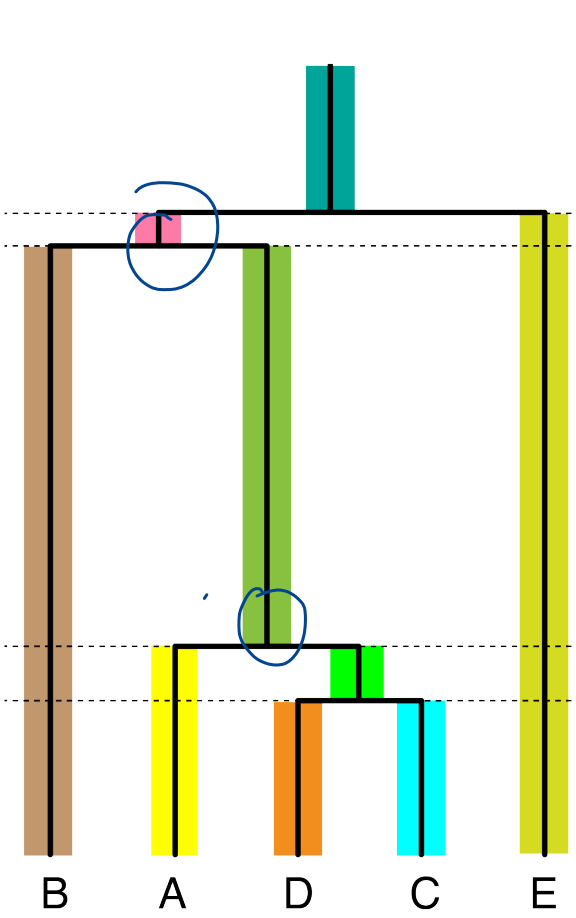
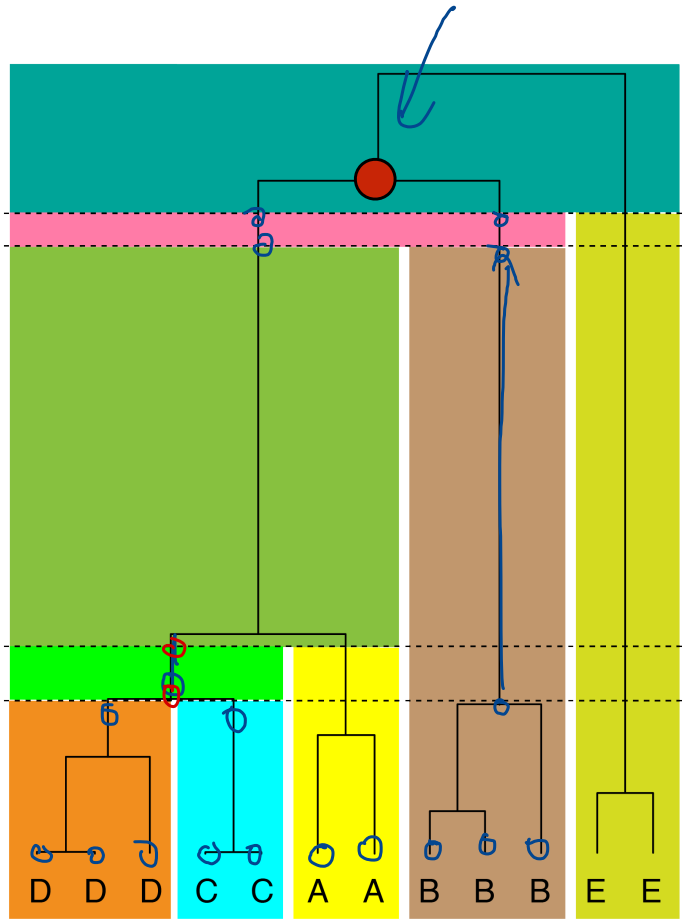
Deep coalescence can cause conflict among gene trees

