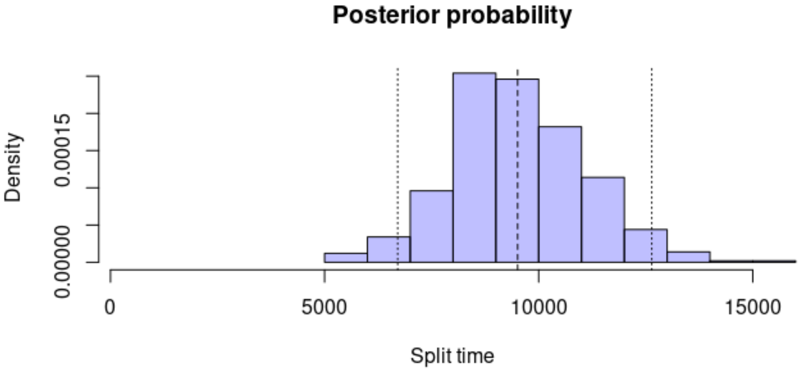
Queen Mary
University of London

Physalia ML

Day1a: introduction to population genetics

Matteo Fumagalli

Teaser


Posterior probability
Density
0.00000
0.00015
0
5000
10000
15000
Split time

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

Gene

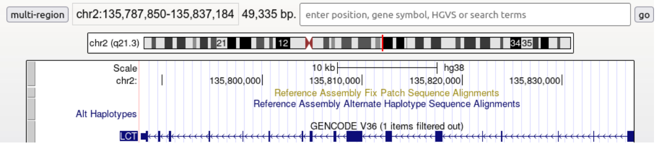
A **gene** can be defined as:

- 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*".



multi-region
chr2:135,787,850-135,837,184
49,335 bp.
enter position, gene symbol, HGVS or search terms
go
chr2 (q21.3)
21
12
34
35
Scale
10 kb
hg38
chr2:
135,800,000
135,810,000
135,820,000
135,830,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
Alt Haplotypes
GENCODE V36 (1 items filtered out)
LCT

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/

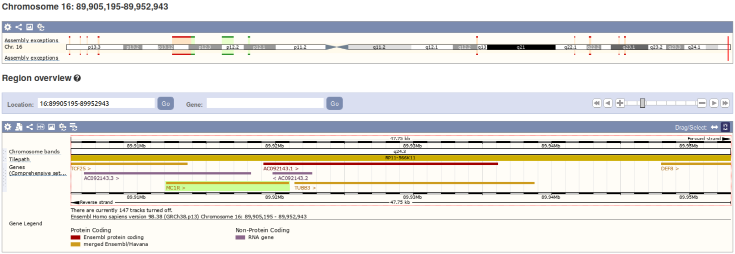
Types of genetic data

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)



Chromosome 16: 89,905,195-89,952,943

Region overview

Location: 16:89905195-89952943 Go

Gene: Go

There are currently 147 tracks turned off.

Ensembl Homo sapiens version 98.38 (GRCh38.p13) Chromosome 16: 89,905,195 - 89,952,943

Gene Legend

Protein Coding
🟥 Ensembl protein coding
🟧 merged Ensembl/Havana

Non-Protein Coding
🟪 RNA gene

Figure 2: *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.

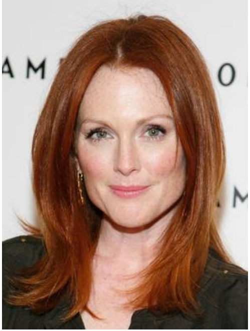


Figure 3: Julianne Moore

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.

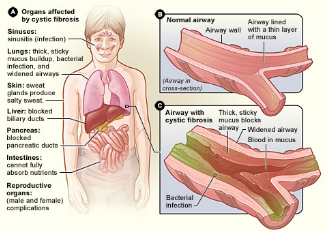

A Organs affected by cystic fibrosis
Sinuses: sinusitis (infection)
Lungs: thick, sticky mucus buildup, bacterial infection, and widened airways
Skin: sweat glands produce salty sweat.
Liver: blocked biliary ducts
Pancreas: blocked pancreatic ducts
Intestines: cannot fully absorb nutrients
Reproductive organs: (male and female) complications
B Normal airway
Airway wall
Airway lined with a thin layer of mucus
(Airway in cross-section)
C Airway with cystic fibrosis
Thick, sticky mucus blocks airway
Widened airway
Blood in mucus
Bacterial infection

Figure 4: 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.

