Queen Mary
University of London

EMBO PopGen

Day2: introduction to population genetics

Matteo Fumagalli

Intended Learning Outcomes

In this session you will learn

- to describe all different types of genetic data
- to demonstrate the relationship between allele and genotype frequencies
- to calculate Hardy-Weinberg Equilibrium proportions

Terminology

Before we dive into the genetic basis of evolution, let's define some important terms, such as:

- Gene
- Phenotype
- Locus
- Allele
- Genotype
- Haplotype

time: 9.30am

Gene

A **gene** can be defined as:

- the segregating and heritable determinant of the phenotype*;
- the fundamental physical and functional unit of heredity, which carries information from one generation to the next one;
- a segment of DNA, composed of a transcribed region and regulatory sequences that make possible transcription and regulation.

* a trait or characteristic of the individual carrier (more on this later)

Gene

For instance, the human *LCT* gene "provides instructions for making an enzyme called lactase. This enzyme helps to digest lactose, a sugar found in milk and other dairy products*".

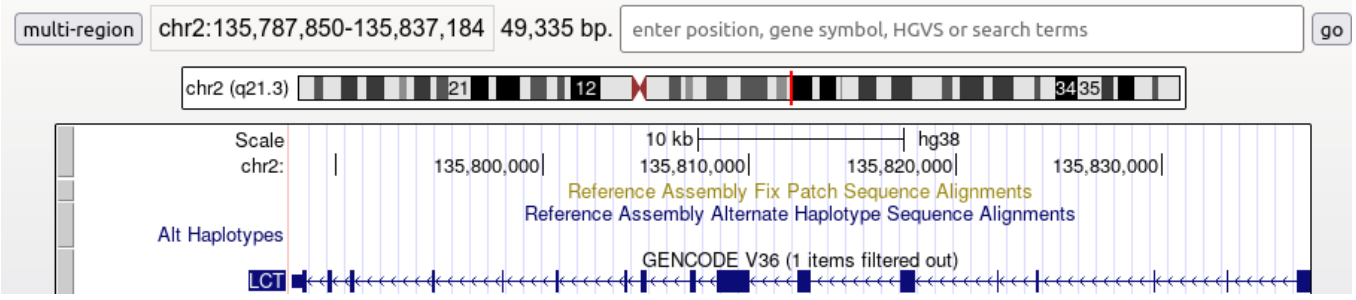


Figure 1: *LCT* gene is located on chromosome 2 in the human genome and spans approx. 50k base pairs.

*`https://pubchem.ncbi.nlm.nih.gov/gene/LCT/human`

Phenotype

A **phenotype** can be defined as a physical/behavioural (etc.) characteristic of an individual.

The genetic component of the phenotype is heritable.

Phenotype

For instance, "lactase is the enzyme that carries out the digestion of the milk sugar lactose. Its expression decreases at some point after the weaning period is over in most mammals and in around 68% of all living adult humans. However, in some humans, particularly those from populations with a history of dairying, lactase is expressed throughout adulthood. This **phenotype** is called lactase persistence"*.

* https://pubmed.ncbi.nlm.nih.gov/24861860/

Lactase persistence

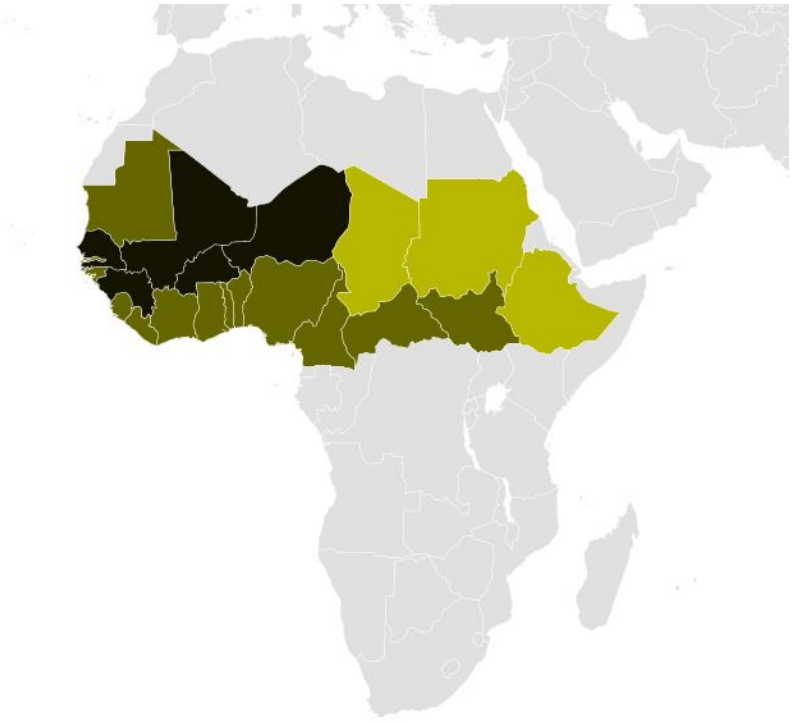


Figure 2: A geographical distribution map of Fula people who show high prevalence of lactase persistence.

https://www.sciencedirect.com/science/article/pii/S0002929714000676

Types of genetic data

The study of the genetic basis of evolution is applicable to all genetic *variants* that can be distinguished by some means and that can be *transmitted* from parents to offspring.

Any variants with these properties are called **alleles**:

- single nucleotide polymorphism,
- insertion/deletion,
- microsatellites.

Single nucleotide polymorphism (SNP)

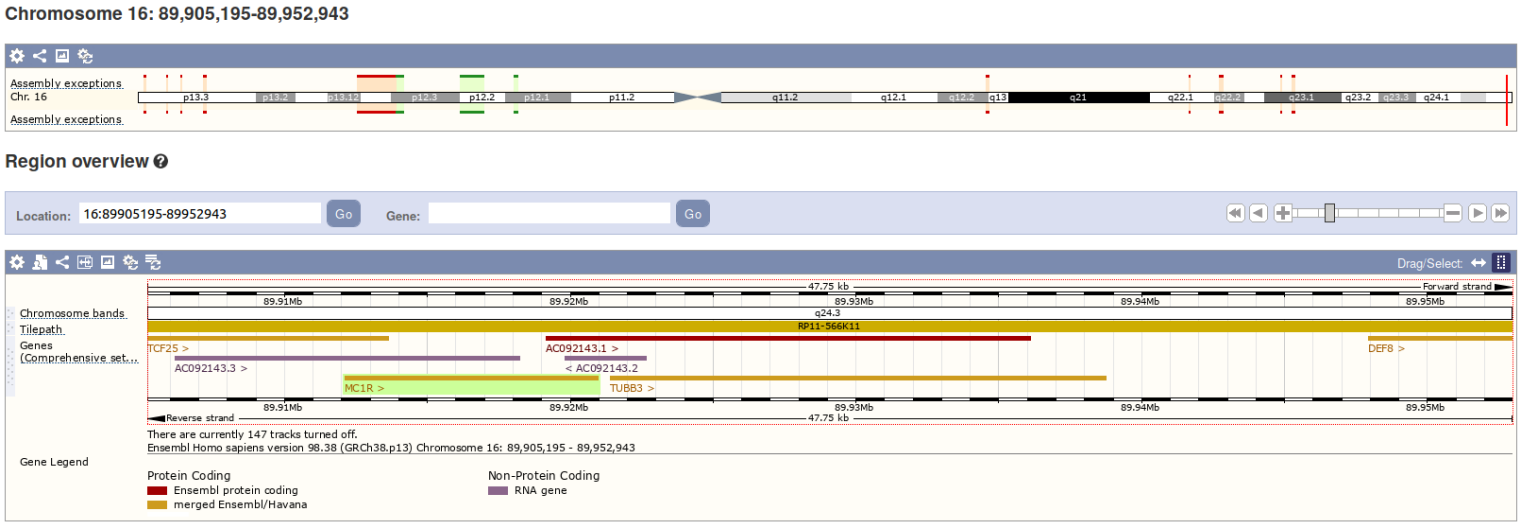


Figure 3: *MC1R* human gene

The C/T variation at position 478 in *MC1R* is an example of a **single nucleotide polymorphism** (SNP, "snip").

Single nucleotide polymorphism (SNP)

MC1R codes for a protein called melanocortin 1 receptor.

Figure 4: Julianne Moore

Individuals with two copies of `T` allele in position 478 of *MC1R* gene tend to have freckles and red hair.

This mutation disrupts the protein and causes an increase of the production of red/yellow pigment melanin instead of brown/black.

Please be aware that most phenotypes are not fully determined by the presence of a single genotype only.

Insertion / deletion (indel)

An **indel** is the insertion or deletion of few nucleotides.

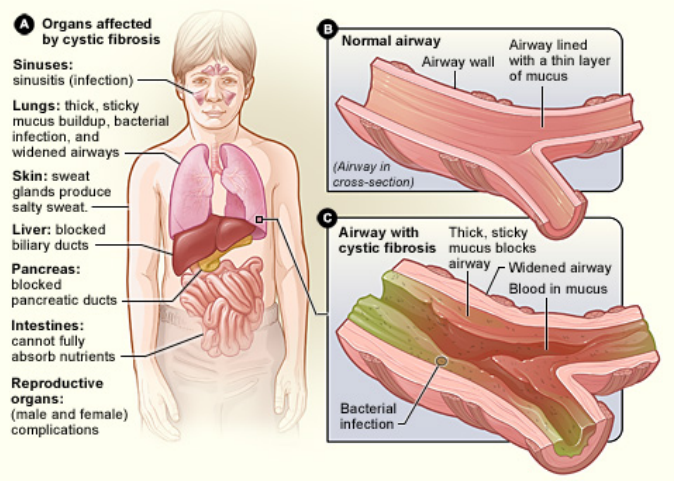


Figure 5: Cystic fibrosis

CFTR gene codes for a transmembrane protein involved in osmotic balance of cells.

Variant ΔF508 has a three-base deletion that results in the absence of the 508th amino acid phenylanine (F).

Microsatellites

DNA replication machinery tends to miscopy repeated sequences in the genome.

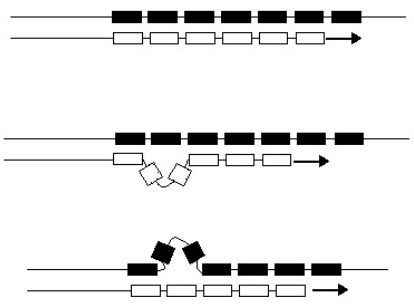


Figure 6: DNA replication errors

e.g. sequence `AGCTGCACACACACACACATGCTG` has `CA` motif repeated seven times, while other individuals may have a different number of copies, thus $(\texttt{CA})_n$.

Simple sequence repeats (SSRs) or **microsatellites** are variants on the number of repeats transmitted during meiosis, with a small possibility of error.

Terminology

Let's introduce some additional concepts that deal with how genetic variation can be summarised. These are:

- **allele**: a distinguishable and heritable quantity (SNP, indel, microsat);
- **locus**: any position (or unit) in the genome with one or more alleles;
- **genotype**: combination of alleles carried by an individual in a particular locus.

Example

An individual has A and G alleles, and therefore has `A/G` genotype, at locus in position 8,789,654 of chromosome 1.

Locus

A **locus** (pl. *loci*) can be defined as a generic position on a chromosome, or the position on a chromosome of a gene or other chromosome marker. Generally speaking, a locus is a location.

Locus

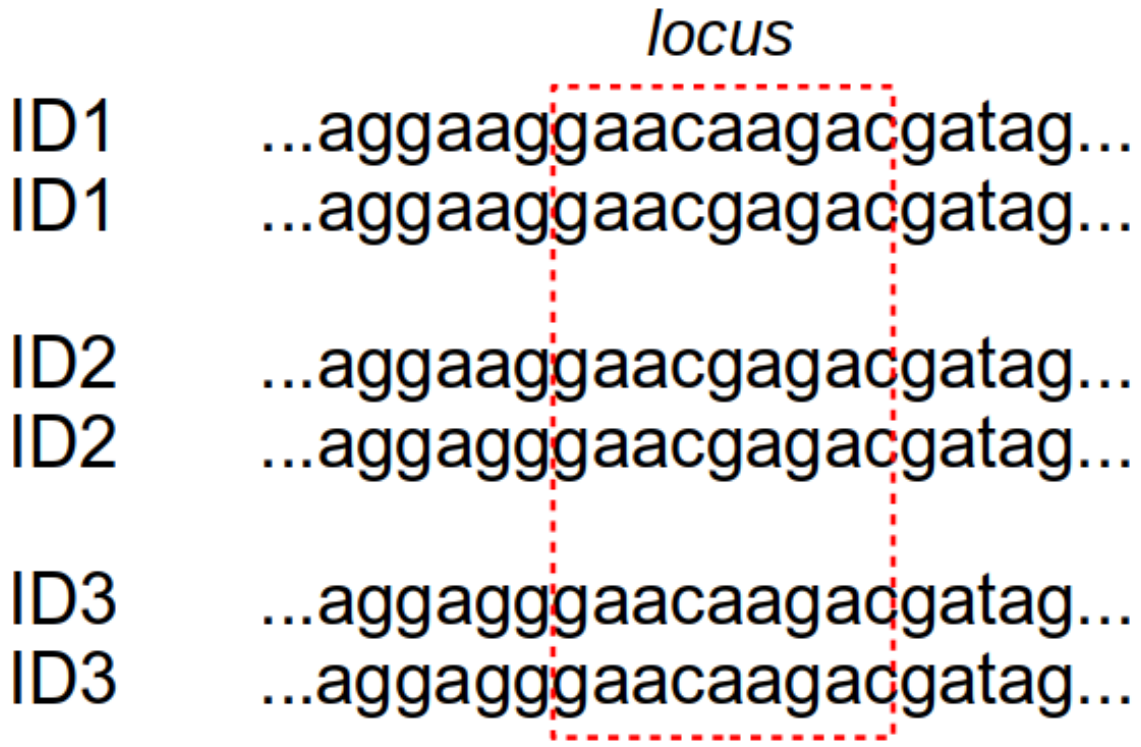


Figure 7: A *locus* of nine base pairs (bp).

Allele

An **allele** is a variant of a gene or locus. Different alleles can lead to different phenotypes.

As we will see later, diploids have two copies of each gene. Therefore, we define homozygote an individual that possesses two copies of the same allele, while heterozygote if it possesses two different alleles.

