The Coalescent

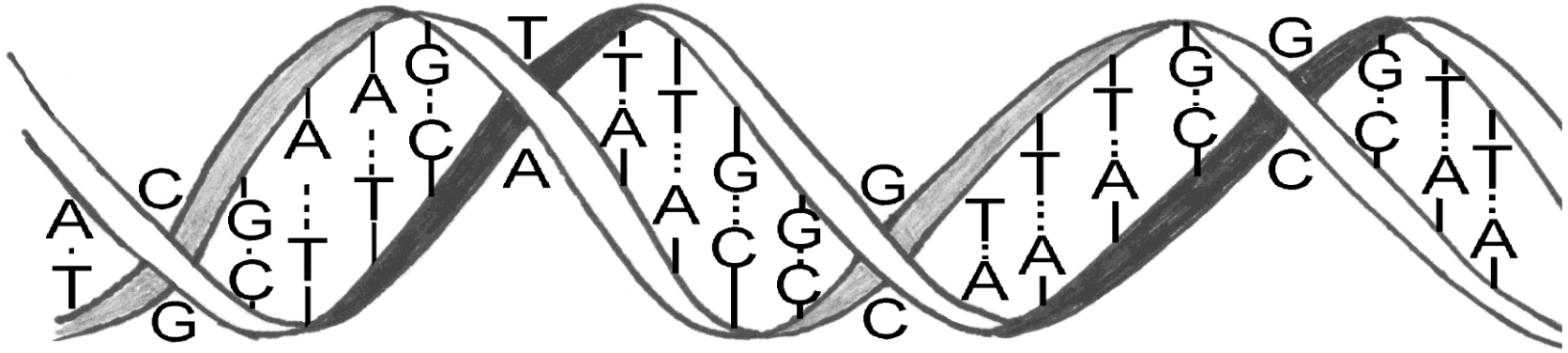
Evolution backward in time

Joachim Hermisson

Mathematics and Biosciences Group

Mathematics & MFPL, University of Vienna

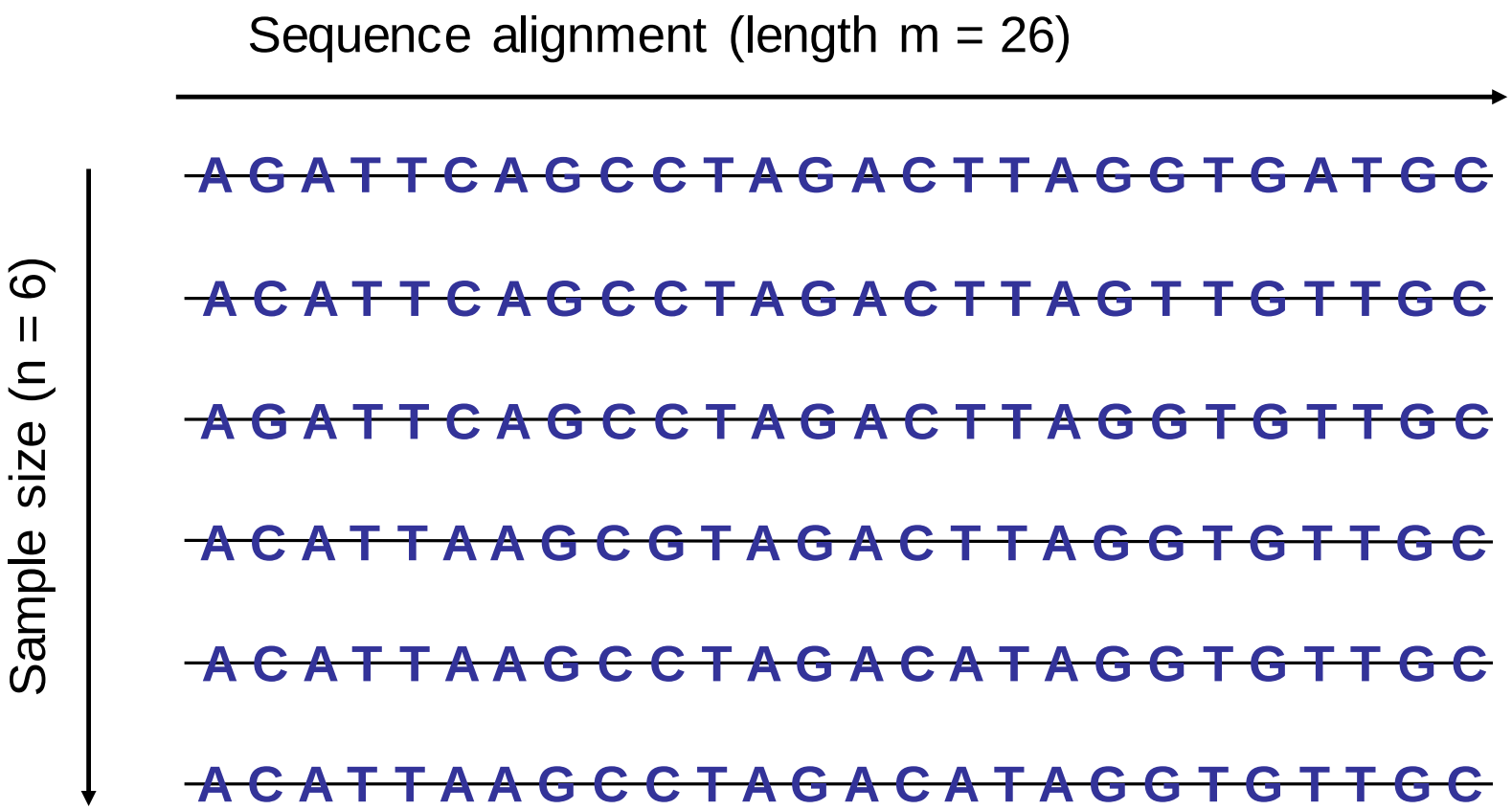
Introduction to the Coalescent
data, data, data, …