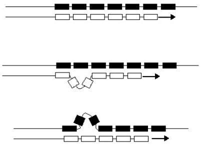


Figure 5: DNA replication errors

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.

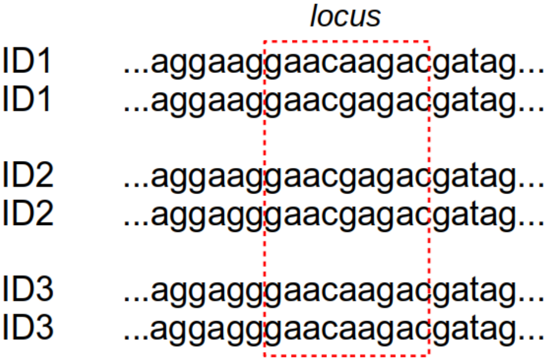
Example

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



locus

```
ID1    ...aggaaggaacaagacgatag...
ID1    ...aggaaggaacgagacgatag...

ID2    ...aggaaggaacgagacgatag...
ID2    ...aggagggaacgagacgatag...

ID3    ...aggagggaacaagacgatag...
ID3    ...aggagggaacaagacgatag...
```

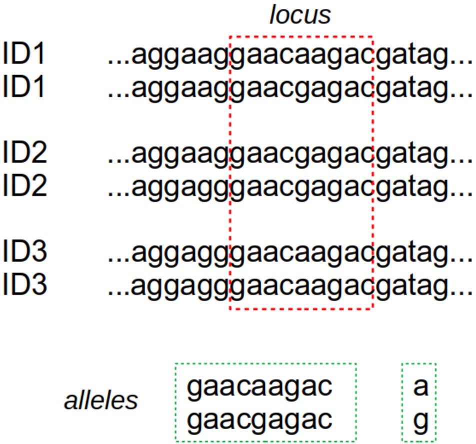
Figure 6: A *locus* of nine base pairs (bp).

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



locus

ID1 ...aggaaggaacaagacgatag...
ID1 ...aggaaggaacgagacgatag...

ID2 ...aggaaggaacgagacgatag...
ID2 ...aggagggaacgagacgatag...

ID3 ...aggagggaacaagacgatag...
ID3 ...aggagggaacaagacgatag...

alleles gaacaagac a
gaacgagac g

Figure 7: 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

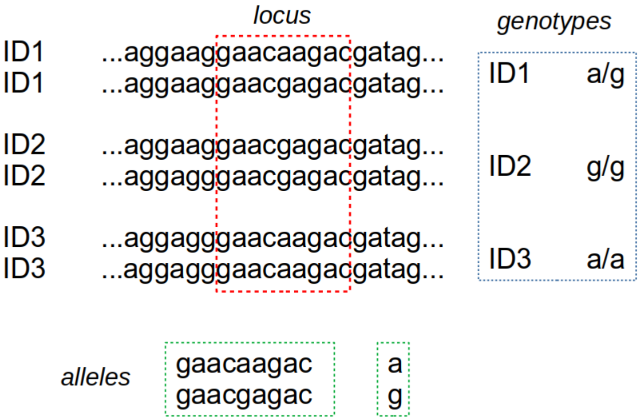
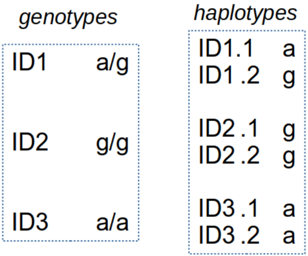

locus
genotypes
ID1 ...aggaaggaacaagacgatag...
ID1 ...aggaaggaacgagacgatag...
ID2 ...aggaaggaacgagacgatag...
ID2 ...aggagggaacgagacgatag...
ID3 ...aggagggaacaagacgatag...
ID3 ...aggagggaacaagacgatag...
ID1 a/g
ID2 g/g
ID3 a/a
alleles
gaacaagac
gaacgagac
a
g

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

Haplotype

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



genotypes

ID1	a/g
ID2	g/g
ID3	a/a

haplotypes

ID1.1	a
ID1 .2	g
ID2 .1	g
ID2 .2	g
ID3 .1	a
ID3 .2	a

Figure 9: 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

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?