Allele

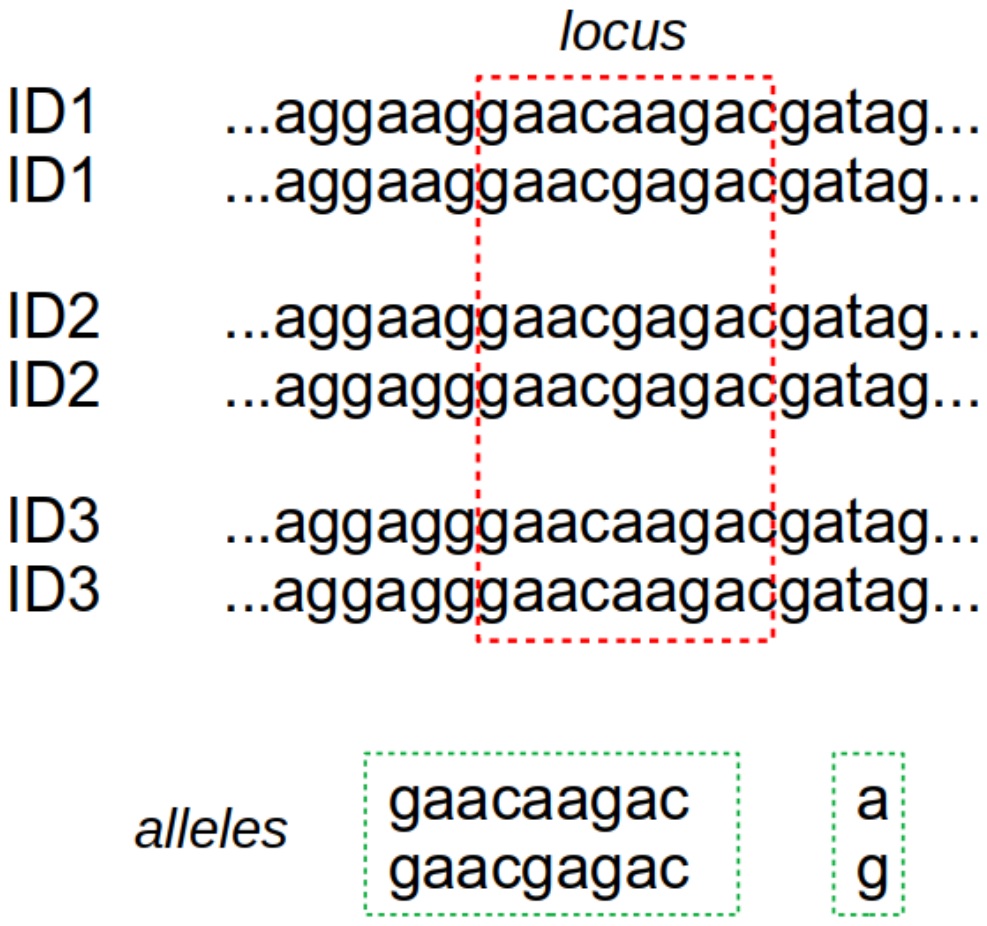


Figure 8: A *locus* of nine base pairs (bp) and two alleles.

Genotype

A **genotype** can be defined as the genetic makeup of an individual (at one or more loci).

It can also be considered as a description of the alleles possessed by an individual.

Genotype

Figure 9: A *locus* of nine base pairs (bp), two alleles and three genotypes.

Haplotype

A **haplotype*** is the series of alleles along the same chromosome.

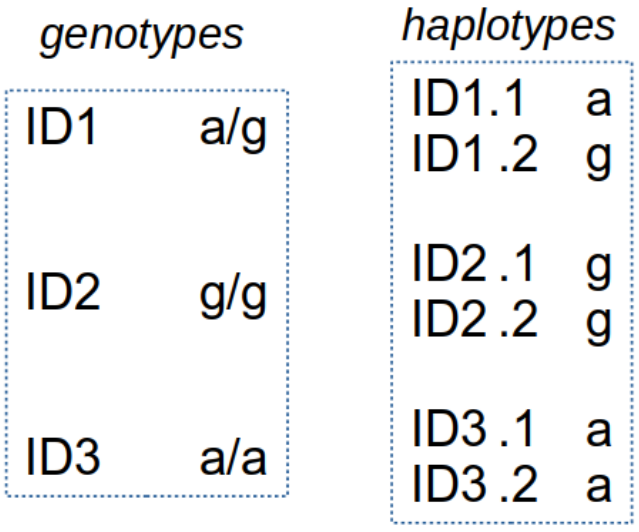


Figure 10: Difference between genotype and haplotype.

* In the literature there are discordant definitions and you will find terms alleles and haplotypes being used interchangeably. In some textbooks you may find that haplotypes refer to chromosomes while allotypes refer to what we define here as haplotypes.

Diploids

What happens if you have multiple copies of each chromosome?

As *diploid* species have two copies of their chromosomes, for a collection of N diploid individuals, there are $2N$ gene copies at each locus, with one or more alleles.

As mutations are rare in most organisms, **di-allelic** models are often used, with at most two alleles at each locus*.

* e.g., at the red-hair *vs.* non-red-hair locus in *MC1R*, most individuals have C, some have T but A and G haven't been observed suggesting a di-allelic model is a valid approximation here.

Terminology

- Gene
- Phenotype
- Locus
- Allele
- Genotype
- Haplotype

time: 9.35am

Evolutionary genetics

Now that all the main terminology has been defined, we can have a closer look at the genetics of evolution.

What is evolutionary genetics?

In evolutionary genetics we are interested in the study genetic **variation** between/within specie/populations. It is both:

- **retrospective**: understanding what determined the current composition of a species/population;
- **predictive**: predicting the future composition of a species/population from its current composition.

Evolution

How do we "measure" evolution?

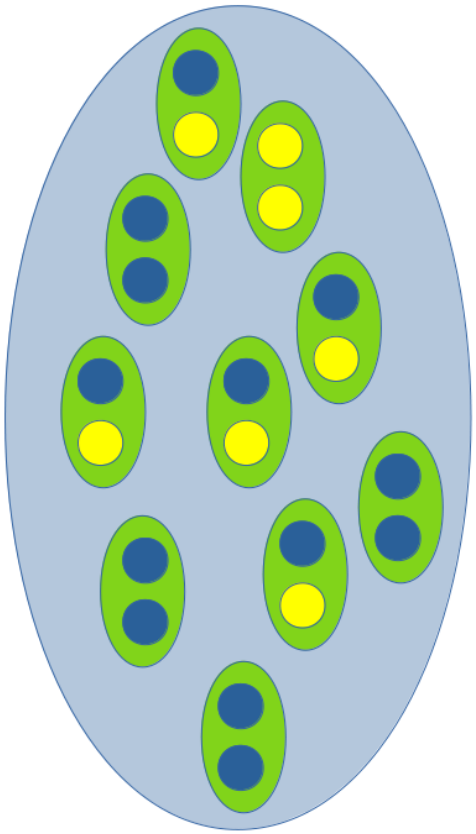
The simplest definition of evolution is "a change in allele frequencies over time".

Therefore, we first need to understand how changes in allele frequencies occur.

But before that, what is an **allele frequency**? How do we calculate them? Is there a concept of **genotype frequency**?

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).

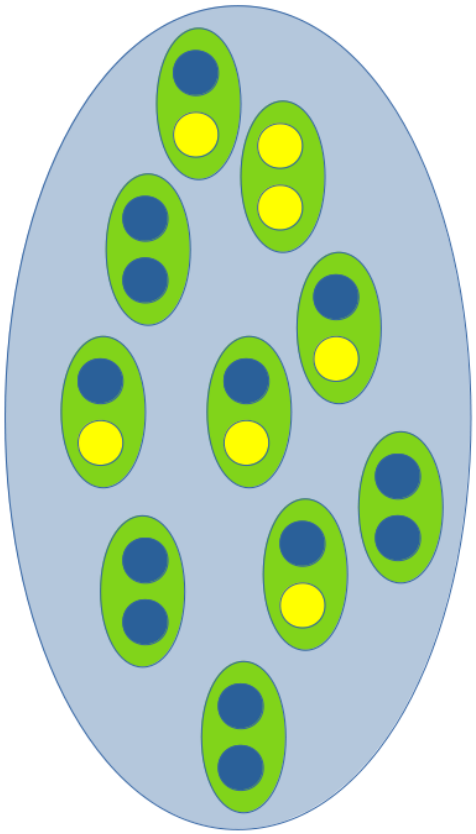


What are the allele frequencies?

How would you calculate, for instance, f_A, the frequency of alleles A (yellow) within this population?
Try to find the answer yourself before moving to the next slide. $f_A = ?$

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).



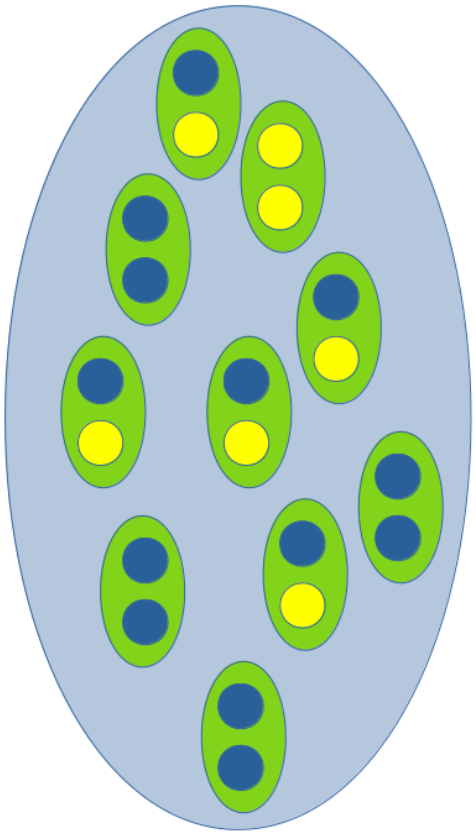
What are the allele frequencies?

$f_A = 7/20 = 0.35$
because we have 7 A alleles out of 20 in total.
How about f_a, the frequency of allele a (blue)?

Try to find the answer yourself before moving to the next slide. $f_a = ?$

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).

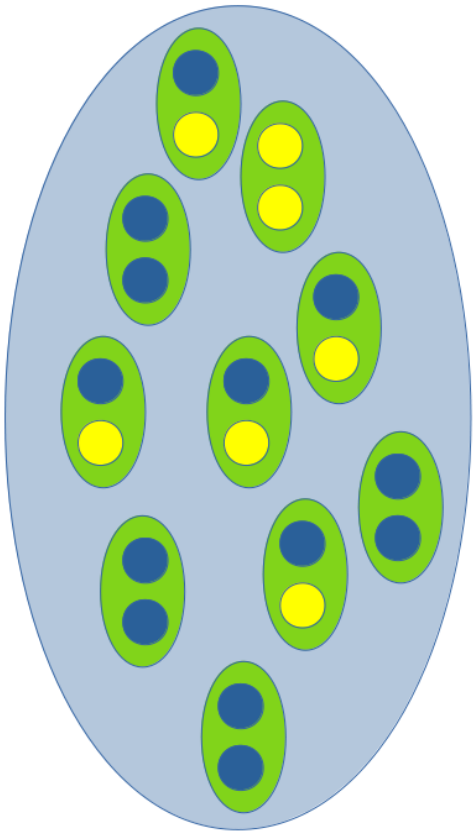


What are the allele frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$
because we have 13 a alleles out of 20 in total.

Do you notice a relationship between f_A and f_a?

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).

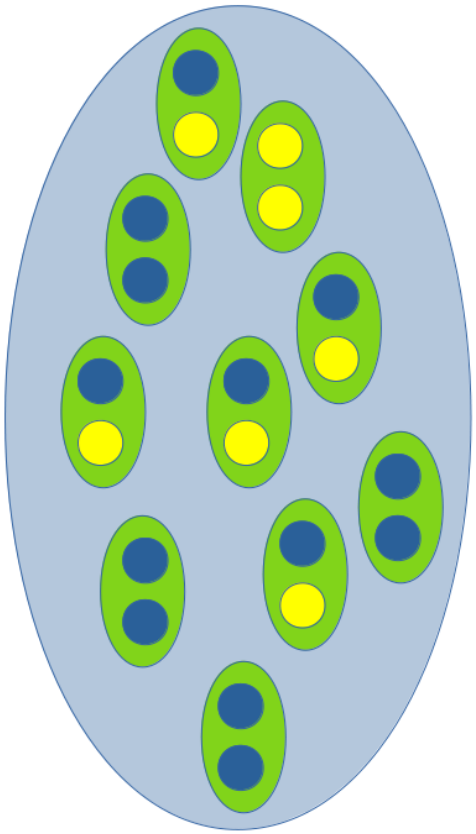


What are the allele frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$
because we have 13 a alleles out of 20 in total.

Do you notice a relationship between f_A and f_a? Their sum is 1! $0.35 + 0.65 = 1$, and this is a true general statement: $f_A + f_a = 1$ for di-allelic variation.

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).



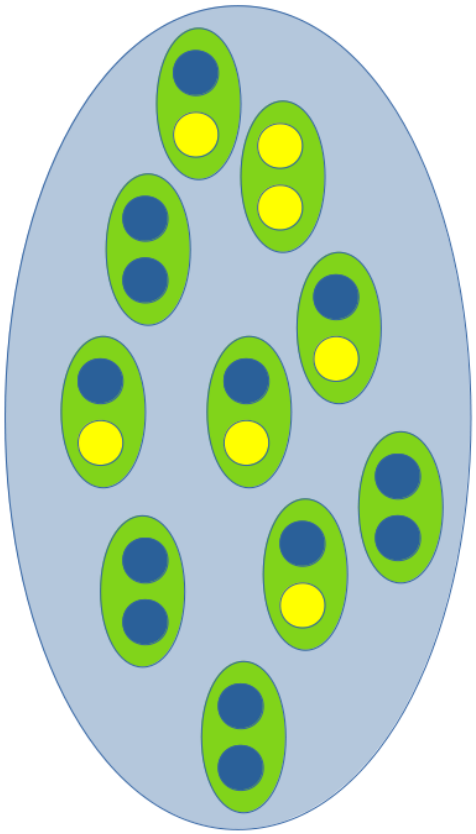
What are the **allele** frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$

What are the **genotype frequencies**? Before answering this question, what are the genotypes in this example? You have diploid individuals with two possible alleles A (yellow) and a (blue).

Try to find the answer yourself before moving to the next slide.

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).



