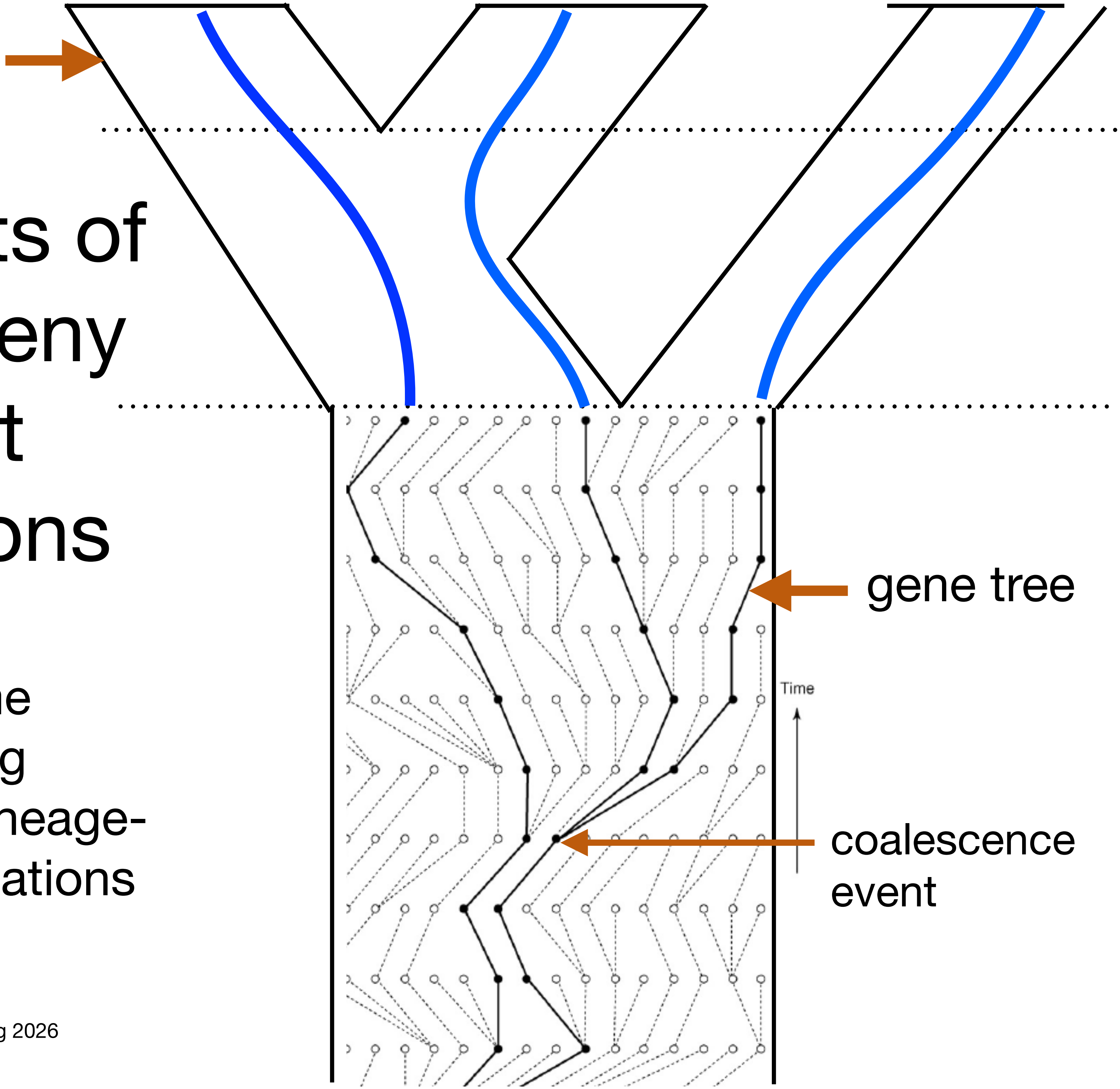
Coalescence

Segments of a phylogeny represent populations

We will assume random mating within these lineage-specific populations



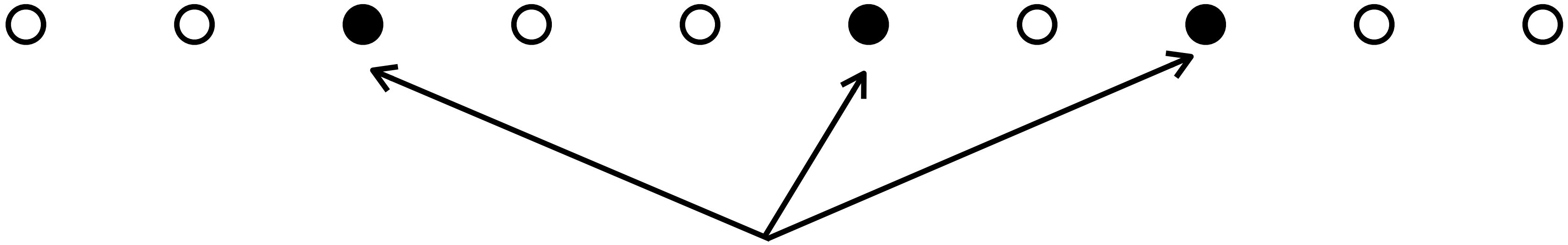
species tree

gene tree

coalescence event

