

# **Week 6 - Genomics and Bioinformatics**

**Alex Mas Sandoval**

**Imperial College  
London**

# Introduction

## Introduction

population genetics



Robin Hammond



Robin Hammond

# 3.2 billion nucleotides





10 million SNPs



# **Population genetics**

**Describe the genetic diversity within and between populations**

**Explain the origin of the observed diversity**

Demography

Natural Selection

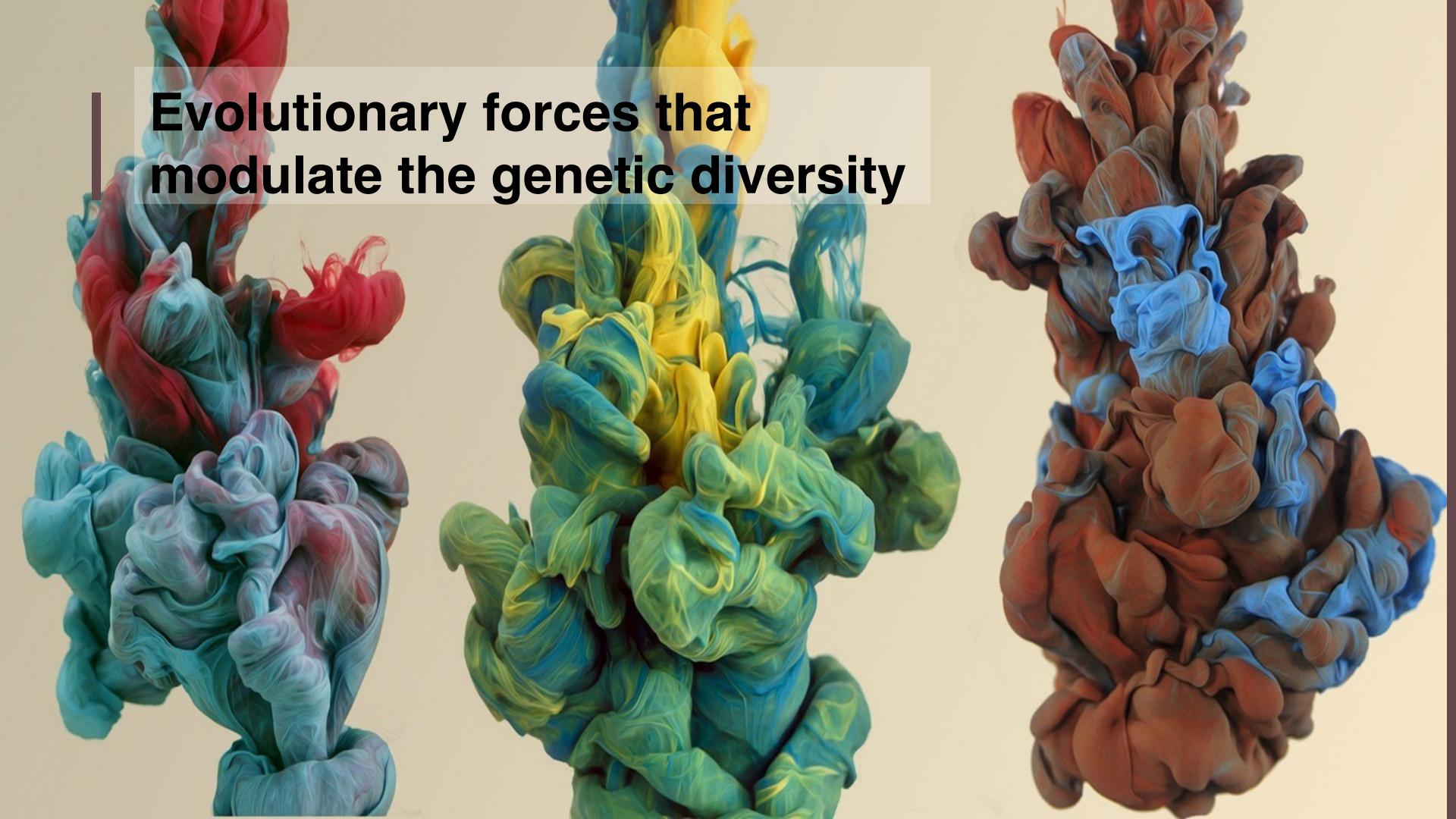
# Population



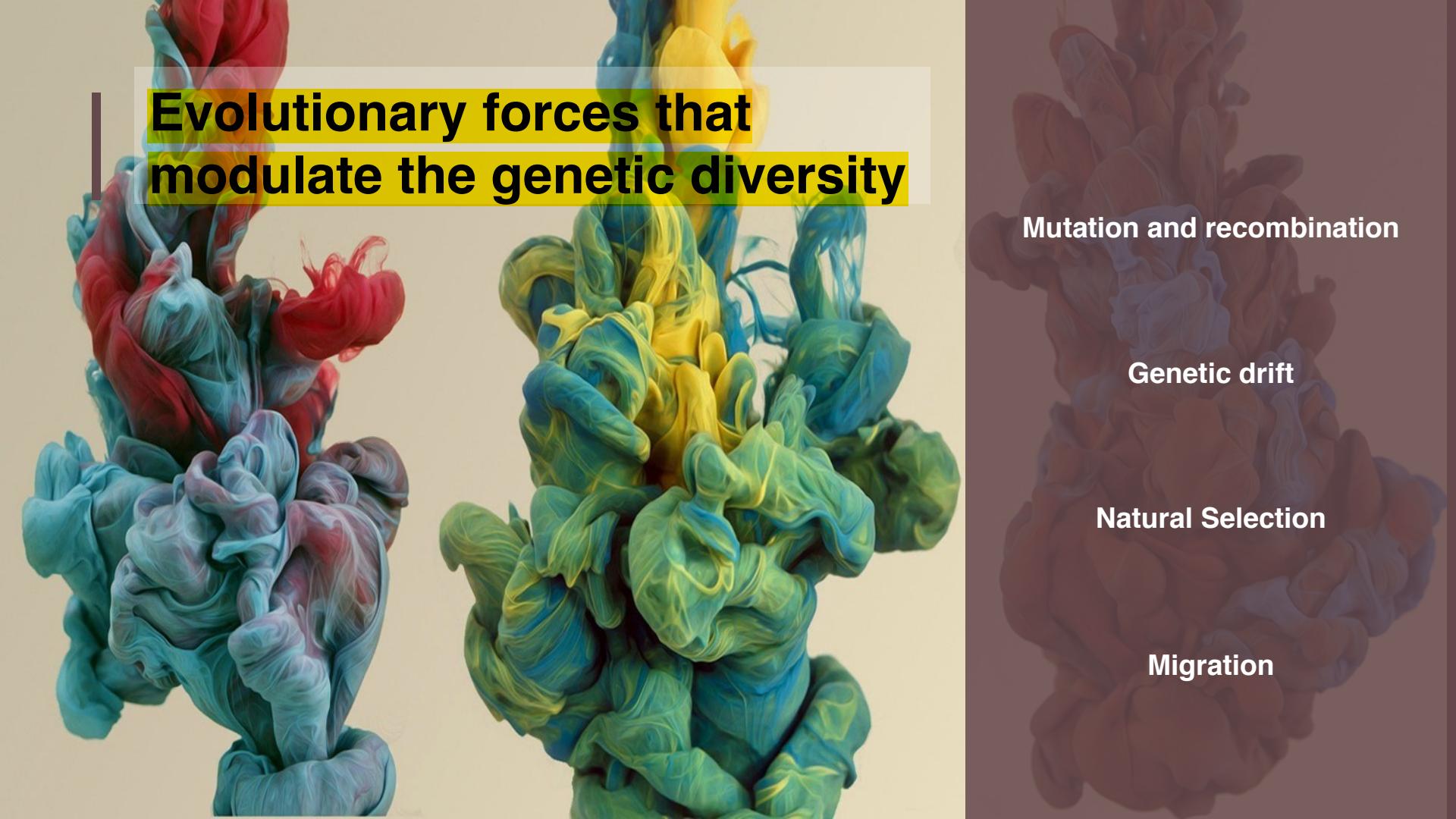
# **Introduction**

**Human genetic diversity**

**Evolutionary forces that modulate the genetic diversity**

The background of the slide features three abstract, organic shapes composed of swirling, translucent colored ink against a light beige background. The left shape is dominated by red and blue. The middle shape is dominated by yellow, green, and blue. The right shape is dominated by orange, red, and blue.

# Evolutionary forces that modulate the genetic diversity

The background of the slide features two abstract, swirling patterns of ink in water. The left pattern is composed of red, blue, and purple ink, creating a complex, organic shape against a light beige background. The right pattern is composed of yellow, green, and blue ink, also forming a dense, flowing mass. Both patterns have a fluid, painterly quality.

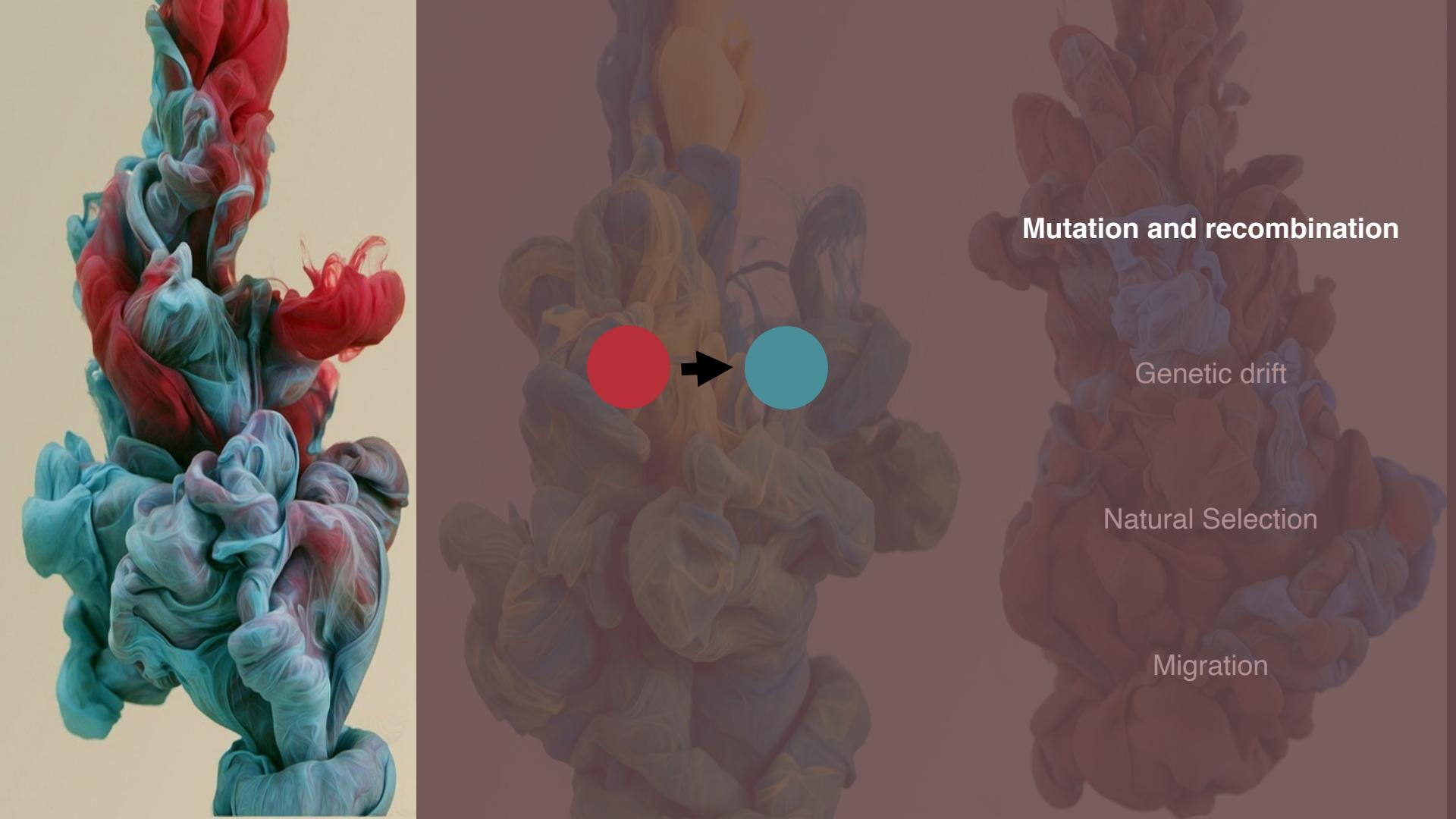
# **Evolutionary forces that modulate the genetic diversity**

**Mutation and recombination**

**Genetic drift**

**Natural Selection**

**Migration**

The background of the image features a complex, organic pattern of swirling, translucent ink droplets in various colors, including red, blue, and yellow, against a light beige or cream-colored surface.