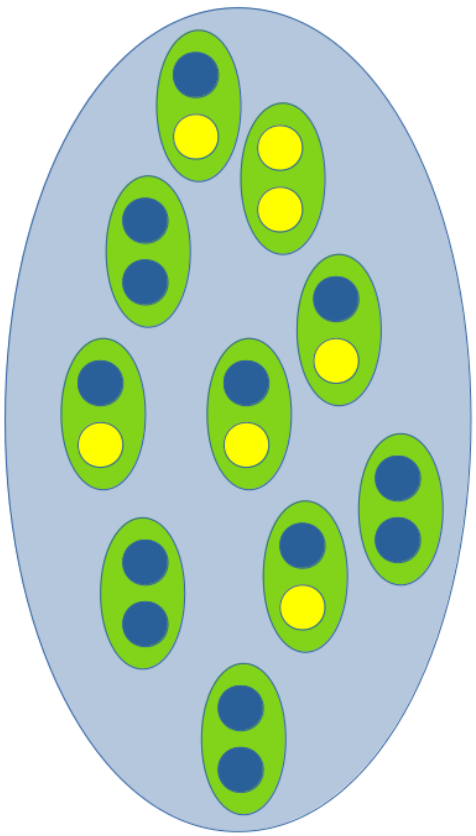
What are the **allele** frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$

What are the **genotype frequencies**?

We have three possible genotypes: {AA, Aa, aa} or {yellow/yellow, yellow/blue, blue/blue}.
What are their frequencies? Count them and check the answer on the next slide.

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).



What are the allele and genotype frequencies?

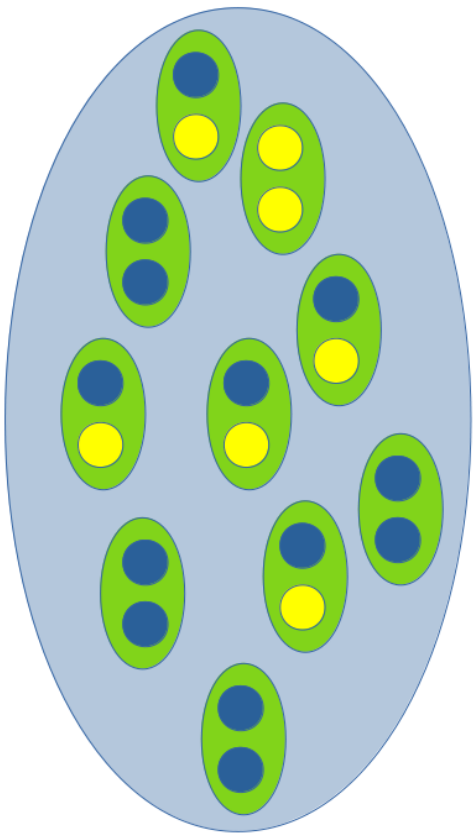
$f_A = 7/20$

$f_a = 13/20$

$f_{AA} =$

Allele and genotype frequency

Let's assume we have a population of $N = 10$ diploid individuals (thus $2N = 20$ gene copies), and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue).



What are the allele and genotype frequencies?

$f_A = 7/20$

$f_a = 13/20$

$f_{AA} = 1/10$

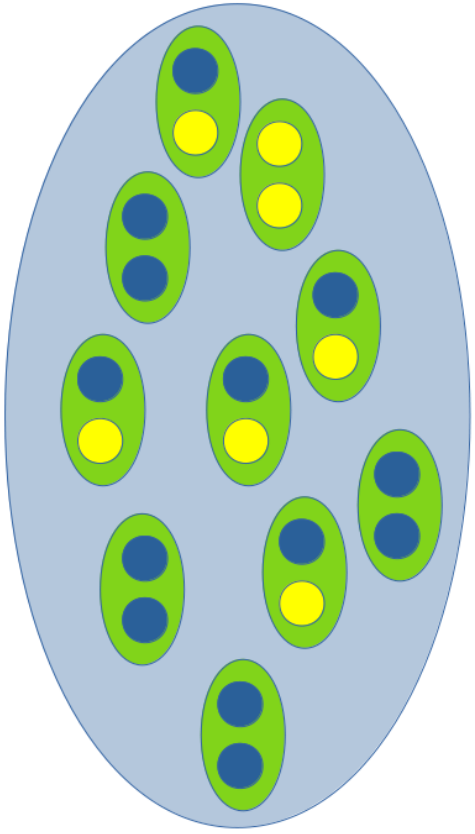
$f_{Aa} = 5/10$

$f_{aa} = 4/10$

We say that AA and aa are **homozygous** individuals and Aa are **heterozygous** individuals.

Note again that $f_{AA} + f_{Aa} + f_{aa} = 1$.

Allele and genotype frequencies

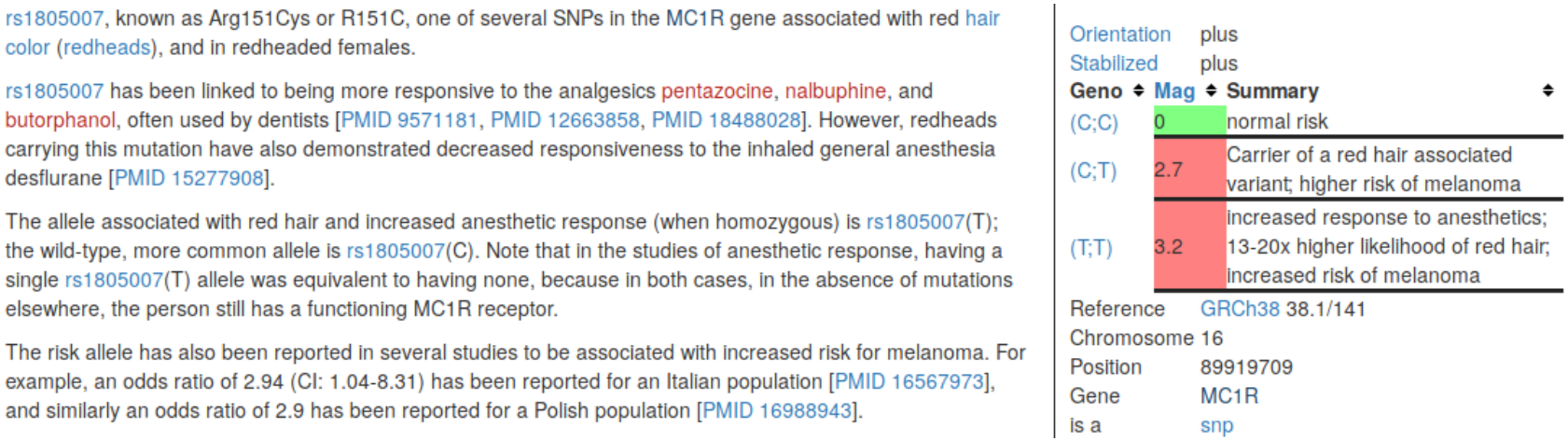


The proportion of heterozygous individuals in the population (f_{Aa}) is called the **heterozygosity**.

The proportion of homozygotes ($1 - f_{Aa} = f_{AA} + f_{aa}$) is the **homozygosity** of the population.

MC1R gene

Here is a challenge for you. The SNPs coded as rs1805007 in the human *MC1R* gene is associated with red hair*.



rs1805007, known as Arg151Cys or R151C, one of several SNPs in the MC1R gene associated with red hair color (redheads), and in redheaded females.

rs1805007 has been linked to being more responsive to the analgesics pentazocine, nalbuphine, and butorphanol, often used by dentists [PMID 9571181, PMID 12663858, PMID 18488028]. However, redheads carrying this mutation have also demonstrated decreased responsiveness to the inhaled general anesthesia desflurane [PMID 15277908].

The allele associated with red hair and increased anesthetic response (when homozygous) is rs1805007(T); the wild-type, more common allele is rs1805007(C). Note that in the studies of anesthetic response, having a single rs1805007(T) allele was equivalent to having none, because in both cases, in the absence of mutations elsewhere, the person still has a functioning MC1R receptor.

The risk allele has also been reported in several studies to be associated with increased risk for melanoma. For example, an odds ratio of 2.94 (CI: 1.04-8.31) has been reported for an Italian population [PMID 16567973], and similarly an odds ratio of 2.9 has been reported for a Polish population [PMID 16988943].

Orientation plus
Stabilized plus

Geno	Mag	Summary
(C;C)	0	normal risk
(C;T)	2.7	Carrier of a red hair associated variant; higher risk of melanoma
(T;T)	3.2	increased response to anesthetics; 13-20x higher likelihood of red hair; increased risk of melanoma

Reference GRCh38 38.1/141
Chromosome 16
Position 89919709
Gene MC1R
is a snp

Figure 11: SNP associated to red hair with alleles C and T

*https://www.snpedia.com/index.php/Rs1805007

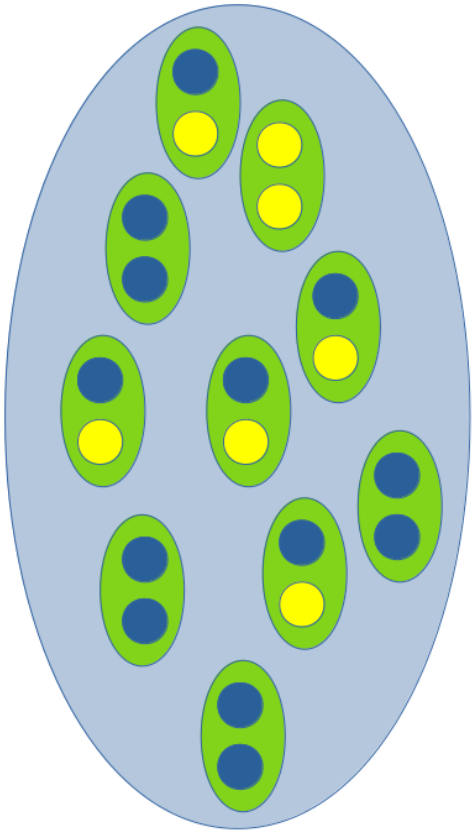
MC1R gene

Assume we obtain a *random* sample of 30 individuals from the population in the UK and find that 25 individuals have genotype CC, 5 individuals have genotype CT and 0 have genotype TT at SNP rs1805007.

1. What are the *estimated* genotype frequencies?
2. What are the homozygosity and heterozygosity in the population?
3. What are the *estimated* allele frequencies? How do you calculate them?
4. Why are these frequencies *estimated* and not *calculated*?

time:9.45am - give until 9.50am

Allele frequencies in time



time:9.50am

We focus on describing the changes of f_A and f_a with time.

If we can describe how we expect allele frequencies to change through time in a population, we have gained important insights of its evolution.

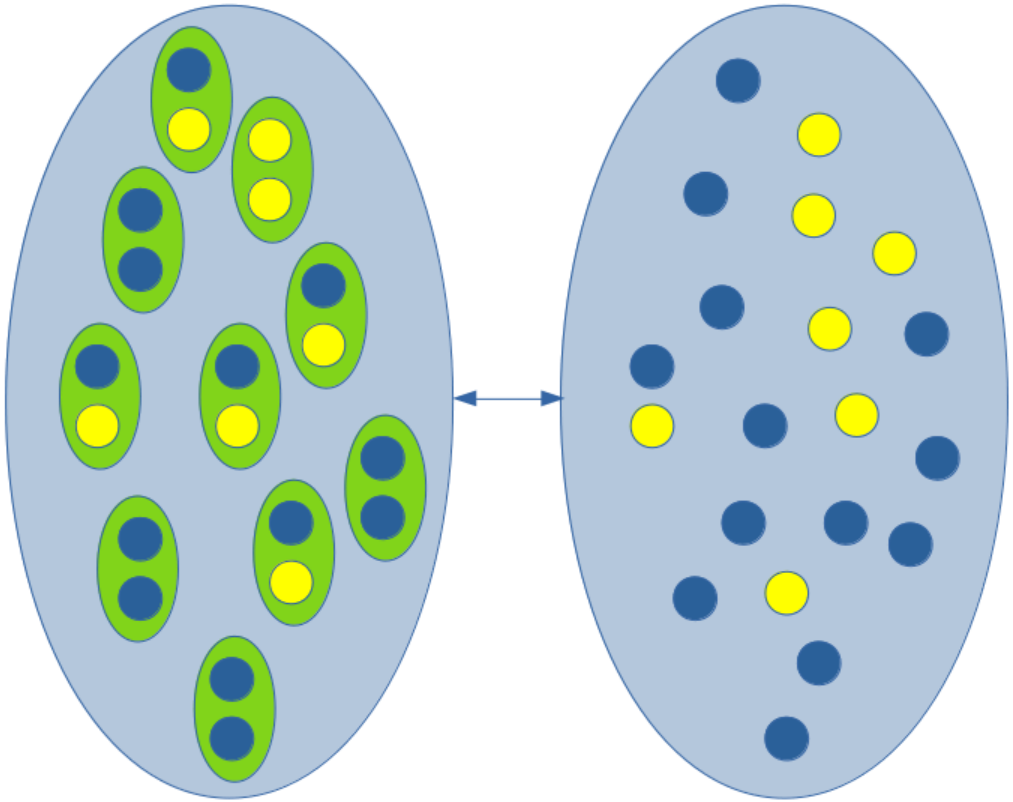
Alleles to genotypes

We can calculate allele frequencies from genotype frequencies.

Can we *predict* genotype frequencies from allele frequencies?

For instance, if we know that the frequency of allele T in position 478 of the human gene *MC1R* gene is 0.08, what proportion of the population is *expected* to have genotype TT?

Can we *predict* genotype frequencies from allele frequencies?
Can we go back and forth between these two figures (genotypes on the left hand side and alleles on the right hand side)?



Can we *predict* genotype frequencies from allele frequencies?
Yes, under these assumptions:

- the organism is diploid
- the locus is di-allelic
- reproduction is sexual
- generations are non-overlapping
- mating is random: individuals mate with each other without regard to their genotype
- populations are "infinite" (very large in size)
- there is no mutation, migration, natural selection or drift*

If these conditions are met, these we can calculate genotype frequencies under **Hardy-Weinberg Equilibrium (HWE)**.

* we will learn what these terms mean later on.

Hardy-Weinberg Equilibrium (HWE)

From the allele frequencies f_A and f_a we can calculate the *expected* genotype frequencies under HWE as following:

Genotype frequencies under HWE			
Genotype	AA	Aa	aa
Frequency	f_A^2	$2f_Af_a$	f_a^2

Hardy-Weinberg Equilibrium (HWE)

From HWE equations we learn that:

1. $f_A^2 + 2f_A f_a + f_a^2 = 1$
2. random mating does not change the allele frequencies in the next generation

In other words, under HWE we do not expect to see *on average* a change in allele frequency from one generation to the next.

