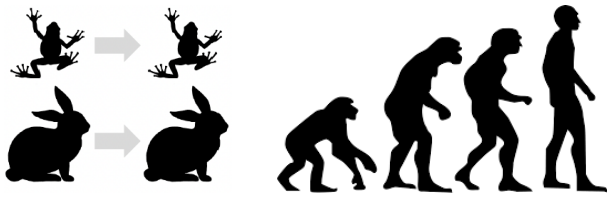
Session 1: Genetic variation

Refers to the differences that exist in the DNA sequence of individuals within a population. These variations can arise from mutations that occur randomly during DNA replication, recombination events during meiosis, or from the mixing of genetic material from different populations.

Fixisim: everything that exists, including all life forms, has not varied along time and everything is today exactly identical to how it was in the past and will be in the future. (Variation → Inconvenience)

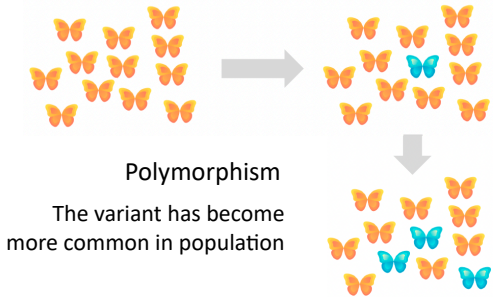
Evolution: change in the heritable characteristics of populations over successive generations. It does not take place in inviduals, but in populations, which change over time. (Variation → Essential)

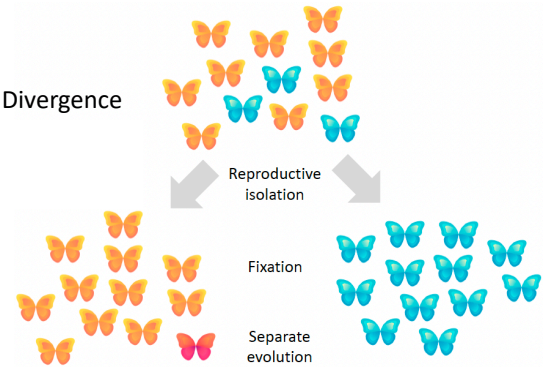


Polymorphism *refers to the presence of multiple versions, or alleles, of a gene within a population.* ***Divergence*** *refers to the process by which populations or species become increasingly different from one another over time.*

Polymorphism: genetic differences <u>within</u> a species that are generated by mutations. Variants with allele frequencies > 5% or > 1% are considered polymorphisms.

Divergence: genetic differences <u>between</u> species.





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

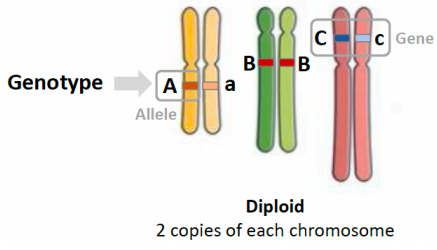
Gene: DNA sequence that codes for an RNA or protein

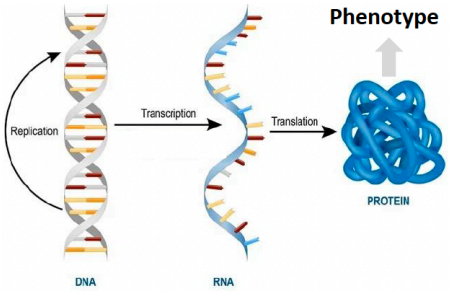
Allele: variant or alternative form of the DNA sequence at a given gene/copy of a gene in a diploid organism

Diploid organism: organism that has two sets of chromosomes in each of its cells. In humans, for example, diploid cells contain 23 pairs of chromosomes, for a total of 46 chromosomes. One set of chromosomes is inherited from the mother and the other from the father.

Haploid organism: organism that has only one set of chromosomes in each of its cells. Each chromosome is represented only once. Haploid organisms can reproduce either sexually or asexually.

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

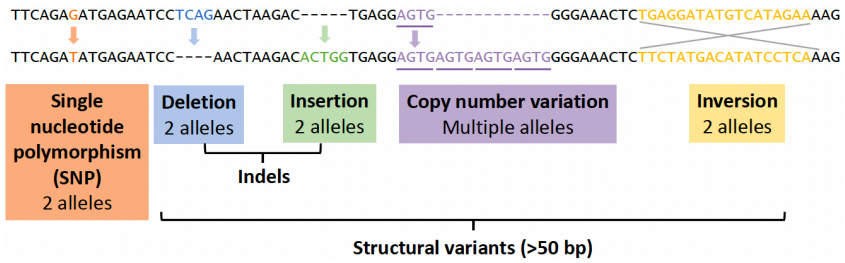




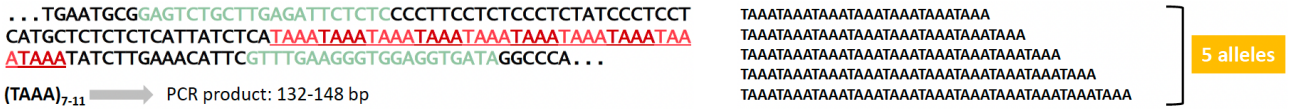
Types of variants (polymorphisms)

- **Single nt polymorphisms** (**SNPs**): variations in a single nt (A, T, C, or G) at a specific location in the DNA sequence. Can occur in both coding and non-coding regions of DNA.
- **Insertion/deletion** (**indels**): involve the insertion or deletion of a small segment of DNA, usually from one to several nts in length. Indels can cause frameshift mutations that alter the reading frame of the gene, potentially leading to the production of a non-functional protein.

- **Copy number variation** (**CNVs**): variations in the number of copies of a particular DNA segment within an individual's genome. Can range in size from a few 100 bp to several megabases
- **Inversion**: involve the reversal of the orientation of a segment of DNA within a chromosome. Can range in size from a few 100 bp to several megabases and can affect gene expression, recombination and chromosome segregation.

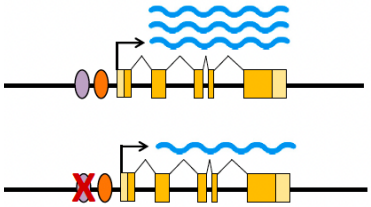


CNV → Microsatellites/Short Tandem Repeats (STRs)
Microsatellites: repeated sequences of 2-5 bp polymorphic in their length (multiallelic variants)



Do all variants have effect on phenotype?
→ Only polymorphisms in **functional elements** may have pehotypical effects
- Polymorphisms **inside** coding regions will NOT always have consequences
- Polymorphisms **outside** coding regions CAN have consequences

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



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

Allele number = k

2 alleles	3 alleles	4 alleles
AA	AA	AA
AB	BB	BB
BB	CC	CC
	AB	DD
	AC	AB
	BC	AC
		AD
		BC
		BD
		CD

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

How many genotypes we can find in alleles.
Exception: X-linked variants

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

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

All allele and genotype frequencies must add 1

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

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

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

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

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

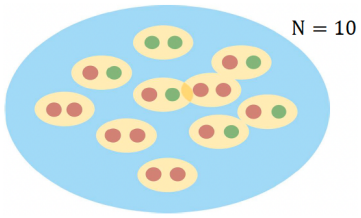
$$\text{Freq}(aa) = \text{Q} = \frac{4}{10} = 0.4$$

Allele frequencies in diploid organisms
2 alleles: A and a

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

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

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



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

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

Allele and genotype frequencies — Counting method

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

Total number of alleles = 2N

ALLELE FREQUENCIES

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

From individual counts with each genotype

From genotype frequencies

Example problem (diapo 22)

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

Genotype	# of individuals	Genotype frequencies	# of Pink alleles	# of Purple alleles
Pink/Pink	170	P = 170/531 = 0.32	170 * 2 = 340	0
Pink/Purple	21	H = 21/531 = 0.04	21	21
Purple/Purple	340	Q = 340/531 = 0.64	0	340 * 2 = 680
Total	531	1	361	701

Allele frequencies
p = 361/1062 = 0.34
q = 701/1062 = 0.66

p + q = 1

Total alleles = 361 + 701 = 1062

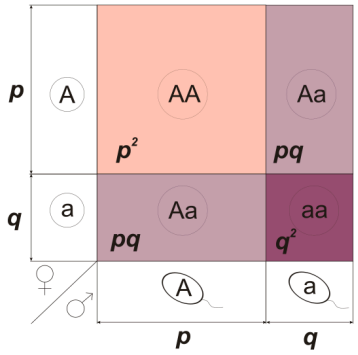
Evolution: change of allele frequencies in a population over time
Hardy-Weinberg equilibrium: provides a null model, a prediction based on a simplified/idealized situation, where no biological processes are acting and genotype frequencies are the result of random combination.
Assumptions:

- Diploid organism
- Sexual reproduction
- Non-overlapping generations
- Random mating
- Equal allele frequencies in both sexes

- Large population size
- No migration, no mutation, no selection

Principles:

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



GENOTYPE FREQUENCIES

P (AA) = p^2
Q (Aa) = 2pq
R (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

PARENTAL MATING FREQUENCIES

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

		Offspring genotype frequencies		
Mating	Frequency	AA	Aa	aa
AA x AA	P^2	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'**

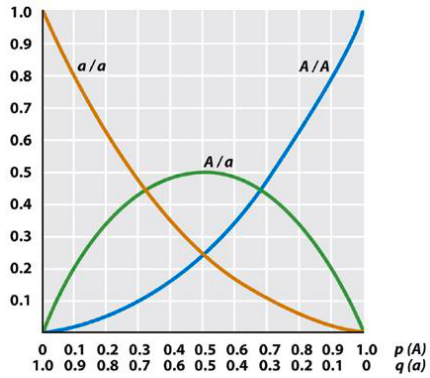
$$P' = P^2 + \frac{2PH}{2} + \frac{H^2}{4} = \left(P + \frac{H}{2}\right)^2 = p^2$$

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

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

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

H-W equilibrium expected genotype frequencies



p = 0.75
q = 0.25

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

H-W equilibrium in X-linked genes

- Genotype frequencies among **females**:
  - AA = p^2
  - Aa = 2pq
  - aa = q^2
- Genotype frequencies in **males**:
  - A = p
  - a = q

Males only have one X chromosome (XY)

		FEMALE GAMETES	
		A (p)	a (q)
MALE GAMETES	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)
	Y	AY (p)	aY (q)

Female offspring

Male offspring

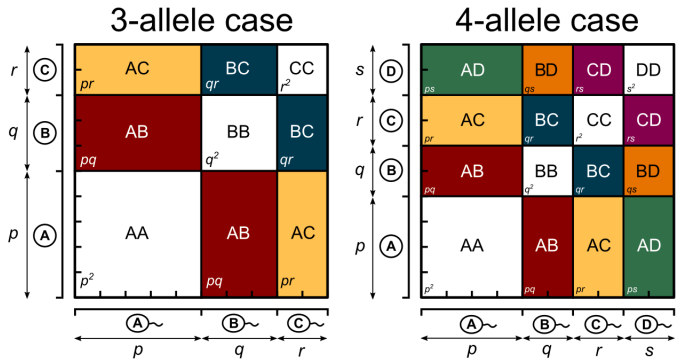
If *a* is a recessive allele there will be **more males** exhibiting the trait than females since the frequency of affected females (q^2) will be smaller than the frequency of affected males (q)

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 1st round of random mating, same allele and frequencies in both sexes; after the **2nd** round of random mating, HWE will be established

HWE 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 HWE

- Null model to analyze the effect of different factors on the genetic composition of a population
- Test if genotype frequencies adjust to expected values > If they don't, one of the assumptions is not true
- Estimation of allele frequencies in case of dominance

Adjustment of genotype frequencies to HWE

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

Estimating allele frequencies with dominance

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

ALLELE FREQUENCIES

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

$Freq(D) = p = 1 - 0.3768 = 0.6232$

GENOTYPE FREQUENCIES

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

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

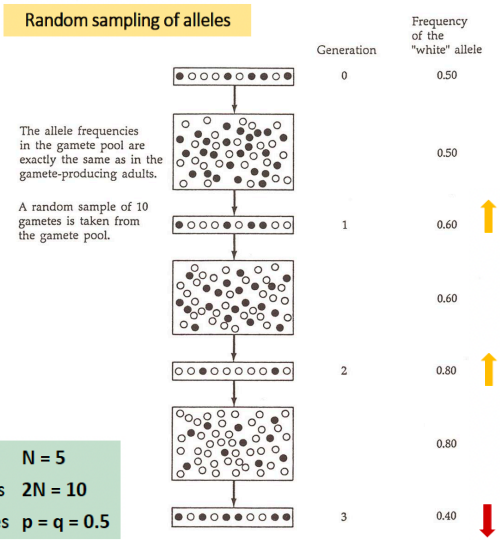
PROPORTION OF HETEROZYGOTES WITHIN Rh+

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

Session 3 and 4: Genetic drift and mutations

Genetic drift: stochastic process from which we can only predict the probability of each possible outcome in the next generation

- **No drift** → allele frequencies equals to 0.5 in all successive generations (*obtain 5 white and 5 black gametes always*)
- **Finite** population size → allele frequencies <u>fluctuate</u> ↑ and ↓



Population size	N = 5
Number of alleles	2N = 10
Allelic frequencies	p = q = 0.5

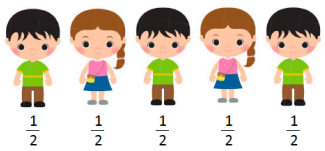
H-W equiibrium assuming a <u>small population size</u>
Binomial distribution is used when:

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

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

Probability of your 5 children being 3 boys and 2 girls



5/10 = ½

$\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

BBBGG
BBGBG
BBGGB
BGBBG
BGBGB
BGGBB
GBBBG
GBBGB
GBGBB
GGBBB

Number of combinations of 3 boys and 2 girls

Probability of being a girl

Probability of being a boy

$$10 \cdot (1/2)^3 \cdot (1/2)^2 = 0.3125$$

Probability of getting 3 heads and 7 tails if you flip a coin 10 times

$\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

Number of combinations with 3 heads and 7 tails

Probability of getting tails

Probability of getting heads

$$120 \cdot (1/2)^3 \cdot (1/2)^7 = 0.1172$$

$$\text{Number of combinations} = \frac{10!}{3!\,(10-3)!} = \frac{10!}{3!\,7!} = \frac{10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{(3\cdot 2\cdot 1)(7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1)} = \frac{10\cdot 9\cdot 8}{3\cdot 2} = \frac{720}{6} = 120$$

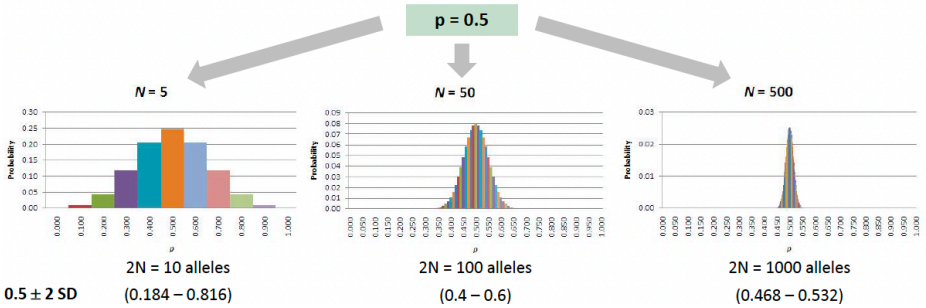
Binomial — Example to calculate (diapo 9)

We have a population of 10 individuals with two alleles in a gene. In a particular generation, the frequency of allele A1 is 0.7 and the frequency of allele A2 is 0.3. Which is the probability of having allele frequencies 0.5 for each allele in the next generation?

