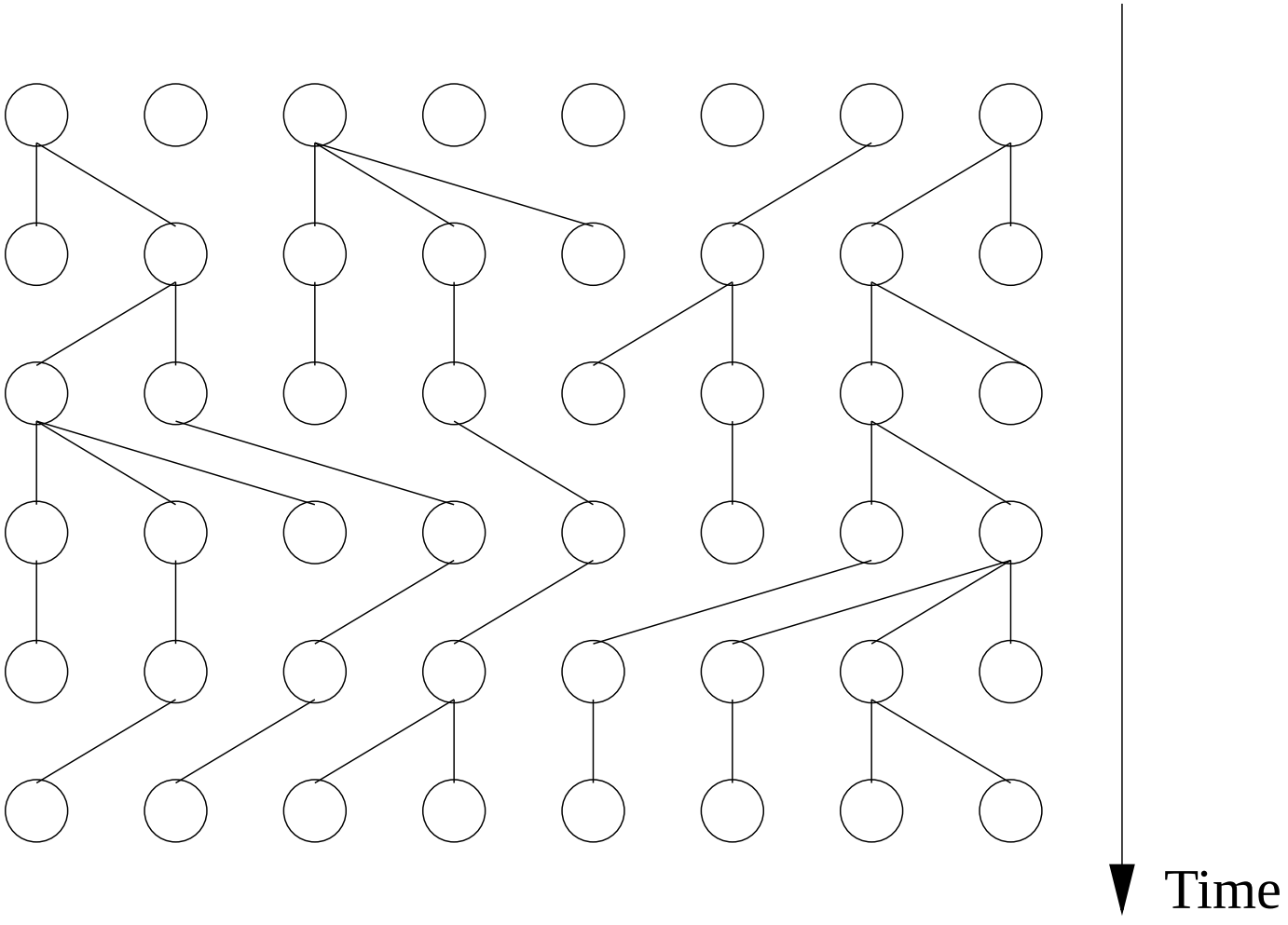
# Statistical population genetics