HWE in *MC1R*

Figure 12: Rupert Grint

With an allele frequency of 0.08 for allele T in the UK population, how many TT homozygotes might we expect under HWE?

HWE in *MC1R*

Figure 12: Rupert Grint

With an allele frequency of 0.08 for allele T in the UK population, how many TT homozygotes might we expect under HWE? It's $0.08^2 = 0.0064$.

Individuals with TT genotype will likely have red hair, but a much larger proportion of the population has red hair.

Deviations from HWE

- Assortative mating: not random with respect to genotype
- Inbreeding: mating of related individuals
- Population structure: sample of individuals from two or more subpopulations
- Natural selection: alleles affect the *fitness** of the carrier

* more on this later time: 10.15am

Inbreeding coefficient

Inbreeding coefficient (F)

Most common statistic to measure deviations from HWE: it describes the degree to which heterozygosity is reduced both in individuals and in populations.

$$F = \frac{2f_A f_a - f_{Aa}}{2f_A f_a} \tag{1}$$

If $F = 0$ the population is in HWE, if $F = 1$

Inbreeding coefficient

Inbreeding coefficient (F)

Most common statistic to measure deviations from HWE: it describes the degree to which heterozygosity is reduced both in individuals and in populations.

$$F = \frac{2f_A f_a - f_{Aa}}{2f_A f_a} \tag{1}$$

If $F = 0$ the population is in HWE, if $F = 1$ there are no heterozygotes.

Inbreeding coefficient

$$f_{Aa} = 2f_A f_a(1 - F) \tag{2}$$

- The proportion of heterozygotes in the population is reduced by a factor of F from that expected under HWE.
- If we know F and the allele frequencies, we can predict genotype frequencies without assuming HWE.

Are there species likely to deviate from HWE?

Self-fertilizing plants

Figure 13: Flower of wild oats (*Avena fatua*)

Genotype frequencies at one locus are:
$f_{AA} = 0.58$, $f_{Aa} = 0.07$, $f_{aa} = 0.35$.

1. What is F, the inbreeding coefficient?
   time: 10:00 - 10:10

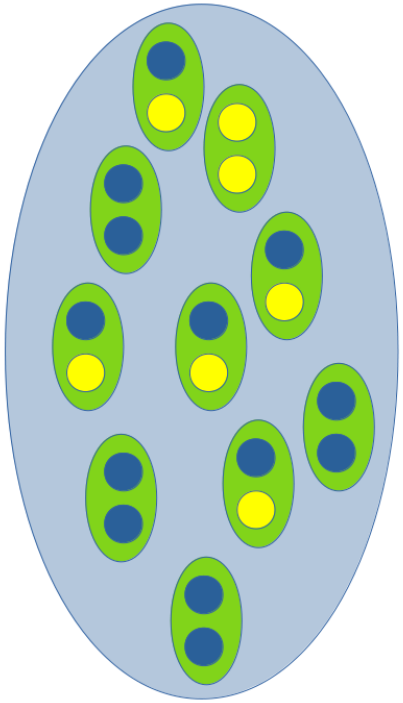
Self-fertilizing plants

Figure 13: Flower of wild oats (*Avena fatua*)

Genotype frequencies at one locus are:
$f_{AA} = 0.58,\ f_{Aa} = 0.07,\ f_{aa} = 0.35.$

1. What is F, the inbreeding coefficient?
   time: 10:00 - 10:10
   $f_A = 0.58 + 0.07/2 = 0.615,$
   $f_a = 1 - 0.615 = 0.385$
   $F = (2 \times 0.385 \times 0.615 - 0.07)/(2 \times 0.385 \times 0.615) = 0.852$
2. Does it deviate from HWE?

Testing for deviations from HWE



- A random sample for a population in HWE may deviate from HWE.
- We need a formal statistical test:

  Null hypothesis: genotype frequencies follow those predicted by HWE

  Alternative hypothesis: genotype frequencies do **not** follow those predicted by HWE

Testing for deviations from HWE

Chi-square test: $\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$

- observed values (O_i, genotype counts)
- expected values (E_i, expected genotype counts under HWE)
- degrees of freedom: $3 - 1 - 1 = 1$

If χ^2 is large enough, we reject the null hypothesis.

Testing for deviations from HWE

$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$ e.g. observed genotypes: $O_{AA} = 20$, $O_{Aa} = 10$,

$O_{aa} = 10$

Testing for deviations from HWE

$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$ e.g. observed genotypes: $O_{AA} = 20$, $O_{Aa} = 10$, $O_{aa} = 10$

- $f_{AA} = 1/2$, $f_{Aa} = 1/4$, $f_{aa} = 1/4$

Testing for deviations from HWE

$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$ e.g. observed genotypes: $O_{AA} = 20$, $O_{Aa} = 10$, $O_{aa} = 10$

- $f_{AA} = 1/2$, $f_{Aa} = 1/4$, $f_{aa} = 1/4$
- $f_A = 1/2 + (1/4)/2 = 5/8$, $f_a = 1/4 + (1/4)/2 = 3/8$

Testing for deviations from HWE

$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$ e.g. observed genotypes: $O_{AA} = 20$, $O_{Aa} = 10$,

$O_{aa} = 10$

- $f_{AA} = 1/2$, $f_{Aa} = 1/4$, $f_{aa} = 1/4$
- $f_A = 1/2 + (1/4)/2 = 5/8$, $f_a = 1/4 + (1/4)/2 = 3/8$
- $E_{AA} = 40 \times (5/8)^2 = 15.625$, $E_{Aa} = 40 \times 2 \times 3/8 \times 5/8 = 18.75$, $E_{aa} = 40 \times (3/8)^2 = 5.625$
- $\chi^2 = ... + ... + ... = 8.711$

Is 8.711 "large enough"?

Testing for deviations from HWE

We need to compare our value 8.711 with a critical value for a chi-square distribution with one degree of freedom.

.995	.99	.975	.95	.9	.1	.05	0.025	.01
0	0	0		0.02	2.71	3.84	5.02	6.63

Do we reject the null hypothesis of HWE?

Testing for deviations from HWE

We need to compare our value 8.711 with a critical value for a chi-square distribution with one degree of freedom.

.995	.99	.975	.95	.9	.1	.05	0.025	.01
0	0	0		0.02	2.71	3.84	5.02	6.63

Do we reject the null hypothesis of HWE?
Yes, since p-value is < 0.05, assuming such threshold for significance (but remember that statistical significance does NOT imply biological significance).
time: 10:15 break

Intended Learning Outcomes

In this session you have learnt

- to describe all different types of genetic data
- to demonstrate the relationship between allele and genotype frequencies
- to calculate Hardy-Weinberg Equilibrium proportions

Queen Mary
University of London

Intended Learning Outcomes

In this session you will learn

- to describe genetic drift,
- to interpret the change of allele frequencies over time,
- to appreciate the effect of population size on drift,
- to define fitness and selection coefficient.

Allele frequencies through time

Evolutionary genetics often focuses on describing the changes of allele frequencies through time.

The three most important factors that cause allele frequencies to change are:

- genetic drift,
- natural selection,
- mutations.

All of these "work" jointly but which one is the strongest force*?

* under most of the circumstances time: 9.45

Genetic drift accounts for most of the genetic differentiation between populations of the same species and between different species.

We need to understand the effect of genetic drift on changes in allele frequency if we wish to gain insights onto the main genetic source of variation between species or populations.

What is genetic drift*?

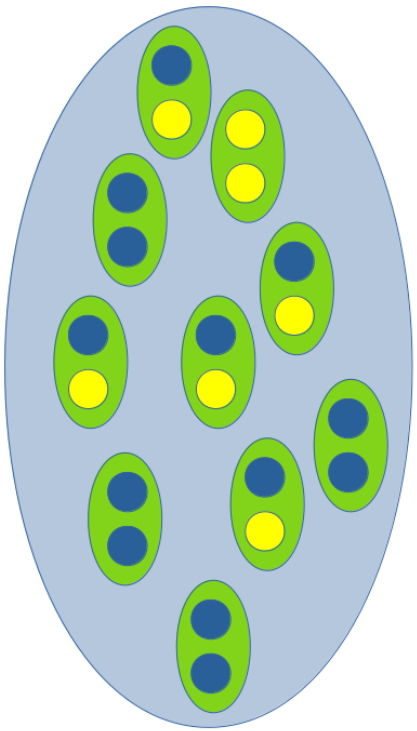
* sometimes also called *genetic draft*

Genetic drift

Genetic drift is the **random** change of allele frequencies in populations of **finite** size.

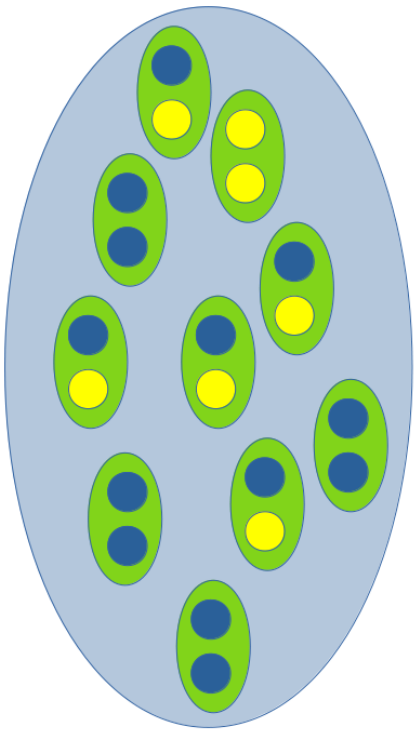
- It describes the process by which allele frequencies change over time due to the effect of random sampling.
- It occurs as a consequence of finite population size.

Genetic drift



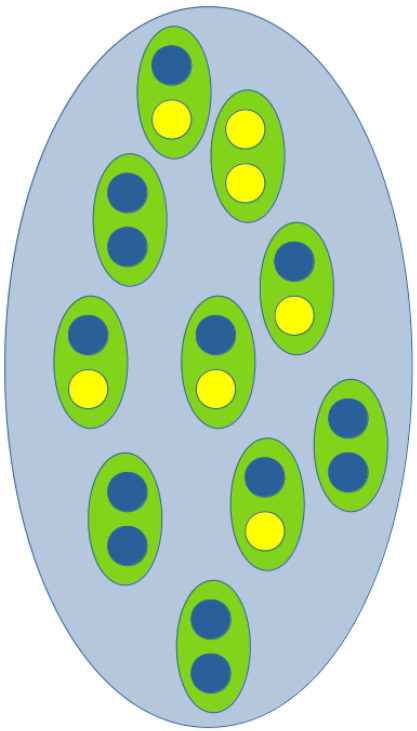
- Some individuals leave many offspring, others fewer, other none.

Genetic drift



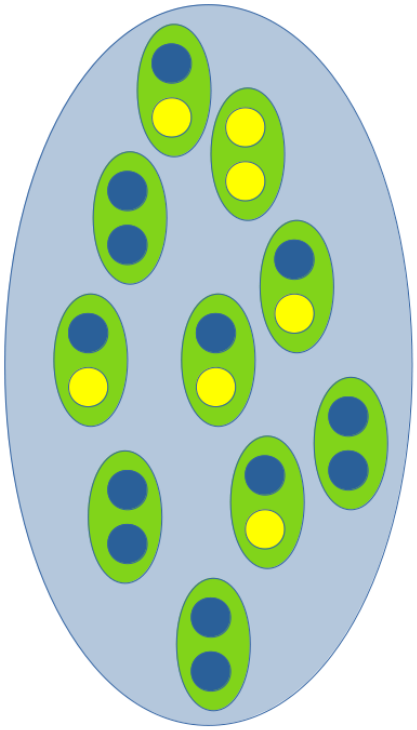
- Some individuals leave many offspring, others fewer, other none.
- Heterozygous individuals will randomly transmit allele A or a.

Genetic drift



- Some individuals leave many offspring, others fewer, other none.
- Heterozygous individuals will randomly transmit allele *A* or *a*.
- It is likely that allele frequencies will *slightly* change from one generation to another.

Genetic drift



- Some individuals leave many offspring, others fewer, other none.
- Heterozygous individuals will randomly transmit allele *A* or *a*.
- It is likely that allele frequencies will *slightly* change from one generation to another.
- Over many generations, this process can produce large changes in allele frequencies

Genetic drift model

The changes in allele frequency due to drift in a population follow a model* which has the following assumptions:

- haploid population
- asexual (no mating)
- discrete generations

The next generation of gene copies (or gametes) is produced by random **sampling with replacement** (independently and with equal probability) gene copies from the previous generation.

* This model is often called Wright-Fisher model.

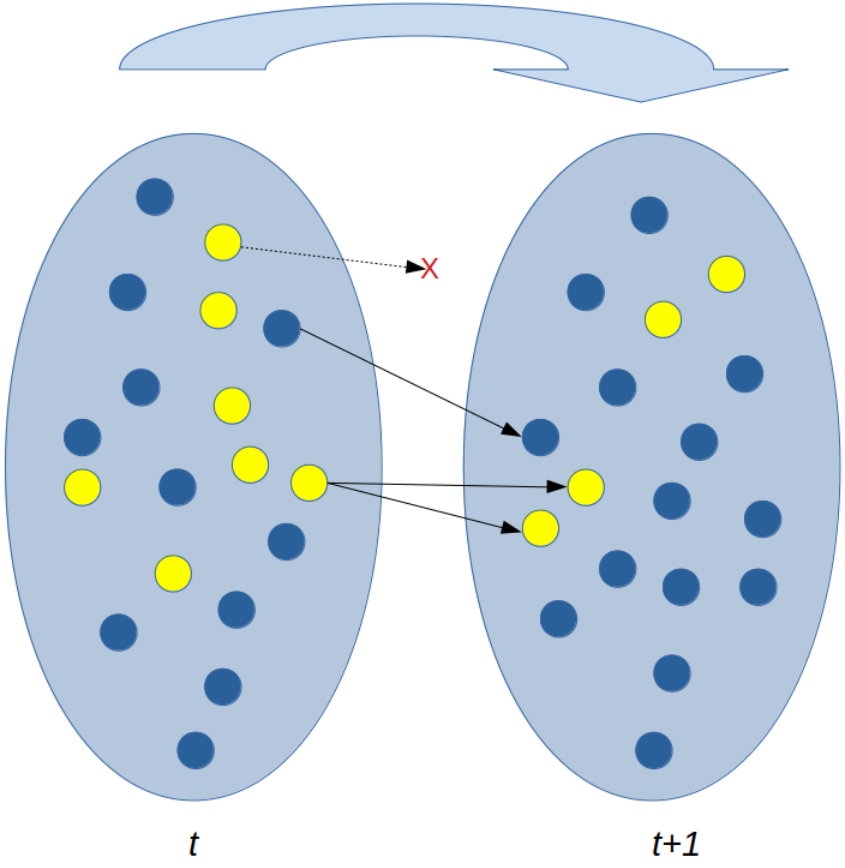


Figure 14: Two generations of genetic drift.

