BLOCK 1: POPULATION GENETICS

GENETIC VARIATION

Fixism vs. evolution

Fixism	Evolution
Everything that exists, including all life forms, has not varied along time and everything is today exactly identical to how it was in the past and will be in the future. Variation = Inconvenience	Change in the heritable characteristics of populations over successive generations Variation = Essential

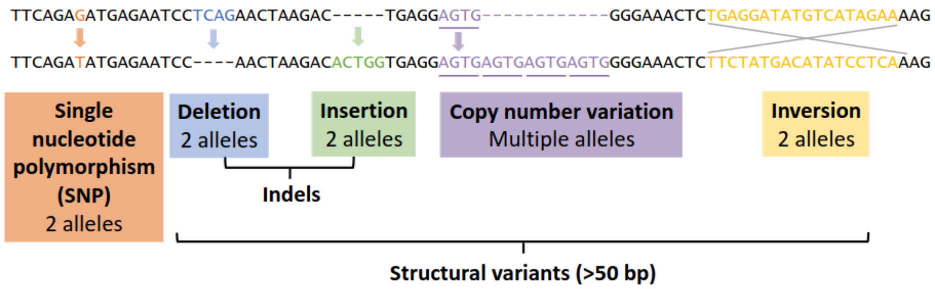
Evolution takes place in populations, which change over time.

Polymorphism: genetic differences within a species, there are 2 possible versions for a particular character for that species.
Generated by mutation, mistakes made randomly when copying DNA.
When the variant has increased the frequency of the population (>5% or >1%) we consider them polymorphisms.

Divergence: genetic differences between species.
Reproductive isolation → Fixation → Separate evolution
When a variance is fixed over a population, the polymorphism has become divergence.

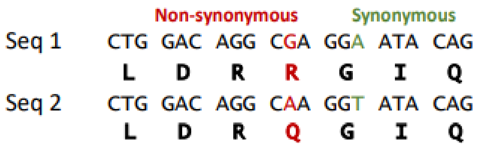
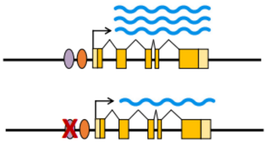
Genotype: set of alleles of an individual at one or many genes/locus/positions of the genome.
Phenotype: morphological, biochemical, physiological or behavioral attributes of an individual (anything that you can measure).
Gene: DNA sequence that codes for an RNA or protein.
Allele: variant or alternative form of the DNA sequence at a given gene / copy of a gene in a diploid organism.
Diploid organisms: organisms with 2 sets of chromosomes ← 2 copies for each chromosome, each one inherited from a different parent.
Population: not an entire species, but a group of individuals of same species living in a geographically restricted area so that any member can potentially mate with any other member

Types of variants


TTCAGAGATGAGAATCCTCAGAACTAAGAC-----TGAGGAGTG-----------GGGAAACTCTGAGGATATGTCATAGAAAAG
TTCAGATATGAGAATCC----AACTAAGACACTGGTGAGGAGTGAGTGAGTGAGTGGGGAAACTCTTCTATGACATATCCTCAAAG
Single nucleotide polymorphism (SNP) 2 alleles
Deletion 2 alleles
Insertion 2 alleles
Copy number variation Multiple alleles
Inversion 2 alleles
Indels
Structural variants (>50 bp)

(CNV) Microsatellites: repeated sequences of 2-5 bp polymorphic in their length (multiallelic variants). Also known as STRs (short tandem repeats)

Do all variants have an effect on phenotype? Only polymorphisms in functional elements (so not only coding but any region that has a function) may have phenotypic effects.

Polymorphisms inside coding regions will not always have consequences	Polymorphisms outside coding regions can have consequences
 Non-synonymous Synonymous Seq 1 CTG GAC AGG CGA GGA ATA CAG L D R R G I Q Seq 2 CTG GAC AGG CAA GGT ATA CAG L D R Q G I Q	

Allele and genotype frequencies

Allele number: # of different alleles in a particular gene.
Genotype frequency: proportion of a given genotype among all individuals in a group.
Allele frequency: proportion of a given allele among all the alleles in a group of individuals.

Allele and genotype numbers in diploid organisms

# of alleles will determine the # of genotypes in a population.

$$Allele\ number\ =\ k$$

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

Exception: x-linked variants.

Allele frequencies in diploid organisms

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

$$Total\ n\ alleles\ =\ total\ n\ of\ individuals(N)\ x\ 2\ =\ 2N$$

Genotype frequencies

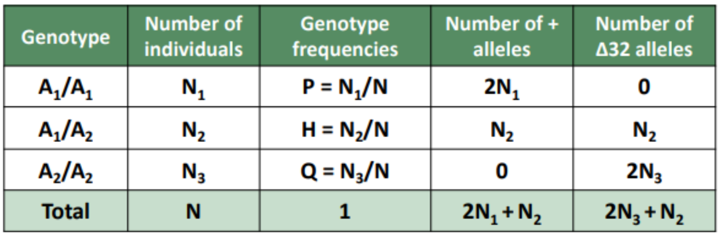
$$Freq(AA)\ =\ P\ =\ \frac{n\ of\ AA\ alleles}{total\ n\ of\ individuals}$$

$$Freq(Aa)\ =\ H\ =\ \frac{n\ of\ Aa\ alleles}{total\ n\ of\ individuals}$$

$$Freq(aa)\ =\ Q\ =\ \frac{n\ of\ aa\ alleles}{total\ n\ of\ individuals}$$

⚠ $p\ +\ q\ =\ 1$ ⚠ $P\ +\ H\ +\ Q\ =\ 1$

Allele and genotype frequencies – Counting method (slide 21)



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

$Total\ number\ of\ alleles\ =\ 2N$

Allele frequencies:

$p\ =\ \frac{2N1+N2}{2N}\ =\ P\ +\ \frac{1}{2}H$ $q\ =\ \frac{2N3+N2}{2N}\ =\ Q\ +\ \frac{1}{2}H$

Evolution: change of allele frequency over time.

Hardy-Weinberg equilibrium

Assumptions: diploid organism, sexual reproduction, non-overlapping generations, random mating, equal allele frequencies in both sexes, large population size, no migration, no mutation, no selection.

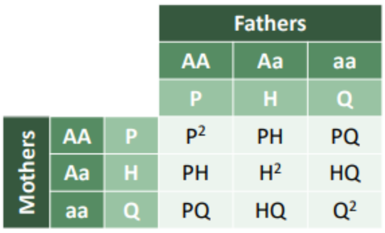
HW = null model, prediction based on a simplified or idealized. Initial situation, nothing is happening.

Hardy-Weinberg principles

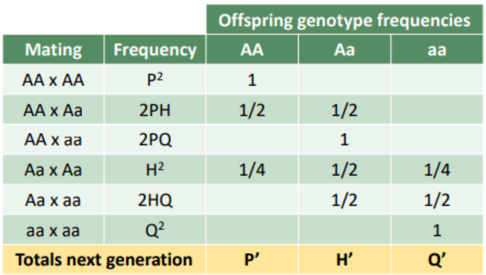
1. Genotype frequencies in a population with random mating are determined by allele frequencies.

P (AA) = p^2 H (Aa) = 2pq Q (aa) = q^2 → $p^2 + 2pq + q^2 = 1$

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



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

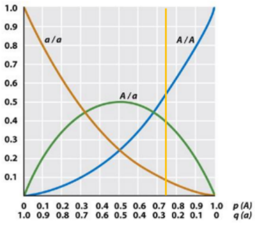


		Offspring genotype frequencies		
Mating	Frequency	AA	Aa	aa
AA x AA	P^2	1		
AA x Aa	2PH	1/2	1/2	
AA x aa	2PQ		1	
Aa x Aa	H^2	1/4	1/2	1/4
Aa x aa	2HQ		1/2	1/2
aa x aa	Q^2			1
Totals next generation		P’	H’	Q’

P’ = p^2 H’ = 2pq Q’ = q^2

Hardy-Weinberg equilibrium expected genotype frequencies

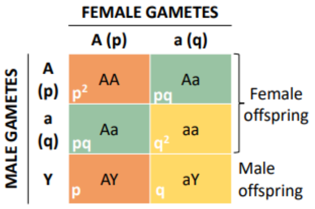
When we have recessive allele the proportion of the allele may be much higher than the observed in the phenotype



P alleles = 0.5 → higher number of heterozygous individuals.

HW equilibrium in X-linked genes

Genotype frequencies:

Females	Males (only 1 X chr)	

AA = p^2
Aa = $2pq$
aa = q^2
	A = p
	a = q	

FEMALE GAMETES

MALE GAMETES

	A (p)	a (q)	
A (p)	AA p^2	Aa pq	Female offspring
a (q)	Aa pq	aa q^2	
Y	AY p	aY q	Male offspring

Males → a → q $q > q^2$
Females → aa → q^2

Generations needed to reach HW equilibrium

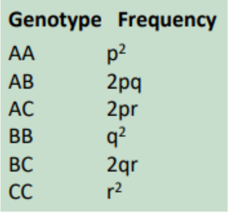
- If allele frequencies are identical in males and females

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

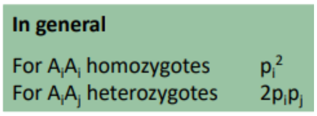
- If allele frequencies are NOT identical in males and females

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

HW equilibrium with multiple alleles



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



In general

For A_iA_i homozygotes	p_i^2
For A_iA_j heterozygotes	$2p_ip_j$

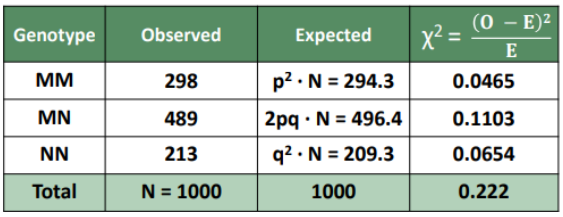
Applications of Hardy-Weinberg equilibrium

- Null model to analyze the effect of different factors on the genetic composition of a population

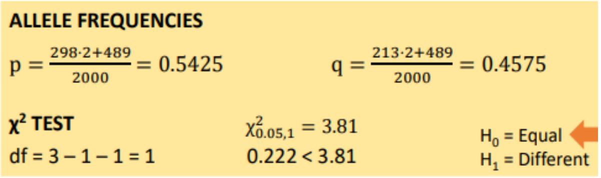
HWE is the null model, when nothing is happening. First thing to do when a population is given to us is to check if it follows the HW equilibrium.

