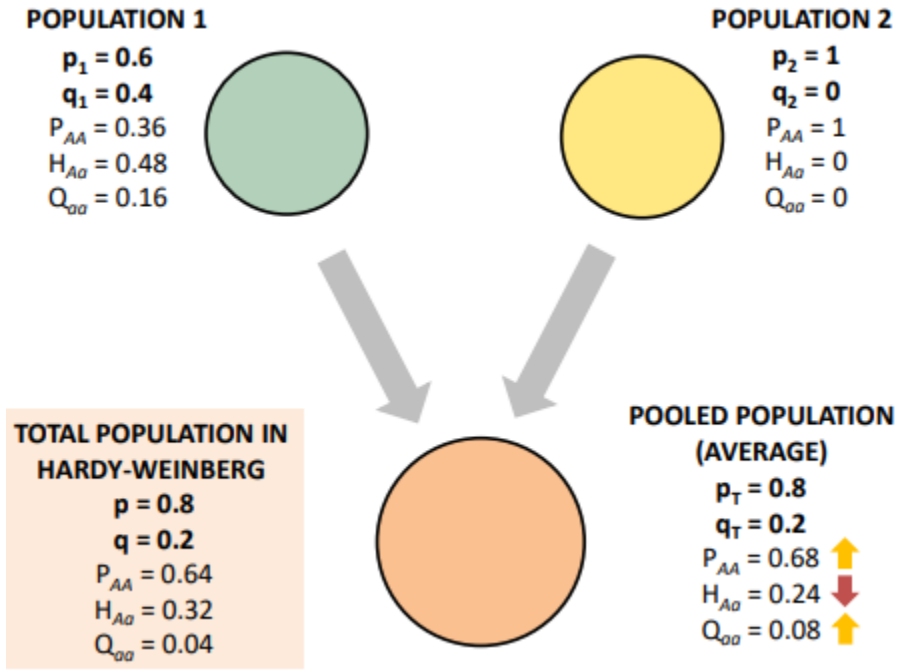
Lluís Giménez Gabarró

Wahlund effect

We are going to compare the differences between having a population made of different subpopulations vs having a unique population where each individual can mate with each other.

Population subdivision generates a deficiency in heterozygotes compared to those expected by HW in a total population with those allele frequencies.

Imagine that we have 2 populations of squirrels that can be Albino (aa) or WT (Aa and AA). Each of these populations are in HWE and thus, by knowing the allele frequencies we can know the genotype frequencies.
What happens if we pool these 2 populations? In this pooled population, the allele frequencies will be the average of the 2 populations.
If we compute the genome frequencies, we should obtain the results in orange. But we actually obtain the values on the right when there are 2 populations with migration.



Thus, we can state that we are not in a HWE. It's not in HWE because there are 2 different populations that differ in allele frequencies.

The proportion of heterozygous individuals in the pooled population (assuming an equal contribution from both) can be calculated using:

$$H_S = \frac{2p_1q_1 + 2p_2q_2}{2} = p_1q_1 + p_2q_2$$

The expected proportion of heterozygous individuals if the pooled population was in HWE is:

$$H_T = 2\left(\frac{p_1+p_2}{2}\right)\left(\frac{q_1+q_2}{2}\right) = \left(\frac{1}{2}\right)(p_1+p_2)(q_1+q_2)$$

$$H_S - H_T = -\left(\frac{1}{2}\right)(p_1-p_2)^2$$

The pooled population will always present a deficit of heterozygous individuals in comparison with the expected frequency under HWE, unless the allele frequency of both populations is the same (this would be the same as being in a HWE). The magnitude of the deficit depends of the allele frequencies:

$$p_1 - p_2 = 0 \Longrightarrow H_S - H_T = 0$$
$$p_1 - p_2 = 1 \Longrightarrow H_S - H_T = -0.5$$

Heterozygosity measurements

$H_{subpopulations}$ = Average expected heterozygosity assuming random mating within each population. So, we compute the proportion of heterozygotes in each population and then we do the average.

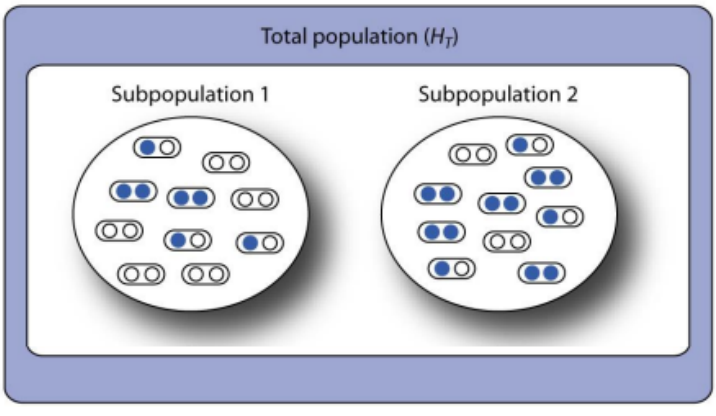
$$H_S = \overline{2pq} = \frac{\sum 2pq}{n}$$

H_{Total} = Expected heterozygosity in the total population. We compute the proportion of heterozygotes considering the average p and average q (allele frequencies)

$$H_T = 2\bar{p}\bar{q}$$

We can see this with an example:

We have 2 populations with different frequencies.



To compute the heterozygosity of the subpopulations:

- I compute the frequency of the alleles in both populations
- Compute the proportion of heterozygotes
- Do the mean

$$p_1 = \frac{13}{20} = 0.65 \qquad p_2 = \frac{7}{20} = 0.35$$

$$q_1 = 1 - p_1 = 0.35 \qquad q_2 = 1 - p_2 = 0.65$$

$$H_S = \frac{2p_1q_1 + 2p_2q_2}{2} = \frac{2 \cdot 0.65 \cdot 0.35 + 2 \cdot 0.35 \cdot 0.65}{2} = 0.455$$

To compute the heterozygosity of the total population:

- I compute the frequency of the alleles as a single population
- Compute the proportion of heterozygotes

$$p_T = \frac{20}{40} = 0.5$$

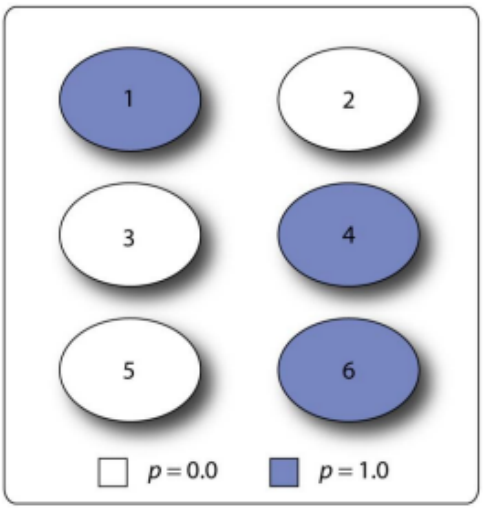
$$q_T = \frac{20}{40} = 0.5$$

$$H_T = 2p_Tq_T = 2 \cdot 0.5 \cdot 0.5 = 0.5$$

The difference in both results can be used to estimate the **fixation index**, which is the measure of how much allele frequencies have diverged among subpopulations.

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

Example. We have 6 populations where 3 populations are fixed for the white allele (p = 0) and 3 populations are fixed for the blue allele (p = 1).



$$H_S = \frac{3\,(2 \cdot 0 \cdot 1) + 3\,(2 \cdot 1 \cdot 0)}{6} = 0$$

$$F_{ST} = \frac{0.5 - 0}{0.5} = 1$$

We obtain a fixation index of 1 (which is the maximum). Meaning that there are highly diverged subpopulations.

If the 6 populations have the same allele frequencies, we will obtain a fixation index of 0. Meaning that there is no subdivision

$$H_S = \frac{6\,(2 \cdot 0.5 \cdot 0.5)}{6} = 0.5$$

$$F_{ST} = \frac{0.5 - 0.5}{0.5} = 0$$

How to know if there is a high divergence:

F_{ST} values	Level of differentiation
0 – 0.05	Low
0.05 – 0.15	Moderate
0.15 – 0.25	High
> 0.25	Very high

In humans, there are more differences between any population and Africans. This is due to the fact that humans originated in Africa and thus they are the ancestrals (more divergence).

Asia, europeans are the fruit of a founder effect from Africans.

The SNPs with very high fixation index are related to local adaptation (by natural selection).

Fixation index will increase by genetic drift

Without migration:

- Allele frequencies change among subpopulations
- Fixation index increases in each subpopulation
- Differentiation continues until one allele is fixed

Here we can see how the fixation index depends on the population size. The smaller the population, the faster the "F" increases.

Increase of fixation index with genetic drift

$$F_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_t$$

Fixation index in generation t in a finite population

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t$$

So, when we have drift and migration, there is an equilibrium:

- Genetic drift differentiates populations
- Migration homogenizes populations

$$F_{t+1} = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_t\right] (1 - m)^2$$

At equilibrium $F_t = F_{t+1} = \hat{F}$

$$\hat{F}_{ST} = \frac{1}{4Nm + 1}$$

At equilibrium, the fixation index does not change.

Migration strongly limits the differentiation between populations.
Meaning that the fixation index decreases rapidly as the number of migrants (N·m) increases independently of population size.

So, mutation is a weak force and migration is a strong force. Just a few individuals going from one population to another is enough to prevent this differentiation between the populations.

How do we know how many individuals are going from one population to another?
If we know the average fixation index we can know it:

$$F_{ST} = \frac{1}{4Nm + 1} \Rightarrow Nm = \frac{1 - F_{ST}}{4F_{ST}}$$

Average F_{ST} = 0.16
Nm = 1.3

Here we are assuming that natural selection is not acting and that we have reached the equilibrium point. Here we only have drift and migration, so it's for neutral variants.

N is always the effective size of the population!

Another way to estimate the migration rate would be by analyzing alleles found in only one population.
Mutation is a factor that can make an allele in one of the populations. If they are isolated, these alleles will remain in those populations.
So, we will have alleles that are private for a single population (when there is no migration).

A negative correlation is expected between the average frequency of private alleles and the values of Nm.

Lluís Giménez Gabarró

Measurements of genetic variation at DNA sequence level

How do we measure DNA polymorphism? One of the ways we can measure genetic variation when we have allele information is to use the proportion of heterozygotes, but this can not be applied to DNA sequences. Thus, we will see other methods.

Proportion of segregating sites
Number of segregating sites per nucleotide. How many positions in my sequence are polymorphic (have more than one variant).
We are looking at the positions that have changed, not the number of changes.

$$p_s = \frac{S}{L}$$

```
Seq1  AGGTATGCTAGAACCCTAGAAAGACACAGAGATAGACAAG
Seq2  AGGTATGCTAGAACCCTAGATAGACACAGAGATAGACAAG
Seq3  AGGTATGCTAGAAACCTAGATAGACACAGAGATAGACAAG
Seq4  AGGTATGCTGGAACCCTAGATAGACACAGAGATAGACAAG
Seq5  AGGTATGCTGGAACCCTAGATAGACACAGAGATAGACAAG
```

$S = 3 \quad L = 40 \quad p_s = 0.075$

This does not take into account the number of sequencing that we are analyzing. If we use more sequences, we will find more polymorphisms. Thus, we need to take this into account using Watterson's O.

Watterson's θ
Measurement of proportion of segregating sites corrected by sample size (take into account the number of sequences).

$$\theta = \frac{p_s}{a}$$

$$a = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}$$

n = number of sequences in the sample

Nucleotide diversity (π)
Average number of nucleotide differences per site between 2 random DNA sequences in the population. In other words, if I take 2 random sequences from a population, which will be the average number of differences?
Also, to normalize this, we will divide it by the total length of the sequence that I analyzed, as we did when computing the proportion of segregating sites (this is why say "per site").

$$\pi = \frac{\delta}{L}$$

δ: Average number of differences between two sequences

L: Length

$$\delta = \frac{\text{Total number of differences}}{\text{Total number of pair comparisons}}$$

How do we calculate it?
Imagine that we have five 500bp sequences and we know each position in which we have a polymorphism. To compute the average number of differences, we need to compare each sequence with all of the other ones and calculate the number of differences that we see (making the matrix).
Then we add the differences and divide it by the number of comparisons (we obtain the average number of differences between the 2 sequences). But this means nothing, because we are not taking into consideration how long the sequences are.

Thus, we can divide it by the number of nucleotides and we will obtain the Nucleotide diversity. **This value can be compared with any other value calculated in other species.**

Nucleotide variability in *Drosophila simulans* Rhodopsin 3 gene

		Position															
		132	142	162	192	198	201	207	240	246	351	354	372	375	405	417	483
Sequences	1	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
	2	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
	3	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
	4	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
	5	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A
Differences		6	6	4	7	4	4	4	4	6	6	4	4	4	6	4	6

L = 500 bp

Sequence	1	2	3	4	5
1					
2	8				
3	9	8			
4	12	7	5		
5	11	10	4	5	

79 differences in total
10 comparisons

$$\delta = \frac{79}{10} = 7.9$$

$$\pi = \frac{7.9}{500} = 0.0158$$

PROPORTION OF SEGREGATING SITES

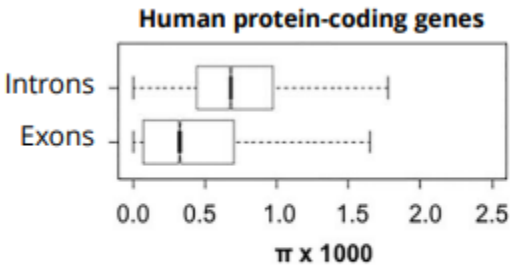
$$p_s = \frac{16}{500} = 0.032$$

WATTERSON'S θ

$$\theta = \frac{0.032}{2.08} = 0.0154$$

$$a = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.08$$

Nucleotide diversity in different positions of the genome
Nucleotide diversity varies among different organisms, different genes, different types of functional positions or along chromosomes. So, depending on the sequence we are analyzing we can get very different values.
If we take into account human protein coding genes, we will get a different value in introns and exons.



Introns average	π ≈ 0.0007
Exons average	π ≈ 0.0003

The value is higher in introns, meaning that there are more polymorphisms in introns. Introns are not that important and therefore they can accept more mutations.

Nucleotide diversity (π) in different species
Arthropods are more diverse than chordata (vertebrates) with plants in an intermediate situation. There is high variation within each species group.

The values of π vary a lot depending on the region of the chromosome you are analyzing.

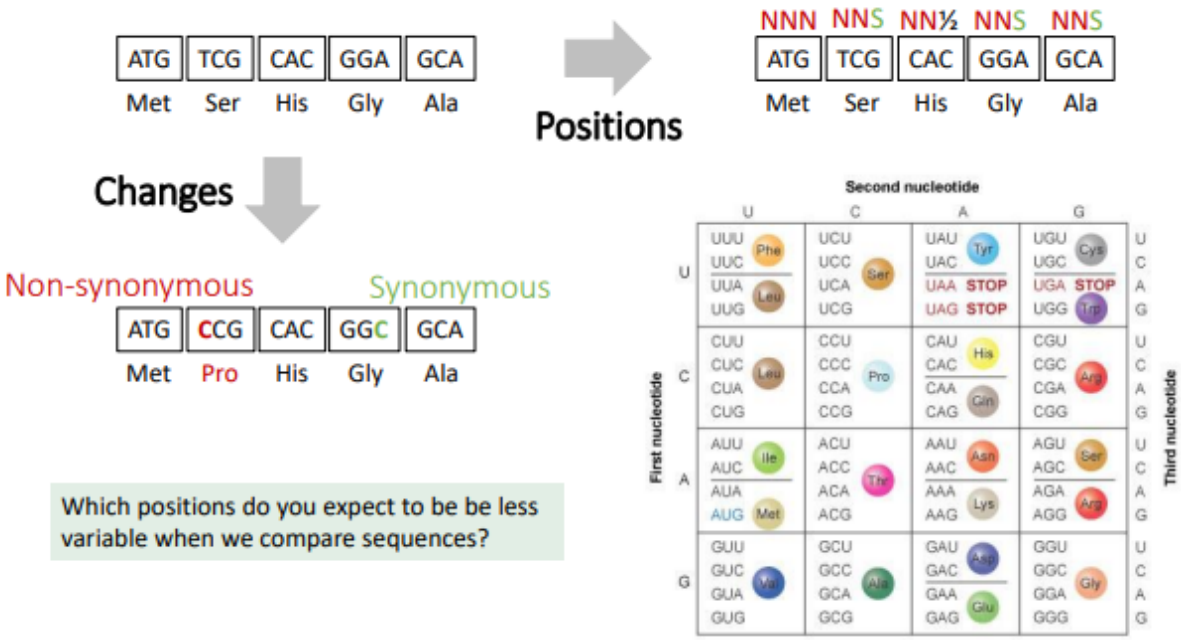
Genetic diversity in metazoans
The diversity of a species is predictable, and is determined in the first place by its ecological strategy.

Short-lived highly fecund species that release high numbers of small eggs into the environment are much more polymorphic than long-lived species that produce a small number of relatively large offspring and provide parental care (k-strategists)
Population size is probably an important factor in determining the level of nucleotide diversity
Large populations have more variation and therefore they are less likely to lose variation because of drift.

Synonymous and non-synonymous positions within coding regions
When we think about coding DNA sequences, we can distinguish 2 types of positions:

- Synonymous
- Non-synonymous

We can take a look at each of the positions of a codon and decide if they are synonymous or not. Example of serine: If we change the 3 position of the codon, we will always obtain a serine.

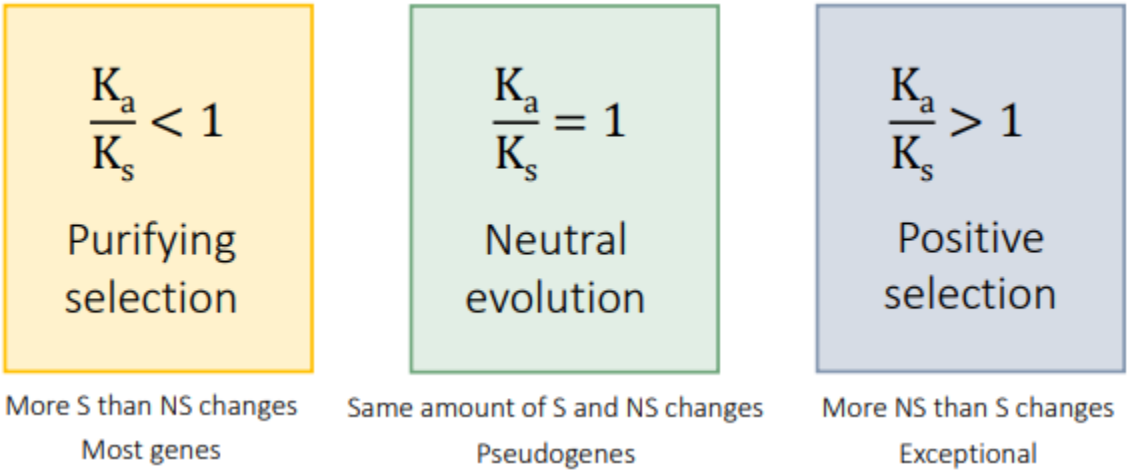


We could calculate π for the synonymous and nonsynonymous positions of a coding region. In this case, which positions do you expect to be more variable? The synonymous should be more variable, because all mutations will be conserved.

Ka/Ks ratio test

Ka = Number of non-synonymous substitution divided by the total number of non-synonymous positions

Ks = Number of synonymous substitutions divided by the total number of synonymous positions.