Gene tree conflict

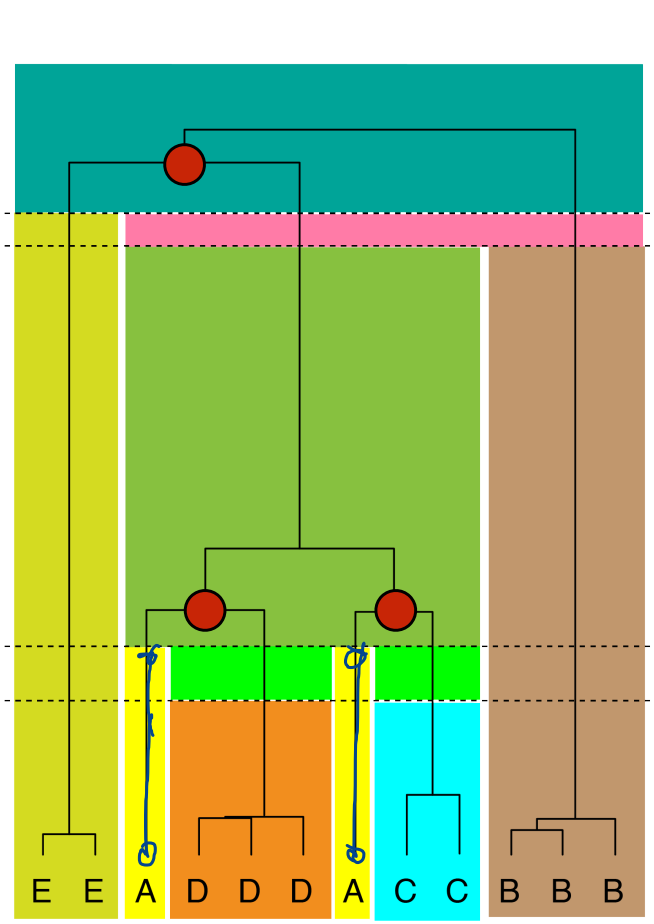


species tree



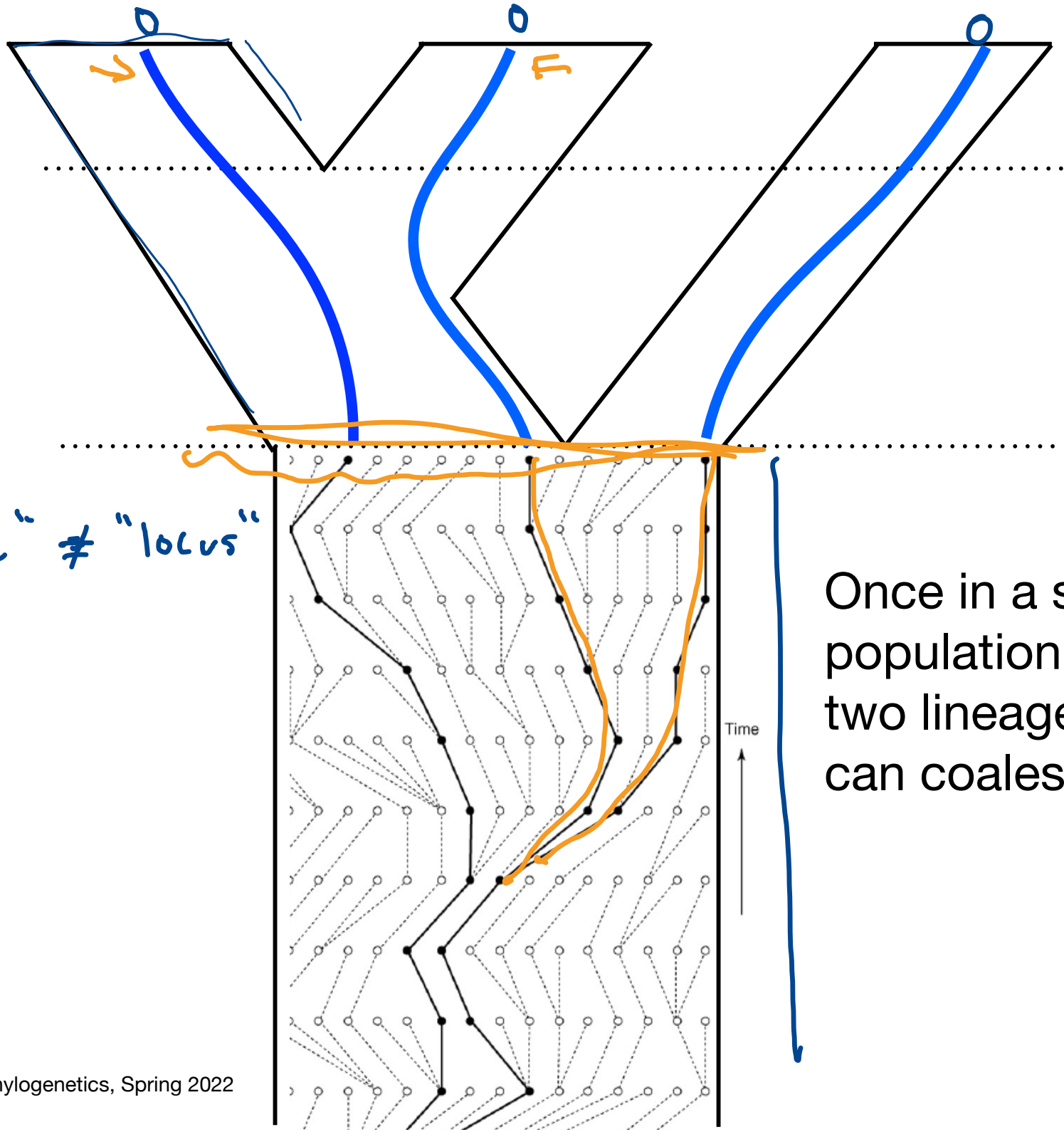
locus gene 3

Gene 3 agrees with the species tree (even though there is one deep coalescence)



locus gene 7

Gene 7 conflicts with the species tree (and thus also with gene 3)



"gene" ≠ "locus"

Once in a single population, any two lineages can coalesce...

Kuhner, M. K. 2009. Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. 24:86-93.

What is coalescence?

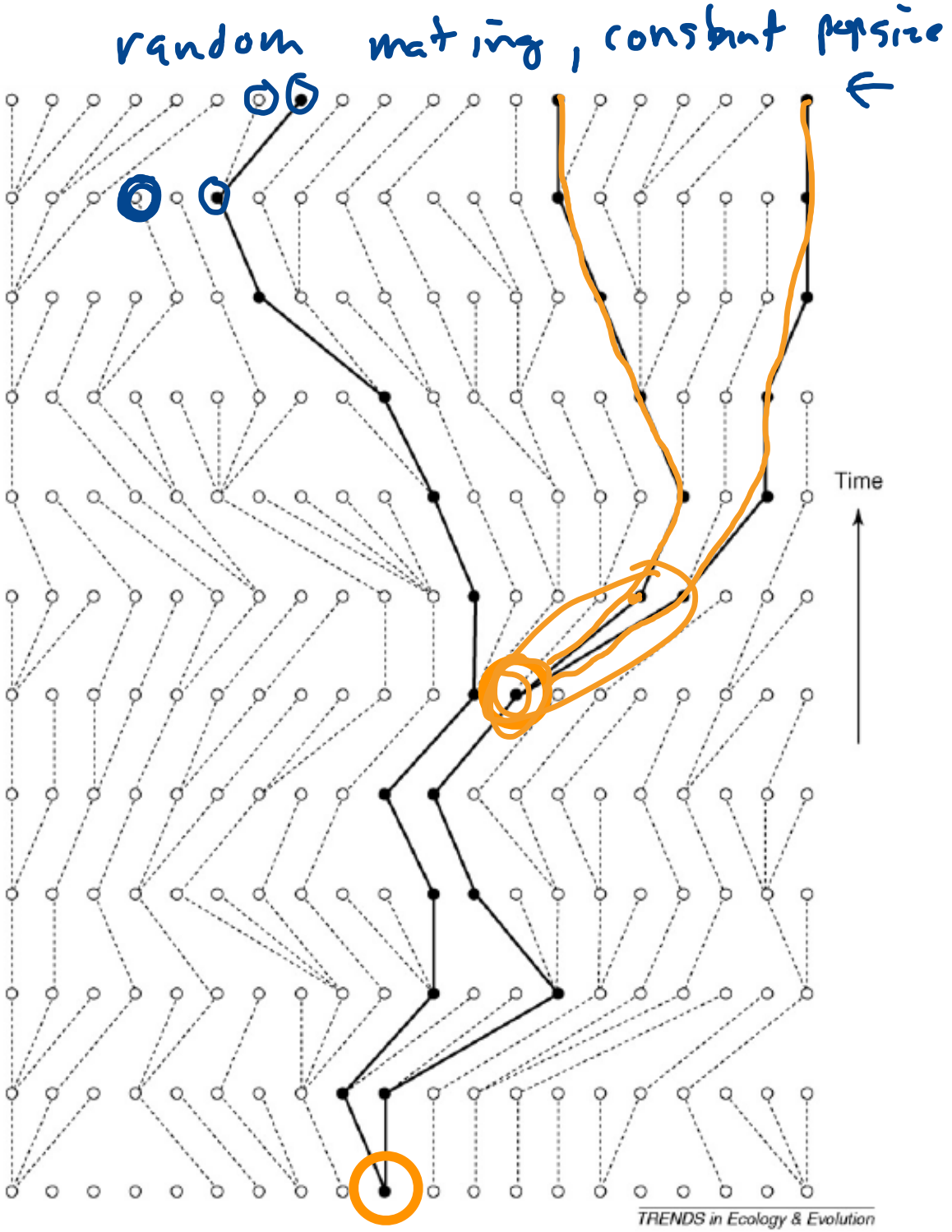
Coalescence

Some individuals **leave no offspring** to the next generation

Therefore (assuming population size remains constant over time), some offspring genes must have been **copied from the same parent** gene.

This merging (looking backwards in time) represents a **coalescence**.

