Evolutionary Thinking 2022
TA session
week 5 – Basis of Population Genetics

Jilong Ma
aujilongm@birc.au.dk

Outline

1. Learning outcome of today

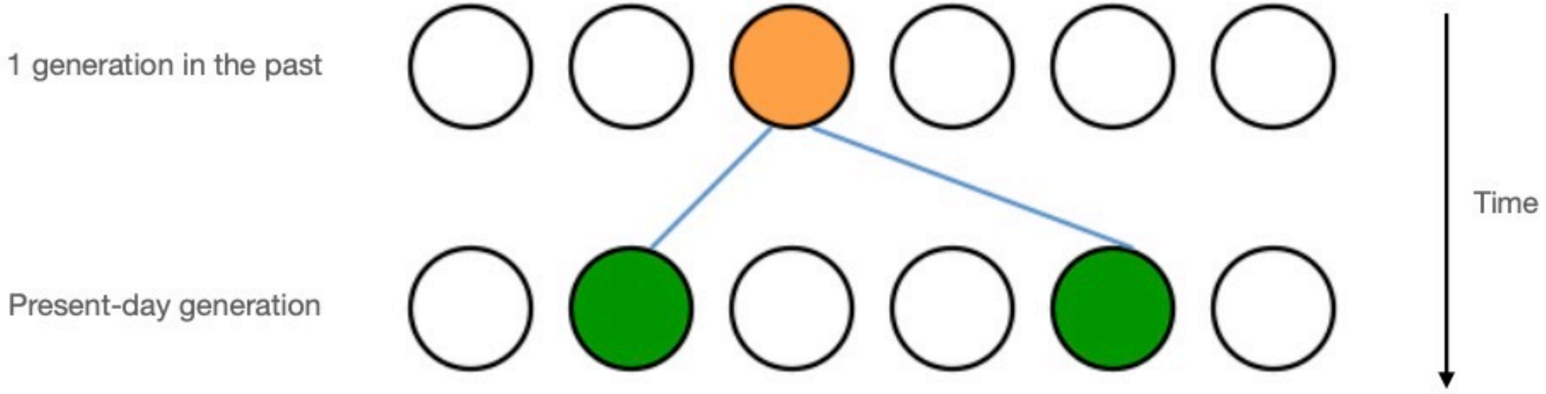
Coalescence Theory (Fri)

– Process, Tree and Tree length, Site Frequency Spectrum

2. R Exercises

Coalescence in a sample of two sequences

P[2 samples have the same parent in the previous generation] =

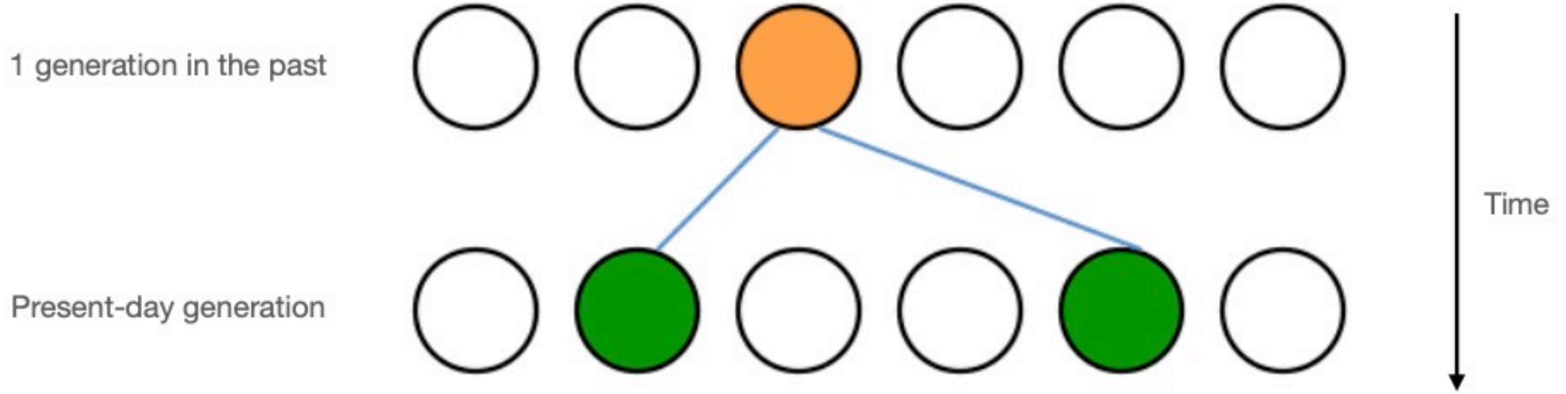


Slides from Fernando Racimo
“Intro to popgen”

Coalescence in a sample of two sequences

P[2 samples have the same parent in the previous generation] =

$$2N\frac{1}{2N}\frac{1}{2N} = \frac{1}{2N}$$

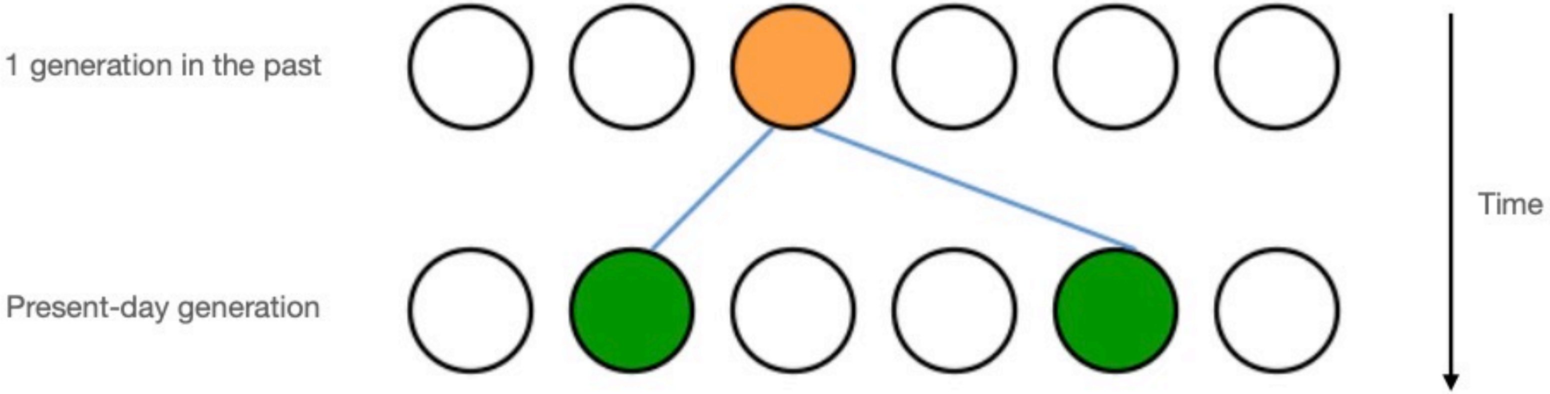


Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples have the same parent in the previous generation] =