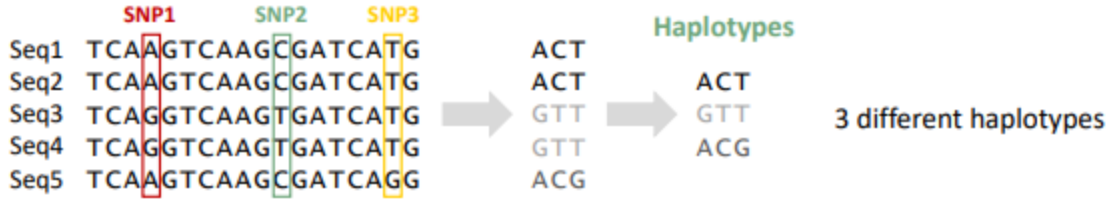
- In the first case, there are more changes in the synonymous positions. Thus, they are under purifying selection.
- In the middle case… This is typical in pseudogenes (genes that have been inactivated by a mutation)
- In the last case, there are more changes in the non-synonymous positions. Thus, they are under positive selection (because we are retaining the changes).

Lluís Giménez Gabarró

Linkage disequilibrium

Haplotypes

Combination of alleles in the same chromosome



Number of possible combinations for n variants

$$2^n$$

n = 2 $2^2 = 4$ combinations

n = 3 $2^3 = 8$ combinations

Surprisingly, we do not find all of these combinations because of the linkage disequilibrium.
For example, from 3 variants we do not see 8 combinations but 3.
This only happens in variants that are in the same sequence (in the same chromosome).
The variants that are in different chromosomes, they are independent from each other.

Variants in the same chromosome can be linked

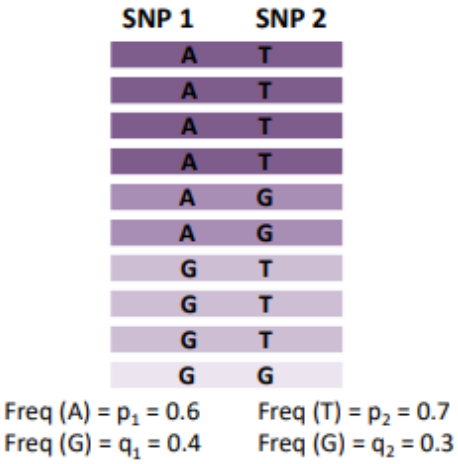
For SNP 1, we have 2 different alleles:

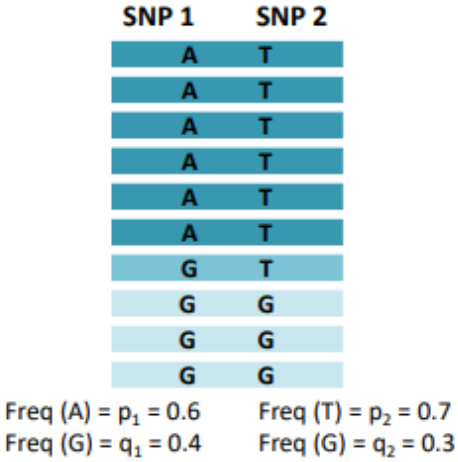
- A and G

For SNP 2 we have 2 different alleles:

- G and T

We can compute the frequency of each allele in each SNP.
Note that we have 10 chromosomes and therefore there are 5 individuals.





As we can see, there are 4 different combinations of alleles.
We can compute their observed and expected frequencies.

			OBSERVED	EXPECTED
A	T	Freq (AT)	0.4	$p_1 \cdot p_2 = 0.42$
A	G	Freq (AG)	0.2	$p_1 \cdot q_2 = 0.18$
G	T	Freq (GT)	0.3	$q_1 \cdot p_2 = 0.28$
G	G	Freq (GG)	0.1	$q_1 \cdot q_2 = 0.12$

			OBSERVED	
A	T	Freq (AT)	0.6	$= p_1 \cdot p_2 + D$
A	G	Freq (AG)	0	$= p_1 \cdot q_2 - D$
G	T	Freq (GT)	0.1	$= q_1 \cdot p_2 - D$
G	G	Freq (GG)	0.3	$= q_1 \cdot q_2 + D$

D = 0.18

We can do the same thing with the blue sample. As we can see, the frequencies of the alleles are the same and therefore we should expect the same expected combinations. But the observed combinations are really different.
This is because there are some variants that are in linkage equilibrium and others in linkage disequilibrium.

For 2 of the combinations I have a higher observed value than expected (there is a value that is added “D”) and in other cases I have a lower observed value than expected (there is a value that is subtracted “D”).

I am trying to see if there is the expected proportion of each haplotype (combination) in the population or not. If I have the expected proportion, then there is no linkage equilibrium .

Linkage equilibrium

Random association

Genotype	Frequency
AB	$P_{AB} = p_A p_B$
Ab	$P_{Ab} = p_A p_b$
aB	$P_{aB} = p_a p_B$
ab	$P_{ab} = p_a p_b$

$D = 0$

The expected frequency of each combination is the product of the involved allele frequencies

Linkage disequilibrium

Correlation between two *loci*

Genotype	Frequency
AB	$P_{AB} = p_A p_B + D$
Ab	$P_{ab} = p_A p_b - D$
aB	$P_{aB} = p_a p_B - D$
ab	$P_{ab} = p_a p_b + D$

$D \neq 0$

ATENTION!
If there are more *AB* and *ab*, there MUST be less *Ab* and *aB*

Note that D can be negative when the proportion of AB and ab is smaller than expected.

Linkage disequilibrium measurement (D)
We use the parameter D to measure the linkage disequilibrium.
D is the difference between the observed frequency of a haplotype and the expected frequency of this haplotype of these alleles were independent.

Genotype	Frequency
AB	$P_{AB} = p_A p_B + D$
Ab	$P_{ab} = p_A p_b - D$
aB	$P_{aB} = p_a p_B - D$
ab	$P_{ab} = p_a p_b + D$

$$D = P_{AB} - p_A p_B$$

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

Linkage disequilibrium relative measurement
The problem with this measure is that D values depend on allele frequencies.
So, we have the same problem as when we were talking about nucleotide diversity. We said that we have 8 differences on average between 2 sequences but we don't know the meaning of that value if we do not know how many positions we are analyzing.
So, here we have the same problem.

Since D depends on the allele frequencies, the same value of D can mean a lot of linkage equilibrium and in other cases the same value can mean little linkage. Thus, we need relative measures that allow us to compare.

To compare them we need to normalize using Dmax or Dmin

$$\text{If } D > 0 \quad D' = \frac{D}{D_{max}}$$

$$\text{If } D < 0 \quad D' = \frac{D}{D_{min}}$$

When **D' = 1,** we have complete LD.
It happens when there are at most 3 of the 4 possible haplotypes present in the population.

The closer to 1, the more Linkage Disequilibrium.
Observed and expected are really different.

	SNP1 A/G	SNP2 C/A
AGAGTTCTGCTCG	A	C
AGGGTTCTGCGCG	G	C
AGGGTTATGCGCG	G	A

AA combination does not exist

The other measure that we use is r^2, which is the correlation coefficient of the allele frequencies (values from 0 to 1). We divide it by the product of all the allele frequencies.

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

When r^2 = 1 we have perfect LD.
It happens if the two variants have the same allele frequencies (they are completely linked).
Only 2 different haplotypes exist.

- When there is an A in the first SNP, there is a C in the second SNP.
- When there is a G in the first SNP, there is an A in the second SNP.

	SNP1 A/G	SNP2 C/A
AGAGTTCTGCTCG	A	C
AGGGTTATGCGCG	G	A

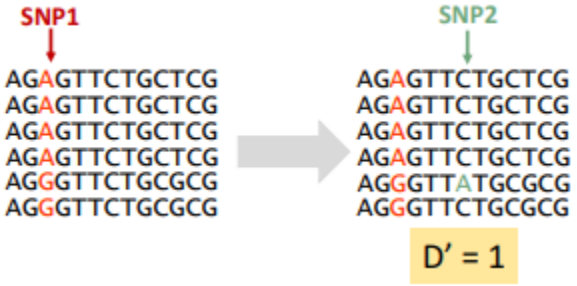
They will have the same allele frequencies, because if the frequency of A is 0.3, the frequency of C is also going to be 0.3

Haplotypes and recombination
What happens when a variant appears for the first time in a DNA sequence. We have the following population of sequences which have a variant on the third position. Then a new variant appears in a single chromosome.
Now we have 2 variants.
We can look at the number of combinations:

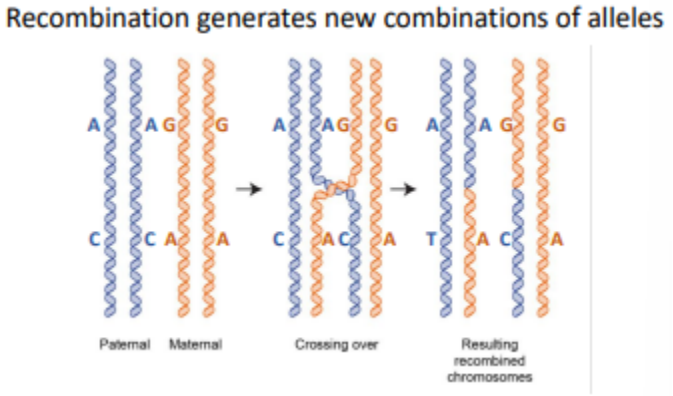
- A and C
- G and A
- G and C

When we have 3 combinations, D' = 1.

When a new variant appears, it is linked to the rest of variants in the chromosome where it originated.



Overtime, if this variant is not eliminated and maybe the allele frequency of the second SNP increases, through recombination we will generate new combinations.

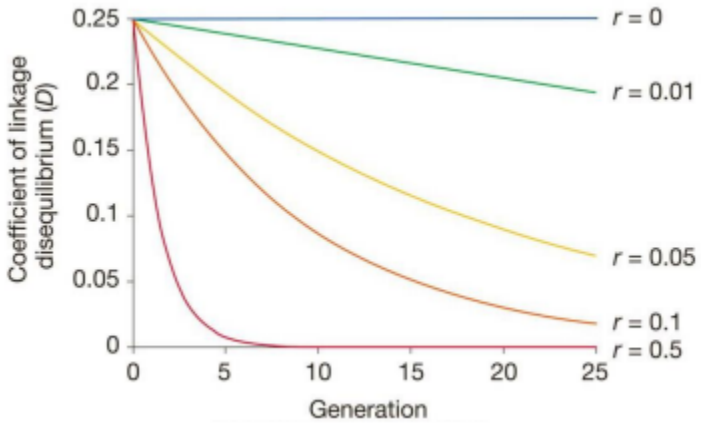


Recombination in heterozygotes breaks the linkage between variants. Thus, overtime, the amount of linkage equilibrium will decrease.
How fast will it decrease? It depends on the rate of recombination.

Linkage disequilibrium decay
Over time, LD between variants will decrease

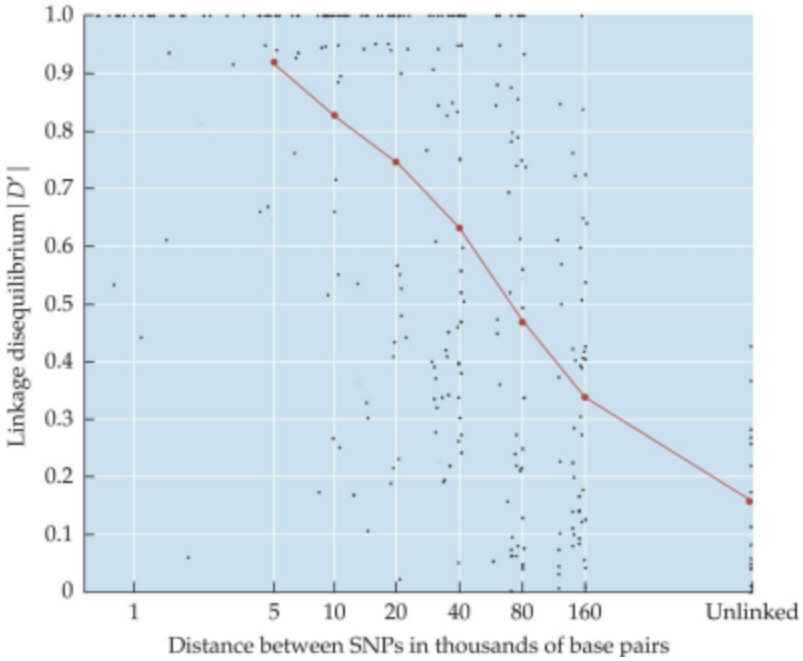
In many generations

$$D_t = (1 - r)^t D_0$$



The higher the recombination rate (r), the faster the linkage between variants will be lost. Thus, we will return faster to the allele frequencies that we expect.

Linkage disequilibrium and distance
LD decreases as the distance between variants increases.
Longer distance implies a higher probability that recombination occurs between the two variants



Applications of linkage disequilibrium

- Genome wide association studies (GWAS)
- Genotype imputation

GWAS
They are used to identify variants linked to a given phenotype. The phenotype is not caused by a mendelian gene. These are more complex characters from which we do not know the reason for that character (for example hate, diabetes… they don't depend on a single gene).

We need 2 groups of people:

- Cases
- Control

The idea is that we have to genotype a lot of variants. We are not genotyping all the variants in this genome, but since these variants are linked to the other ones that are close, if we genotype enough SNPs, we will control all variants.
So, we will be able to find the variants that are associated with this phenotype.

The variants that we are going to find are probably not the ones causing the phenotype, but when you detect a variant it is probably linked to another variant (not detected) that is responsible for the phenotype.

GWAS allows us to find variants associated with the phenotype of the case group. The identified variants most likely are not the causal variants responsible for that phenotype, but they will be in LD with the causal variant, so this variant must be located close to the ones identified by the GWAS.

Association mapping
Association between a SNP on human chromosome 16 and type 1 diabetes

	Cases	Controls	Total
A	1300 / 1363.6	2109 / 2045.4	**3409**
G	700 / 636.4	891 / 954.6	**1591**
Total	**2000**	**3000**	**5000**

Observed
Expected

$$P(A, case) = P(A) \cdot P(case) = \frac{3409}{5000} \cdot \frac{2000}{5000} \cdot 5000 = 1363.6$$

$$P(A, control) = P(A) \cdot P(control)$$

$$P(G, case) = P(G) \cdot P(case)$$

$$P(G, control) = P(G) \cdot P(control)$$

$\chi^2 = 15.537$
df = 4 -1 -2 = 1
$P < 8 \cdot 10^{-5}$

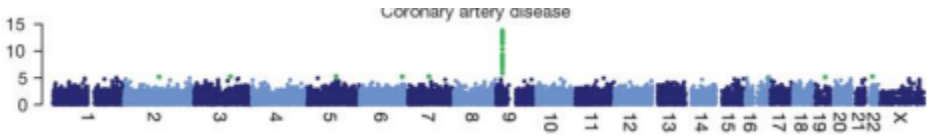
Observed and expected are different. There is an association.
The SNP is closely linked to the variant that is causing the disease.
Allele G is a risk factor for diabetes.

Allele G is a risk factor for diabetes because it is in a higher proportion than expected. Maybe it is not the causal but it is linked to the causal.

GWAS
Identification of genes involved in common complex human diseases

The X axis represents the chromosomes and each dot represents the result of the previous test for one SNP. In some places, there is a higher pick. Meaning that the SNPs are very associated to the phenotype (very significant).
As we said, this does not mean that these variants are causing the disease but they indicate that in that particular region, there are some variants involved in increasing the risk of having that disease.

So, now we know which is the region involved in the disease.



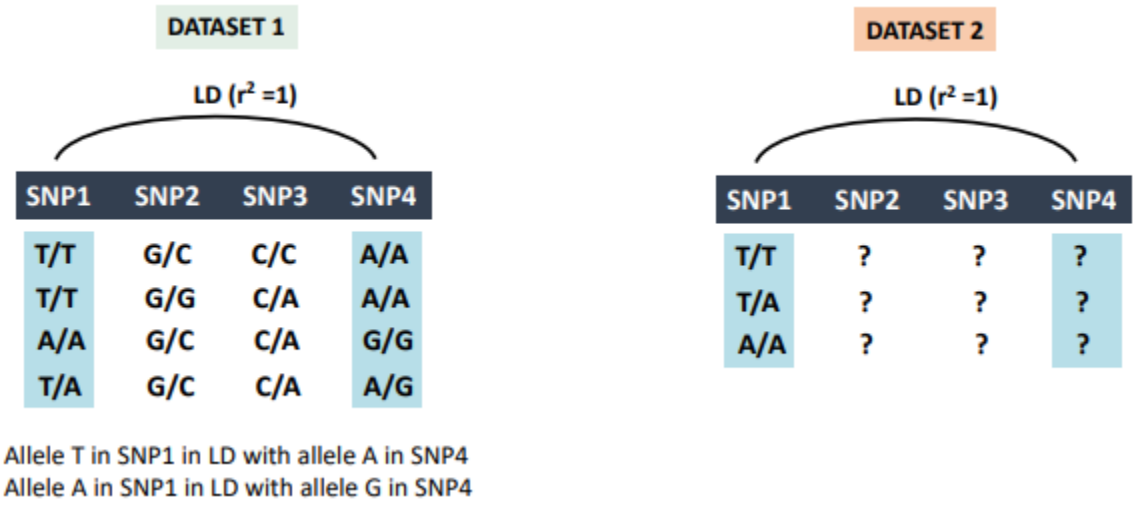
We can only do this because the variants are in linkage disequilibrium with other variants. If they were independent, we would have a hit on the causal variant.

LD and genotype imputation

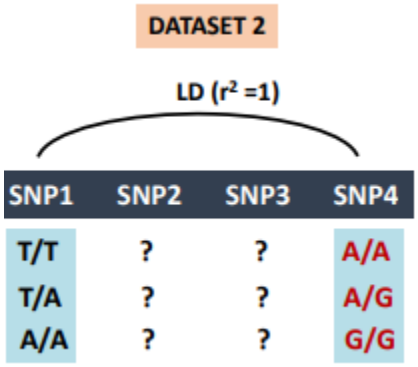
Another application of LD is genotype imputation. We can deduce genotypes without actually genotyping.

Imagine that we have dataset 1, where we genotype a lot of SNPs (each row is an individual). Then we check the LD for these SNPs. Imagine that SNP1 is in LD with SNP4 with a value of 1. Meaning that only 2 of the 4 combinations are possible.

Thus, we can deduce the SNP4 of another dataset.