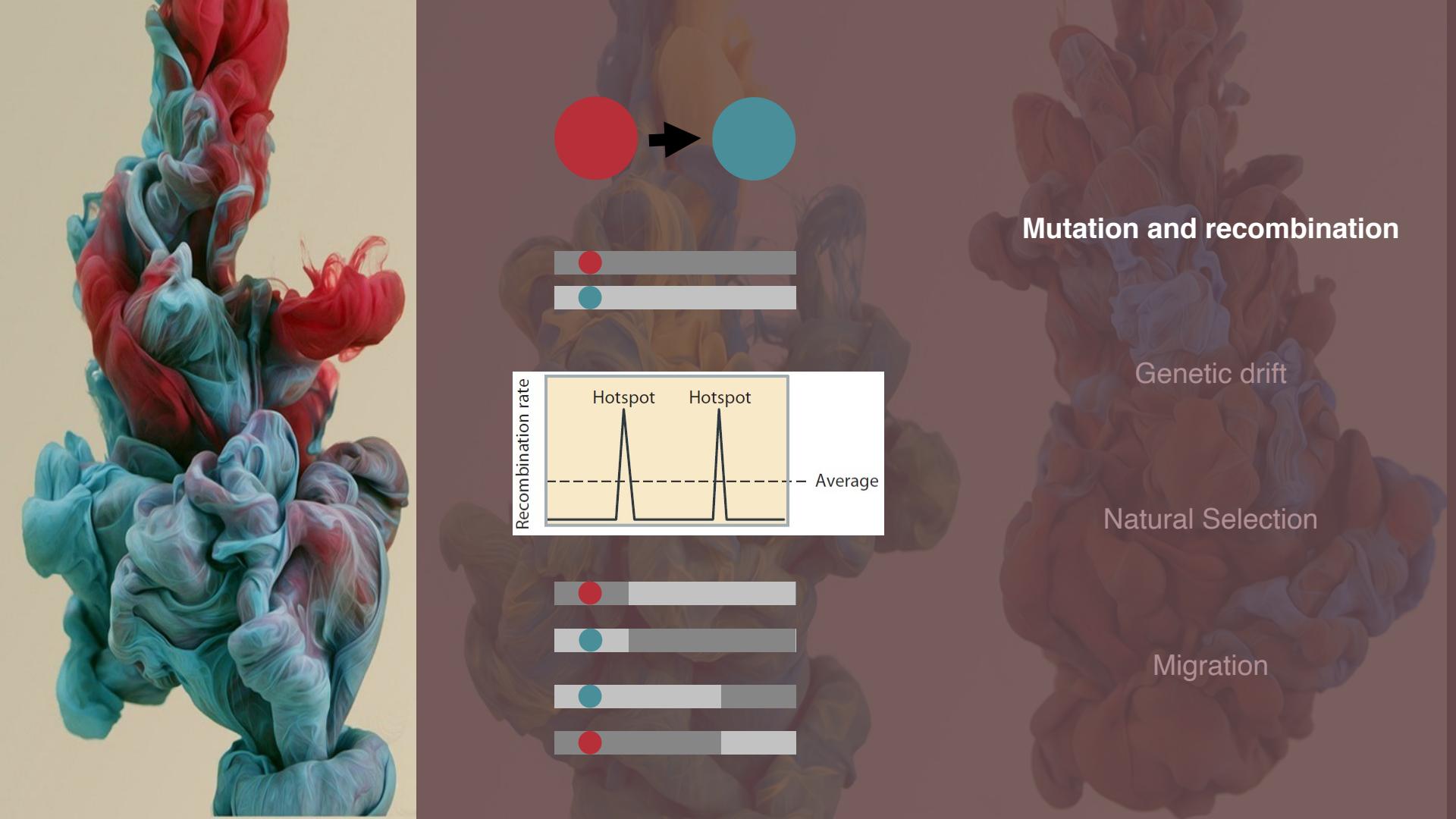
**Mutation and recombination**

Genetic drift

Natural Selection

Migration



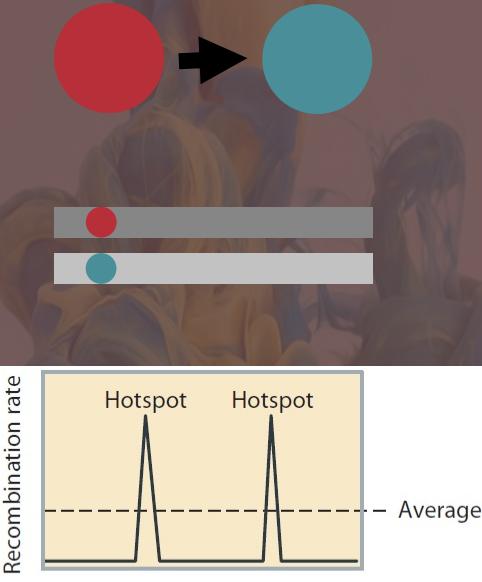


**Mutation and recombination**

Genetic drift

Natural Selection

Migration



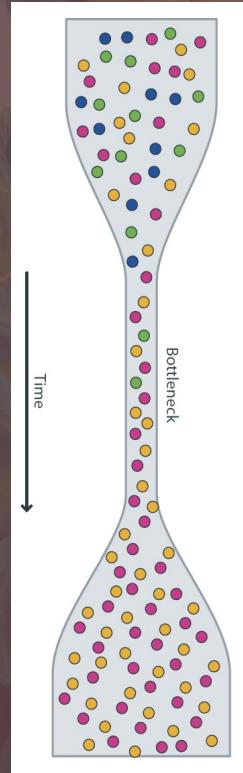


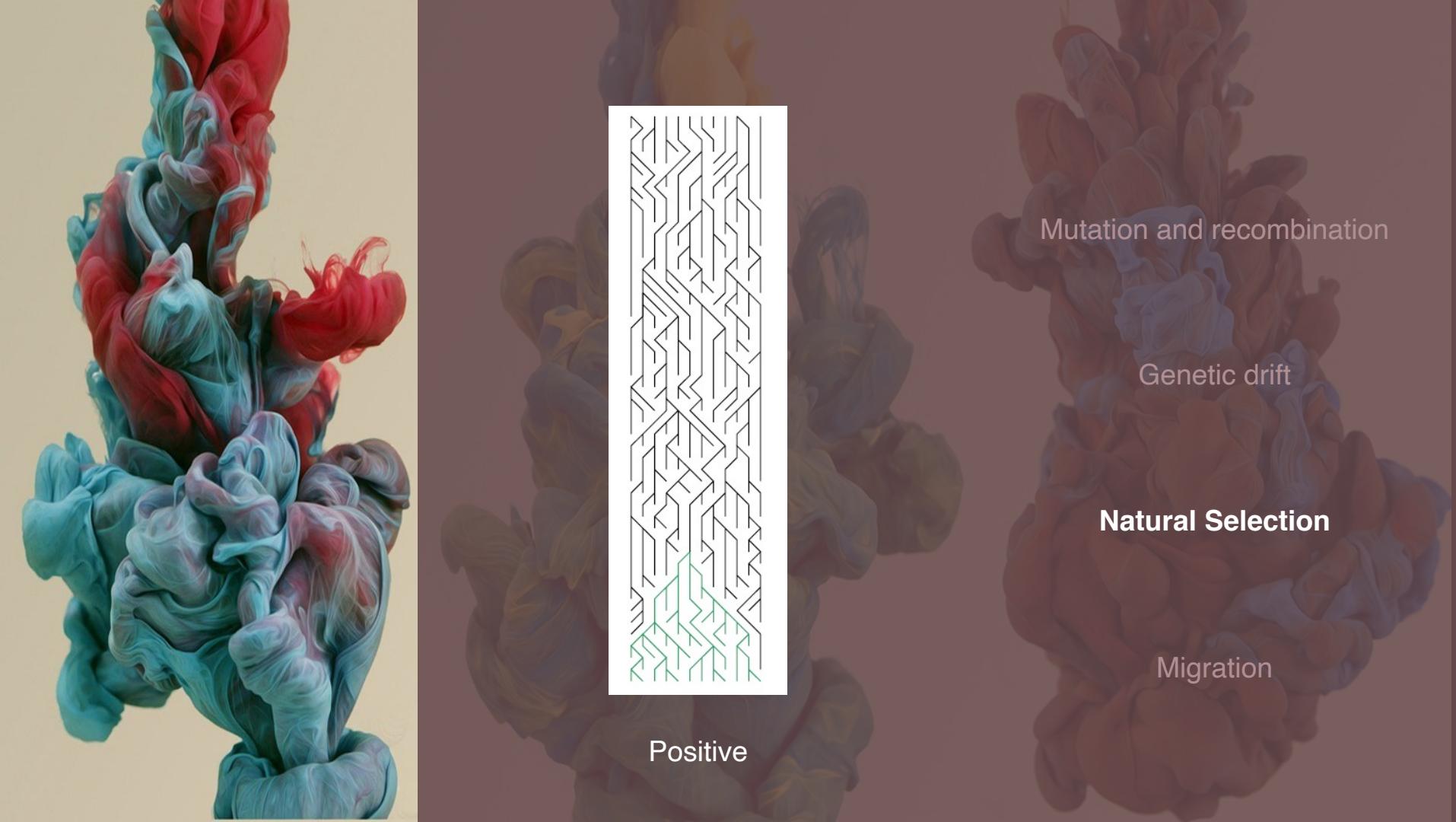
Mutation and recombination

Genetic drift

Natural Selection

Migration





Positive

Mutation and recombination

Genetic drift

**Natural Selection**

Migration



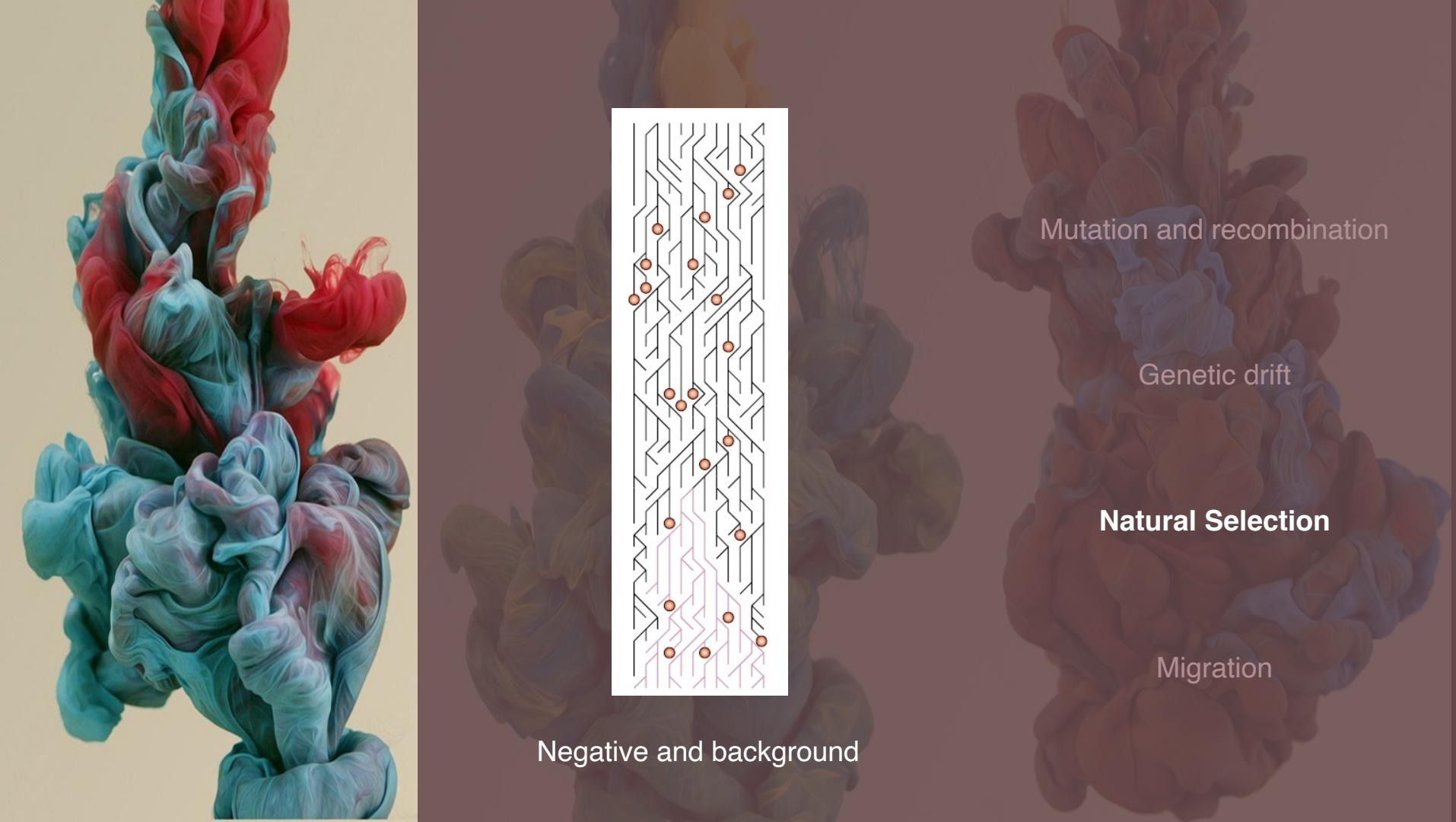
Balancing

Mutation and recombination

Genetic drift

Natural Selection

Migration



Mutation and recombination

Genetic drift

**Natural Selection**

Migration



Mutation and recombination

Genetic drift

Natural Selection

Migration

# Week Objectives

Understand the theory of population genomics and its applications to estimate evolutionary parameters

Describe how genomic data is generated and the challenges associated to its analysis

Implement scripts in R to perform evolutionary inferences from genomic data

# 1. Alleles and genotypes

## Intended Learning Outcomes

In this lecture you will learn

- to describe all different types of genetic data
- to demonstrate the mathematical relationship between allele and genotype frequencies
- to calculate inbreeding coefficients and test for deviation from Hardy-Weinberg Equilibrium from genomic data with R

## Genomic variability

Population genetics assesses the genetic diversity through the analysis of the **polymorphic loci/segregating sites** in the genome, which are genomic regions that can present variability between individuals. They can be:

- single nucleotide polymorphism **SNP**
- insertion/deletion **indel**
- microsatellites

Each of the genetic variants, that a polymorphic locus can present is called an **allele**. Alleles are inherited from parents to offspring

# Single nucleotide polymorphism (SNP)

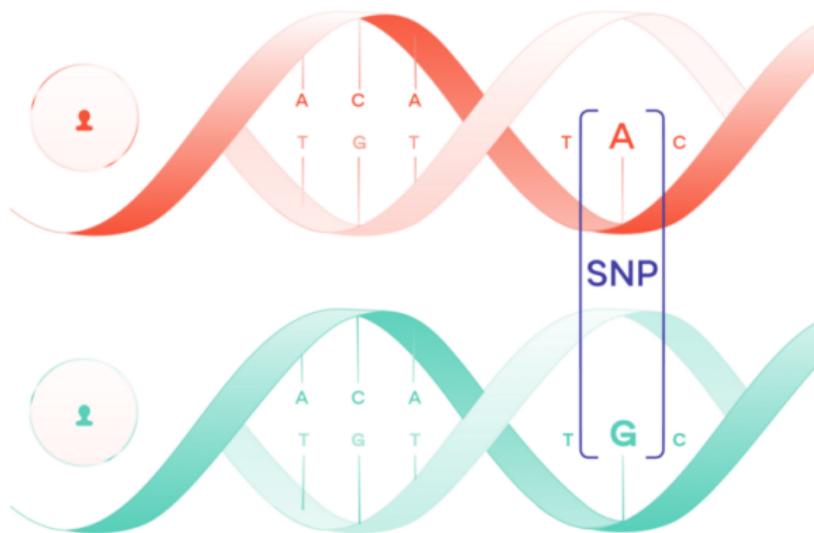


Figure 1: a SNP

# Single nucleotide polymorphism (SNP)

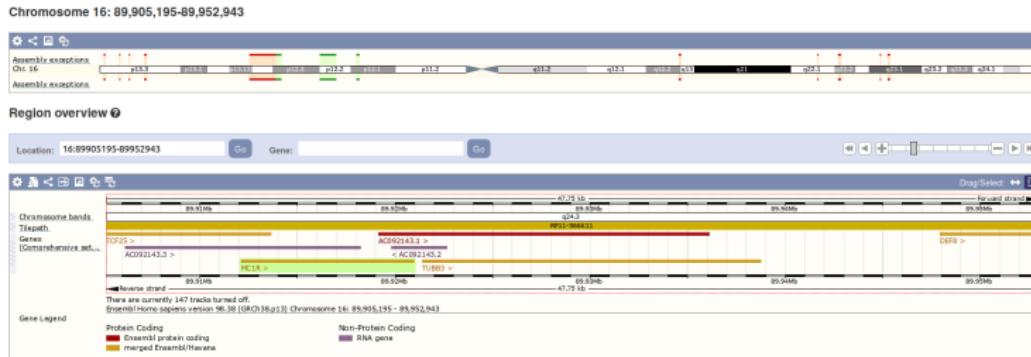


Figure 2: *MC1R* human gene

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.