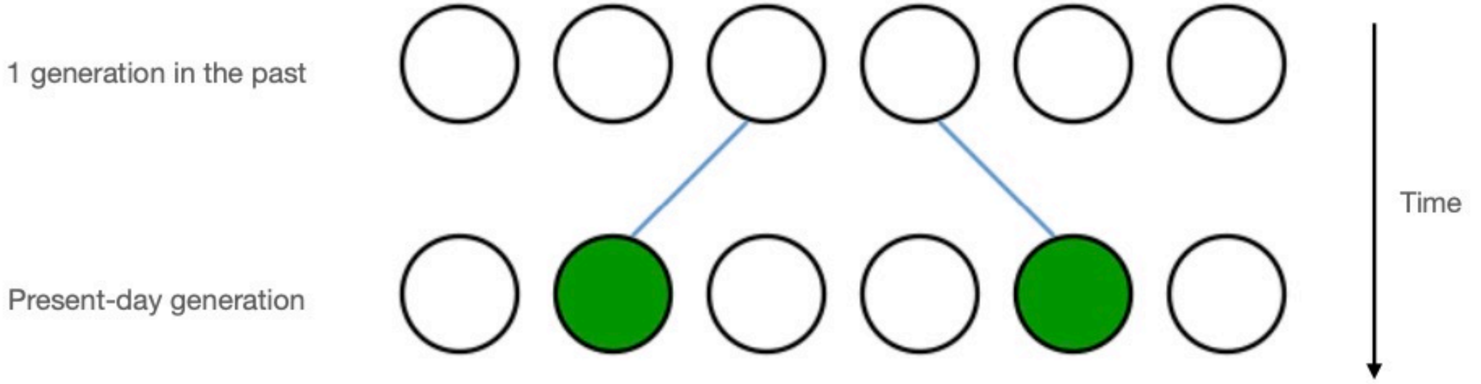
$$2N\frac{1}{2N}\frac{1}{2N} = \frac{1}{2N}$$



Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples do not have the same parent in the previous generation] =

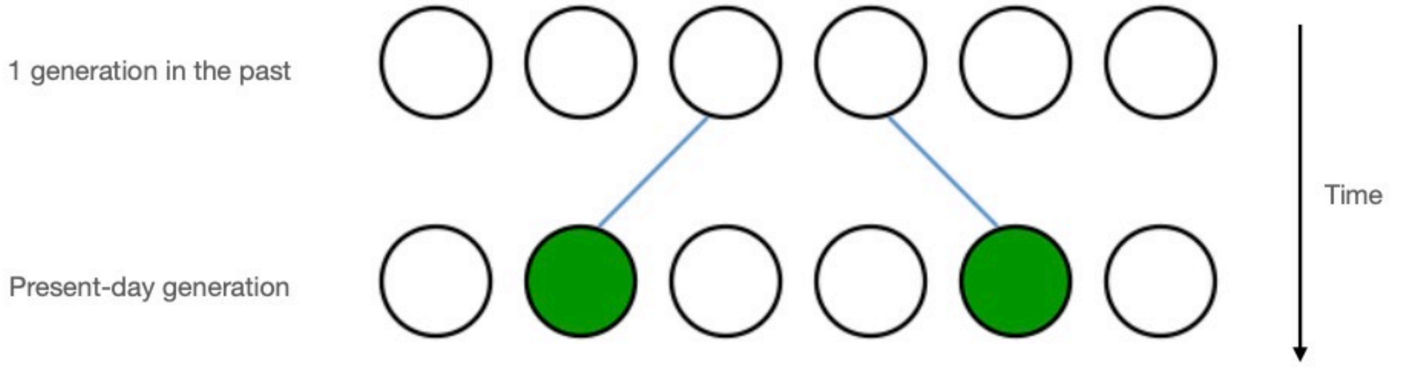


Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples do not have the same parent in the previous generation] =

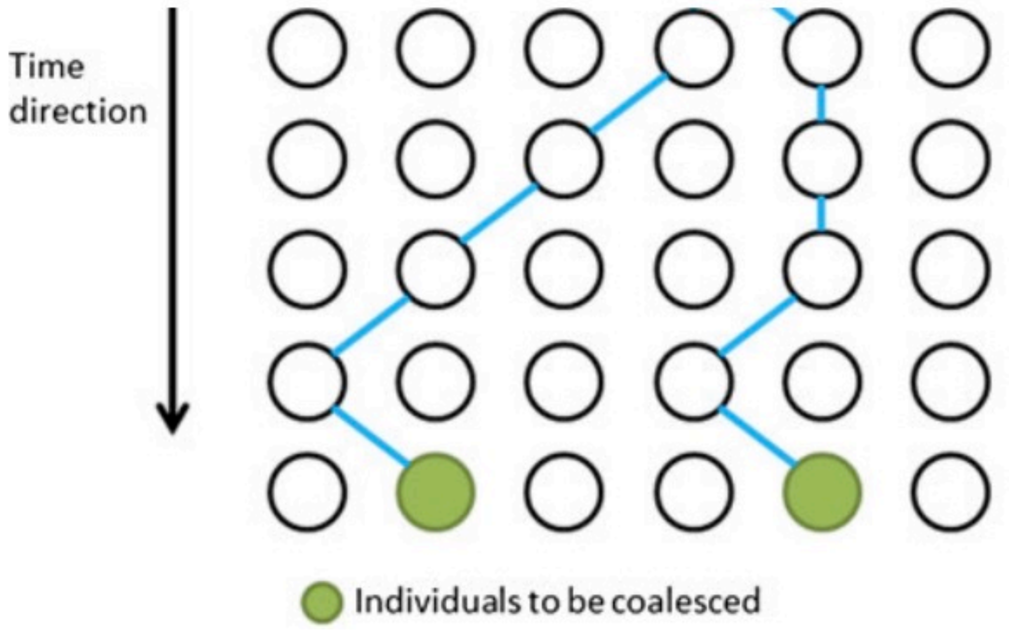
$$1 - \frac{1}{2N}$$



Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples do not find a common ancestor in **r** generations] =