10 individuals → 20 alleles
Number of combinations with 5 A_1 and 5 A_2 alleles = 20! / 10! * (20 - 10)! = 184 756
P = 184 756 * $(0.7)^{10}$ * $(0.3)^{10}$ = 0.0308

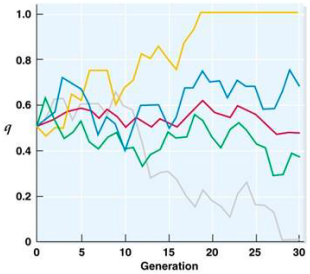
$$P = \frac{2N!}{k!\,(2N-k)!}\,p^k\,q^{2N-k}$$

Decrease in sampling error with increasing population size
The breadth of the distribution narrows as population size ↑ due to a ↓ in a sampling error

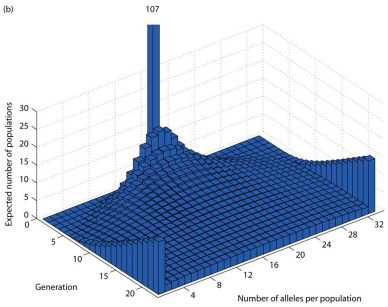


Allele frequencies will change by chance in population of all sizes, but the amount of change due to **sampling error ↓** as **population size ↑**

Allele frequencies over time in populations of different sizes



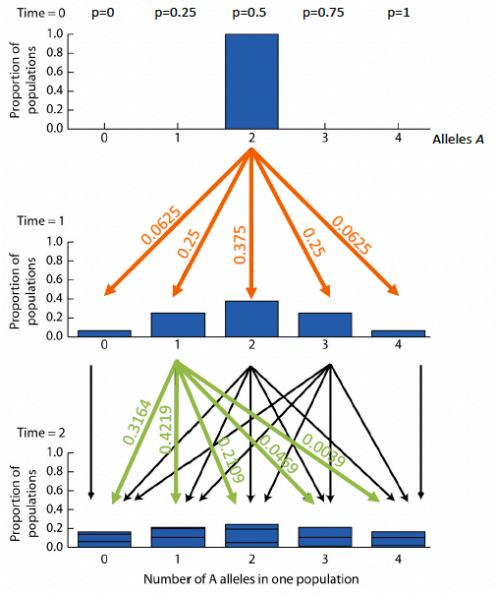
- Lines correspond to independent populations/replicates
- Random changes from one generation to the next
- Allele frequencies that reach upper or lower axis represent cases of **fixation** or **loss**



Wright-Fisher model: describes genetic drift in populations by tracking changes in allele frequencies over generations through random sampling, assuming:

- Infinite populations
- Constant population 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

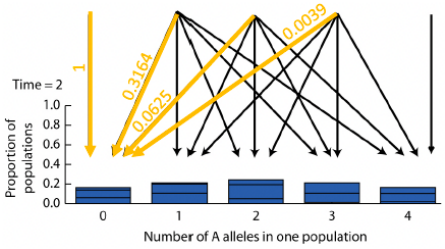
Markov chain model: used to simulate how allele frequencies change over time in a population undergoing random genetic drift, based on probabilities of sampling different alleles at each generation.



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	$1 \cdot (0.5)^4 \cdot (0.5)^0 = 0.0625$

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

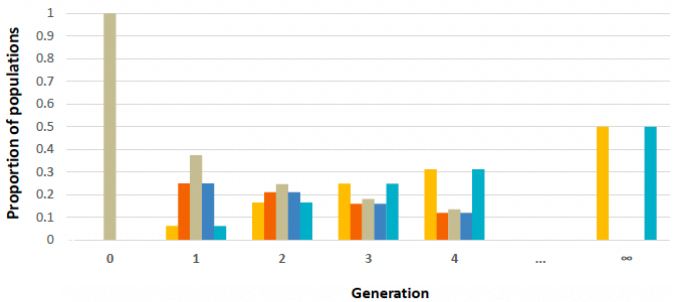
Number of alleles *A* in generation t+1 \ Number of alleles *A* in generation t	0	1	2	3	4
0	1	0.3164	0.0625	0.0039	0
1	0	0.4219	0.25	0.0469	0
2	0	0.2109	0.375	0.2109	0
3	0	0.0469	0.25	0.4219	0
4	0	0.0039	0.0625	0.3164	1



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

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 (**P**)
- Mean heterozygosity ↓ with time (**H**)
- Variance ↑ with time (**Var(p)**)



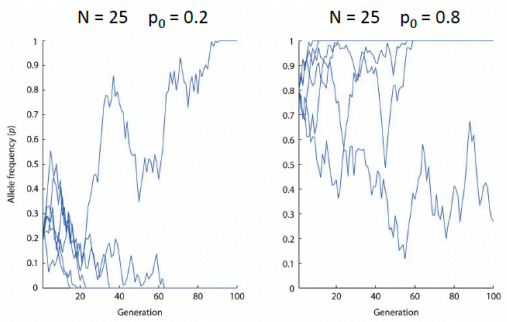
Number of alleles *A* \ Generation	0	1	2	3	4	...	∞
0 (p=0)	0	0.0625	0.1660	0.2490	0.3117	...	0.5
1 (p=0.25)	0	0.25	0.2109	0.1604	0.1205	...	0
2 (p=0.5)	1	0.375	0.2461	0.1813	0.1356	...	0
3 (p=0.25)	0	0.25	0.2109	0.1604	0.1205	...	0
4 (p=1)	0	0.0625	0.1660	0.2490	0.3117	...	0.5
$\bar{p}$	0.5	0.5	0.5	0.5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

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)

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

More populations go to loss of the allele

P_{fix} = 0.2



More populations reach fixation

P_{fix} = 0.8

Probability of fixation (P_{fix}) of a neutral allele is equal to its initial frequency in the population **p_0**

$P_{fix} = p_0$

A new allele with $p_0 = \frac{1}{2N}$ is more likely to be lost than fixed

Genetic drift causes a reduction in heterozygosity

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

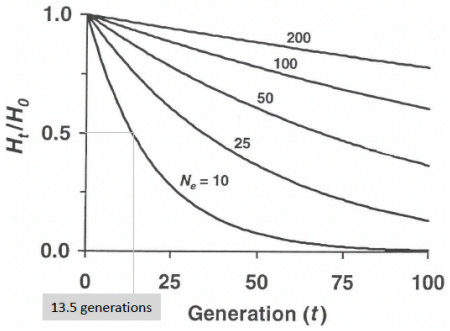
where **N** is the population size, **t** is the number of generations, H_0 represents the initial heterozygosity of the population and H_t represents the heterozygosity of the population after t generations

When one allele is fixed, H = 0

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

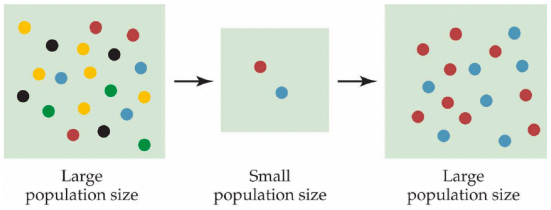
Consequences of genetic drift → Loss of alleles → Loss of heterozygous (proportion of H_0 will ↓)

Decay of heterozygosity under genetic drift



Decay is slower or faster, depending on the size N

Reductions in population size



Genetic drift acts more **quickly** to reduce genetic variation in small populations. The resulting population can have:

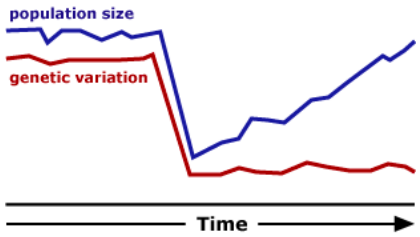
- Reduced **variation** and reduced **ability** to adapt to new selection pressures
- A non-random sample of the genes in the original population

Bottleneck effect occurs when a population's size is reduced by a natural disaster 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)

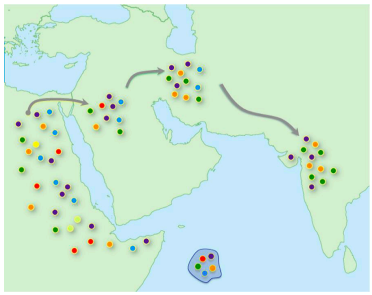
Founder effect occurs when a new colony is started by a few members of the original population

- Consequences: **substantial loss** in genetic diversity, and **rapid divergence** between source and founder populations



Bottleneck

Founder effect



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.

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

Most deviations will ↓ the effective population size

N_e is a usually much **smaller** than N_c

Factors that can contribute to this difference:

1. Different number of males and females
2. Fluctuations in population size
3. Variation in the number of offspring among individuals (some individuals contribute more/lees in offspring)
4. Bottlenecks
5. Overlapping generations

1) N_e in a population unequal sex ratio for autosomic genes

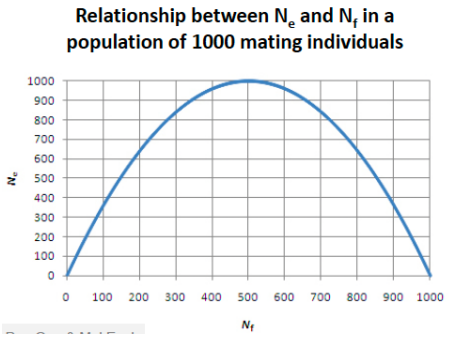
$$N_e = \frac{4N_m N_f}{N_m + N_f}$$

N_m = number of males
N_f = number of females

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

N ≈ 100.000
N_f = 1
N_m >>> 1

N_e = 4



2) Fluctuations in population size

- Populations can show regular cycles of ↑ and ↓ 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

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

1/Ne = 3 / (1/100+1/10+1/100)

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

Important ideas

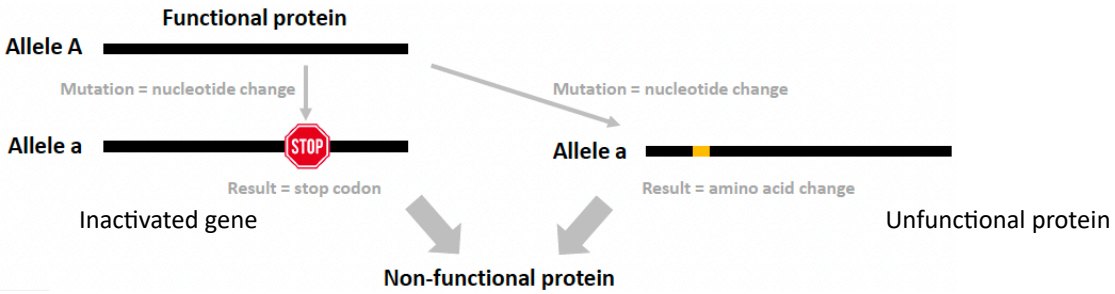
- Allele frequencies change **randomly** due to **sampling error**
- The direction of the change is unpredictable (allele frequencies will randomly ↑ and ↓ over time)
- **Cumulative behavior** (each generation allele frequency will tend to deviate more and more from the initial frequency, and probability of fixation ↑ with time)
- The amount of change due to sampling error ↓ as the population size ↑ (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 allel is equal to its initial **frequency**
- **Heterozygosity** will ↓ 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 (**N_c**)
- What changes allele frequencies → Natural selection, genetic drift, mutation and migration

Mutations

Mutation as an evolutionary force that changes allele frequencies

Mutation: source of all genetic variation. It introduces new alleles in populations. It is any permanent change in an organism's DNA (from nt substitutions to large SV) and is the result of unrepaired damage in DNA and errors during DNA replication or repair.