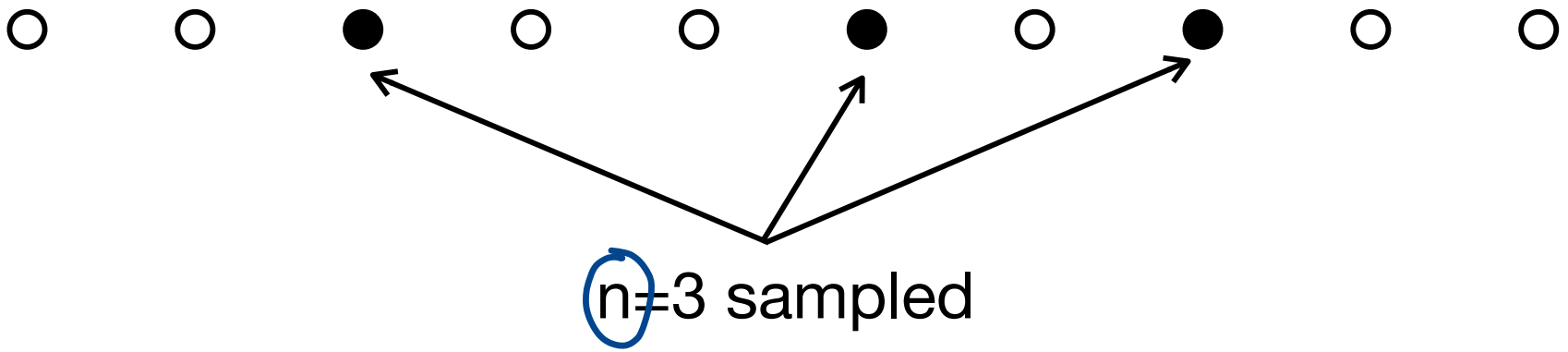
All genes sampled must coalesce by some time in the past, and this history of coalescence is the **gene genealogy** (gene tree).



Kuhner 2009

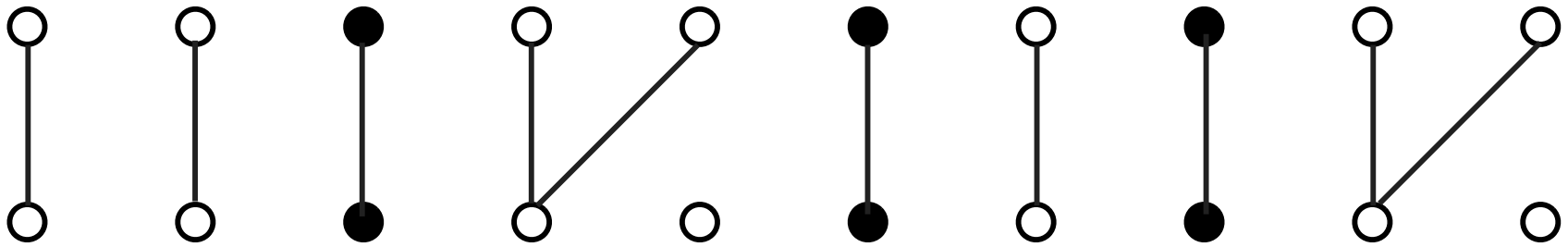
Understanding coalescence

N=10 haploid individuals in a population today



The coalescent process

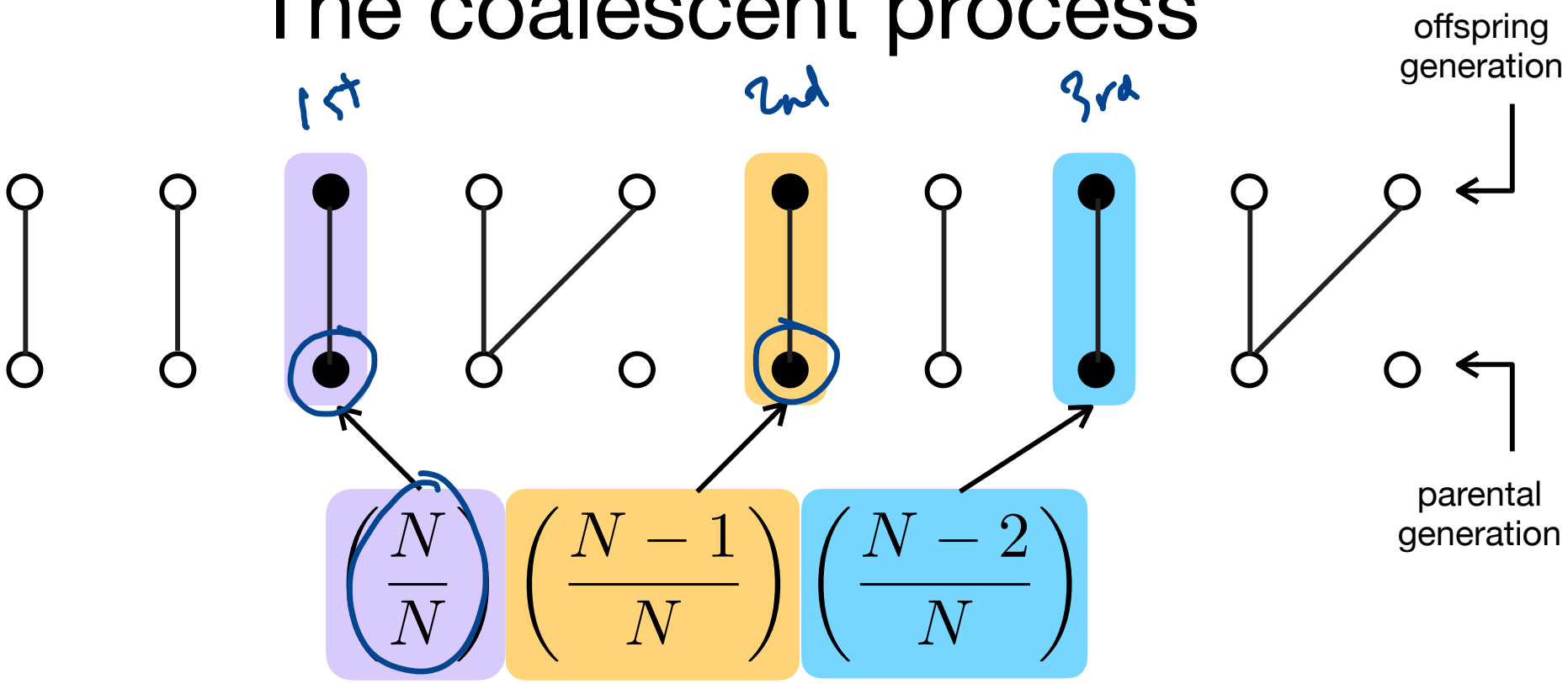
N=10 haploid individuals in a population today



N=10 haploid individuals in previous generation

Each *sampled* gene had a distinct ancestor, **no** coalescent events affected our *sampled* genes

The coalescent process

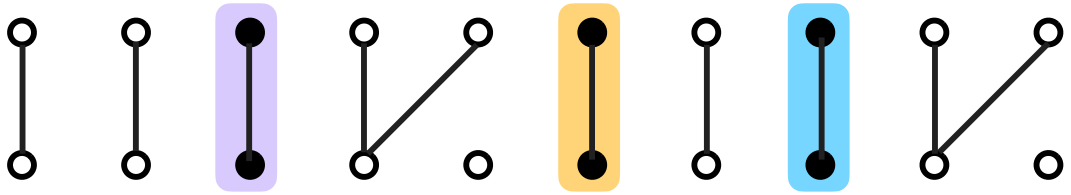


1st gene considered must have had a parent

2nd gene considered can have any parent except the one already taken by 1st gene

3rd gene considered can have any parent except the 2 already taken by 1st and 2nd genes

Probability that all n=3 sampled genes had *distinct* parents



The coalescent process

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

following $n = 3$ lineages

$$= 1 - \frac{1}{N} - \frac{2}{N} + \frac{2}{N^2}$$

Can ignore terms like this if N is large

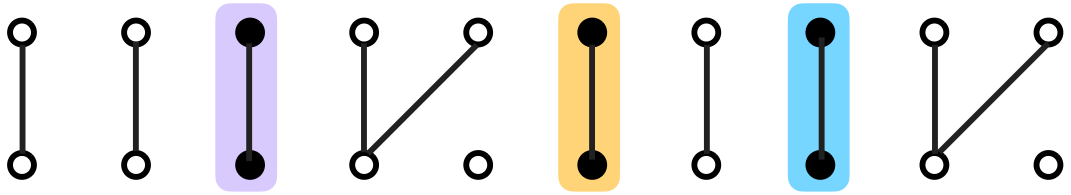
pr.(no coal.)