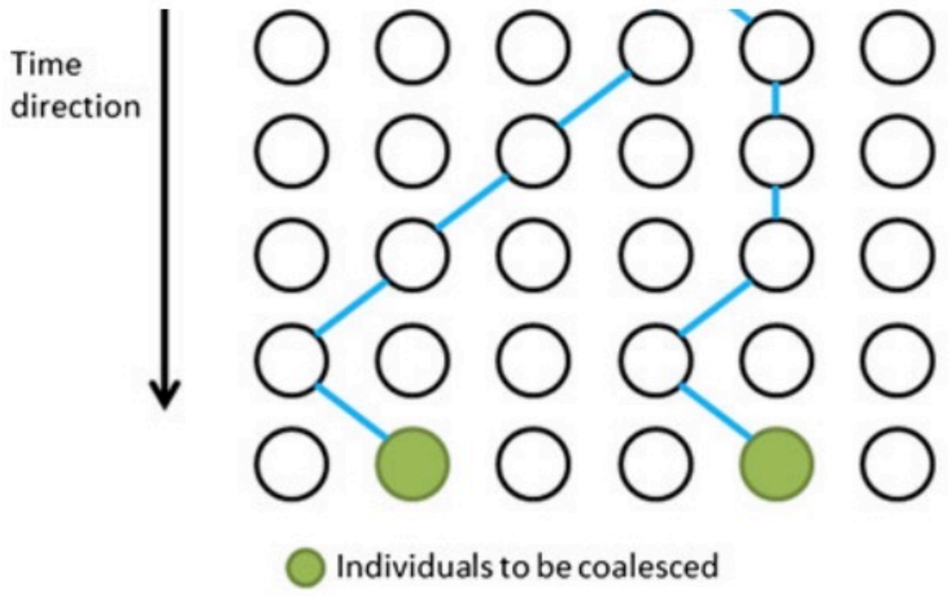


Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples do not find a common ancestor in **r** generations] =

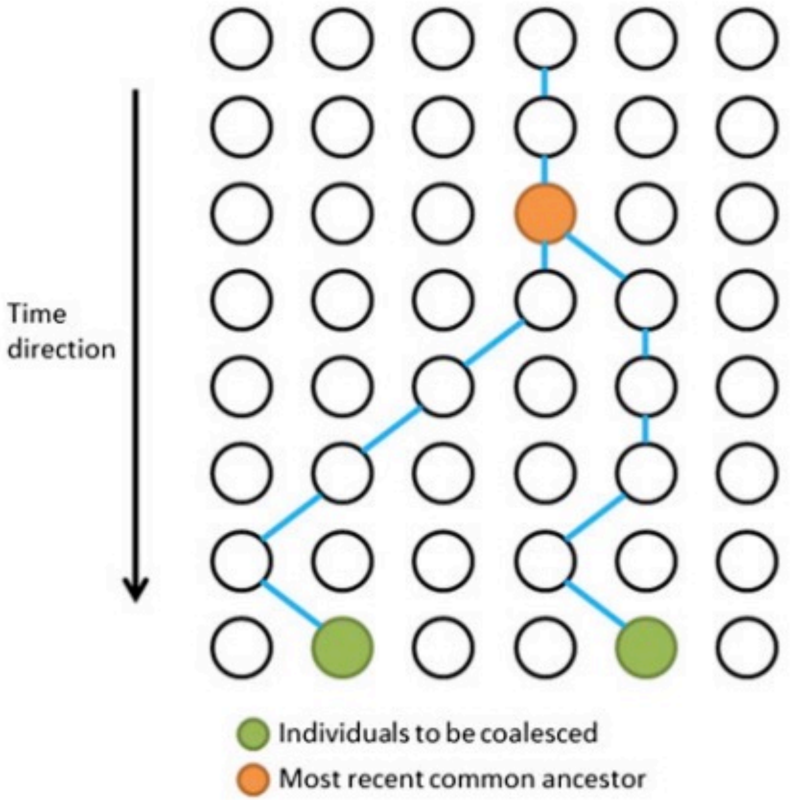
$$\left(1 - \frac{1}{2N}\right)^r$$



Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

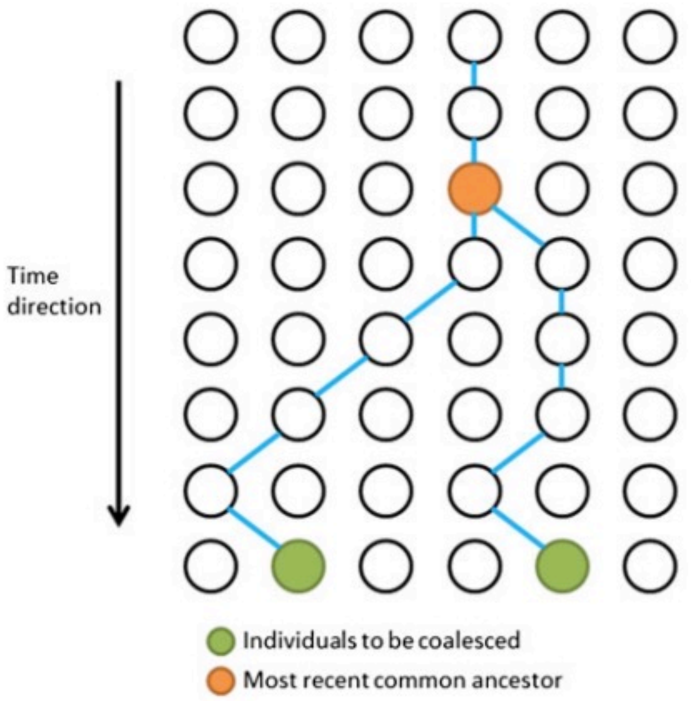
P[2 samples find a common ancestor in **exactly r** generations] =



Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples find a common ancestor in **exactly r** generations] =