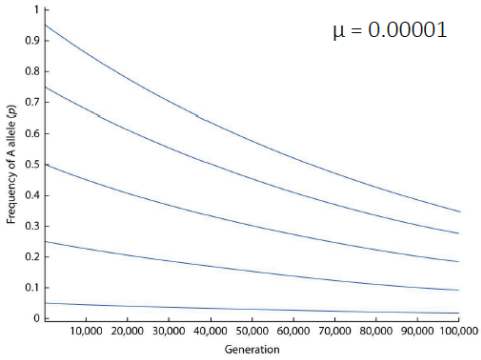
- In the germinal line (cells that give rise to sperm and eggs) are transmitted to **offspring**, but somatic mutations (body cells) are **not**
- At phenotypical level, mutation can be considered **recurrent**
- At molecular level, most mutations are **unique**



Irreversible mutation

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

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



Changes in allele frequency due to mutation alone occur over very long time scales

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

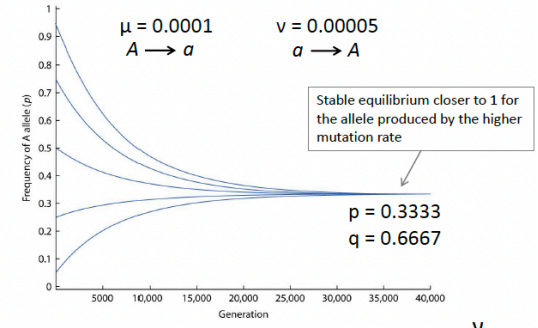
Reversible mutation

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

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

μ > ν

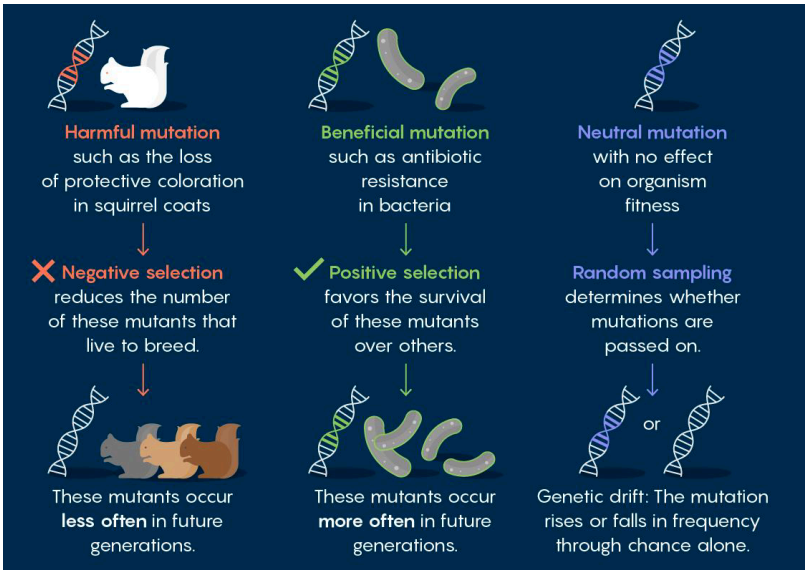
- The rates of mutation from wild type to a novel allele (**forward mutations**) are nearly a factor of 10 more common than mutations from a novel allele to wild type (**reverse mutations**)
- This asymmetry occurs because there are more ways mutation can cause a normal allele **to malfunction**, than ways to exactly restore that function once it is disrupted

Neutral theory of molecular evolution: majority of mutations that occur at the molecular level are **neutral**, meaning they do **not** affect the fitness of the organism. The **rate of fixation** of neutral mutations is largely determined by genetic drift (random fluctuation of gene frequencies in a population due to chance events). Therefore, neutral mutations that are not subject to natural selection can **accumulate** over time, leading to the gradual evolution of genetic diversity within populations and between different species.

Models of molecular evolution

→ Genome evolution depends more on natural selection or on genetic drift?

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

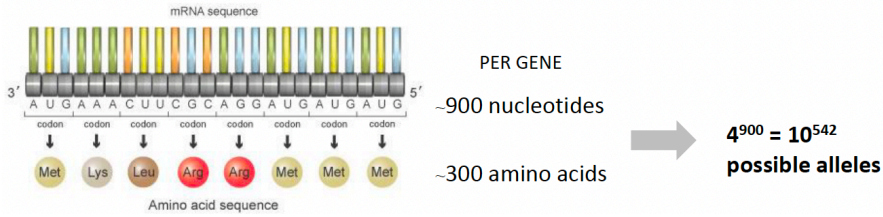


Selectionist models propose that most genetic changes arise due to natural selection and are therefore adaptive. Neutralist models propose that most genetic changes arise due to random genetic drift and are therefore neutral in terms of their effect on an organism's fitness. These models of evolution predict different proportions of beneficial and neutral mutations.

Evolution occurs primarily by natural selection

Allelic frequencies change mainly by genetic drift

Infinite alleles model



PER GENE

~900 nucleotides

~300 amino acids

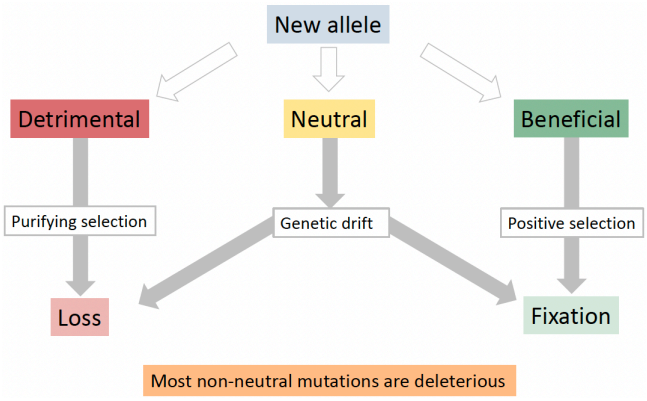
$4^{900} = 10^{542}$ **possible alleles**

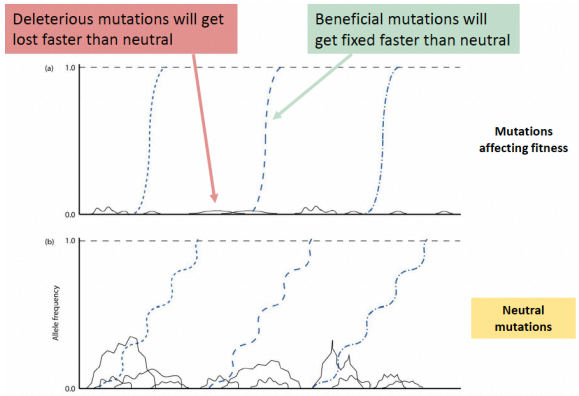
Every time a mutation occurs, it creates a **new** allele

In a finite population

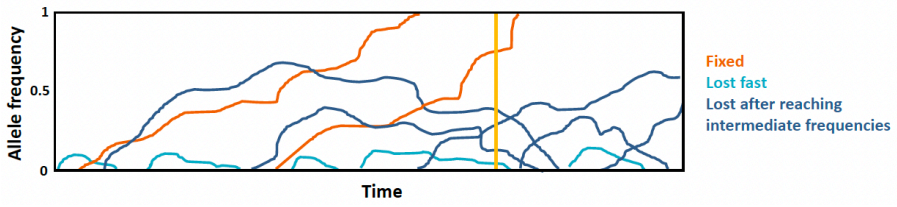
Mutation creates alleles ⟷ **Genetic drift** eliminates alleles

Mutation effects



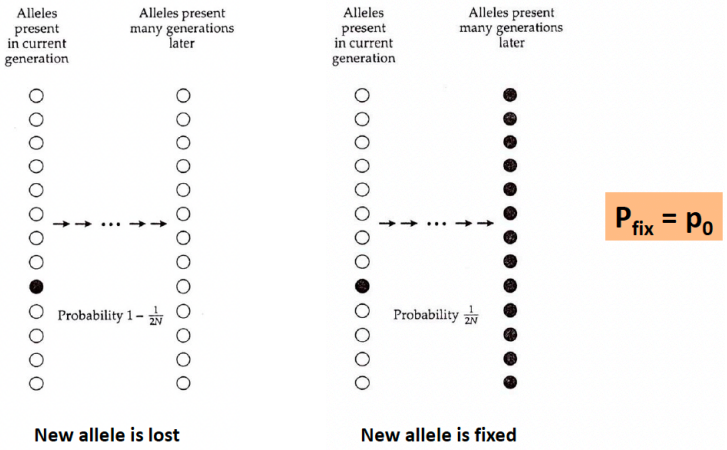


Most genetic variation is maintained in populations simply due to the **random allele-frequency** walk that new mutations take before reaching either fixation or loss



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

Fixation or loss of an allel in a finite population



$P_{fix} = p_0$

Neutral evolution rate

- **Mutation rate (μ)**: rate at which changes are incorporated in a nt sequence during DNA replication and reparations processes
- **Substitution rate (K)**: rate at which new mutations are fixed in the population

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

NEUTRAL MUTATIONS

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

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

$$\boxed{K = \mu}$$

EXAMPLE

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

$N = 10000$ individuals $\qquad 2N = 20000$ alleles

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

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

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

Average time to fixation or loss

Average time to fixation

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

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

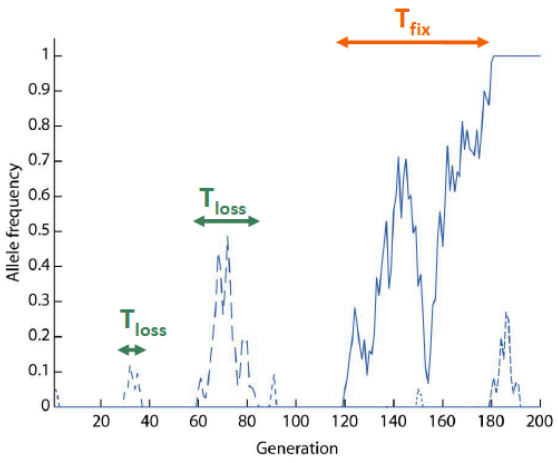
Average time to loss

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

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

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

New alleles introduced every 30 generations into a population of N_e = 10



New allele

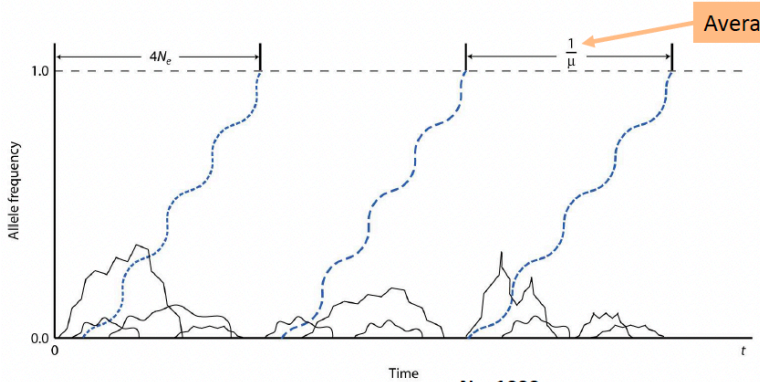
N = 1000000
$p = 5 \cdot 10^{-7}$

T_{fix} = 4000000 generations

T_{loss} = 29 generations

Polymorphism under netural 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 a much longer time than the mutations that end in loss



High levels of polymorphism will result from:

- High mutation rate
- Large population size (= low genetic drift)
- Intermediate levels of mutation and genetic drift

NEW ALLELES

$P_{fix} = \frac{1}{2N}$ $\quad T_{fix} = 4N$ $\quad T_{loss} = 2\ln(2N)$

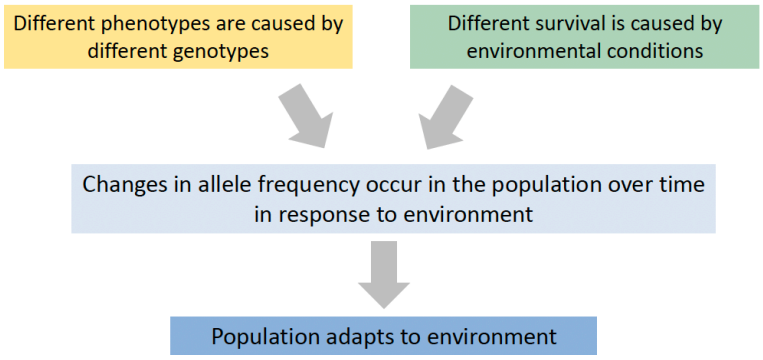
N = 1000

$P_{fix} = \frac{1}{2000} = 0.0005$ $\quad T_{fix}$ = 4000 generations $\quad T_{loss}$ = 15 generations

Session 2: 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 ↑ in frequency of the favorable traits in the population over generations. Natural selection:

- Requires existing heritable variation in a group
- Requires a casual relationship between genotype and number of offspring
- Depends on the environment
- Results in a greater adaptation of organisms to their environment over time



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

Basic selection model

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

Absolute fitness (**W**): measurement of the ability to survive of each genotype

Relative fitness (**w**): ability of one genotype to survive relative to another genotype taken as reference

→ Fitness depends on the environment

Genotype frequencies

$\bar{w}$ = average fitness

	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)	$p^2 w_{AA}$	$2pq w_{Aa}$	$q^2 w_{aa}$	$\bar{w} = p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}$
Adults (normalized)	$\frac{p^2 w_{AA}}{\bar{w}}$	$\frac{2pq w_{Aa}}{\bar{w}}$	$\frac{q^2 w_{AA}}{\bar{w}}$	1

Allele frequencies

Average fitness of allele A

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

Average fitness of allele a

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

Relative fitness values, and not their magnitudes (absolute values), determine the change in allele and genotype frequencies

↓

We will use **relative fitness values** to calculate genotype and allele frequencies in the next generation

In equilibrium allele frequencies do not change $\Delta p = p' - p = 0$

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

$\bar{w}_A > \bar{w}$	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

Types of selection

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

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

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

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

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 – s

Relative fitness values (w)

AA	AB	BB
1	1 – hs	1 – s

Possible values for s

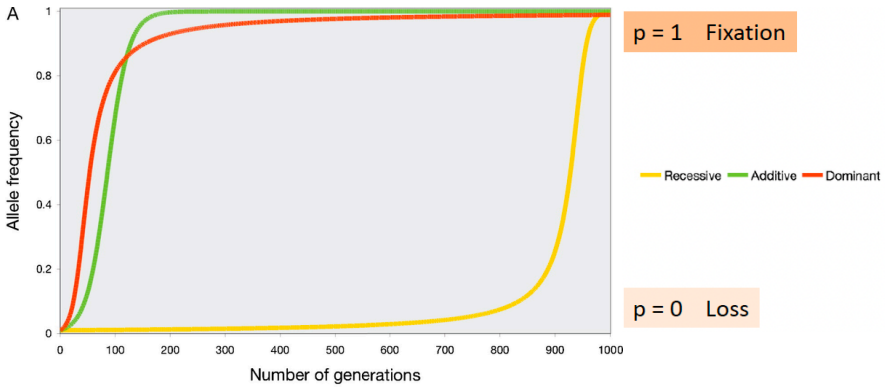
s	w	Result
0	1	No selection
1	0	Lethal
0 < s < 1	0 < w < 1	Some degree of natural selection