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

sum of natural numbers up to n-1

Probability of no coalescence in 1 generation given:

- n current sampled lineages (in this case n=3)
- N constant and somewhat large (in this case N=10)



The coalescent process

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

following $n = 3$ lineages

$$= 1 - \frac{1}{N} - \frac{2}{N} + \frac{2}{N^2}$$

Can ignore terms like this if N is large

pr. (no. coal.)

$$\approx 1 - \frac{\binom{n}{2}}{N}$$

number of ways of choosing 2 things out of n things

Probability of no coalescence in 1 generation given:

- n current sampled lineages (in this case n=3)
- N constant and somewhat large (in this case N=10)

$\binom{n}{2}$ = no. ways of choosing 2 out n things

= $\frac{n!}{2!\,(n-2)!}$ ← no. ways of rearranging n things

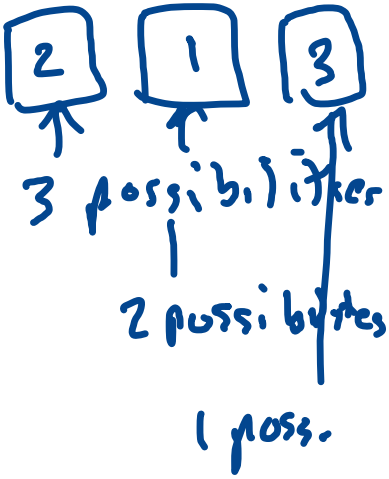
$= \binom{3}{2} = \frac{3!}{2!\,1!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1)}$