Massive accumulation of DNA sequence data

- *1980's:* 3-4 years PhD projects to sequence a single gene (some 1000 base pairs)
- *1990 – 2003:* *Human Genome Project* (~ 3 10^9 (3 billion) bases) expected: 3 billion $, final: ~ 300 Mio $
- *since 2010:* *1000 Genome Project* 4000 $ – 10000 $ per genome, soon < 1000 $
- *today*: extended to 2500 (25 x 100), completed May 2013 1000 genomes also for *Drosophila*, *Arabidopsis* …

Patterns of Evolution
”Summary Statistics”



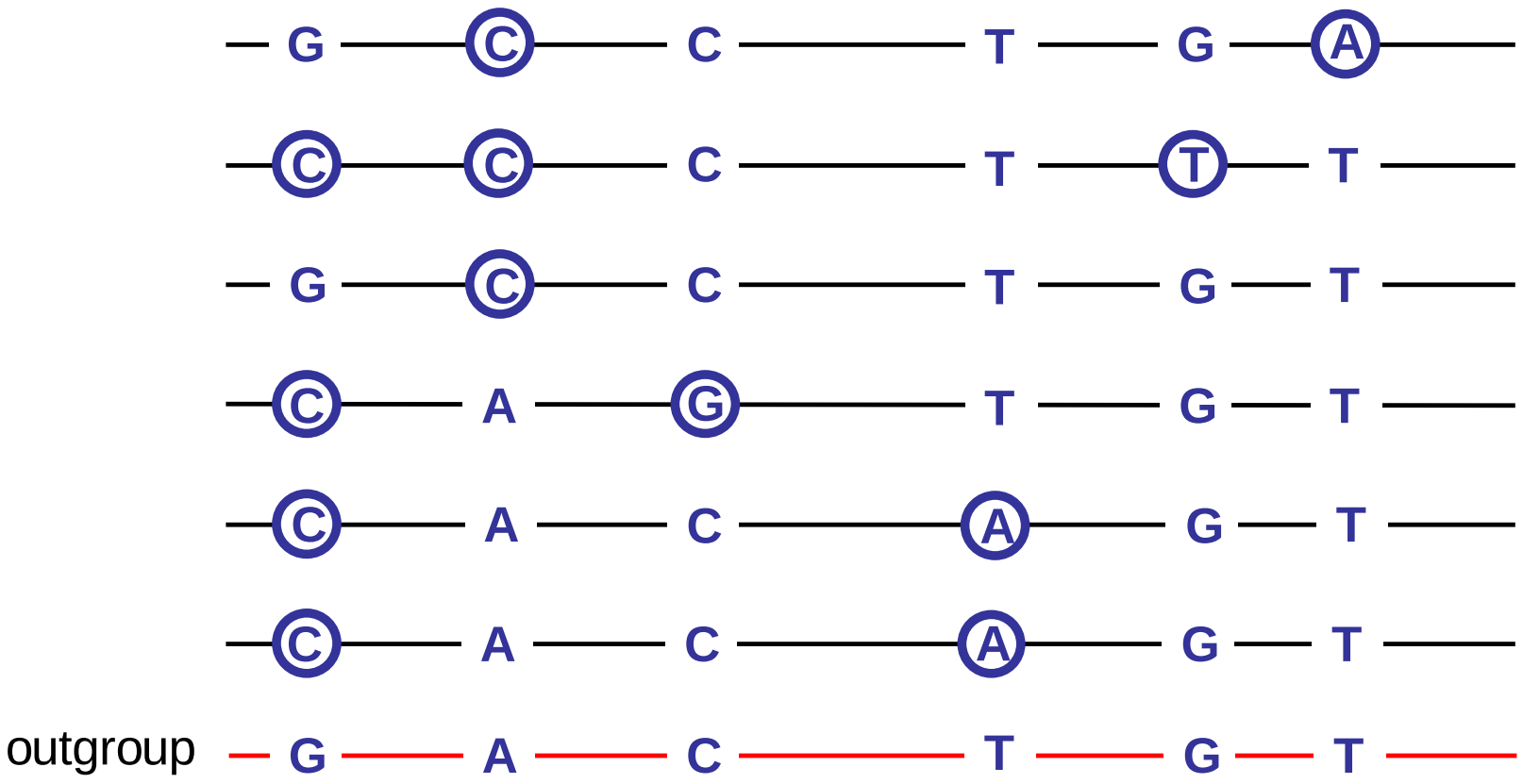
$4^{(6\times26)} = 8.3 \times 10^{93}$

Patterns of Evolution
”Summary Statistics”