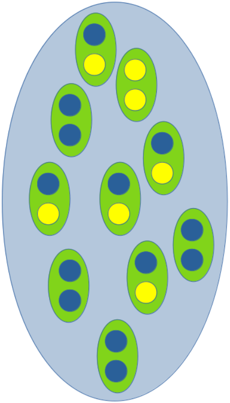
The simplest definition of evolution is "a change in allele frequencies over time".

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).

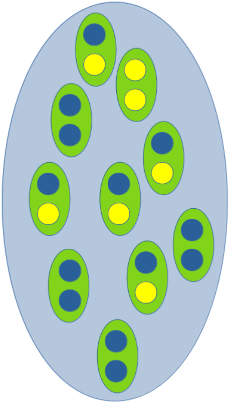


What are the allele frequencies?

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).



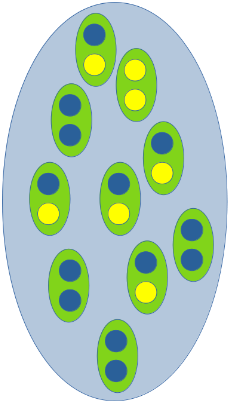
What are the allele frequencies?

$f_A = 7/20 = 0.35$
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).

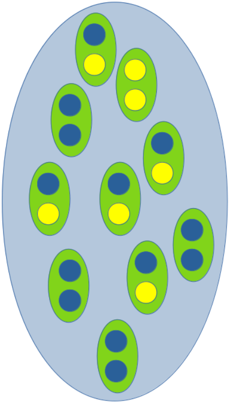


What are the allele frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$
because we have 13 a alleles out of 20 in total.

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).

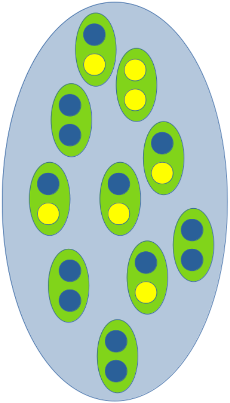


What are the allele frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$
because we have 13 a alleles out of 20 in total.

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).



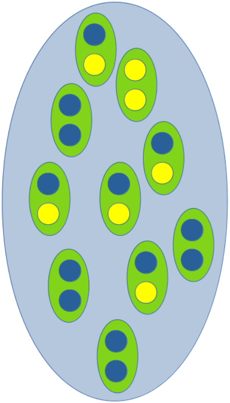
What are the **allele** frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$

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).



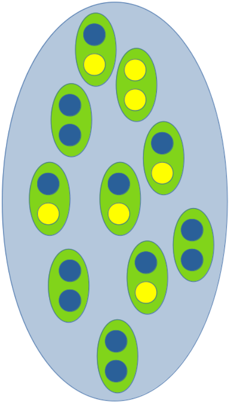
What are the **allele** frequencies?
$f_A = 7/20 = 0.35$
$f_a = 13/20 = 0.65$

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?

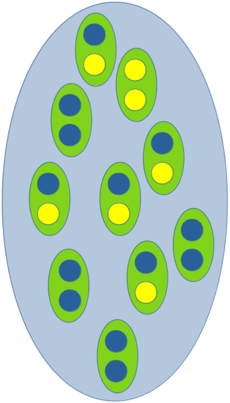
$f_A = 7/20$

$f_a = 13/20$

$f_{AA} =$

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).



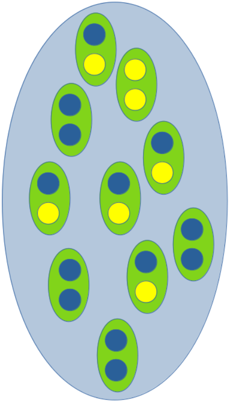
What are the allele and genotype frequencies?

$f_A = 7/20$
$f_a = 13/20$

$f_{AA} = 1/10$
$f_{Aa} = 5/10$
$f_{aa} = 4/10$

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.

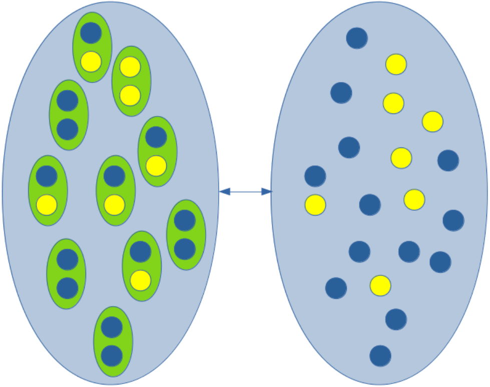
Alleles to genotypes

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.