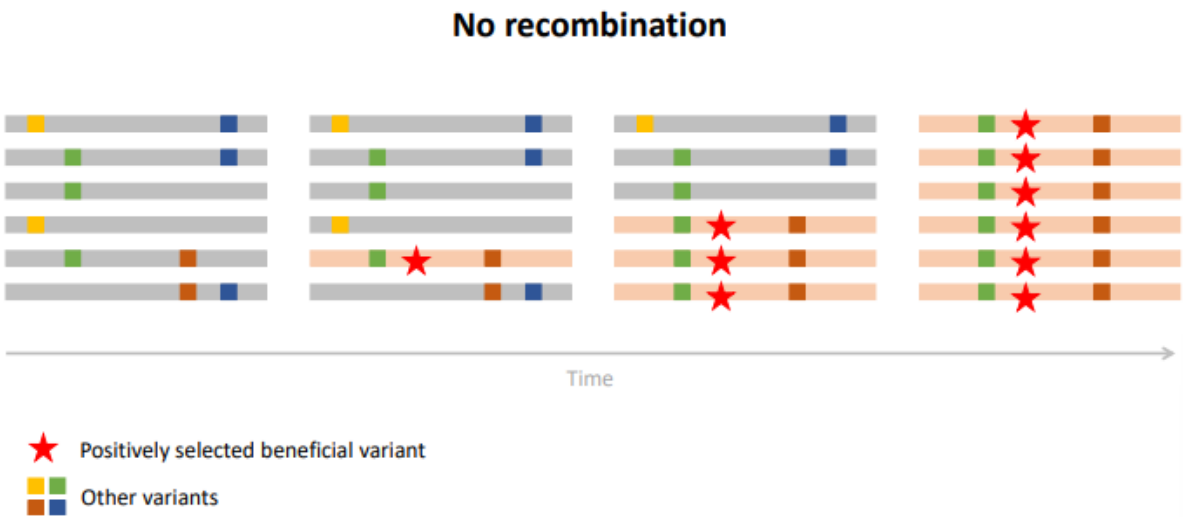
Allele T in SNP1 in LD with allele A in SNP4
Allele A in SNP1 in LD with allele G in SNP4



Thanks to LD between variants, we can complete the missing genotypes

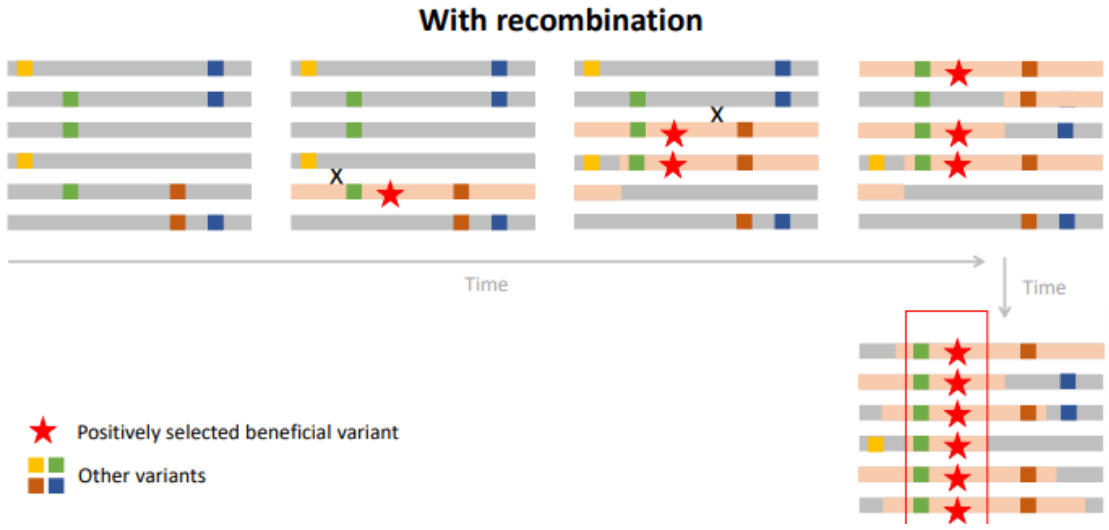
Selective sweep - no recombination
Another consequence of linkage disequilibrium.
Imagine we have a case without recombination.
We have chromosomes with different variants represented in different colors.

If a new positively selected beneficial variant occurs, it will be selected. Its frequency will increase. Thus, we will be increasing the frequency of the chromosome that has this variant. All chromosomes will be equal. Meaning that not only the beneficial variant will be selected but all the variants that are in that chromosome.



Selective sweep - with recombination
What happens when we have recombination?
We said that recombination destroys LD.

In this case, the beneficial variant will also be selected but due to recombination we will not obtain the same chromosomes. Note that the close variants (green and red star) will be maintained together because of LD.

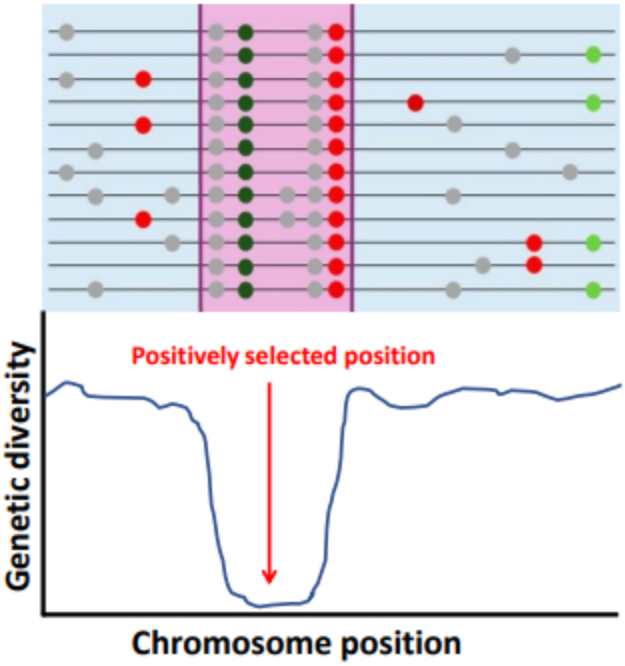


Detection of a selective sweep

Selective sweep: Reduction of measured diversity in the surroundings of a positively selected mutation

1. A new beneficial mutation appears
2. It rapidly becomes the most common variant in the population
3. Nearby positions also become more frequent because they are not physically independent

Genetic hitchhiking: Occurs when an allele changes frequency not because it itself is under natural selection, but because it is near another allele that is undergoing a selective sweep



Lluís Giménez Gabarró

Neutrality Tests (Kimura)

Idea that most polymorphisms are selectively neutral. We are going to test if natural selection is acting or not.
We have seen that genes can evolve by genetic drift, where natural selection does not act.

How can we test this neutral hypothesis?
The neutral theory can be used as a null model against specific occurrences of natural selection, and in some cases, natural selection can be detected (eg, providing evidence for positive selection; when a new selectively advantageous mutation –polymorphic in a population –is driven by selection).

Very important because it may provide evidence for adaptation at the molecular level, and help to elucidate genotype–phenotype relationships. The new availability of large (massive) genomic data sets (aka, NGS data) has invigorated the field of molecular population genetics and spurred new controversies regarding the causes of molecular evolution. Large samples of Single Nucleotide Polymorphisms (SNPs), microsatellites and DNA sequence data are currently being obtained in humans and other organisms. Using this new NGS genomics data (under the appropriate statistical framework), it is possible to identify regions that have undergone positive selection, and therefore we to determine the causes for species-specific phenotypic differences (eg, adaptation to altitude in human), or identify regions currently under selection (eg, by the presence of some disease-causing mutations).

Tests of neutrality provide us with a powerful tool for developing hypotheses regarding function from genomic data. They can be divided into two categories:

- Tests based on the pattern and levels of intraspecific variability (polymorphism)
- Tests based on comparisons of intraspecific and interspecific (divergence variability) between different classes of mutations (eg, such as nonsynonymous and synonymous mutations.

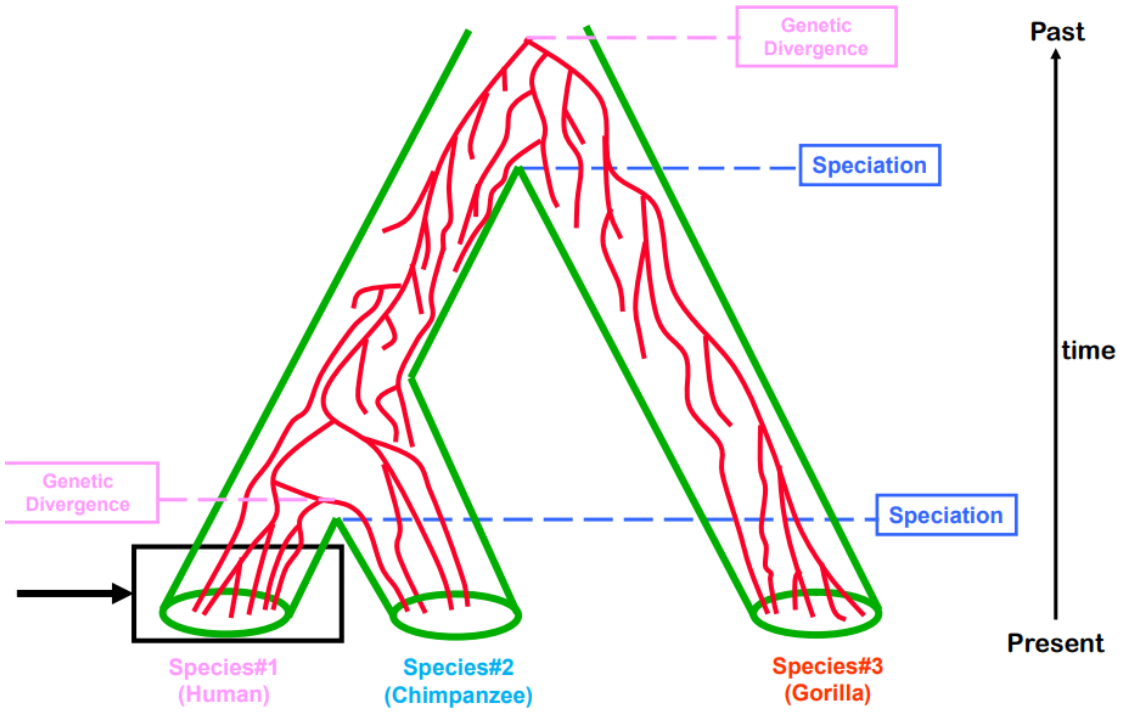
Using comparative analysis of DNA sequences, we can detect which amino acid change of a certain gene has been positively selected to do a certain thing (capture O2 more efficiently, for example)

Test	Compares
Tests based on allelic distribution and/or level of variability	
Tajima's *D*	The number of nucleotide polymorphisms with the mean pairwise difference between sequences
Fu and Li's *D*, *D**	The number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants
Fu and Li's *F*, *F**	The number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences
Fay and Wu's *H*	The number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies
Tests based on comparisons of divergence and/or variability between different classes of mutation	
d_N/d_S , K_a/K_s	The ratios of non-synonymous and synonmyous nucleotide substitutions in protein coding regions
HKA	The degree of polymorphism within and between species at two or more loci
MK	The ratios of synonymous and non-synonymous nucleotide substitutions in and between species

HKA, Hudson–Kreitman–Aguade; MK, McDonald–Kreitman.

Tajima's D is a neutrality test that attempts to determine if a particular set of nucleotide sequences are or are not compatible with the null hypothesis.

Note that the genetic divergence is older than the speciation event.

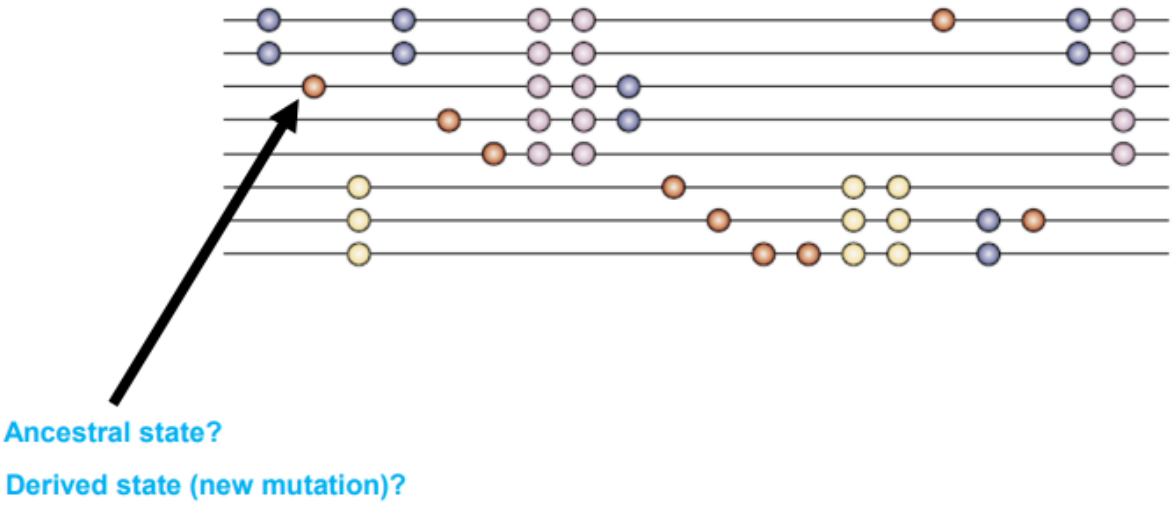


Which are the requirements to use information from the neutrality test to test for the behavior of some nucleotide sequences (if they are compatible or not with the neutral theory).

SFS (Site Frequency Spectrum)
Statistics that summarizes the distribution of derived allele frequencies in a sample of DNA sequences.
It provides useful information about genetic variation within and among populations and it can be used to make population genetic inferences. There are 2 SFS:

- **Unfolded:** Represent the SFS when you can use information of the ancestral variant.
- **Folded:** Represent the SFS when you can't use information of the ancestral variant and which is the derived variant.

Unfolded
We have a MSA of 20 Polymorphic sites out of 1000 nucleotides.
The different balls represent the different variants.

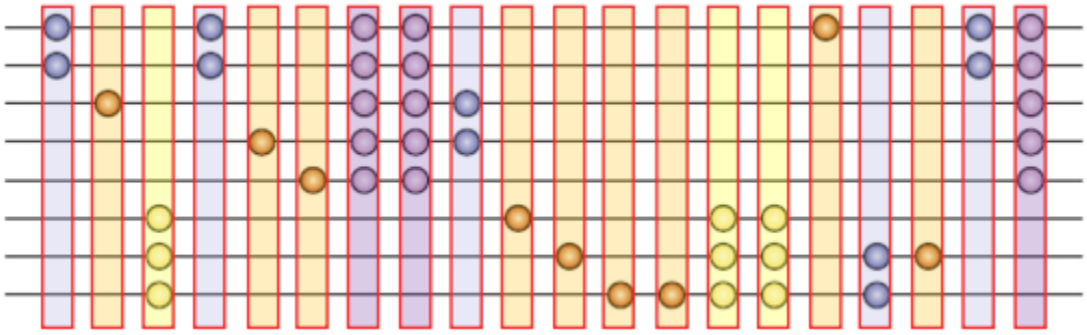


How do we know if it is an ancestral variant or a derived variant? Meaning that the ancestral variant is an A and the derived variant a T or the other way.
The answer is we do not know.

If we have information about a single species, we do not know if a certain variant is a mutation that occurred recently or it represents the ancestral state.

But, if we have information from the genealogy, we can know it by just checking if the ancestral species has this variant or not.

The derivate state is a Singleton Variant, which represents a variant (mutation) that you can see only once in a subset of DNA sequences.



1D / 7A (1 derived state / 7 ancestral)
Singleton. 9 cases

2D / 6A (2 derived / 6 ancestral)
Dobleton. 5 cases

3D / 5A (3 cases)

4D / 4A (0 cases)

5D / 3A (3 cases)

6D / 2A (0 cases)

7D / 1A (0 cases)

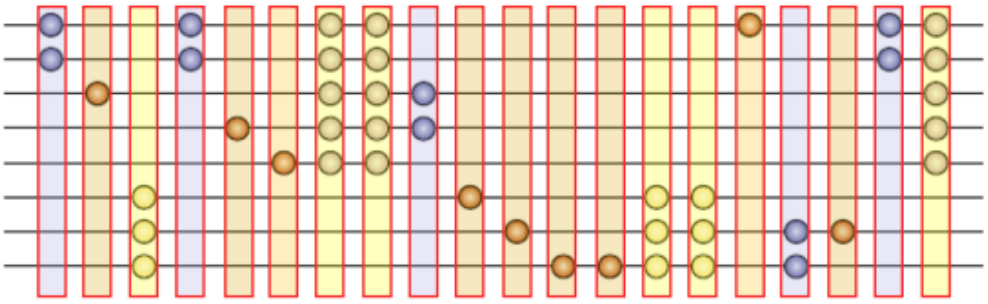
8D / 0A (0 cases; Fixed cases)

Here we can also see 2 derived mutations (dobleton) across the 8 sequences (thus, there are 6 ancestral variants).
Also 3 derived mutations across the 8 sequences (thus, there are 5 ancestral variants).

We just need to know the number of derived mutations and the rest are going to be ancestral variants.

Folded
If we don't have the genealogy, we do not have the information to know if one particular state is ancestral or derived, we use the Folded SFS.

If we work with 8 sequences, there can be at maximum 4 options. Because we will join the cases in which there are 3 derived and 5 ancestrals and the other way.



1 / 7 (1 state / 7 alternative state): 9 cases

2 / 6 (2 states / 6 alternative). 5 cases

3 / 5 (3 states / 5 alternative). 6 cases

4 / 4 (4 states / 4 alternative). 0 cases)

Tajima's D test

Tajima's D is computed as the difference between two measures of genetic diversity: the mean number of pairwise differences and the number of segregating sites
To compute the Tajima's D we need to know the concept of heterozygosity.

θ, heterozygosity under the mutation-drift equilibrium ($\theta = 4N_e\mu$)

θ_W, Watterson θ. Estimator of θ based on the number of segregating sites ($\theta_W = S/a_n$)

where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$

$\theta_T(\pi)$, Estimator of θ based on the nucleotide diversity ($\theta_T(\pi) = k$))

where k, is the average number of nucleotide differences

Under the neutral model, the values of θw and θT are equal to θ.

Thus, we can compute θ for the average number of segregating sites (θw) and for the average number of nucleotide differences (θT). Then we can compare them.
If there are many differences, then we reject the null hypothesis.

Aims: To determine if a particular region (eg a gene) is evolving neutrally

H_0: Neutral model (neutral evolution)

$k = S/a_n$

The pattern of polymorphisms (the frequency) identified in the multiple sequence alignments is the expected under de neutral model (equilibrium mutation-genetic drift)

H_1: Non-neutral evolution (such as natural selection, other factors)

$k \neq S/a_n$

No. The pattern does not follow such expectations.

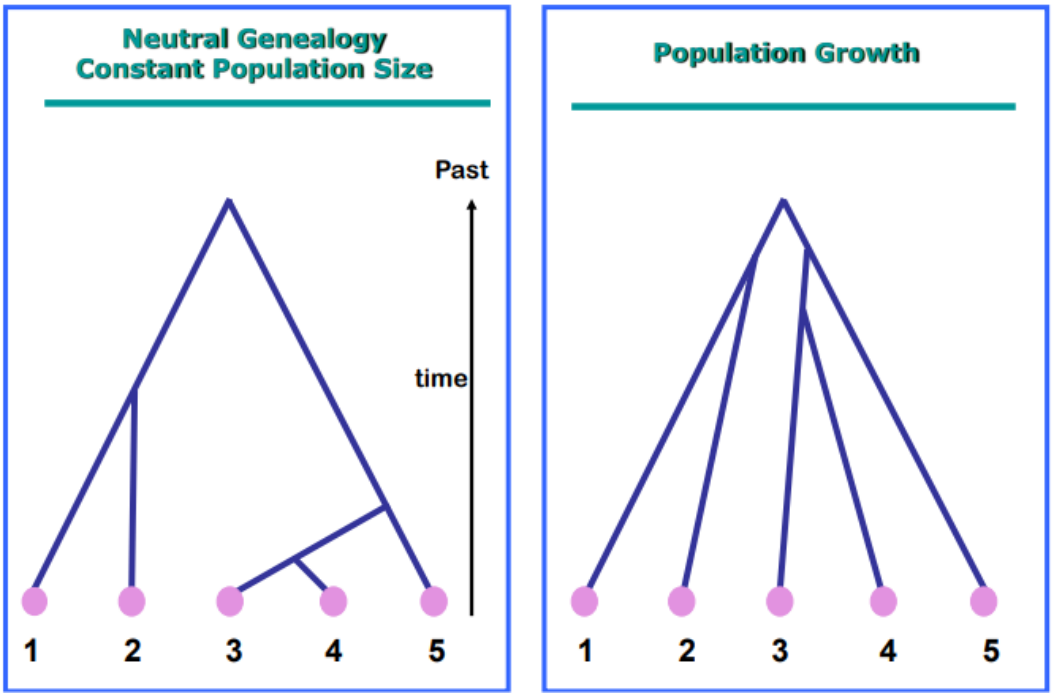
Why? There are some (but unknown) evolutionary factors affecting patterns of polymorphisms. Positive selection? Negative selection? Migration? Demographic factors? In this case we will need to perform additional analysis/experiments to determine the specific event.