- Test if genotype frequencies adjust to expected values > If they do not, one of the assumptions is not true

Adjustment of genotype frequencies to Hardy-Weinberg



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



ALLELE FREQUENCIES

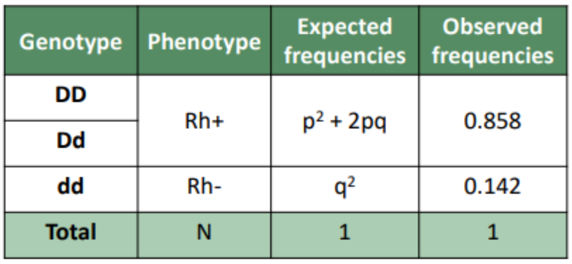
$p = \frac{298 \cdot 2 + 489}{2000} = 0.5425$ $q = \frac{213 \cdot 2 + 489}{2000} = 0.4575$

χ^2 TEST
df = 3 – 1 – 1 = 1

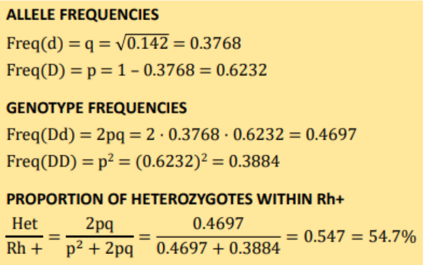
$\chi^2_{0.05,1} = 3.81$
0.222 < 3.81

H_0 = Equal
H_1 = Different

- Estimation of allele frequencies in case of dominance



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



ALLELE FREQUENCIES

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

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

GENOTYPE FREQUENCIES

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

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

PROPORTION OF HETEROZYGOTES WITHIN Rh+

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

- Test alternative models of inheritance

- Forensic DNA profiling

In forensic DNA (police tests) → to identify people we need to use variants STRs
We identify people using unique variants that are unique for each person. These variants are microsatellites.

Compute the frequency of each genotype, we get the probability to find that combination. Finally we multiply all the results obtained.
Amelogenin is not a microsatellite, is a sex test → (X, Y)

GENETIC DRIFT

Hardy-Weinberg equilibrium

Assumptions: diploid organism, sexual reproduction, non-overlapping generations, random mating, equal allele frequencies in both sexes, large population size, no migration, no mutation, no selection.

Changing this and observing → what happens if our population is small?
Imagine a population as a group of alleles

Populations of finite size

Pool of gametes → select 10 alleles to be the next allele in the generation.
We randomly select this so it could be as the og or could be different. Allele frequency now has changed in this generation (from 5/5 white black to 6 whites and 4 black). Repeat this step again (8 white and 2 black). Fluctuating by chance by process of random sampling

In a small pop just because it's small it will have random sampling → genetic drift.

No drift → Allele frequencies are always the same in all the generations
Finite population size → Allele frequencies fluctuate up and down.

Genetic drift is a stochastic process → We can only predict the probability of each possible outcome in the next generation.

Binomial distribution

- 2 possible outcomes of a trial.
- Probability of each outcome remains the same across all trials.
- All trials are independent of each other.

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

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

k = number of individuals * 2 (number of alleles) * allele frequency

Allele frequencies over time in populations of different sizes

Random changes from one generation to the next

Process of fluctuation of the alleles will end when it either reaches 1 (fixed = everyone has it) or 0 (lost = no one has it), because then the population stays that way forever

Wright-Fisher model for genetic drift

Assumptions: infinite populations, constant size N, random mating, isolated populations (no migration), no mutation, all individuals contribute equally to the infinite pool of gametes, each generation is formed by a random sample of 2N gametes from the previous generation

To get results we experiment many times on many populations so that in the end we can use the proportions on all populations → count how many times each outcome happens

Markov chains (slide 16)

Calculate each population based on the previous generation

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

Genetic drift will cause allele fixation

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.

Variance increases with time.

At the end half of the population will have one allele and the other half the other.

Fixation index: (F_{ST}) measures how far a group of populations is into the genetic drift process (which ends in the equilibrium with all populations with a fixed allele).

$F_{ST} = \frac{Var(t)}{Var(\infty)} = 1 - \left(1 - \frac{1}{2N}\right)^t$

Probability of fixation of a neutral allele

Probability of fixation of a neutral allele is equal to its initial frequency in the population.

$P_{fix} = p_0$

A new allele with p_0 = 1/2N is more likely to be lost than fixed.

New allele = p = 1/2N

Average time to fixation = $T_{fix} \approx 4N$

Average time to loss = $T_{loss} \approx 2 \cdot \ln(2N)$

Genetic drift causes a reduction in heterozygosity

Smaller the population, stronger the effect of drift, and faster that heterozygosity will be lost.

Heterozygosity = observed proportion of heterozygotes in a population

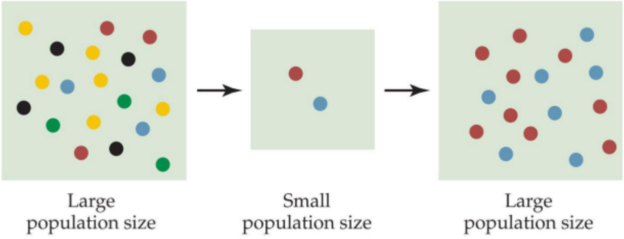
$H_t = H_0\left(1 - \frac{1}{2N}\right)^t$ $\frac{H_t}{H_0} = \left(1 - \frac{1}{2N}\right)^t$

Heterozygosity declines by a factor of $1 - \frac{1}{2N}$ every generation due to drift

If a population evolves only with genetic drift it will 100% lose an allele eventually so we are losing variation → reducing proportions of heterozygotes

Reductions in population size

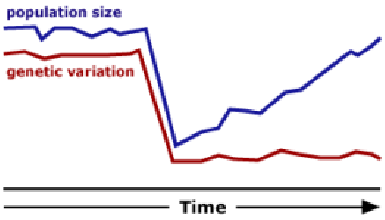
Genetic drift acts more quickly to reduce genetic variation in small populations



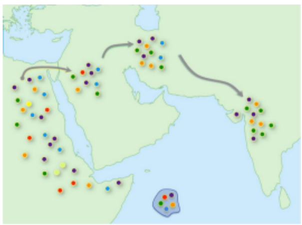
The resulting population can have:

1. Reduced variation and reduced ability to adapt to new selection pressures.
2. A non-random sample of the genes in the original population.

Bottleneck



Founder effect



Bottleneck: occurs when a population's size is reduced for at least one generation. Consequences: long-term reduction of genetic variation (even if the bottleneck does not last for many generations and the population regains its previous size)

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

1. Substantial loss in genetic diversity.
2. Fast divergence between source and founder populations.

Consequences of genetic drift on a small population

Reduced genetic diversity levels due to:

1. Limited input (only 3 founders) and absence of gene flow from neighboring populations until 2008.
2. Severe inbreeding and drift reduced further the limited diversity.

Effective population size (N_e)

Ideal population ($N_c = N_e$)

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

Most deviations will decrease the effective population size

H_0 = is 1 (the 100%)

Census population (N_c): total number of individuals in a population
Effective population size (N_e): individuals that actively participate 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.

Factors that can contribute to this difference:

1. Different numbers of males and females.

N_e in a population unequal sex ratio for autosomal genes.

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

2. Fluctuations in population size.

Populations can show regular cycles of increase and decrease spanning a number of years.
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.

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

3. Variation in the number of offspring among individuals
4. Bottlenecks
5. Overlapping generations

Genetic drift - Important ideas (slide 35)

- 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.

MUTATIONS

Hardy-Weinberg equilibrium
Assumptions: diploid organism, sexual reproduction, non-overlapping generations, random mating, equal allele frequencies in both sexes, large population size, no migration, no mutation, no selection.

Mutation as an evolutionary force that changes allele frequencies
Mutation is the source of all genetic variation, introducing new alleles in populations.

Mutation: any permanent change in an organism’s DNA (from nucleotide substitutions to large structural variants) and is the result of unrepaired damage in DNA and errors during DNA replication or repair.
Mutations in the germinal line are transmitted to offspring but somatic mutations are not.

Phenotypic level → mutation can be considered recurrent.
Molecular level → most mutations are unique.

Irreversible mutation

Rate: An allele gets mutated into another allele. It would happen from time to time.
A → a (μ)

$$p_t = p_0(1 - \mu)^t$$

As time passes, the frequency changes, but very slowly.
In the end, all A will become a.

Reversible mutation

(v) A ⇄ a (μ)   μ > v

Assume that A is the normal allele and a the one that causes the mutation. The mutation rate is always to be higher going from A to a

In equilibrium: $\hat{p} = \frac{\nu}{\mu+\nu}$

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

When a long time is considered, it doesn't matter where we started, we'll end up in equilibrium → both alleles in the population.

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 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.**

NATURAL SELECTION

Hardy-Weinberg equilibrium

Assumptions: diploid organism, sexual reproduction, non-overlapping generations, random mating, equal allele frequencies in both sexes, large population size, no migration, no mutation, no selection.

Natural selection

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.
*Mutation almost no impact bc it’s too slow

- Requires existing heritable variation in a group, important and necessary

- Requires a causal relationship between genotype and number of offspring
- Depends on the environment
- Results in a greater adaptation of organisms to their environment over time

Natural selection

Only the variants that have phenotypic consequences will evolve through natural selection. The environment (with different survival rates) is necessary for this process → Population will adapt to the environment.

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

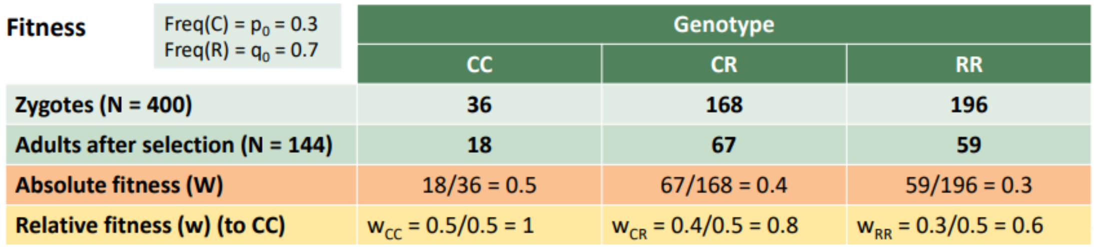
Gamete → Zygote → Adult → Parents → …
The selection occurs in the development step, from zygote to adult.

Basic selection model - Fitness

Fitness: overall ability of an organism to survive and reproduce. Contribution (in number of offspring) of an individual to the next generation. Depends on the environment

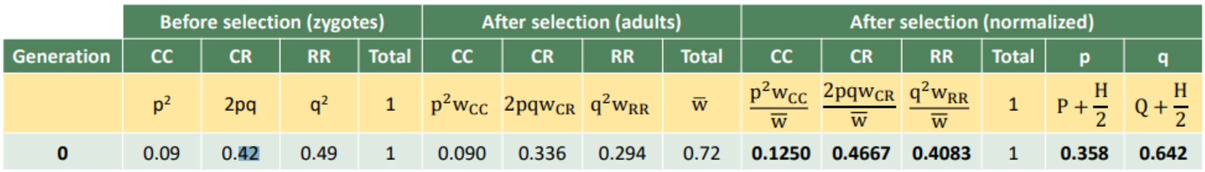
Absolute fitness (W): measurement of the ability to survive of each genotype. Ideal W = 1.
Relative fitness (w): ability of one genotype to survive relative to another genotype taken as reference
Average fitness of the population ($\overline{w}$)

Calculations based on genotypes
The fitness of a genotype is the average fitness of all the individuals with that genotype



Fitness Freq(C) = p_0 = 0.3 Freq(R) = q_0 = 0.7	Genotype		
	CC	CR	RR
Zygotes (N = 400)	**36**	**168**	**196**
Adults after selection (N = 144)	**18**	**67**	**59**
Absolute fitness (W)	18/36 = 0.5	67/168 = 0.4	59/196 = 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

Basic selection model - General model



	Before selection (zygotes)				After selection (adults)				After selection (normalized)					
Generation	CC	CR	RR	Total	CC	CR	RR	Total	CC	CR	RR	Total	p	q
	p^2	2pq	q^2	1	p^2w_{CC}	$2pqw_{CR}$	q^2w_{RR}	$\overline{w}$	$\frac{p^2w_{CC}}{\overline{w}}$	$\frac{2pqw_{CR}}{\overline{w}}$	$\frac{q^2w_{RR}}{\overline{w}}$	1	$P+\frac{H}{2}$	$Q+\frac{H}{2}$
0	0.09	0.42	0.49	1	0.090	0.336	0.294	0.72	**0.1250**	**0.4667**	**0.4083**	1	**0.358**	**0.642**

Relative fitness values, and not their magnitudes (absolute values), determine the change in allele and genotype frequencies.
The average fitness of the population always increases by action of natural selection.