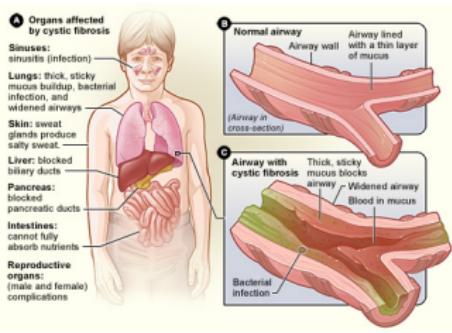
Figure 3: Julianne Moore

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.

# Insertion / deletion (indel)

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



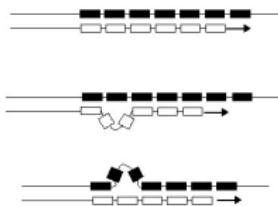
*CFTR* gene codes for a transmembrane protein involved in osmotic balance of cells.

Variant  $\Delta F508$  has a three-base deletion that results in the absence of the 508th amino acid phenylalanine (F).

Figure 4: Cystic fibrosis

## Microsatellites

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



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

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

Figure 5: DNA replication errors

## Terminology

*locus*

ID1	...aggaag <ins>gaacaagacgatag...</ins>
ID1	...aggaag <ins>gaacgagacgatag...</ins>
ID2	...aggaag <ins>gaacgagacgatag...</ins>
ID2	...aggagg <ins>gaacgagacgatag...</ins>
ID3	...aggagg <ins>gaacaagadgatag...</ins>
ID3	...aggagg <ins>gaacaagacgatag...</ins>

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

## Terminology

	<i>locus</i>
ID1	...aggaagg <i>gaacaagac</i> gatag...
ID1	...aggaagg <i>gaacgagac</i> gatag...
ID2	...aggaagg <i>gaacgagac</i> gatag...
ID2	...aggagg <i>gaacgagac</i> gatag...
ID3	...aggagg <i>gaacaaga</i> gatag...
ID3	...aggagg <i>gaacaagac</i> gatag...

alleles

gaacaagac  
gaacgagac

a  
g

## Terminology

		<i>locus</i>	<i>genotypes</i>
ID1	...	aggaaggagaacaaggacgatag...	ID1 a/g
ID1	...	aggaaggagaacgagacgatag...	ID2 g/g
ID2	...	aggaaggagaacgagacgatag...	ID3 a/a
ID2	...	aggaggagaacaaggacgatag...	
ID3	...	aggaggagaacaaggacgatag...	
ID3	...	aggaggagaacaaggacgatag...	

*alleles*

gaacaagac	a
gaacgagac	g

A **genotype** is

- the combination of alleles carried by an individual in a specific locus
- alternatively, the combination of alleles at multiple loci that confer a phenotype

## Terminology

genotypes		haplotypes	
ID1	a/g	ID1.1	a
ID2	g/g	ID1.2	g
ID3	a/a	ID2.1	g
		ID2.2	g
		ID3.1	a
		ID3.2	a

A **haplotype** is

- The allele or alleles, at one or multiple consecutive loci, from each of the homologous chromosomes (Commonly used for all the loci of each of the homologous chromosomes).
- alternatively, the DNA fragment inherited in block due to the lack of recombination between two recombination points

## Terminology

**Ploidy** is the amount of homologous chromosomes of a species.

**Diploid** species have a ploidy of two copies of its chromosomes.

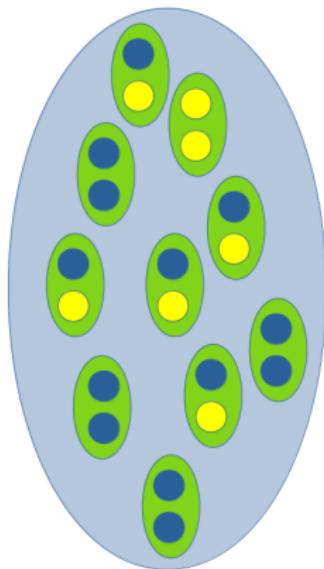
For a collection of  $N$  diploid individuals, there are  $2N$  gene copies at each locus.

As mutations are rare in most organisms, most of polymorphisms are **biallelic**. 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 biallelic model is a valid approximation here.

## Alleles and genotypes

Population of  $N = 10$  individuals,  $2N = 20$  gene copies, and a total of 7 copies of allele A (yellow) and 13 copies of allele a (blue)

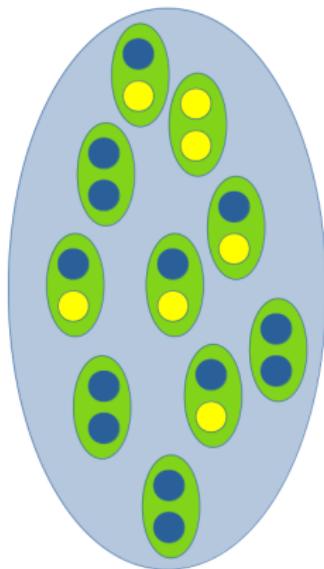


What are the allele and genotype frequencies?

$$f_A =$$

## Alleles and genotypes

Population of  $N = 10$  individuals,  $2N = 20$  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?

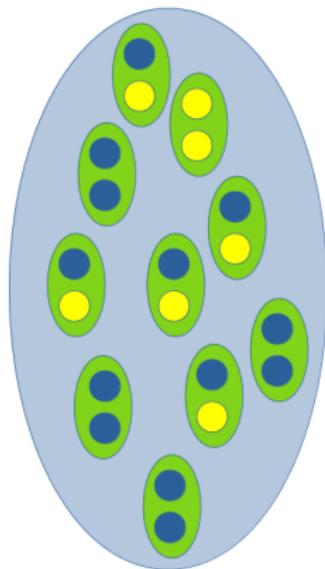
$$f_A = 7/20$$

$$f_a = 13/20$$

$$f_{AA} =$$

## Alleles and genotypes

Population of  $N = 10$  individuals,  $2N = 20$  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?

$$f_A = 7/20$$

$$f_a = 13/20$$

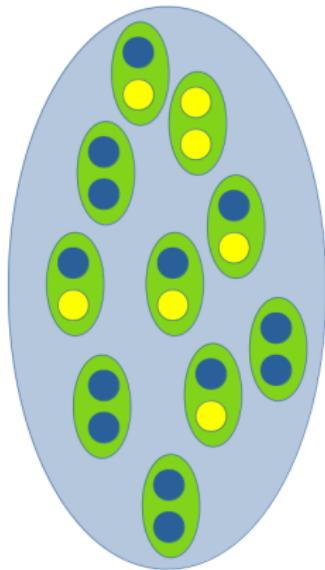
$$f_{AA} = 1/10$$

$$f_{Aa} = 5/10$$

$$f_{aa} = 4/10$$

$AA$  and  $aa$  are homozygous individuals and  $Aa$  are heterozygous individuals.

## Allele frequencies



With  $N$  diploid individuals and  $A$  and  $a$  alleles:

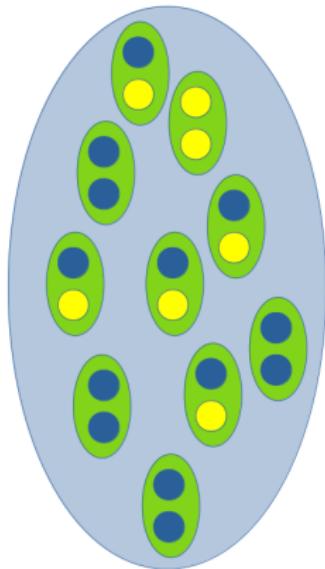
$$f_A = \frac{N_A}{2N} \quad (1)$$

$$f_a = \frac{N_a}{2N} \quad (2)$$

where  $N_A$  and  $N_a$  are number of  $A$  and  $a$  alleles segregating in the population, respectively.

Note that  $f_A + f_a = 1$ .

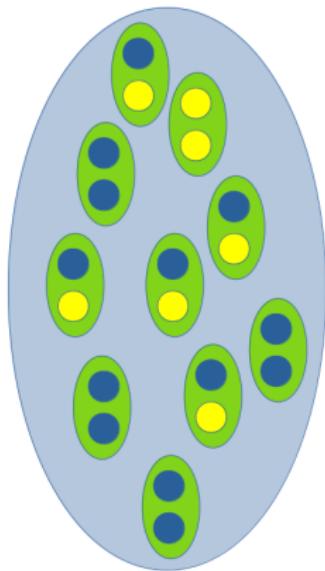
## Allele frequencies



Much population genetics focuses 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.

## Genotype frequencies



In a biallelic locus there are three possible genotypes:  $AA$ ,  $Aa$  and  $aa$ .

$$f_{AA} = \frac{N_{AA}}{N} \quad (3)$$

$$f_{Aa} = \frac{N_{Aa}}{N} \quad (4)$$

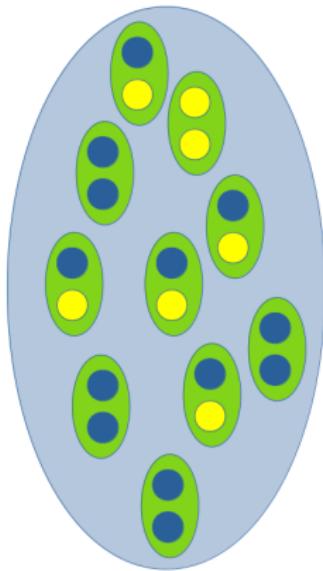
$$f_{aa} = \frac{N_{aa}}{N} \quad (5)$$

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

## Allele and genotype frequencies

Can you calculate allele frequencies from genotype frequencies?

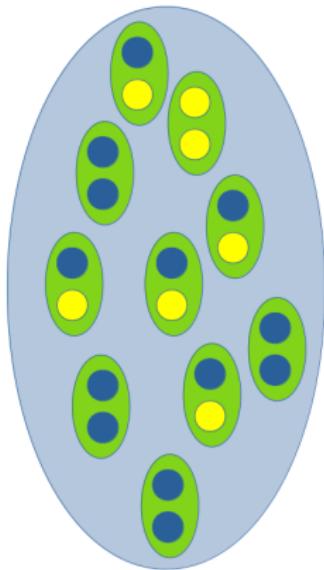
$$f_A =$$



## Allele and genotype frequencies

Can you calculate allele frequencies from genotype frequencies?

$$f_A = \frac{2N_{AA} + N_{Aa}}{2N} = f_{AA} + \frac{f_{Aa}}{2} \quad (6)$$



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

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].

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

Orientation plus  
Stabilized plus  
Reference GRCh38 38.1/141  
Chromosome 16  
Position 89919709  
Gene MC1R  
Is a snp

Figure 7: SNP associated to red hair with alleles C and T

## *MC1R* gene

Assume we obtain a *random* sample of 30 individuals from the UK and find 25 individuals of genotype  $CC$ , 5 individuals with  $CT$  and 0 with  $TT$ .

What are the *estimated* genotype and allele frequencies?

## *MC1R* gene

Assume we obtain a *random* sample of 30 individuals from the UK and find 25 individuals of genotype  $CC$ , 5 individuals with  $CT$  and 0 with  $TT$ .

What are the *estimated* genotype and allele frequencies?

$$f_{CC} = 25/30 = 0.833, f_{CT} = 5/30 = 0.167, f_{TT} = 0/30 = 0$$

## *MC1R* gene

Assume we obtain a *random* sample of 30 individuals from the UK and find 25 individuals of genotype  $CC$ , 5 individuals with  $CT$  and 0 with  $TT$ .

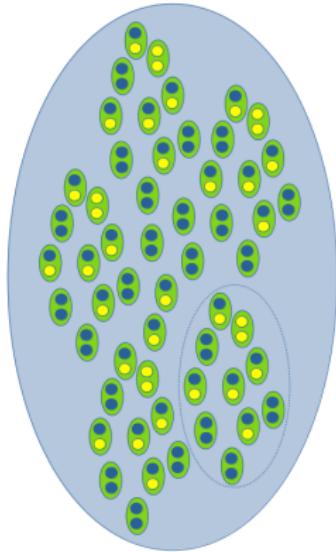
What are the *estimated* genotype and allele frequencies?

$$f_{CC} = 25/30 = 0.833, f_{CT} = 5/30 = 0.167, f_{TT} = 0/30 = 0$$

$$f_C = 0.833 + 0.167/2 = 0.917, f_T = 1 - 0.917 = 0.083$$

Why *estimated*?

## Sample vs. population



We cannot know the true genotype/allele frequency in the entire population but we aim at estimating it from a random sample of individuals.

We calculate the sample allele frequency as a proxy for the unknown population allele frequency.