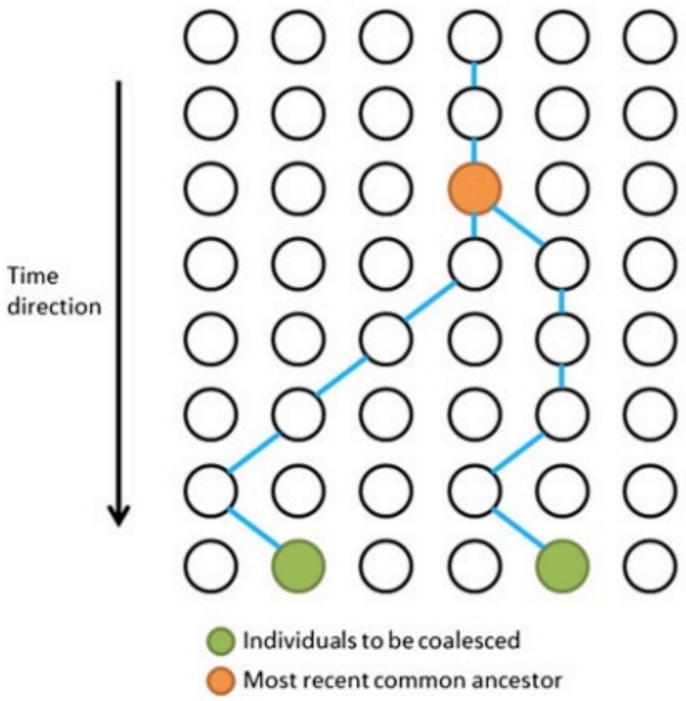


$$\left(1-\frac{1}{2N}\right)^{r-1}\frac{1}{2N}$$

Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples find a common ancestor in **exactly r** generations] =

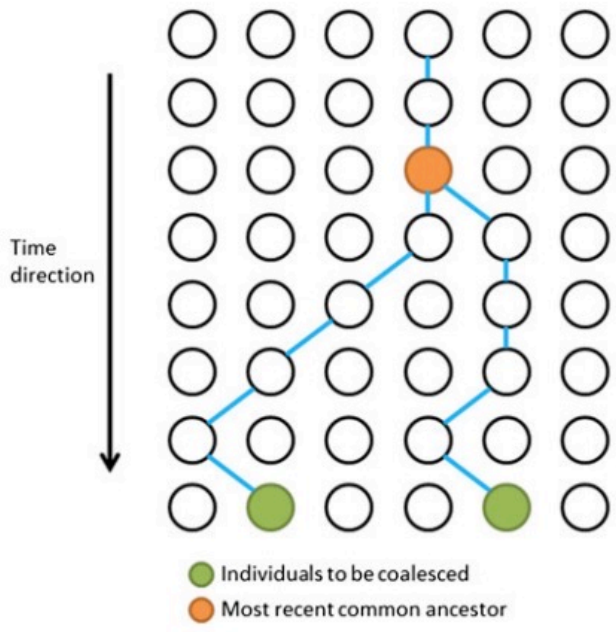


$$\left(1 - \frac{1}{2N}\right)^{r-1} \frac{1}{2N}$$

Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples find a common ancestor in **exactly r** generations] =



$$\left(1-\frac{1}{2N}\right)^{r-1}\frac{1}{2N}$$

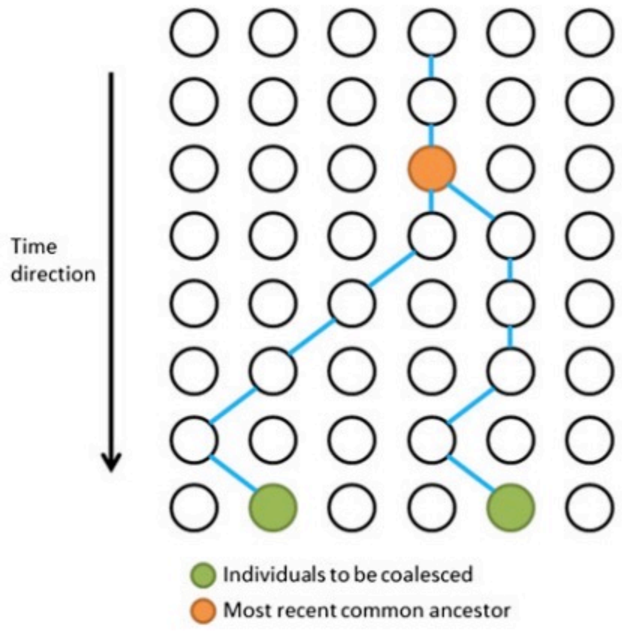
$$\approx \frac{1}{2N}e^{-r/(2N)}$$

This approximation is very good when the population size is large (2N > ~1000)

Slides from Fernando Racimo "Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples find a common ancestor in **exactly r** generations] =



$$\left(1 - \frac{1}{2N}\right)^{r-1} \frac{1}{2N}$$

$$\approx \frac{1}{2N} e^{-r/(2N)}$$

This approximation is very good when the population size is large (2N > ~1000)

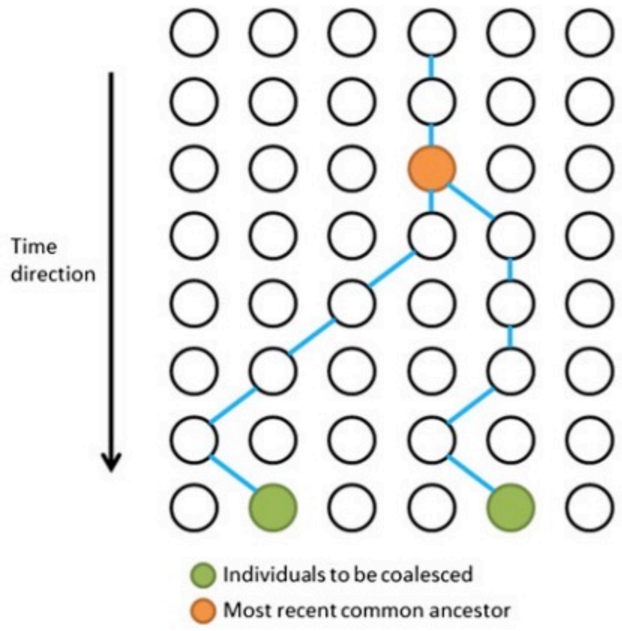
$$= e^{-t}$$

If we measure time in units of 2N generations (1 t = 2N r)