$= 3$

$\binom{n}{2} = \frac{(n)(n-1)(n-2)\cdots}{(2)(1)(n-2)\cdots}$

$= \frac{n^2 - n}{2}$

1, 2, 3
1, 3, 2
2, 1, 3
2, 3, 1
3, 1, 2
3, 2, 1



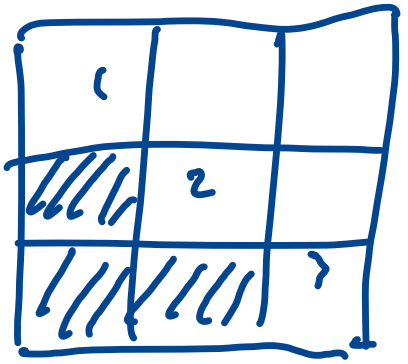
$3 \cdot 2 \cdot 1 = 3!$

$1 + 2 + \dots + n-1$

$n = 3$ $\quad 1+2$

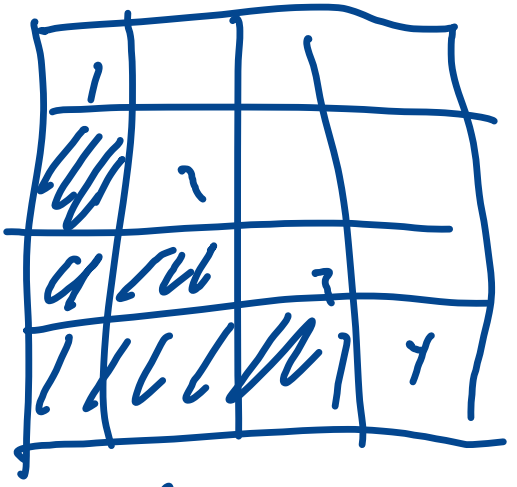
$n=3$

3



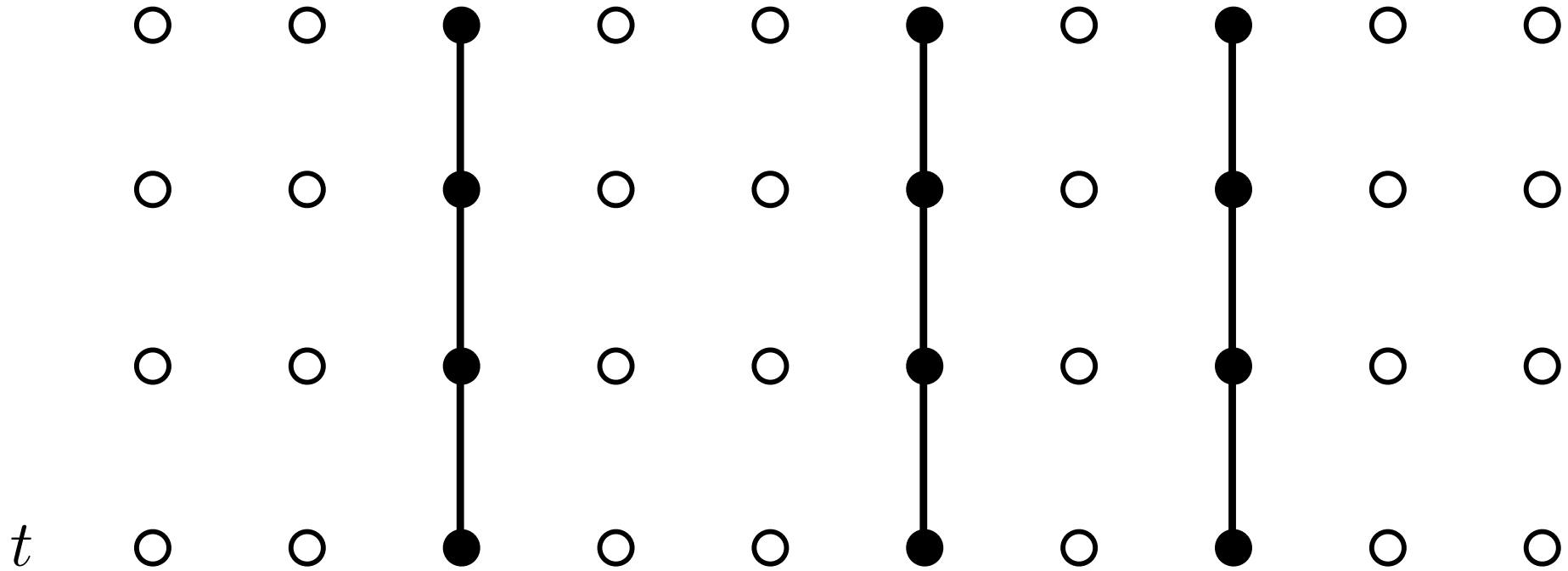
$\frac{n^2-n}{2} = 1+2$

$n=4$



$\frac{n^2-n}{2} = 1+2+3$

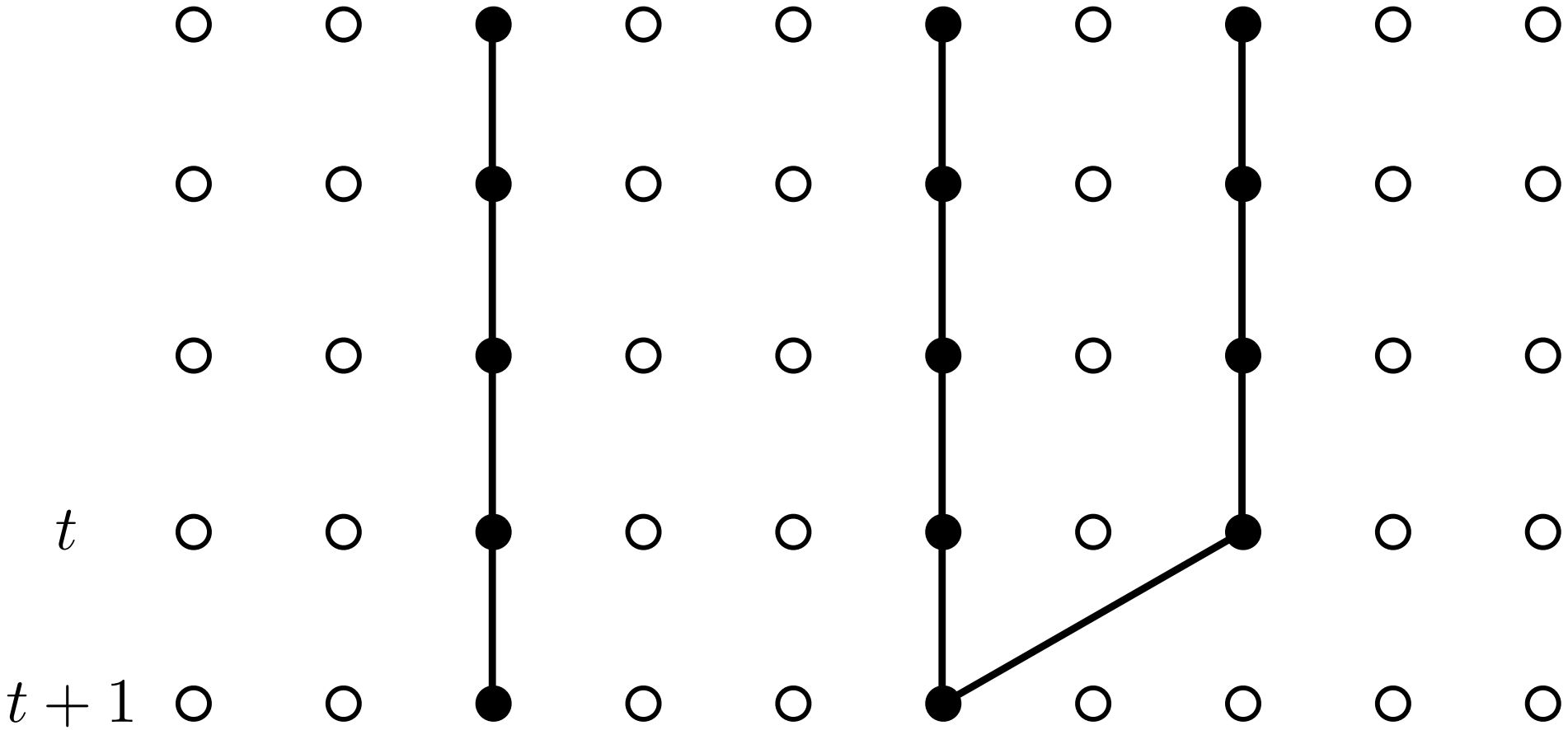
The coalescent process



$$\Pr(\text{no coalescence by gen. } t) = (1-p)^t$$

$$\text{where} \quad p = \frac{\binom{n}{2}}{N}$$

The coalescent process



$$\Pr(\text{coalesce at gen. } t+1) = (1-p)^t p \quad \text{where} \quad p = \frac{\binom{n}{2}}{N}$$

The coalescent process

discrete generations $(1-p)^t p$

geometric distribution with probability of success

$p = \binom{n}{2} / N$

If many generations are considered, can model coalescence as a continuous time process; each generation becomes a point on a continuous time axis.

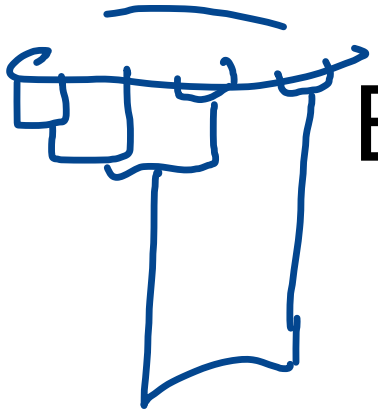
continuous time $(e^{-\lambda})^t \lambda$

exponential distribution with rate

$\lambda = \binom{n}{2} / N$

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

Expected time until coalescence: $\frac{1}{\lambda} = \frac{N}{\binom{n}{2}}$



Expected time until next coalescence

$$\frac{1}{\lambda} = \frac{N}{\binom{n}{2}}$$

longer waits in **larger populations**

shorter waits if **more lineages**

Special case: 2 lineages

$n = 2$

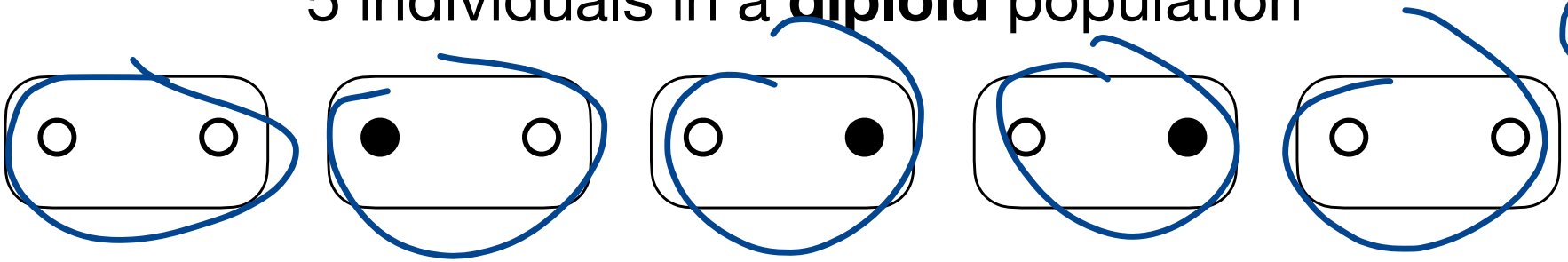
$$\frac{N}{\binom{n}{2}} = \frac{N}{\binom{2}{2}} = N$$