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

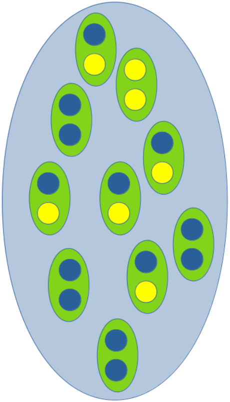
Queen Mary
University of London

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.

Allele frequencies in time



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.

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

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*?

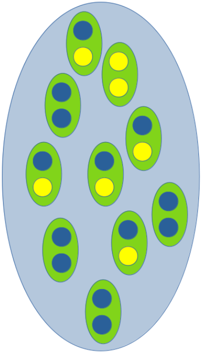
* sometimes also called *genetic draft*

Genetic drift

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

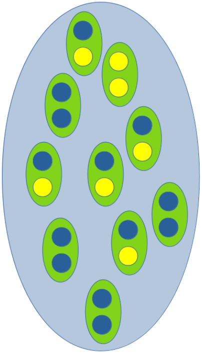
- It describes the process by which allele frequencies change over time due to the effect of random sampling.
- It occurs as a consequence of finite population size.

Genetic drift



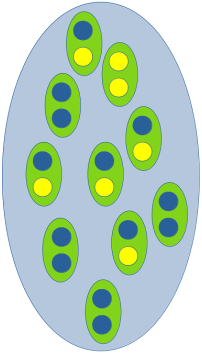
- Some individuals leave many offspring, others fewer, other none.

Genetic drift



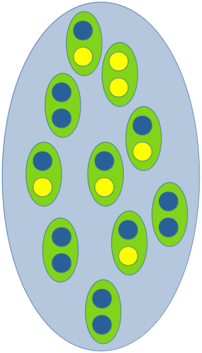
- Some individuals leave many offspring, others fewer, other none.
- Heterozygous individuals will randomly transmit allele *A* or *a*.

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.

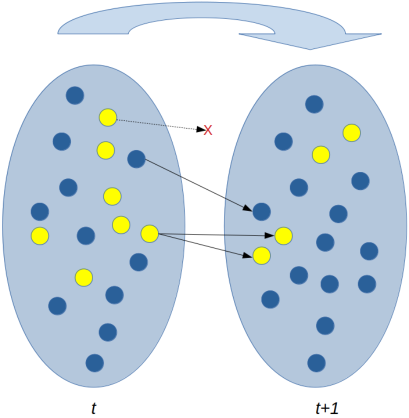


Figure 10: Two generations of genetic drift.

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

note to lecturer: jupyter-notebook: drift (1)

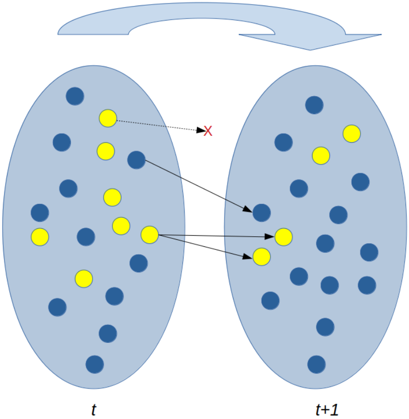


Figure 11: Two generations of genetic drift.

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?

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

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)

- 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)

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.

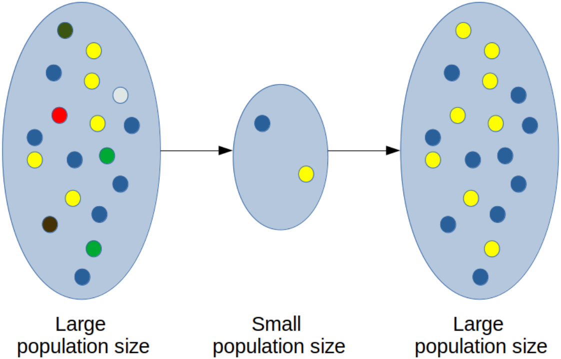
Genetic drift works much faster in small populations than in large populations.

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**.

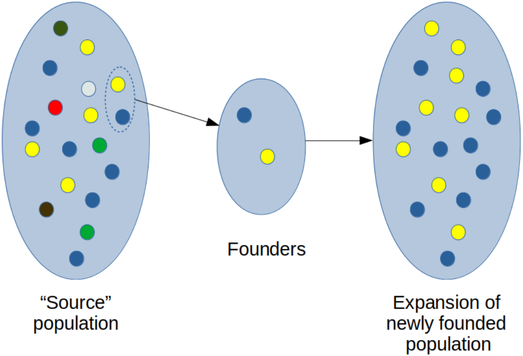
Bottleneck

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.