Favorable allele = A

Dominance	h	Fitness AA	Fitness AB	Fitness BB	Fitness in heterozygotes
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

Since selection acts on phenotypes, the rate of change in allele frequency (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 = 0.1
p_0 = 0.01



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

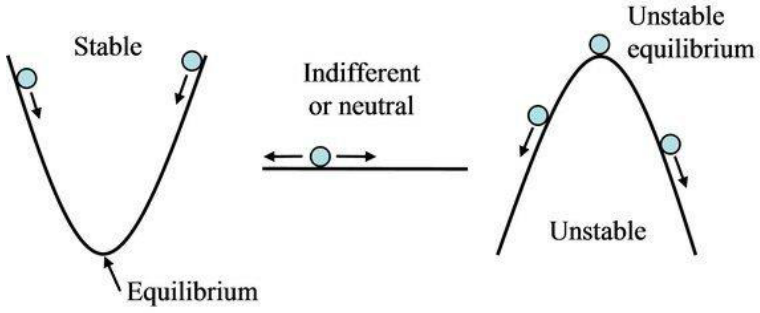
$w_{AA} = 1 - s_1$
$w_{Aa} = 1$
$w_{aa} = 1 - s_2$

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

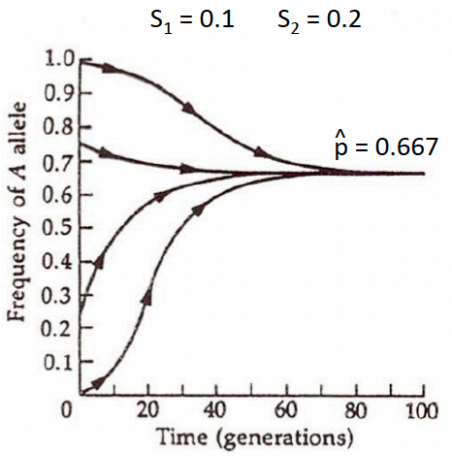
Equilibrium

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

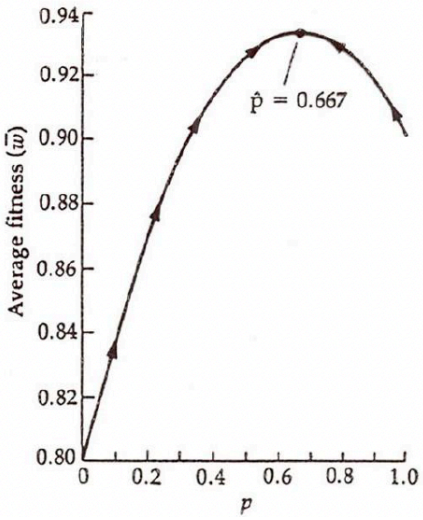
Types of equilibrium



Stable equilibrium → allele frequencies converge to an equilibrium value irrespective of initial frequencies



Average fitness is at a maximum at equilibrium



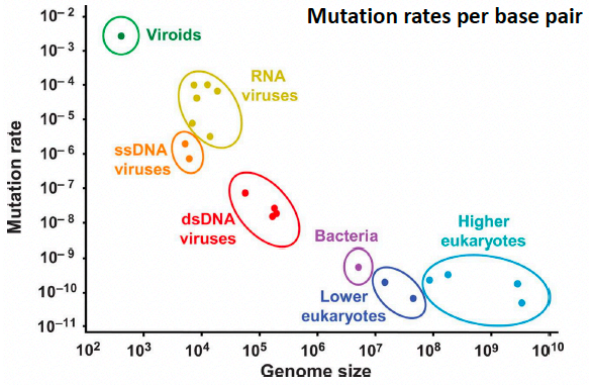
Underdominance or heterozygote inferiority
→ Polymorphism can be maintained under certain equilibrium allele frequencies, but any deviation from these frequencies will lead to fixation of one of the alleles, so it is an extremely rare phenomenon

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 rates are generally ↓

Mutation rates per gene = 10^{-4} – 10^{-6}



Mutation-selection balance

***A* = favored allele**
p(*A*) ≈ 1

$A \xrightarrow{\mu} a$

***a* = harmful recessive allele**
q(*a*) ≈ 00

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

$$\hat{q} = \frac{\mu}{sh}$$

Session 4: Migration and population structure

Two different populations of the same species can have different allele frequencies and even different alleles. How does this happen?

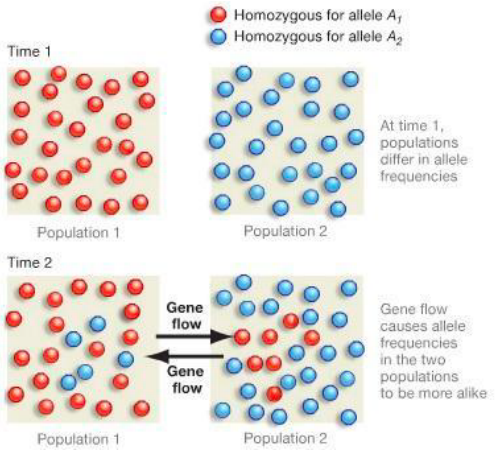
- **Genetic drift**: allele frequencies will change randomly in each population
- **Selection**: different alleles may be favored in different environments
- **Mutation**: different alleles will be generated by mutation in isolated populations

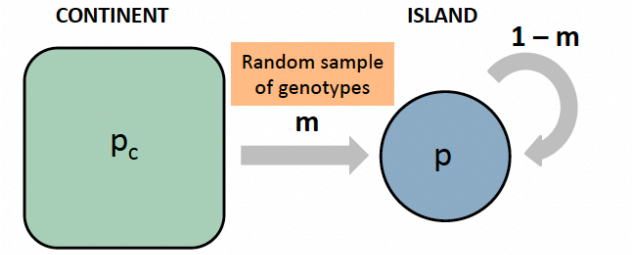
Migration: movement of individuals among populations. It causes gene flow or transfer of genetic material from one population to another. Also it limits the genetic divergence that can occur among subpopulations.

Continent-island model

Population genetics model used to describe gene flow and genetic drift in a 2-population system. In this model, one population is considered to be the mainland/"**continent**" and the other population is an isolated "**island**". The mainland population is usually **larger** and has a **higher** degree of gene flow compared to the island population, which is usually **smaller** and has a **lower** degree of gene flow due to its isolation.

The continent-island model can be used to study how gene flow and genetic drift affect the genetic diversity of populations and how these factors can lead to the divergence of populations over time. It is used to understand the genetic structure and diversity of fragmented populations.





Large population where allele frequency does not change

Smaller population where a proportion of the individuals are replaced by migrants from the large continent population each generation

Random mating in both populations
No genetic drift
No natural selection
No mutation

m = probability that a randomly chosen allele in any subpopulation comes from a migrant

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

ALLELE FREQUENCY IN THE ISLAND

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

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

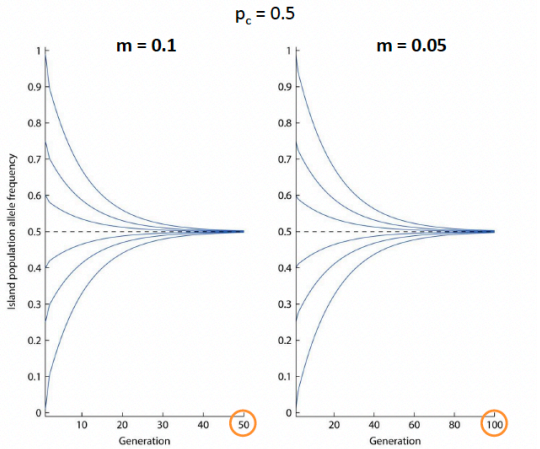
If $p_0 < p_c$ then $p_0 - p_c < 0$ and allele frequency in the island will INCREASE

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

When $t \to \infty$, $p_t \to p_c$

Over time, the island population is increasingly composed of migrants from the continent and allele frequency in the island approaches that on the continent.

The time required depends on the proportion of continental alleles moving to the island each generation, but not on the difference in allele frequencies.



Equilibrium is reached more slowly when there is less migration

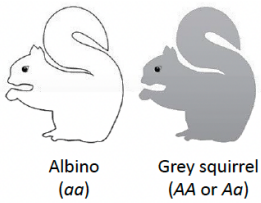
Migration rates (m_{ij})	Donor population (i)				
Recipient population (j)	A	B	C	…	Z
A	m_{AA}	m_{BA}	m_{CA}	…	m_{ZA}
B	m_{AB}	m_{BB}	m_{CB}	…	m_{ZB}
C	m_{AC}	m_{BC}	m_{CC}	…	m_{ZC}
…	…	…	…	…	…
Z	m_{AZ}	m_{BZ}	m_{CZ}	…	m_{ZZ}

$$p_1 = m_{AA}\,p_{0A} + m_{BA}\,p_{0B} + \ldots + m_{ZZ}\,p_{0Z} = \sum m_{ij}\,p_{0i}$$

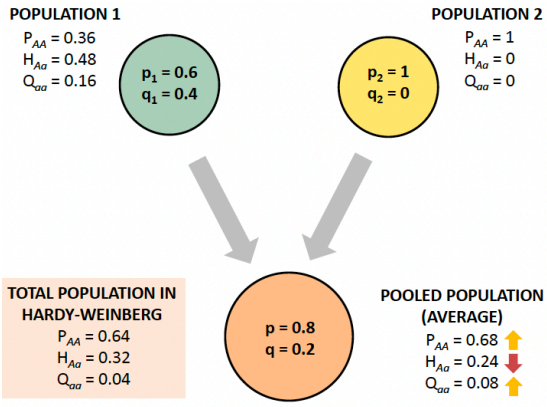
$$p_2 = m_{AA}\,p_{1A} + m_{BA}\,p_{1B} + \ldots + m_{ZZ}\,p_{1Z} = \sum m_{ij}\,p_{1i}$$

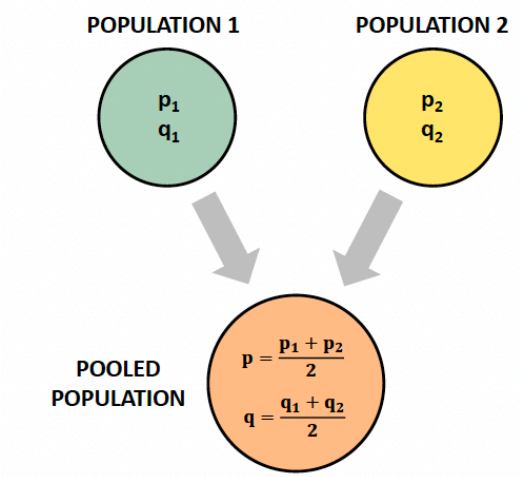
General model

Wahlund effect: an apparent deviation from HWE is observed in a population that is actually in equilibrium. This can occur when 2 or more subpopulations within a larger population have different allele frequencies for a given gene, and these subpopulations are not randomly mixed during mating. As a result, the genotype frequencies in the combined population deviate from the expected values under HWE.



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





The magnitude of the defficit depends on allele freqs:

$p_1 - p_2 = 0 \Rightarrow H_S - H_T = 0$

$p_1 - p_2 = 1 \Rightarrow H_S - H_T = -0.5$

Proportion of heterozygous individuals in the pooled population (assuming an equal contribution from both)

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

Expected proportion of heterozygous individuals if the pooled population was in Hardy-Weinberg equilibrium

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

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

This difference is always ≤ 0

The pooled population will always present a deficit of heterozygous individuals in comparison with the expected frequency under Hardy-Weinberg equilibrium

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

Division in subpopulations will cause a ↓ in the proportion of heterozygotes

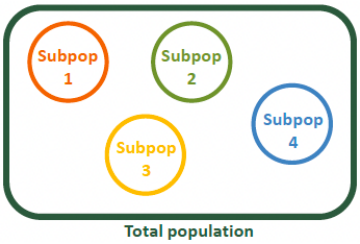
Heterozygosity measurements

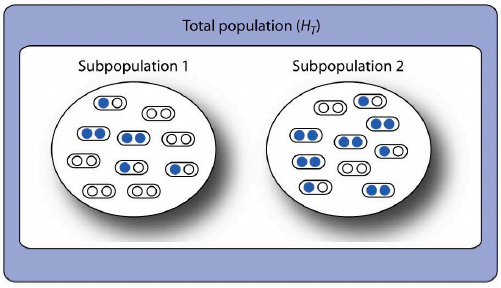
- H_S → average expected heterozygosity assuming random mating <u>within each population</u>
- H_T → expected heterozygosity in the <u>total population</u>

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

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

S = subpopulations
T = total





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

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

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

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

Heterozygosities

$$H_S = \frac{2p_1q_1+2p_2q_2}{2} = \frac{2\cdot0.65\cdot0.35 + 2\cdot0.35\cdot0.65}{2} = 0.455$$

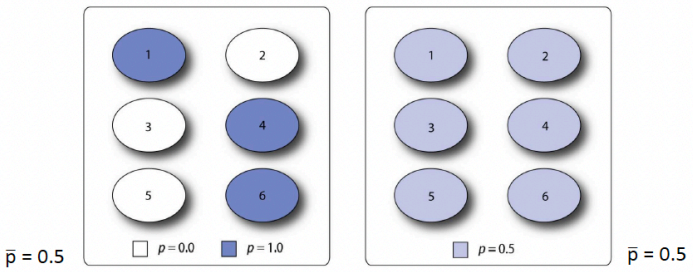
$$H_T = 2p_Tq_T = 2 \cdot 0.5 \cdot 0.5 = 0.5$$