In the formula:

- θw = S. It is the number of segregating sites
- θT = π. It is the average number of nucleotide differences (per site)
- k is the average number of nucleotide differences (per region)
- n is the sample size (number of DNA regions)

$$D = \frac{k - S/a_n}{\sqrt{Var(k - S/a_n)}}$$

Here we have 2 genealogies of 5 sequences within species (human form example):

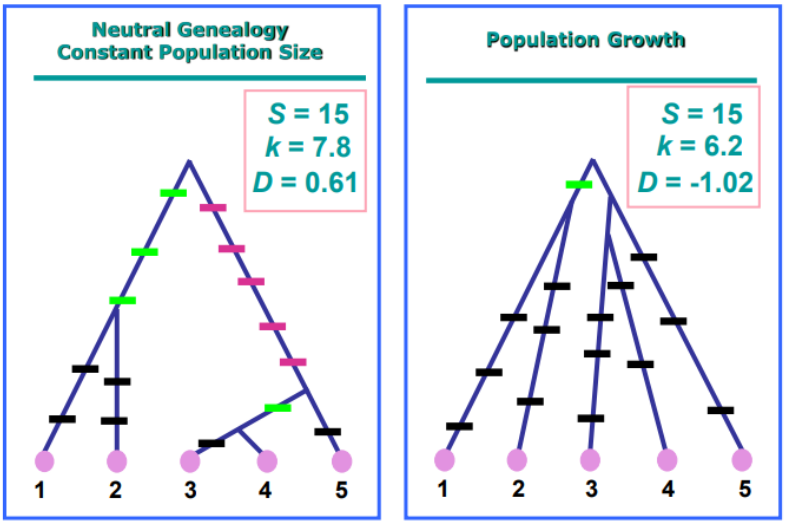


The first one is the expected genealogy.
The second is specific for a population that has expanded recently (population growth).

Imagine that we have 15 segregating sites in both genealogies and:

- Black represent Singletons
- Green represent doubletons
- Pink represent 3

To know if they are singletons, doubletons… we need to look at the species affected by the mutation.
We can compute the number of segregating sites (S = 15), the average number of nucleotide differences (per site) and Tajima's D.



If there is a population growth, the topology shape affects the SFS.

We have seen that when there is an expansion, there are way more singletons. Thus, by comparing the number of singletons, doubletons… between the 2 cases, we can accept or reject if the population is under a neutral model or not.

In the last example, the 2 values of D are statistically different?

> But (in our example), for $n = 5$; $D_{obs} = 0.61$ or $D_{obs} = -1.02$ are significant?
>
> We need to know the distribution (the confidence intervals, CI) of D
> eg, for $n = 5$, the CI (at 95%) are: (-1.269; 1.834)
>
> Therefore, $D_{obs} = 0.61$ or $D_{obs} = -1.02$ are not significant

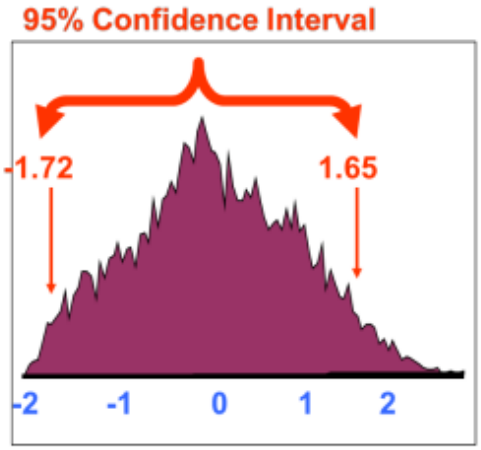
Thus, we can not reject the null hypothesis.

The distribution of D depends on "n", the theta (heterozygosity) value and the recombination value (the distribution is very sensitive to the recombination).

- It is not the same considering regions of a non-recombining molecule and considering autosomal regions that exhibit recombinations.

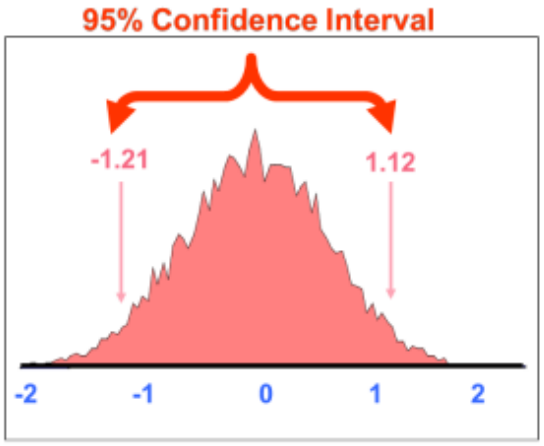
Another example: $D_{obs} = -1.45$ $n = 10$; $\theta = 12$

$n = 10$; $\theta = 12$; Replicates = 10,000
No recombination, $R = 0$



Not significant

$n = 10$; $\theta = 12$; Replicates = 10,000
Recombination, $R = 15$

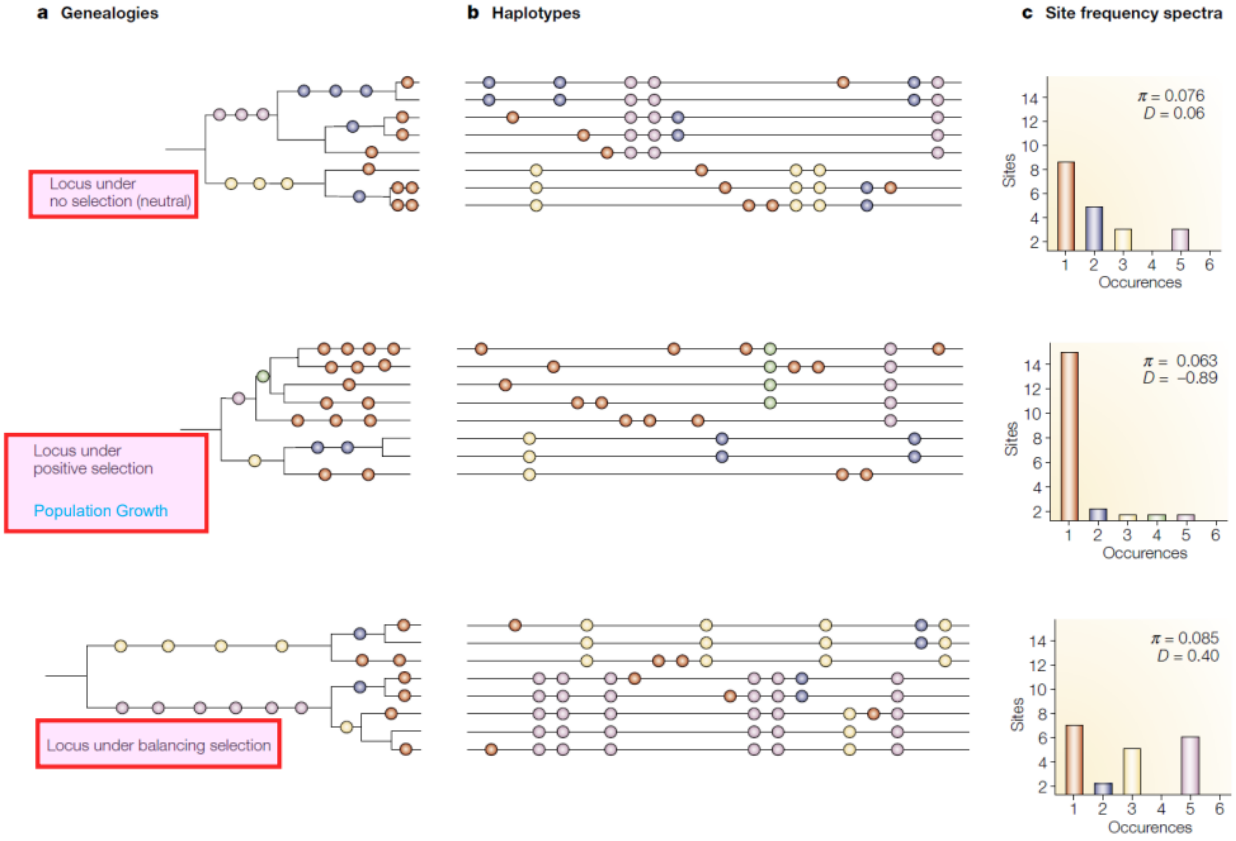


Significant: ($D_{obs} \neq 0$)

Interpreting Tajima's D

- Tajima's D lower than 0: There is an excess of singletons (excess of low frequency polymorphisms relative to expectation). This is related to bottlenecks, selective sweep (which indicate population size expansion) or purifying selection.
- Tajima's D higher than 0: There is a low level of both low and high frequency polymorphisms, indicating a decrease in population size and/or balancing selection.

Value of Tajima's D		Biological interpretation
Tajima's D=0	Observed variation similar to expected variation	Population evolving as per mutation-drift equilibrium. No evidence of selection
Tajima's D<0	Rare alleles present at high frequencies (excess of rare alleles)	Recent selective sweep, population expansion after a recent bottleneck, linkage to a swept gene
Tajima's D>0	Rare alleles present at low frequency (lack of rare alleles)	Balancing selection, sudden population contraction



We can use other tests. Like the tests based on comparisons of divergence and/or variability between different classes of mutation:

- HKA

Neutral Model

In the absence of natural selection:

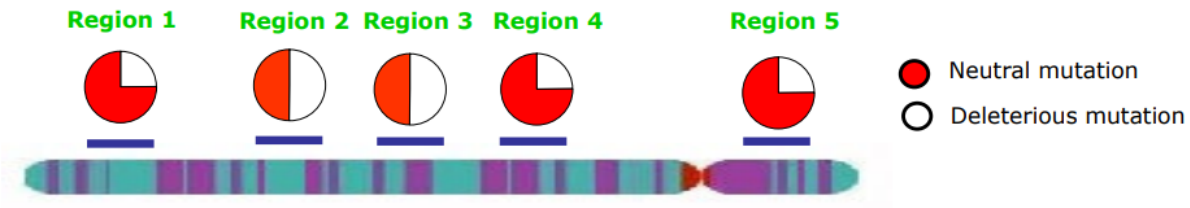
- The levels of polymorphisms within species (intraspecific data; within species) and the levels of divergence (interspecific data; among species) are proportional to the neutral mutation rate.

Polymorphism: $H(\theta) = 4N\mu$

Divergence: $K = 2T\mu$

This is the basis of developing a neutrality test that tries to check if this proportion between divergence and polymorphism occurs or not.

Imagine that we have 5 genomic regions across one chromosome:



The variability levels of the different regions can be different. Because different regions (loci; genes) from the same species can exhibit distinct neutral mutation rates, since different regions can exhibit different functional constraint levels (different levels of the impact of the deleterious mutations).

Even so, the relationship between polymorphism and divergence must be constant for all regions across species.

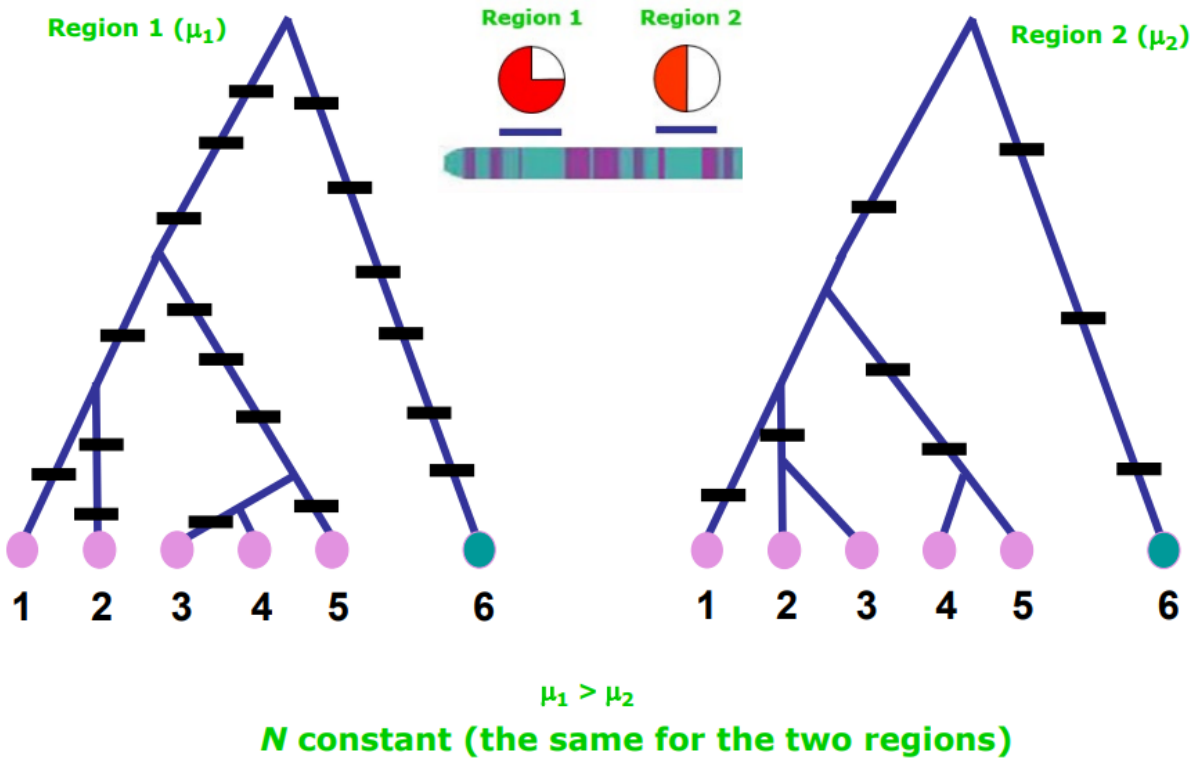
If there is a proportionality of the neutral mutation rate and polymorphism/divergence, then we expect that the regions that have higher levels of variability within species, also have a greater number of nucleotide differences between species.

In this representation we have 2 regions from 5 individuals of the same species and an outgroup.

Region 1 is one gene and region 2 is another gene.

If region 1 is more variable, then the levels of variability within species will be higher than region 2. Also the number of fixed differences between species will be higher.

We can see that in region 1 there are more mutations, thus the regions will be more different (within species and with different species).



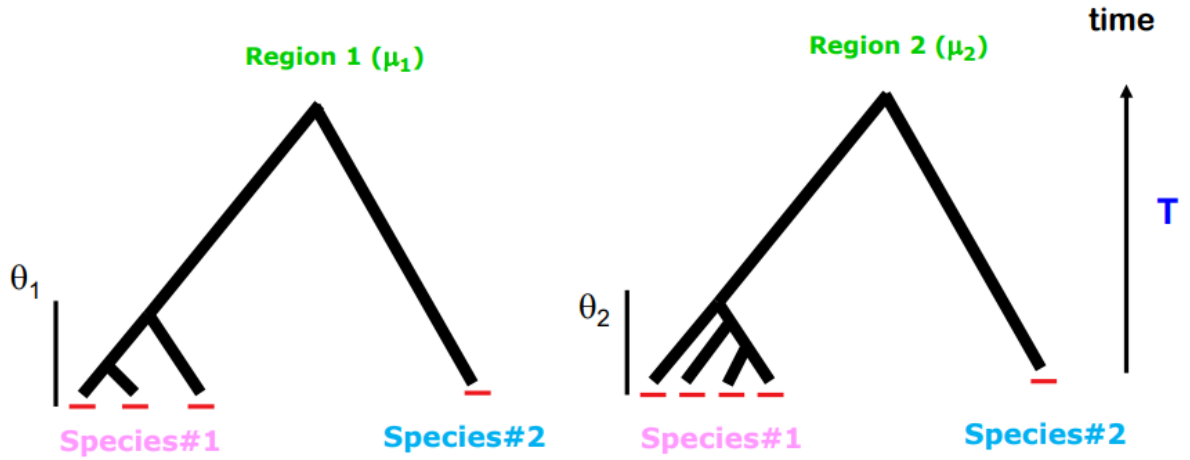
If there is a higher number of polymorphism, there will be a higher difference.

The HKA test compares the polymorphism levels within species, with those of divergence across species, from two (in this case) or more genomic regions.
The neutral model (H0) is rejected if the goodness of fit test between polymorphism and divergence is statistically significant (rejects proportionality).

Example: Imagine that we have 2 loci from 2 different species. Locus 1 has a great divergence and a high polymorphism. Locus 2 has a great divergence but a low polymorphism.
Locus 1 is under neutral selection
Locus 2 is under positive selection.

θ1 represents the nucleotide variation in region 1 of species 1
θ2 represents the nucleotide variation in region 2 of species 1

The divergence is the same in both cases.



$\theta_1 = 4N\mu_1$

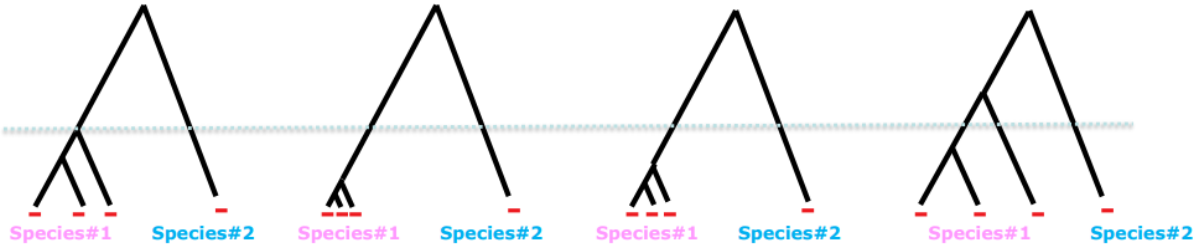
$K_1 = 2T\mu_1$

$$\frac{\theta_1}{K_1} = \frac{\theta_2}{K_2} = \frac{2N}{T}$$

$\theta_2 = 4N\mu_2$

$K_2 = 2T\mu_2$

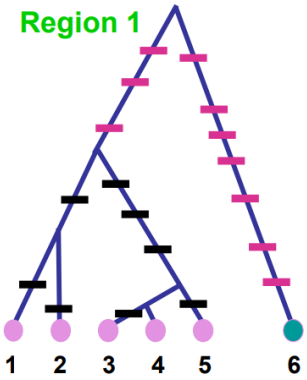
The polymorphism pattern and levels are very sensitive to particular selective scenarios.

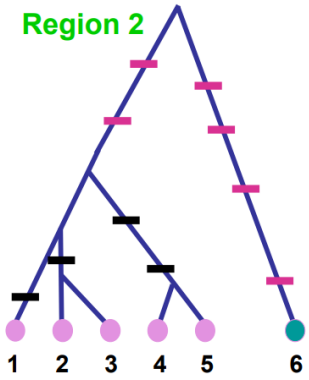


Neutral	Selective Sweep	Background Selection	Balancing Selection
π	π	π	π

The size of pi represents the value that it should have.

In black we have the within species mutations.
In pink we have fixed differences.

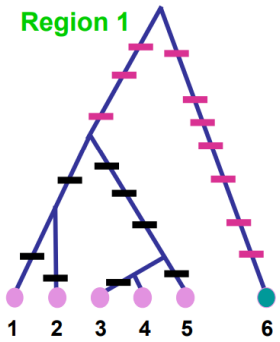


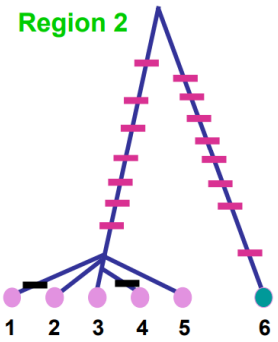


	Region 1	Region 2
Polymorphism	8	4
Divergence	10	6

$8/10 \approx 4/6$

This is a neutral evolving region (there is no evidence for selection). The values are more or less equivalent. So, we don't have enough information to say that one particular region evolves differently.





	Neutral Region Region 1	Region 2
Polymorphism	8	2
Divergence	10	15

$8/10 \not\approx 2/15$

Putative evidence for positive selection (selective sweep).

HKA: Goodness of Fit Test

We don't normally use the contingency table. To compute the test it is needed to build a similar contingence table with the expected values …that are estimated from the data.
The estimated number of mutations (S) for each of the L regions, and their variance.
The estimated divergence between species (D) (between species A and B), and its variance.