only polymorphic sites …

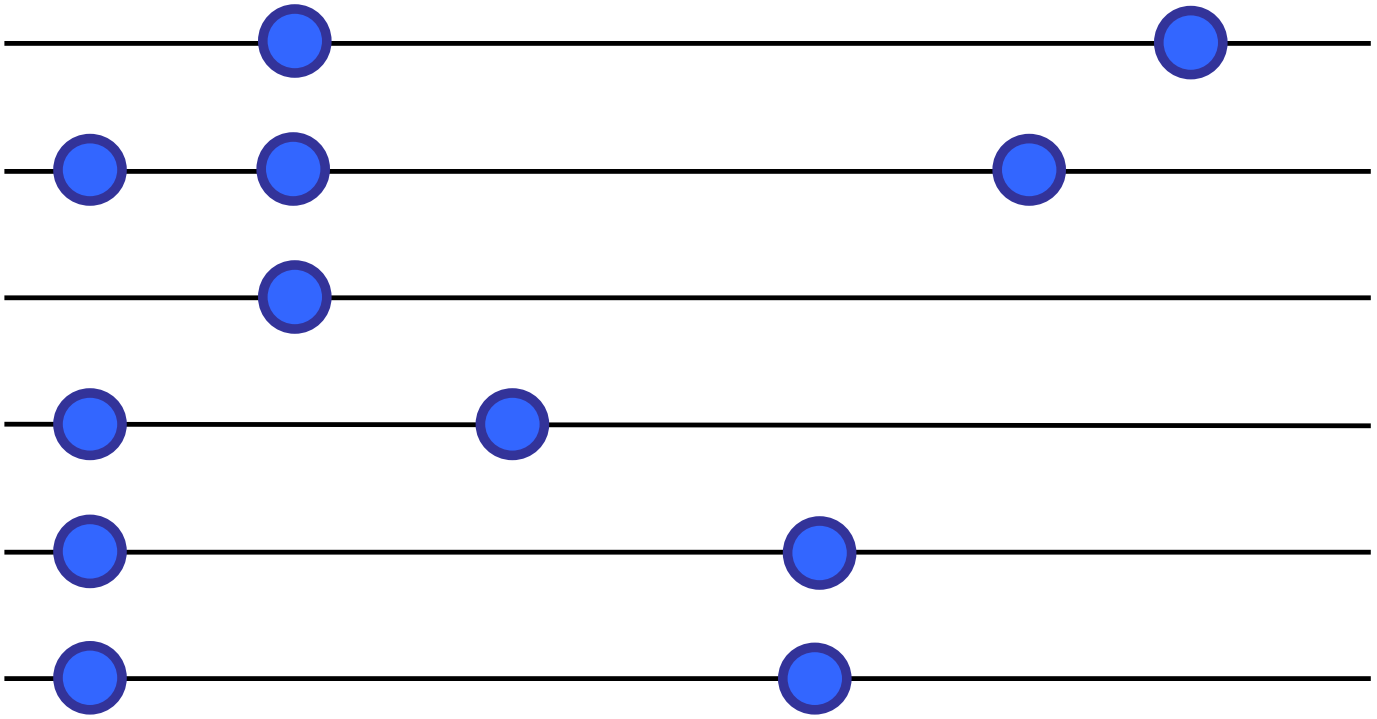
Patterns of Evolution
”Summary Statistics”

compare with outgroup …



Patterns of Evolution
”Summary Statistics”

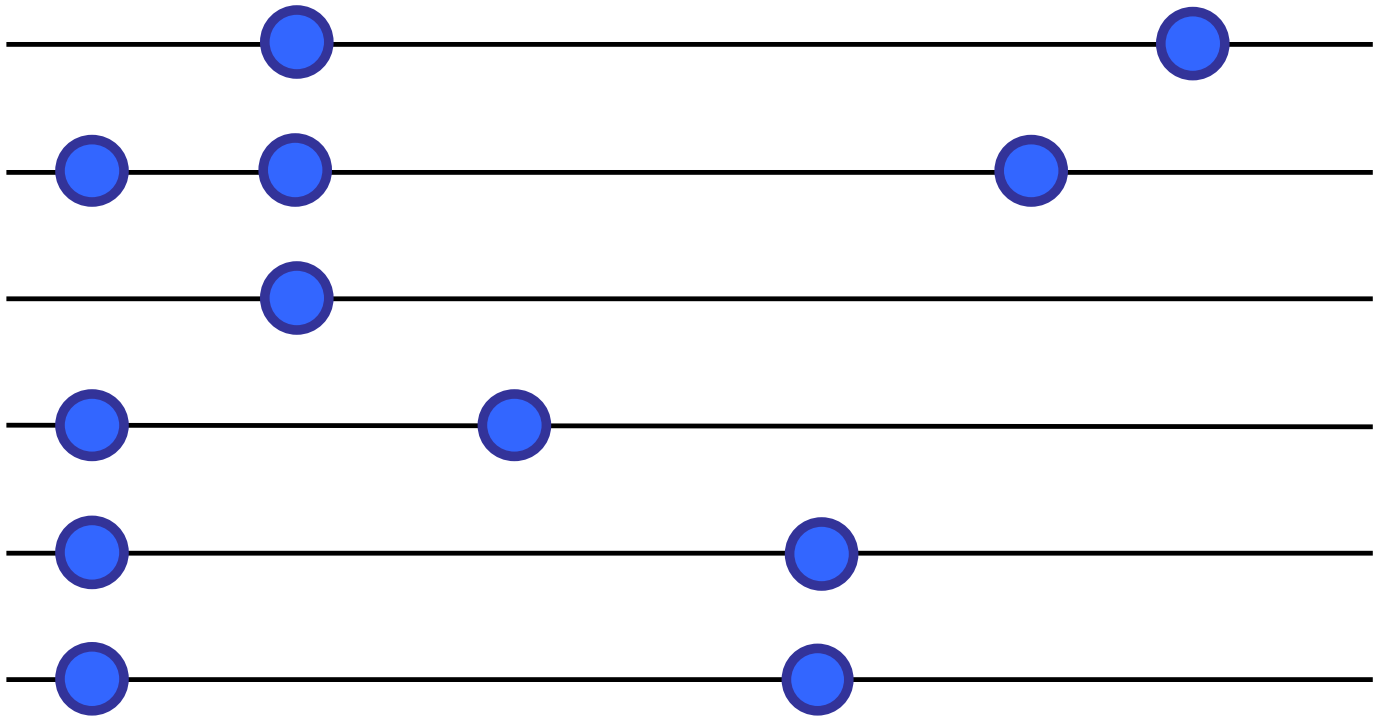
forget about molecular state …



(assumes *infinite sites mutation* model)