What is the *expected* allele frequency in the next generation under this model? Do the experiment yourself!

note to lecturer: jupyter-notebook: drift (1) : time: 10:55 - until 11:10

: time: 10:55 - until 11:10

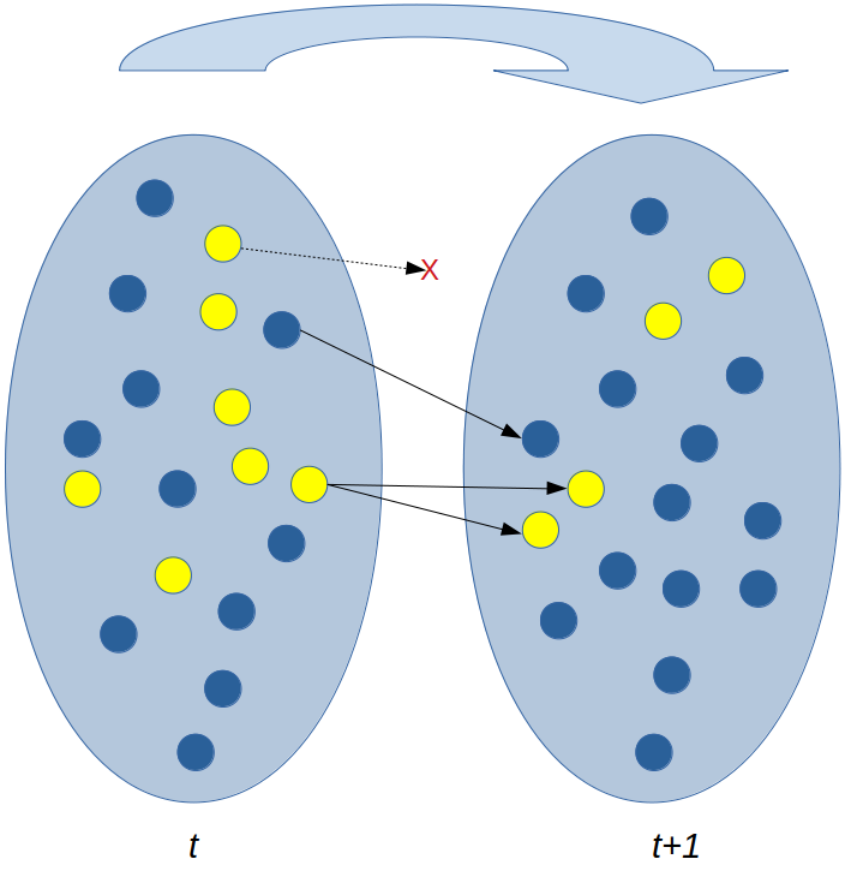


Figure 15: Two generations of genetic drift.

The *expected* allele frequency in the next generation is equal to the allele frequency in the current generation.

- By pure chance we might sample a particular allele more or less often, causing the allele frequency to *slightly* change from one generation to the next.
- Nevertheless, the *expected* allele frequency in the next generation is equal to the allele frequency in the current generation.

Why are these two apparently contradictory statements both true?

Genetic drift model

- The distribution of offspring in generation $t+1$ is given by a **binomial** distribution
- Under the Wright-Fisher model, we can easily characterise the change in allele frequency mathematically.

e.g. what is the probability that any gene copy in generation $t+1$ is A?

Expected allele frequency

What is the probability that any gene copy in generation $t+1$ is A?

$$E[f_A(t+1)] = 2Nf_A(t)/2N = f_A(t) \tag{3}$$

The **expected** allele frequency in generation $t+1$ is equal to the allele frequency in generation t.

What happens if we repeat the sampling with replacement scheme over **many** generations?
Do allele frequencies change over time? Do they get lost or fixed and with what probality?
What is the effect of the initial allele frequency? note to the lecturer:

jupyter-notebook: drift (2) : time: 11:15 - until 11:40

note to the lecturer: jupyter-notebook: drift (2) : time: 11:15 - until 11:40

- At each generation, allele frequency might change a bit.
- Small changes add up and, after many generations, allele frequency may have changed significantly.

Many small changes may result in large evolutionary changes over sufficiently long periods of time.

- Allele frequency may increase or decrease with equal probabilities.
- In some cases, allele has become fixed ($f = 1$) or is lost ($f = 0$).

When an allele first has become fixed or is lost, its frequency cannot change anymore*.

* if we ignore the effect of mutation; in fact, in the absence of recurrent mutation, it can be shown mathematically that a new allele must eventually become fixed or be lost.

Effect of population size

1. How **fast** can genetic drift change allele frequencies?
2. Does it depend on the population size?
3. If so, would genetic drift be stronger in a small population or in a large population?

note to the lecturer: jupyter-notebook: drift (3) : time: 11:45 - until 12:00

Effect of population size

note to the lecturer: jupyter-notebook: drift (3) : time: 11:45 - until 12:00

Effect of population size

Large changes in allele frequency are unlikely to happen in large populations, but they happen more easily by chance in small populations.

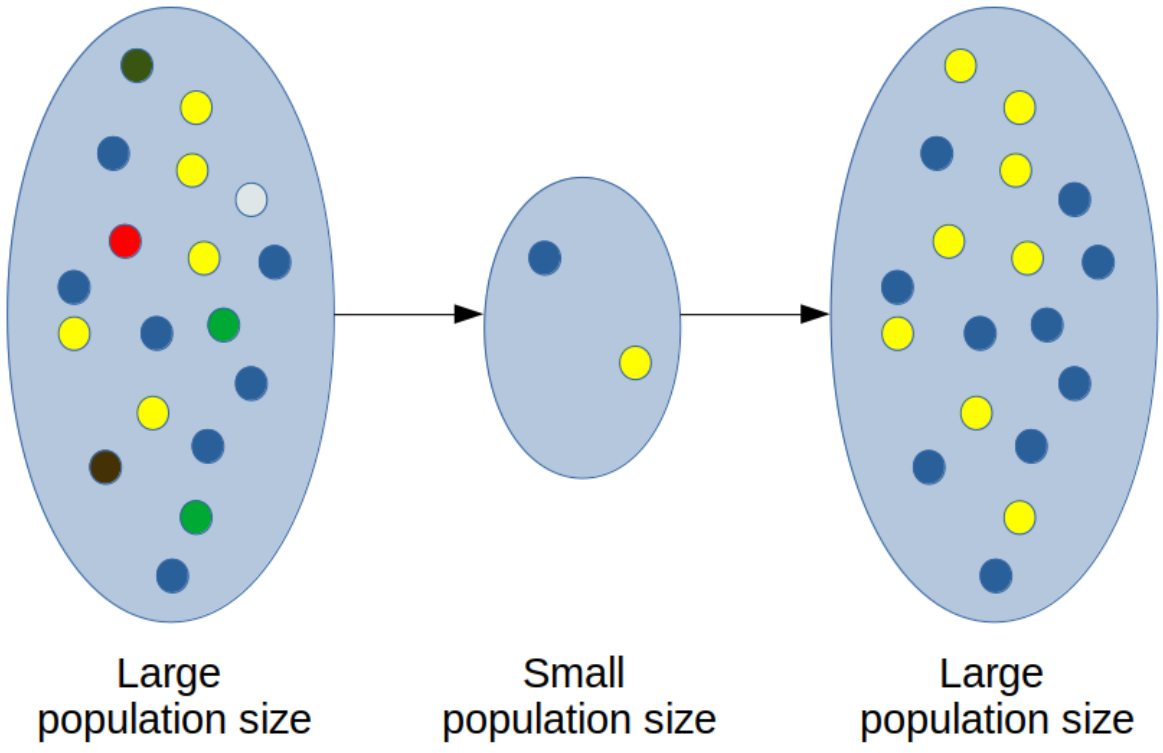
Genetic drift works much faster in small populations than in large populations.

The effect of population size on genetic drift has important implications for our understanding of natural populations. Why?

Drift can be particularly strong under a population **bottleneck**.

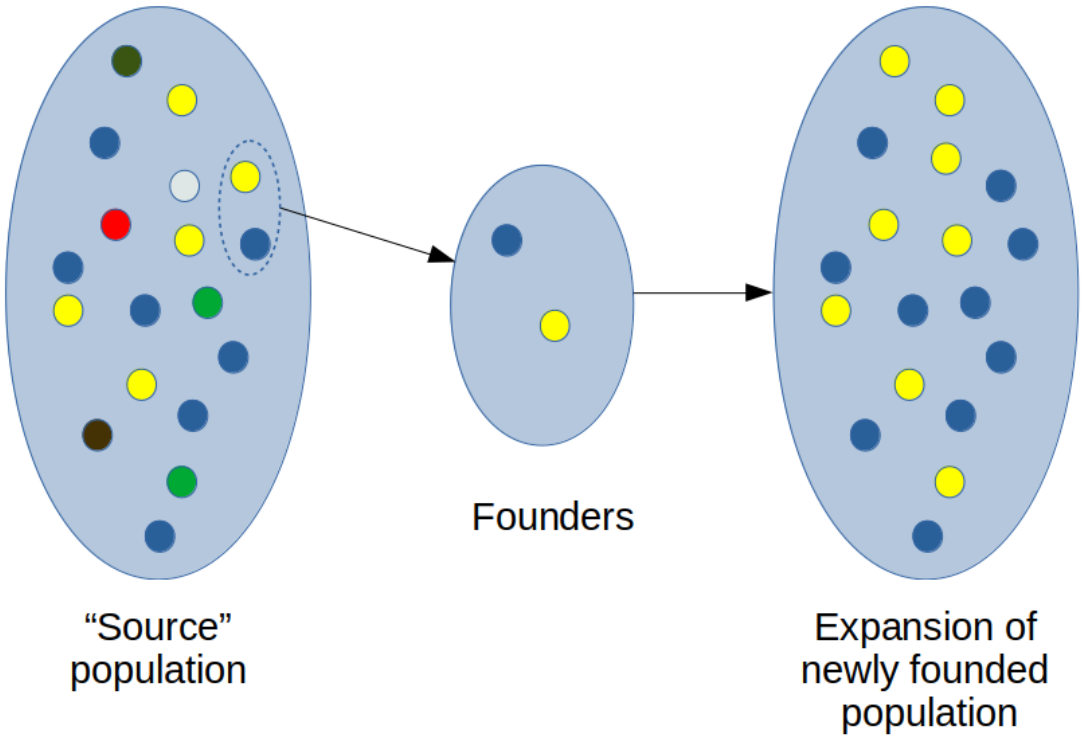
Bottleneck

Short period of time when the population size is very small and many alleles become either fixed or lost in the population. As a consequence, much of the population genetic variation is lost.

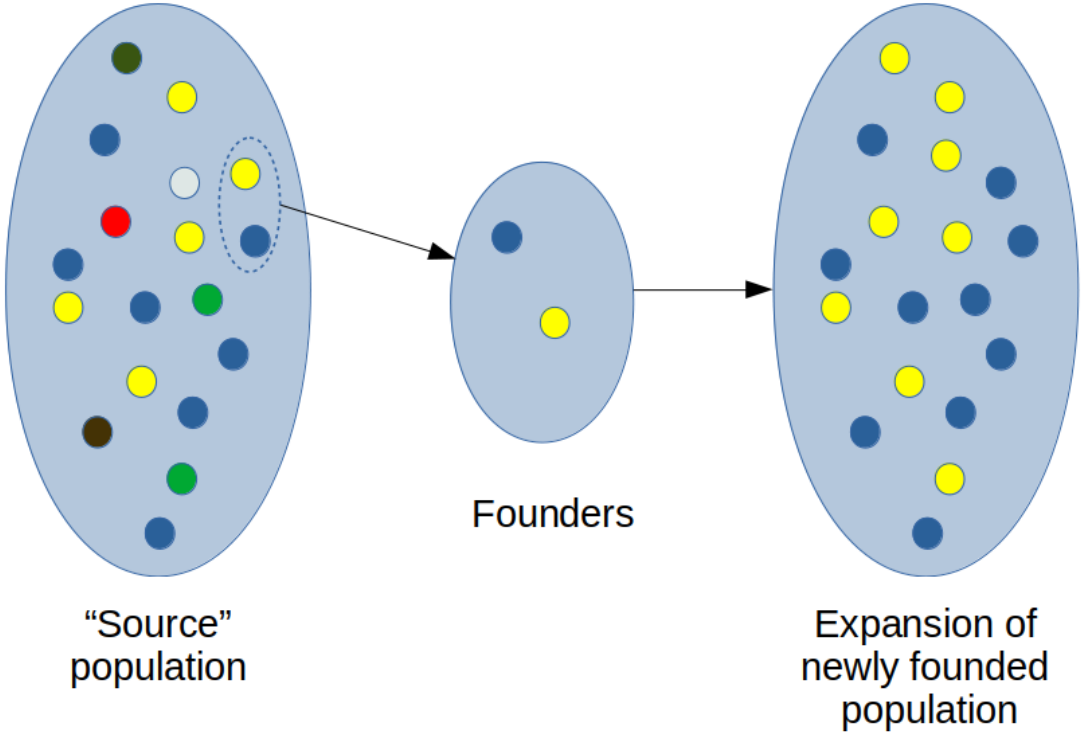


Founder effect

Reduction in variability caused by a bottleneck in the population size during the founding of a new population.



Founder effect



Genetic divergence after speciation may be helped along by the **strong** effects of genetic drift in the founders of a population.

Genetic drift

These models are an extremely simplified cartoon of "real" life. We could make it more realistic by allowing for:

- two sexes,
- population size to change over time,
- number of offspring per individual to vary,
- ...

However, these modifications make little difference to the process of drift. The key fact is always true:

genetic drift causes allele frequencies to change in a random fashion over time.