The coalescent process

N=10 haploid individuals in a population today

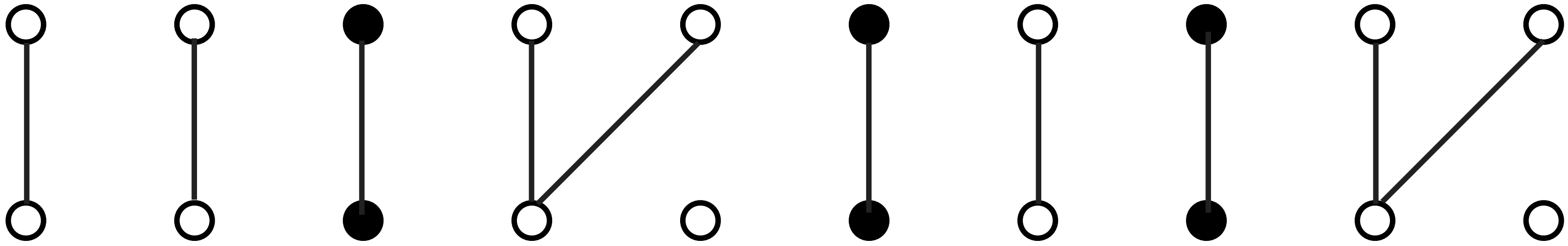


n=3 sampled

Kingman 1982

The coalescent process

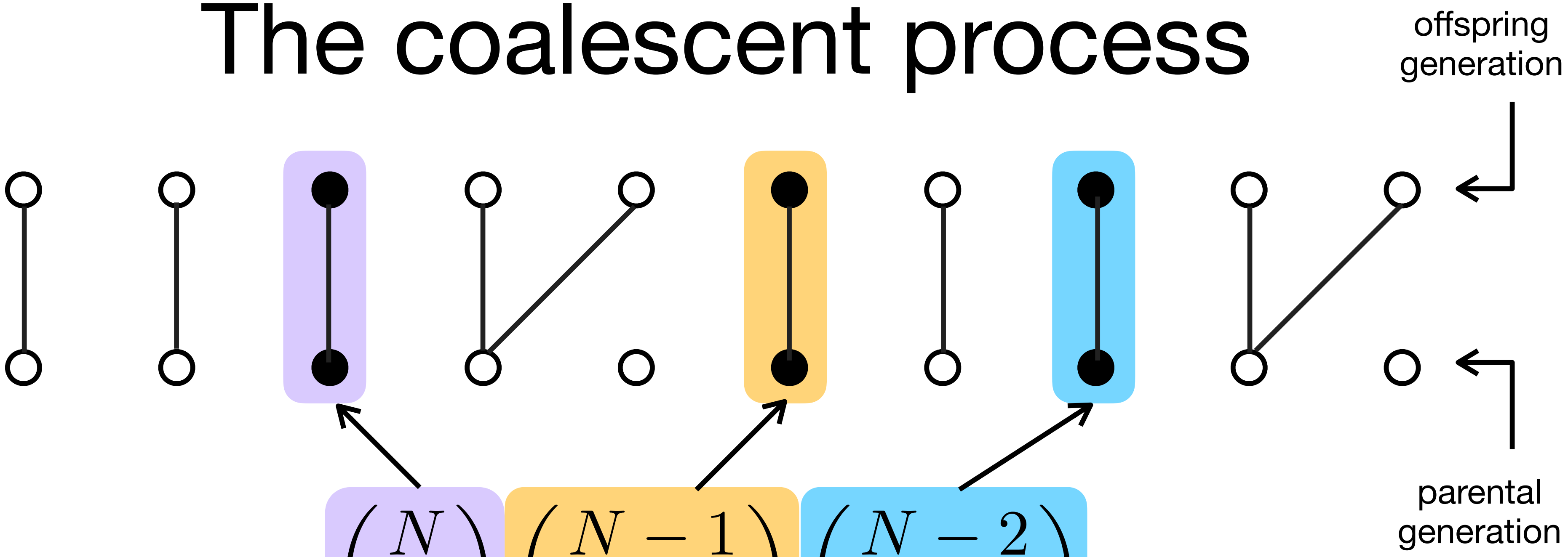
N=10 haploid individuals in a population today