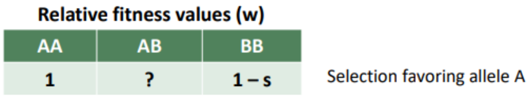
Types of selection

- Directional selection: selection will cause the fixation of advantageous alleles and the loss of deleterious alleles.

- Purifying selection: new allele has negative effects and is selectively removed from the population.
- Positive selection: 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

Selection coefficient

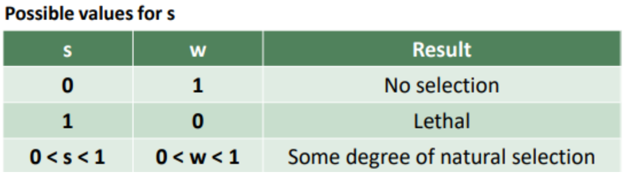
Selection coefficient (s): reduction in fitness of a given genotype when compared to another ($0 \leq s \leq 1$)



Relative fitness values (w)

AA	AB	BB
1	?	1 – s

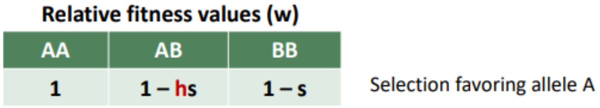
Selection favoring allele A



Possible values for s

s	w	Result
0	1	No selection
1	0	Lethal
0 < s < 1	0 < w < 1	Some degree of natural selection

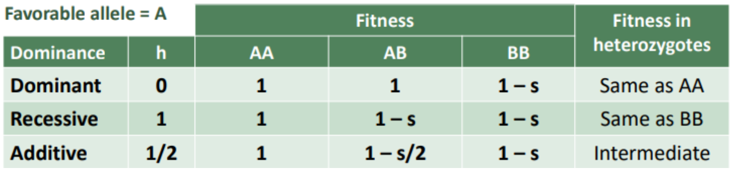
Degree of dominance (h): parameter that modulates the selection coefficient in heterozygotes depending on the dominance relationship of the alleles.



Relative fitness values (w)

AA	AB	BB
1	1 – hs	1 – s

Selection favoring allele A



Favorable allele = A

		Fitness			Fitness in heterozygotes
Dominance	h	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.

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

Overdominance or heterozygote superiority

Heterozygous genotype has a greater fitness that either homozygote

$w_{AA} = 1 - s_1$

$w_{Aa} = 1$

$w_{aa} = 1 - s_2$

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

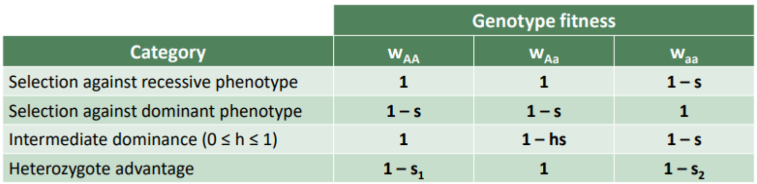
$\Delta p = 0$

$\hat{p} = \frac{s_2}{s_1+s_2}$

Stable equilibrium: allele frequencies converge to an equilibrium value irrespective of initial frequencies.

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

General categories of relative fitness values



	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 ≤ h ≤ 1)	1	1 – hs	1 – s
Heterozygote advantage	$1 - s_1$	1	$1 - s_2$

Mutation facts

- Mutation is the source of all genetic variation
- A mutation is any permanent change in an organism's DNA (from nucleotide substitutions to large structural variants)
- Mutation is the result of unrepaired damage in DNA and errors during DNA replication or repair
- Mutation rates are generally low

Mutation-selection balance
Mutation will generate non-functional alleles with negative effects on fitness
(funct. allele, high freq.) A → a (μ) (non-funct. recessive allele, very low freq.)

- Mutation: ↑ a, ↓ A
- Purifying selection: ↓ a, ↑ A

At equilibrium: Δp = 0

- Selection against a recessive allele. h = 0. $\hat{q} = \sqrt{\frac{\mu}{s}}$
- Selection against a partially dominant allele. h > 0. $\hat{q} = \frac{\mu}{sh}$

A = favored allele p(A) ≈ 1
a = harmful recessive allele q(a) ≈ 00

MIGRATION

Hardy-Weinberg equilibrium
Assumptions: diploid organism, sexual reproduction, non-overlapping generations, random mating, equal allele frequencies in both sexes, large population size**,** no migration, no mutation, no selection

Population differentiation
2 different populations of the same species can have different allele frequencies and even different alleles.

- Genetic drift: allele frequencies change randomly in each population
- Selection: different alleles may be favored in different environments

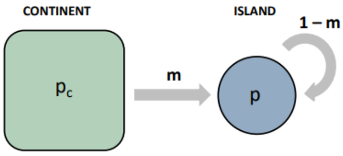
- Mutation: different alleles will be generated by mutation in isolated populations

Migration facts

Migration: movement of individuals among populations.
Causes gene flow or transfer of genetic material from one to the other.
Limits the genetic divergence that can occur among subpopulations.
⚠️ migration from the receiving population pov

Continent-island model

Large population → allele frequencies don't change.
Small population→ a proportion of the individuals are replaced by migrants from the large continent population each generation



m = percentage of alleles that come from another population.

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

$p_1 = p_0(1 - m) + p_c m$

p0= island p

Substituting the alleles on island by the ones on the continent → island will have the same allele frequency as in the continent.

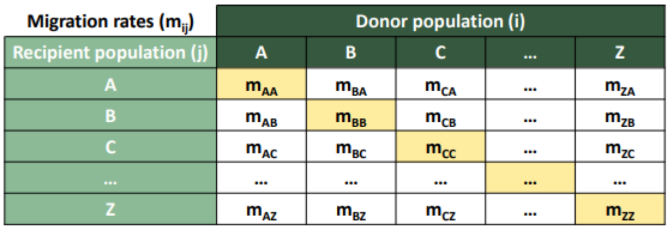
The frequency can change in any direction:

$\Delta p = p_1 - p_0 =- m(p_0 - p_c)$

- $p_0 > p_c$ → $p_0 - p_c > 0$ → allele frequency in the island will decrease
- $p_0 < p_c$ → $p_0 - p_c < 0$ → allele frequency in the island will increase

It doesn't matter where we start on the island, we'll end up at the same frequency as the continent (it will take different time depending on the migration rate).
Equilibrium when having the same allele frequency in the island than in the continent.

General model



Migration rates (m_{ij})	Donor population (i)				
Recipient population (j)	A	B	C	...	Z
A	m_{AA}	m_{BA}	m_{CA}	...	m_{ZA}
B	m_{AB}	m_{BB}	m_{CB}	...	m_{ZB}
C	m_{AC}	m_{BC}	m_{CC}	...	m_{ZC}
...	...	...	...		...
Z	m_{AZ}	m_{BZ}	m_{CZ}	...	m_{ZZ}

$p_1 = m_{AA}p_{0A} + m_{BA}p_{0B} + ... + m_{ZZ}p_{0Z} = \Sigma m_{ij}p_{0i}$

$p_2 = m_{AA}p_{1A} + m_{BA}p_{1B} + ... + m_{ZZ}p_{1Z} = \Sigma m_{ij}p_{1i}$

Wahlund effect

2 different populations → calculate p and q
Calculate the genotype frequencies for each population

- Calculate the average between the genotypes (observed)
- Calculate the HWE genotype frequencies (expected)

Lack of heterozygous genotypes because we are probably counting 2 pop that are different as if they were the same.

If our population is made up of different populations, we will always find fewer heterozygotes than we expect, only if there is no migration (as migration homogenizes everything, ending with all the populations having the same allele frequency).

Estimating population subdivision: Fixation index (F_{ST})

Division in subpopulations will cause a decrease in the proportion of heterozygotes.

Heterozygosity measurements:

- H_s = average expected heterozygosity assuming random mating within each population.

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

- H_T = expected heterozygosity in the total population.

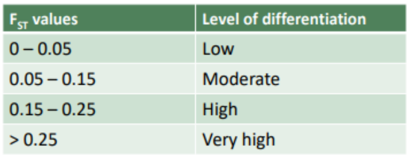
$$H_T = 2\,\bar{p}\,\bar{q}$$

S = subpopulations
T = total

F_{ST} is a measure of how much allele frequencies have diverged among subpopulations.

Is the difference between the expected heterozygosity of the total population and the average expected heterozygosity of subpopulations.

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



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

F_{ST} in different species

Some species show a very great differentiation level among populations

Most of human diversity is found within groups and not between groups

F_{ST} in human populations

Average level of population differentiation is low

There are several hundred thousand SNPs with large allele frequency differences in each population comparison.

The most highly differentiated sites are enriched for non-synonymous variants, indicative of the action of local adaptation

F_{ST} will increase by genetic drift in finite populations

Without migration:

- Allele frequencies change among subpopulations
- F_{ST} increases in each subpopulation
- Differentiation continues until one allele is fixed

Increase of F_{ST} 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:

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

Migration-genetic drift balance

Genetic drift (differentiates)

Migration (homogenizes)

Migration limits strongly the differentiation between populations

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

MOLECULAR POPULATION GENETICS

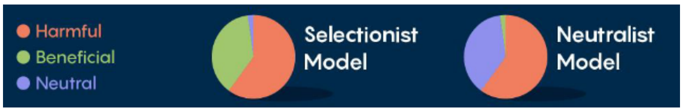
NEUTRAL THEORY OF MOLECULAR EVOLUTION

Models of molecular evolution

Selection affects the frequency of harmful or beneficial mutations, but mutations with neutral consequences could survive and fix purely by chance.