Figure 8: A random sample of individuals from a population

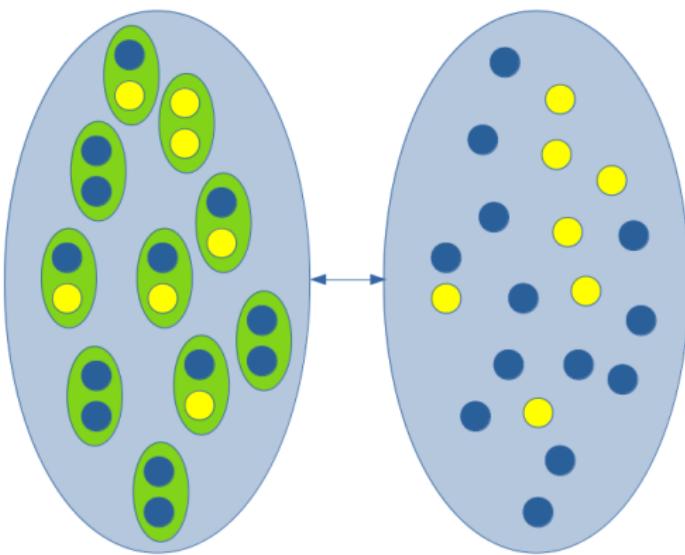
## Alleles to genotypes

We can calculate allele frequencies from genotype frequencies.

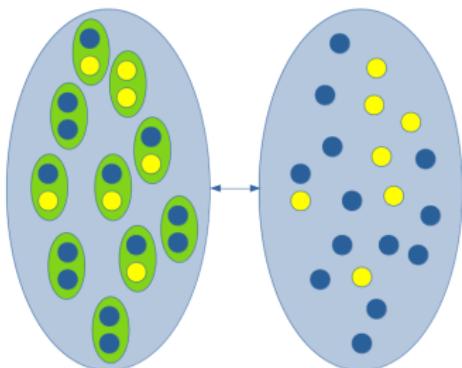
Can we *predict* genotype frequencies from allele frequencies?

e.g. knowing that the frequency of  $T$  in position 478 of the  $MC1R$  gene is 0.08, what proportion of the population is expected to have genotype  $TT$ ?

## Alleles to genotypes



## Alleles to genotypes



### Assumptions

- random mating: individuals have the same probability to mate with each other, independently of the genotype
- males and females have equal allele frequency
- biallelic locus

# Hardy-Weinberg Equilibrium (HWE)

---

Genotype frequencies under HWE

---

Genotype	$AA$	$Aa$	$aa$
Frequency	$f_A^2$	$2f_A f_a$	$f_a^2$

---

## Hardy-Weinberg Equilibrium (HWE)

---

Genotype frequencies under HWE

---

	Female gamete A <i>freq</i> $f_A$	Female gamete a <i>freq</i> $f_a$
Male gamete A <i>freq</i> $f_A$	AA $f_A^2$	Aa $f_A f_a$
Male gamete a <i>freq</i> $f_a$	aA $f_a f_A$	aa $f_a^2$

---

Table 1: Punnett square

## Hardy-Weinberg Equilibrium (HWE)

- expected homozygosity is  $f_A^2 + f_a^2$
- expected heterozygosity is  $2f_A f_a$

## Hardy-Weinberg Equilibrium (HWE)

- expected homozygosity is  $f_A^2 + f_a^2$
- expected heterozygosity is  $2f_A f_a$
- $f_A^2 + 2f_A f_a + f_a^2 = 1$

## Hardy-Weinberg Equilibrium (HWE)

- expected homozygosity is  $f_A^2 + f_a^2$
- expected heterozygosity is  $2f_A f_a$
- $f_A^2 + 2f_A f_a + f_a^2 = 1$
- random mating does not change the allele frequencies in the next generation:  
$$f'_A = f_A^2 + 2f_A f_a / 2 = f_A^2 + 2f_A(1 - f_A) / 2 = f_A$$

## HWE in *MC1R*



Figure 9: Rupert Grint

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

## HWE in *MC1R*



Figure 9: Rupert Grint

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

$$\text{It's } 0.08^2 = 0.0064.$$

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

## Deviations from HWE

**Assortative mating.** Mating not independent from the genotype due to:

# Deviations from HWE

**Assortative mating.** Mating not independent from the genotype due to:

- Population structure

*Geographical barriers*



Figure 10: Galapagos

*Socioeconomic and cultural barriers*



Figure 11: Santa Fe, CDMX - Johny Miller

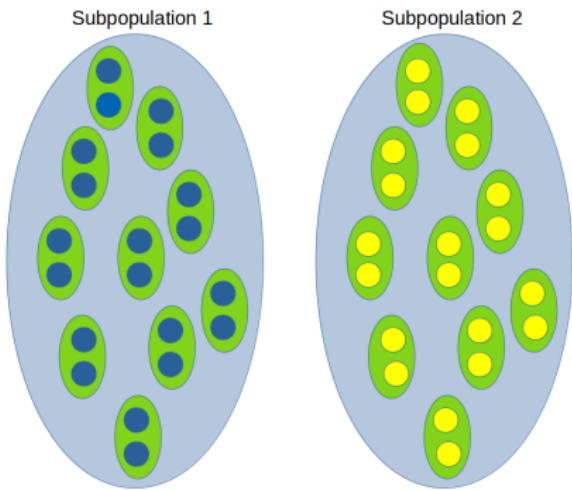
- Active mate choice by similarity



Figure 12: Merrill2019

# Deviations from HWE

## Assortative mating



If there is population subdivision, genotype frequencies deviate from HWE.

# Deviations from HWE

## Natural Selection

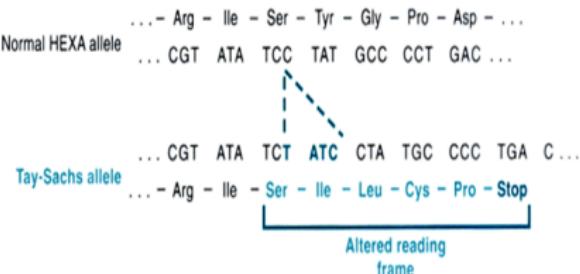


Figure 13: Insertion in *HEXA* gene associated with Tay-Sachs disease

## Deviations from HWE

### 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.

$$F = \frac{2f_A f_a - f_{Aa}}{2f_A f_a} \quad (7)$$

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

## Deviations from HWE

### 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.

$$F = \frac{2f_A f_a - f_{Aa}}{2f_A f_a} \quad (7)$$

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

## Deviations from HWE

$$f_{Aa} = 2f_A f_a (1 - F) \quad (8)$$

- 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?

## Deviations from HWE



Figure 14: Flower of wild oats  
(*Avena fatua*)

Genotype frequencies at one locus are:

$$f_{AA} = 0.58, f_{Aa} = 0.07, f_{aa} = 0.35.$$

- ➊ What is  $F$ , the inbreeding coefficient?

## Deviations from HWE



Figure 14: Flower of wild oats  
(*Avena fatua*)

Genotype frequencies at one locus are:

$$f_{AA} = 0.58, f_{Aa} = 0.07, f_{aa} = 0.35.$$

- ➊ What is  $F$ , the inbreeding coefficient?

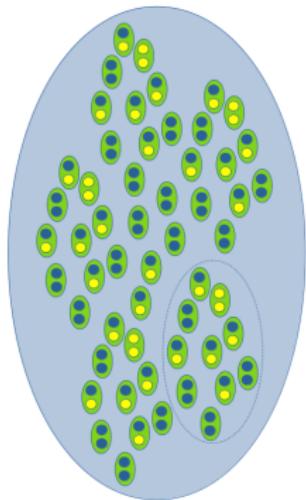
$$f_A = 0.58 + 0.07/2 = 0.615,$$

$$f_a = 1 - 0.615 = 0.385$$

$$F = (2 \times 0.385 \times 0.615 - 0.07) / (2 \times 0.385 \times 0.615) = 0.852$$

- ➋ Does it deviate from HWE?

## 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

## 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  $\chi^2$  is large enough, we reject the null hypothesis.

## Deviations from HWE

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

## Deviations from HWE

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

$O_{aa} = 10$

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

## Deviations from HWE

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

$O_{aa} = 10$

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

## Deviations from HWE

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

$$O_{aa} = 10$$

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

Is 8.711 "large enough"?

## 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?

## 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).

## Intended Learning Outcomes

### Alleles and genotypes

In this lecture you have learnt

- to describe all different types of genetic data
- to demonstrate the mathematical relationship between allele and genotype frequencies
- to calculate inbreeding coefficients and test for deviation from Hardy-Weinberg Equilibrium from genomic data using R

## 2. Genetic drift

### Intended Learning Outcomes

In this lecture you will learn

- to describe the Wright-Fisher model of genetic drift
- to calculate expected allele frequencies
- to appreciate the effect of population size on drift

## Allele frequencies through time

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

The most important factors that cause allele frequencies to change are

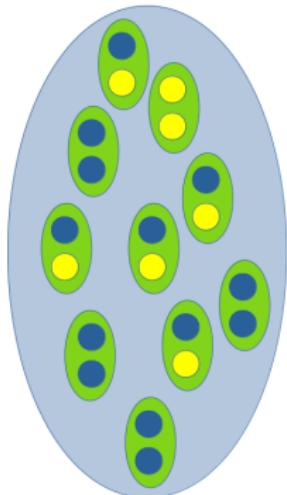
- mutation
- natural selection
- migration
- **genetic drift**

## Genetic drift

The **random** change of allele frequencies in populations of **finite** size.

## Genetic drift

The **random** change of allele frequencies in populations of **finite** size.



- 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 change from one generation to another
- over many generations, this process can produce large changes in allele frequencies

## Wright-Fisher model



Figure 15: Sewall Wright



Figure 16: R.A. Fisher

## Wright-Fisher model

Assumptions:

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

In a population of size  $2N$ , gene copies are transmitted from generation  $t$  to  $t + 1$  by random sampling with replacement (independently and with equal probability) of the gene copies in generation  $t$ .

## Wright-Fisher model

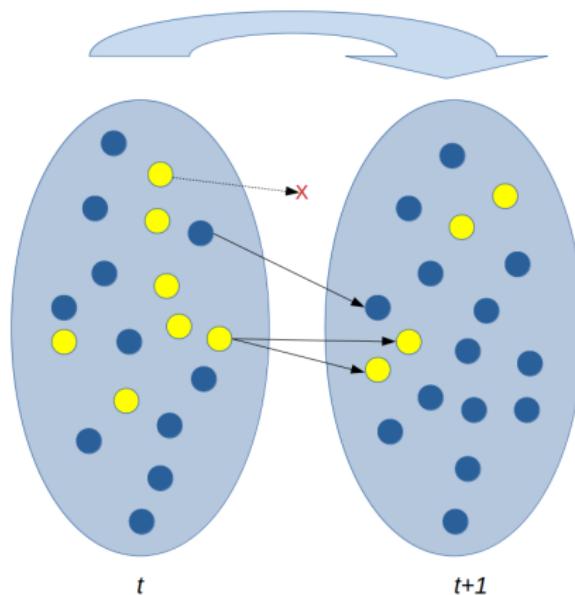


Figure 17: Two generations of a Wright-Fisher population.

## Wright-Fisher 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  $A$ ?

R script

## Expected allele frequency

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

$$E[f_A(t + 1)] = 2Nf_A(t)/2N = f_A(t) \quad (9)$$

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

## Drift in the Wright-Fisher model

What happens when we repeat the Wright-Fisher sampling scheme over **many** generations?

R plots

## Drift in the Wright-Fisher model

- 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.

## Drift in the Wright-Fisher model

- allele frequency may increase or decrease with equal probabilities
- in some cases, allele has become fixed  $f_A(t) = 1$  or lost  $f_A(t) = 0$

When an allele first has become fixed or lost, its frequency cannot change anymore (e.g. if  $f_A(t) = 0$  then  $f_A(t + 1) = 0$ )

Is it always true?

## Drift in the Wright-Fisher model

- allele frequency may increase or decrease with equal probabilities
- in some cases, allele has become fixed  $f_A(t) = 1$  or lost  $f_A(t) = 0$

When an allele first has become fixed or lost, its frequency cannot change anymore (e.g. if  $f_A(t) = 0$  then  $f_A(t + 1) = 0$ )

Is it always true? Yes if we assume no recurrent mutation:  
in the absence of recurrent mutation, it can be shown  
mathematically that an allele must eventually become fixed or lost.

## Effect of population size

How **fast** can genetic drift change allele frequencies?  
It depends on the population size,  $N$ .

R plots

## Effect of population size

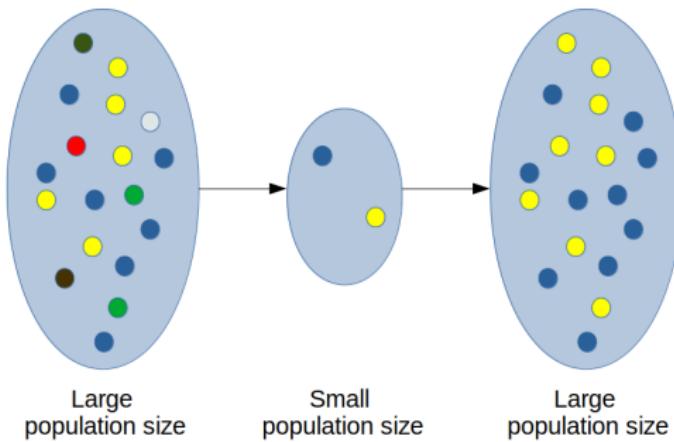
Large changes in allele frequency are unlikely in large populations, but happen more easily by chance in small populations.

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.

## Bottleneck in population size

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.