N=10 haploid individuals in previous generation

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

The coalescent process



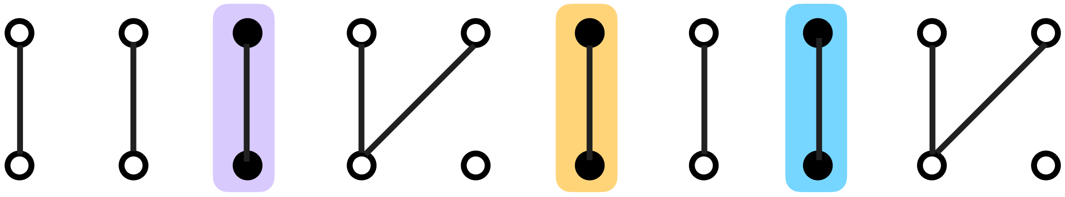
$$\left(\frac{N}{N}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)$$

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

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

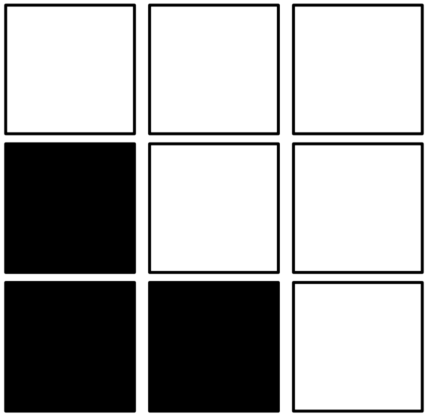
Can ignore terms like this if N is large

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

sum of natural numbers up to n-1

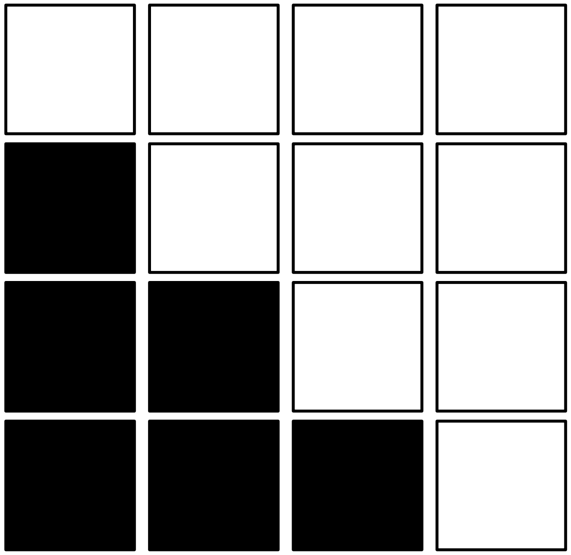
Probability of no coalescence in 1 generation given:

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



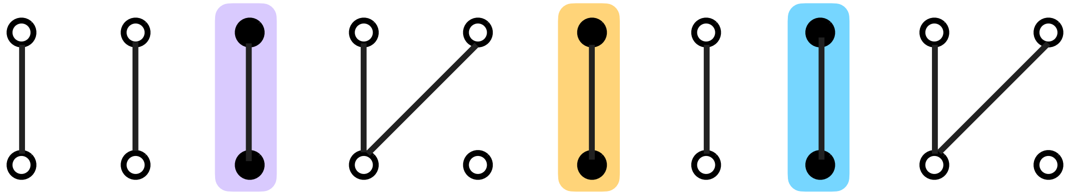
$$1 + 2 = \frac{3^2 - 3}{2}$$

$$= \binom{3}{2}$$



$$1 + 2 + 3 = \frac{4^2 - 4}{2}$$

$$= \binom{4}{2}$$



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

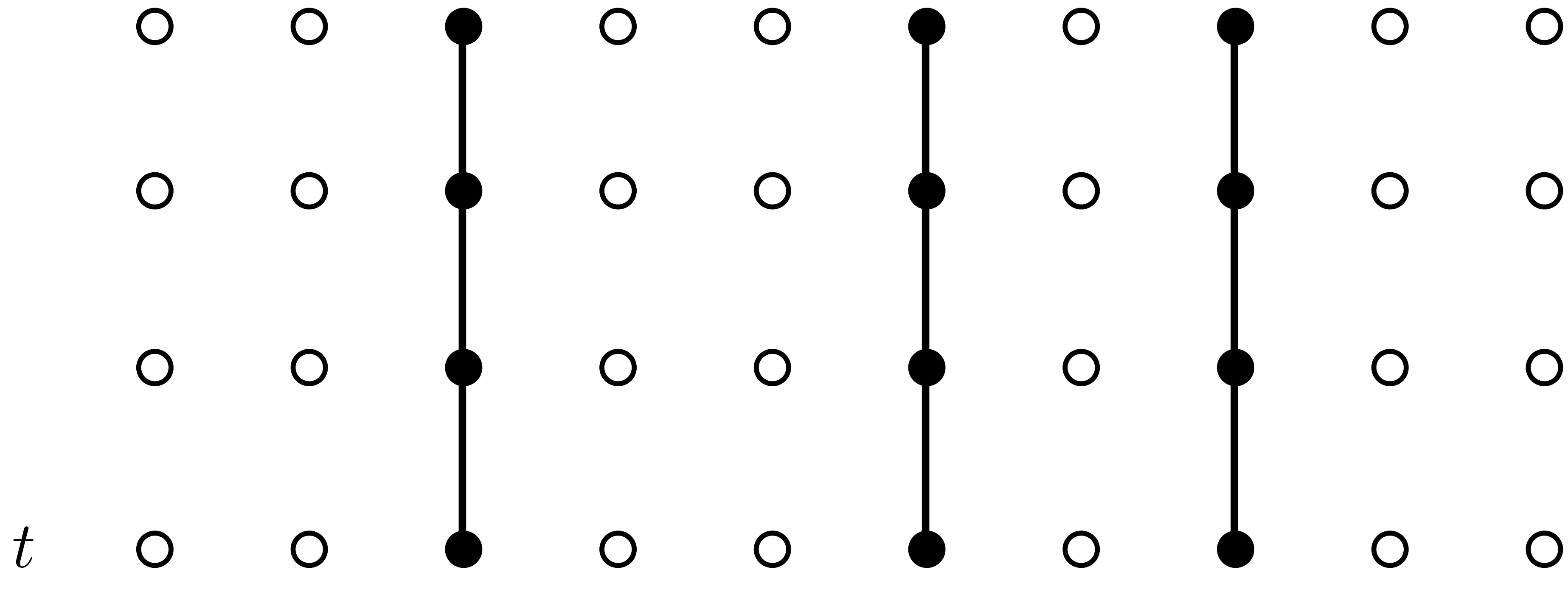
$$\approx 1 - \frac{\binom{n}{2}}{N}$$

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)