Slides from Fernando Racimo
“Intro to popgen”

Coalescence in a sample of two sequences

P[2 samples find a common ancestor in **exactly r** generations] =



$$\left(1-\frac{1}{2N}\right)^{r-1}\frac{1}{2N}$$

$$\approx \frac{1}{2N}e^{-r/(2N)}$$

This approximation is very good when the population size is large (2N > ~1000)

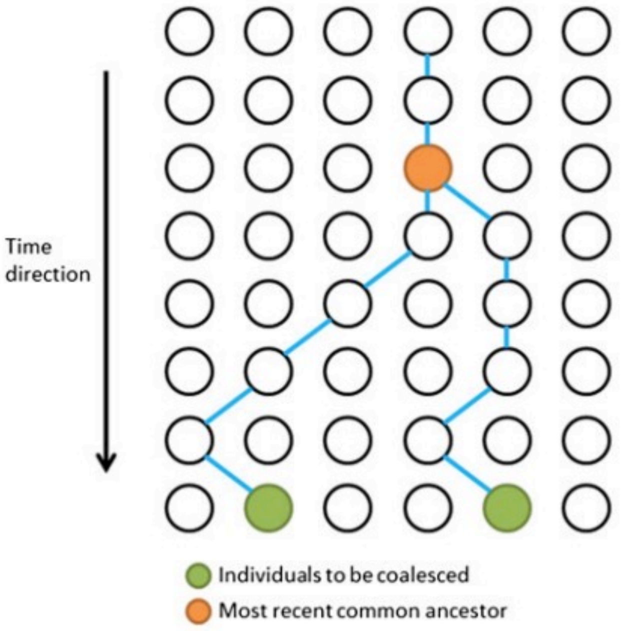
$$= e^{-t}$$

If we measure time in units of 2N generations (1 t = 2N r)

Slides from Fernando Racimo
"Intro to popgen"

Coalescence in a sample of two sequences

P[2 samples find a common ancestor in **exactly r** generations] =



$$\left(1 - \frac{1}{2N}\right)^{r-1} \frac{1}{2N}$$

$$\approx \frac{1}{2N} e^{-r/(2N)}$$

This approximation is very good when the population size is large (2N > ~1000)

$$= e^{-t}$$

This is an **exponential distribution** with rate 1

If we measure time in units of 2N generations (1 t = 2N r)

Slides from Fernando Racimo
"Intro to popgen"

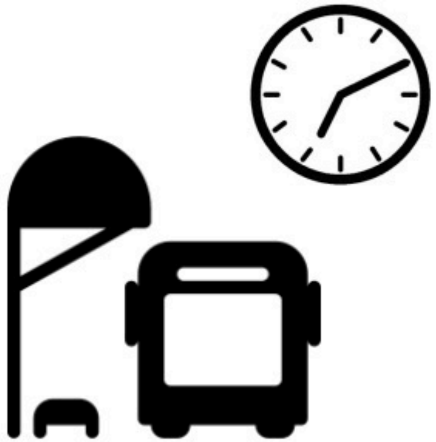
The expected waiting time

- Used to model **waiting times**
  - "What is the probability that I have to wait less than 30 minutes till the next bus arrives?"

Slides from Fernando Racimo
"Intro to popgen"

The expected waiting time

- Used to model **waiting times**
  - "What is the probability that I have to wait less than 30 minutes till the next bus arrives?"
- One parameter: rate (λ)
  - The higher the rate, the less I will have to wait



Slides from Fernando Racimo
"Intro to popgen"

The expected waiting time

One parameter: λ

$$E[T] = \frac{1}{\lambda}$$

Slides from Fernando Racimo
"Intro to popgen"

The expected waiting time

One parameter: λ

$$E[T] = \frac{1}{\lambda}$$

The expected waiting time is the inverse of the rate

If buses arrive at an average rate of λ=4 per hour,
I expect to wait 1/4 of an hour for the next bus, on average

Slides from Fernando Racimo
"Intro to popgen"

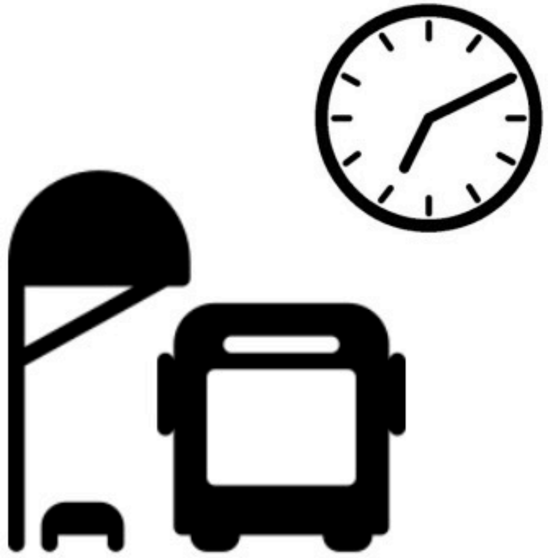
The expected waiting time

"Buses arrive at a rate of λ per hour"

$$f(t) = \lambda e^{-\lambda t}$$

"The expected waiting time for the next bus is $1/\lambda$ hours"

$$E[T] = \frac{1}{\lambda}$$

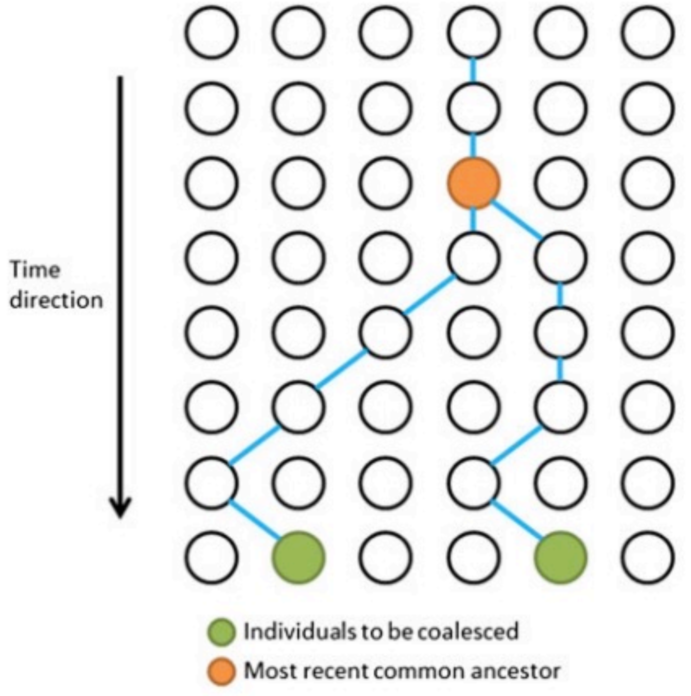


Slides from Fernando Racimo
"Intro to popgen"