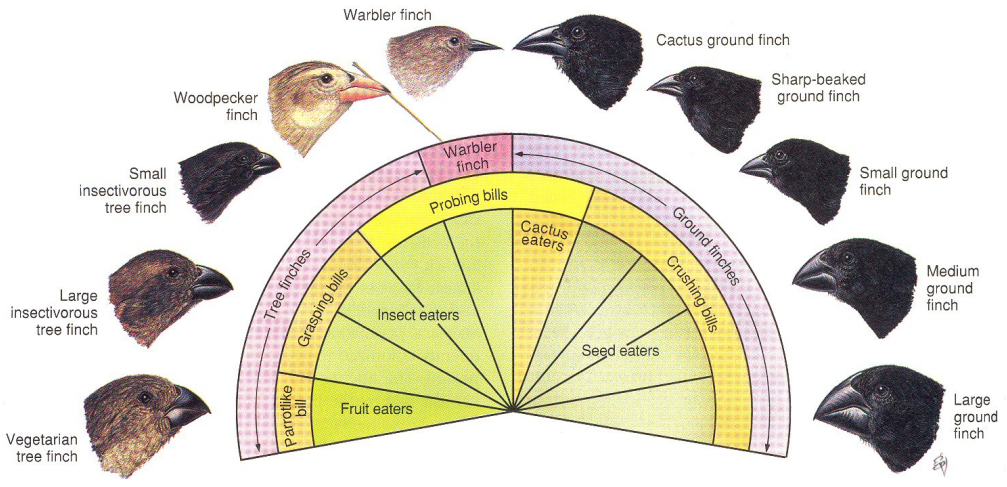
So far so good

- random mating
- genetic drift

What if different alleles affect survival? time: 12:10

Natural selection

- heritable traits that increase the **fitness** become more common in the population
- mutations evolve accordingly to their **effect on the fitness** of the carrier
- **functionality** is the prerequisite for selection to be effective



Fitness

We know that selection occurs because different individuals have different fitness, but what exactly do we mean by **fitness**?

Fitness

The word fitness in an evolutionary context can be defined as:

> the expectation of the number of descendant genes at the same stage of the life cycle in the next generation.

In the next part, I will use *fitness* to indicate a property of genotypes, not of alleles or phenotypes.

Absolute and relative fitness

- **Absolute fitness**: measured as the change in abundance of a genotype from one generation to the next (assuming infinite population size, no mutation, and non-overlapping generations). In other words, the fittest genotypes will increase in abundance compared to less fit genotypes.

Absolute and relative fitness

- **Absolute fitness**: measured as the change in abundance of a genotype from one generation to the next (assuming infinite population size, no mutation, and non-overlapping generations). In other words, the fittest genotypes will increase in abundance compared to less fit genotypes.
- **Relative fitness**: calculated by dividing all fitness values by the largest value, meaning the fittest genotype always has a relative fitness of 1.

Fitness and selection

- Fitness is a property of a particular genotype.
- Selection is a process leading to different expectations of how gene copies are transmitted to the next generation.

If different individuals of a population have different fitness then we say that selection is operating.

If different individuals have the same fitness then we say that there is no selection, or equivalently, that the population is evolving **neutrally**.

Fitness and selection

We need to consider the **fitness** of each *genotype*.

For instance, for one locus with two alleles A and a in a diploid, each genotype has its fitness:

- ω_{AA} for genotype AA
- ω_{Aa} for genotype Aa
- ω_{aa} for genotype aa

Fitness and selection

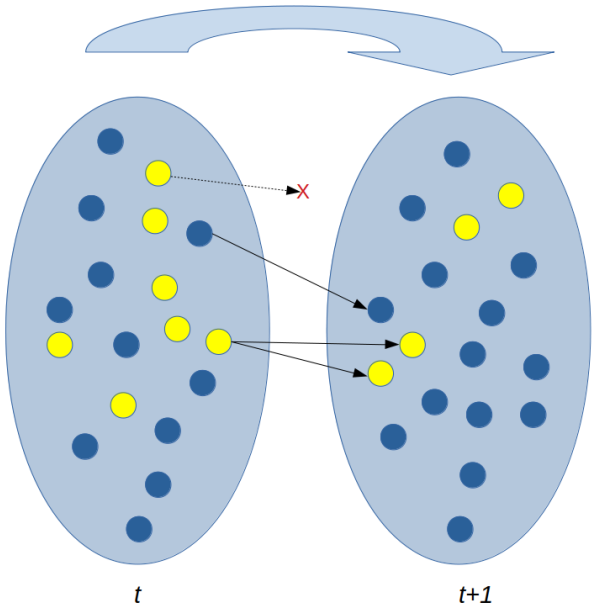
The strength of selection (or **selection coefficient**), often represented by the symbol s, is defined from the fitness values.

For example, if Aa is not the fittest genotype then the selection coefficient against heterozygotes can be interpreted as the deficit from a relative fitness of 1, so that

$$\omega_{Aa} = 1 - s$$

What is the effect of selection on changes in allele frequency?

Selection and drift combined



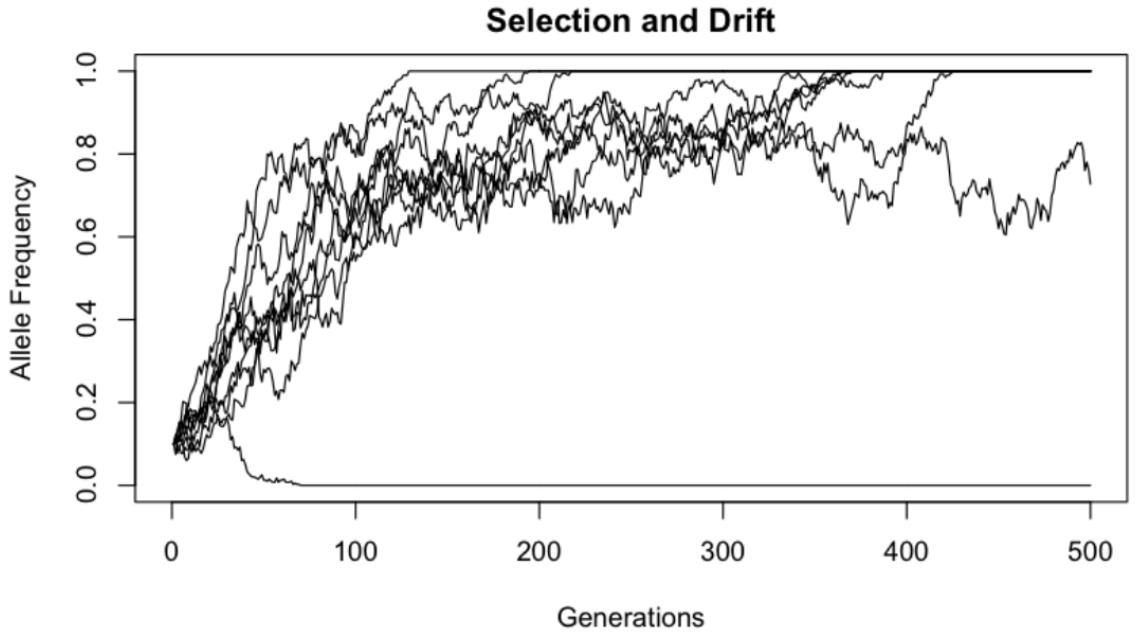
With drift only, all individuals had the same fitness.

The effect of high fitness is to make an individual more likely to be the parent of offspring in the next generation.

Nevertheless, it is still possible that a fit individual will have no offspring.

Selection and drift combined

If allele A belongs to a genotype with high fitness, its frequency will still drift but with a "tendency" towards fixation.



Is it still possible that the A allele is lost? : jupyter-notebook: selection (3) time: 12:20 - 10mins QandA

Intended Learning Outcomes

In this session you have learnt

- to describe genetic drift,
- to interpret the change of allele frequencies over time,
- to appreciate the effect of population size on drift,
- to define fitness and selection coefficient.

Intended Learning Outcomes

In this session you will learn

- to describe all different types of natural selection,
- to appreciate the effect of novel mutations on allele frequency,
- to understand the concept of gene flow.

So far we have learnt how random mating and genetic drift change allele frequencies in time.
What if

- different alleles affect survival?
- new mutations arise?
- migrants move between populations?

In other words, what is the effect of

- **selection**
- mutation
- gene flow

on the change of allele frequency?

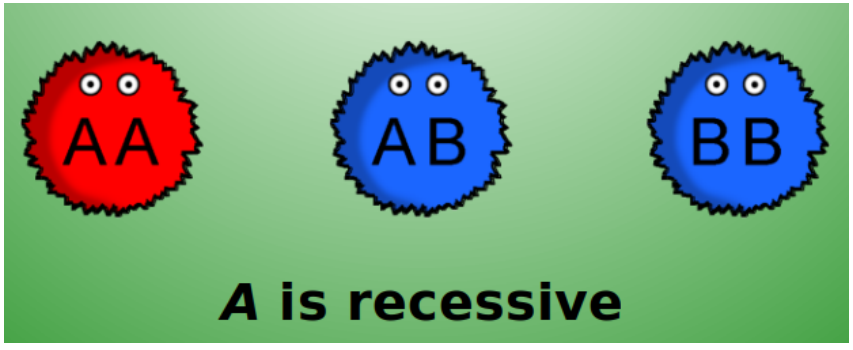
Types of selection

If an allele is **dominant** then the heterozygote has the same *phenotype* as the homozygote for the dominant allele.

Types of selection

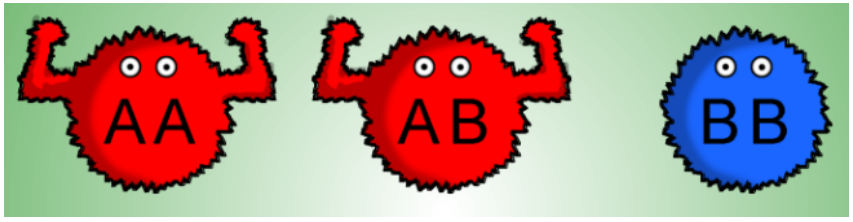
If an allele is **dominant** then the heterozygote has the same *phenotype* as the homozygote for the dominant allele.

If an allele is **recessive** then the heterozygote has the same *phenotype* as the other homozygote.



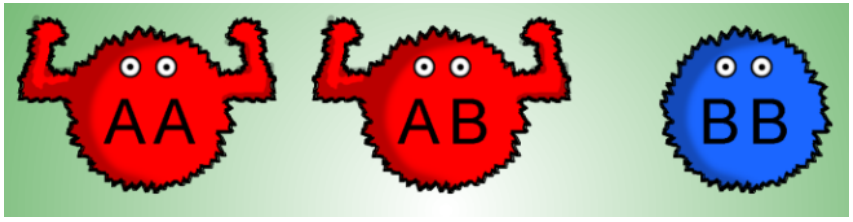
Types of selection

If A is **dominant** then the heterozygote has the same *fitness* as the homozygote for that allele.

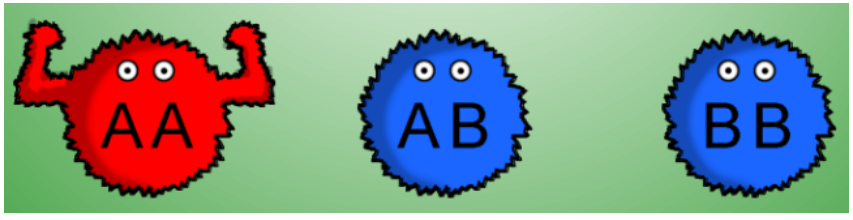


Types of selection

If A is **dominant** then the heterozygote has the same *fitness* as the homozygote for that allele.



If an allele is **recessive** then the heterozygote has the same *fitness* as the other homozygote.



The **selection coefficient** s is defined as $\frac{\omega_a}{\omega_A} = 1 - s$ for haploids with alleles A and a, with A being dominant.

The **selection coefficient** s is defined as $\frac{\omega_a}{\omega_A} = 1 - s$ for haploids with alleles A and a, with A being dominant.

$f_A(t)$ depends on the product between s (selection coefficient, or "strength") and t (time).
If s is small, then the value of t to generate a certain change in f_A is *inversely* proportional to s.

If $s > 0$ then A-bearing individuals have the advantage, the opposite is true for $s < 0$.

jupyter-notebook: selection (1)

Additive selection

What if a recessive allele is not totally masked by the dominant allele?

If A is dominant but the traits have **additive** fitness then the heterozygote has a fitness value that is intermediate between the two homozygotes*.

* You do not need to know any more than that. This enters an area called *quantitative genetics* and there are entire courses on that topic.

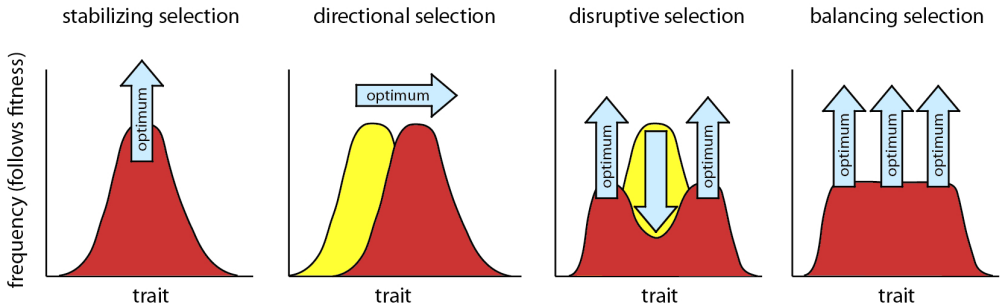
The types of selection just described can be defined as:

- additive selection: $\omega_{Aa} = 1 - s$; $\omega_{aa} = 1 - 2s$
- dominant advantageous allele: $\omega_{AA} = \omega_{Aa}$
- recessive advantageous allele: $\omega_{Aa} = \omega_{aa}$

with A being the advantageous allele.

If the selection coefficient does not change in time, we have several possible scenarios:

- **directional** selection (additive, dominant or recessive): one allele is favoured,
- **heterozygote advantage**: a special case of balancing selection where multiple alleles are maintained in the population,
- **heterozygote disadvantage**: a special case of disruptive selection where the extremes of a trait are favoured,
- stabilising selection: variation is reduced.



If the direction of selection changes over time, we have fluctuating selection.

Directional selection

A is the **advantageous allele** if $\omega_{aa} \leq \omega_{Aa} \leq \omega_{AA}$.

f_A will increase every generation and eventually reach 1 (i.e. **fixation** of A and **loss** of a).

