

# EMBO PopGen

## Coalescent theory

Matthias Steinrücken

adapted from Matteo Fumagalli (EMBO, 2024)

Department of Ecology and Evolution, University of Chicago  
Department of Human Genetics, University of Chicago

Day 2b, June 24, 2025



# Intended Learning Outcomes

## Coalescent theory

In this session you will learn:

- Describe principles and assumptions of coalescent 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.

## Motivation

Example question: On the X chromosomes, two Europeans differ, on average, at 0.08% of sites, while individuals from African populations differ at 0.12% of sites.

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

## Motivation

Example question: On the X chromosomes, two Europeans differ, on average, at 0.08% of sites, while individuals from African populations differ at 0.12% of sites.

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

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

# Coalescence

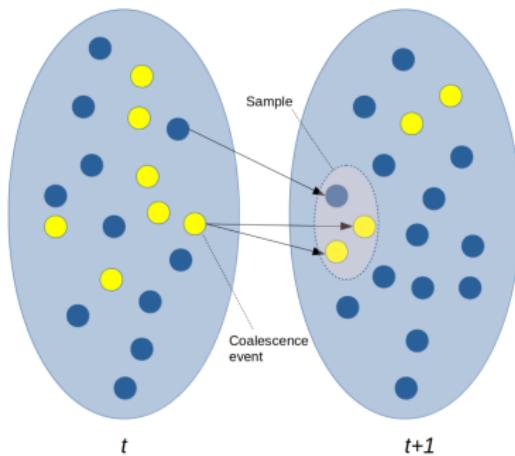


Figure 1: Tracking the ancestry of a sample between two generations.

## Coalescence

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

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

# Coalescent tree

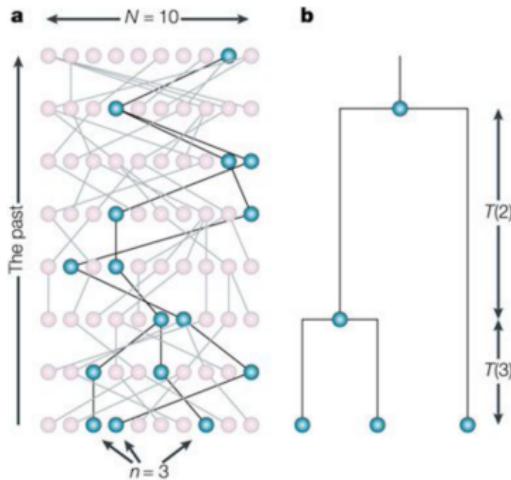


Figure 2: Ancestry of three samples.

## Coalescent tree

The ancestry of an individual gene copy is represented by a lineage.

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

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

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

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

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:

$$\mathbb{P}(\dots) = [1 - 1/(2N)]^{r-1} [1/(2N)]$$

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 (w.p.  $1/2N$ ).

## Coalescence in a sample of two gene copies

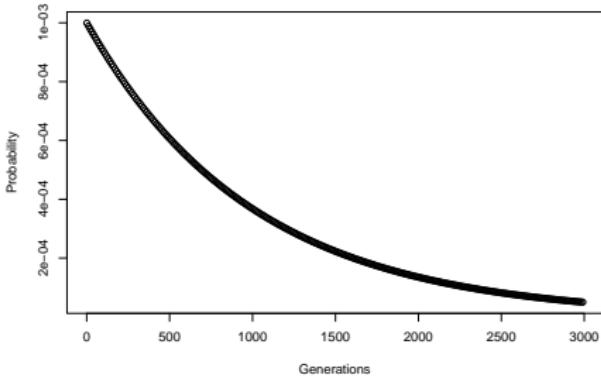


Figure 3: Distribution of TMRCA for  $n = 2$  in population of size  $N = 500$ .

## 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  $2N$  generations. In other words, there is a constant rate of coalescence of 1 per  $2N$  generations.