Expected waiting time until next coalescence = N

Diploid vs haploid

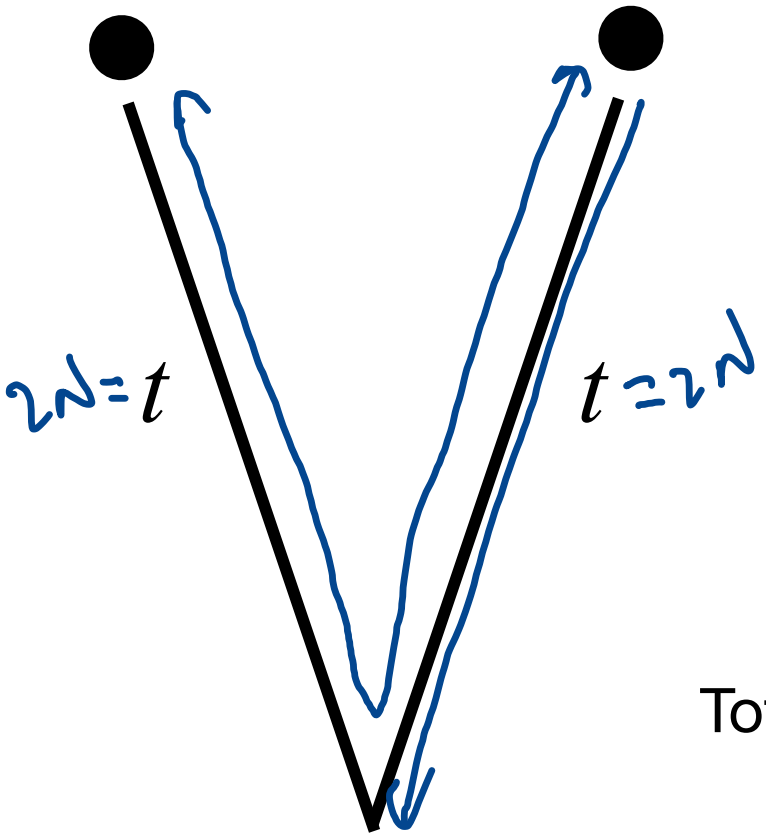
10 individuals in a **haploid** population

5 individuals in a **diploid** population



By convention, N = number of individuals (whether haploid or diploid), but it is the **number of gene copies** (10) **that matters** for coalescence.

Theta



If time to coalescence is t,
then **total path** is $2t$

Population size is N, but there are $2N$ genes if organism is **diploid**

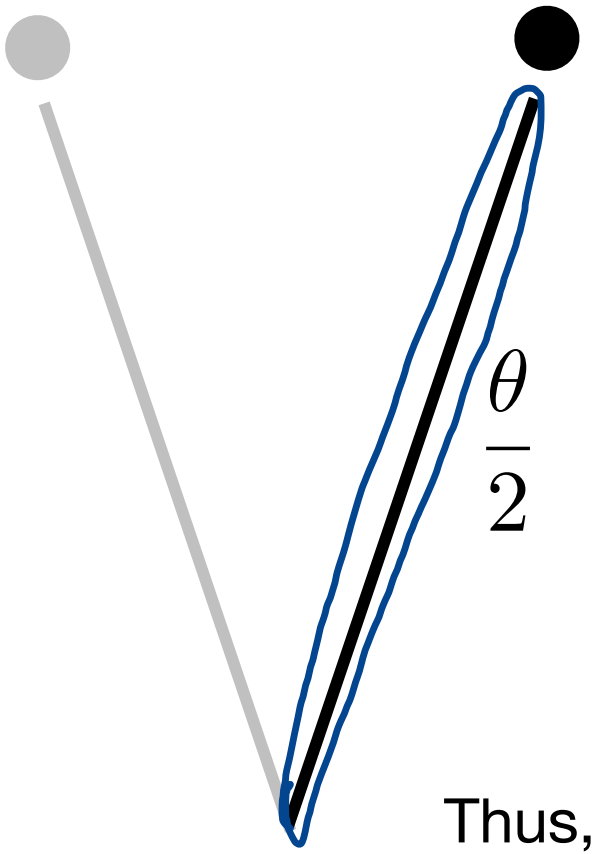
$$E[t] = 2N$$

Total time along path between two sampled genes in a diploid is thus ***4N***

If substitution rate is μ, expected number of substitutions is

$$\theta = 4N\mu$$

time × mutats/time = total mutations



Theta

$\theta = 4 N_e \mu$

Expected number of substitutions for one edge in a gene tree

Thus, estimated theta is twice the edge length (expected number of substitutions) as estimated on a gene tree

STOPPED HERE 2024-03-07

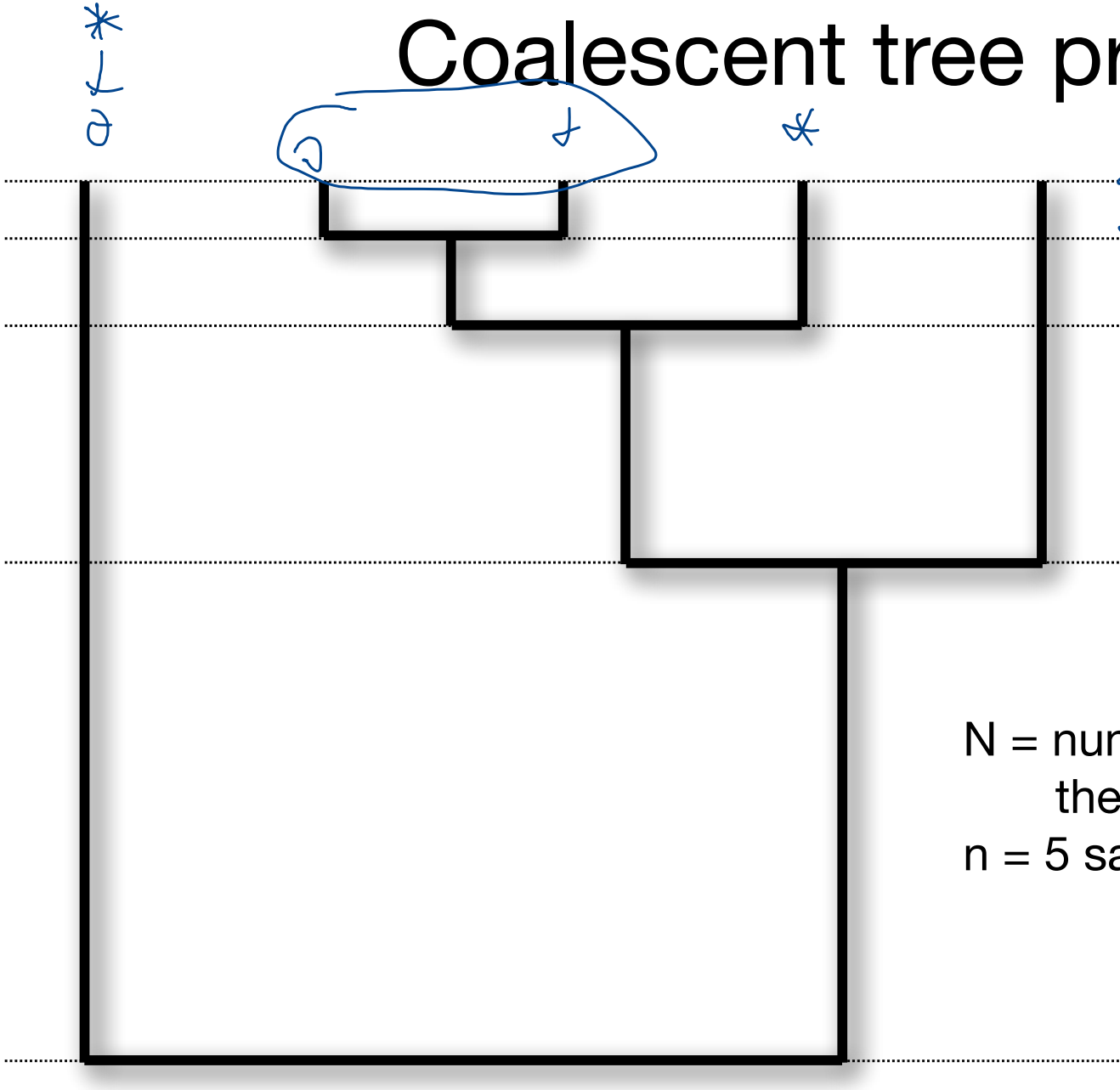
Effective population size

The **effective population size N_e** is the size of a **randomly mating population** that would behave the same way as the population under study (with census size N)