The expected waiting time

"A coalescent event between 2 lineages occurs at a rate of 1/2N per generation"

$$f(t) = \frac{1}{2N} e^{-t/(2N)}$$



Slides from Fernando Racimo
"Intro to popgen"

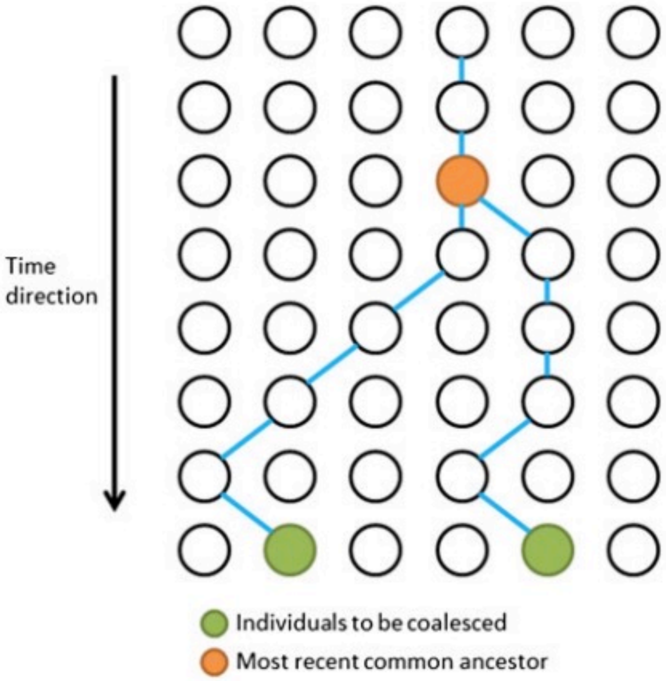
The expected waiting time

"A coalescent event between 2 lineages occurs at a rate of 1/2N per generation"

$$f(t) = \frac{1}{2N} e^{-t/(2N)}$$

"The expected waiting time for a coalescent event is 2N generations"

$$E[T] = 2N$$

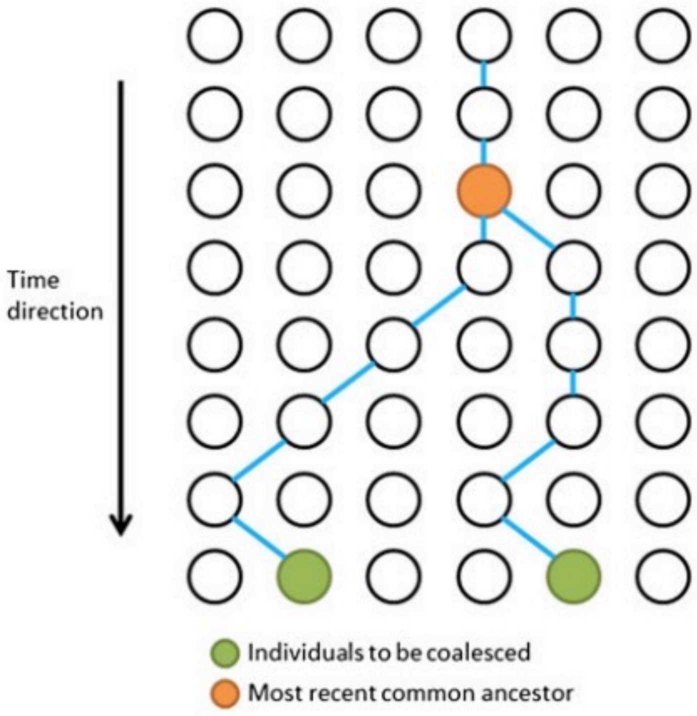


Slides from Fernando Racimo
"Intro to popgen"

The expected waiting time

"A coalescent event between 2 lineages occurs at a rate of 1 per coalescent unit (1 unit = 2N generations)."

$$f(t) = e^{-t}$$



Slides from Fernando Racimo
"Intro to popgen"

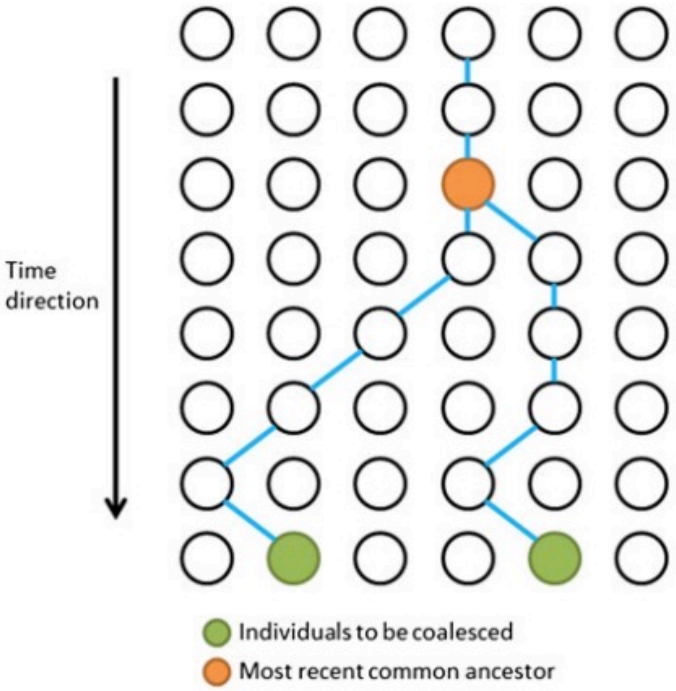
The expected waiting time

"A coalescent event between 2 lineages occurs at a rate of 1 per coalescent unit (1 unit = 2N generations)."