## Coalescence in large populations

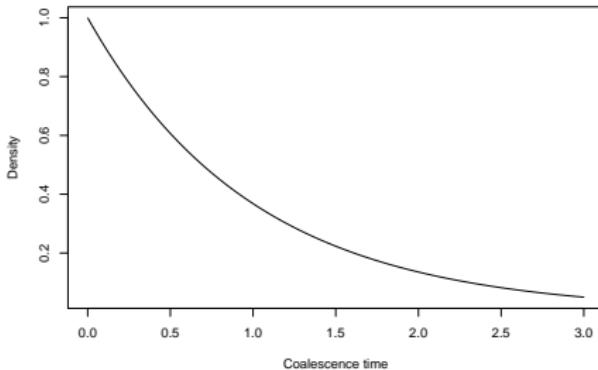


Figure 4: Exponential distribution of TMRCA for  $n = 2$ . Time  $t = 1$  corresponds to  $2N$  generations.

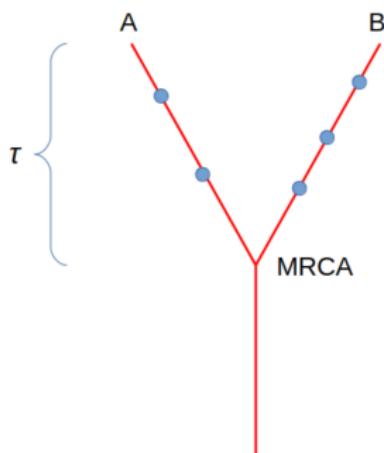
## Coalescence in large populations

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

## Coalescence in large populations

- The coalescence process in a large randomly mating diploid population with two sexes is the same as that in the simple haploid model.
- Once we have a convenient description of the genealogy, then it is easy to derive various properties of our sample.

## Coalescence in large populations



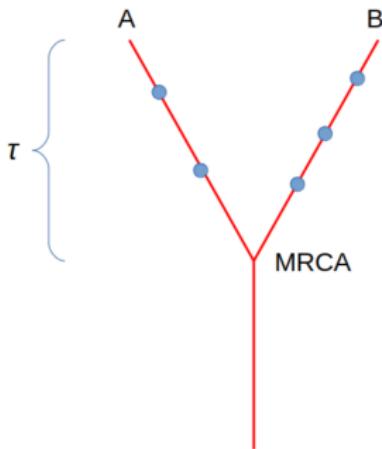
We expect  $\mu r$  mutations in  $r$  generations. If we measure time by  $2N$  generations, that is  $t = r/(2N)$ , we expect  $2N\mu t$  mutations on a lineage of length  $t$ .

Since  $\mathbb{E}[t] = 1$  and there are two lineages, the expected number of mutations separating two gene copies is

$$\theta = 4N\mu$$

which is a simple relationship between the amount of genetic variability, mutation rate, and population size.

## Coalescence in large populations



The expected number of mutations occurring on 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 coalescent tree and a subsequent process in which mutations are distributed across the lineages of the tree at rate  $\theta/2$ .

## Infinite Sites Model

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



Figure 5: The sequence is infinitely long so that the chance of two mutations hit the same site is essentially zero.

## Infinite Sites Model

The sites at which some of the individuals differ are called **segregating sites** or **single nucleotide polymorphisms** (SNPs).

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, 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	gagac	gatag
Sequence 3	aggaa	ggaac	gagac	gatag
Sequence 4	aggag	ggacc	gagac	gatag
Sequence 5	aggag	ggacc	gagac	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 6: 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 7: 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  $\pi$ .

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

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 8: What is the value of  $\pi$ ?

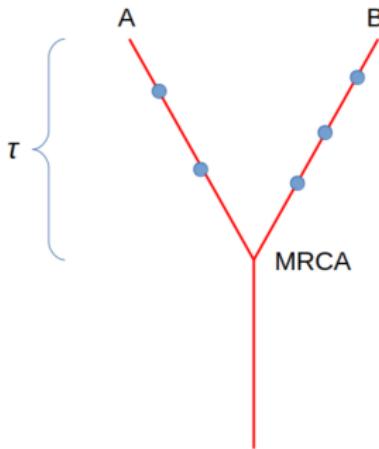
## Tajima's estimator

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

Figure 8: What is the value of  $\pi$ ?