The rate of change in allele frequency depends on s, selection coefficient.
Even small s can change allele frequency substantially over many generations.

jupyter-notebook: selection (2)

Heterozygote advantage

If $\omega_{Aa} > \omega_{AA}, \omega_{aa}$ we define

$$\frac{\omega_{aa}}{\omega_{Aa}} = 1 - s_{aa}$$
$$\frac{\omega_{AA}}{\omega_{Aa}} = 1 - s_{AA}$$

f_A will tend to the same value regardless of its initial frequency.
In fact, selection won't eliminate either allele (it is a special case of **balancing selection**).

Heterozygote disadvantage

If $\omega_{Aa} < \omega_{AA}, \omega_{aa}$, it results in the fixation of one of the two alleles (which one?).

It is an example of **disruptive selection** that removes low-frequency alleles.

We have learnt how **selection** changes allele frequencies in time. You also know how **genetic drift** changes allele frequencies in time.

What is the combined effect of selection and drift?

jupyter-notebook: selection (3)

For a population of size N, the fixation probability u of a new mutation with selection coefficient s can be defined as:

- strongly deleterious, if $2Ns << -1$ then $u \approx 0$
- nearly neutral, if $-1 < 2Ns < 1$ then $u \approx 1/(2N)$
- strongly advantageous, if $2Ns >> 1$ then $u \approx 2s$

Strongly advantageous mutations are not necessarily fixed. Slightly deleterious alleles have a small but non-zero chance of being fixed.

Whether an allele is strongly selected or nearly neutral depends on both the selection coefficient and the population size.

What is the effect of

- selection
- **mutation**
- gene flow

on the change of allele frequency?

Mutations

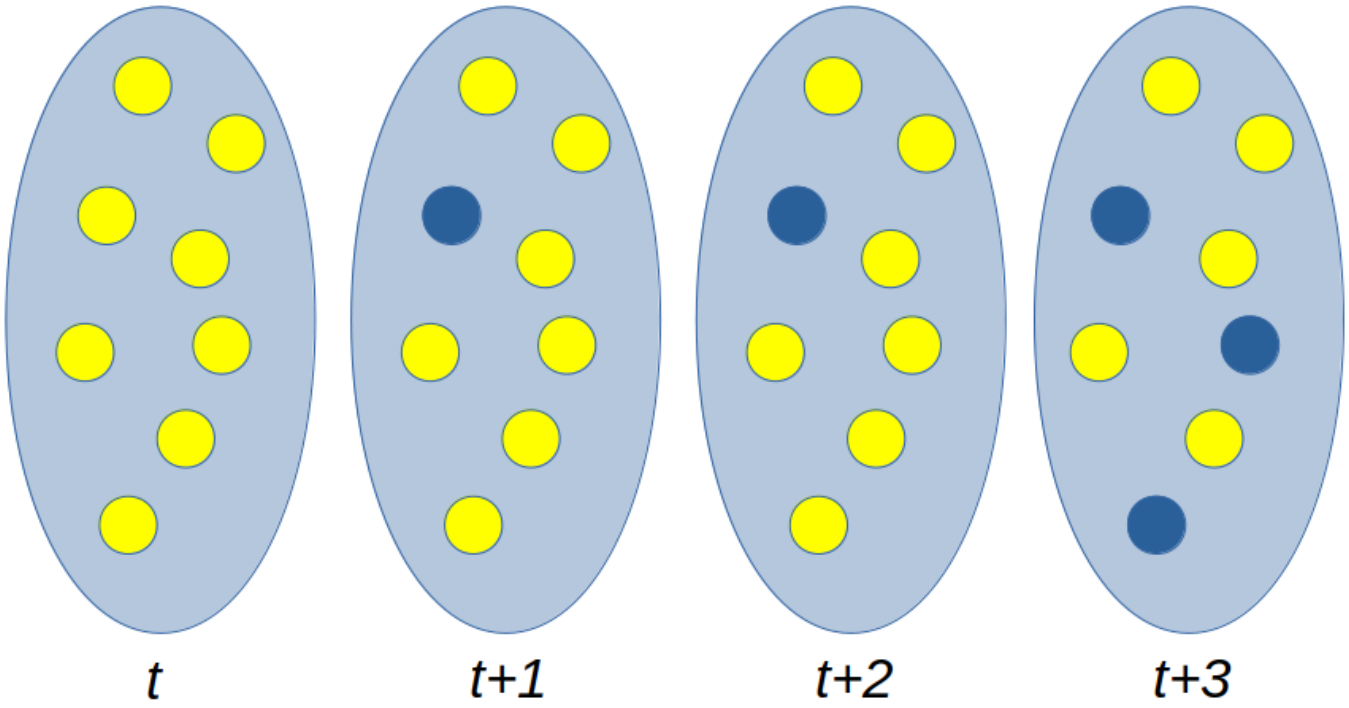
New mutations arise to produce new genetic variation that genetic drift can act on:

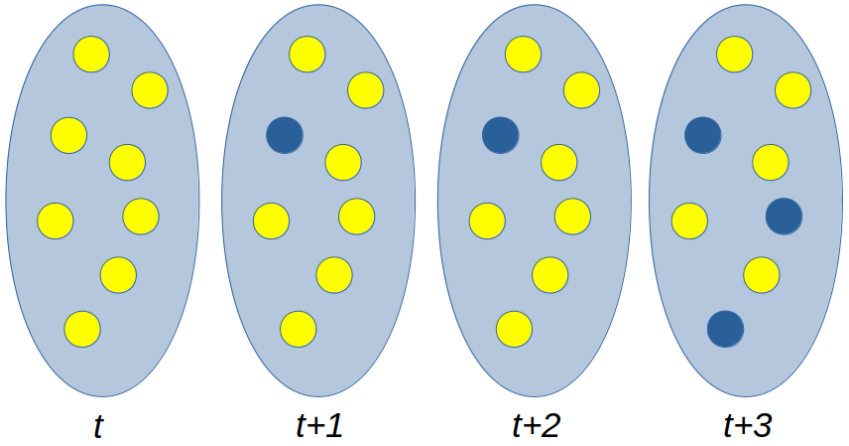
- deletions,
- insertions,
- translocations,
- point mutations.

Any of these mutations can be represented with a di-allelic model (e.g. presence/absence) if we assume that multiple mutations cannot occur in the same location*.

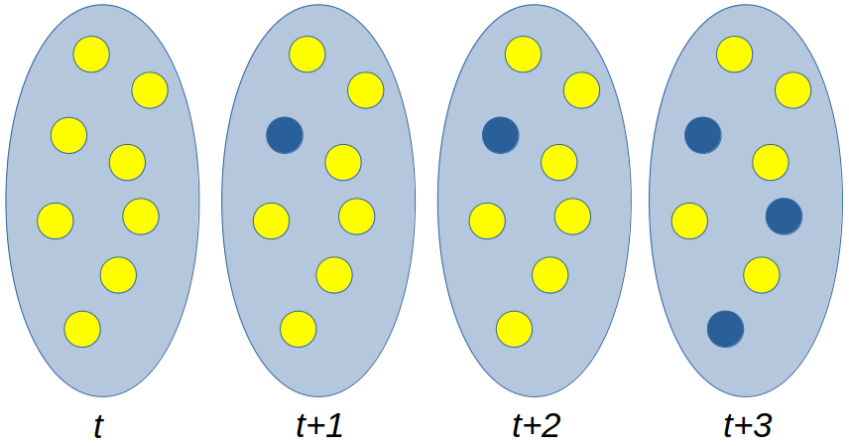
* This assumption is also called the *infinite site model.*

Assume that the yellow *a* allele in each individual randomly mutates to blue *A* with probability μ (called the **mutation rate**) in each generation.



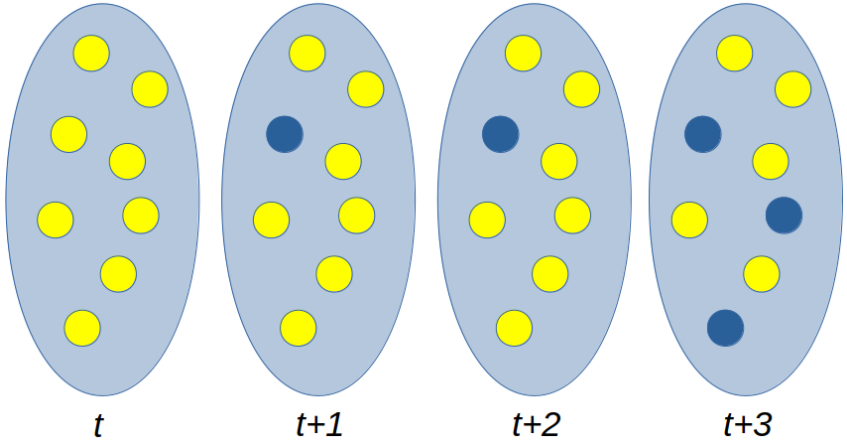


What is the expected allele frequency at the next generation given the current allele frequency and the mutation rate?

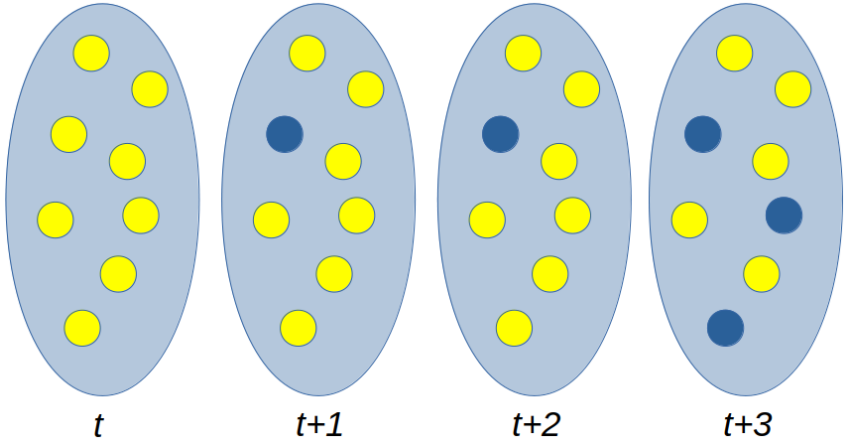


What is the expected allele frequency at the next generation given the current allele frequency and the mutation rate?

In the absence of other forces (e.g. genetic drift and selection), an equilibrium will be reached at:
mutation rate to blue / (mutation rate to blue + mutation rate to yellow)



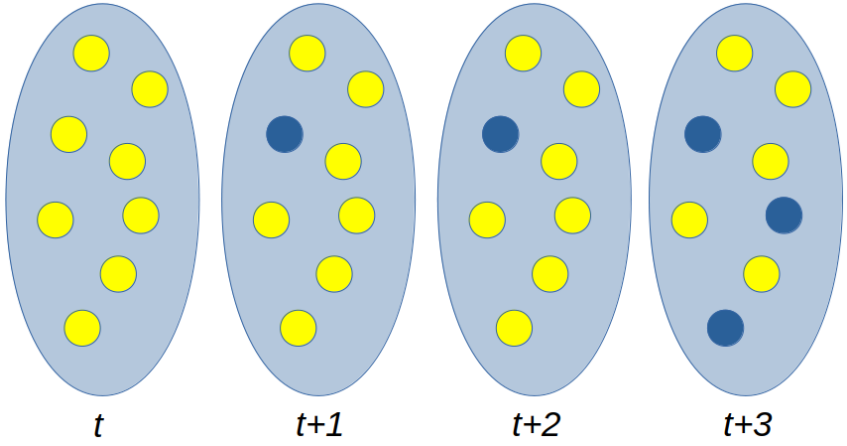
What is the expected allele frequency at the next generation $E[f_A(t+1)]$ given the current allele frequency $f_A(t)$ and the mutation rate μ?



What is the expected allele frequency at the next generation $E[f_A(t+1)]$ given the current allele frequency $f_A(t)$ and the mutation rate μ?

$$E[f_A(t+1)] = f_A(t) + \mu f_a(t) \tag{4}$$

Why?

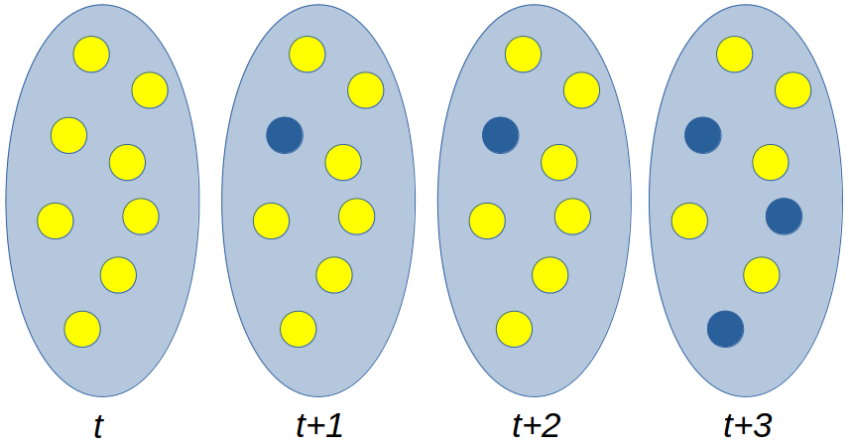


What is the expected allele frequency at the next generation $E[f_A(t+1)]$ given the current allele frequency $f_A(t)$ and the mutation rate μ?

$$E[f_A(t+1)] = f_A(t) + \mu f_a(t) \tag{4}$$

Why?

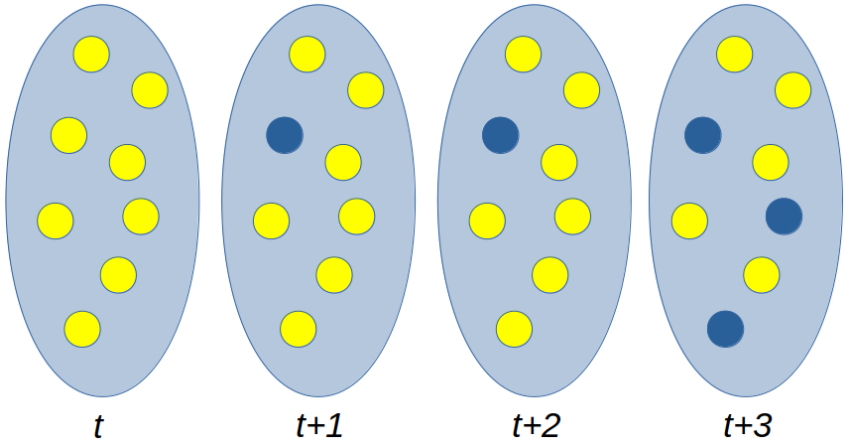
f_A will increase by the number of alleles a that mutate to A. This number is given by the rate μ times the frequency of a alleles to be mutated to A.