## Bottleneck in population size



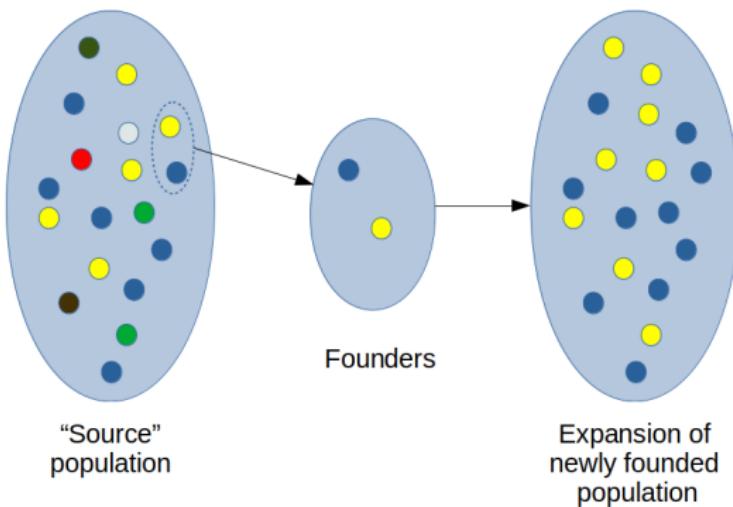
Figure 18: Northern elephant seal (*Mirounga angustirostris*)

Sea elephants were hunted nearly to extinction to a population of just 2-20 individuals. Today the population rebounded to 175,000 individuals. From historical (before hunting) and modern samples, their genetic diversity\* was reduced from 0.90 to 0.41.

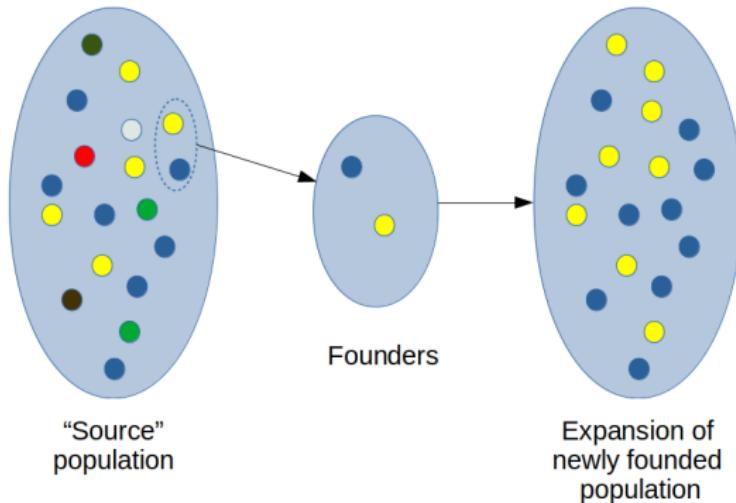
\* more in this later

## 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.

## Intended Learning Outcomes

### Genetic drift

In this lecture you have learnt

- to describe the Wright-Fisher model of genetic drift
- to calculate expected allele frequencies
- to appreciate the effect of population size on drift

# Intended Learning Outcomes

## Mutation

In this lecture you will learn

- to appreciate the effects of novel mutations on allele frequencies
- to describe the concepts of mutation and substitution rate
- to calculate divergence times using the molecular clock from genomic data with R

## Mutations

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

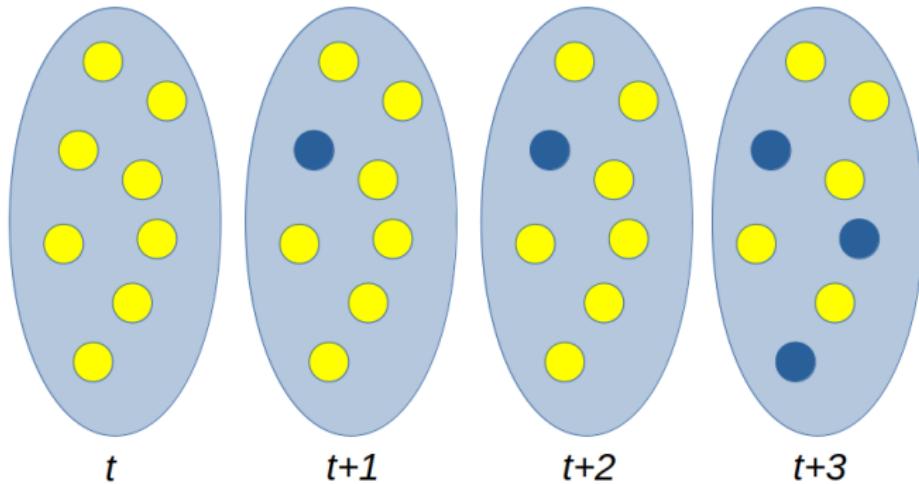
- deletions
- insertions
- translocations
- point mutations

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\*.

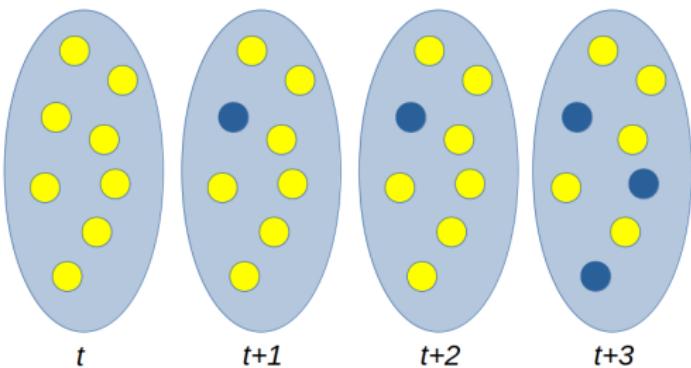
\* more on this later

## Effect of mutations on allele frequency

Assume that the  $a$  allele in each individual randomly mutates to  $A$  with probability  $\mu$  (**mutation rate**) in each generation.

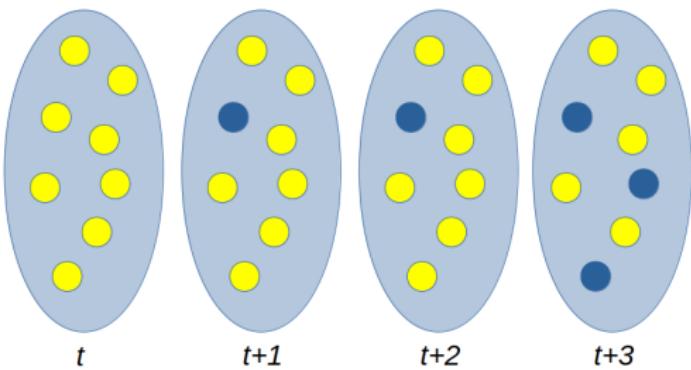


## Effect of mutations on allele frequency



What is the  $E[f_A(t + 1)]$  given  $f_A(t)$  and  $\mu$ ?

## Effect of mutations on allele frequency



What is the  $E[f_A(t + 1)]$  given  $f_A(t)$  and  $\mu$ ?

$$E[f_A(t + 1)] = f_A(t) + \mu f_a(t) \quad (10)$$

R plots

## Effect of mutations on allele frequency

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

$$E[f_A(t+1)] = (1 - \mu_{A \rightarrow a})f_A(t) + \mu_{a \rightarrow A}f_a(t) \quad (11)$$

R plots

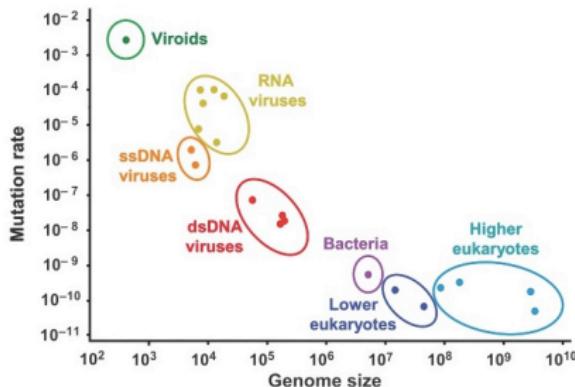
## Effect of mutations on allele frequency

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

$$f_A = \frac{\mu_{a \rightarrow A}}{\mu_{a \rightarrow A} + \mu_{A \rightarrow a}} \quad (12)$$

R plots

## Mutation rate



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

## Probability of fixation

The probability that an allele of frequency  $1/(2N)$  goes to fixation is  $1/(2N)$ .

$Pr(\text{fixation of allele A}) = N_A \times (1/2N) = f_A(t)$  at generation  $t$ .

In the absence of selection and mutation, the probability of fixation of an allele is simply its allele frequency.

## Rate of substitution

Rate at which mutations accumulate **between species or populations**.  
Substitution refers to mutations that have gone to fixation.

Assume:

- mutation rate  $\mu$ : in each generation  $\mu$  new mutations occur in each gene copy (e.g. per site, per gene, ...)
- $2N$  gene copies

## Rate of substitution

Rate at which mutations accumulate **between species or populations**.  
Substitution refers to mutations that have gone to fixation.

Assume:

- mutation rate  $\mu$ : in each generation  $\mu$  new mutations occur in each gene copy (e.g. per site, per gene, ...)
- $2N$  gene copies

The expected number of mutations each generation that eventually will go to fixation is

$$2N\mu \times 1/(2N) = \mu$$

The rate of substitution is simply the mutation rate.

## Molecular clock

If

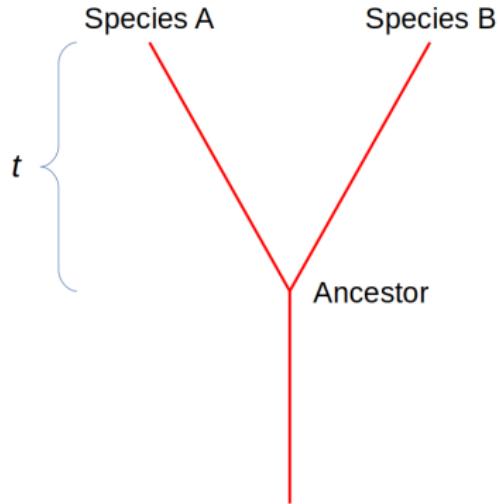
- no selection
- "low" mutation rate (not affecting allele frequencies much)
- constant mutation rate

then the rate of substitution should be constant *in time*.

Mutations can be used to date divergence between species.

How?

## Molecular clock



The expected number of nucleotide differences separating sequences of the same genes in the two species is  $E[d_{AB}] = 2\mu t_{AB}$ .

Therefore

- $t_{AB} = \frac{d_{AB}}{2\mu}$  or

- $\mu = \frac{d_{AB}}{2t_{AB}}$

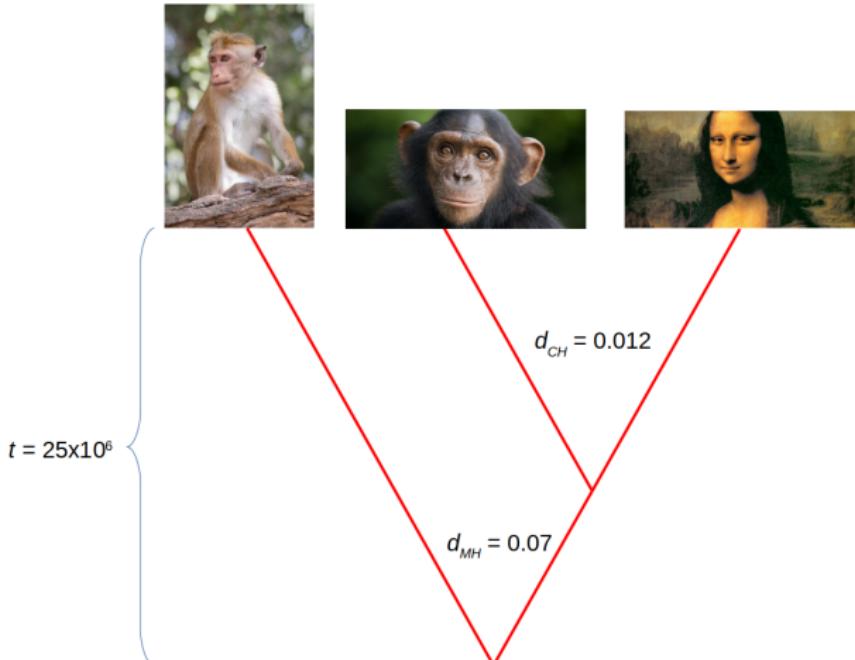
## Molecular clock

Caveats:

- it depends on an estimate of  $\mu$
- it assumes no natural selection acting upon
- it assumes that mutation rate is constant and equal among different species

Not very realistic but a good approximation for closely related species.

## Dating human-chimpanzee divergence time



## Dating human-chimpanzee divergence time

$$\mu = 0.07 / (2 \times 25 \times 10^6) = 1.4 \times 10^{-9} \text{ per site per year}$$

$$t = 0.012 / (2 \times 1.4 \times 10^{-9}) = 4.3 \text{ million years ago}$$

Compatible but shorter than expected:

- generation time has changed and therefore the rate per year changed
- effect of natural selection
- change in mutation rate
- errors in estimating nucleotide divergence
- error in estimating human-macaque divergence time
- ...

## Intended Learning Outcomes

### Mutation

In this lecture you have learnt

- to appreciate the effects of novel mutations on allele frequencies
- to describe the concepts of mutation and substitution rate
- to calculate divergence times using the molecular clock from genomic data with R

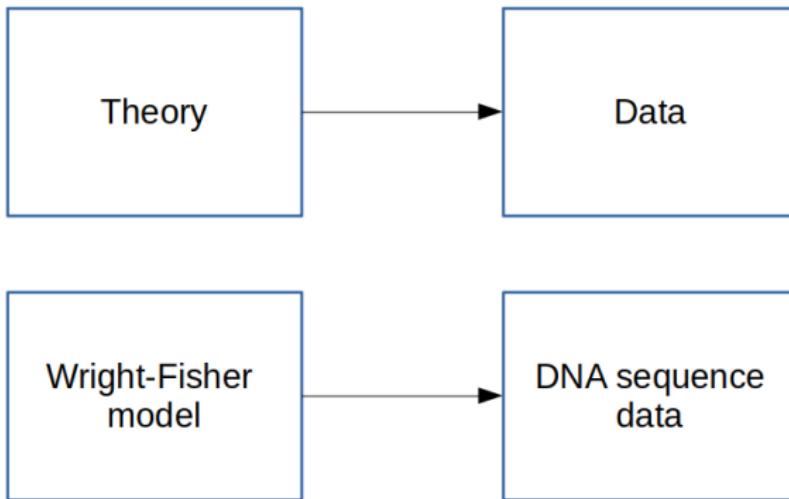
## 03. Coalescence theory

### Intended Learning Outcomes

In this lecture you will learn to

- describe principles and assumptions of the coalescence theory
- discuss the infinite sites model
- provide estimators of  $\theta$  and effective population size
- measure genetic variability with summary statistics and the site frequency spectrum with R

## Motivation



## Motivation

e.g. on the X chromosomes, two Europeans differ, on average, in 0.08% of sites, while individuals from African populations differ in 0.12% of sites.