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

## Coalescence in large populations



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

$$\mathbb{E}[d_{i,j}] = \theta$$

$$\mathbb{E}[\pi] = \theta$$

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

## Watterson's estimator

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

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

$$\mathbb{E}[\hat{\theta}_W] = \theta$$

## Watterson's estimator

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

Figure 9: 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 9: 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?

## Effective population size ( $N_e$ )

The number of individuals in the 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

- Expected heterozygosity,
- Expected number of pairwise differences ( $\approx \theta$ ).
- Rate of coalescence.
- ...

## Effective population size ( $N_e$ )

E.g. “A population with an effective size of 200 with respect to heterozygosity harbours the same amount of heterozygosity as a Wright-Fisher population of 200 individuals.”

The true number of individuals in the population can be very different from its effective population size! (Often a lot larger.)

## Effective population size



Figure 10: 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 in offspring.

$N_e$  is equal to harmonic mean if sizes fluctuate.  
→ Smaller sizes have more impact.

## Effective population size ( $N_e$ )



The effective population size with unequal sex ratio is

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

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

## Summary statistics

Possible summaries of DNA sequence data are:

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

but they do not 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 proportions of “1” alleles with a frequency of  $1/5$ ,  $2/5$ ,  $3/5$  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  $4/5$ .

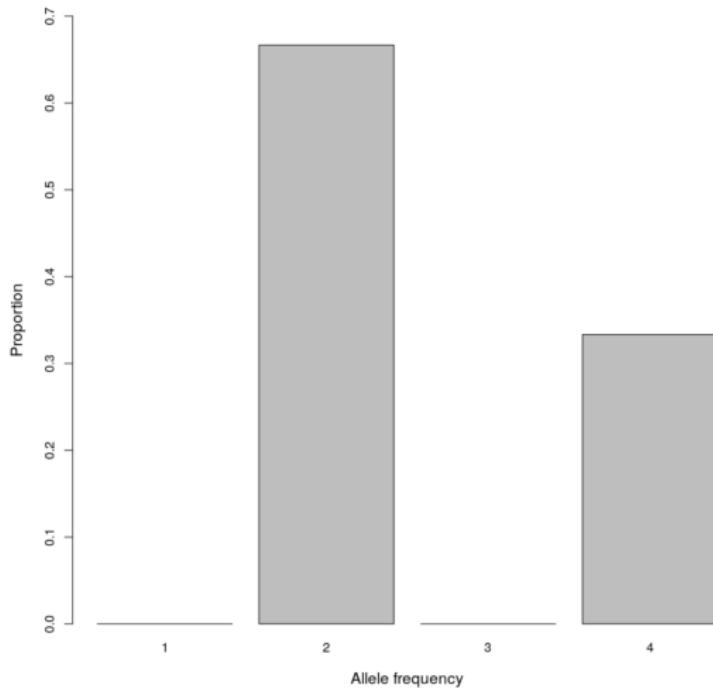
The proportions of “1” alleles with a frequency of  $1/5$ ,  $2/5$ ,  $3/5$  and  $4/5$  in the sample are  $f_1 = 0$ ,  $f_2 = 2/3$ ,  $f_3 = 0$  and  $f_4 = 1/3$ .

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

for a sample of  $n$  haploid individuals.