F_{ST} measures differentiation among subpopulations

$H_S = \overline{2pq}$

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

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



$\bar{p} = 0.5$ $\bar{p} = 0.5$

$H_S = \frac{3\,(2 \cdot 0 \cdot 1) + 3\,(2 \cdot 1 \cdot 0)}{6} = 0$ $H_S = \frac{6\,(2 \cdot 0.5 \cdot 0.5)}{6} = 0.5$

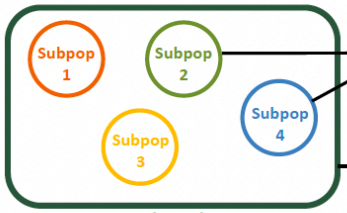
Highly diverged subpopulations

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

No subdivision

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

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



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

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

Calculation	Definition	Cause of deviation
$F_{ST} = \frac{H_T - H_S}{H_T}$	Difference between the expected heterozygosity of the total population and the average expected heterozygosity of subpopulations	Allele frequency divergence among subpopulations

F_{ST} in human populations

- Average level of population differentiation is ↓
- 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 ↑ by genetic drift in finite populations

Without migration:

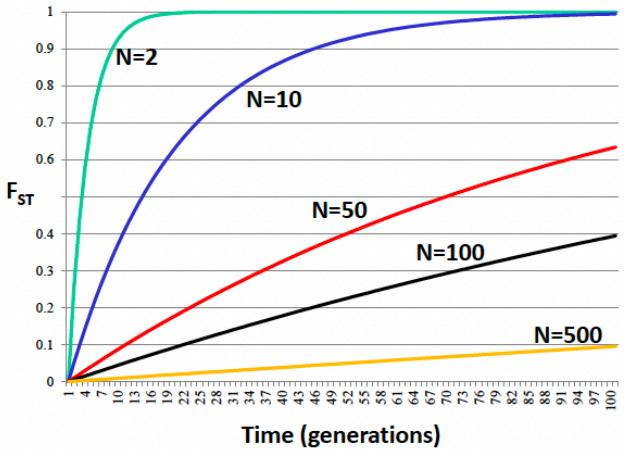
- Allele frequencies change among subpopulations
- F_{ST} increases in each subpopulation
- Differentiation continues until one allele is fixed

Increase of 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

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



When alleles are fixed, F = 1

Migration-genetic drift balance

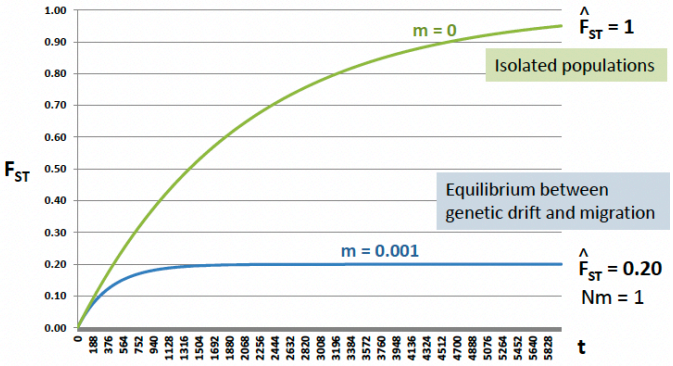
Genetic drift differentiates ⟷ **Migration** homogenizes

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

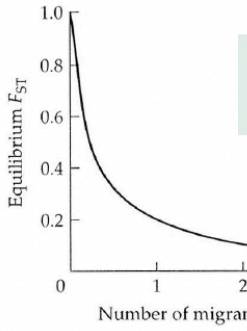
At equilibrium $F_t = F_{t+1} = \hat{F}$

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

Increase in fixation index (F_{ST}) in a population N = 1000



Migration limits strongly the differentiation between populations



$\hat{F}$ decreases rapidly as the number of migrants increases independently of population size

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

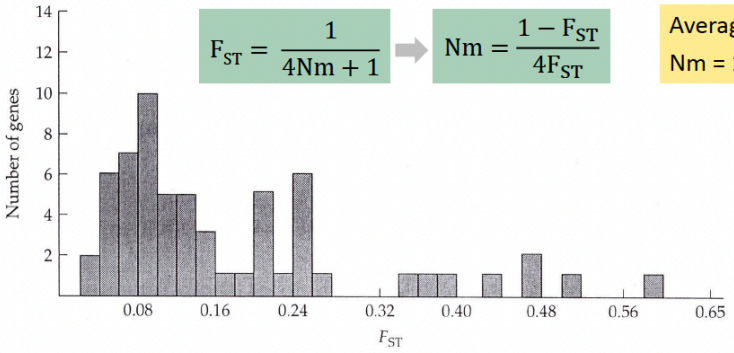
Nm = absolute number of migrants into a subpopulation per generation

Assumptions

- Natural selection is **not** operating and divergence between populations is exclusively caused by **genetic drift**
- Populations have arrived at genetic drift — **migration equilibrium**

Estimation of the migration rate (Nm)

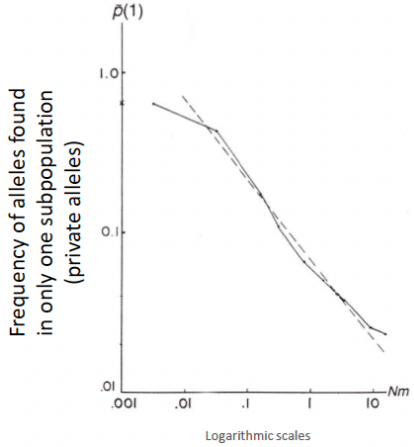
Distribution of F_{ST} values for 61 genes among natural populations of *Drosophila melanogaster*



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

Average F_{ST} = 0.16
Nm = 1.3

Excluding variants with high F_{ST} values
Average F_{ST} = 0.09 Nm = 2.53



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

No migration ⟹ private allele stays in single population
Migration ⟹ private allele will get to other populations

Factors that can influence the level of genetic differentiation and genetic flow **BETWEEN populations**.

↑ Differentiation and ↓ Genetic Flow:

1. **↓ dispersal**: when individuals have limited dispersal abilities or restricted movement, they are less likely to migrate and mix genes with other populations.
2. **↑ distance**: greater geographic distance between populations hinders gene flow because it becomes more difficult for individuals to migrate and reproduce with individuals from distant populations, resulting in increased genetic differentiation.
3. **Strong barrier**: a physical barrier, such as a mountain range or large body of water, can impede the movement of individuals between populations.
4. **Long divergence time**: if populations have been isolated for a long time, they accumulate genetic differences through independent evolutionary processes.
5. **Small population**: in small populations, genetic drift has a greater impact, causing random fluctuations in allele frequencies and leading to genetic differentiation. Additionally, smaller populations have a reduced number of potential migrants, limiting genetic flow.
6. **Adaptive differences**: when populations face different selective pressures in their respective environments, adaptive differences can arise. Selection favors different genetic variants in each population.

↓ Differentiation and ↑ Genetic Flow:

1. **↑ dispersal**: increased dispersal ability allows individuals to migrate more readily, facilitating gene flow between populations. This leads to a greater exchange of genetic material, reducing genetic differentiation (population become more equal between themselves).
2. **↓ distance**: when populations are closer to each other, individuals are more likely to migrate and reproduce with individuals from neighboring populations.
3. **Weak barrier**: a permeable barrier allows individuals to move more freely between populations.
4. **Short divergence time**: if populations have been recently separated or have experienced recent gene flow, genetic differences have had less time to accumulate, resulting in reduced genetic differentiation.
5. **Large population**: in larger populations, genetic drift has a lesser impact, as there is a larger pool of potential migrants. This facilitates gene flow.

* Key concepts

- **Dispersal**: movement of individuals from one location to another. It can be active (e.g., flying, swimming) or passive (e.g., wind dispersal of seeds).
- **Distance**: physical separation between populations or geographic locations. Greater distances between populations ↓ the likelihood of gene flow, as individuals have to traverse longer distances to interact and reproduce with individuals from other populations.
- **Barrier**: any physical or ecological feature that impedes or restricts the movement of individuals between populations. Examples of barriers include mountain ranges, rivers, deserts, and habitat fragmentation.
- **Divergence**: process by which populations become genetically and phenotypically distinct from each other over time. It occurs when populations accumulate genetic changes through various mechanisms like mutation, genetic drift, natural selection, and genetic isolation. Divergence can lead to the formation of new species or subspecies.
- **Divergence time**: length of time that has passed since two populations or species shared a common ancestor. It represents the point in evolutionary history where genetic divergence began. Longer divergence times result in greater genetic differentiation between populations due to the accumulation of genetic changes over time.
- **Population**: group of individuals of the same species that occupy a specific geographic area and have the potential to interbreed.

Session 5: Molecular population genetics

Measuring DNA polymorphism

Measuremnts of genetic variation at DNA sequence level

- **Proportion of segregating sites** → number of segregating sites per nt

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

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

S = 3 L = 40 $p_s = 0.075$

- **Watterson's Θ** → measurement of proportion of segregating sites corrected by sample size

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

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

n = number of sequences in the sample

- **Nt diversity (π)** → average number of nt differences per site between two random DNA sequences in the population

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

δ: Average number of differences between two sequences

L: Length

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

π in different positions of the genome

Nt diversity varies among different organisms, genes, types of functional positions or along chromosomes.

π in different species

- Arthropods are more diverse than chordata with plants in an intermediate situation
- There is ↑ variation within each species group

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 ↑ numbers of small eggs into the environment (**r-strategists**) are much more polymorphic than long-lived species that produce a ↓ number of relatively large offspring and provide parental care (**K-strategists**).

Population size is an important factor in determining the level of π

Large populations are **less** likely to lose variation

Synonymous and non-synonimous positions within coding regions

ATG	TCG	CAC	GGA	GCA
Met	Ser	His	Gly	Ala

Positions

NNN	NNS	NN½	NNS	NNS
ATG	TCG	CAC	GGA	GCA
Met	Ser	His	Gly	Ala

Changes

Non-synonymous Synonymous

ATG	CCG	CAC	GGC	GCA
Met	Pro	His	Gly	Ala

Which positions do you expect to be less variable when we compare sequences?

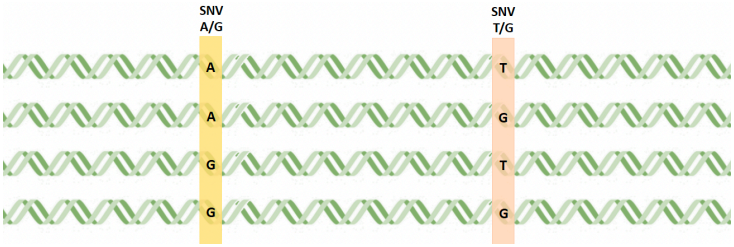
K_a/K_s ratio test

- **K_a**: number of non-synonymous substitutions per non-synonymous position
- **K_s**: number of synonymous subtitutions per synonymous position

$\frac{K_a}{K_s} < 1$	$\frac{K_a}{K_s} = 1$	$\frac{K_a}{K_s} > 1$
Purifying selection	Neutral evolution	Positive selection
More S than NS changes	Same amount of S and NS changes	More NS than S changes
Most genes	Pseudogenes	Exceptional

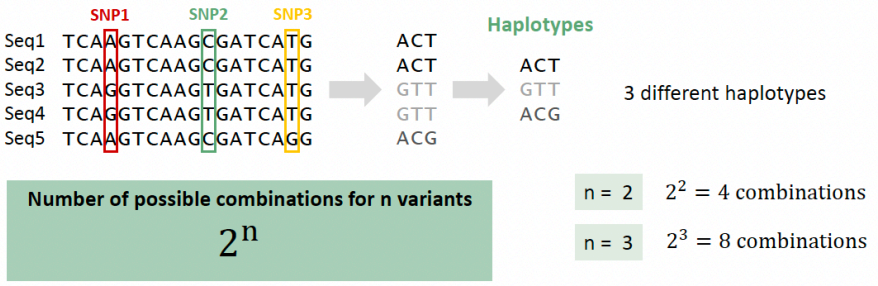
Linkage disequilibrium

Two different variant positions in the same chromosome



Haplotype

Combination of alleles in the same chromosome



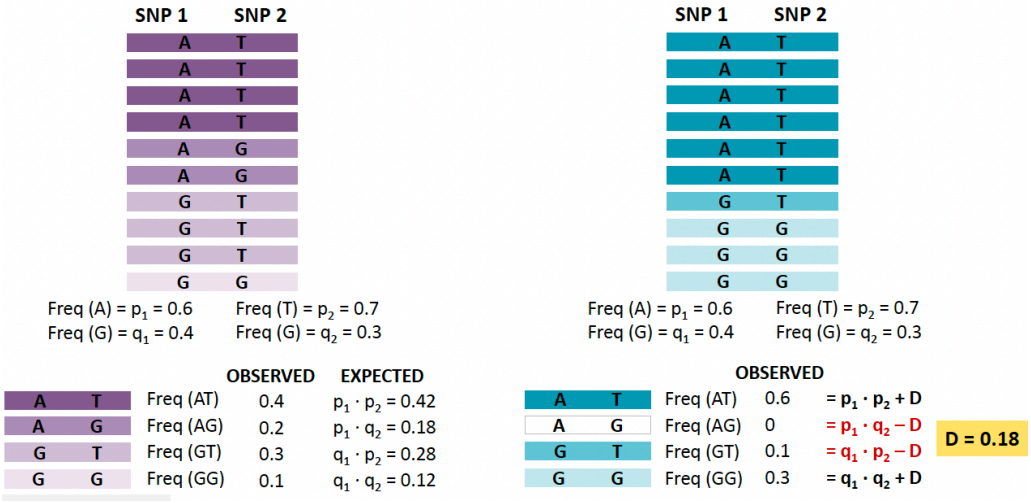
Number of possible combinations for n variants

$$2^n$$

n = 2 $2^2 = 4$ combinations

n = 3 $2^3 = 8$ combinations

Variants in the same chromosome can be linked



Linkage equilibrium/disequilibrium

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 ≠ 0

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

Linkage disequilibrium measurement (**D**)

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	$P_{ab} = p_a p_b + D$

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

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

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.