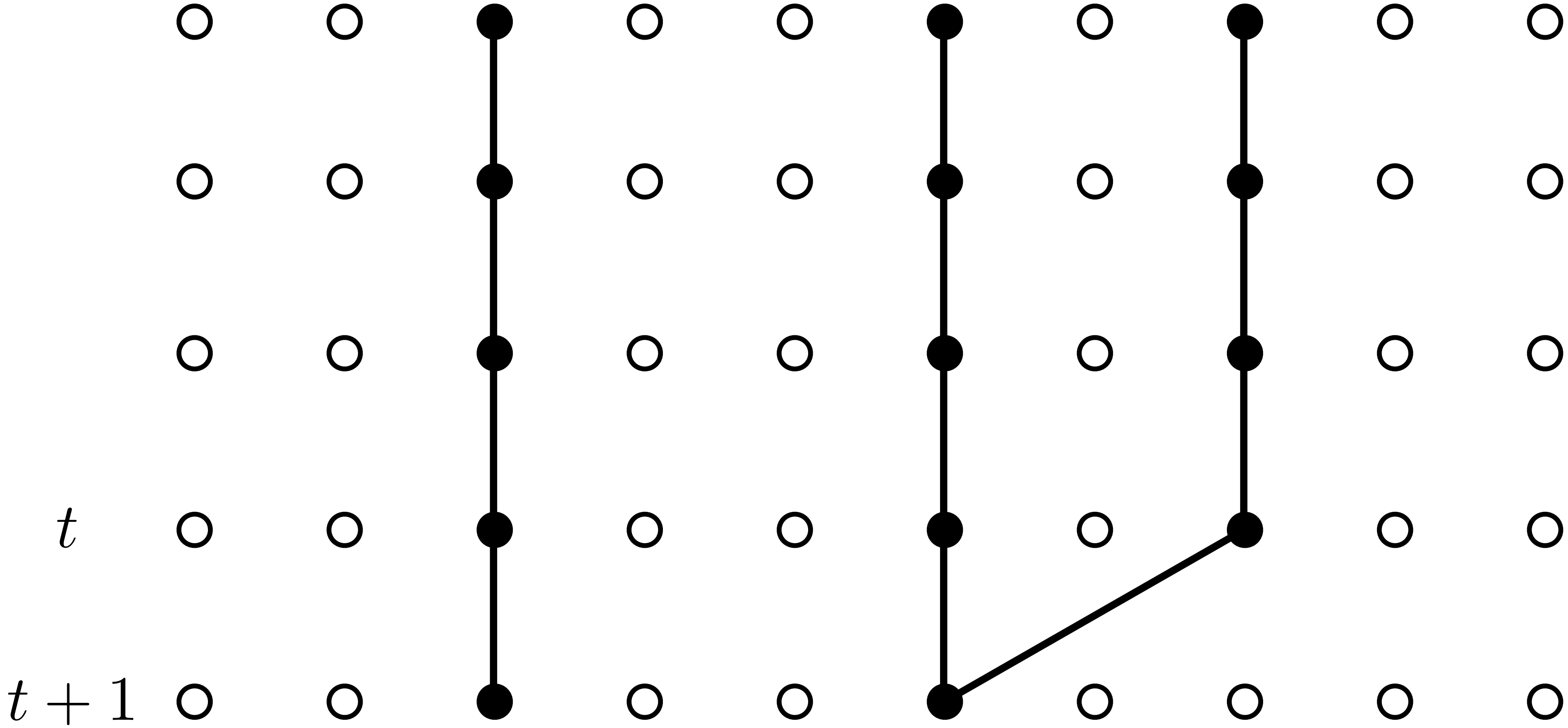
The coalescent process



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

where $p = \dfrac{\binom{n}{2}}{N}$

The coalescent process



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

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 $\left(e^{-\lambda}\right)^t \lambda$