The goodness of fit:

$$X^2 = \sum_{i=1}^{L} \frac{S_i^A - \widehat{\mathbf{E}}[S_i^A]}{\widehat{\mathbf{Var}}[S_i^A]} + \sum_{i=1}^{L} \frac{S_i^B - \widehat{\mathbf{E}}[S_i^B]}{\widehat{\mathbf{Var}}[S_i^B]} + \sum_{i=1}^{L} \frac{D_i - \widehat{\mathbf{E}}[D_i]}{\widehat{\mathbf{Var}}[D_i]}$$

Where i, denote the genomic region (i = 1, ..., L); j (j = 1, 2) denote the two species

The statistic X^2 follows a *chi*-square distribution with 2L -2 degrees of freedom

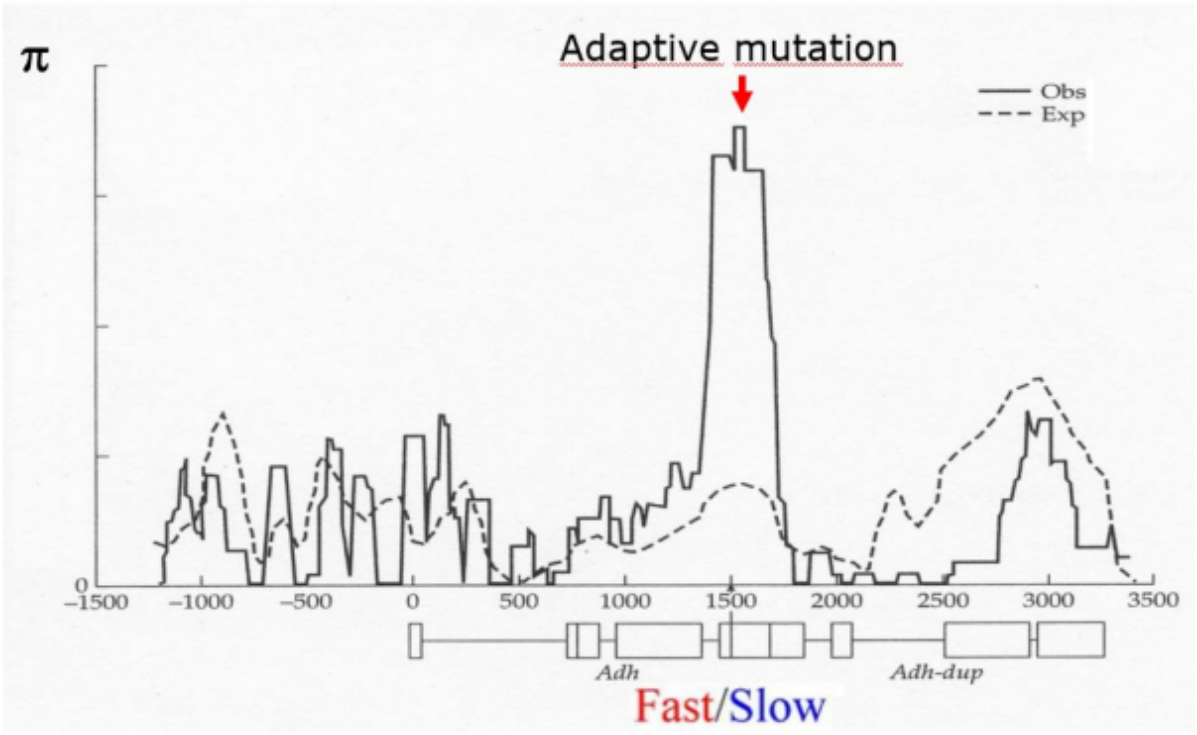
Adh gene: Example

Nucleotide diversity of 11 alleles (4750 bp), encompassing the Adh and Adh-dup genes of D. melanogaster Analysis of 5 Fast alleles and 6 Slow alleles (Adh alleles differing by a amino acid polymorphism) (Kreitman & Hudson 1991).
They asked if there is some evidence that natural selection is acting or not.

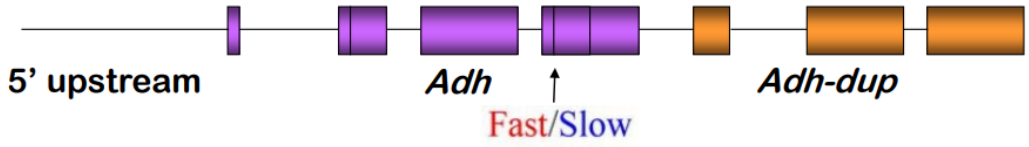
Which is the evolutionary meaning of such amino acid polymorphism? Is it neutral? Or has it been shaped by natural (positive) selection?

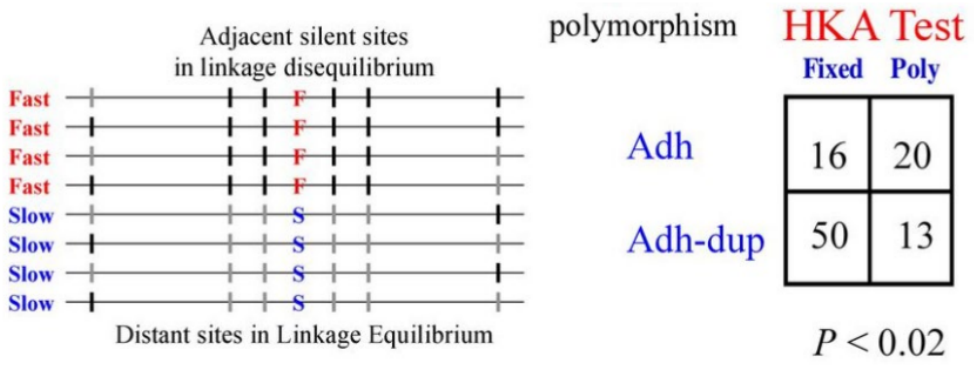
They computed the values of π for the different positions of the alleles. In one particular part of the gene, there is a peak of variation where the observed values are larger than expected.

Sliding-Window Plot (nucleotide diversity over the *Adh* region)



But is this significant?
We can use the HKA test using information of the Adh gene compared to another gene (Adh-dup in this case).



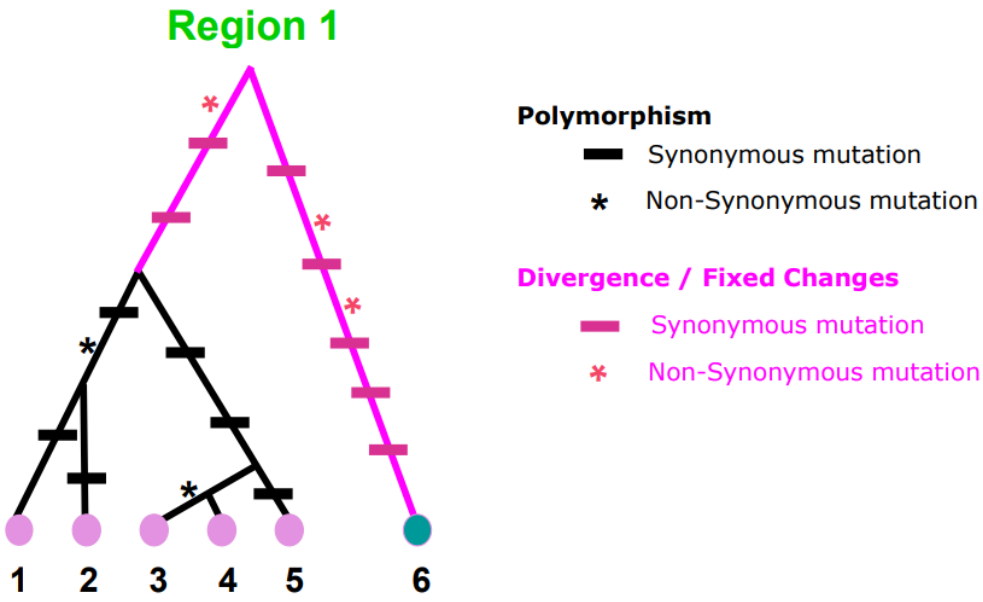


Non-neutral evolving region (consistent with balancing selection)

They computed the number of fixed and polymorphic variants across the Adh and Adh-dup and they found that the number of changes was significant. Thus, there is evidence against Neutral selection.

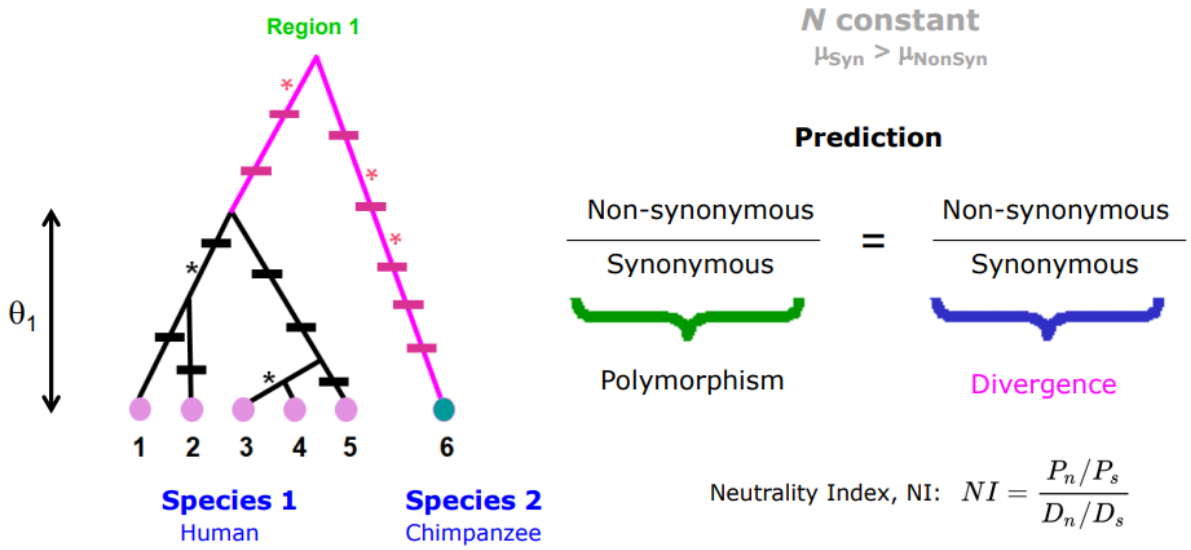
McDonald-Kreitman

It's another test that uses information within and between species. In this case they don't use information from different regions. It only uses information from a particular gene.



In one particular region we have information from individuals of a single species (human) and different species (chimp). We also know the number of synonymous and nonsynonymous mutations.
Obviously, the effect of a nonsynonymous mutation is more relevant.

The MK test compares the number of polymorphisms (within species) with the number of fixed differences (between species), for one particular gene, and for two different classes of mutations (synonymous and nonsynonymous).



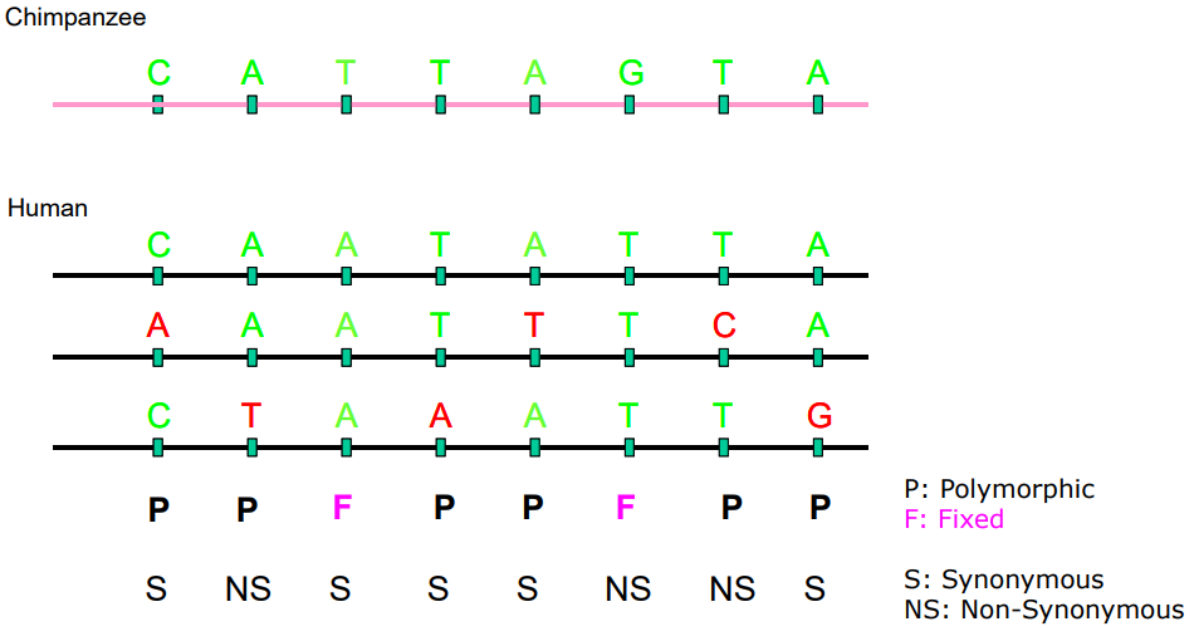
So, we will check the correlation between synonymous and nonsynonymous mutations within species vs synonymous and nonsynonymous mutations between species .

Assuming that population size is constant, usually the number of synonymous mutations will be higher than the non-synonymous.

But, under the neutral model, the ratio should correlate within different species.

If the NI <1, then it's evolving under positive selection. Since it means that there is an excess of nonsynonymous divergence, which is expected under positive selection.
If the NI > 1, then it's evolving under negative selection.

We have 1 sequence from Chimp and 3 human sequences.



The first position is polymorphic because in humans there are 2 possible states.
The third position is fixed because all the human sequences have the same state.

Then we check if they are synonymous or not.

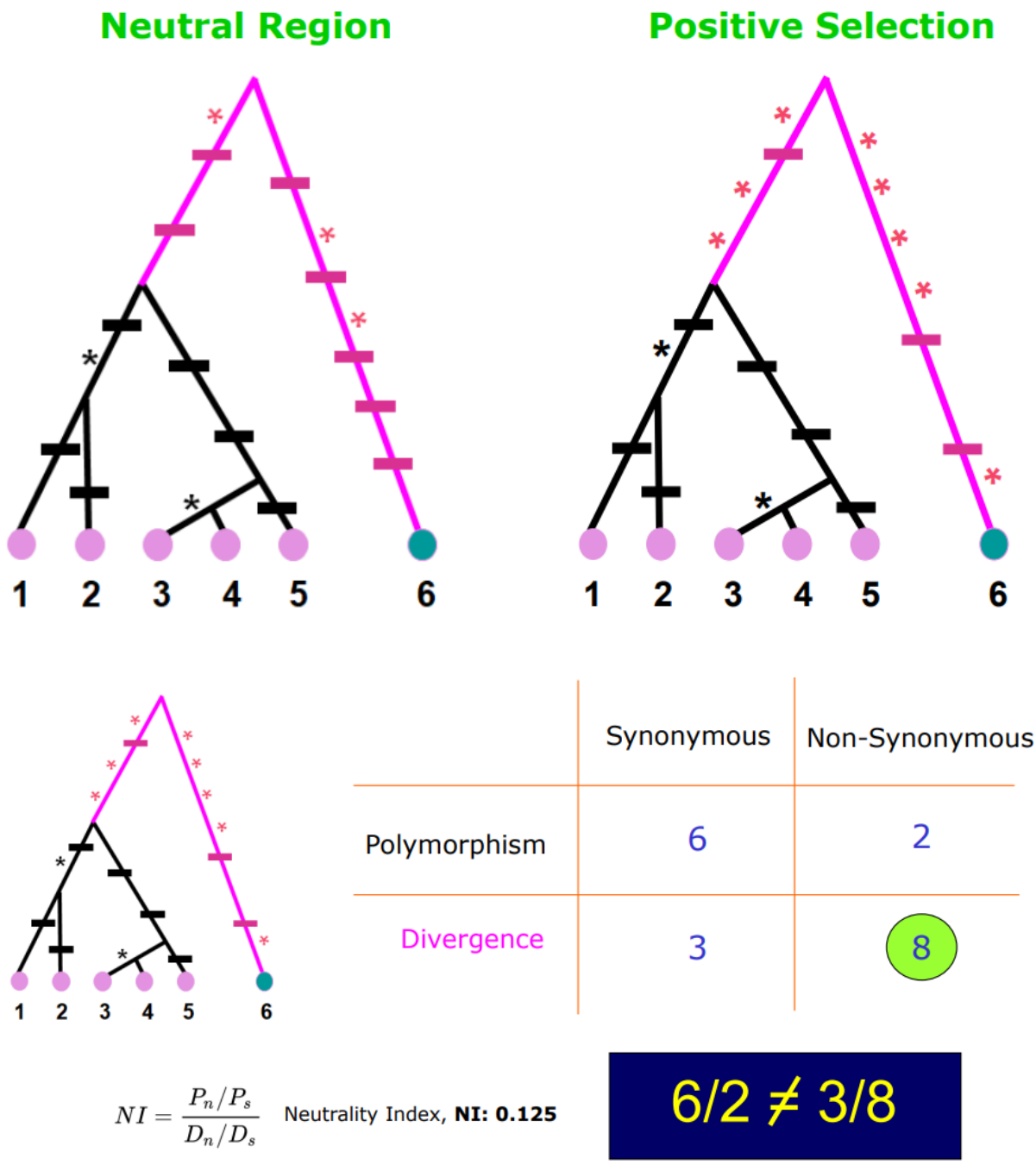
	Synonymous	Non-Synonymous
Polymorphism	6	2
Divergence	7	3

$$NI = \frac{P_n/P_s}{D_n/D_s}$$ Neutrality Index, **NI: 0.777**

$$6/2 \approx 7/3$$

Neutral evolving region (there is no evidence for selection).

In this case.



Non-neutral evolving region (there is evidence for positive selection). Because there are a lot of non-synonymous mutations.

HKA Test: Analyze two (or more) genomic regions, so these regions can exhibit different genealogies, and therefore, important variation across genes.

MK Test: Study linked (within the same gene) nucleotide positions, and therefore there is only a single genealogy.

Lluís Giménez Gabarró

Molecular Clock

Population Genetics: Is a branch of genetics (part of evolutionary biology) that deals with genetic differences within and between populations (intraspecific level), and study phenomena such as adaptation, speciation, and population structure.

Molecular Evolution: Is a part of evolutionary biology that deals with the process of molecular change, change in the DNA, RNA or protein sequence composition across generations (intra- and interspecific levels).
It uses the principles of evolutionary biology and population genetics to explain patterns of variation: the rates and impact of single nucleotide changes, the genetic basis of speciation, adaptation, evolution of development, the origins of new genes, the ways that evolutionary forces influence genomic and phenotypic changes, the analysis of the impact of neutral vs. non-neutral evolution (impact of natural selection).

There are two links: the main link of these concepts (PopGenetics & MolEvol) is that both can be interpreted using the neutral theory of molecular evolution; but, in addition, both are using the same empirical data DNA/protein sequences.

Genetic divergence

We have information about the multiple alignment in pairs between different species about the alpha-globin.

α-globin (141 aa; human)

% Observed amino acid differences

	Shark	Carp	Newt	Chicken	Echidna	Kang	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp	0.90		53.2	51.4	53.6	50.7	47.9	48.6
Newt	0.95	0.76		44.7	50.4	47.5	46.1	44.0
Chicken	0.91	0.72	0.59		34.0	29.1	31.2	24.8
Echidna	0.93	0.77	0.70	0.42		34.8	29.8	26.2
Kang	0.81	0.71	0.64	0.34	0.43		23.4	19.1
Dog	0.84	0.65	0.62	0.37	0.35	0.27		16.3
Human	0.76	0.67	0.58	0.28	0.30	0.21	0.18	

35 aa diff.

Corrected amino acid distance

The value 24.8 means that between human and chicken there is 24.8% of differences. Thus there are 35 different amino acids between humans and chicken.

The lower part of the matrix represents the corrected number of amino acid changes between pairs of sequences.
Thus, between chicken and human, the correct aa distance is 0.28 (which is always higher or equal to the % of observed aa differences).