- Harmful selection → negative selection
- Beneficial mutation → positive selection
- Neutral mutation → random sampling (no effect)

Selectionist vs. neutralist models



Selectionist model (evolution by natural selection). Neutral mutations are only a small part.
Neutralist Model (change of allele frequencies by genetic drift). Changes the proportion of neutral variants (in respect of the selectionist model)

Neutral theory of molecular evolution

Genetic drift is the primary evolutionary process that dictates the fate of new mutations that have no effect on fitness. 2 predictions:

Polymorphism	Divergence
The amount of polymorphism for sequences sampled within a population of one species	Degree and rate of divergence among sequences sampled from separate species

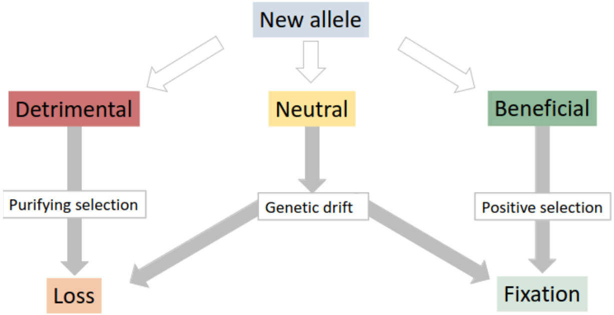
Neutral theory represents the null model in molecular evolution

Mutation-drift balance - infinite alleles model

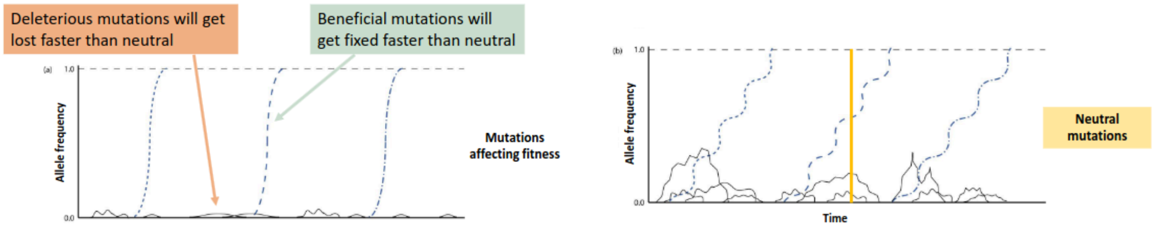
Assumes that every time a mutation occurs it creates a new allele.
Mutation (creates alleles) ↔ Genetic drift (eliminates alleles)
Equilibrium (neutral theory)

Mutation effects

Most non-neutral mutations are deleterious.



Neutral theory of molecular evolution

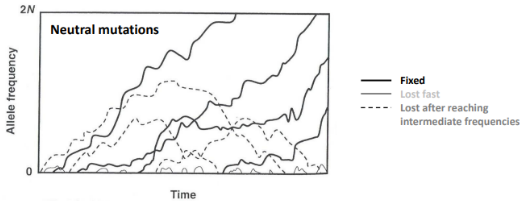


It will take a longer time for neutral mutations to get fixed or lost
In the meantime, we can see them as polymorphic variants in the population
When an allele gets fixed (p = 1) it becomes divergence with other populations or species

Most of the polymorphic variants are going to be neutral, since beneficial/deleterious alleles eliminate/fix rapidly (poques vegades podrem veure els 2 al·lels junts).

When a variant reaches 1, it stops being polymorphism and becomes divergent (when we have a difference between populations or species), becoming a position that differentiates the species.

Most genetic variation is maintained in populations simply due to the random allele-frequency walk that new mutations take before reaching either fixation or loss.



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

Polym. is the neutral variant in the process of becoming lost or fixed.

Polymorphism under neutral theory
Very few new alleles will fix, but those segregate for a much longer time than the mutations that end in loss and will be considered polymorphisms

High levels of polymorphism will result from:

- High mutation rate: lots of variants being generated
- Large population size (low genetic drift)
- Intermediate levels of mutation and genetic drift

Divergence under neutral theory
Mutation rate (μ): rate at which changes are incorporated in a nucleotide sequence during DNA replication and reparation processes: rate we are fixing this new allele.
Substitution rate (K): rate at which new mutations are fixed in the population

$$\frac{substitutions}{generations} = \frac{mutations}{generation} \cdot P(fixation) \qquad K = 2N\mu \cdot \frac{1}{2N} = \mu$$

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

The more divergence time between species, the more differences accumulate by the fixation of neutral alleles by genetic drift at a constant rate.
If a sequence does not change over time, it is not neutral but essential.

MEASURING DNA POLYMORPHISM

Measurements of genetic variation at DNA sequence level

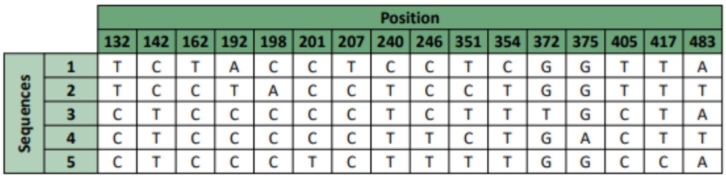
Proportion of segregating sites	Watterson's θ

Number of segregating sites per nucleotide
$p_s = \frac{S}{L}$
L = length
S = segregating sites
	Measurement of the proportion of segregating sites corrected by sample size
	$\theta = \frac{p_s}{a}$
	$a = 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n-1}$

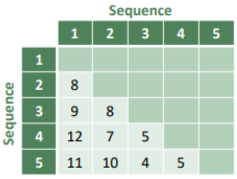
n = number of sequences in the sample

Nucleotide diversity (π): average number of nucleotide differences per site between two random DNA sequences in the population.
δ = Average number of differences between 2 sequences (# different/# comparisons)
$\Pi = \frac{\delta}{L}$

Nucleotide diversity (π)



		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



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

First table: variable positions.
Second table: # differences on every possible pair of sequences.

Nucleotide diversity in different positions of the genome

Nucleotide diversity varies among different organisms, different genes, different types of functional positions or along chromosomes
π is much lower in exons than in introns because in exons many positions are not neutral so if we change it, then it will be affected by selection.

Nucleotide diversity (π) in different species

There is high variation within each species group.
Low variation on species will make it difficult for them to adapt when the environment may change. That's why endangered species have low variation.

Genetic diversity in metazoans

Diversity of species is predictable, and is determined in the first place by its ecological strategy.
Population size is probably an important factor in determining the level of nucleotide diversity
Large populations are less likely to lose variation

Synonymous and nonsynonymous positions within coding regions

- Synonymous: changing a nucleotide doesn't change the aa.
- Non-Synonymous: changing a nucleotide changes the aa.
- "Half synonymous": changing a nucleotide changes the aa depending on the nt we change to.

We expect to see more variation on the synonymous bc they are neutral positions, as changing the nucleotide does not “matter”. Important positions or changes will be non-synonymous.

Jukes and Cantor correction

Divergence (D) is a minimum estimate because there may be positions that were initially different and are equal now due to a substitution

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

K = # substitutions/position

D = divergence, proportion of different nucleotides.

K_a/K_s ratio test

Tell us how the sequence is evolving.

K_a = number of nonsynonymous substitutions per nonsynonymous position

K_s = number of synonymous substitutions per synonymous position

- $K_a/K_s < 1$: purifying selection → more S than NS changes, it happens in most genes
- $K_a/K_s = 1$: neutral evolution → same S than NS changes, these are pseudogenes
- $K_a/K_s > 1$: positive selection → more NS than S, it happens exceptionally

LINKAGE DISEQUILIBRIUM

Two different variant positions in the same chromosome

Although having two different variants on each mutation, we can have four different chromosome sequences.

Haplotypes

Combination of alleles in the same chromosome. Number of possible combinations for n variants: 2^n

Variants in the same chromosome can be linked

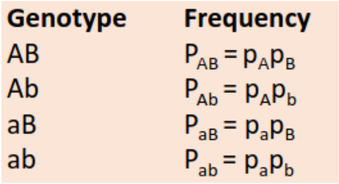
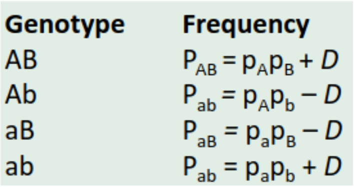
Expected freq → multiply paired frequencies

Comparing the observed with the expected in this example we can see that the values are very similar, in the purple example.

When we expect more or less frequencies, there’s linkage equilibrium or disequilibrium.

D value will be the same for all of them.

Linkage equilibrium/disequilibrium

<table>
<tr><th>Linkage equilibrium</th><th>Linkage disequilibrium</th></tr>
<tr><td>Random association. D = 0

<table>
<tr><th>Genotype</th><th>Frequency</th></tr>
<tr><td>AB</td><td>$P_{AB} = p_A p_B$</td></tr>
<tr><td>Ab</td><td>$P_{Ab} = p_A p_b$</td></tr>
<tr><td>aB</td><td>$P_{aB} = p_a p_B$</td></tr>
<tr><td>ab</td><td>$P_{ab} = p_a p_b$</td></tr>
</table></td>
<td>Correlation between 2 loci. D ≠ 0

<table>
<tr><th>Genotype</th><th>Frequency</th></tr>
<tr><td>AB</td><td>$P_{AB} = p_A p_B + D$</td></tr>
<tr><td>Ab</td><td>$P_{ab} = p_A p_b - D$</td></tr>
<tr><td>aB</td><td>$P_{aB} = p_a p_B - D$</td></tr>
<tr><td>ab</td><td>$P_{ab} = p_a p_b + D$</td></tr>
</table></td></tr>
</table>

⚠️ if there are more AB and ab, there must be less Ab and aB

Linkage disequilibrium measurement (D)

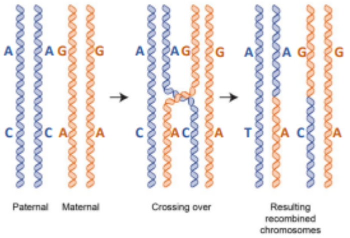
We use parameter D to measure linkage disequilibrium.

D is the difference between the observed frequency of a haplotype and the expected frequency of this haplotype if these alleles were independent.

D value is not comparable as it depends on the allele frequencies.

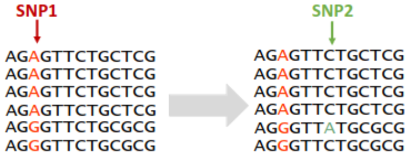
$$D = P_{AB} - p_A p_B$$

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



Haplotypes and recombination

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



When having a new variant we only have 3 different chr. seq, the 4th one will appear by recombination. This will make D decrease, make the expected close to the observed.

Recombination in heterozygotes breaks the linkage between variants.

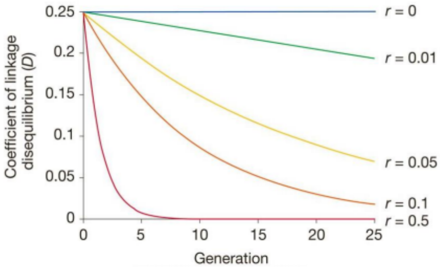
Recombination destroys D, as it generates new combinations.

Linkage disequilibrium decay

Over time, LD between variants will decrease

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

The higher the recombination rate (r), the faster the linkage between variants will be lost and the less linkage disequilibrium we will find.