- Obligate outcrossing: $N_e > N$
- Fluctuation in population size: $N_e <$ average N
- Biased sex ratios: $N_e < N$
- Inbreeding: $N_e < N$

Bottom line: we are always estimating N_e rather than N

Coalescent tree prior



$t_0 = 0$

$t_1 \quad E[t_1 - t_0] = N/10$

$t_2 \quad E[t_2 - t_1] = N/6$

$t_3 \quad E[t_3 - t_2] = N/3$

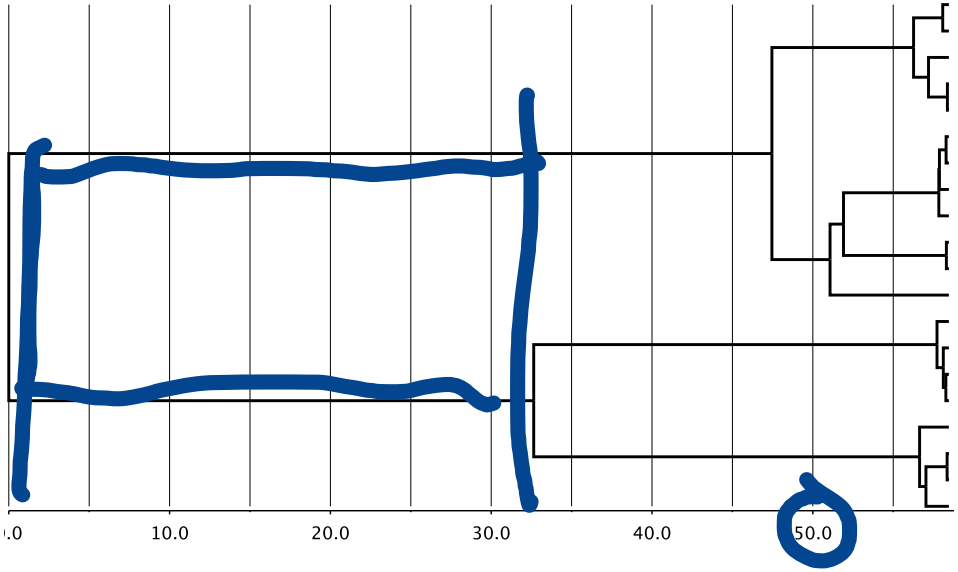
$t_4 \quad E[t_4 - t_3] = N$

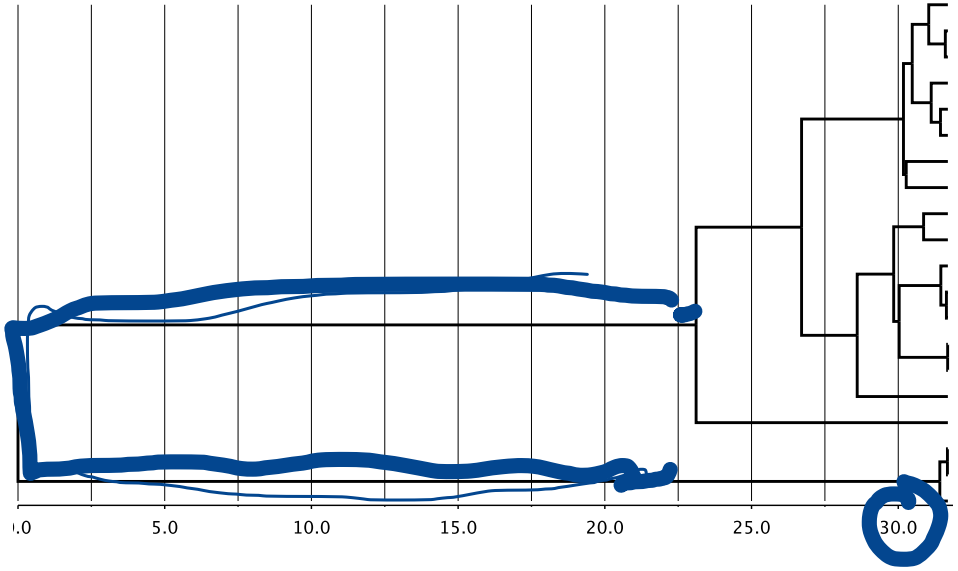
N = number of gene copies in the entire population