Patterns of Evolution
Summary statistics based on segregating sites

- number of segregating sites and allele frequencies

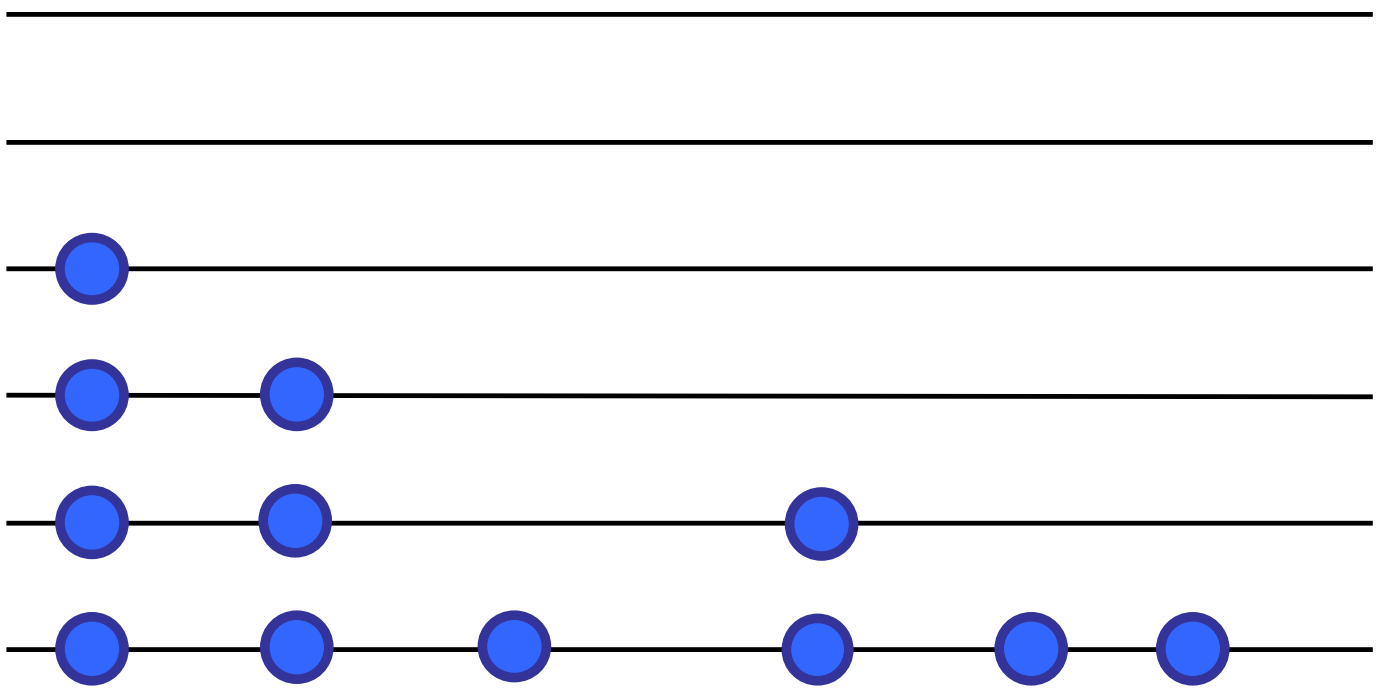


mutation *"size"*:

4 **3** **1** **2** **1** **1**

Patterns of Evolution
Summary statistics based on segregating sites

- number of segregating sites and allele frequencies
  - associations not important (“molecular bean bag“)

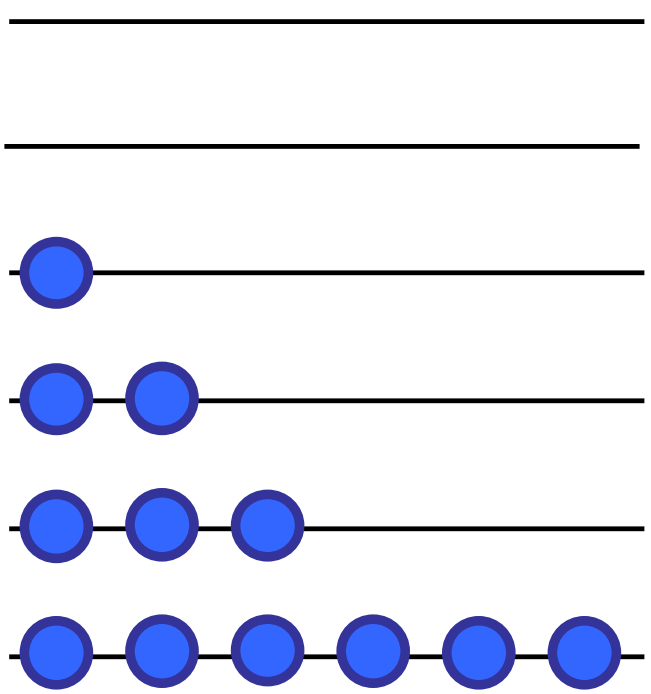


mutation “*size*“:

4 **3** **1** **2** **1** **1**

Patterns of Evolution
Summary statistics based on segregating sites

- number of segregating sites and allele frequencies
  - associations not important (“molecular bean bag“)