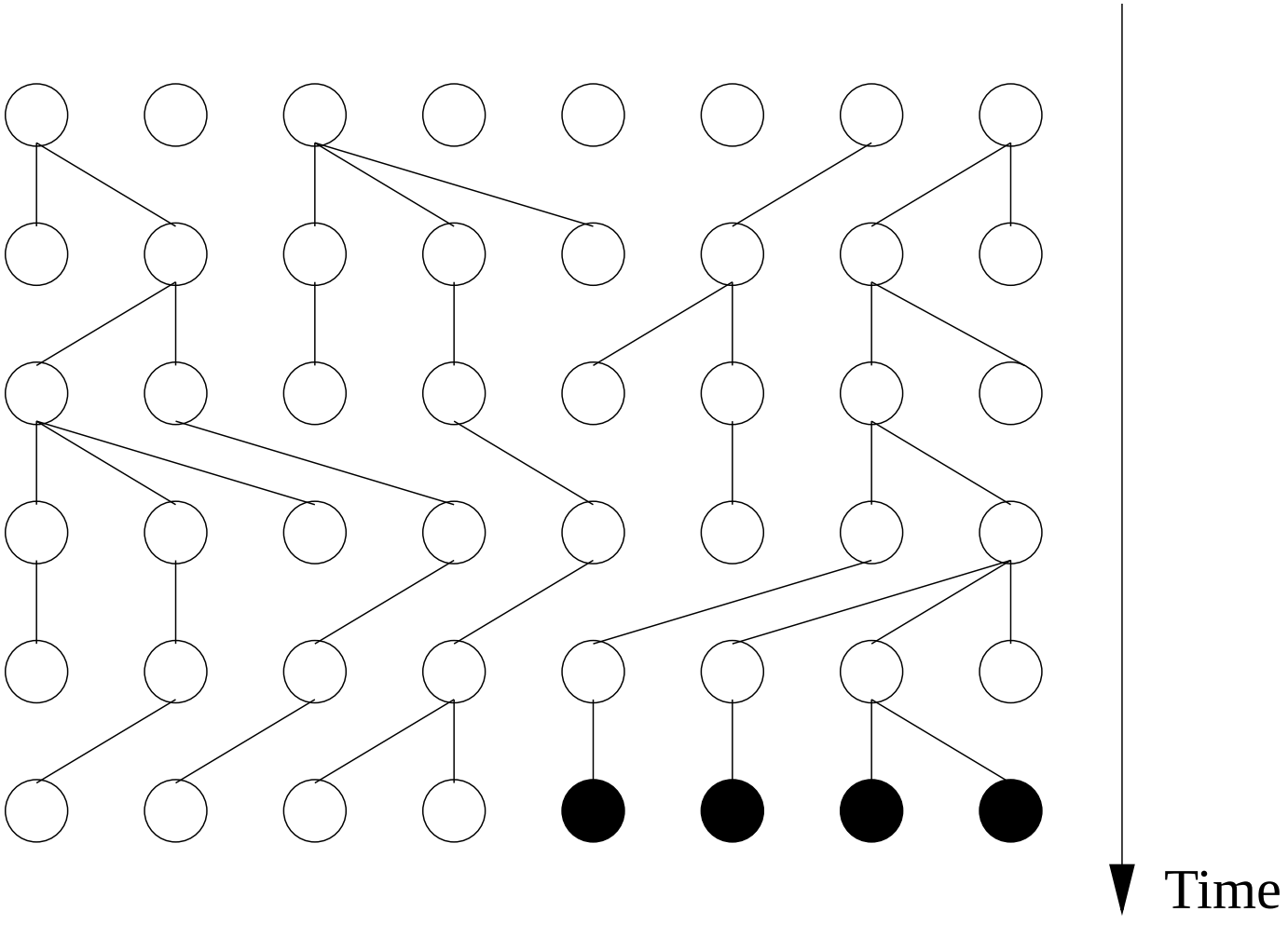
## *Lecture 4: Derivation of the coalescent*