Linkage disequilibrium and distance

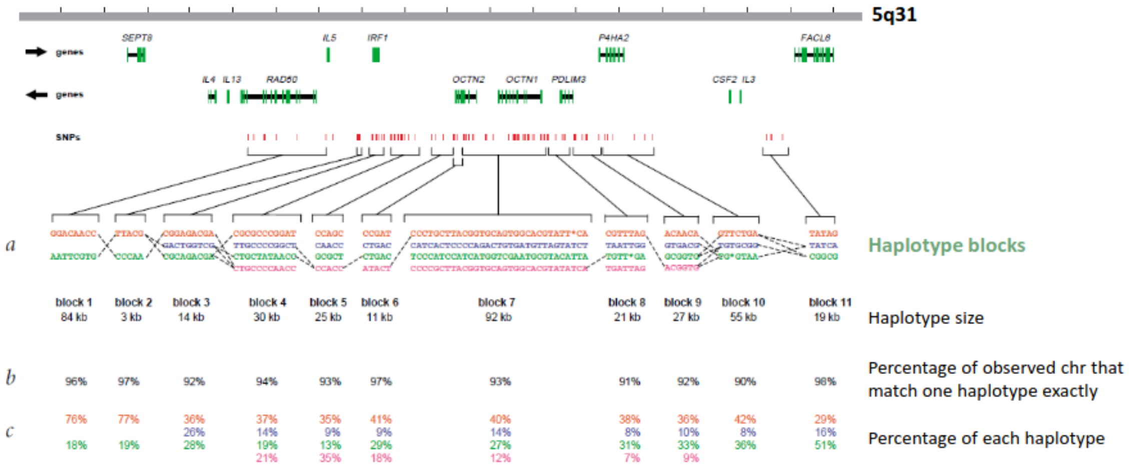
LD decreases as the distance between variants increases.

Longer distance implies a higher probability that recombination occurs between the two variants.

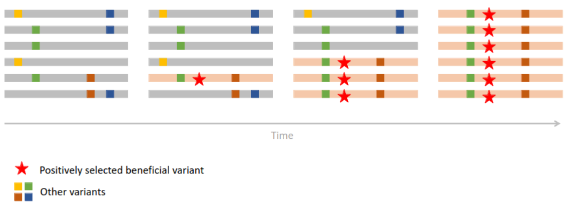
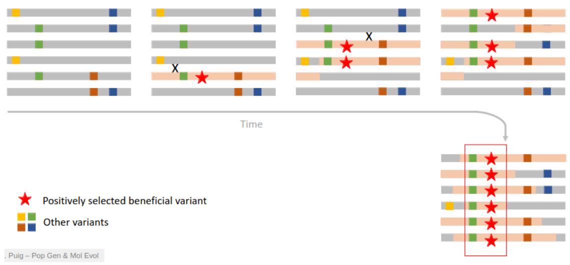
Haplotype blocks in the human genome

Analysis of haplotype structure of 500 kb using 103 common SNP (minor allele frequency > 5%)

Haplotype blocks are sequences of variants.



Selective sweep

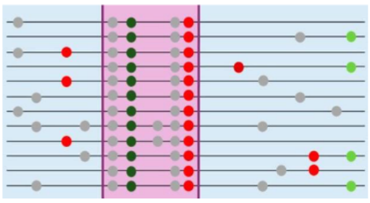
No recombination	With recombination
 Time Positively selected beneficial variant Other variants	 Time Positively selected beneficial variant Other variants Puig – Pop Gen & Mol Evol
All chr will be positively selected, so it will increase in frequency. We will fix all the chr at the end.	Something similar will happen but less dramatic Not the whole chr will be fixed. At the end we will have the selected variant on all chr but just the variants close to it will be also fixed together, not all the chr.

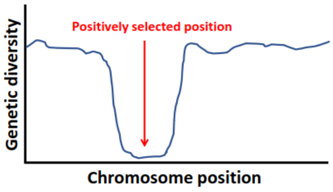
Detection of a selective sweep

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 (they are in linkage disequilibrium with the selected variant)

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.





Selection destroys variation, so the genetic diversity/variation will drop on the positively selected position.

This region includes variants, linked ones. But it is impossible to know which is the good one, the positively selected variant.

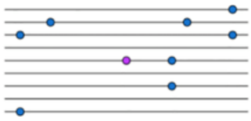
BLOCK 2: MOLECULAR EVOLUTION

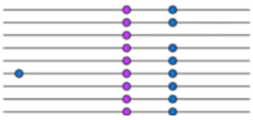
MOLECULAR ADAPTATION AND NEUTRALITY TESTS

The Neutrality Tests
Use neutral theory as a null model.
Important as it may provide evidence for adaptation at the molecular level, and help to elucidate genotype-phenotype relationships.
Use neutrality test to infer natural selection from sequence data

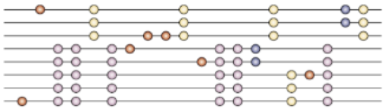
2 types of tests:

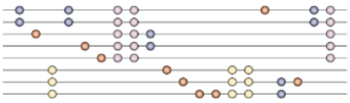
- Levels of variation: how many mutations do you have in a population level in a specific gene





- Patterns of variation: how the mutations are distributed, their frequency

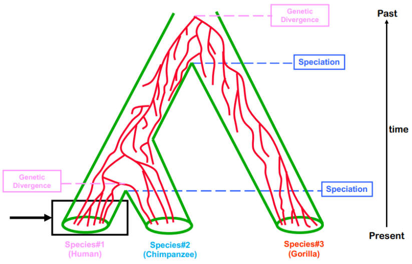




Combining both we can link the genotype with the phenotype

Tajima's D
Neutrality test that attempts to determine if a particular set of nucleotide sequences are or are not compatible with the null hypothesis.
Test based on comparisons of divergence and/or variability between different classes of mutation.

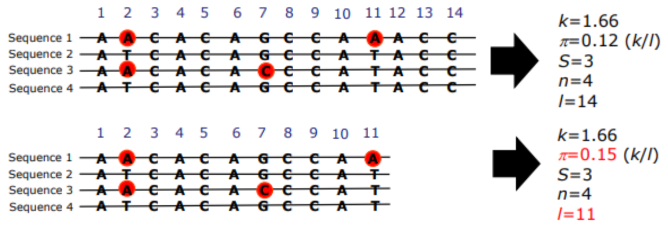
A particular gene can be differentiated before the speciation time.



Intraspecific Data: Neutrality Tests
Computed from a MSA from a gene within species

- k = average # of nucleotide differences (per region)
- π = average number of nucleotide differences (per site) → k / # positions.
- S = # segregating sites (in a DNA region)
- n = sample size (number of DNA regions)

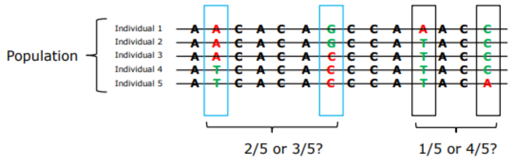
Once we have this MSA we can calculate statistics to compute variability
Segregating sites: 2, 7, 11 (variable positions)
π: important bc all the other statistics can be the same but it varies. We take into account the length.



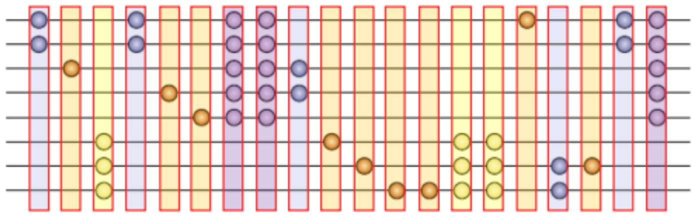
SFS, Site (or allele) Frequency Spectrum

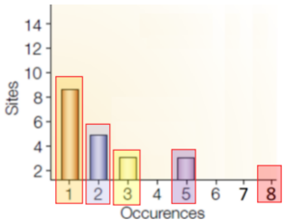
Summarizes the distribution of allele frequencies in a sample of DNA sequences. 2 SFS:

- Unfolded: using information of the ancestral variant (ancestral allele), we can estimate the frequency bc we know the ancestral state.



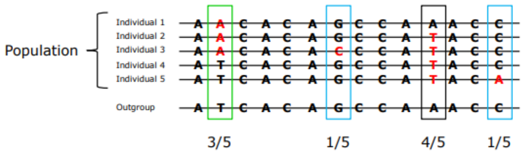
Singleton variant: only individuals carry it in a specific position.



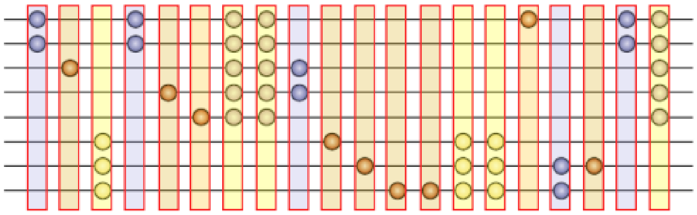


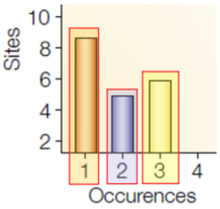
We need to know the number of derived mutations, the remaining ones will be ancestral variants.

- Folded: without information of the ancestral variant, we need to consider all together.



In this case as we don't have a reference we count states.





Heterozygosity values

θ = heterozygosity under the mutation-drift equilibrium. $\theta = 4N_e\mu$

θ_w = Watterson θ. Estimator of θ based on the number of segregation sites. $\theta_w = \frac{S}{a_n}$

$\theta_w(\pi)$ = Estimator of θ based on the nucleotide diversity. $\theta_w(\pi) = k$

Under the neutral model - mutation-drift equilibrium

$\theta = 4N_e\mu = \theta$ $\theta_w = \frac{S}{a_n} = \theta$ $\theta_w(\pi) = k = \theta$

If we have different values it means that something is happening (pop increase/decrease, bottleneck, …) or natural selection → Tajima’s D test

Tajima’s D test

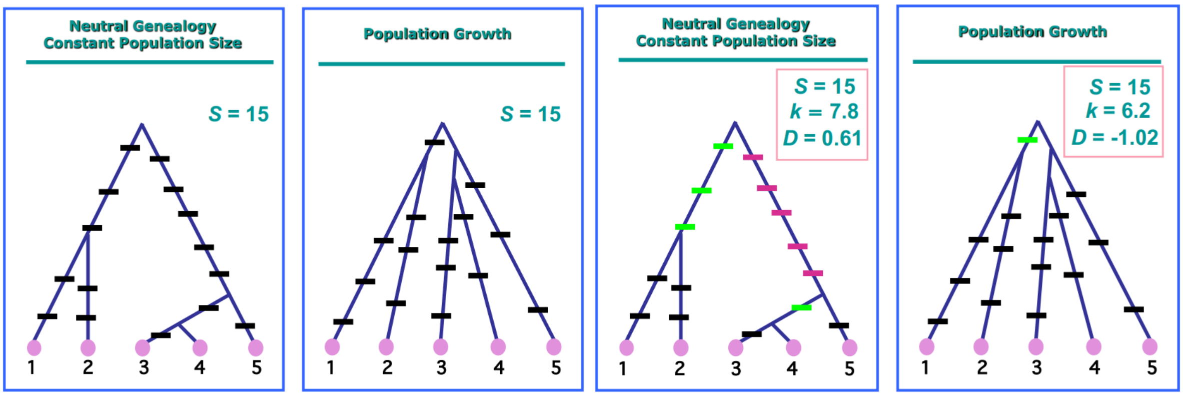
Aims to determine if a particular region is evolving neutrally (D=0). $D = \frac{k-S/a_n}{\sqrt{Var(k-S/a_n)}}$

- H_0: Neutral model (neutral evolution). $k = \frac{S}{a_n} = \theta$

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

- H_1: Non-neutral model. $k \neq \frac{S}{a_n}$

Relationship between individuals in the present. See how these individuals are related to each other.