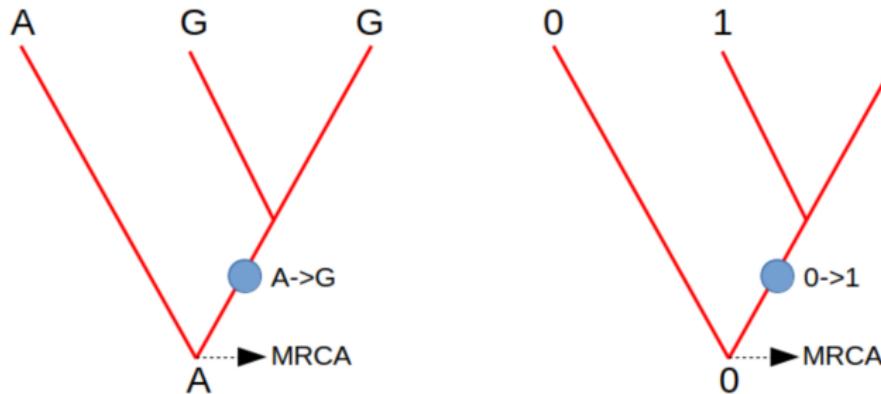
# The Site Frequency Spectrum (SFS)

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



# Alleles

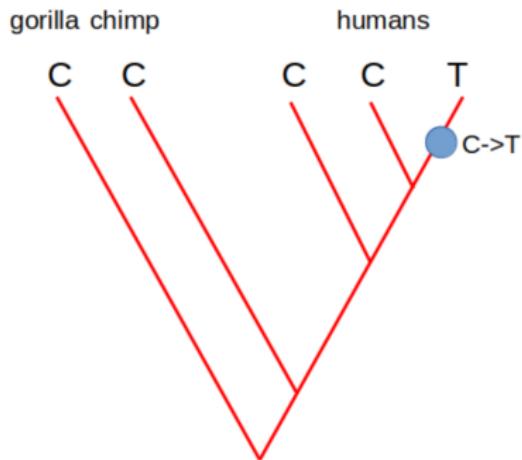
- **Ancestral** allele is the allele found in the MRCA of the sample.
- **Derived** allele (or mutated) is an allele that is not ancestral.



# 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.



# 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

$$\mathbb{E}[f_i] = \frac{1/i}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

with  $i = 1, 2, \dots, n - 1$ .

# Expected Site Frequency Spectrum

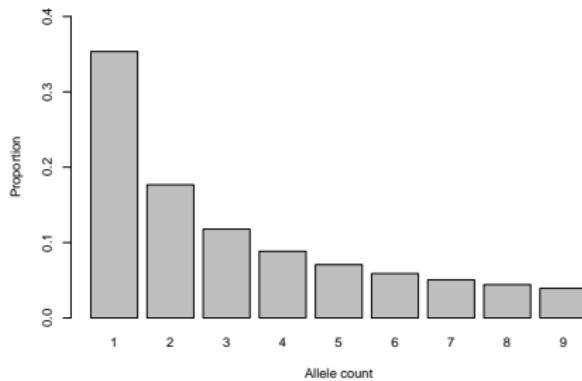


Figure 11: The expected SFS for  $n = 10$ .

# Alleles

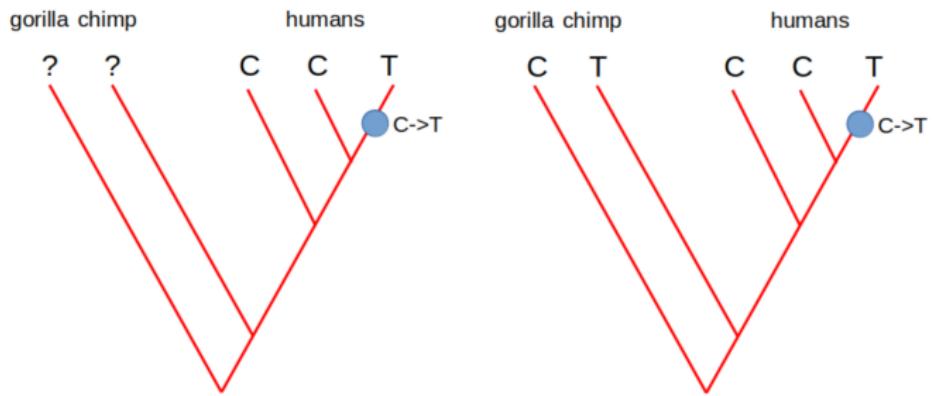


Figure 12: Uncertain ancestral allele.