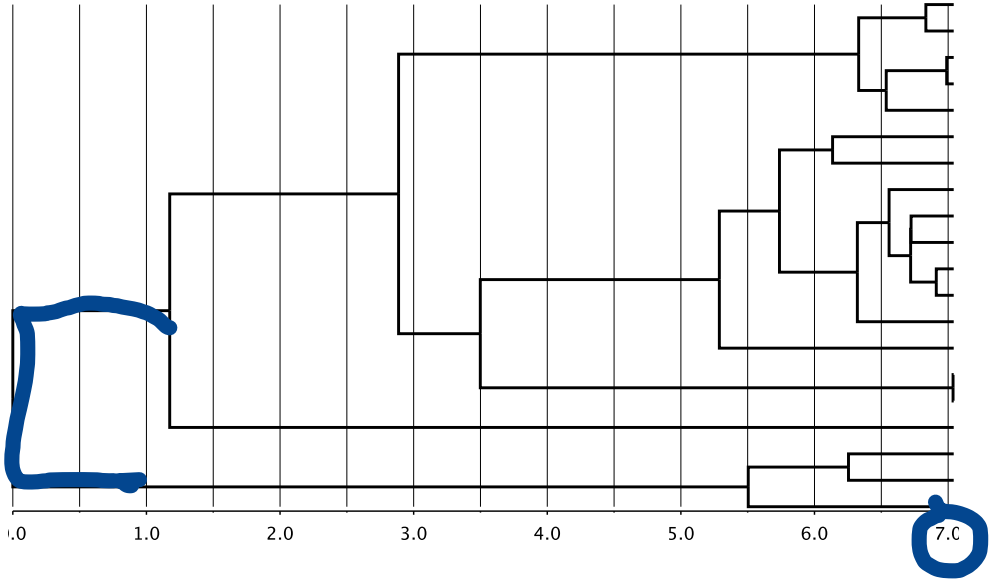
n = 5 sampled gene copies

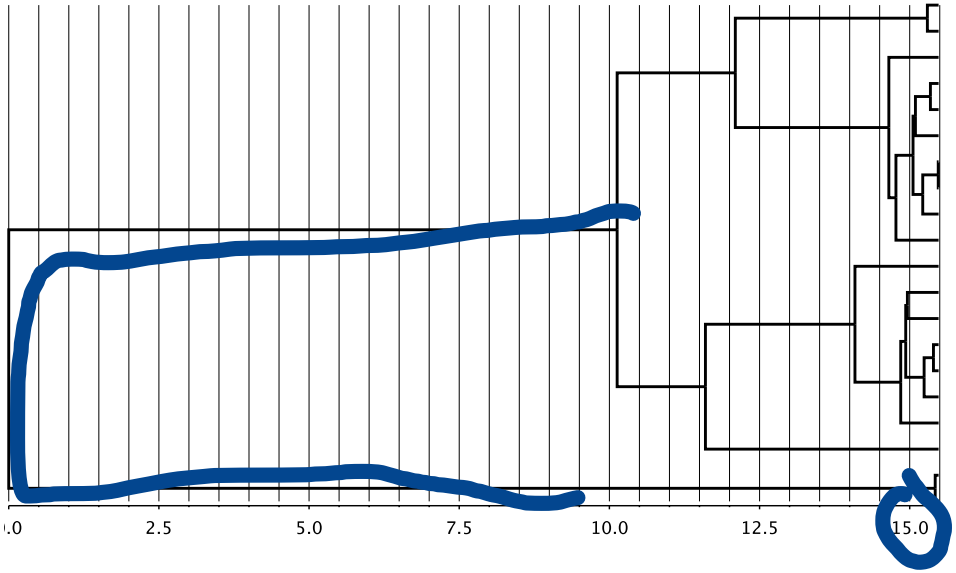
coalescence times shorter near the present

Examples of coalescent trees







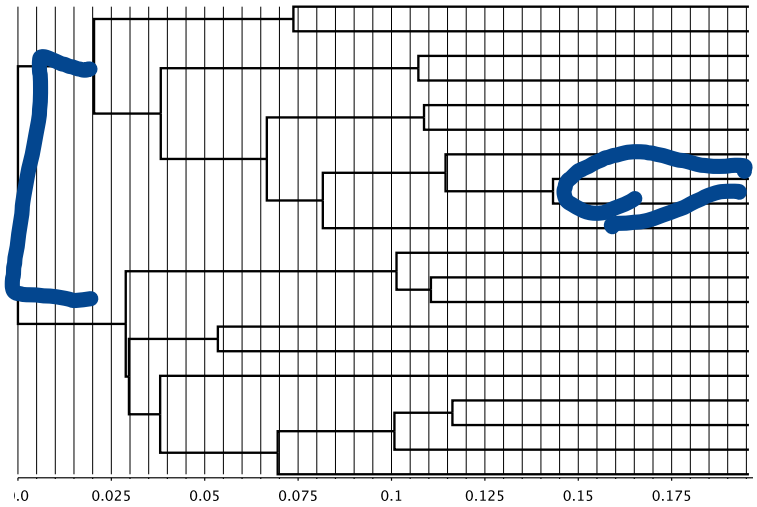


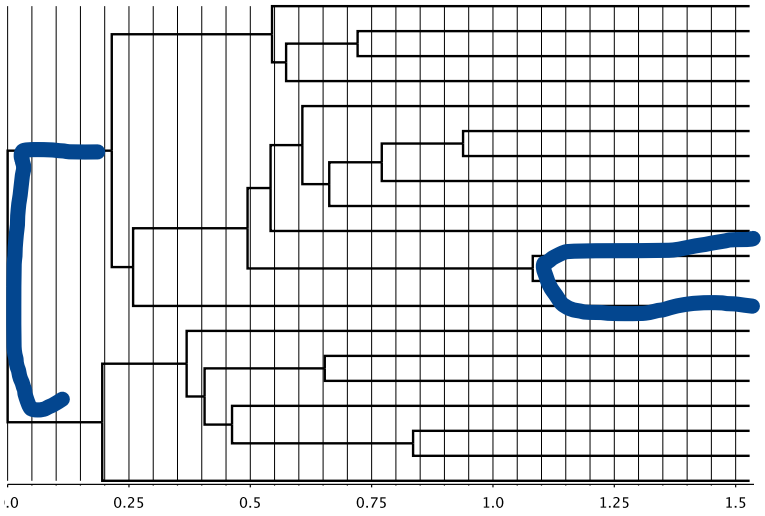
depth of trees simulated with same N quite variable

skyline analyses

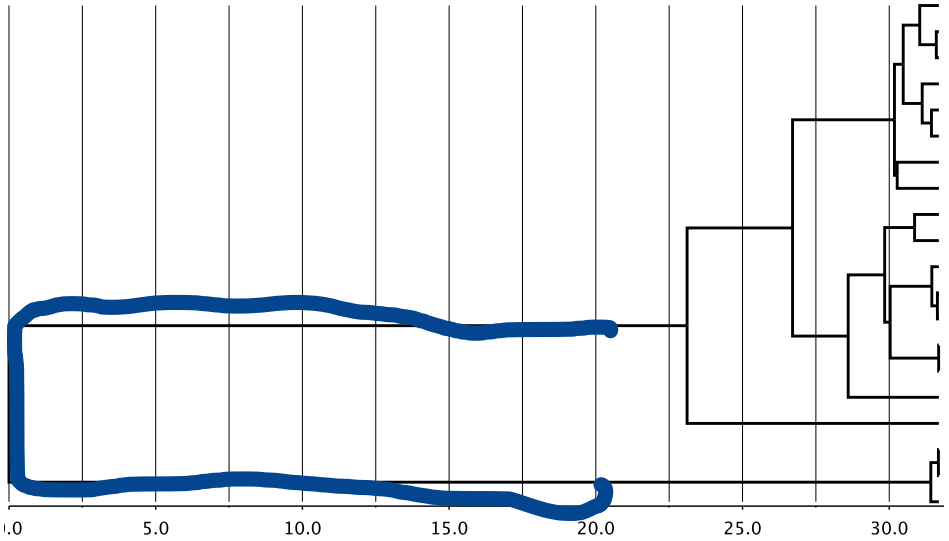
Exponential growth reduces compression near present

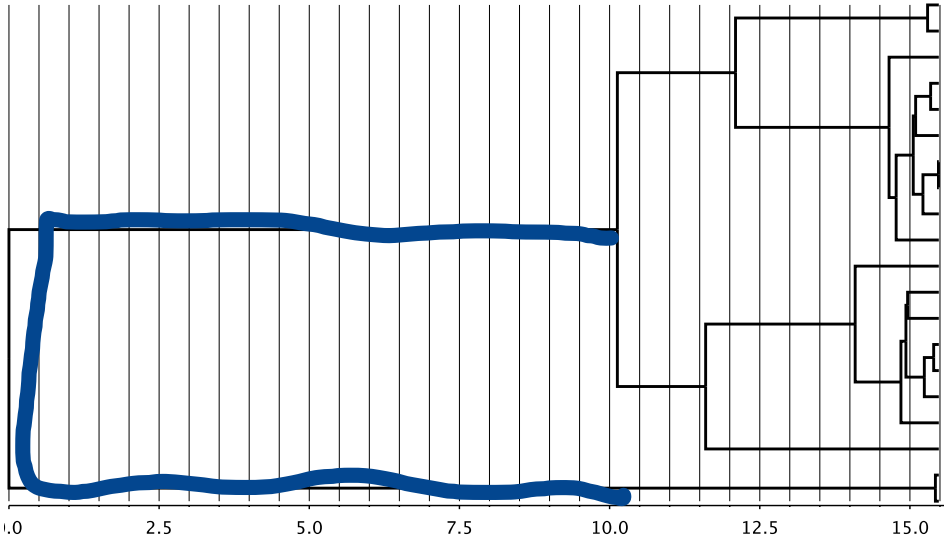
Exponential growth





Constant population size





exponential growth makes edge lengths more even