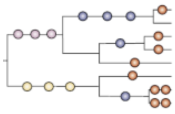
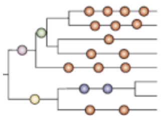
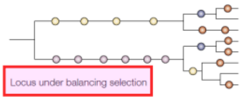
When is D significant? When it is statistically different from zero.
We need to know the distribution of D (the CI). If the value falls inside the interval, D is not significant

The distribution of D approximates the beta distribution. Depends on:

- n
- θ
- Recombination value (since the distribution is very sensitive to the recombination)

Interpreting Tajima’s D

- Tajima's D = 0: population evolving as per mutation-drift equilibrium. No evidence of selection.
- Tajima's D < 0: excess of low frequency polymorphisms relative to expectation, indicating population size expansion (f.e. after bottleneck or selective sweep) and/or purifying selection.
- Tajima's D > 0: low levels of both low and high frequency polymorphisms, indicating a decrease in population size and/or balancing selection

No selection - neutral	Positive selection	Balancing selection
		

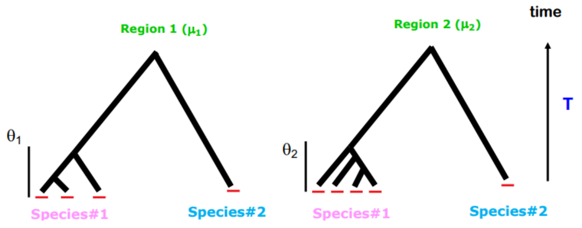
The Neutral Model

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: $K = 2T\mu$

Relationship between polymorphism and divergence must be constant for all regions (loci) across species.

The Hudson-Kreitman and Aguadé (HKA) Test

The HKA test compares the polymorphism levels within species, with those of divergence across species, from two (in this case) or more genomic regions.



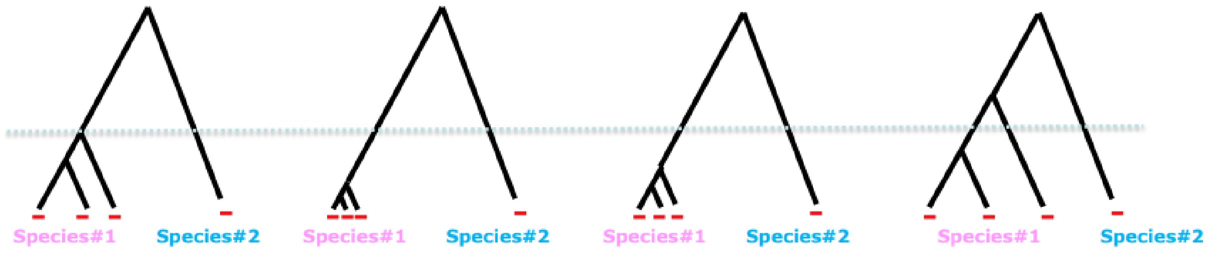
$$\theta_1 = 4N\mu_1 \qquad \theta_2 = 4N\mu_2$$

$$K_1 = 2T\mu_1 \qquad K_2 = 2T\mu_2$$

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

Compare relationships divergence-polymorphism, if the relation is constant → evolving under neutrality.

Neutral | Selective Sweep | Background selection | Balancing selection



Make a table polymorphism and divergence, then calculate the relationships and conclude.

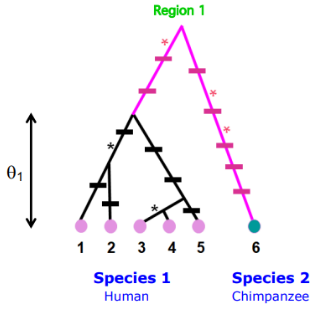
The estimated number of mutations (S) for each of the L regions, and their variance. The estimated divergence between species (D) (between species A and B), and its variance.

Goodness of fit test:

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

The McDonald & Kreitman (MK) Test

Compare 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 (NS and S).



Prediction

$$\underbrace{\frac{\text{Non-synonymous}}{\text{Synonymous}}}_{\text{Polymorphism}} = \underbrace{\frac{\text{Non-synonymous}}{\text{Synonymous}}}_{\text{Divergence}}$$

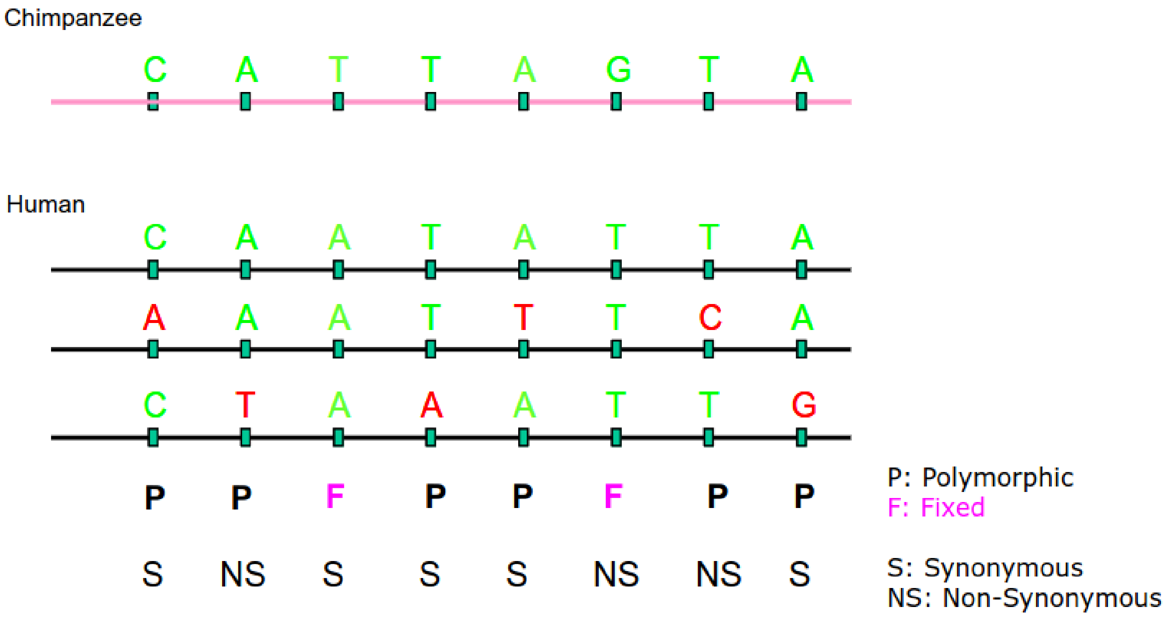
Neutrality index: $NI = \frac{P_n/P_s}{D_n/D_s}$

Check the correlation between S and NS mutations within species vs S and NS between species.
Assuming constant population size, # S > NS

- NI < 1: positive selection. Excess of NS divergence
- NI > 1: negative selection.

What happens if pos selection is operating ?
HKA is for searching positive selection intraspecific population → within a species
The test compares polymorphism (genetic variation within a population) and divergence (genetic variation between species) at multiple loci to test the neutral theory of molecular evolution, which posits that most evolutionary changes are the result of genetic drift rather than natural selection.

MK is a neutrality test at a DIVERGENCE LEVEL, comparing synonymous and non-synonymous substitutions between species.



How many polymorphic are synonymous sites and how many divergences are synonymous

Then compute contingency table, calculate expected values and compare obs vs exp
if it is similar → region is evolving under neutrality
higher ratio of non-synonymous to synonymous fixed differences compared to polymorphisms, suggesting advantageous mutations have become fixed → Positive selection
MK only one region but can be with more species looking for signal of pos selection in divergence level

HKA and MK Tests
HKA Test: Analyze two (or more) genomic regions, so these regions can exhibit different genealogies, and therefore, important variation across genes. Searches for pos selection at population level and on

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

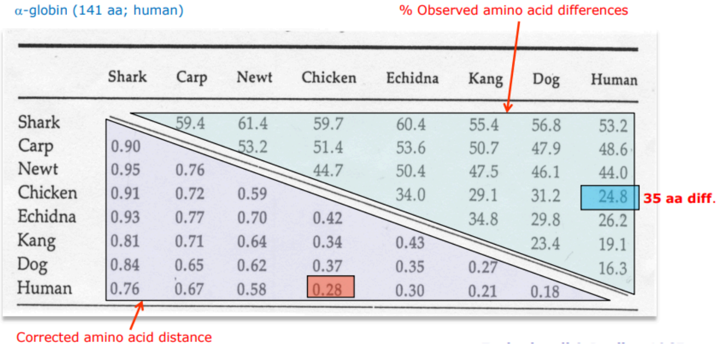
THE MOLECULAR CLOCK AND THE NEUTRAL THEORY OF MOLECULAR EVOLUTION

Population genetics: branch of genetics that deals with genetic difference within and between populations (intraspecific level). Study: adaptation, speciation and population structure.
Molecular evolution: deals with the process of molecular change. Explain patterns of variation.

Both can be interpreted using the neutral theory of molecular evolution. In addition, both are using the same empirical data DNA/protein sequences.

THE MOLECULAR CLOCK

Genetic divergence



α-globin (141 aa; human)

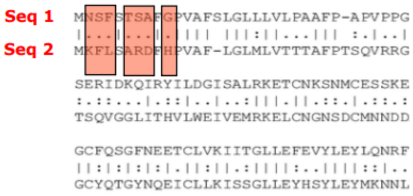
% Observed amino acid differences

	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

Homologous sequences



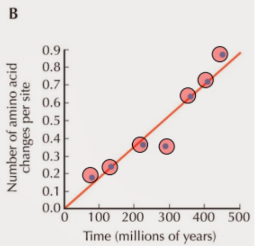
```
Seq 1  MNSFSTSAFGPVAFSLGLLLVLPAAFP-APVPPG
       |...|...|.||||  |||:||...|||  :.|..|
Seq 2  MKFLSARDFHPVAF-LGLMLVTTTAFPTSQVRRG

       SERIDKQIRYILDGISALRKETCNKSNMCESSKE
       :.::...|.::|..|..:|||.||.::.|.::.:
       TSQVGGLITHVLWEIVEMRKELCNGNSDCMNNDD

       GCFQSGFNEETCLVKIITGLLEFEVYLEYLQNRF
       ||:|:|:|:|.||:||.:|||||:..||||::|..
       GCYQTGYNQEICLLKISSGLLEYHSYLEYMKNNL
```

Length = 141aa Differences = 35aa % observed aa differences = 24.8%
24.8% can be interpreted as an aa distance

The lower part of the matrix represents the corrected aa distances. Chicken-human aa corrected distance is 0.28 (always >= observed aa differences)

The molecular clock

Linear relationship between number of amino acid changes accumulated and the time since the divergence of the species: constant rate of accumulation = molecular clock



The rate of amino acid (or nucleotide) substitution

K = # aa replacements per site
T = divergence time (how long ago the 2 seq split from the common ancestor)
$r = \frac{K}{2T}$ aa substitutions per year
Substitution rates among species vary

Divergence at different gene regions

$d_N(K_A)$ - $d_S(K_S)$
Substitution rates can also be computed for DNA sequences.