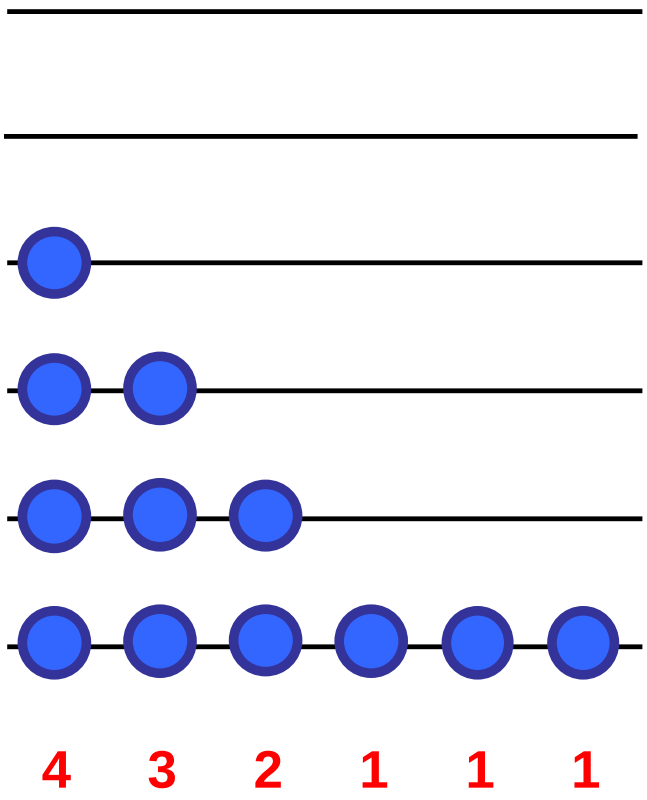
mutation “*size*“: **4 3 2 1 1 1**

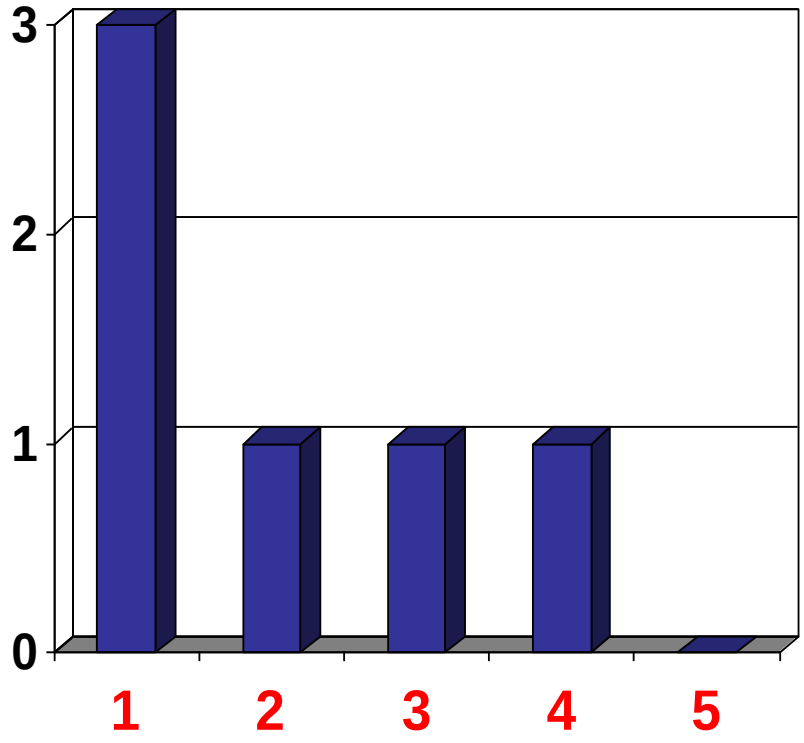
- genome position does not matter

Patterns of Evolution

Summary statistics based on segregating sites

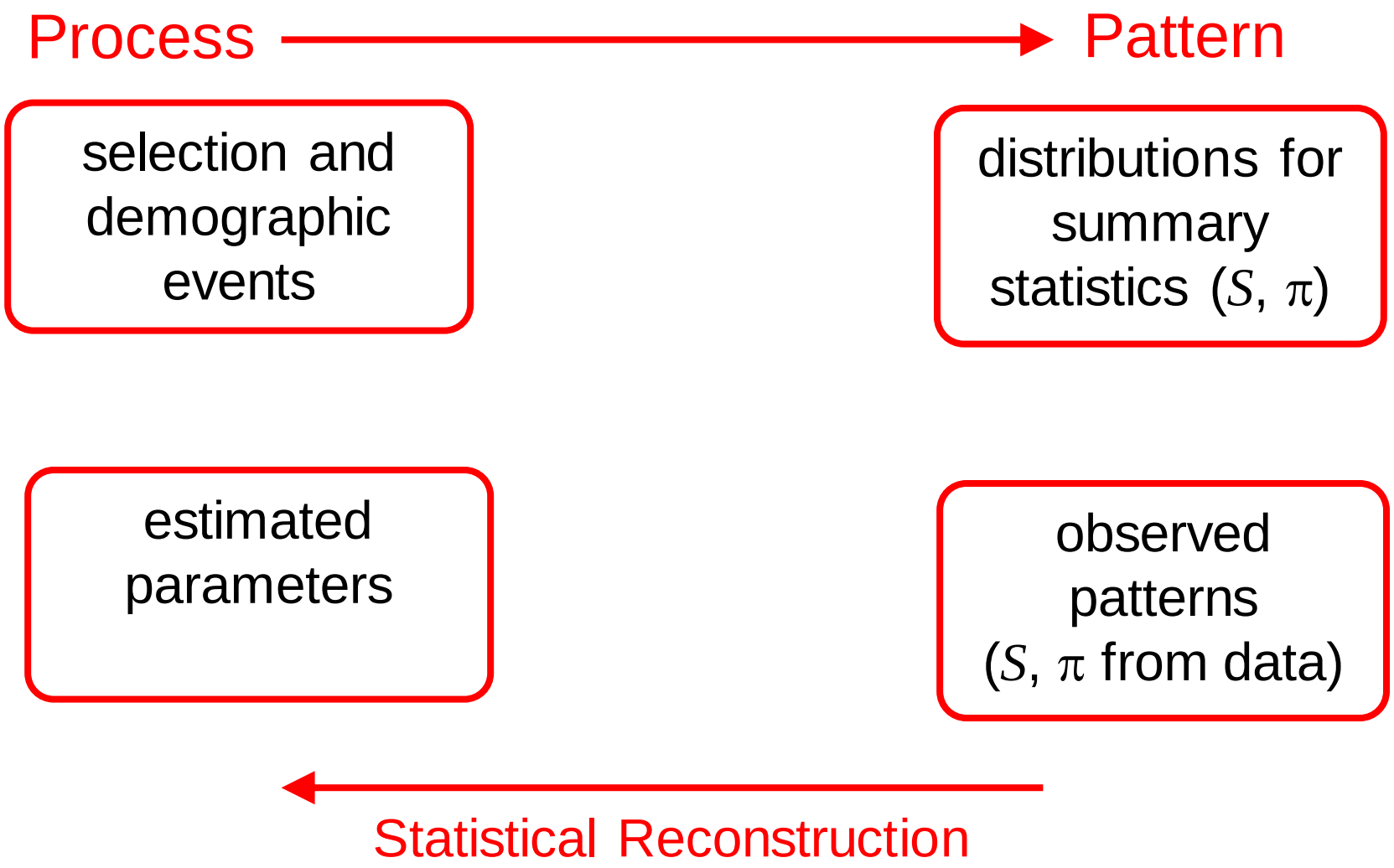