How can we compute the corrected aa distance from the information about the observed aa distance?

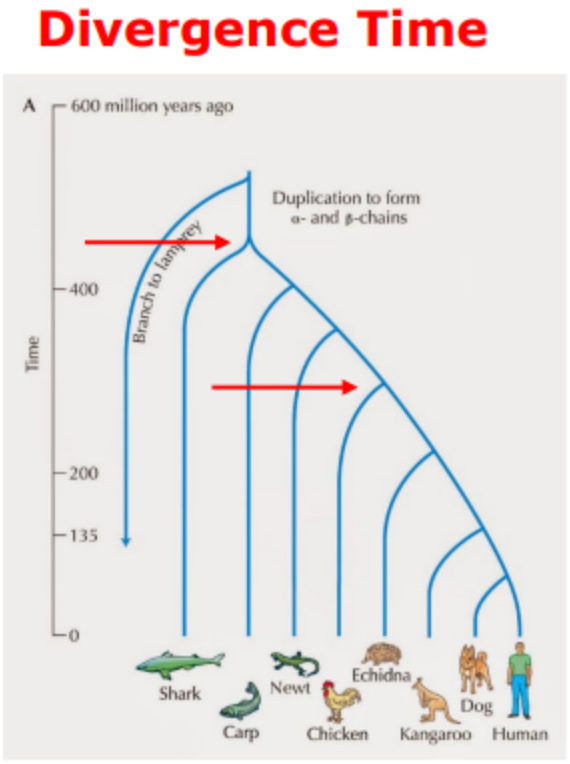
α-globin (141 aa; human). Pairwise human-chicken

```
Seq 1  MNSFSTSAFGPVAFSLGLLLVLPAAFP-APVPPGEDSKDVAAPHRQPLTS
       |...|...|. ||| |||:||...||| :.|..|:.::| ..|:| |:.:
Seq 2  MKFLSARDFHPVAF-LGLMLVTTTAFPTSQVRRGDFTED-TTPNR-PVYT

       SERIDKQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKD
       :.::...|.::|..|..:||||.||.::.|.::.|||||||.||::...|
       TSQVGGLITHVLWEIVEMRKELCNGNSDCMNNDDALAENNLKLPEIQRND

       GCFQSGFNEETCLVKIITGLLEFEVYLEYLQNRF-ESSEEQARAVQMSTK
       ||:|:|:|:|.||:||.:||||:..||||::|.. ::.:::||.:|..|:
       GCYQTGYNQEICLLKISSGLLEYHSYLEYMKNNLKDNKKDKARVLQRDTE

       VLIQFLQKKAKNLDAITTPDPTTNASLLTKLQAQNQWLQDMTTHLILRSF
       .||...::.|:|..|..|.|.:||.|..||::|.:||:..|...||:|.
       TLIHIFNQEVKDLHKIVLPTPISNALLTDKLESQKEWLRTKTIQFILKSL

       KEFLQSSLRALRQM
       :|||:.:||:.||.
       EEFLKVTLRSTRQT
```

Length: 141 AAs **% Observed AA Differences: 35/141 = 24.8% AAs (0.248)**

Differences: 35 AAs **Corrected Value (distance): 0.28 (equal or greater than %Differences)**

Using fossil records, we know that the most common ancestor between chicken and human is 290 mya.

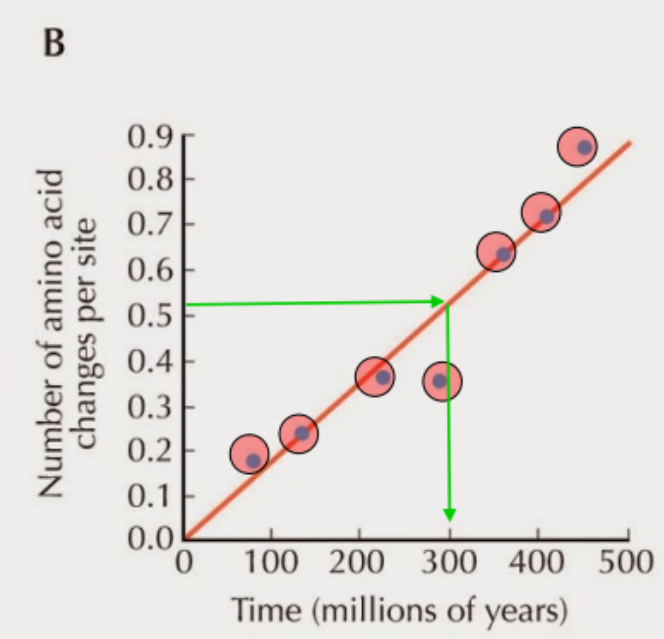


Combining both pieces of information:

- We use the average of the corrected aa distance (genetic distance)
- The divergent time is the same for all species

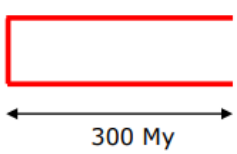
	Shark	Carp	Newt	Chicken	Echidna	Kang	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp	0.90		53.2	51.4	53.6	50.7	47.9	48.6
Newt	0.95	0.76		44.7	50.4	47.5	46.1	44.0
Chicken	0.91	0.72	0.59		34.0	29.1	31.2	24.8
Echidna	0.93	0.77	0.70	0.42		34.8	29.8	26.2
Kang	0.81	0.71	0.64	0.34	0.43		23.4	19.1
Dog	0.84	0.65	0.62	0.37	0.35	0.27		16.3
Human	0.76	0.67	0.58	0.28	0.30	0.21	0.18	
Avg ⟨*K*⟩	0.87	0.71	0.63	0.35	0.36	0.24	0.18	
Time	450	410	360	290	225	135	80	

We can do a regression



We can see that there is a clear relationship between the genetic distance and the divergence distance.

Then we can infer other data (if we know the regression)



K, the number of amino-acid replacements per site. $K = 0.52$

T, Divergence time (how long ago the two sequences split from their common ancestors). $T = 300 \times 10^6$ years

Rate of substitution, r
The number of amino acid (or nucleotide) substitutions that occur per site per year.

$$r = \frac{K}{2T}$$

If K = 0.52 and T = 300 My

Rate of amino acid (or nucleotide) substitution:
$0.52 / (2 * 300 \times 10^{6}) = 0.87 \times 10^{-9}$ substitutions per site and per year

The time for 1% of changes to occur ($K = 0.01$):
$0.01 / (2 * 0.87 \times 10^{-9}) = 5.8 \times 10^{6}$ years

We need 5.8 My so that the alpha globin changes 1%.

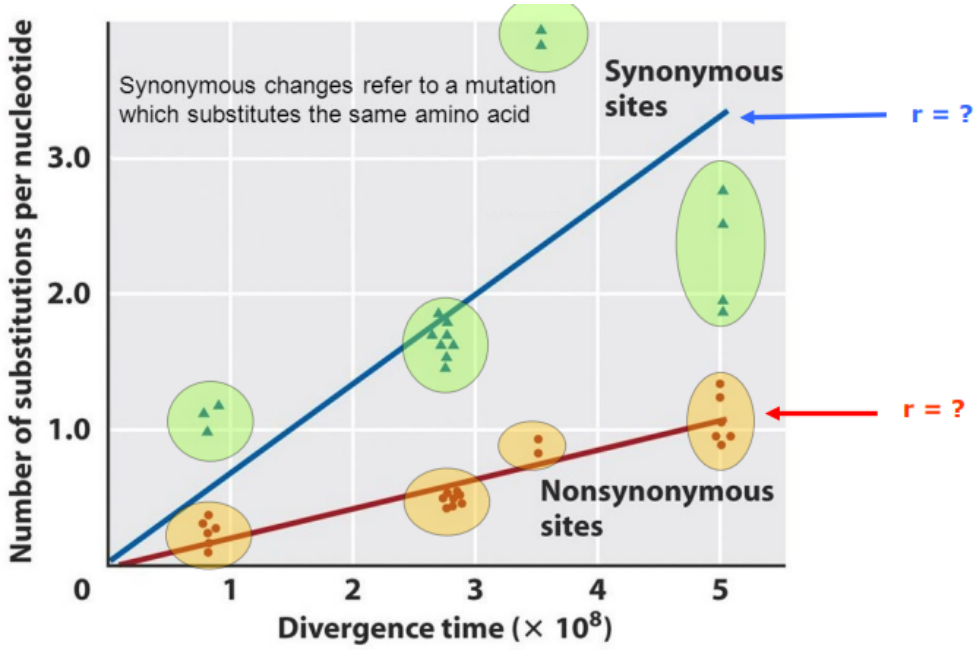
It's important to know that the aa substitution rate is very variable (3 orders of magnitude).

dN(Ka) - dS(Ks)

Refers to a number of methods that compare synonymous (silent, do not change the specified amino acid) and nonsynonymous (amino acid replacement) substitutions (i.e. in the coding region) to infer the action of natural selection.

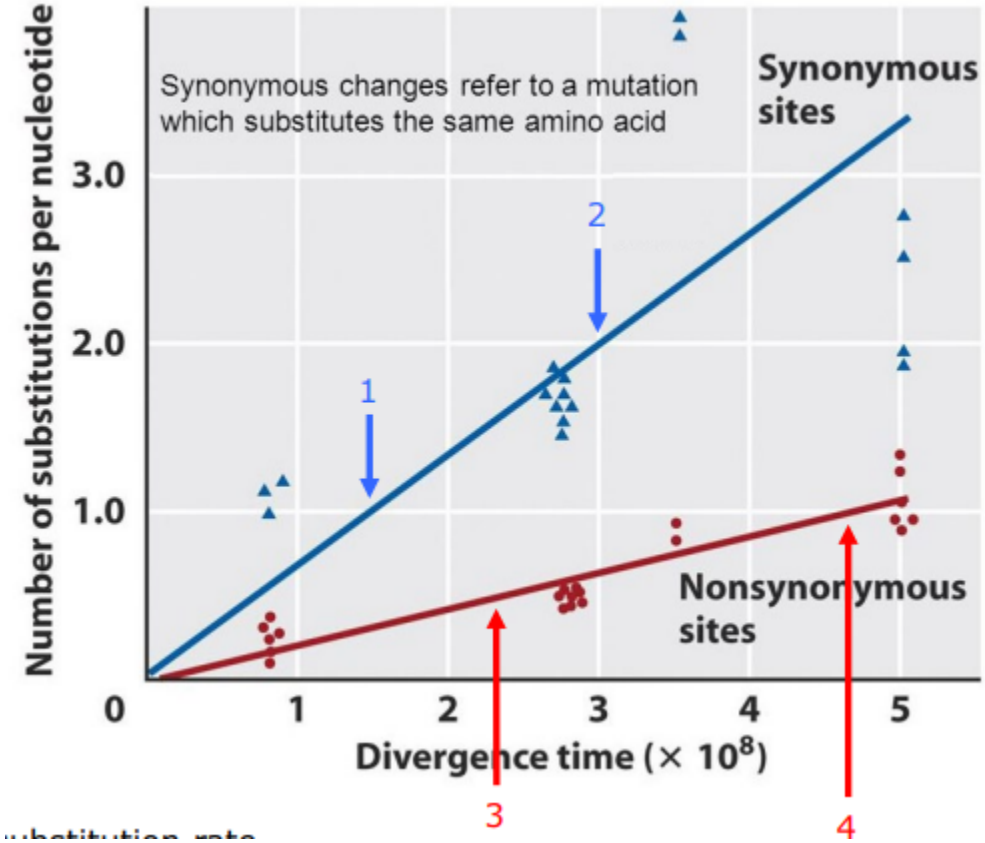
Note that a change in the first or second aa of a codon will always produce a non-synonymous mutation and in the third position sometimes.

Rate of nucleotide substitution
The synonymous substitution rate is higher than nonsynonymous.



Lluís Giménez Gabarró

Exercise



Compute the synonymous substitution rate of point 1 and 2
Compute the nonsynonymous substitution rate of point 3 and 4
Which are the units?
Which is the time to occur a 1% of nonsynonymous changes?

Synonymous substitution rate (SR):

Point 1: $1 / (2 * 1.5 \times 10^{8}) = 3.33 \times 10^{-9}$ synonymous substitutions per site and per year

Point 2: $2 / (2 * 3 \times 10^{8}) = 3.33 \times 10^{-9}$ synonymous substitutions per site and per year

Nonsynonymous substitution rate (NSR):

Point 3: $0.5 / (2 * 2.3 \times 10^{8}) = 1.09 \times 10^{-9}$ nonsynonymous substitutions per site and per year

Point 4: $1 / (2 * 4.6 \times 10^{8}) = 1.09 \times 10^{-9}$ nonsynonymous substitutions per site and per year

The time for 1% of nonsynonymous changes to occur ($K = 0.01$):

$0.01 / (2 * 1.09 \times 10^{-9}) = 4.6 \times 10^{6}$ years

We can see that it's a molecular clock because we obtain the same values (it's a regression)

SR (synonymous) and NSR (nonsynonymous) are as expected by the molecular clock hypothesis.
SR are usually higher that NSR
Great variation of NSR across genes but SR are more conserved

Substitution rates (per nucleotide site and per 10^9 years)

Gene	Pseudogene	Functional genes		
		Position 1	Position 2	Position 3
Mouse ψα3	5.0	0.75	0.68	2.65
Human ψα1	5.1	0.75	0.68	2.65
Rabbit ψβ2	4.1	0.94	0.71	2.02
Goat $\psi\beta^x$ and ψ^z	4.4	0.94	0.71	2.02
Average	4.7	0.85	0.70	2.34

How would you explain these results? Pseudogenes evolve under neutral evolution.
A change in position 1 and 2 causes a non-synonymous mutation.
A change in position 3 can cause a synonymous mutation.
All these genes have a functional constraint and thus it needs to be conserved.

The higher the functional constraint, the lower the evolutionary rate.

The molecular clock hypothesis

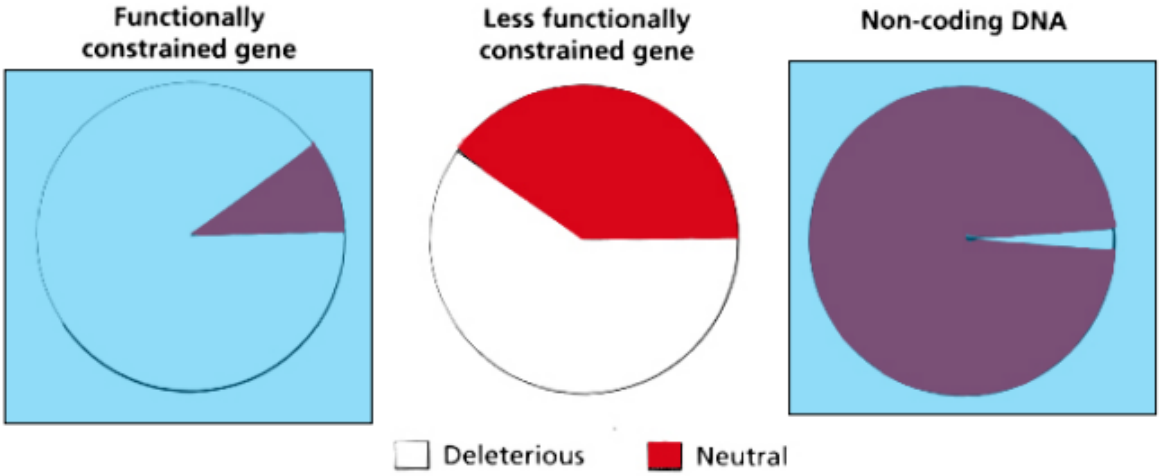
A given DNA/Protein sequence evolves at a relatively constant rate (like a clock).

- Even though it is evolving within very different organisms (human, mice, Drosophila, Arabidopsis, etc) (Different size, generation time, offspring number, etc)
- E.g. α-globin = 1.7 x 10-9 amino acid changes per year

Evolutionary rates vary greatly between genes (or regions within a gene)

- α-globin = 1.7 x 10-9 amino acid changes per year
- Fibrinopeptides = 9 x 10-9 amino acid changes per year
- Histone H4 = 0.01 x 10-9 amino acid changes per year
- Pseudogenes = 4.7 x 10-9 nucleotide substitutions per year

Functional constraints and substitution rates



The neutral theory of molecular evolution

The molecular clock

Substitution rate (for a given protein; DNA region; DNA fragment) -> Constant
Substitution rate among different proteins (DNA regions; DNA fragments; etc) -> Variable (among proteins)

The Neutral theory

At the molecular level, the majority of evolutionary changes are caused by random genetic "drift" through mutations that are selectively neutral.
Describes cases in which positive selection (caused by beneficial mutations) are not strong to outweigh random events.
This is an ongoing process which gives rise to genetic polymorphisms, and later the divergence between species.

The majority of evolutionary change (at the molecular level) is caused by random drift (genetic drift) of selectively neutral mutants.
Genetic Drift: Refers to random changes in allele frequency (especially in small populations) that can lead to the preservation or loss of particular variants. Thus, the molecular clock would be explained by the random fixation of neutral substitutions

The prevailing idea (in the 1960s, 1970s) was that the predominant evolutionary force underlying amino acid or nucleotide substitutions was natural (positive) selection.
Thus, the molecular clock would indicate a constant rate of adaptive substitutions. It is hard to explain how and why adaptive substitutions would occur in a clock-like manner.

Both at the intraspecific level (polymorphism; mutations segregating within species) As well as at the interspecific level (divergence; substitutions fixed over time).

Polymorphism
Neutral mutation is an ongoing process which gives rise to genetic polymorphisms.

Substitution
The complete replacement (fixation) of one allele previously most frequent in the population with another allele that originally arose by mutation.
The neutral theory predicts the rate at which allelic substitutions occur, and thereby the rate at which divergence occurs.
Predicting the substitution rate for neutral alleles requires knowing the probability that an allele becomes fixed in a population and the number of new mutations that occur each generation.