- $d_N(K_A)$ = # NS changes/# NS sites
- $d_S(K_s)$ = # S changes/# S sites

Compare S and NS substitutions to infer in the action of natural selection.

- $d_N \sim d_S$ = neutral evolution
- $d_N < d_S$ = negative selection
- $d_N > d_S$ = positive selection

It's very important in which position of the codon the mutation falls → first / middle / last

The rate of nucleotide substitution

S substitution rate is always higher than NS

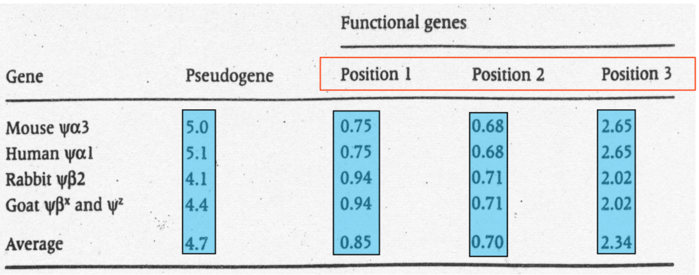
Molecular clock because we obtain the same values (it's a regression)

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

Pseudogene: gene not longer functional. They evolve under neutral evolution

Mutation in position 1, 2 → NS mutation

Mutation in position 3 → can cause a S mutation



		Functional genes		
Gene	Pseudogene	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 ψβ^x and ψ^z	4.4	0.94	0.71	2.02
Average	4.7	0.85	0.70	2.34

Mutations falling in pseudogenes or S mutations will be neutral so they can all pass on to the next gen

The molecular clock hypothesis

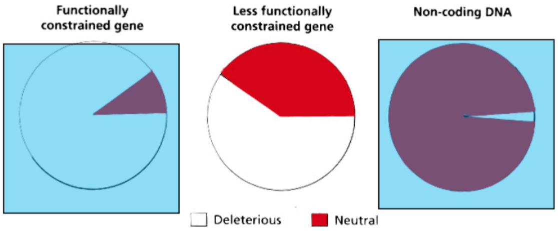
Given a DNA/protein sequence evolves at a relatively constant rate.

Evolutionary rates vary greatly between genes.

The stronger the functional constraints the slower substitution rate

Substitution rates correlate with the intensity of purifying selection

Hypothesis: mutation rate introduces mostly deleterious and neutral mutations. However, the first are removed by purifying (negative) selections. Most of the observed variation is selectively natural



The neutral theory of molecular evolution

They thought the predominant evolutionary force was neutral (positive) selection.

Molecular clock → constant rate of adaptive substitutions

Molecular clock and evidences of functional constraints are compatible with the neutral theory of molecular evolution. Fundamentals:

- Molecular level: most of observed variation (aft. purifying selection) is neutral, by genetic drift.
- Slightly deleterious or beneficial mutations that cannot counteract the effect of genetic drift and behave as effectively neutral.
- Polymorphism and divergence are two pictures of the same neutral process. Neutral mutations segregate as transient polymorphism until they are fixed or lost.

- Positive selection is rare and episodic. Beneficial mutations do not contribute significantly to the observed variation.

Polymorphism: neutral mutation is an ongoing process which gives rise to genetic polymorphisms.
Substitution: 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 variants requires knowing the probability that a variant becomes fixed in a population and the number of new mutations that occur each generation.

Most mutations that occur through time are deleterious, there is a little neutral and very little positive.

Neutral site: DNA position at which all alleles are selectively equivalent:

- Mutation rate per generation and per gamete is μ
- Expected number of new mutations (per generation) in the whole population is: $2N\mu$ (there are 2N gametes)
- Initial frequency of all new mutations in a diploid population of N individuals is $1/2N$. Mutation only occurs in one chromosome of the individual

Given that there is no selection, all alleles have the same probability of fixation (freq=1). This probability is simply its relative frequency. $p = 1/2N$

Substitution rate per generation, r (or k in Kimura) = number of new mutations * p(fixation)
$r = k = 2N\mu \cdot 1/2N = \mu$
⚠️ Only true when all mutations are neutral
Substitution rate = mutation rate → neutral mutation accumulates in a regular rate.

Neutral fraction:

- f_0 → fraction of neutral mutations ($1-f_0$, deleterious): fraction of a gene that can "tolerate" mutations because they are neutral.
- $f_0 = 1$ → all mutations are neutral

$\mu_{neural} = f_0\mu_{tot}$

Under the neutral theory: substitution rate is equal to the mutation rate in the neutral fraction.
$r = \mu$ neutral

MODELS OF DNA SEQUENCE EVOLUTION

The phylogenetic Analysis

1. Sampling
2. DNA Extraction
3. DNA Sequencing (homologous sequences)
4. MSA → Statistical Inference (relies on probabilities)
5. Phylogenetic Tree Inference → Models of DNA/Protein evolution → How to choose the model? Phylogenetic algorithm/method (NJ, ML, Bayesian, etc)
6. Evaluation of the inferred topology → Robustness of the phylogenetic tree (bootstrap, etc)
7. Evolutionary interpretation

Need models of DNA evolution: To infer phylogenetic trees from molecular (DNA/prot) information.

- Perform sequence alignments
- Estimate divergence
- Simulate sequence evolution
- Detect natural selection

Main objectives of molecular phylogeny are:

- Determine a hierarchical relationship between species according to their relationship evolutionary
- Estimate the divergence time between species (time to the most recent common ancestor)

SEQUENCES → ALIGNMENT → genetic divergence (or distance) → PHYLOGENY

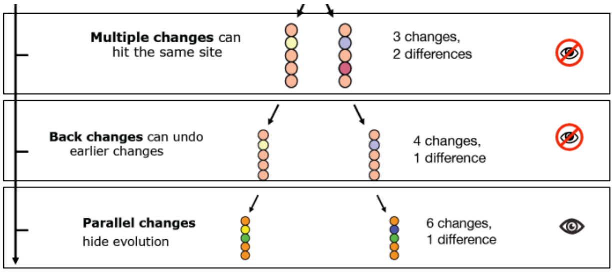
Genetic distances (aa substitution per site)
Measure of the genetic divergence between species or between populations within a specie:

- Upper diagonal → % observed aa differences
- Corrected aa distance (to solve the problem of multiple hits)

k = genetic distance → # aa replacements per site (usually higher than the uncorrected one)

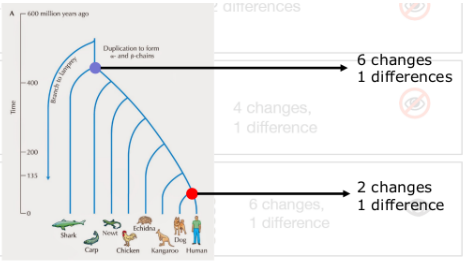
Mutations & Multiple Substitutions
Transitions: purine → purine, pyrimidine → pyrimidine.
Transversions: purine → pyrimidine, …

Nucleotide changes


Multiple changes can hit the same site
3 changes,
2 differences
Back changes can undo earlier changes
4 changes,
1 difference
Parallel changes hide evolution
6 changes,
1 difference

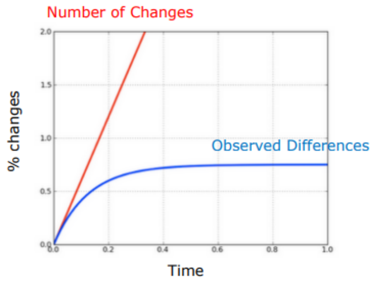
Multiple substitutions: differences in the same position (reality and observed is different→ we only see the last change and the ancestor is unknown).
Back changes: can undo earlier changes.
Parallel changes: hide evolution (parallel mutations in both descendants)

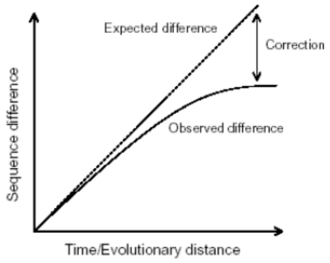
The longer since the split from the common ancestor the more discrepancy between the observed differences and the real changes



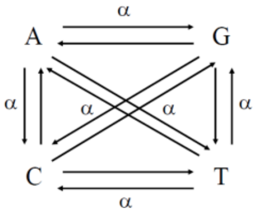
Models of nucleotide substitution

Rely on mathematical models that account for the nucleotide substitution process (since we cannot repeat the evolutionary history).





Jukes-Cantor (JC69) one parameter model assumes that all nucleotides have the same probability of changing



$$M = \begin{array}{ccccc} A & G & C & T & \\ 1-3\alpha & \alpha & \alpha & \alpha & A \\ \alpha & 1-3\alpha & \boxed{\alpha} & \alpha & G \\ \alpha & \alpha & 1-3\alpha & \alpha & C \\ \alpha & \alpha & \alpha & 1-3\alpha & T \end{array}$$

M_{ij} is the probability to change from i (row) to j (column). in the example, G → C

Kimura 2 parameter → K2P assumes that each nucleotide has equal probability to be substituted by any of the other 3 in a fixed period of time. (but transitions α more often than transversions β), α > β

$$M = \begin{array}{ccccc} A & G & C & T & \\ 1-\alpha-2\beta & \alpha & \beta & \boxed{\beta} & A \\ \alpha & 1-\alpha-2\beta & \beta & \beta & G \\ \beta & \beta & 1-\alpha-2\beta & \alpha & C \\ \beta & \beta & \alpha & 1-\alpha-2\beta & T \end{array}$$

General Time-Reversible Model (GTR)

The most general time reversible model, allows various instantaneous rates of substitution between each of the 6 nucleotide pairs. a = A↔C, b = A↔G, c = A↔T, d = C↔G, e = C↔T, and f = G↔T

	A	C	G	T
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	$c\pi_A$	$e\pi_C$	$f\pi_G$	—

More complex models

Real data often shows variability of substitution rates between sites, even some sites might be completely invariant (purifying selection). 3 nucleotides of a particular codon usually exhibit different substitution rates. Also, heterotachy (subst. rates vary along time)

Number of Nucleotide Substitutions (per site between 2 sequences) (k)

Can be estimated from the # differences between the 2 sequences.

Genetic distance: measure of genetic dissimilarity between species (divergence) or individuals (polymorphism)

Si hi ha molt temps (divergence), el número de differences és probable que sigui més petit que el número real de substitutions (multiple hits).