Xavier Didelot
Dept of Statistics, Univ of Oxford
didelot@stats.ox.ac.uk

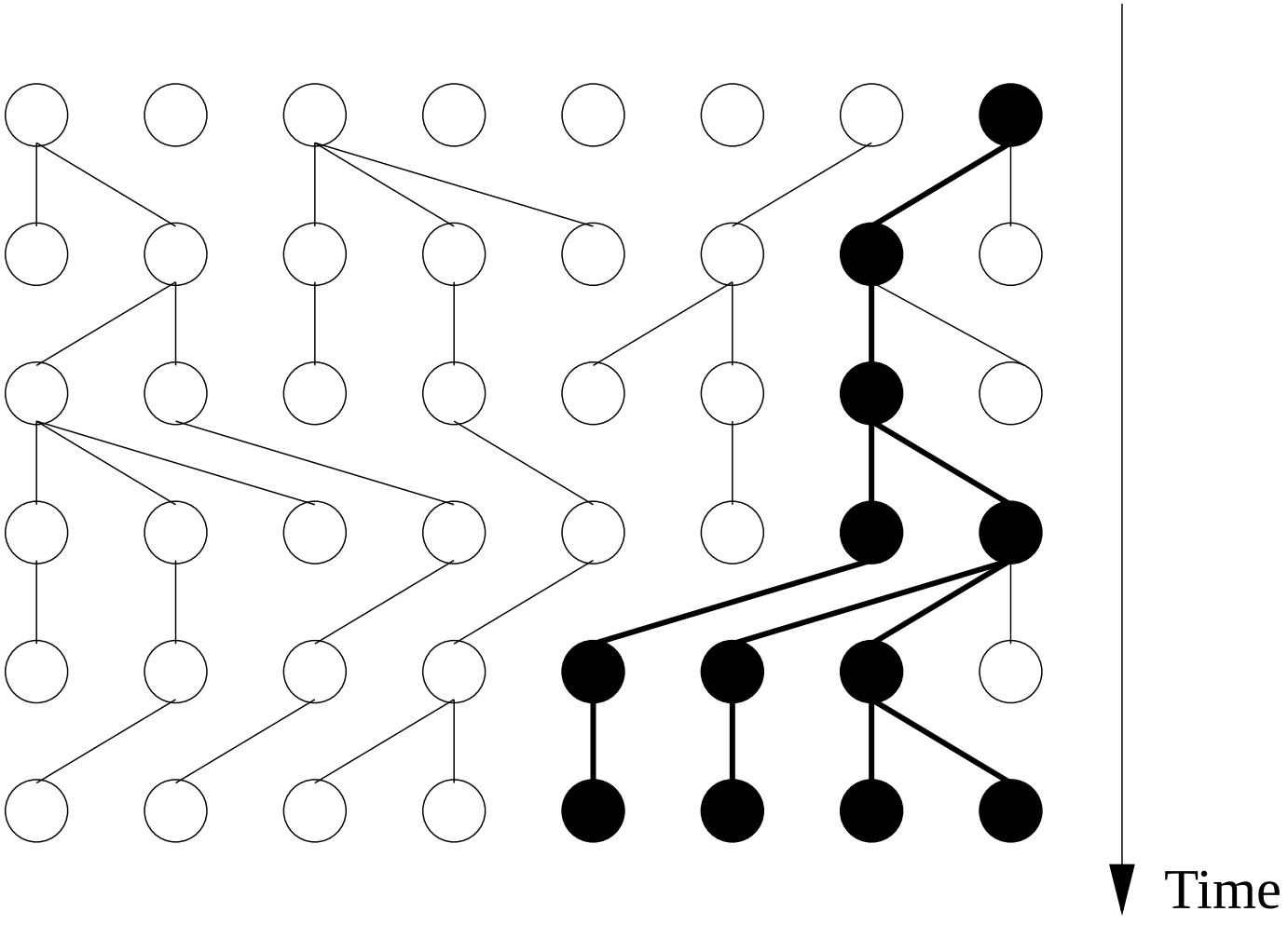
# Gene genealogies



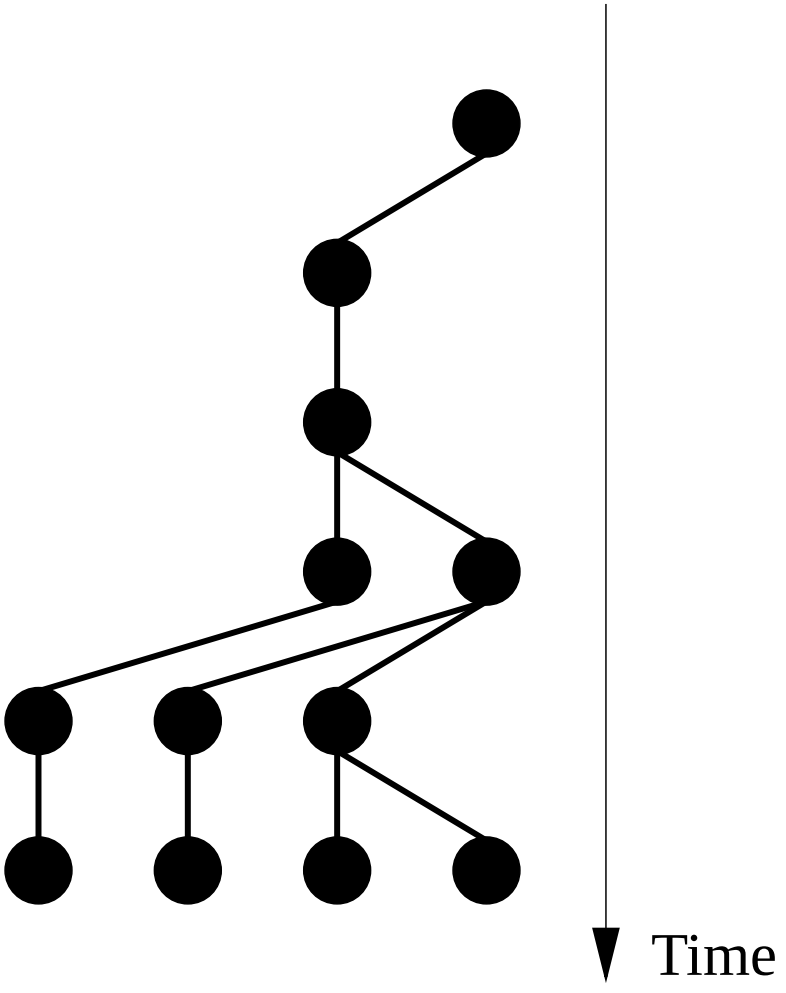
# Gene genealogies



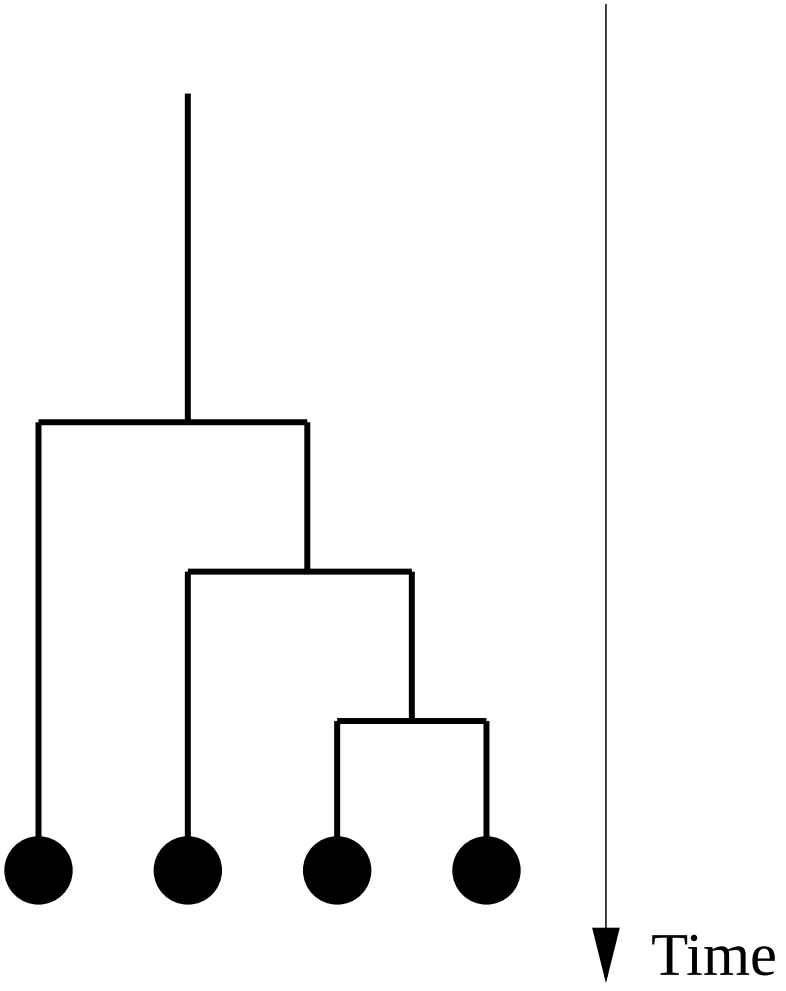
# Gene genealogies



# Gene genealogies



# Gene genealogies



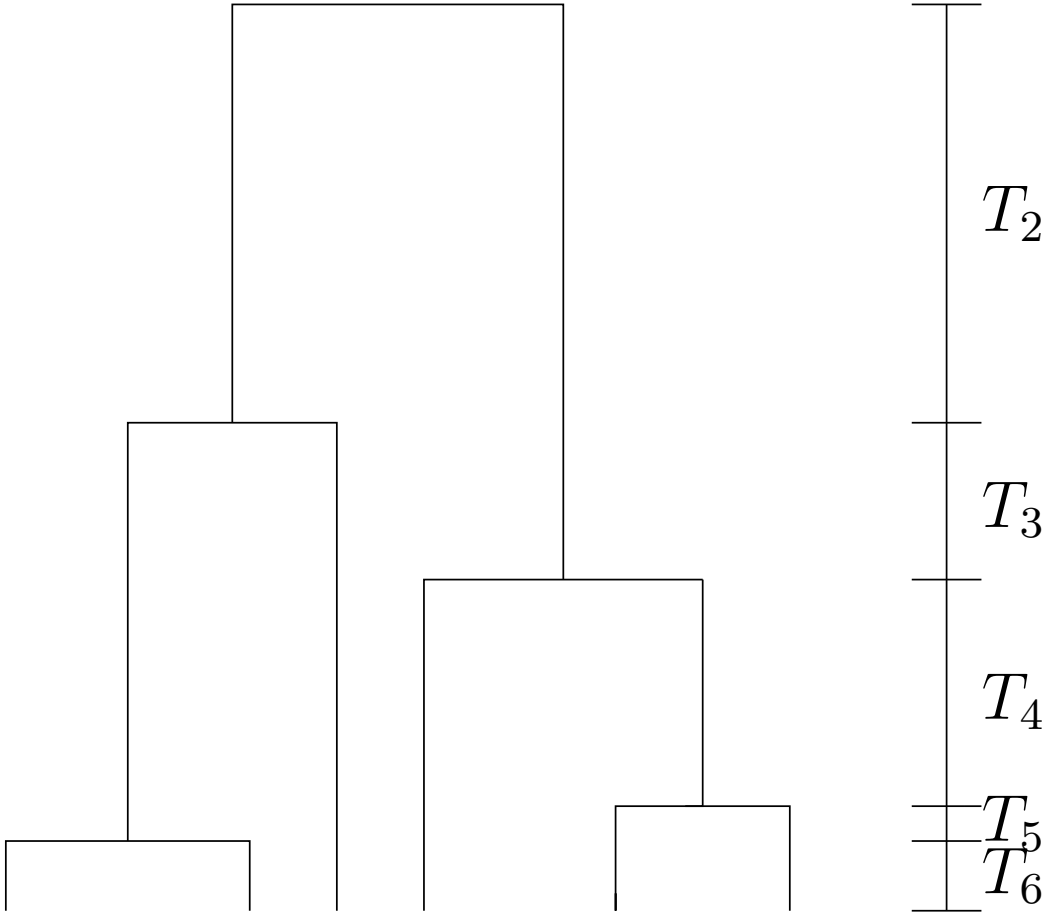
# Gene genealogies

- So far, we have been looking at the evolution of a population **forward in time**. Given the current state of the population, we have been wondering what will happen in the future.
- We now do the opposite, ie. look **backward in time**.
- We focus on **a sample** of the whole current population.
- If we look at the ancestry of the members of the sample, we get a tree which is called **the genealogy** of the sample.
- A genealogy has two components: a **branching order** (sometimes called structure or topology) and the **ages of the internal nodes**
- In the previous slide, if the individuals are labeled 1 to 4 from left to right, the branching order could be noted (1,(2,(3,4))) and the ages of internal nodes (1,2,3).

# The coalescent

- The **coalescent** (Kingman, 1982) is a model describing the **genealogy** of a sample of $n$ individuals.
- In the coalescent, branching order and internal ages are **independent**.
- In the coalescent, the branching order is always **binary**.
- Any two branches are equally likely to find a common ancestor so that **all branching orders are equally likely**.
- The internal ages are defined by the times $T_n$ during which there are $n$ ancestors.