Site Frequency Spectrum





Patterns of Evolution

Reconstruction of evolutionary history



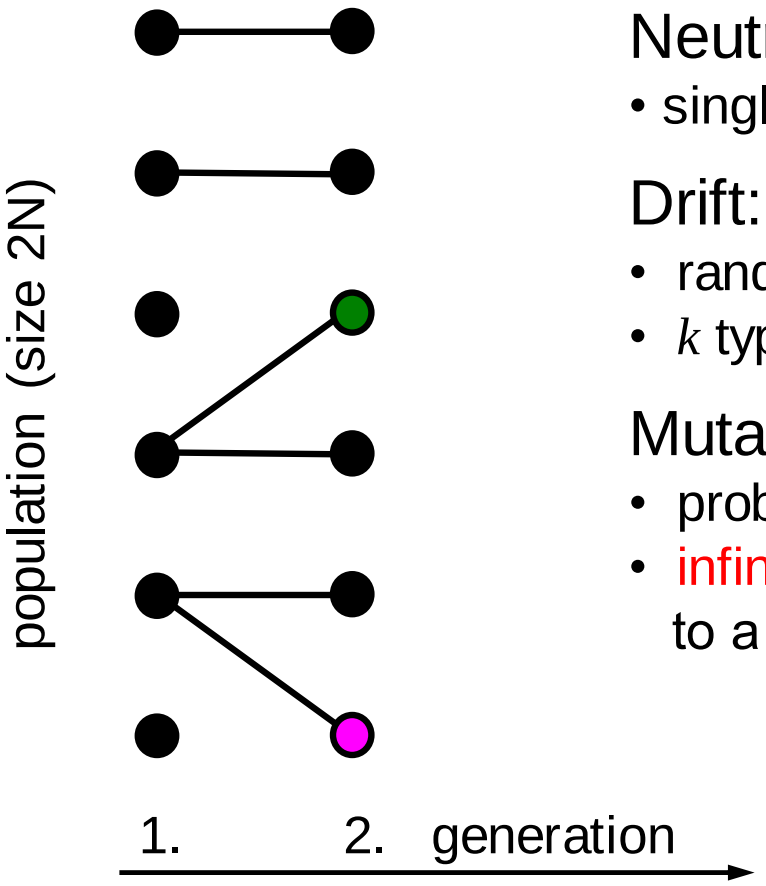
Patterns of Evolution

Reconstruction of evolutionary history

How does pure randomness look like ?