Jukes and Cantor correction: we can estimate K from p (proportion of differences between two sequences)

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

Changes, Substitutions & Rates

We can compute the speed of the substitution rate from k and the divergence time

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

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

MODELS OF PROTEIN SEQUENCE EVOLUTION

BLAST

Performs the alignment. Score:

- Match
- Mismatch
- Gap → creation or extension

Scoring matrix: BLOSUM62 (based on aa changes and their possible impact).

Substitution Matrices

Contains values proportional to the probability that one aa mutates into a different aa.
Matrix constructed by assembling a large and diverse sample of verified pairwise alignments of aa.

PAM: Point Accepted Mutations

PAM1 compares seq with no more than 1% of divergence (per 100 aa)
PAM10 → 10 mutations per 100 aa.
PAM250 → 250 mutations per 100aa. For very different sequences, ~20% aa identity.
PAM250 = PAM1^250 (extrapolation)

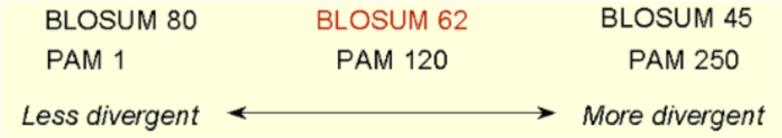
Log-odds scoring matrix: PAM250

$$s_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_{i,j}}\right)$$

$q_{i,j}$: observed frequency of substitution (i to j)
$p_{i,j}$: observed frequency of amino acids i and j

BLOSUM matrices

Based on local multiple alignments of distantly related proteins. They are not extrapolated from comparisons of closely related proteins.
BLOSUM62 → sequences sharing up to 62% similarity



BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250

Less divergent ⟷ More divergent

BLOSUM → based on similarity
PAM → based on divergence

Twilight zone

MOLECULAR PHYLOGENETICS CONCEPTS AND THE TREE THINKING

Molecular phylogenetics

The study of relationships among organisms by using molecular data.
Reconstruction of the evolutionary history of organisms
Investigation of the mechanisms of evolution

Phylogenetic Tree Reconstruction

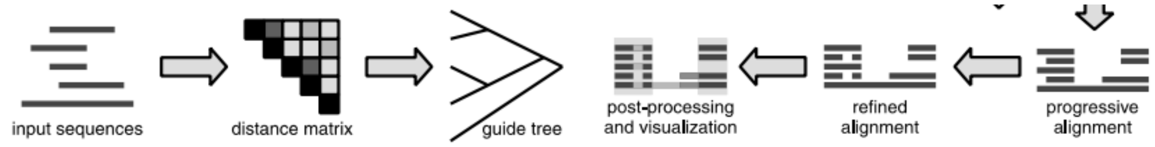
1. Sampling - DNA extraction - DNA sequencing

If we don't sample appropriately we can get erroneous conclusions
Increasing the # markers adds information (but also risky, conflicting signals)

2. MSA

How do we do these alignments?
Pairwise alignments → dynamic programming algorithms (find the optimal alignment).
MSA → we cannot do that (DP is very computationally demanding)

Use heuristic approaches (try to find a solution close to the real one)
Based on progressive alignment.
Input sequence → distance matrix → guide tree
We can align the sequences that are close in the tree and progressively add more sequences in the order provided in the tree.
When the score is not improved anymore we get the final alignment.
Probably is not the most optimal alignment, but it's a good approach (close to the optimal).



3. Phylogenetic tree inference

From the alignment we can infer the homologous positions.
Inference method

- Individual genes: get the gene tree, not the species tree
- Super matrix: concatenate the alignments
- Models

Phylogenetic algo/method:

- NJ
- distances
- ML: probabilistic based on substitution models
- bayesian: probabilistic based on substitution models

Once we have the phylogeny we can test it for the uncertainty

- Bootstrap
- Posterior probability

OTU (Operational Taxonomic Unit) = leaf = external node
Root gives us the path of the evolution

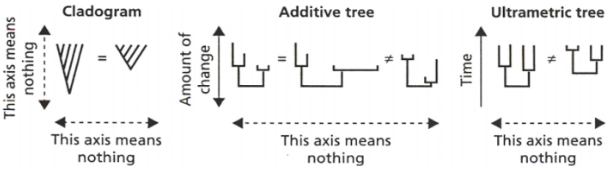
Dichotomy: del node certain 2 branques
Polytomy: del node certain +2 branques

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; members do not form a natural clade

Ingroup: group we are interested in
Outgroup: used to root the phylogeny

Tree topology: branching pattern

- Cladogram
- Additive tree: y axis amount of change
- Ultrametric tree: y axis time



Unscaled: branches not proportional to the # changes
Scaled: branches proportional to the # changes

Rooted: MRCA (most recent common ancestor) is identified
Unrooted: only specifies the relationship among taxa, without reference to the direction of the evolutionary time (the common ancestor is unspecified)

$$U_N = \frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad R_N = \frac{(2n-3)!}{2^{n-2}(n-2)!} = (2n-3)U_n$$

n = # taxa

Bayesian or ML: algorithm has to check for all possible trees and select the one with the most likely probability (too many trees)

UPGMA

Simplest way to reconstruct tree (we don't even use it)

- Rooted tree
- Constant evolutionary rates (ultrametric)

NJ

- Unrooted
- Not constant evolutionary rate

Bootstrapping

Measures the robustness of the tree
Generate alignments from the original one to see if all the positions in the alignment are consistently supporting the tree → we get a consensus tree.

MOLECULAR ADAPTATION AND FUNCTIONAL DIVERGENCE

PART 1: ANALYSIS OF SELECTIVE FORCES ACTING ON PROTEIN CODING GENES: CODON SUBSTITUTION MODELS

- Deleterious → rarely fixed (negative selection)
- Neutral or nearly neutral → lost or fixed (genetic drift)
- Beneficial → rapid fixation (positive selection) → molecular adaptation

Signatures: type, age, strength

Codon based MSA
Input: alignment of a coding region
Unit of evolution = codon
$\omega = d_N/d_s$

Based on the presence of S and NS substitutions

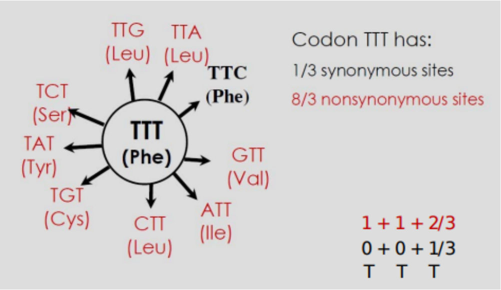
- $\omega = 1$: neutral evolution
- $\omega < 1$: purifying (negative) selection
- $\omega > 1$: diversifying (positive) selection

Only works when estimating 2 divergences → independent ev lineages

In pairwise comparisons or in a phylogenetic context:
Count synonymous (S) and non-synonymous (N) sites
Count synonymous (s) and non-synonymous (n) substitutions
Apply some correction (e.g. correct for multiple hits

Codon substitution models

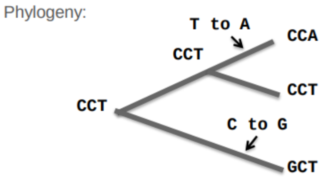
1. Counting sites



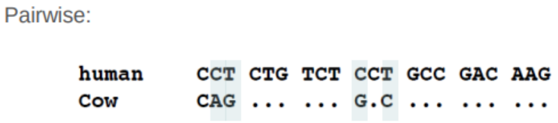
Codon TTT has:
1/3 synonymous sites
8/3 nonsynonymous sites

1 + 1 + 2/3
0 + 0 + 1/3
T T T

2. Counting differences



```
human  CCA
rat    ..T
cow    G.T
```



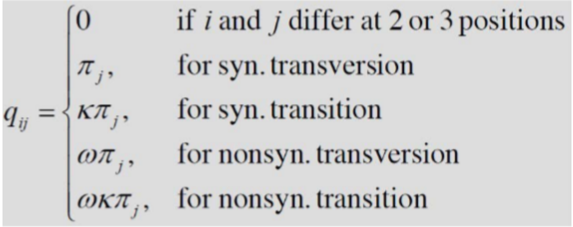
```
human    CCT CTG TCT CCT GCC GAC AAG
Cow      CAG ... ... G.C ... ... ...
```

3. Applying corrections

Correction for multiple hits have been based on nucleotide substitution models and are invalid but errors will be low if sequence divergence is low
Ignoring the transition/transversion rate bias leads to underestimation of S (underestimation of ω)
Codon usage bias has the opposite effect and can be more important (false positives of positive selection) for some genes not all S are used with the same frequency.
Computing the ω ratio with these methods can be tricky!

<u>Markov model of codon evolution</u>

1. Estimate parameters (ω and K)
2. Correct multiple hits
3. Weight of evolutionary pathways between codons



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

Maximum likelihood estimation
Likelihood function:

$$L_h(CCC, CCT) = \sum_k \pi_k p_{kCCC}(t_0) p_{kCCT}(t_1)$$

Maximum likelihood estimation

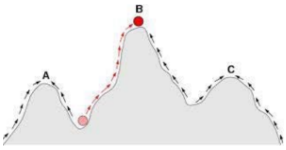
$$L = L_1 \times L_2 \times L_3 \times \ldots \times L_N = \prod_{h=1}^{N} L_h$$

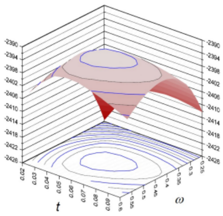
Log likelihood

$$\ell(t, \kappa, \omega) = \ln\{L\} = \ln\{L_1\} + \ln\{L_2\} + \ln\{L_3\} + \ldots + \ln\{L_N\} = \sum_{h=1}^{N} \ln\{L_h\}$$

Do it for all the codons in the alignment and multiply probabilities

Numerical hill-climbing algorithm to maximize the likelihood function





Maximum likelihood estimation of parameters:
ω, κ, τ
π_S = empirical?

Models for heterogeneous selection pressure: lineages
Selection pressure varies across lineages

Estimate omega independently in several lineages. Then, we can infer where positive/negative selection is acting → branch models (use codon substitution models to estimate omega in different branches).
Podem veure que la pressió selectiva pot variar entre les diferents branches d’un phylogenetic tree

Model 0: all branches under the same constraint
Model 1: lineages are evolving under different omegas

$$L_h(CCC, CCT) = \sum_k \pi_k p_{kCCC}(t_0) p_{kCCT}(t_1)$$

pkcc is omega1

We can compute the log likelihood of both and compare which model is better
LR statistic = 2(log L M1 - logL M0) → follows a chisq distribution

Selection pressure varies along sites

a. Use prior information to partition sites and test selection on a particular partition.
   Estimate the omega ratio for each site
b. Use a statistical distribution to model omega variation (we don't know the specific omega value of each site but we can obtain a posteriori estimate from the fitted distribution)

At the end you can have p of different parts of the protein. If you consider all the codons in the gene evolving with the same omega value, maybe you can’t identify the positions evolving under positive selection.