Linkage disequilbrium relative measurements (**D'**)

D values depends on allele frequencies.
It is a % over total, so **0 < D' < 1**.
To compare them we need to normalize using D_{max} or D_{min}

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

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

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

AA combination does not exist

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

Linkage disequilbrium relative measurements (r^2)

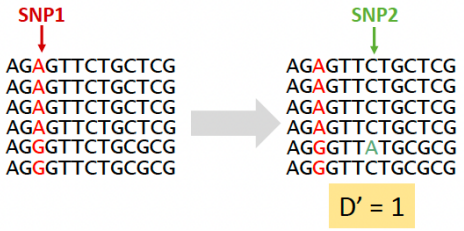
r^2 is the correlation coefficient of the allele frequencies (values from 0 to 1) $r^2 = \frac{D^2}{p_A p_a p_B p_b}$

When $r^2 = 1$ we have perfect LD.
It happens if the two variants have the same allele frequencies. Only 2 different haplotypes exist.

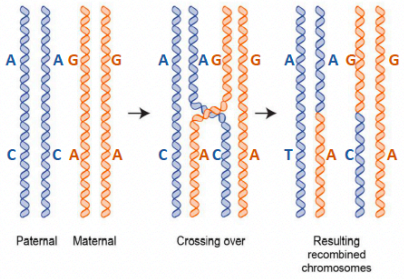
	SNP1 A/G	SNP2 C/A
AGAGTTCTGCTCG	A	C
AGGGTTATGCGCG	G	A

Haplotypes and recombination

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



Recombination generates new combinations of alleles



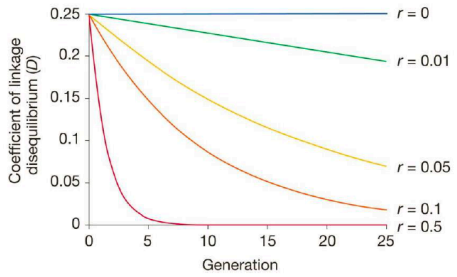
Recombination in heterozygotes breaks the linkage between variants

Linkage disequilibrium decay

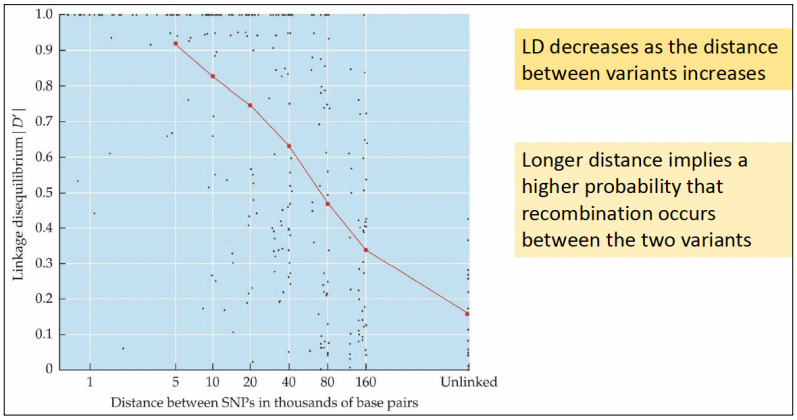
Over time, LD between variants will decrease

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



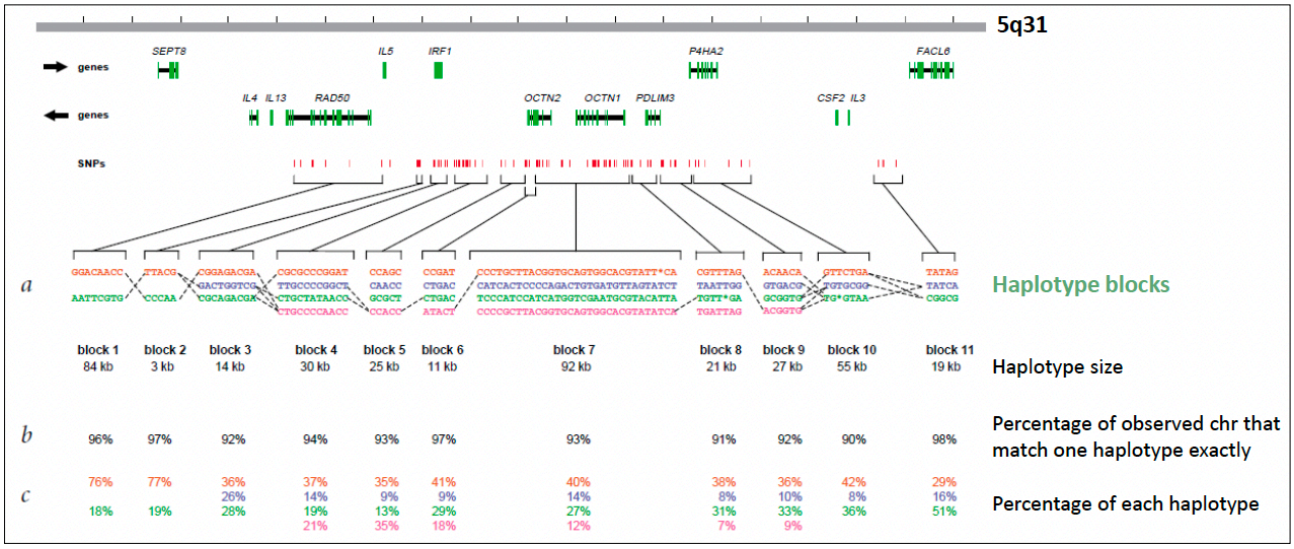
LD decreases as the distance between variants increases

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

Linkage desequilibrium and distance

Haplotype blocks in the human genome

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



Applications of linkage disequilibrium

→ Genomie-wide association studies (GWAS)

- A large number of markers covering all the genome is genotyped in case and control groups that we want to compare.
- 500 000 SNPs in a $3*10^9$ bp genome results in an average spacing between SNPs of 6 kb, enough so that at least some SNPs will be in LD with the casual variant.
- **GWAS** allow us to find variants associated to the phenotype of the case group. The identified variants most likely are not the casual 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 SNV on human chromosome 1 (rs6679677) and type 1 diabetes

	Cases	Controls	Total
AA	57 / 33.6	27 / 50.4	**84**
AC	562 / 433.2	521 / 649.8	**1083**
CC	1381 / 1533.2	2452 / 2299.8	**3833**
Total	**2000**	**3000**	**5000**

Freq A (cases) = 0.169
Freq A (controls) = 0.096

Allele A is a risk factor for diabetes

$$P(AA, case) = P(AA) \cdot P(case) = \frac{84}{5000} \cdot \frac{2000}{5000} \cdot 5000 = 33.6$$

$$P(AG, case) = P(AA) \cdot P(case)$$
$$P(GG, case) = P(GG) \cdot P(case)$$
$$P(AG, control) = P(AA) \cdot P(control)$$
$$P(AG, control) = P(AA) \cdot P(control)$$
$$P(GG, control) = P(GG) \cdot P(control)$$

$\chi^2 = 116.167$
df = 6 -1 - 3 = 2
Threshold ($\alpha=1 \cdot 10^{-7}$) = 32.24
$P = 5.43 \cdot 10^{-26}$

Observed and expected are different. There is an association.
The SNP is closely linked to the variant that is causing the disease.

→ LD and genotype imputation

DATASET 1

LD (r^2 =1)

SNP1	SNP2	SNP3	SNP4
T/T	G/C	C/C	A/A
T/T	G/G	C/A	A/A
A/A	G/C	C/A	G/G
T/A	G/C	C/A	A/G

DATASET 2

LD (r^2 =1)

SNP1	SNP2	SNP3	SNP4
T/T	?	?	?
T/A	?	?	?
A/A	?	?	?

DATASET 2

LD (r^2 =1)

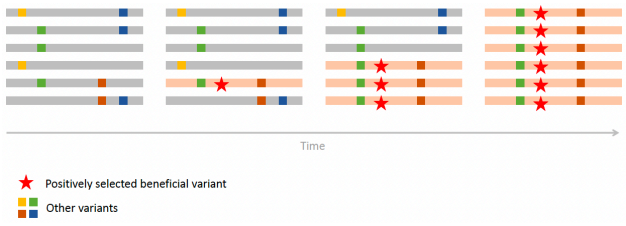
SNP1	SNP2	SNP3	SNP4
T/T	?	?	A/A
T/A	?	?	A/G
A/A	?	?	G/G

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

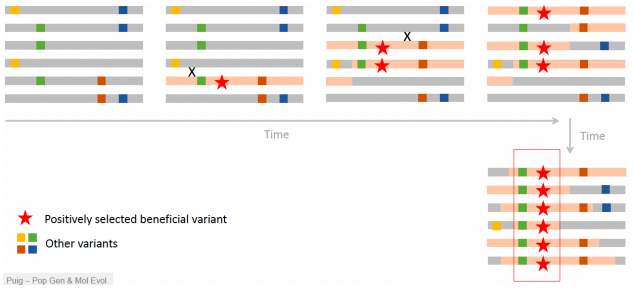
Thanks to LD between variants we can complete the missing genotypes

Selective sweep — no recombination

No recombination

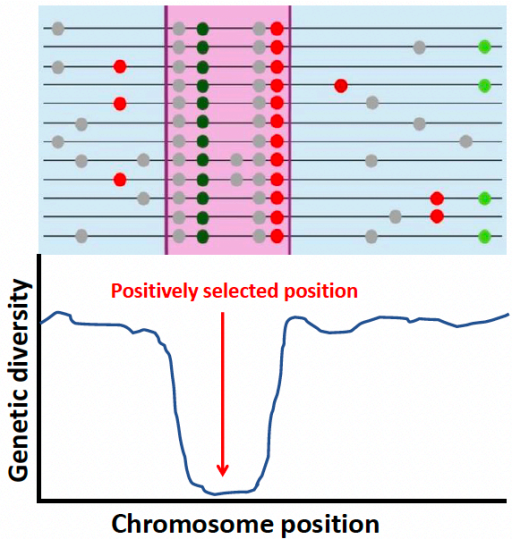


With recombination



Detection of a selective sweep

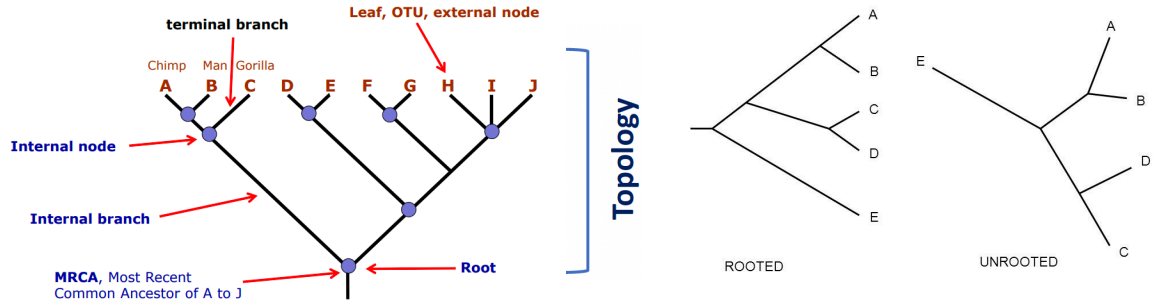
- **Selective sweep**: reduction of measured diversity in the surroundings of a positively selected mutation
  1. A new beneficial mutation appears
  2. It rapidly becomes the most common variant in the population
  3. Nearby positions also become more frequent because they are not physically independent
- **Genetic hitchhiking**: occurs when an allele changes frequency not because itself is under natural selection, but because it is near another allele that is undergoing a selective sweep.



Part B: Molecular Evolution

Looking at common traits allows the possibility of inferring the evolutionary relationships of organisms
→ A **phylogeny**

Main workflow in molecular phylogenetics
Sequencing/finding **orthologous genes** → **Alignment** (finding homology within genes) → **Tree inference**

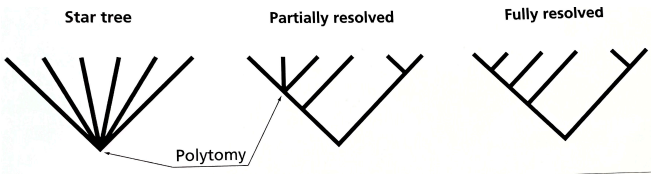


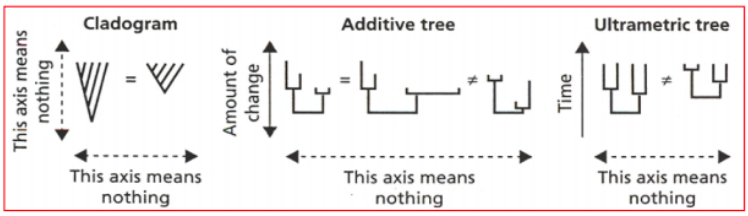
How to fit a tree into a computer → Newick format

Trees may show different degrees of resolution:

- **Soft polytomy**: refers to a branch point in the tree where the exact relationships between the descendants are not fully resolved.
- **Hard polytomy**: refers to a branch point where multiple branches emerge simultaneously, indicating uncertainty or lack of resolution.

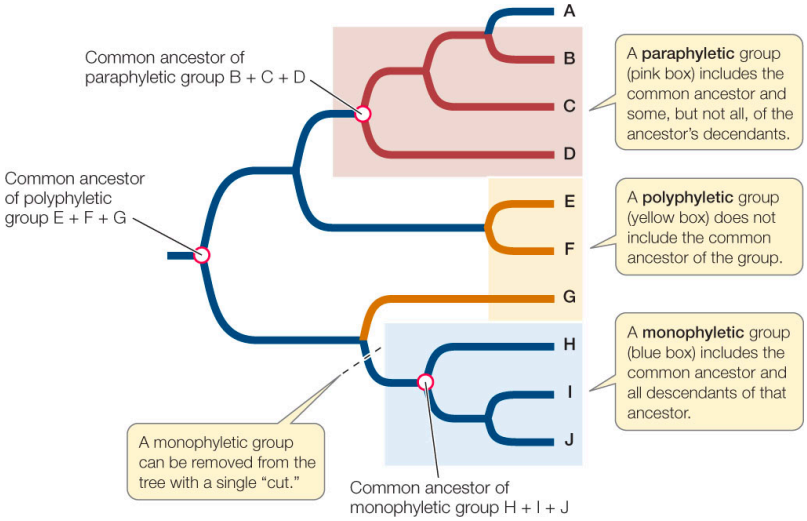
Dichotomy: node that is divided in 2 species
Polytomies: failure to resolve the branching order. Reason why the branch divides into 3 or more species





- **Cladogram**: just common ancestry
- **Phylogram/additive tree**: it has brantch lengths (e.g. amount of change)
- **Ultrametric tree**: tips equidistant from root (Y axis represents time or relative time)

Types of clades: monophyletic, polyphyletic and paraphyletic



Paraphyletic: group of organisms that includes an ancestral species and some, but **not all**, of its descendants.