# Coalescent tree



# Discrete-time coalescent

**Definition** (Discrete-time coalescent)**.**

*In the discrete version of the coalescent model,* $T_n$ *is Geometric with parameter* $\frac{n(n-1)}{2M}$*:*

$$f_{T_n}(t) = \left(1 - \frac{n(n-1)}{2M}\right)^{t-1} \left(\frac{n(n-1)}{2M}\right)$$

# Coalescent from the WF

**Theorem** (Genealogy in a WF population)**.**

*The genealogy of a sample in a WF population is approximately given by the coalescent model when* $M$ *is large.*

# Coalescent from the WF

**Proof.** Consider a sample of $n$ genes in a population evolving under the WF model with constant population size $M$. Let $P_{n,k}$ denote the probability that these $n$ genes have $k$ ancestors in the previous generation. Then:

$$P_{n,n} = \frac{M-1}{M}\frac{M-2}{M}\ldots\frac{M-n+1}{M} = \prod_{i=1}^{n-1}\left(1-\frac{i}{M}\right)$$

$$= 1 - \sum_{i=1}^{n-1}\frac{i}{M} + \mathcal{O}(M^{-2}) = 1 - \frac{n(n-1)}{2M} + \mathcal{O}(M^{-2})$$

and:

$$P_{n,n-1} = \frac{n(n-1)}{2}\frac{1}{M}\frac{M-1}{M}\frac{M-2}{M}\ldots\frac{M-n+2}{M}$$

$$= \frac{n(n-1)}{2M}\prod_{i=1}^{n-2}\left(1-\frac{i}{M}\right) = \frac{n(n-1)}{2M} + \mathcal{O}(M^{-2})$$

# Coalescent from the WF

Therefore we have:

$$P_{n,k} = \mathcal{O}(M^{-2}) \text{ for all } k \in [1..n-2]$$

- When the population is large, we can ignore the probability of having more than one coalescent event per generation or having non-binary events.
- At each generation, a population of $n$ genes has probability $\frac{n(n-1)}{2M}$ to have a coalescent event.
- The waiting time until the next coalescent event is therefore approximately Geometric with parameter $\frac{n(n-1)}{2M}$.
- It is furthermore obvious that any two genes are equally likely to coalesce in the Wright-Fisher model, as it is in the coalescent model. □

# Coalescent from Moran model

**Theorem** (Genealogy in a Moran population)**.**

*The genealogy of a sample in a Moran population is exactly given by the coalescent model (with time rescaled by a factor* $2/M$*)*

# Coalescent from Moran model

**Proof.** Because of the definition of the Moran model, it is clear that only 0 or 1 coalescent event can happen per time step and that non-binary events can not occur, so that $P_{n,k} = 0$ for $k < n - 1$, just as in the Wright-Fisher with infinite population size. Let us calculate $P_{n,n-1}$:

$$P_{n,n-1} = \frac{M-1}{M}\frac{n}{M}\frac{n-1}{M-1} = \frac{n(n-1)}{M^2}$$

and therefore:

$$P_{n,n} = 1 - P_{n,n-1} = 1 - \frac{n(n-1)}{M^2}$$

# Coalescent from Moran model

- At each Moran time step, the sample has probability $\frac{n(n-1)}{M^2}$ to coalesce
- The time until the first coalescent event is Geometrically distributed with parameter $\frac{n(n-1)}{M^2}$.
- This is the same as in the discrete coalescent model, if we rescale time by a factor $2/M$.
- It is furthermore obvious that any two genes are equally likely to coalesce in the Moran model as it is in the coalescent. □

# Continuous-time coalescent

- In the discrete-time version of the coalescent, $T_n$ is geometric with parameter $\frac{n(n-1)}{2M}$
- If the time is now measured in units of $M$ generations (also called **coalescent units of time**), we have:

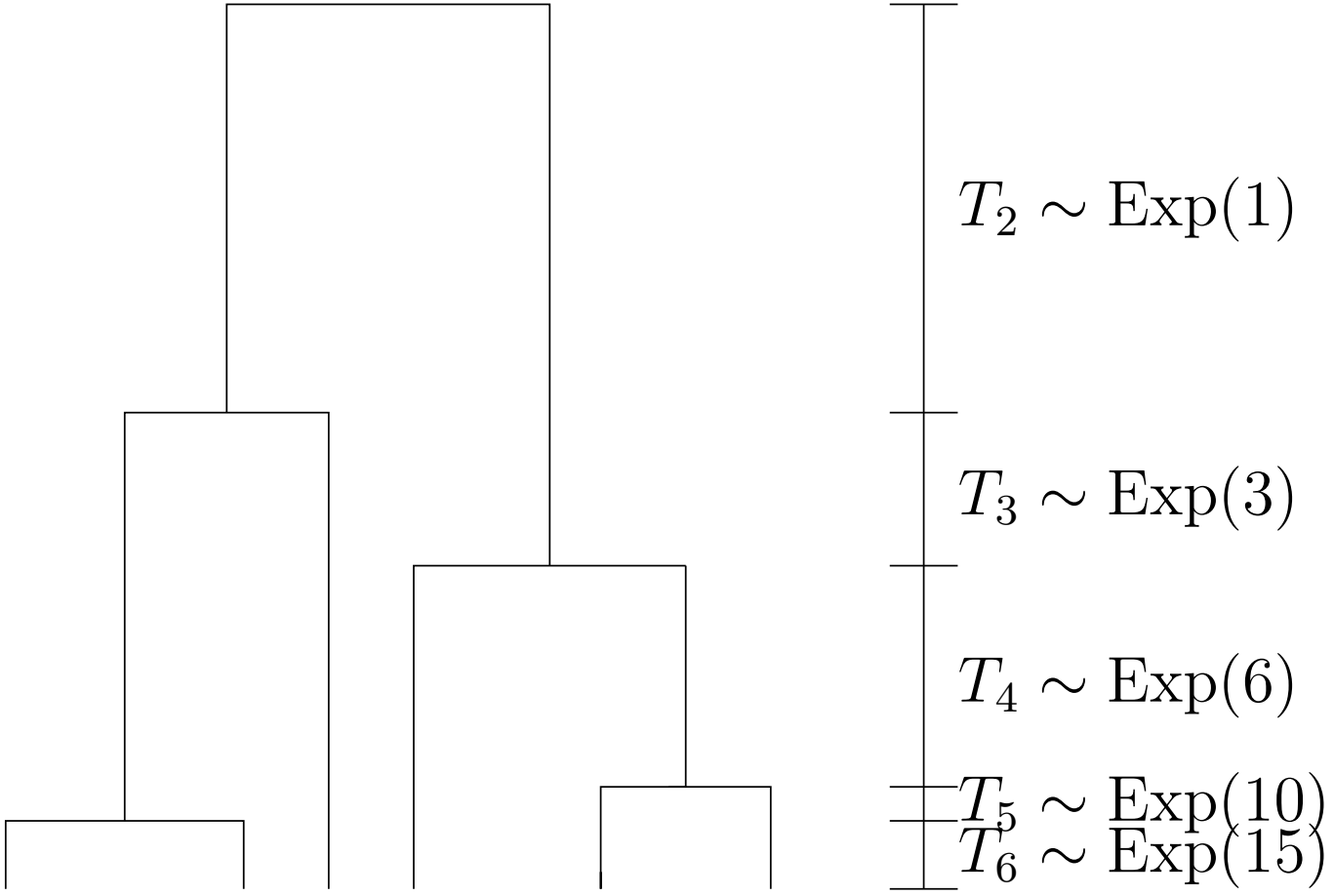
$$\mathbb{P}(T_n > t) = \left(1 - \frac{n(n-1)}{2M}\right)^{Mt} \xrightarrow[M\to\infty]{} \exp\left(\frac{-tn(n-1)}{2}\right)$$

- We deduce the continuous version of the coalescent:

  **Definition** (Continuous-time coalescent)**.**
  *In the continuous version of the coalescent, $T_n$ is Exponentially distributed with parameter $\frac{n(n-1)}{2}$:*

$$f_{T_n}(t) = \frac{n(n-1)}{2} \exp\left(-\frac{tn(n-1)}{2}\right)$$

# Continuous-time coalescent



# Summary

- From now on we will only refer to the **continuous-time coalescent**.
- The coalescent describes **genealogies** in the Moran model and in the WF model for large populations.
- The WF and Moran models scale against the coalescent in **the same way** that they did against each other based on their heterozygosity decay and their fixation time.
- The coalescent is by far the **most important** object in modern population genetics.