What do these numbers tell us *about* the two populations?

## Motivation

e.g. on the X chromosomes, two Europeans differ, on average, in 0.08% of sites, while individuals from African populations differ in 0.12% of sites.

What do these numbers tell us *about* the two populations?

We use **coalescence theory**, which is based on Wright-Fisher model, to consider the genealogy history of a sample and make inferences about populations instead of modelling changes of allele frequencies forward in time.

## Coalescence

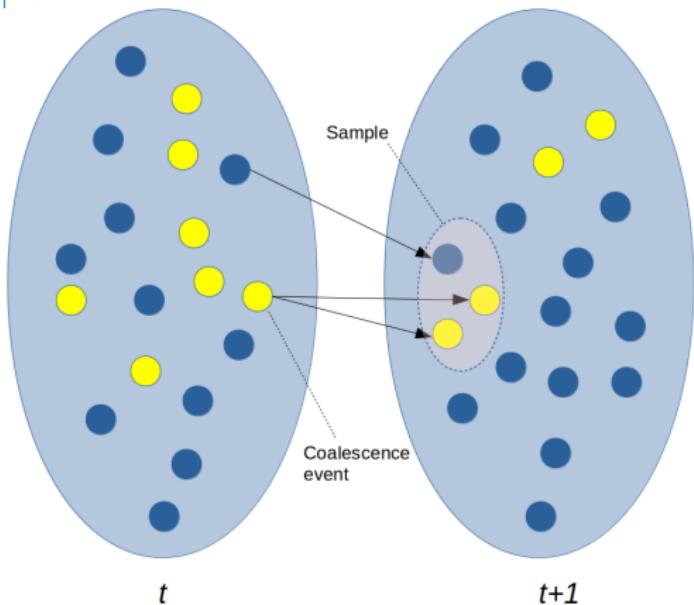


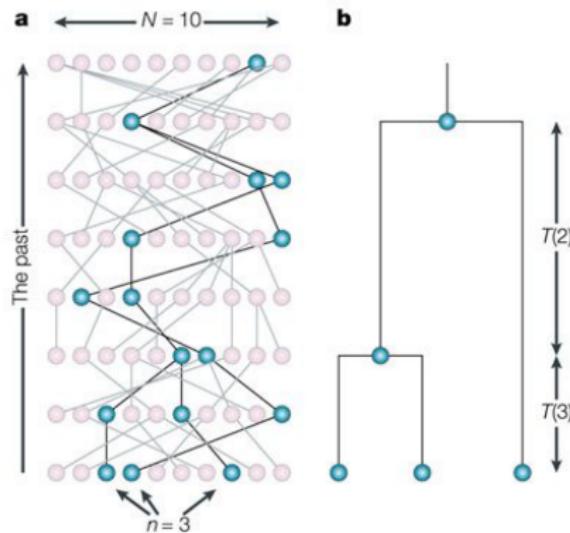
Figure 19: Tracking the ancestry of a sample between two generations.

## Coalescence

If two individual gene copies have the same parent in the previous generation, we say that the **ancestral lineage** representing these two individuals have **coalesced**.

They have a **common ancestor** and a **coalescent event** has occurred.

## Coalescence tree



Nature Reviews | Genetics

Figure 20: Ancestry of three samples.

## Coalescence tree

The ancestry of an individual gene copy is represented by a line (or edge).

The time until two lineages find a **most recent common ancestor (MRCA)** is called **coalescence time**.

How can we find the coalescence time?

## Coalescence in a sample of two gene copies

What is the probability of two samples having the same parent in the previous generation?

As there are  $2N$  potential parents "chosen" with equal probability, the probability of two individuals having the same parent in the previous generation is

## Coalescence in a sample of two gene copies

What is the probability of two samples having the same parent in the previous generation?

As there are  $2N$  potential parents "chosen" with equal probability, the probability of two individuals having the same parent in the previous generation is  $1/(2N)$ .

## Coalescence in a sample of two gene copies

What is the probability of two samples having the same parent in the previous generation?

As there are  $2N$  potential parents "chosen" with equal probability, the probability of two individuals having the same parent in the previous generation is  $1/(2N)$ .

The probability that two gene copies did NOT have the same parent in the previous generation is

## Coalescence in a sample of two gene copies

What is the probability of two samples having the same parent in the previous generation?

As there are  $2N$  potential parents "chosen" with equal probability, the probability of two individuals having the same parent in the previous generation is  $1/(2N)$ .

The probability that two gene copies did NOT have the same parent in the previous generation is  $1 - 1/(2N)$ .

## Coalescence in a sample of two gene copies

What is the probability of two samples having the same parent in the previous generation?

As there are  $2N$  potential parents "chosen" with equal probability, the probability of two individuals having the same parent in the previous generation is  $1/(2N)$ .

The probability that two gene copies did NOT have the same parent in the previous generation is  $1 - 1/(2N)$ .

The probability that two gene copies did not have the same parent in the past  $r$  generations is

## Coalescence in a sample of two gene copies

What is the probability of two samples having the same parent in the previous generation?

As there are  $2N$  potential parents "chosen" with equal probability, the probability of two individuals having the same parent in the previous generation is  $1/(2N)$ .

The probability that two gene copies did NOT have the same parent in the previous generation is  $1 - 1/(2N)$ .

The probability that two gene copies did not have the same parent in the past  $r$  generations is  $[1 - 1/(2N)]^r$ .

## Coalescence in a sample of two gene copies

The probability of not finding any common ancestor in generation  $r - 1$  but then finding the first common ancestor in generation  $r$  is

## Coalescence in a sample of two gene copies

The probability of not finding any common ancestor in generation  $r - 1$  but then finding the first common ancestor in generation  $r$  is

$$Pr(\dots) = [1 - 1/(2N)]^{r-1} [1/(2N)] \quad (13)$$

This equation gives us the probability distribution of the time to the MRCA in a sample of size  $n = 2$ . This is a geometric random variable: the probability distribution of the number of Bernoulli trials needed to get one success.

R scripts

## Coalescence in large populations

- If we consider the limit of an infinitely large population, calculations simplify but we can still consider the effect of genetic drift.
- It is convenient to measure time in  $2N$  generations, by setting  $r = 2Nt$  with  $t$  measuring time in  $2N$  generations.

## Coalescence in large populations

- If we consider the limit of an infinitely large population, calculations simplify but we can still consider the effect of genetic drift.
- It is convenient to measure time in  $2N$  generations, by setting  $r = 2Nt$  with  $t$  measuring time in  $2N$  generations.

The probability that two gene copies do not find a common ancestor in  $2Nt$  generations becomes

$$[1 - 1/(2N)]^{2Nt} \rightarrow e^{-t} \text{ as } N \rightarrow \infty$$

## Coalescence in large populations

As  $N$  becomes large, the distribution of the coalescence times follows an **exponential distribution** with mean 1.

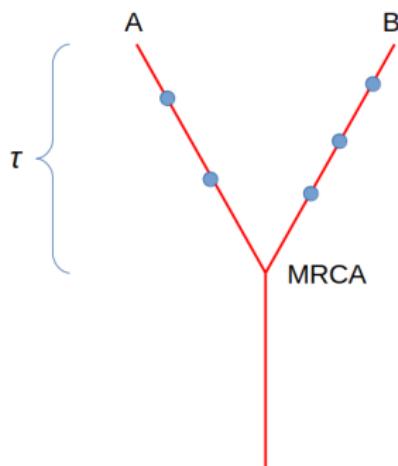
As time is measured in  $2N$  generations, the mean (expected) time to coalescence is actually 1, so  $1 * 2N$  generations. In other words, there is a constant rate of coalescence of 1 per  $2N$  generations.

R scripts

## Coalescence in large population

- The random process of following the lineages backward in time until a most recent common ancestor has been found is called a **coalescence process**.
- If the coalescence rate is 1 per  $2N$  generations, it is intuitive to understand that the expected coalescence time of two alleles (the time until the coalescent event occurs) is  $2N$  generations (although there is considerable variability in the coalescence times).

## Genetic variability and population size



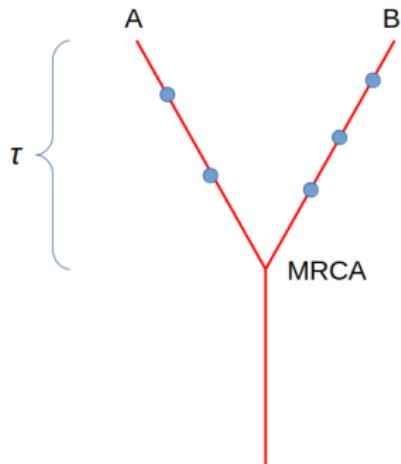
We expect  $\mu r$  mutations in  $r$  generations. If we measure time by  $2N$  generations, that is  $t = r/(2N)$ , we expect  $2N\mu t$  mutations on a lineage of length  $t$ .

Since the expected time of coalescence is  $E[t] = 1$  and there are two lineages, the expected number of mutations separating two gene copies is

$$\theta = 4N\mu \quad (14)$$

which is a simple relationship between the amount of genetic variability and population sizes.

## Genetic variability and population size



The expected number of mutations occurring in a lineage during any time interval of length  $\tau$  is  $2N\mu\tau = \tau\theta/2$ .

As such, we can think of the data generated by a coalescence process producing a coalescence tree and a subsequent process in which mutations are distributed across the lineage of the tree at rate  $\theta/2$ .

## Infinite Sites Model

Each new mutation creates a new variable site, i.e. that each new mutation hits a new site in the sequence, such that no site experiences more than one mutation.



Figure 21: The sequence is infinitely long so that the chance of two mutations hit the same site is essentially zero.

## Infinite Sites Model

The sites in which some of the individuals differ are called **segregating sites**. **Single nucleotide polymorphisms** (SNPs) if involve only one nucleotide.

Sequence 1	aggaa ggacc aagac gatag
Sequence 2	aggaa ggaac gagac gatag
Sequence 3	aggaa ggaac gagac gatag
Sequence 4	aggag ggacc gagac gatag
Sequence 5	aggag ggacc gagac gatag

Under the infinite sites model (no back-mutations) we can deduce which mutations occurred in the ancestry of a sample of sequences.

## Infinite Sites Model

The model does not distinguish between different nucleotides and does not care about invariable sites.

Sequence 1	aggaa	ggacc	aagac	gatag
Sequence 2	aggaa	ggaac	<b>gagac</b>	gatag
Sequence 3	aggaa	ggaac	<b>gagac</b>	gatag
Sequence 4	aggag	ggacc	<b>gagac</b>	gatag
Sequence 5	aggag	ggacc	<b>gagac</b>	gatag

Sequence 1	0	0	0
Sequence 2	0	1	1
Sequence 3	0	1	1
Sequence 4	1	0	1
Sequence 5	1	0	1

Figure 22: Data as a binary matrix of the variable sites.

## Infinite Sites Model

- Labelling with zeros and ones is arbitrary.
- Good approximation if the rate of mutation is low.
- DNA sequences with different mutations are different **haplotypes**.

0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

Figure 23: How many DNA sequences? How many haplotypes?

## Tajima's estimator

We want an estimate of  $\theta = 4N\mu$  under the infinite sites model from the expected number of mutations separating two individuals based on the DNA sequences obtained from data.

## Tajima's estimator

We want an estimate of  $\theta = 4N\mu$  under the infinite sites model from the expected number of mutations separating two individuals based on the DNA sequences obtained from data.

Data can be summarised as the **average number of pairwise differences**, or  $\pi$ .

$$\pi = \frac{\sum_{i < j} d_{i,j}}{n(n - 1)/2} \quad (15)$$

with  $n$  sequences,  $d_{i,j}$  number of differences between sequence  $i$  and  $j$ .

## Tajima's estimator

0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

Figure 24: What is the value of  $\pi$ ?

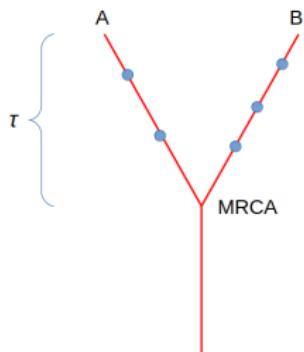
## Tajima's estimator

0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

Figure 24: What is the value of  $\pi$ ?

$$\pi = (2 + 2 + 2 + 2 + 0 + 2 + 2 + 2 + 2 + 0) / (5 \times 4/2) = 1.6$$

## Tajima's estimator



The expected number of nucleotide differences between two sequences is the expected number of mutations,  $\theta = 4N\mu$ .

$$E[d_{i,j}] = \theta \quad (16)$$

$$E[\pi] = \theta \quad (17)$$

$\hat{\theta}_T = \pi$  is called **Tajima's estimator** of  $\theta$ .

## Effective population size

The number of individuals in a Wright-Fisher model that would produce the same amount of genetic drift as in the real population.

The amount of genetic drift can be measured as

- the expected heterozygosity
- expected number of pairwise differences
- rate of coalescence
- ...

## Effective population size ( $N_e$ )

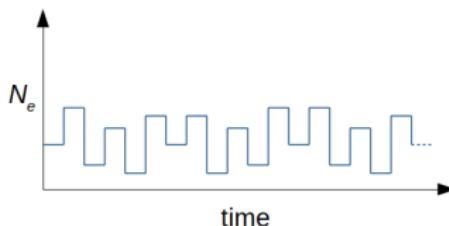
The true number of individuals in the population can be very different from its effective population size!

## Effective population size



**Figure 25:** The effective population size of the Chinook salmon (*Oncorhynchus tshawytscha*) has been estimated to be very low, possibly because the population size fluctuates between years and high variance offspring.

## Effective population size ( $N_e$ )



If a population fluctuates between sizes  $N_1, N_2, \dots, N_k$  at a proportion  $p_1, p_2, \dots, p_k$  of the time, the coalescent effective population size is the harmonic mean:

$$N_e = \frac{1}{p_1/N_1 + p_2/N_2 + \dots + p_k/N_k} \quad (18)$$

which is smaller than the arithmetic mean and gives more weight to smaller sizes.