- Null-model of the evolutionary theory

Patterns of Evolution
Wright-Fisher model



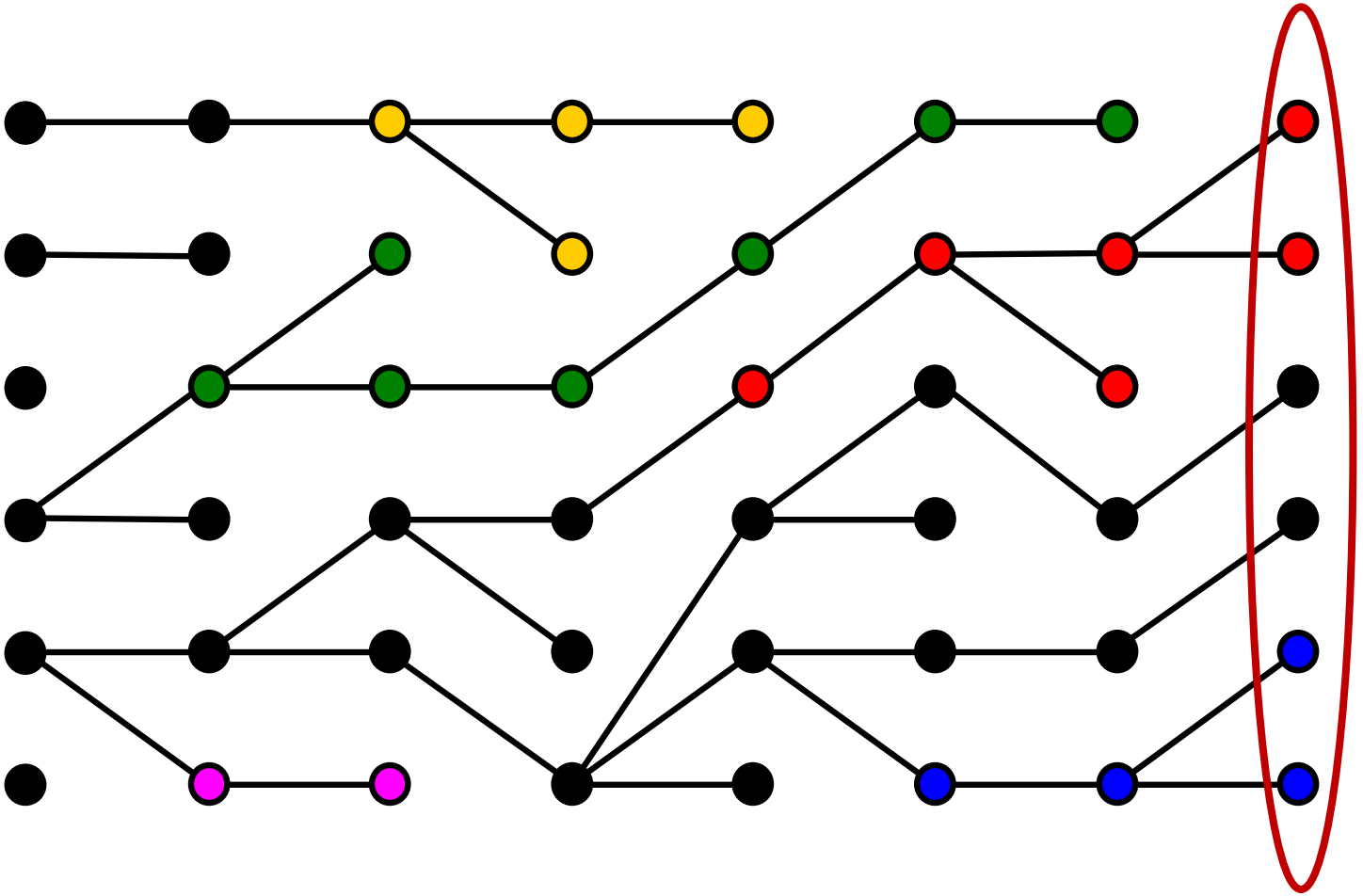
Neutral genetic variation
- single locus, multiple alleles

Drift:
- random sampling of parents
- k types: multinomial offspring distribution

Mutation:
- probability u for each offspring
- infinite alleles model: every mutation leads to a new allele (“new color”)