$$f(t) = e^{-t}$$

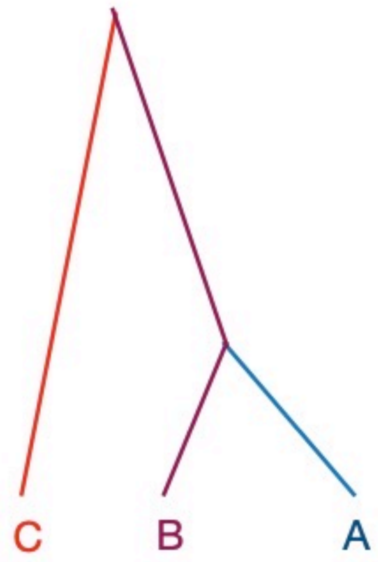
"The expected waiting time for a coalescent event is 1 coalescent unit"

$$E[T] = 1$$



Slides from Fernando Racimo
"Intro to popgen"

Tree with 3 samples



Slides from Fernando Racimo
“Intro to popgen”

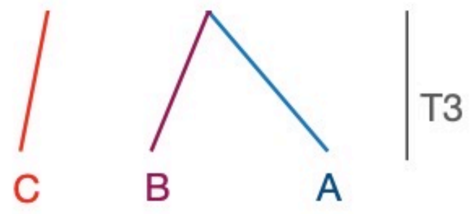
Tree with 3 samples

Slides from Fernando Racimo
"Intro to popgen"

Tree with 3 samples

T3 = time while there are 3 lineages

$$\textbf{Rate: } \lambda_{T3} = 1 * \binom{3}{2} = 3$$



Slides from Fernando Racimo
“Intro to popgen”

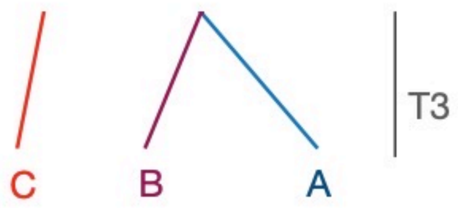
Tree with 3 samples

T3 = time while there are 3 lineages

Rate: $\lambda_{T3} = 1 * \binom{3}{2} = 3$

…because there are 3 “competing pairs” of lineages fighting for the opportunity to coalesce (each at rate 1): A+B, B+C, A+C

Expected time: $E[T3] = \frac{1}{\binom{3}{2}} = \frac{1}{3}$



Slides from Fernando Racimo
“Intro to popgen”

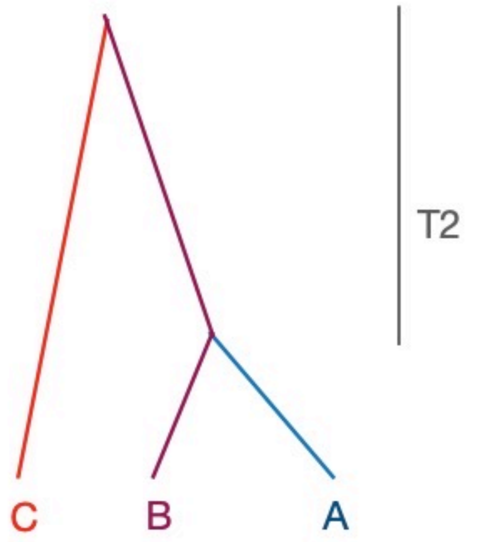
Tree with 3 samples

T2 = time while there are 2 lineages

$$\text{Rate: } \lambda_{T2} = \frac{1}{\binom{2}{2}} = 1$$

...because there is just one possible pair
that can coalesce (at rate 1)

Expected time: $E[T2] = 1$

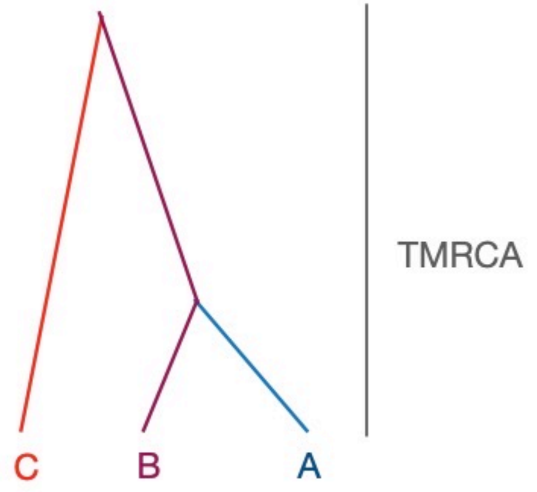


Slides from Fernando Racimo
"Intro to popgen"

Tree with 3 samples

$E[T_{MRCA}] = T_2 + T_3 = 1 + 1/3$ coalescent units

$E[T_{MRCA}] = (1 + 1/3) * 2N$ generations

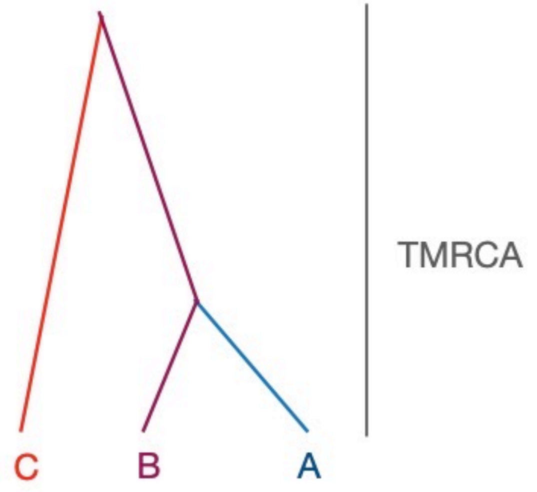


Slides from Fernando Racimo
"Intro to popgen"

Tree with 3 samples

$$E[T_{MRCA}] = T_2 + T_3 = 1 + 1/3 \text{ coalescent units}$$

$$E[T_{MRCA}] = (1 + 1/3) * 2N \text{ generations}$$

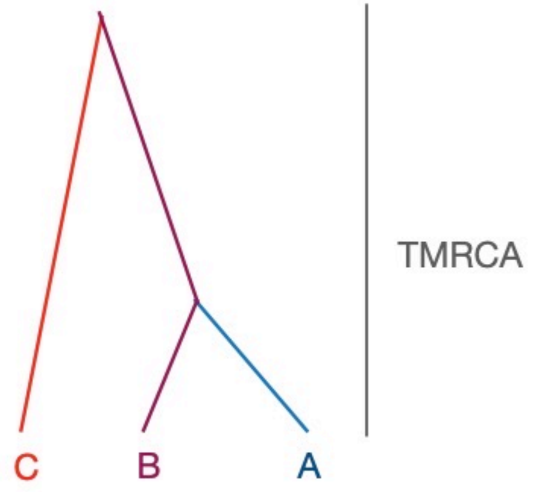


Slides from Fernando Racimo
“Intro to popgen”