Under the neutral theory:
The substitution rate is equal to the mutation rate

$$r = \mu \text{ (also denoted as } K = \mu)$$

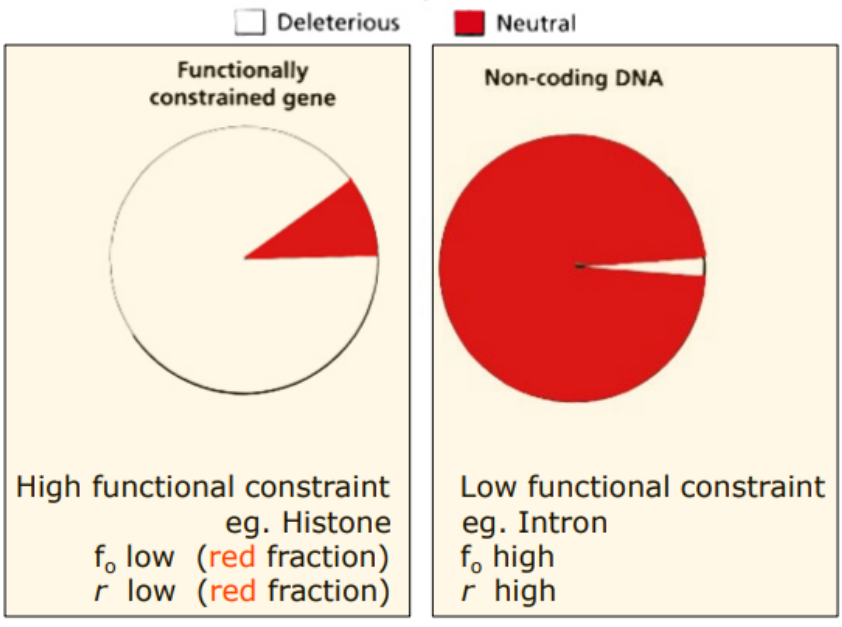
This equation could explain the molecular constant rates in evolution

But, how to explain that different genes could exhibit different evolutionary rates? And that these rates were related to their functional limitation? This concept is not explained by a "putative neutral mutation rate"

Neutral and deleterious mutations

f_o, the fraction of neutral mutations (1- f_o, deleterious)
$f_o = 1$, all mutations are neutral

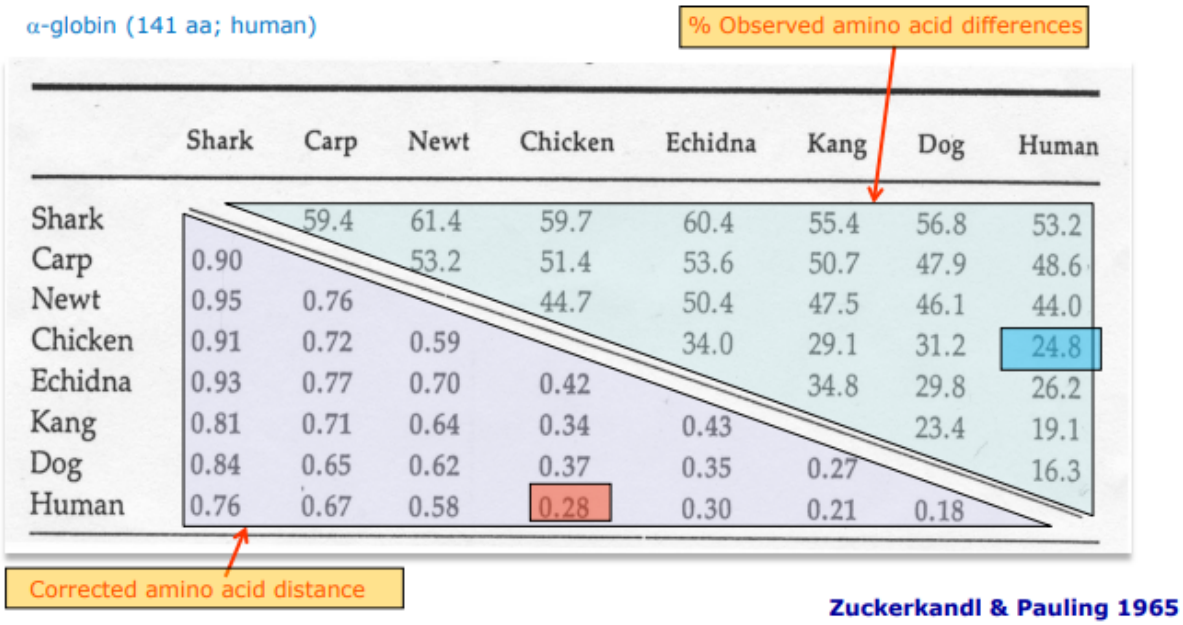
$$\mu_{neutral} = f_0 \, \mu_{tot}$$



So, the neutral mutation rate is just a fraction of the total mutation rate. This explains why different genes evolve in a different manner.

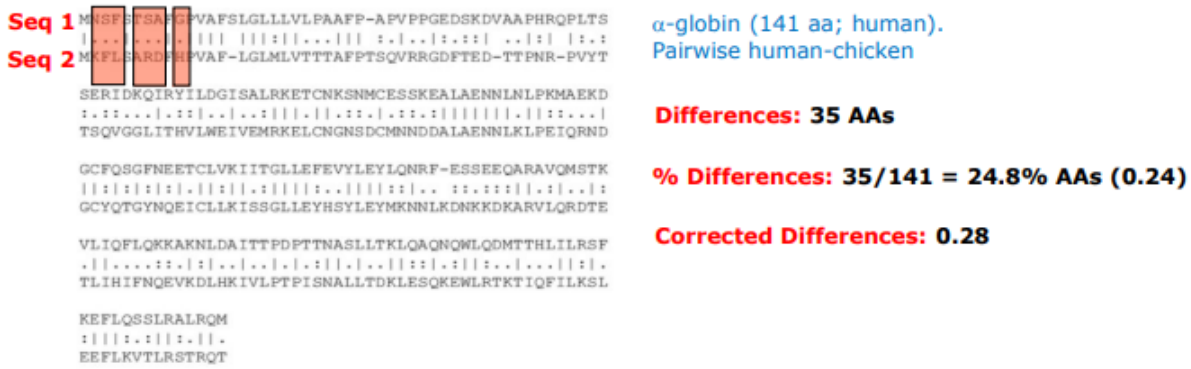
Genetic divergence (amino acid substitution per site)

As we said, on top we have the % of observed aa differences between a certain gene of different species. But we normally use the corrected aa change or corrected nucleotide distance.



	Shark	Carp	Newt	Chicken	Echidna	Kang	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp	0.90		53.2	51.4	53.6	50.7	47.9	48.6
Newt	0.95	0.76		44.7	50.4	47.5	46.1	44.0
Chicken	0.91	0.72	0.59		34.0	29.1	31.2	24.8
Echidna	0.93	0.77	0.70	0.42		34.8	29.8	26.2
Kang	0.81	0.71	0.64	0.34	0.43		23.4	19.1
Dog	0.84	0.65	0.62	0.37	0.35	0.27		16.3
Human	0.76	0.67	0.58	0.28	0.30	0.21	0.18	

A genetic distance is a measure of the genetic divergence between species or between populations within a species



K, (is a genetic distance) the number of amino-acid replacements per site

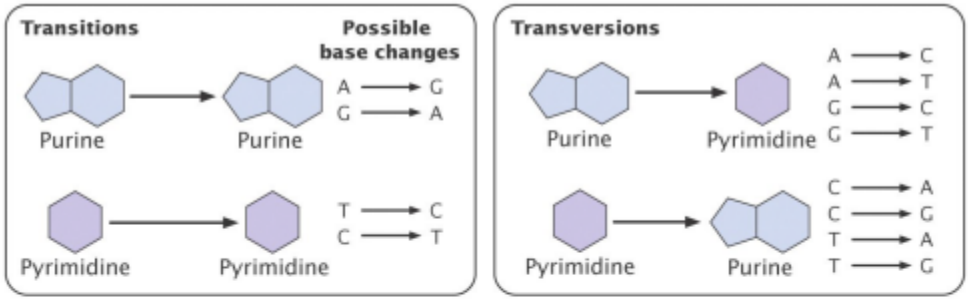
Uncorrected value, (eg, the fraction of AA differences) = 0.24

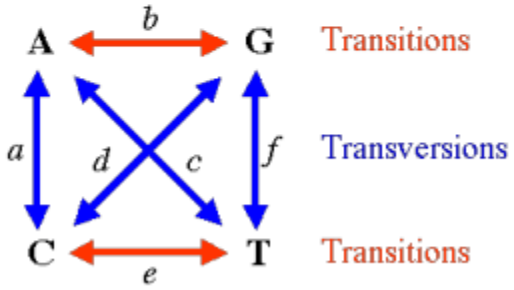
***K*, Corrected** = 0.28 (equal, but usually greater than the uncorrected value)

We need to correct the distance. We can't use the number of observed differences.

Lluís Giménez Gabarró

Mutations and Multiple Substitutions





Reversible Model

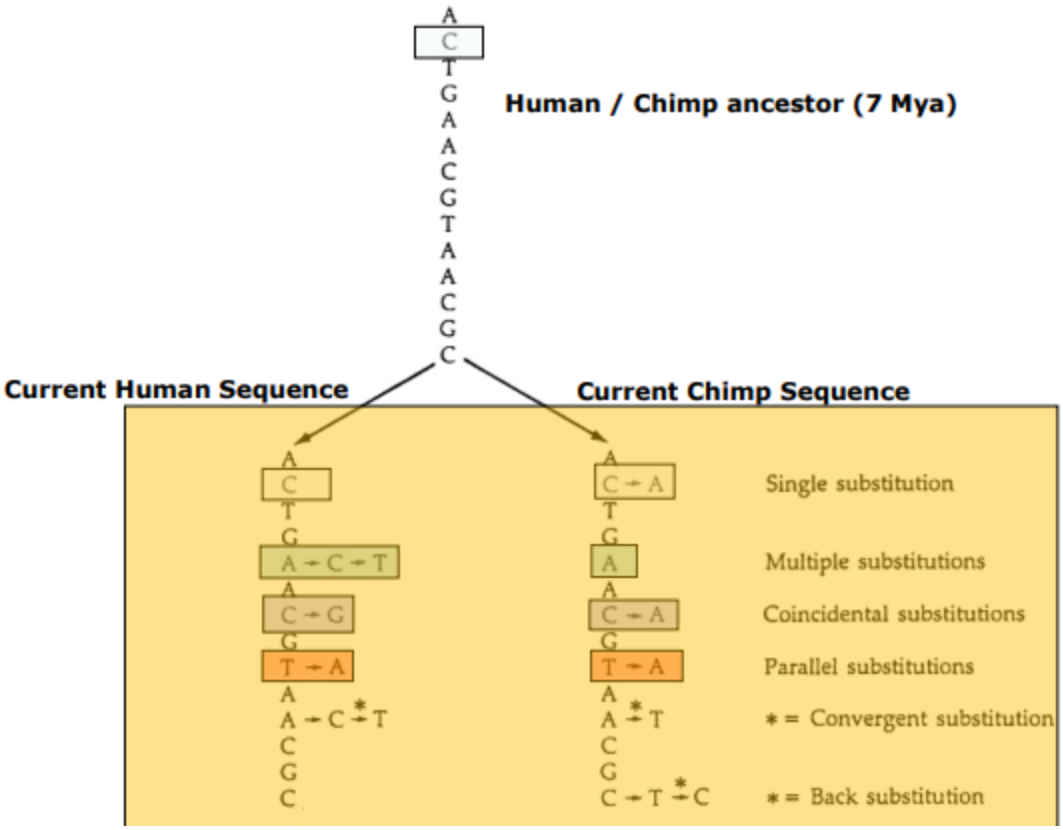
As we can see, there are 4 possible transitions and 8 possible transversions (the main nitrogen base is mutated). This is a reversible model because the probability of going from A to T is the same as going from T to A.

When we have 2 sequences, we can see that there has been a mutation. But we do not know which strand has been mutated.

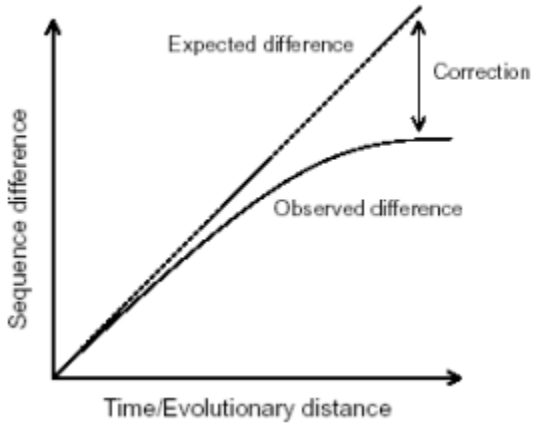
Let's suppose that we know which are the first 20 nucleotides of an intron of the human/chimp ancestor.
Now we try to determine the different mutations that have occured in the human and chimp lineages.

As we can see, many things can happen:

- Single substitution
- Multiple substitutions
- Coincidental substitutions
- Parallel substitutions
- Convergent substitutions
- Back substitutions



Thus, the number of differences that we can see is actually lower than the number of real mutations. Thus we need to make a correction.

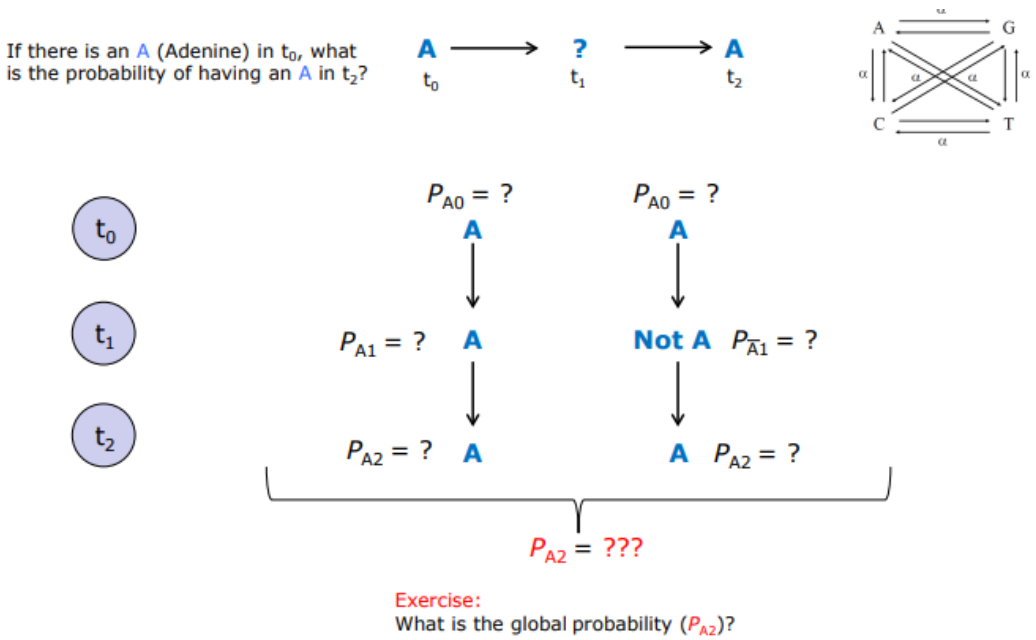


So, the correction is the difference between expected and observed differences.

Lluís Giménez Gabarró

Models of nucleotide substitution

We cannot repeat evolutionary history to study the nucleotide substitution process; we rely on mathematical models that account for the nucleotide substitution process.
To study these dynamics we must make several assumptions about the probability of substitution of one nucleotide by another.
The Jukes-Cantor (JC69) one parameter model (1969) assumes that each nucleotide has equal probability to be substituted by any of the other three in a fixed period of time.

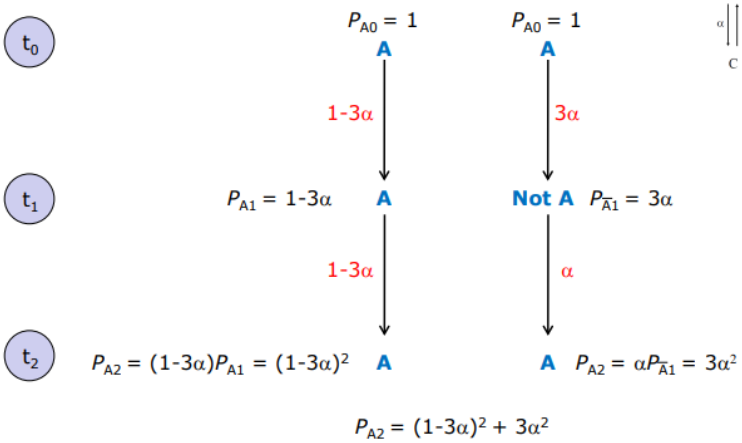


We can compute this by looking at all the possibilities:

- Probability of not changing of nucleotide in T1 and T2
- Probability of changing to any other nucleotide in T1 and returning to A in T2.

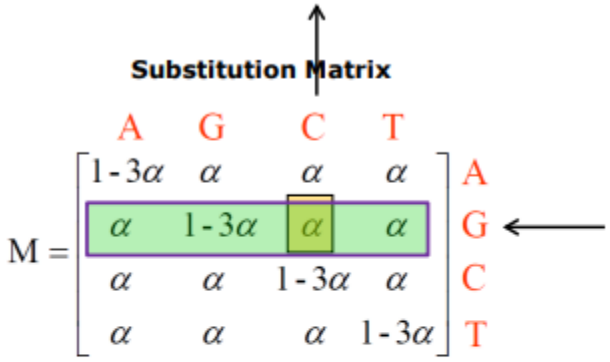
Thus, we just need the transition probability matrix.

- Probability of A to A is 1-3α. Because α is the probability of changing to any other nucleotide that is not the same.
- Probability of changing to any other nucleotide is 3α, because there are 3 possible nucleotides to which you can change.
- Probability of returning to A is α.



Then we add the probabilities.

We can also see this is the substitution matrix:



M_{ij} probability to change nucleotide i (row) to j (column), per unit of time (the sum over in each row should be 1)

M_{GC} probability of substitution of G (row) to C (column) (G→C):

$M_{GC} = \alpha$

True Empirical Data

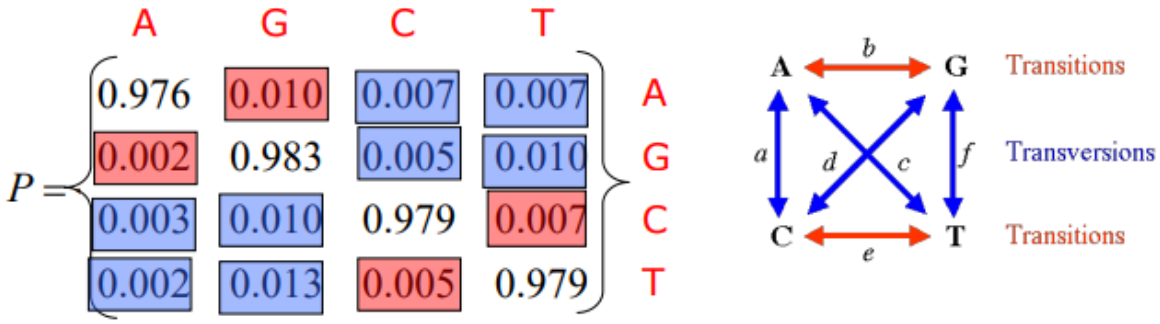
$$P = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} 0.976 & 0.010 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.010 \\ 0.003 & 0.010 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{pmatrix} \end{array}$$

P_{AC} = 0.007 (probability that a site that started with A had nucleotide C, at the end of a time interval).

P_{CC} = 0.979 (probability that a C remains unchanged), per unit of time. That is, there has been no (apparent) change at this site.

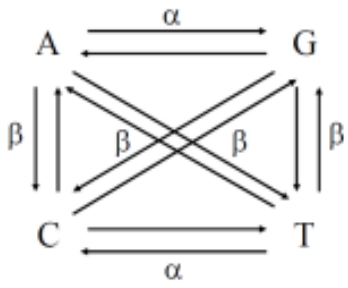
The rows of the matrix sum to 1 (all possible states are represented, one of has to be chosen)

We know that the number of transitions should be higher than the number of transversions.



The Jukes-Cantor model assumes that each nucleotide has equal probability to be substituted by any of the other three in a fixed period of time.
However, it is often observed (real data) that transitions occur more frequently than transversions. To explain this, we use the Kimura two-parameter (K2P) model which has two parameters α (for transitions) and β (transversions). Usually α > β

Note that JC69 is simple a particular case of K2P in which α = β



$$M = \begin{array}{c} \begin{matrix} A & G & C & T \end{matrix} \\ \begin{bmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{bmatrix} \end{array} \begin{matrix} A \\ G \\ C \\ T \end{matrix}$$

M_{ij} probability of substitution of nucleotide i (row) to j (column) per unit time.

M_{AT} probability of substitution of A→T = β

General Time-Reversible model (GTR)
An important property of real sequences (not accounted for JC69 or K2P models) is that the frequencies of the four bases are often different. The equilibrium frequency of the bases are denoted by πA, πC, πG and πT.
The GTR is the most general time-reversible model. GTR allows variable instantaneous rates of substitution between each of the six nucleotide pairs: a = A <-> C, b = A <-> G, c = A <-> T, d = C <-> G, e = C <-> T, and f = G <-> T.
The Q (substitution matrix) represents the instantaneous rate of change from base i to base j. The diagonal elements are omitted for clarity.

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & — & a\pi_C & b\pi_G & c\pi_T \\ C & a\pi_A & — & d\pi_G & e\pi_T \\ G & b\pi_A & d\pi_C & — & f\pi_T \\ T & c\pi_A & e\pi_C & f\pi_G & — \end{array}$$

We have dealt with models assuming that changes are equally likely across positions, and among phylogenetic lineages.
Real data often show variability of substitution rates between sites; even some sites might be completely invariant (affected differentially by purifying selection).
Additionally, the three nucleotides of a particular codon (coding region) usually exhibit different substitution rates.
And, it is also possible that substitution rates vary along time (heterotachy).
There are a number of models and methods dealing with that!