If mutations occur in both directions, e.g. mutations occur at rate $\mu_{a \to A}$ from a to A and $\mu_{A \to a}$ from A to a, then

$$E[f_A(t+1)] = (1 - \mu_{A \to a}) f_A(t) + \mu_{a \to A} f_a(t) \tag{5}$$

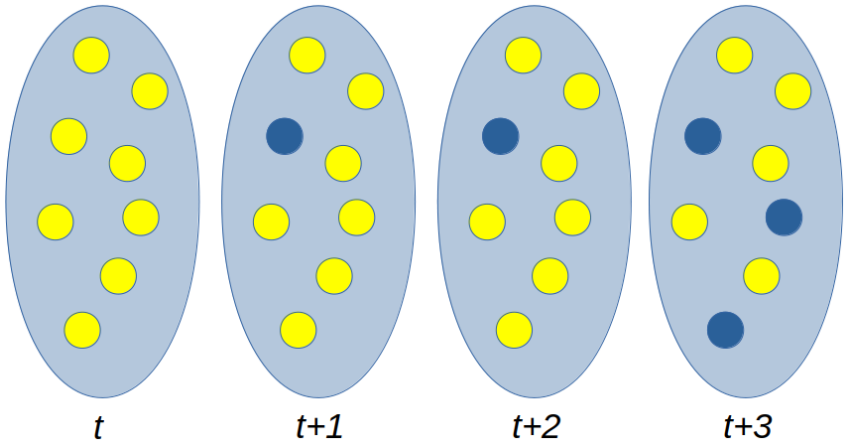
Why?



If mutations occur in both directions, e.g. mutations occur at rate $\mu_{a\to A}$ from a to A and $\mu_{A\to a}$ from A to a, then

$$E[f_A(t+1)] = (1-\mu_{A\to a})f_A(t) + \mu_{a\to A}f_a(t) \qquad (5)$$

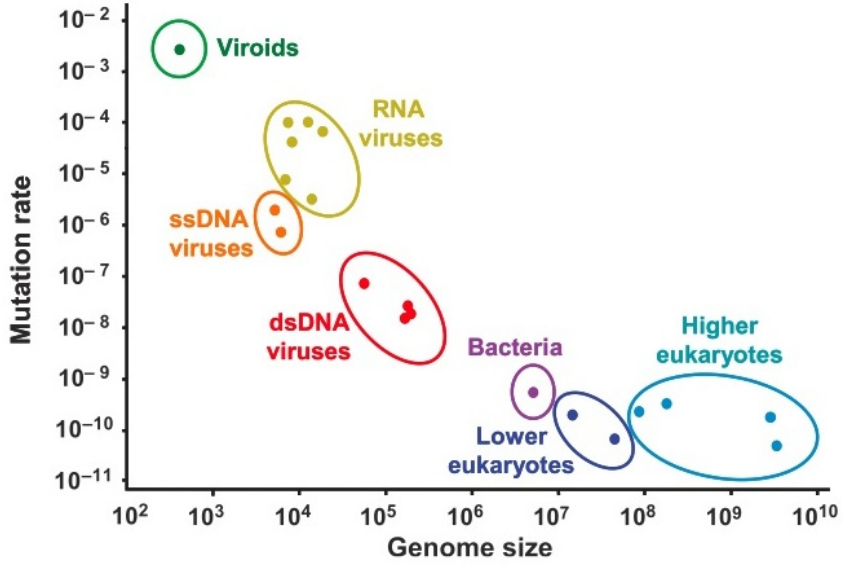
Why? f_A will increase by the number of alleles a that mutate to A but decrease by the number of alleles A that mutate to a.



In the absence of other forces (e.g. genetic drift and selection), we can show that an equilibrium will eventually be established at:

$$f_A = \frac{\mu_{a \to A}}{\mu_{a \to A} + \mu_{A \to a}} \tag{6}$$

Mutation rate



It is important to note that:

- mutation is a weak force in higher organisms,
- with no genetic drift (and selection), it takes a long time for the allele frequency to reach equilibrium,
- we can often ignore recurrent mutations.

What is the effect of

- selection
- mutation
- **gene flow**

on the change of allele frequency?

Population subdivision

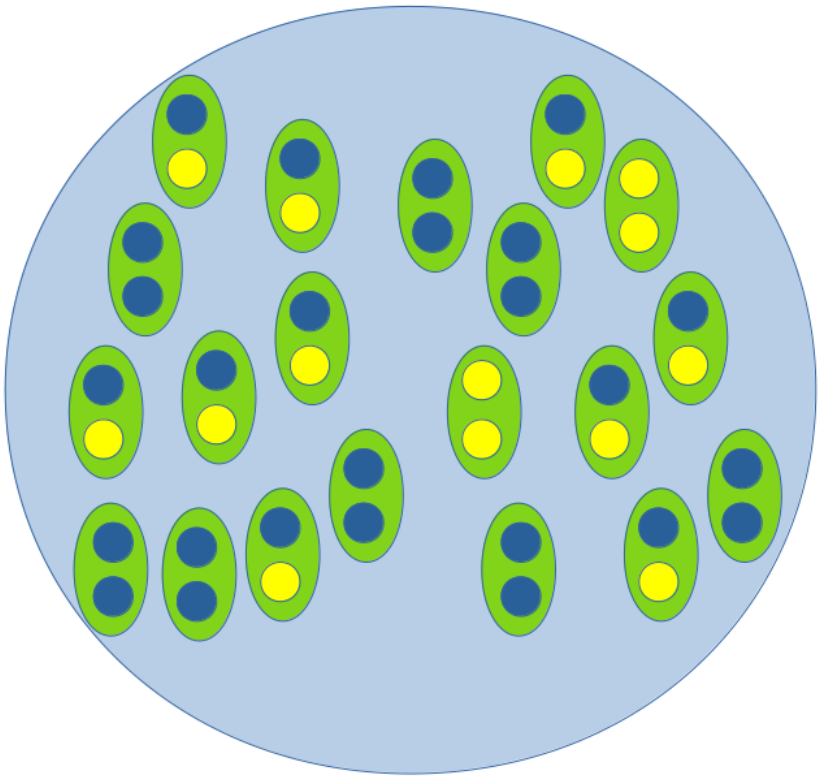
There is population subdivision, or **structure**, when the population is not randomly mating because of geographic or social structure.

Population subdivision is important to

- understand the effects of drift and natural selection,
- plan conservation strategies for rare or endangered species.

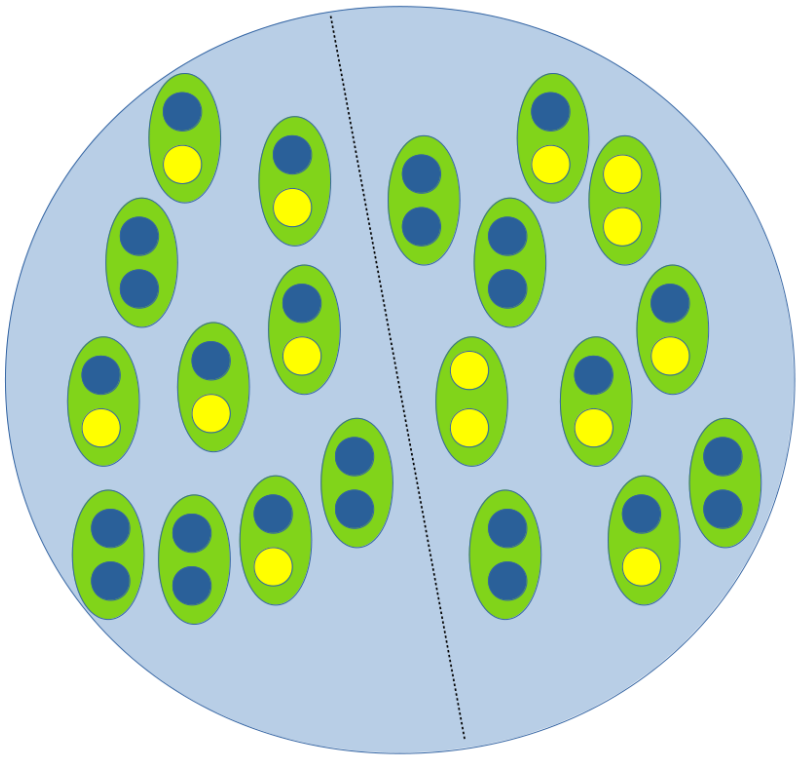
Population subdivision

Let's assume we have a population comprising of a certain number of individuals (i.e. diploid genotypes)



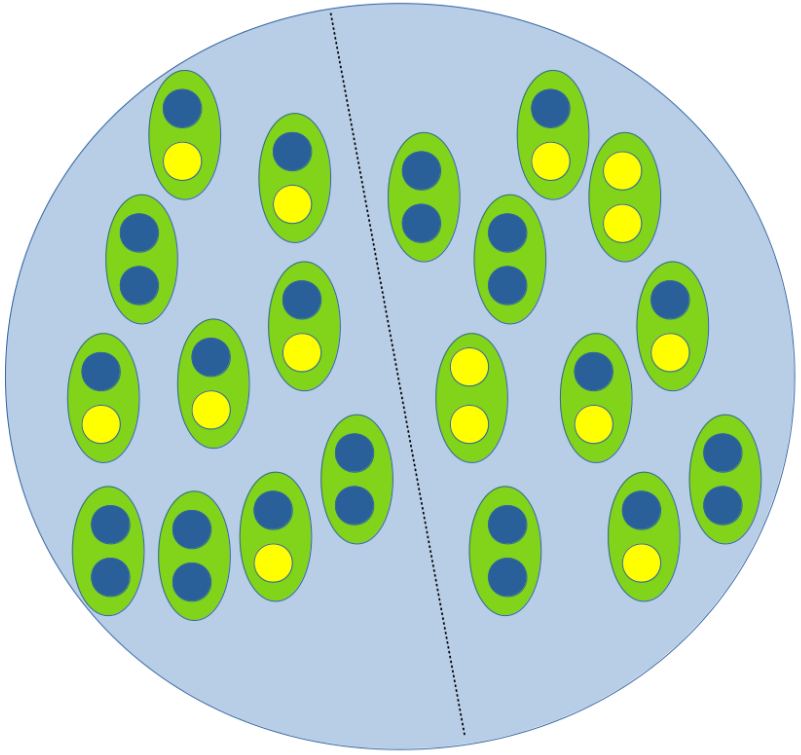
Population subdivision

At some point in time, a geographical/social barrier (dashed line in the figure) may prevent mating between individuals belonging to the two different groups/regions (left and right of the dashed line).



Population subdivision

The two populations will experience separate genetic drifts and their allele frequency will change accordingly.



What if some individuals can move from one population to another? What will the effect on the allele frequency be?

The model of genetic drift can be extended to include the effect of migration (and therefore of gene flow) and the allele frequency will change accordingly.

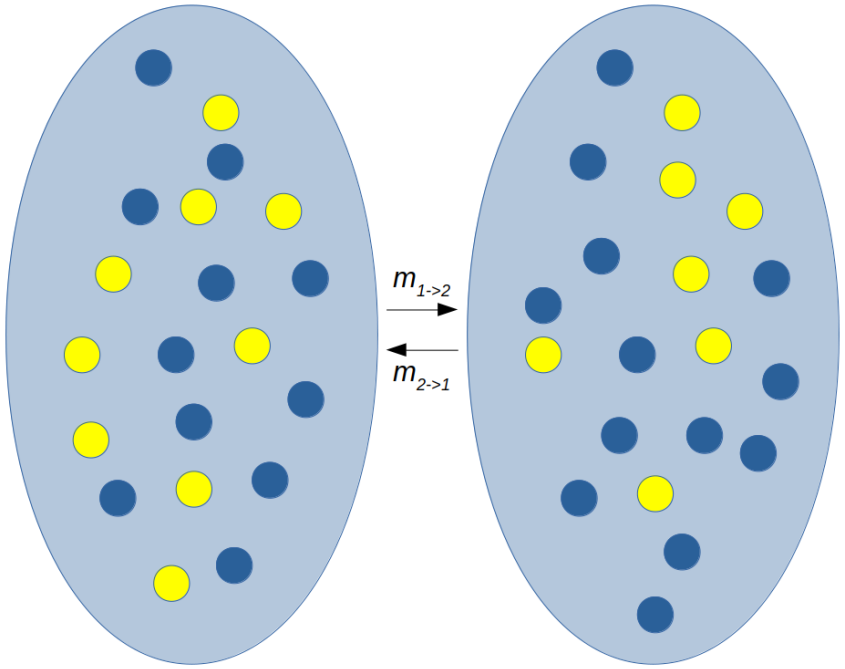
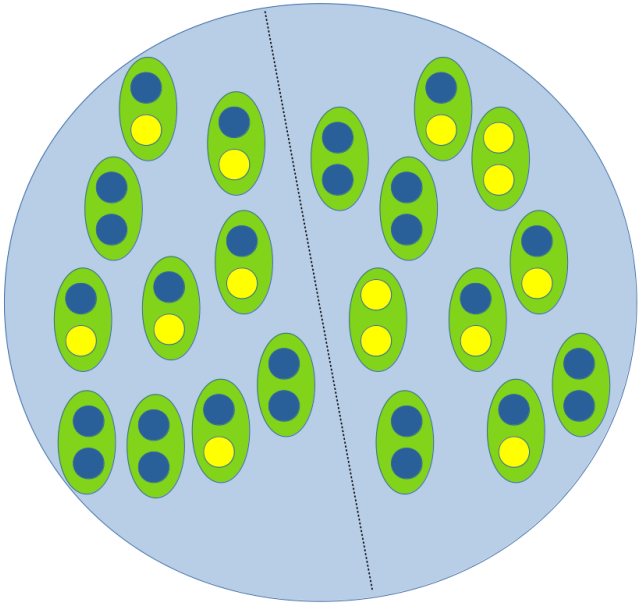


Figure 16: An individual from one population is replaced with an individual from the other with probability m (migration rate).

The model of genetic drift can include gene flow.



jupyter-notebook: subdivision

Intended Learning Outcomes

In this session you have learnt

- to describe all different types of natural selection,
- to appreciate the effect of novel mutations on allele frequency,
- to understand the concept of gene flow.