Polyphyletic: group of organisms that includes multiple evolutionary lineages that do **not** share a common ancestor.

Monophyletic/clade: group of organisms that includes an ancestral species and all of its descendants, but no other organisms.

Orthologs: genes that diverge through <u>speciation</u> (their MRCA is a speciation event)
Paralogs: genes that diverge through <u>duplication</u> (their MRCA is a duplication event)

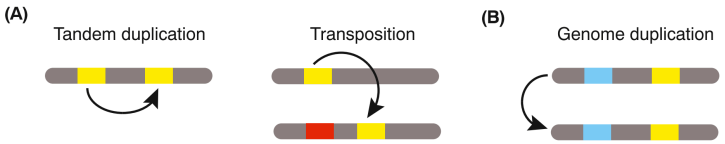
MRCA: Most Recent Common Ancestor

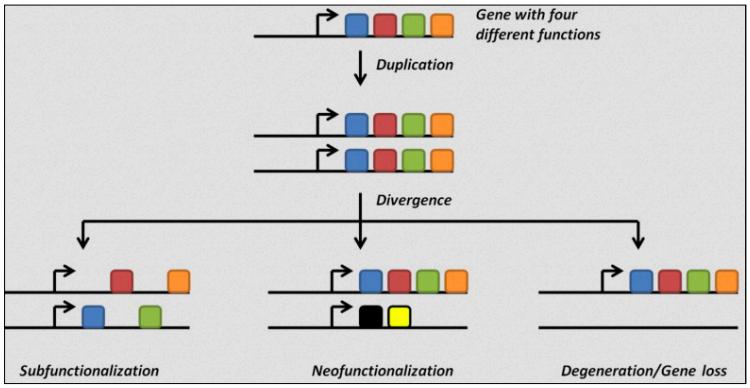
Methods to infer orthology:

- **Functional characterization**: involves examining the biological functions and properties of genes. If two genes from different species perform similar functions or participate in the same biological pathways, it suggests that they are orthologous. (e.g. functional assays, GEA and phenotypic studies)
- **Sequence similarity**: involves comparing nt or Aa sequences of genes from different species. Orthologous genes tend to have ↑ sequence similarity, indicating their shared evolutionary origin.

- **Expert curation**: involves manual curation and annotation of genes by domain experts. These experts analyze data sources, including experimental evidence, literature and DBs to identify orthologous relationships. (e.g. functional annotations, gene structure, evolutionary relationships, and experimental validation to determine orthology).
- **Synteny**: conservation of gene order and arrangement in different species. Orthologous genes often maintain their relative positions in the genome, even after speciation events. By comparing gene order across species, orthologous genes can be inferred. Particularly for closely related species.
- **Phylogeny**: involves constructing evolutionary trees based on sequence data. By comparing the phylogenetic relationships of genes across different species, orthologous relationships can be determined. Orthologs are expected to cluster together in the phylogenetic tree, forming monophyletic groups. (e.g. ML, BI).

Gene duplication: a gene is copied, resulting in two or more identical or similar gene sequences
- Gene duplicaties account for 8-20% of the genes in eukaryotic genomes
- Rates of gene duplication are estimated at between 0.2-2% per gene per million years





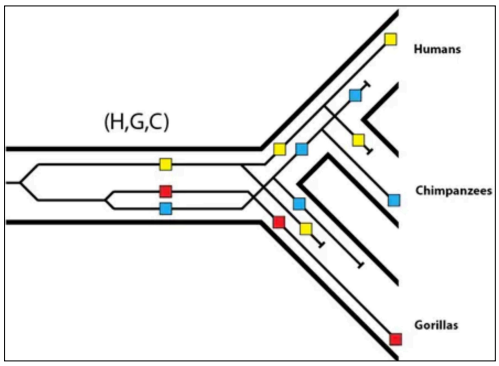
Neofunctionalization: where one copy of the gene retains its original function while the other copy evolves a new function.
Subfunctionalization: the two copies of the gene divide the functions of the original gene between them.
Degeneration: loss or inactivation of one copy of a duplicated gene, resulting in a reduction in gene dosage or complete loss of the function of that gene.

Do always orthologous genes recover species tree? No
Reasons that explain gene tree/species tree **discordance**:
- **Incomplete lineage sorting**: alleles inherited from a common ancestor may not be completely sorted out within the descendant species. As a result, different alleles can be found in different species, leading to incongruence between the genealogy of alleles and the species tree.

  Gene trees depend on how alleles are sorted among lineages
- **Molecular convergence**: different genes may independently evolve similar sequences or traits due to selective pressures, leading to discordance between gene trees and the species tree. (e.g. wings in bats and birds, which have different genetic and anatomical origins, is an example of convergent evolution at the molecular level).

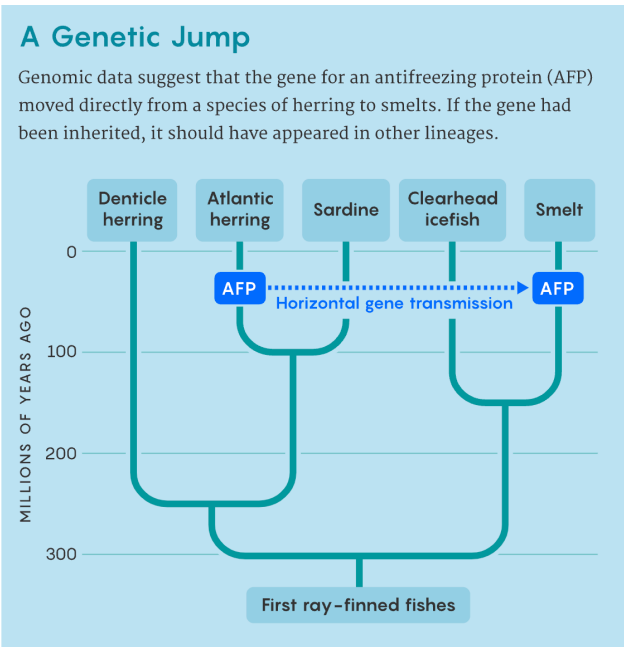
  Natural selection can shape genes in similar ways
- **Hybridization**: occurs when individuals from different species or distinct genetic lineages interbreed and produce offspring with mixed genetic characteristics. It can result in the exchange of genetic material between species, leading to the formation of hybrid genomes. Hybrids may exhibit a combination of traits from both parent species or display unique characteristics.

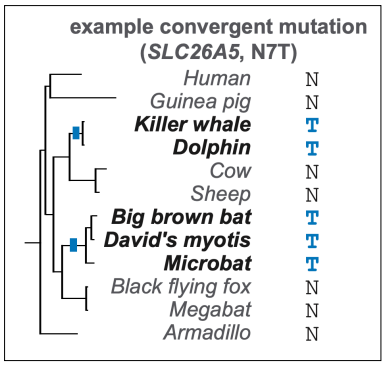
  Genes can move through the boundaries of closely related species
- **Horizontal Gene Transfer**: refers to the transfer of genetic material between different organisms that are not parent and offspring. Unlike vertical gene transfer (inheritance from parent to offspring), HGT involves the movement of genetic material across species boundaries.
- **Phylogenetic errors**: biases in phylogenetic analysis (e.g. incorrect models, inadequate data, inappropriate methods) can lead to inaccuracies in gene tree estimation and discordance with the species tree.



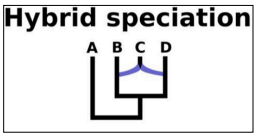
Incomplete lineage sorting

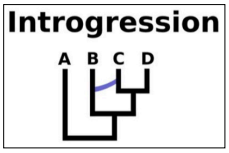
Horizontal Gene Transfer





Molecular convergence

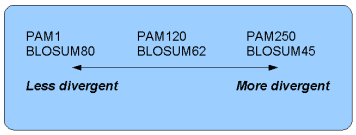




Hybridization

Alignment

- All sequences have the same length
- For each column, at least one of the positions is not a gap
- They typically have matches, mismatches and gaps (indels)



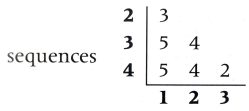
PAM (Point Accepted Mutation) and **BLOSUM** (BLOcks SUbstitution Matrix): scoring matrices used in BI to quantify Aa substitution probabilities in sequence alignments for protein sequence analysis. PAM matrices are based on evolutionary models, while BLOSUM matrices are derived from observed substitutions.

Gene tree inference

Methods to reconstruct the phylogenetic tree:

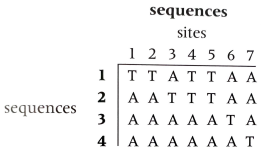
→ Distance based

- UPGMA (Unweighted Pair Group Method with Arithmetic Mean)
- Negihbor-Joining



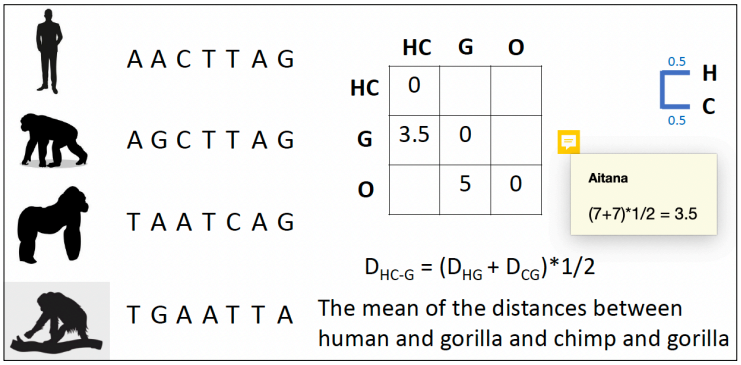
→ Character based

- Maximum parsimony (minimize number of changes)
- Maximum likelihood
- Bayesian inference



UPGMA: assumes that the substitution (evolution) rate is constant across lineages. It clusters the species that have the lowest value in the UPGMA matrix and then the matrix is reconstructed.

- If sequences i and j have 100 nts and there is no gaps, the MSA has 100 positions (sites or columns)
- 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 NJ approach



NJ: does not assume that the substitution rate is constant across lineages. Produces unrooted trees

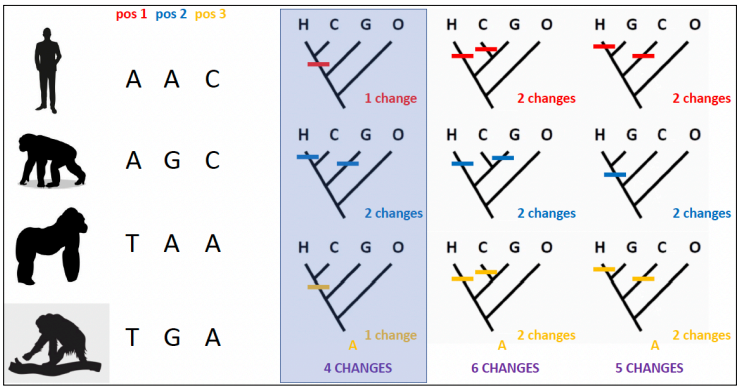
- Finds the shortest (minimum evolution) tree by finding neighbors that minimize the total length of the tree. Shortest pairs are chosen to be neighbors and then joined in distance matrix as one OTU

Maximum parsimony

The phylogeny that requires the smallest number of character changes is most likely to be correct.

Problem: the number of potential trees grows exponentially as 'n' ↑.

Therefore, we need a search algorithm (heuristic method). But, this does not guarantee that we find the best tree.



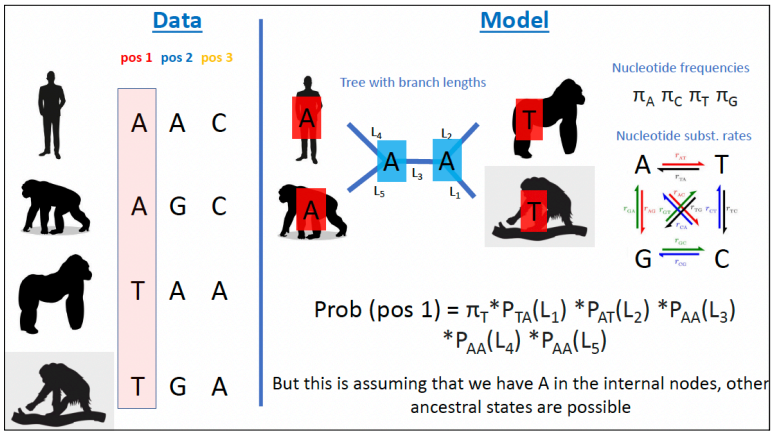
Maximum likelihood

The likelihood of a model is the probability of observing the data given the model.

likelihood = P(data|model)

The most likely model is the one that maximizes the probability of observing the data.

→ The most likely **tree** is the one that maximizes the probability of observing the **alignment**.