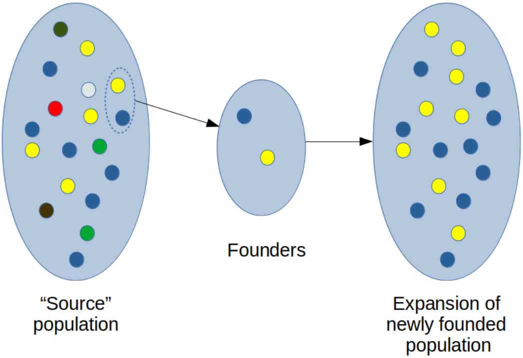


Founder effect

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.

Population subdivision

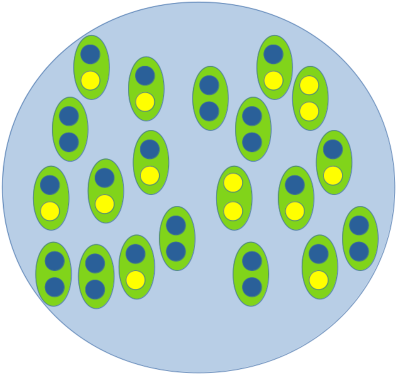
There is population subdivision, or **structure**, when the population is not randomly mating because of geographic or social structure.

Population subdivision is important to

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

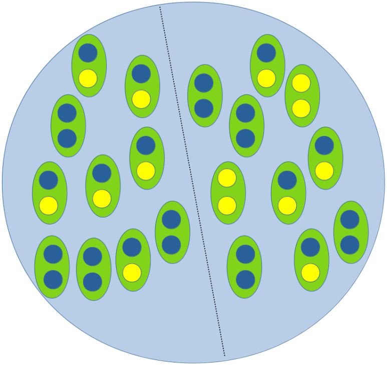
Population subdivision

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



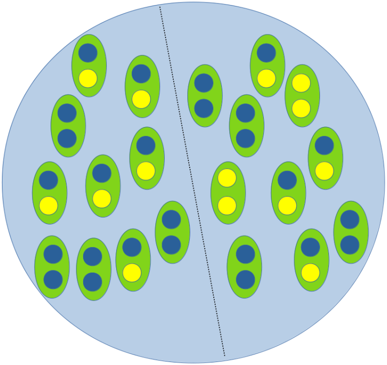
Population subdivision

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.

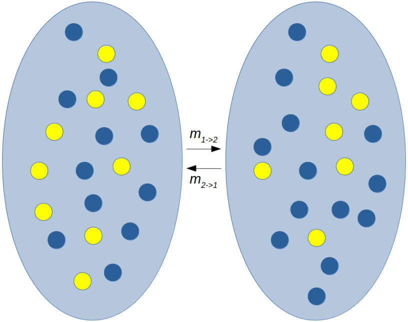
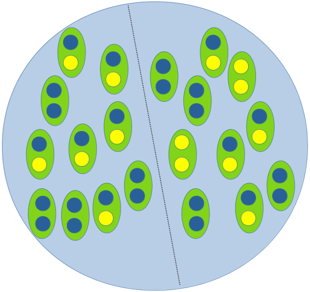


Figure 12: An individual from one population is replaced with an individual from the other with probability m (migration rate).

The model of genetic drift can include gene flow.



Intended Learning Outcomes

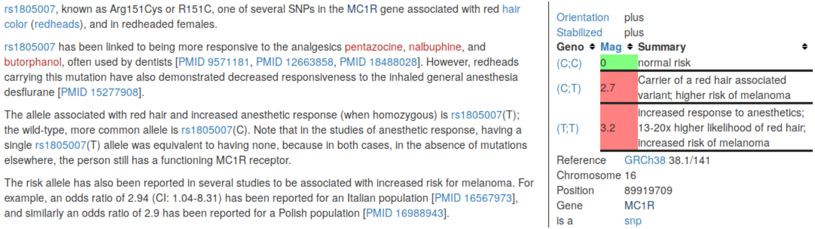
In this session you have learnt

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

Queen Mary
University of London

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 13: 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*?