## The Site Frequency Spectrum (SFS)

If no information on the ancestral allele is available, we can fold the frequency spectrum.

The **folded frequency spectrum**  $f^*$  is obtained by adding together the frequencies of the derived and ancestral alleles.

$$f_j^* = f_j + f_{n-j} \text{ for } j < n/2 \text{ and}$$

$$f_j^* = f_j \text{ for } j = n/2$$

only defined for values of  $f_j^* \leq n/2$ .

## The folded SFS

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

$$\vec{f} = (f_1 = 0, f_2 = 2/3, f_3 = 0, f_4 = 1/3)$$
$$\vec{f}^* =$$

## The folded SFS

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

$$\vec{f} = (f_1 = 0, f_2 = 2/3, f_3 = 0, f_4 = 1/3)$$
$$\vec{f}^* = (f_1^* = 1/3, f_2^* = 2/3)$$

## Expected folded SFS

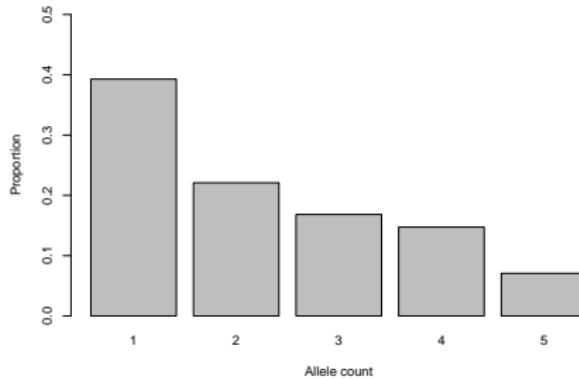


Figure 13: The expected folded SFS for  $n = 10$ .

## Tree shape and population size

Measured in number of generations, the expected coalescence time for  $k$  lineages is  $2N/[k(k - 1)]$ .

Constant population size



Increasing population size



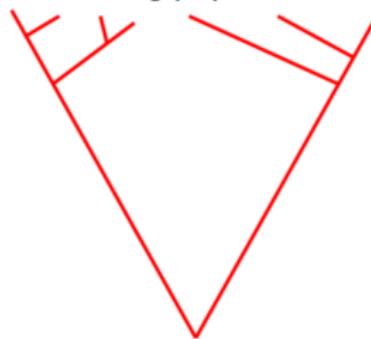
## Tree shape and population size

Measured in number of generations, the expected coalescence time for  $k$  lineages is  $2N/[k(k - 1)]$ .