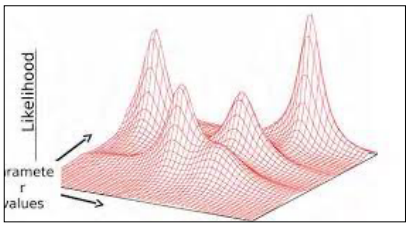
Calculation of the likelihood of the tree (model) for pos 1

Calculation of the probability of the alignment give the model

Assuming that all sites are independent, we can multiply the probability of each site to get the probability of the alignment given the model:

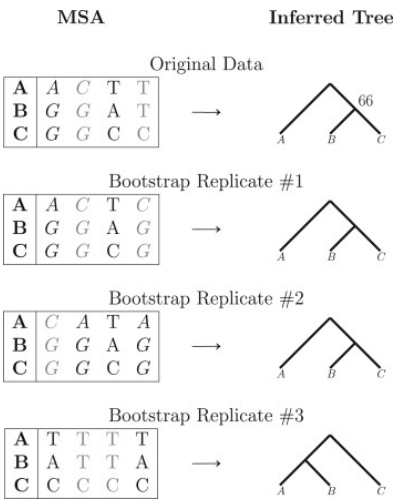
Prob (alignment) = Prob(pos1) * Prob(pos2) * Prob(pos3)

Likelihood of the model = Prob(pos1) * Prob(pos2) * Prob(pos3)



Hill climbing algorithm: optimization algorithm that iteratively improves a solution by making incremental adjustments in a step-by-step manner, always moving towards a better or higher-valued solution in the search space.

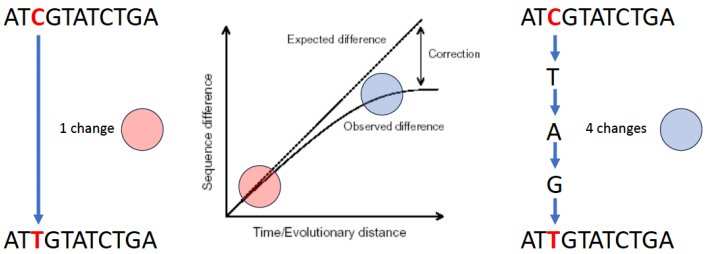
Bootstrap: statistical resampling technique used to assess the **robustness** and **reliability** of estimates or hypotheses. Involves creating multiple bootstrap samples by randomly sampling with replacement from the original data, and then analyzing each sample to generate a distribution of statistics or to test hypotheses.



Applications of phylogenetics

- **Reconstruct the Tree of Life**: phylogenetics helps to understand the evolutionary relationships among organisms and construct the Tree of Life, revealing patterns of diversification and evolutionary history.
- **Biodiversity discovery:** phylogenetic analysis aids in identifying and classifying new species, exploring species relationships, and understanding patterns of biodiversity across ecosystems.
- **Epidemiology**: phylogenetics assists in tracing the transmission and spread of infectious diseases, identifying sources of outbreaks, and understanding patterns of disease evolution. (e.g. COVID-19)
- **Cancer research**: phylogenetic analysis of tumor samples helps in studying the evolution and progression of cancer, identifying driver mutations, and guiding personalized treatment strategies.
- **Human history**: phylogenetics provides insights into human evolution, migration patterns, and population genetics, helping to understand our evolutionary origins and genetic diversity.
- **Forensics**: phylogenetic techniques can be used in forensic investigations to analyze DNA sequences and determine genetic relatedness, aiding in identifying individuals and establishing familial relationships.

Models of evolution



How sure we are that the changes we count are the changes that really took place?
→ To correct, we need to assume a **substitution model**

Stochastic process (mutation) → evolution is modeled as a random process where genetic mutations occur, leading to genetic variation and subsequent evolutionary change.
The probability of future states of the process depends **ONLY** upon the present state (**memoryless**)
We need a model to describe this process → **Markov process**

Components of a Markov process

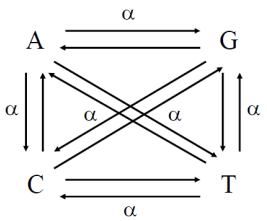
1. State space (A, T, C or G)
2. Initial state (an initial distribution of probabilities across the space state) if A (1, 0, 0, 0)
3. Transition probabilities that capture the probability of going from one state to another

With a Markov process we can simulate DNA evolution, and compute the probability of any nt any time

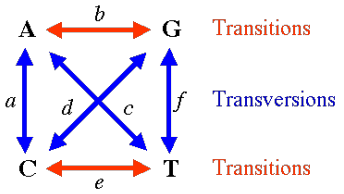
A → A → T → T → T → C → C → C → C → C

What is the probability of observing A, T, C or G at any time in the future?
We can just simulate this process for a large number of steps and estimate the frequency at which we see each of the nts, but there is another more elegant way → Quick R simulation

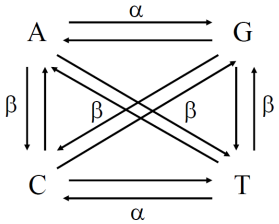
Jukes and Cantor one parameter model assumes that each nt has equal probability to be substituted by any of the other 3. It predicts equal gene frequencies (¼).



How realistic is Jukes and Cantor?
DNA structure predicts that (in general) **transitions** will be more probable than transversions.



Kimura two-parameter (**K2P**) model has two parameters α (for transitions) and β (transversions). Usually α > β.



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.

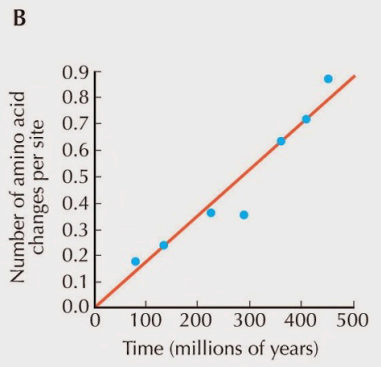
Accomodating rate variation among sites
→ Rates of evolution might be highly variable within on gene (Intra-gene rate variation: example w/ 16S)
- We can acommodate this variation with an additional parameter: γ (for example GRT + γ)

Nucleotide substitution models — Summary
- **Jukes and Cantor Model** (JC69): assumes equal base frequencies and a single rate for all possible substitutions. It is a simple and symmetric model. The key assumption is that the probability of a nt transitioning to any of the other three nts is the same. However, this model is unrealistic for actual DNA sequences due to its oversimplified assumptions.
- **Kimura Two-Parameter Model** (K80): improves upon the JC69 model by incorporating unequal base frequencies and different transition/transversion rates. It assumes that **transitions** (purine to purine or pyrimidine to pyrimidine) and **transversions** (purine to pyrimidine or vice versa) occur at different rates. This model provides a better approximation of nt substitution patterns.
- **General Time Reversible Model** (GTR): more complex and flexible compared to the previous models. It allows for different substitution rates between nts and also considers the variation in base frequencies. The GTR model estimates the rates of substitution for different nt pairs and takes into account the overall composition of the sequence. It is widely used in phylogenetic analysis due to its flexibility.
- **Hasegawa-Kishino-Yano Model** (HKY): variation of the GTR model that assumes equal base frequencies but allows for different transition/transversion rates. It is a popular model for DNA sequence analysis, particularly estimating evolutionary distances and constructing phylogenetic trees. The HKY model strikes a balance between complexity and computational efficiency.

These nt substitution models are used to estimate evolutionary distances between sequences, infer phylogenetic relationships, and study evolutionary processes. More advanced and sophisticated models have been developed over time, such as the GTR+Gamma and GTR+Invariant models, which consider additional parameters to account for rate variation across sites.

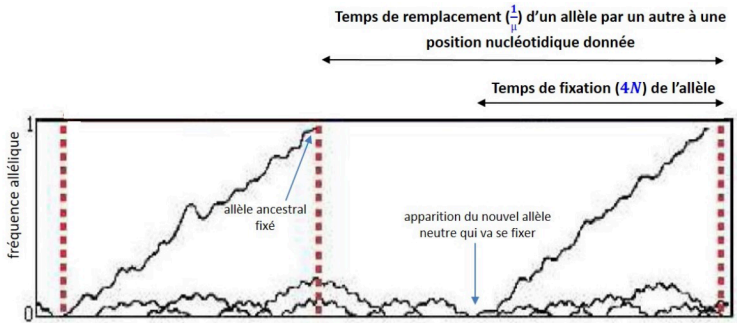
Molecular clock: concept that assumes a relatively constant rate of nt substitution over time, providing a method to estimate evolutionary divergence and date common ancestors.

The number of substitutions between protein sequences of different species are roughly proportional to the time since species divergence.



Calibrating the molecular clock involves using known reference points, such as fossil records or historical events, to estimate the rate of molecular evolution and establish meaningful time scales.

Neutral theory of molecular evolution
For a given gene in a population, mutations take place at **2Nμ**
→ (2N = number of copies of the gene in a diploid population, μ = mutation rate of the gene)
Under drift, each new mutation has a probability of fixation = its frequency (p_0) = **1/2N**
The rate of fixation (replacement) of neutral alleles = **2Nμ * 1/2N = μ**



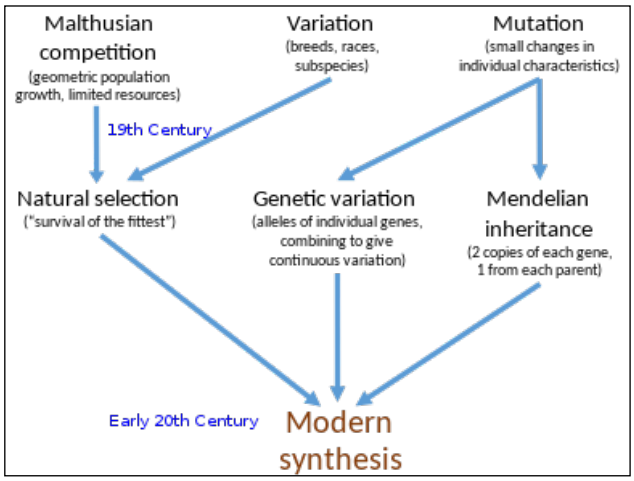
From this, a constant substitution rate is **expected** as observed in the molecular clock

Assumptions:
1. Effective population size (N_e) is constant
2. Homogeneous population density
3. New mutations have no differences in fitness → **Neutral**

From this, we **get** a constant substitution rate as observed in the molecular clock

Modern/neo-Darwinian synthesis: integration of Mendelian genetics with Darwinian evolution. It combines natural selection, genetic variation, and heredity to explain the process of evolution and the diversity of life on Earth.

Panselectionism: view that all traits, behaviors, or characteristics observed in organisms are a result of natural selection. It suggests that every aspect of an organism has been shaped by adaptive forces.



Natural selection acting on genetic variation is the main mechanism of evolutionary change
→ All variation has a purpose

If natural selection is the main force shaping phenotypes, why is there so much variation in natural populations? While natural selection is a powerful force, other factors like genetic drift, mutation, gene flow, and environmental variation can introduce and maintain genetic diversity within populations. Additionally, not all variation is necessarily subject to strong selection, and some traits may be influenced by complex interactions or trade-offs.

Mapping variation refers to the process of identifying and characterizing genetic or phenotypic differences across individuals or populations within a species.
Molecular variation in wild populations was ↑ than expected, can this variation be explained by natural selection? High molecular variation observed in wild populations can be influenced by various factors, including natural selection. Natural selection can act on genetic variation, favoring certain traits or alleles that provide a selective advantage in specific environments. However, other factors such as genetic drift, mutation, and gene flow can also contribute to observed variation in wild populations.

Cost of selection
If this variation was maintained by natural selection, the cost would be huge

Alleles	Frequency	Fitness
A	p	1
a	q	1-s

Pr(deads) ~ qs
Pr(survivors) ~ 1 - qs

Cost per generation ~ qs/1 - qs

Rate variation across genes can indicate the presence of natural selection. Genes experiencing **positive selection** may exhibit accelerated evolutionary rates, while **negative selection** can lead to reduced rates of evolution in certain genes.

Tests of neutrality
→ **dN/dS ratio** (ω/nonsynonymous-to-synonymous substitution rate ratio): used to evaluate selective pressures on protein-coding genes. It compares the rate of nonsynonymous (Aa-changing) substitutions (dN) to the rate of synonymous (silent) substitutions (dS). A ratio > 1 suggests positive or diversifying selection, while a ratio < 1 indicates negative or purifying selection.

We can take advantage to fact that (in protein coding genes) nt substiutions can be:
- **Non-synonymous**: nt substitutions that change the Aa → Advantageous, deleterious, neutral
- **Synonymous**: nt substitutions that do not change the Aa → Neutral

The ratio between synonymous and non-synonymous differences (dN/dS) **estimates how much** "non-neutral" (**non-synonymous**) changes occurred relative to "neutral" (**synonymous**) changes.

- dN/dS = 1 → neutral evolution (no selection on non-synonymous sites)
- dN/dS < 1 → purifying selection (non-synonymous sites are selected against)
- dN/dS > 1 → positive selection (non-synonymous favoured by natural selection)

→ **McDonald-Kreitmann test**: compares the ratio of nonsynonymous to synonymous substitutions within a species (**within-species**) to the ratio of fixed nonsynonymous to synonymous substitutions between species (**between-species**). Significant differences between the two ratios can indicate positive selection or other departures from neutrality, suggesting that natural selection has influenced the evolution of the gene.

- **Adaptative** alleles are expected to be polymorphic for a very short period of time.
- **Neutral** alleles are expected to be polymorphic for longer periods.

- **Within species**: most of the polymorphism will be neutral (it is very unlikely to sample adaptive alleles in a population, when they are still polymorphic)
- **Between species**: we can see advantageous mutations as they accumulate when we compare sequences between species

Under neutrality ($D_n/D_s = P_n/P_s$), the ratio of nonsynonymous to synonymous variation **within** a species (**P**) is equal the ratio of nonsynonymous to synonymous variation **between** species (**D**). Deviations from this imply some **kind of selection**.

→ **Tajima's test**: attempts to determine if a particular set of nt sequences are or are not compatible with the H_0 (neutral model). Examines the distribution of genetic variants **within** a population. It compares the # of segregating sites with the average number of pairwise differences. Deviations from the neutral expectation can indicate departures from neutrality, suggesting the influence of natural selection, population expansions, or other evolutionary processes.

- **Π** = average pairwise differences
- **S** = variable sites

Under a mutation-drift equilibrium:

the expected value of π is S/a_n where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$

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

Tajima's D:

Observed π - Expected π

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.