## Effective population size ( $N_e$ )



The effective population size with unequal sex ratio is

$$N_e = \frac{4N_m N_f}{N_m + N_f} \quad (19)$$

which is smaller than  $N_m + N_f$ .

## Interpreting estimates of $\theta$

---

$\pi$  on autosomes

---

Mandenka	0.00120
Biaka	0.00121
San	0.00126
Han	0.00081
Basque	0.00087
Melanesians	0.00078

---

## Interpreting estimates of $\theta$

---

$\pi$  on X chromosomes

---

Mandenka	0.00099
Biaka	0.00095
San	0.00085
Han	0.00058
Basque	0.00071
Melanesians	0.00066

---

## Watterson's estimator

$$\hat{\theta}_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}} \quad (20)$$

with  $S$  segregating sites and  $n$  samples.

$$E[\hat{\theta}_W] = \theta \quad (21)$$

## Watterson's estimator

0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

Figure 26: What is the value of  $\hat{\theta}_W$ ?

## Watterson's estimator

0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

Figure 26: What is the value of  $\hat{\theta}_W$ ?

$$\hat{\theta}_W = 3/(1 + 1/2 + 1/3 + 1/4) = 1.4$$

but before we obtained  $\hat{\theta}_T = 1.6$ .

Why?

## Summary statistics

Possible summaries of DNA sequence data are:

- the number of segregating sites ( $S$ )
- the average number of pairwise differences ( $\pi$ )

but they don't provide much information regarding **allele frequencies**.

## The Site Frequency Spectrum (SFS)

### SFS

The SFS is obtained by tabulating the sample allele frequencies of all mutations.

0 0 0  
0 1 1  
0 1 1  
1 0 1  
1 0 1

## The Site Frequency Spectrum (SFS)

### SFS

The SFS is obtained by tabulating the sample allele frequencies of all mutations.

0 0 0  
0 1 1  
0 1 1  
1 0 1  
1 0 1

The "1" alleles have frequencies  $2/5$ ,  $2/5$  and  $4/5$ .  
The proportion of sites with the allele frequencies equal to  $1/5$ ,  $2/5$ ,  $3/5$  and  $4/5$  in the sample are

## The Site Frequency Spectrum (SFS)

### SFS

The SFS is obtained by tabulating the sample allele frequencies of all mutations.

0 0 0  
0 1 1  
0 1 1  
1 0 1  
1 0 1

The "1" alleles have frequencies  $2/5$ ,  $2/5$  and  $4/5$ .

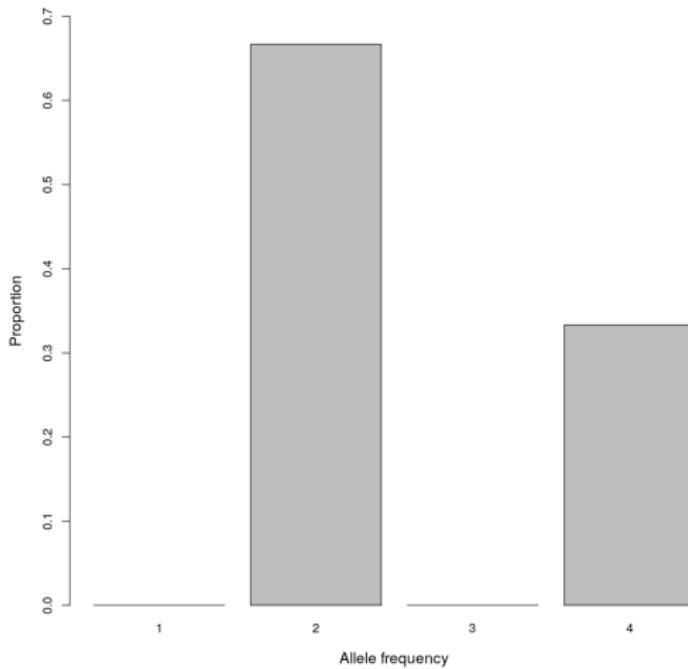
The proportion of sites with the allele frequencies equal to  $1/5$ ,  $2/5$ ,  $3/5$  and  $4/5$  in the sample are  $f_1 = 0$ ,  $f_2 = 2/3$ ,  $f_3 = 0$  and  $f_4 = 1/3$ .

$$\vec{f} = (f_1, f_2, \dots, f_{n-1})$$

for a sample of  $n$  haploid individuals.

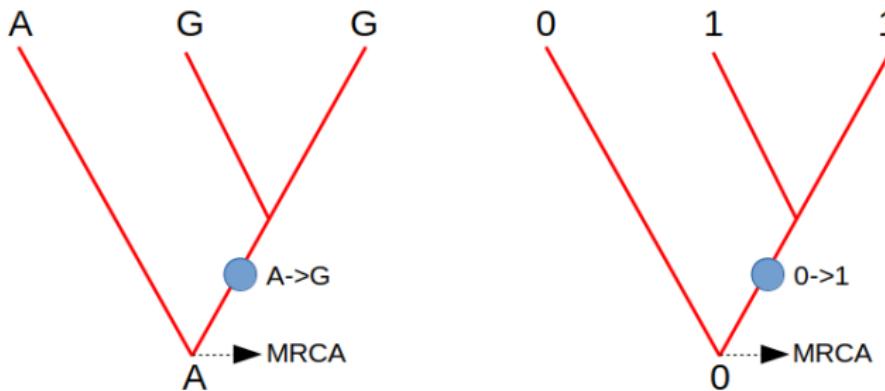
# The Site Frequency Spectrum (SFS)

0 0 0  
0 1 1  
0 1 1  
1 0 1  
1 0 1



## Alleles

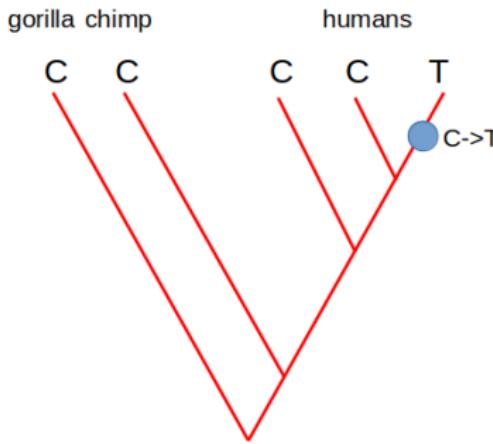
- **ancestral** allele is the allele found in the MRCA of the sample.
- **derived** allele (or mutated) is an allele that is not ancestral.



## Alleles

The ancestral allele is often inferred using **outgroups**.

e.g. if  $C/T$  polymorphism in humans and primate have  $C$ , then  $C$  is likely to be the ancestral allele.



# Alleles

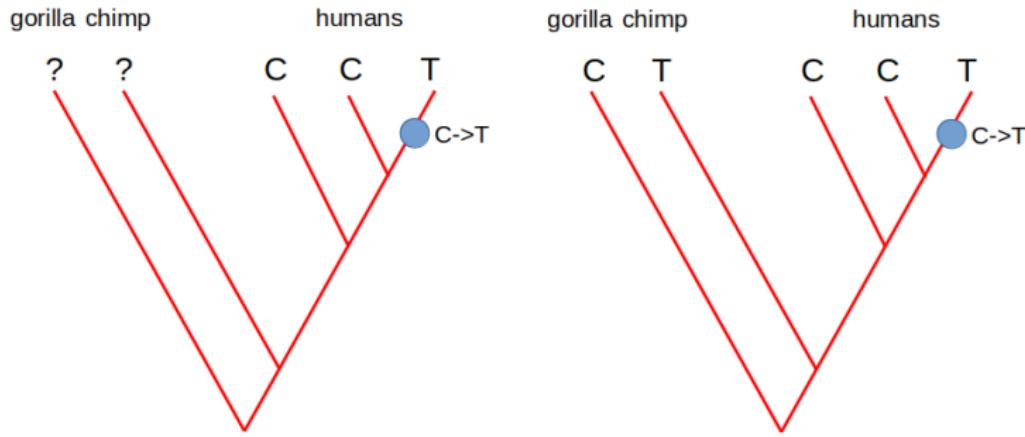


Figure 27: Uncertain ancestral allele.

## The folded SFS

If no information on the ancestral allele is available, the alternative allele is the allele with lowest allele frequency.

- allele in highest frequency: 0
- allele in lowest frequency: 1

The maximum possible Minor Allele Frequency (MAF) is 0.5

A frequency of 0.8 of an uncertain derived allele becomes a MAF of 0.2

The SFS calculated using MAF is called **folded SFS( $f^*$ )**

## The folded SFS

0 0 0  
0 1 1  
0 1 1  
1 0 1  
1 0 1

$$\vec{f}^* =$$

## The folded SFS

0 0 1

0 1 0

0 1 0

$\vec{f}^* =$

1 0 0

1 0 0

## The folded SFS

0 0 1

0 1 0

0 1 0       $\vec{f}^* = (f_1^* = 1/3, f_2^* = 2/3)$

1 0 0

1 0 0

## The Site Frequency Spectrum

- $S$  and  $\pi$  can be calculated directly from  $\vec{f}$  but the opposite is not true.
- Alleles segregating at frequency of  $1/n$  are called **singletons**.
- The expected SFS under the standard coalescence model with infinite sites mutations is

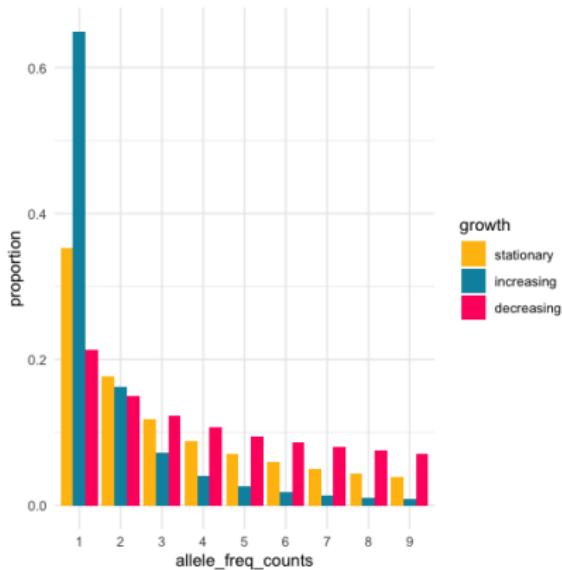
$$E[f_i] = \frac{1/j}{\sum_{k=1}^{n-1} \frac{1}{k}} \quad (22)$$

with  $j = 1, 2, \dots, n - 1$

R script

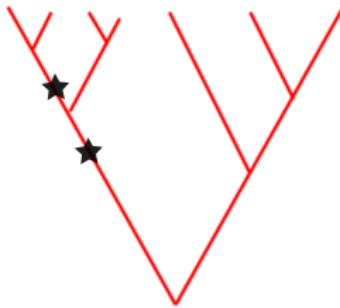
## The Site Frequency Spectrum

- Increasing population size leads to more singletones
- Decreasing population size leads to more mid-frequency counts

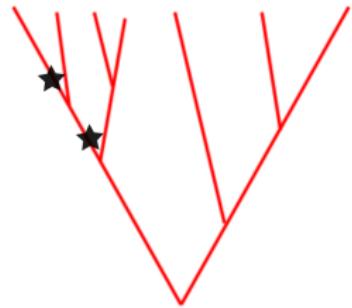


# Tree shape and population size

Constant population size

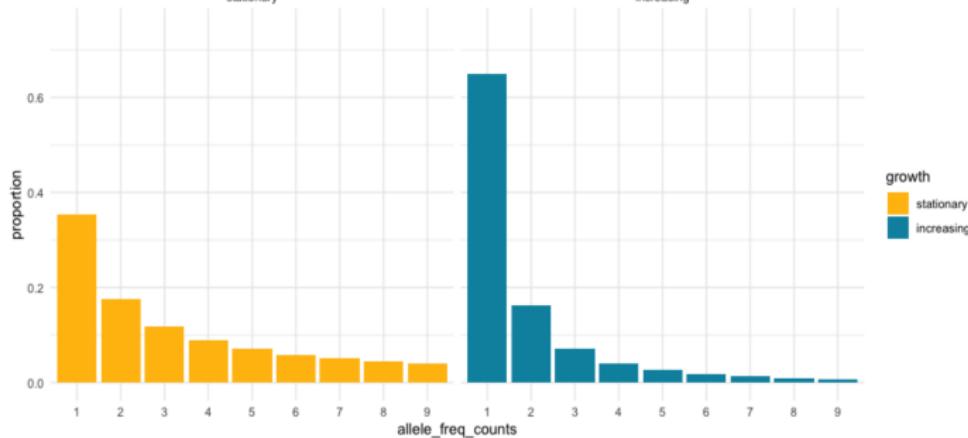


Increasing population size

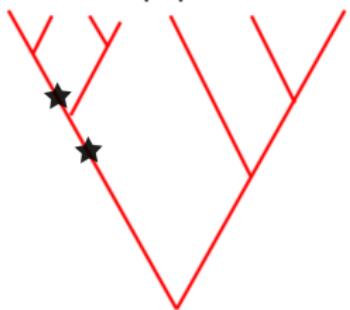


stationary

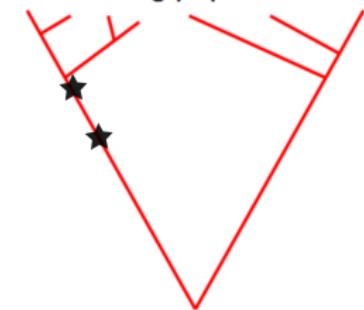
increasing



Constant population size

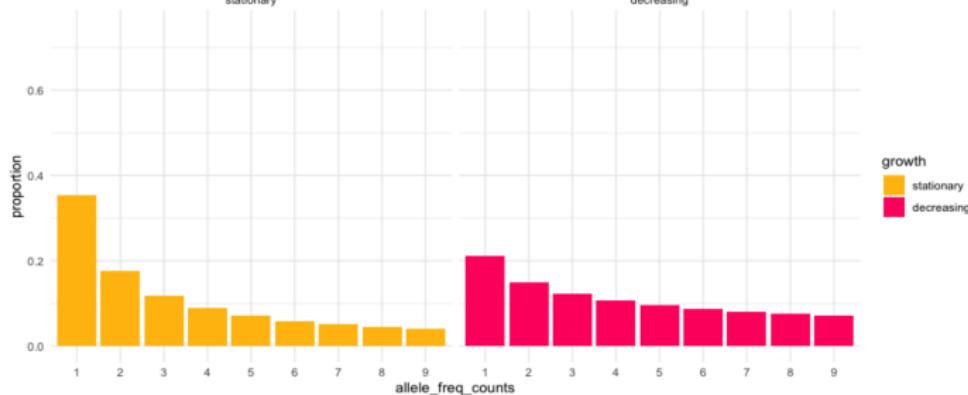


Decreasing population size



stationary

decreasing



## Intended Learning Outcomes

### Coalescence theory

In this lecture you have learnt to

- describe principles and assumptions of the coalescence theory
- discuss the infinite sites model
- provide estimators of  $\theta$  and effective population sizes
- measure genetic variability with summary statistics and the site frequency spectrum with R