Tree with 3 samples

$$E[T_{MRCA}] = T_2 + T_3 = 1 + 1/3 \text{ coalescent units}$$

$$E[T_{MRCA}] = (1 + 1/3) * 2N \text{ generations}$$



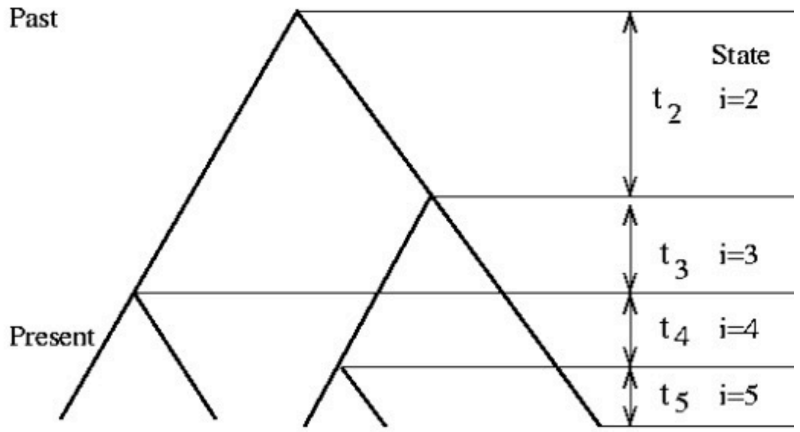
Slides from Fernando Racimo
“Intro to popgen”

Tree with 3 samples

$$E[T_{MRCA}] = T_2 + T_3 + \ldots + T_n$$

$$E[T_{MRCA}] = \sum_{i=2}^{n} \frac{1}{\binom{i}{2}} = 2\left(1 - \frac{1}{n}\right)$$

where n is our sample size



Slides from Fernando Racimo
"Intro to popgen"

Variance of an exponential distribution

One parameter: λ

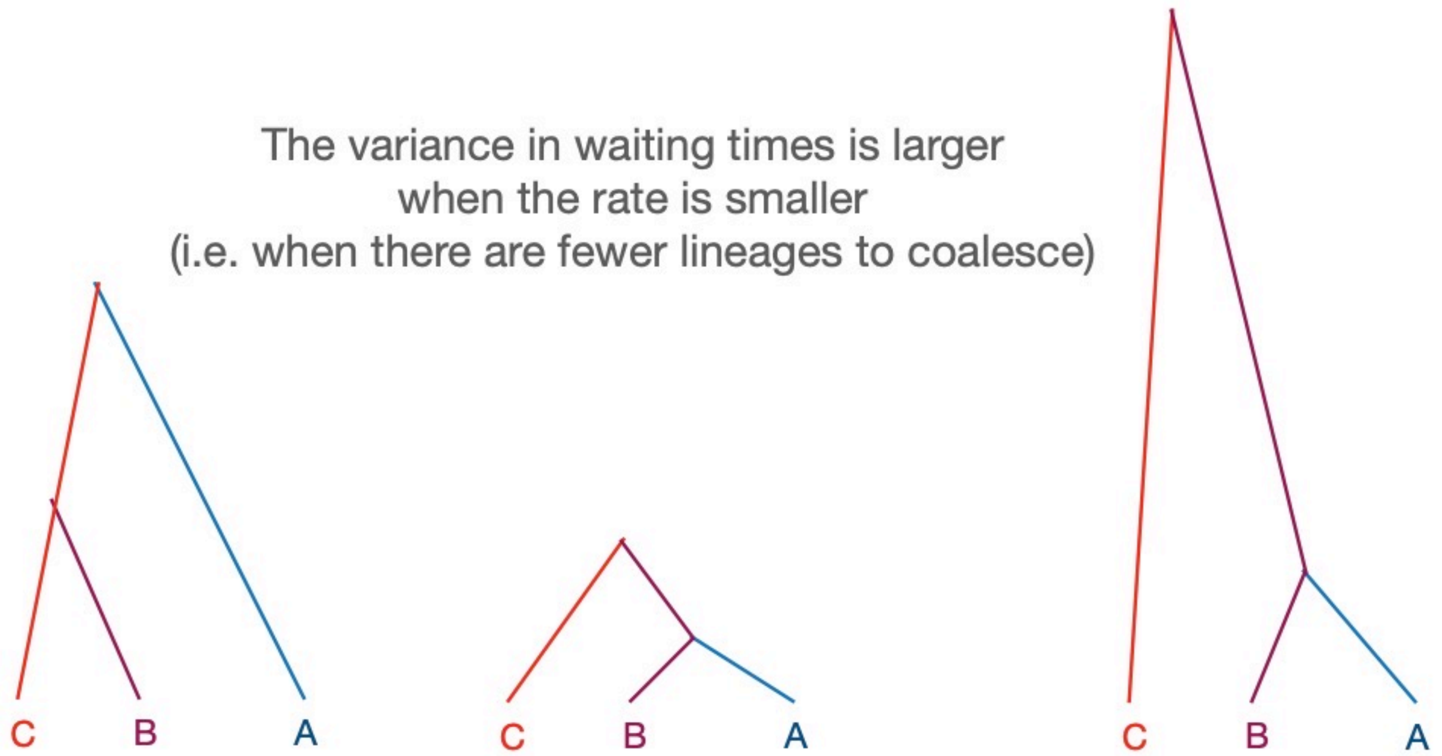
$$E[T] = \frac{1}{\lambda}$$

$$Var[T] = \frac{1}{\lambda^2}$$ → The variance in waiting times is larger when the rate is smaller

Slides from Fernando Racimo
"Intro to popgen"

Variance in coalescence times

The variance in waiting times is larger
when the rate is smaller
(i.e. when there are fewer lineages to coalesce)



Slides from Fernando Racimo
“Intro to popgen”

Exercises

R exercises

Left from Wednesday
Chapter 1:
1.1-1.4
Chapter 2:
2.1-2.3
2.6-2.9
Chapter 3:
3.1-3.8

AARHUS
UNIVERSITY