Constant population size



Decreasing population size



# Intended Learning Outcomes

## Coalescent 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.

# 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.

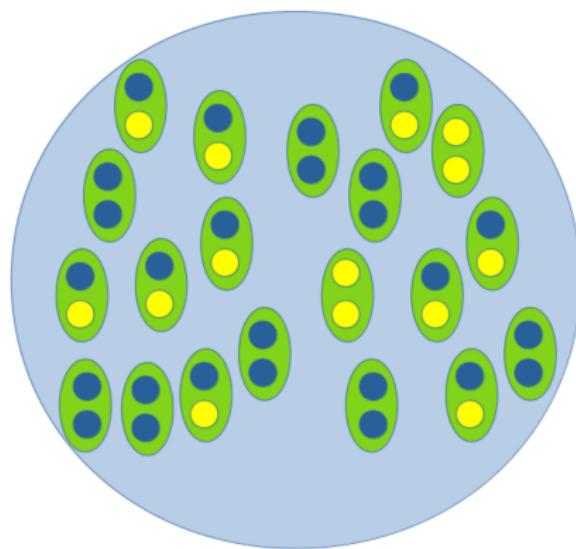
## Population subdivision

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

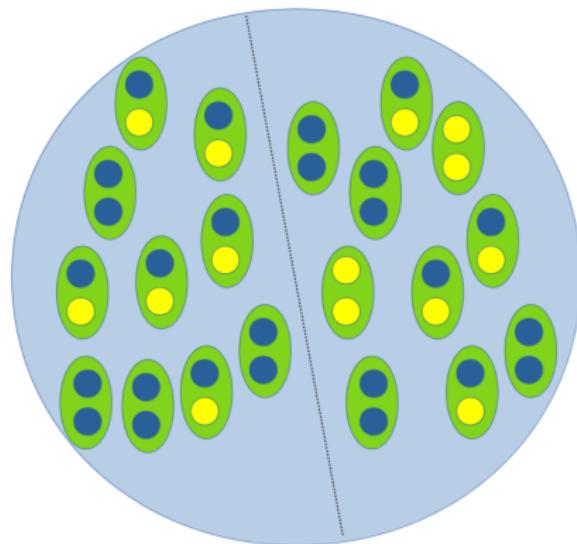
Population subdivision is important to

- Understand the effects of drift and natural selection.
- Plan conservation strategies for rare or endangered species.

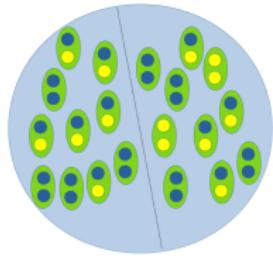
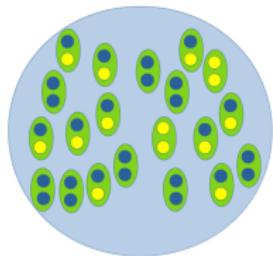
## Population subdivision



## Population subdivision



## Allele frequencies in a subdivided population



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

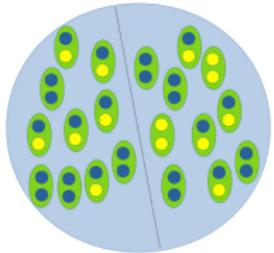
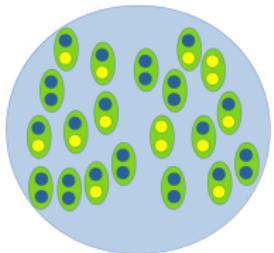
The average frequency of allele  $A$  when pooling the two subpopulations is

$$f_A = \frac{2N_1 f_{A1} + 2N_2 f_{A2}}{2N_1 + 2N_2}$$

if  $N_1 = N_2$

$$f_A = \frac{f_{A1} + f_{A2}}{2}$$

## Allele frequencies in a subdivided population



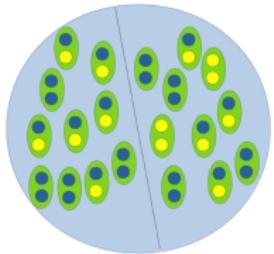
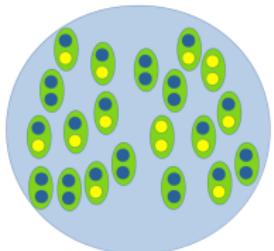
The proportion of heterozygous individuals is

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

which is the expected heterozygosity when both populations are sampled.

*S* in  $H_S$  stands for “in the subdivided population”

## Allele frequencies in a subdivided population



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

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

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

## Heterozygosity in a subdivided population

After some rearrangements we have

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

and

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

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

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

## Quantifying population subdivision

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

$F_{ST}$  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 14: 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