## Intended Learning Outcomes

### Population subdivision

In this lecture you will learn to

- quantify the effect of population subdivision on allele frequencies and heterozygosity
- calculate measures of population genetic differentiation
- discuss divergence models

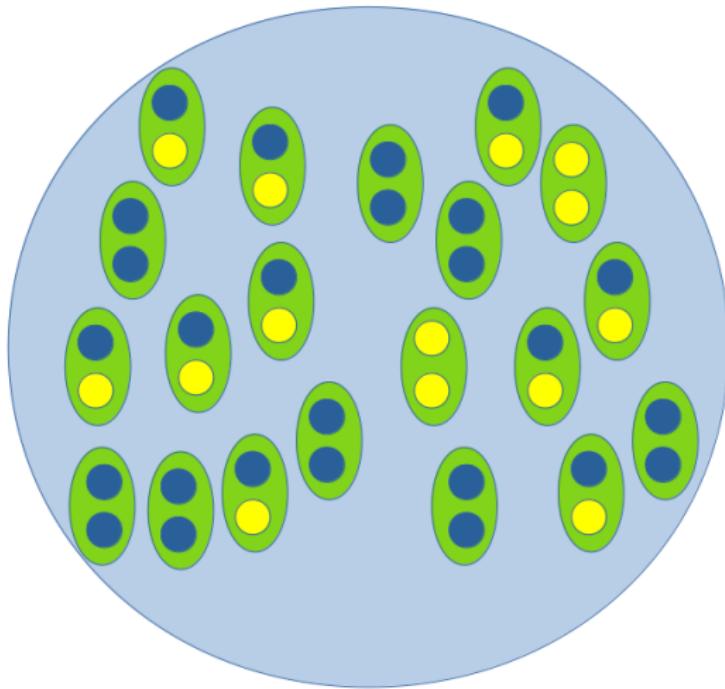
## Population subdivision

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

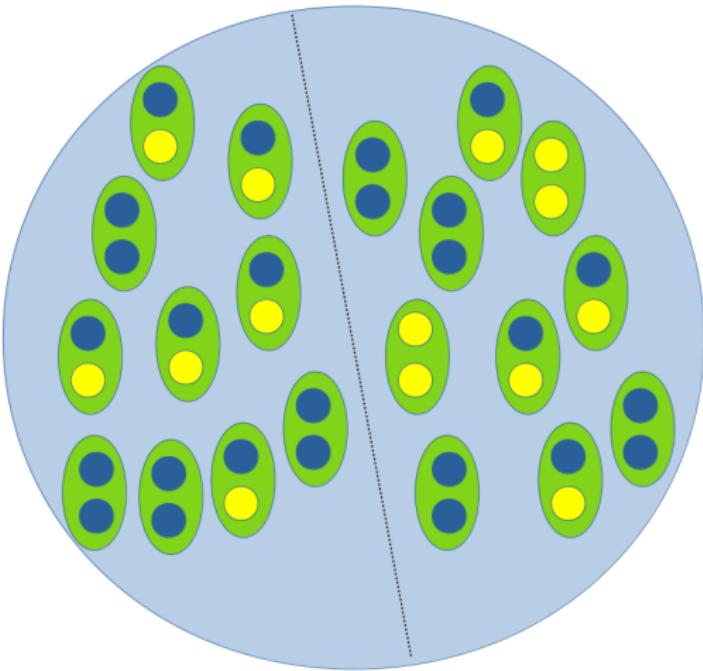
Population subdivision is important to

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

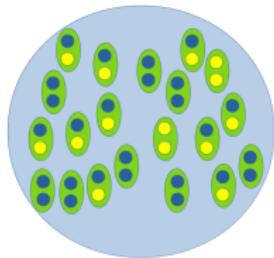
## Population subdivision



## Population subdivision



## Allele frequencies in a subdivided population



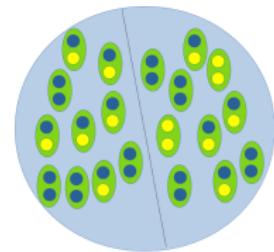
Assume two subpopulations, each one in HW equilibrium with  $N_1$  and  $N_2$  individuals, respectively.

The average frequency of allele  $A$  when pooling the two subpopulations is

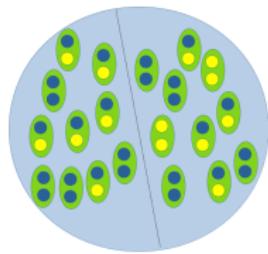
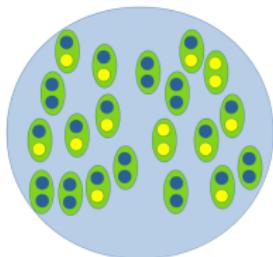
$$f_A = \frac{2N_1 f_{A1} + 2N_2 f_{A2}}{2N_1 + 2N_2} \quad (23)$$

If  $N_1 = N_2$

$$f_A = \frac{f_{A1} + f_{A2}}{2} \quad (24)$$



## Heterozygosity in a subdivided population



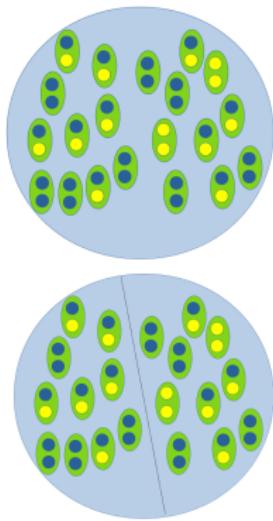
The expected proportion of heterozygous individuals considering subpopulations is

$$H_S = \frac{2f_{A1}(1 - f_{A1}) + 2f_{A2}(1 - f_{A2})}{2} \quad (25)$$

which is the expected heterozygosity when both subpopulations are sampled separately.

*S* in  $H_S$  stands for "in the subdivided population"

## Heterozygosity in a subdivided population



However, the expected proportion of heterozygous individuals in a population with frequency  $f_A$  is

$$H_T = 2 \frac{f_{A1} + f_{A2}}{2} \left(1 - \frac{f_{A1} + f_{A2}}{2}\right) \quad (26)$$

$T$  in  $H_T$  stands for "in the total (pooled)  
population"

## Heterozygosity in a subdivided population

After some rearrangements we have

$$H_S = f_{A1}(1 - f_{A1}) + f_{A2}(1 - f_{A2}) \quad (27)$$

and

$$H_T = f_{A1}(1 - f_{A1}) + f_{A2}(1 - f_{A2}) + \delta^2/2 \quad (28)$$

with  $\delta = |f_{A1} - f_{A2}|$ .

## Heterozygosity in a subdivided population

$$H_T = H_S + \delta^2/2$$

- If  $\delta = 0$  then

## Heterozygosity in a subdivided population

$$H_T = H_S + \delta^2/2$$

- If  $\delta = 0$  then  $H_T = H_S$  and the total (pooled) population is also in HWE.
- If  $\delta >> 0$  then

## Heterozygosity in a subdivided population

$$H_T = H_S + \delta^2/2$$

- If  $\delta = 0$  then  $H_T = H_S$  and the total (pooled) population is also in HWE.
- If  $\delta >> 0$  then  $H_T > H_S$  and

## Heterozygosity in a subdivided population

$$H_T = H_S + \delta^2/2$$

- If  $\delta = 0$  then  $H_T = H_S$  and the total (pooled) population is also in HWE.
- If  $\delta >> 0$  then  $H_T > H_S$  and the total (pooled) population contains fewer heterozygous individuals than expected given the pooled allele frequency.

### Wahlund effect

The decrease of heterozygosity in a subdivided population compared to a randomly mating one with the same (total) allele frequency.

## Quantifying population subdivision

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad (29)$$

## Quantifying population subdivision

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad (29)$$

$F_{ST}$  has a range defined as

- if  $\delta = 0$  then  $H_T = H_S$  and  $F_{ST} = 0$
- if  $\delta \gg 0$  then  $F_{ST} \approx 1$

$F_{ST}$  can be calculated for more than two subpopulations.

## $F_{ST}$ : population genetic differentiation



Figure 28: Humpback whales in the Pacific and Atlantic have strong genetic differentiation ( $F_{ST} > 0.4$ ) while populations in the North Atlantic have low differentiation ( $F_{ST} \approx 0.04$ ).

## Wright-Fisher model with migration

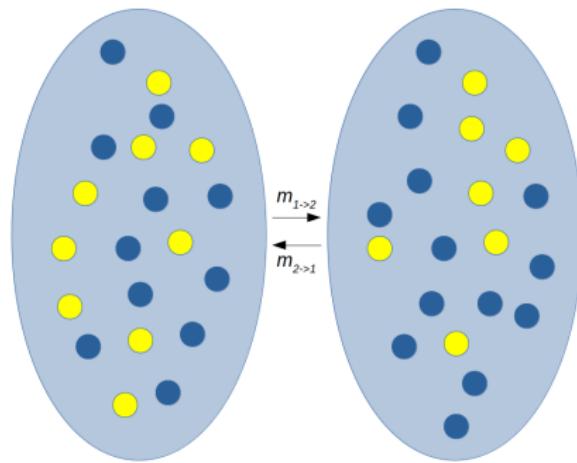


Figure 29: An individual from one population is replaced with an individual from the other with probability  $m$  (migration rate).

## $F_{ST}$ and migration rates

Using the coalescence theory assuming an infinite sites model, we can derive that

$$F_{ST} = \frac{1}{1 + 4Nm_T} \quad (30)$$

with  $m_T$  being the total number of migrants.

Rscript

## "Island" model

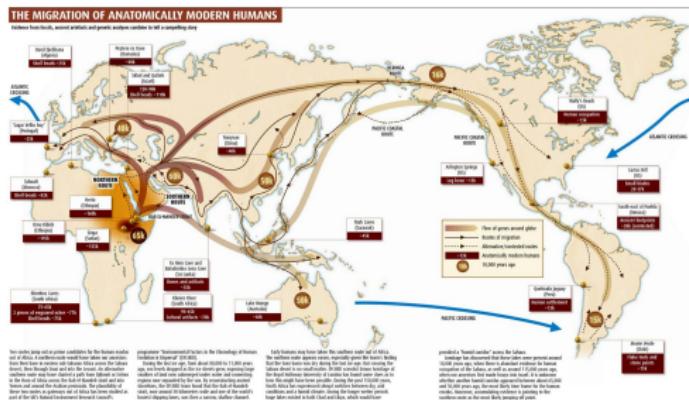
It assumes that populations have been subdivided for a very long time so that an equilibrium has been established and that then there is ongoing **gene-flow**.

It is not a realistic model for some species.

## ”Island” model

It assumes that populations have been subdivided for a very long time so that an equilibrium has been established and that then there is ongoing gene-flow.

It is not a realistic model for some species.



## Divergence model

It describes populations diverging from common ancestral populations without subsequent gene-flow.

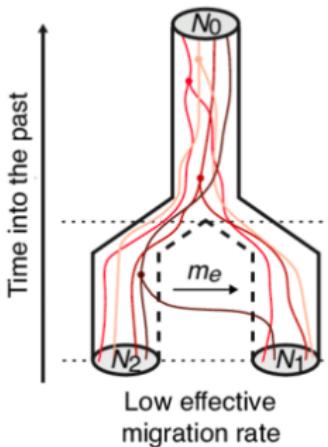


Figure 30: TMRCA overestimates the divergence time.

## Isolation by distance

The degree of population subdivision increases with geographical distance.

- migration rate is a linear function of geographical distance
- migration occurs only between adjacent populations (stepping-stone models)
- series of divergence events (sequential colonisation)

## Isolation by distance

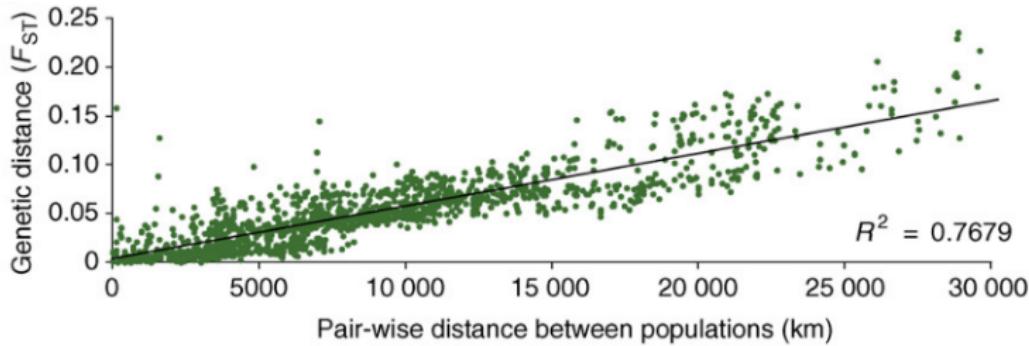


Figure 31: Isolation by distance in human populations.

## Intended Learning Outcomes

### Population subdivision

In this lecture you have learned to

- quantify the effect of population subdivision on allele frequencies and heterozygosity
- calculate measures of population genetic differentiation
- discuss divergence models

## Intended Learning Outcomes

### Demographic inference

Through examples from the literature, in this lecture you will learn to

- make inferences on population history from genomic data with R
- interpret results from population genetics studies
- design a genomic study
- (discuss the effect of demography on signatures of natural selection)

!

Please do not redistribute. Some materials may be protected by copyright.