exponential distribution with rate

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

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

Expected time until next coalescence

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

longer waits in **larger populations**

shorter waits if **more lineages**

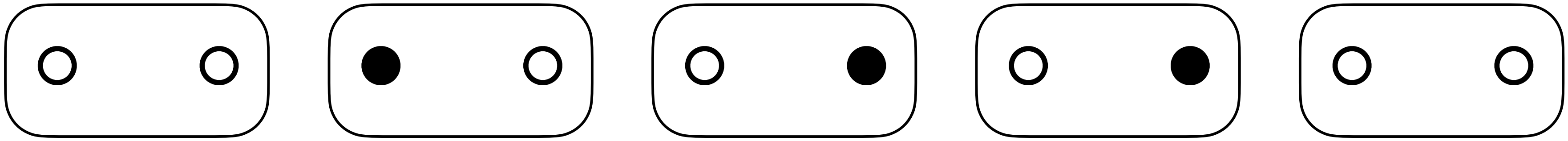
Special case: 2 lineages

Expected waiting time until next coalescence = *N*

Diploid vs haploid

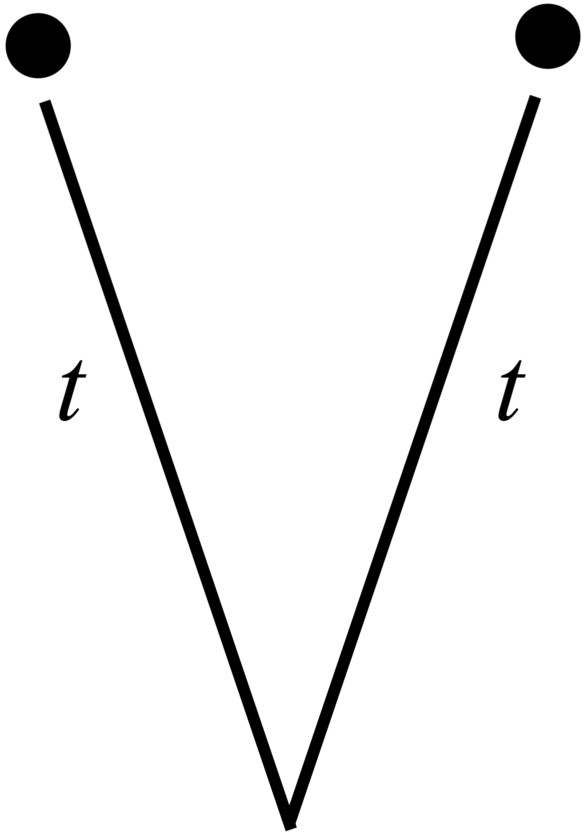
10 individuals in a **haploid** population

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 2*t*

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

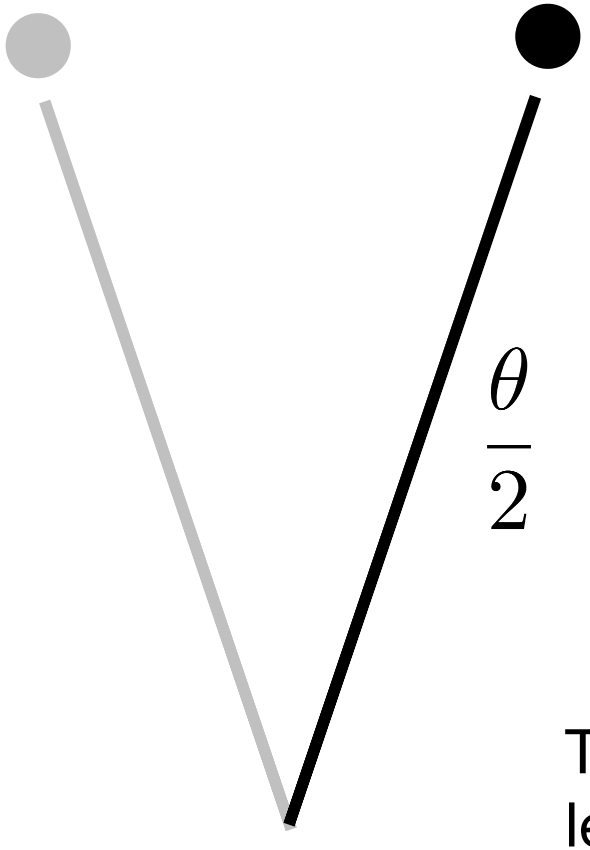
$$E[t] = 2N$$

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

If the mutation rate is *μ*, expected number of mutations is

$$\theta = 4N\mu$$

Theta



Expected number of mutations for one edge in a gene tree

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

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)

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

Bottom line: we are always estimating N_e rather than N