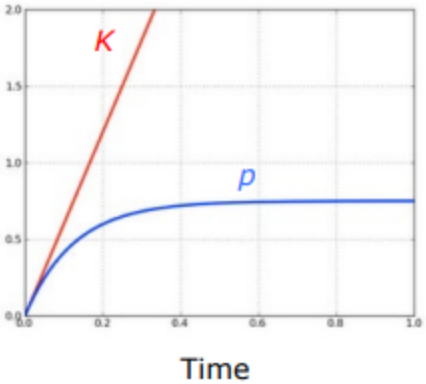
Number of Nucleotide Substitutions

The number of nucleotide substitutions per site between 2 sequences (K) –or divergence- can be estimated from the number of differences between the two sequences. K is a genetic distance measure.
If the degree of divergence between the two sequences is substantial, the observed number of differences is likely to be smaller than the actual number of substitutions due to multiple substitutions (multiple hits) at the same site.
There are many methods to correct this effect, e.g. by the Jukes-Cantor model.
Nevertheless, we do not know the divergence time (we cannot estimate α), but we can estimate K from p (proportion of differences between two sequences).

Jukes and Cantor correction: $K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$

A genetic distance is a measure of the genetic dissimilarity between different species (divergence), or individuals of the same species (polymorphism). It can be used to compare the genetic relationships among individuals or species (i.e, applying phylogenetic methods).

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$



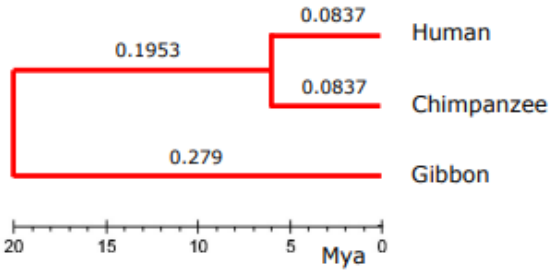
What is the range of *p*?
What is the range of *K*?

p = 0.01 -> *K* = 0.01001 (99% identical)
p = 0.05 -> *K* = 0.0517
p = 0.10 -> *K* = 0.1073
p = 0.15 -> *K* = 0.1674
p = 0.20 -> *K* = 0.233
p = 0.30 -> *K* = 0.383
p = 0.40 -> *K* = 0.536
p = 0.50 -> *K* = 0.824
p = 0.60 -> *K* = 1.21 (40% identical)
p = 0.70 -> *K* = 2.03
p = 0.74 -> *K* = 3.24
p = 0.749 -> *K* = 4.97
p = 0.75 -> *K*: not applicable

P goes from 0 to 1
K goes from 0 to inf
But we can not use the formula when p > 0.75

Gen A; 1200 bp

$$r = \frac{K}{2T} \qquad K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$



Nucleotide Differences Between Human & Chimpanzee: 180 bp

Proportion of Differences Between Human & Chimpanzee: p = 180 / 1200 = 0.15

Number of Nucleotide Substitutions per site Between Human & Chimpanzee: K = 0.1674

Nucleotide Substitution Rate Between Human & Chimpanzee: r = 0.1674 / 2*6,000,000 = 1.395×10^{-8}
Nucleotide Substitutions per nucleotide site and per year

Nucleotide Differences Between Human & Gibbon: 473 bp

Proportion of Differences Between Human & Gibbon: p = 473 / 1200 = 0.394

Number of Nucleotide Substitutions per site Between Human & Gibbon: K = 0.558

Nucleotide Substitutions Rate Between Human & Gibbon: r = 0.558 / 2*20,000,000 = 1.395×10^{-8}

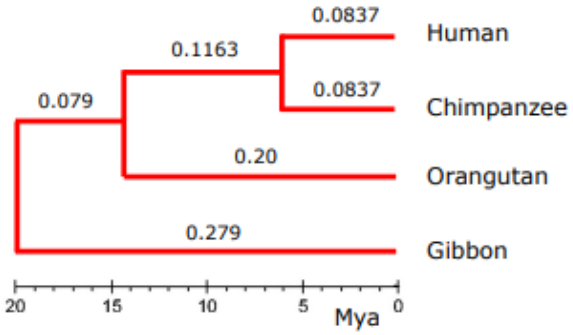
We can see that “r” is the same because it is corrected. Thus, we can know the divergent time between 2 species (without fossil record) using the formulas all the way around.
We just need to know “K”, which can be found by comparing 2 sequences.

Gen A; 1200 bp
$r = 1.395 \times 10^{-8}$

What is the divergence time between Orangutan and Human?

Nucleotide Differences Between Human & Orangutan: 372 bp ➡ p = 0.31; K = 0.40

Number of nuc. Substitutions per site Between Orangutan & Human-Orangutan ancestor: $K/2$ = 0.20



Divergence time Between Orangutan & Human: T = 0.4 / 2*1.395×10^{-8} = ~14.300.000 years

Lluís Giménez Gabarró

Molecular phylogenetics

The study of evolutionary relationships among organisms by using molecular data:

- Reconstruction of the evolutionary history of organisms
- Investigation of the mechanisms of evolution

We need molecular phylogenies to understand how changes are occuring during evolution. It is the global map where you can see all the mutations and events that happen through evolution.

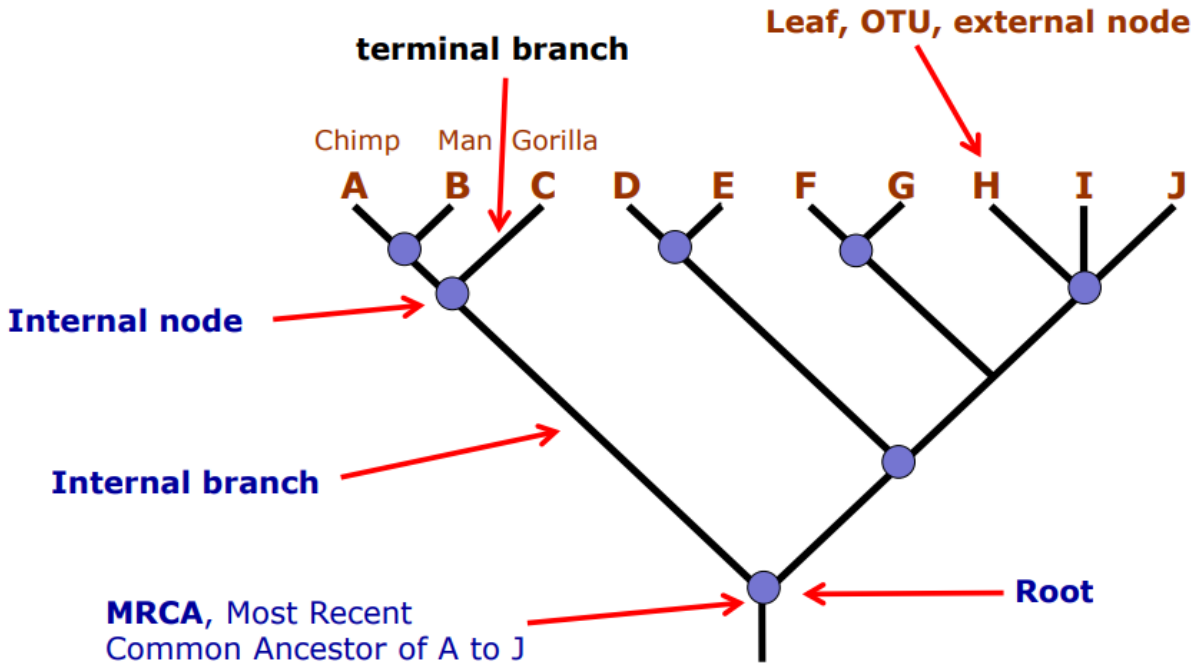
Using molecular phylogenetics we have seen that eukaryotes come from bacterias. We can't see this using other phylogenetic methods.

The step to reconstruct a phylogenetic tree are:

- DNA extraction (sampling)
- DNA sequencing
- MSA
- Tree inference

If we don't choose the correct taxa, we will reach erroneous conclusions.
Increasing the number of markers adds information but can also raise the risk of introducing conflicting signals.

Here we have the different parts of a tree:



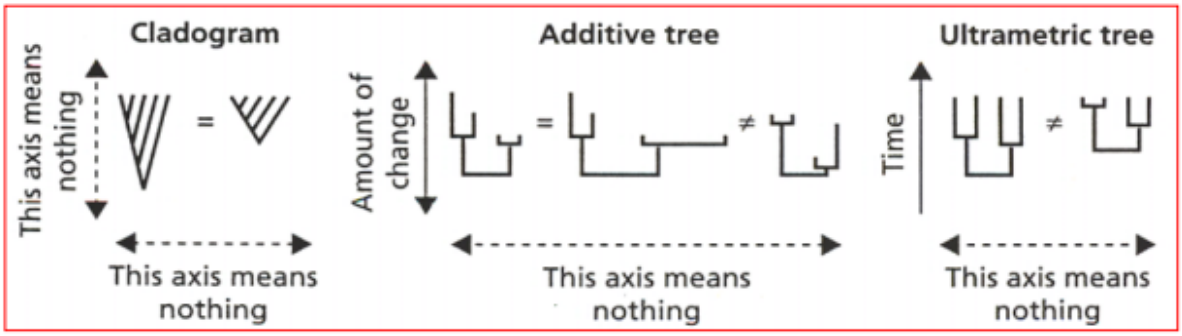
Dichotomy: Node that is divided in 2 species.
Polytomies: Failure to resolve the branching order. Reason why the branch divides into 3 or more species (H+I+J).
Sister taxa: Descendant of the same node (dichotomy):

- D and E are a sister taxa
- The clade of species F+G is the sister group to the clade of species H+I+J

Monophyletic: All taxa within the group derive from a single common ancestor
Paraphyletic: The group contains some, but not all, of the descendants from a common ancestor common; members do not form a natural clade

Unscaled trees: Branch lengths are not proportional to the number of changes
Scaled trees: Branch lengths are proportional to the number of nucleotide (or amino acid) changes

An outgroup is a species/lineage/gene that is not descended from the most recent common ancestor of the ingroup species that we are considering.



Methods to reconstruct the phylogenetic tree:

- Distance Matrix methods (UPGMA, NJ)
- Maximum Parsimony methods (minimize the number of changes)
- Maximum Likelihood methods
- Bayesian methods

UPGMA

Assumes that the substitution (evolution) rate is constant across lineages.So it will cluster the species that have the lowest value in the UPGMA matrix and then it will reconstruct the matrix…

	1	2	3	4	5	6	7	8	9	10	11
1. D. melanogaster											
2. D. pseudoobscura	0.154										
3. S. lebanonensis	0.236	0.234									
4. S. albovittata	0.307	0.298	0.324								
5. D. crassifemur	0.292	0.299	0.326	0.099							
6. D. mulleri	0.256	0.233	0.252	0.254	0.256						
7. D. affinidisjuncta	0.307	0.288	0.303	0.209	0.212	0.205					
8. D. heteroneura	0.305	0.283	0.306	0.204	0.224	0.205	0.038				
9. D. mimica	0.288	0.276	0.294	0.192	0.215	0.205	0.063	0.055			
10. D. adiastola	0.315	0.300	0.304	0.207	0.222	0.208	0.063	0.062	0.059		
11. D. nigra	0.313	0.277	0.286	0.201	0.212	0.208	0.097	0.088	0.079	0.097	

d_{ij}, genetic distance between sequence ***i*** and ***j***

d_{ij} could be K_{ij} (the number of substitutions per site between sequence ***i*** and ***j***) corrected by Jukes and Cantor.

For example, if sequences ***i*** and ***j*** have 100 nucleotides, and there is no gaps; the MSA has 100 positions (sites or columns).
Let me suppose that there are 20 differences (between ***i*** and ***j***)

$p = 20/100 = 0.2$
d_{ij} (**=** K_{ij}) = 0.233

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

An UPGMA tree is always rooted and ultrametric.
An assumption of the algorithm is that the molecular clock is constant for sequences in the tree.
If there are unequal substitution rates, the tree may be wrong. While UPGMA is simple, it is less accurate than the neighbor-joining (NJ) approach.

NJ

Does not assume that the substitution rate is constant across lineages
Produces unrooted trees

NJ, finding the shortest (minimum evolution) tree by finding neighbors that minimize the total length of the tree. Shortest pairs are chosen to be neighbors and then joined in distance matrix as one OTU

Methods to infer selective pressures in genomics studies

$K_s \simeq d_s \rightarrow$syn subs per syn site

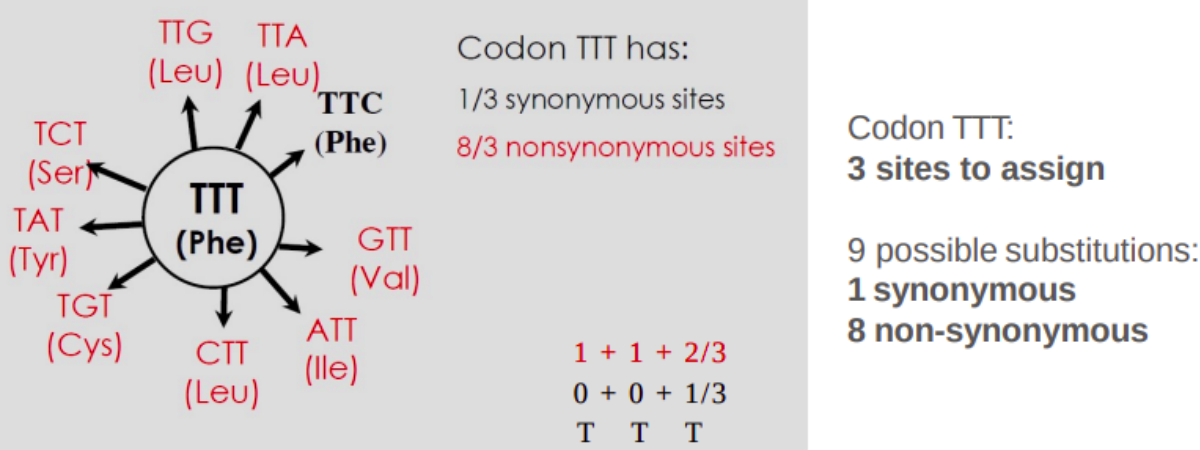
$K_a \simeq d_n \rightarrow$non syn subs per non syn site

$\omega = d_n / d_s$

The parameter "w" measures the type and strength of natural selection:

- If $\omega = 1$ non syn mutations were fixed in the same rate than syn mutations → neutral evolution
- $\omega < 1 \rightarrow d_s > d_n$ some syn mutations weren't fixed due to neg selection
- $\omega > 1 \rightarrow$diversifying (positive) selection

Counting sites method → based on all possible changes of a codon → evaluation of syn / non syn



Count substitutions (differences)
Comparison between 2 sequences or more given phylogenetic info
Apply corrections: as models are based on nucleotide subs and we're trying to apply something to codons,
Could lead to underestimation of ω 'cause most of syn subs are in transition
Codon usage bias: ignoring highly expected genes →false positive selection

Markov model of codon evolution → subs model for codons
One step allows calculation of ω & k
correction for multiple hits
weight evolutionary pathways between codons in phylogeny

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at 2 or 3 positions} \\ \pi_j, & \text{for syn. transversion} \\ \kappa\pi_j, & \text{for syn. transition} \\ \omega\pi_j, & \text{for nonsyn. transversion} \\ \omega\kappa\pi_j, & \text{for nonsyn. transition} \end{cases}$$

probability of syn transversion
same taking into account bias

Synonymous:

CTC (Leu) → CTG (Leu) π_{CTG}

TTG (Leu) → CTG (Leu) $\kappa\pi_{CTG}$

Non-synonymous:

GTG (Val) → CTG (Leu) $\omega\pi_{CTG}$

CCG (Pro) → CTG (Leu) $\kappa\omega\pi_{CTG}$

Likelihood function to calculate just as any MMC → calculation for whole sequence
Result: probability of getting the analyzed sequence

Max likelihood function calculation by finding proper parameters for ω / k … by different algorithms
K is like a control for bias, it's calculated before
Found ω → compare < > = 1 → pos neg neutral selection over average ω between sequences of phylogenetic tree

Based on this idea → use different pressure varies across lineages (each ω can have different weight)
Selection pressure varies along sites (binding sites for enzymes, …)

Bayes empirical Bayes

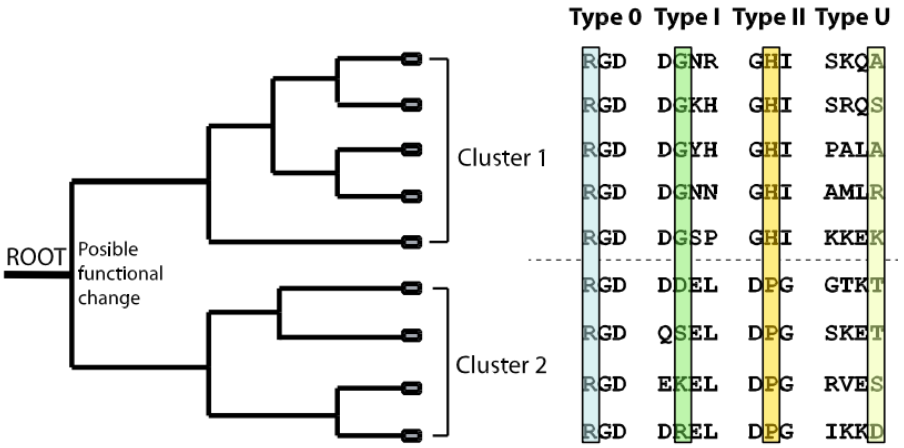
Prior info → to partition sites → different ω → statistical distribution to model ω variation
Estimation of the distribution, not the best / worst codon
We get a list of candidates based on probabilities, not robust distribution

P value reflects how much positive selection was calculated for each tissue / physiological process
If studied tissues are exposed to environment circumstances that favor changes
Good reason to use this method

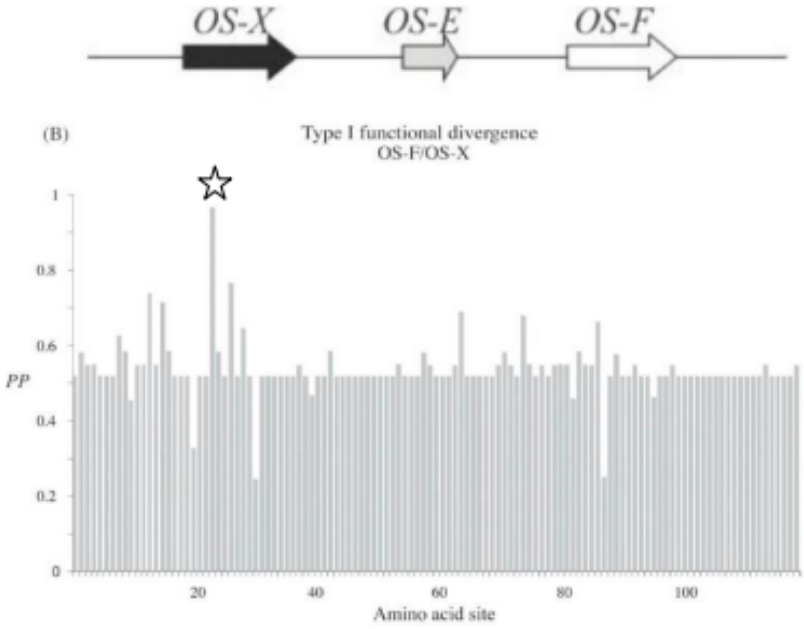
Genes that are not functionally annotated →mRNA (it's useful but there's no explicit function linked)

Method functional divergence in pros

Based on 2 homolog pros that start evolving independently till they get speciated // or not
Prot alignment for pros of the same family in rel species
Look for patterns to infer importance of amino acids by species



Create max likelihood model based on Θ →calculate functional divergence probability



To study gene families
Orthologs
etc