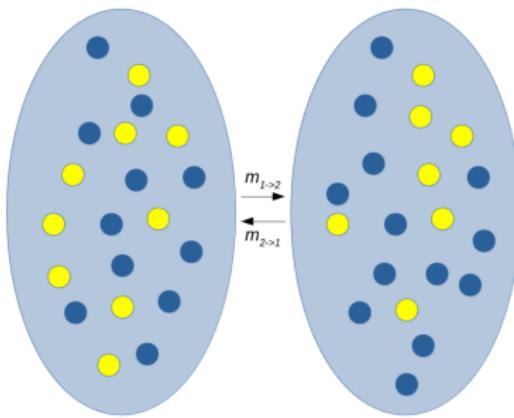


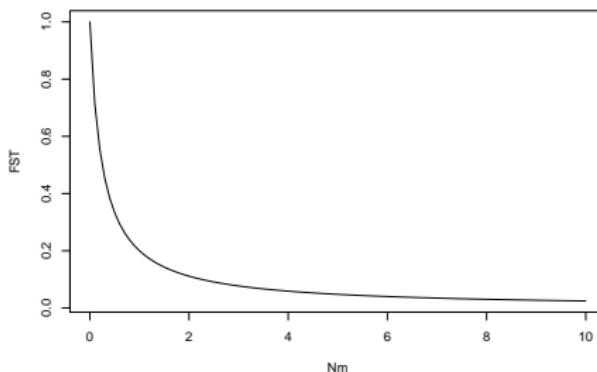
Figure 15: An individual from one population is replaced with an individual from the other with probability  $m$  (migration rate).

## $F_{ST}$ and migration rates

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

$$F_{ST} = \frac{1}{1 + 4Nm_T}$$

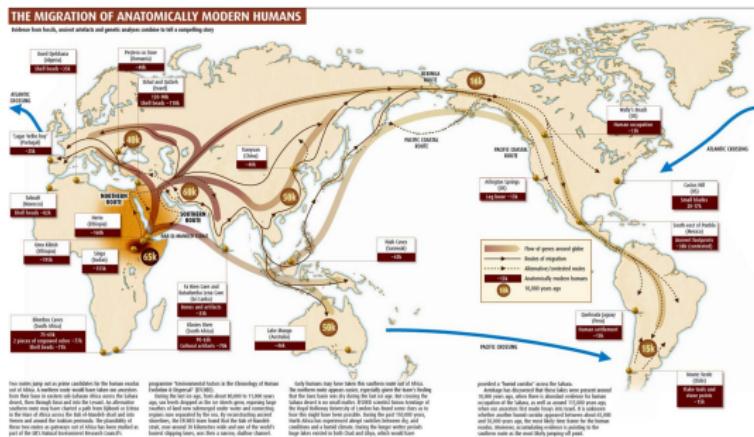
with  $m_T$  being the total number of migrants.



# “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.

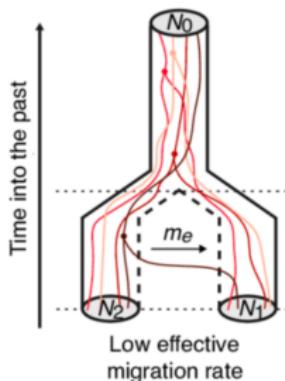


Figure 16: TMRCA overestimates the divergence time.

## 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

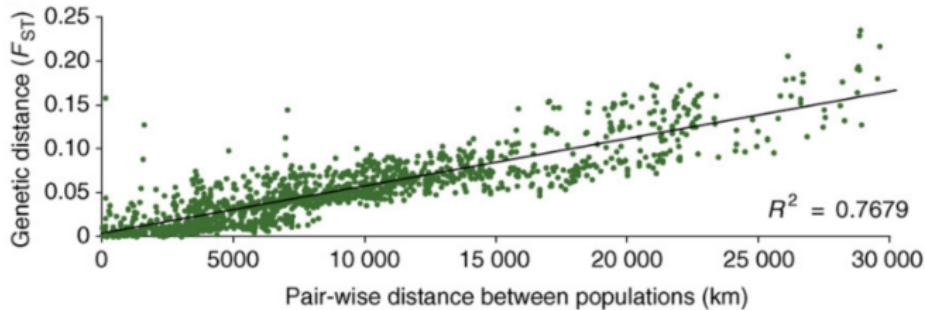


Figure 17: Isolation by distance in human populations.

# 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.