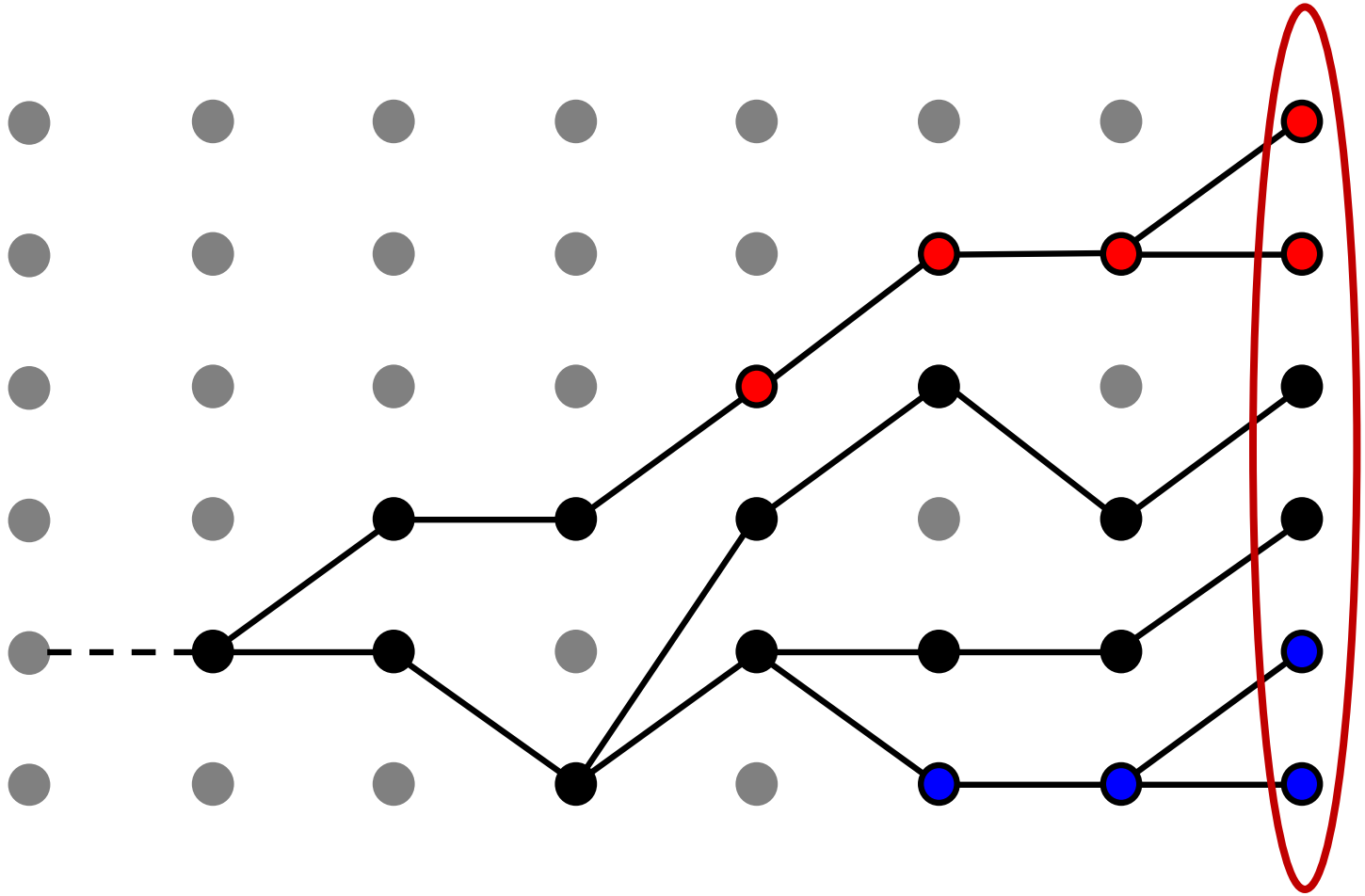
Patterns of Evolution

Wright-Fisher model



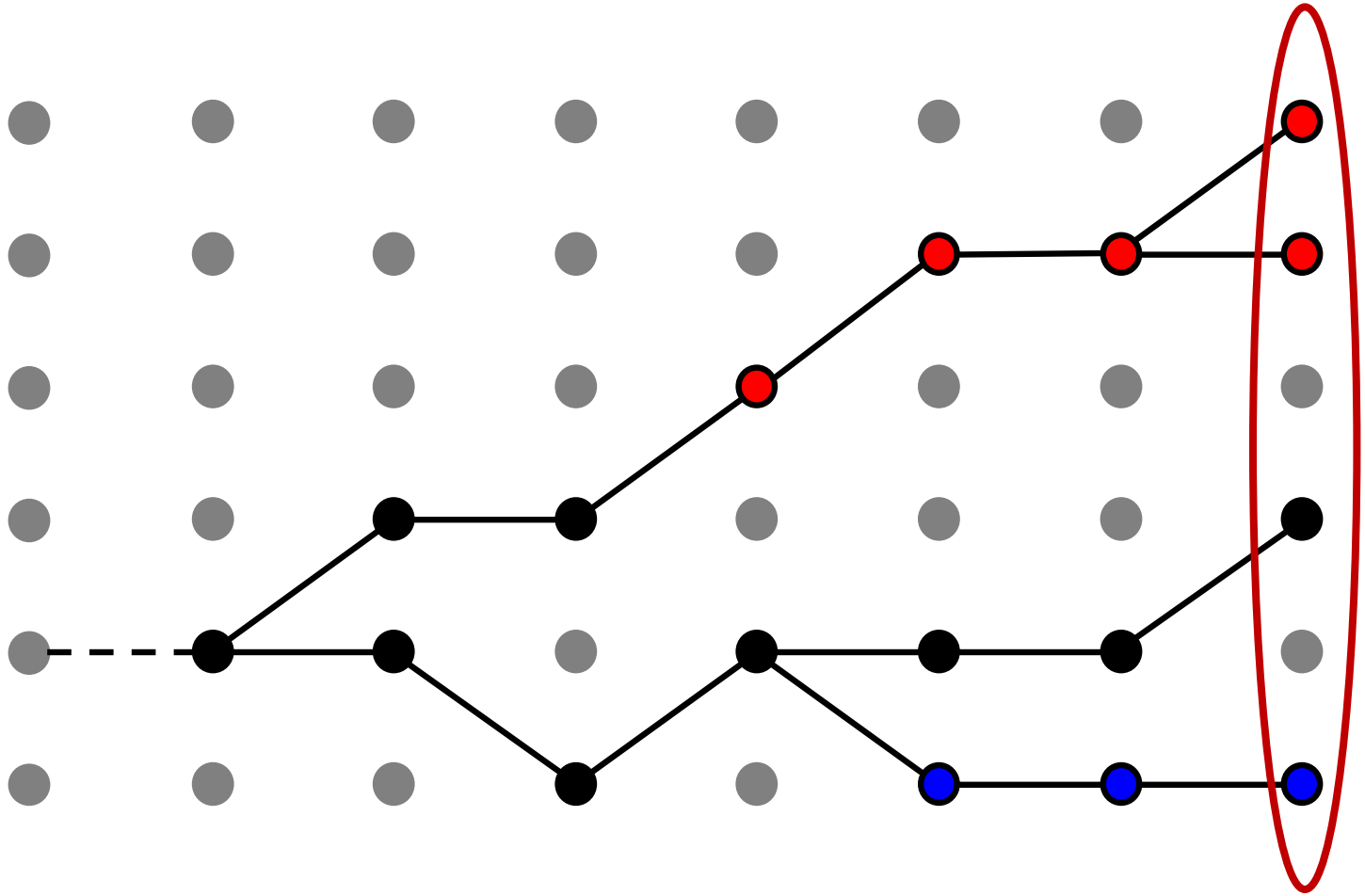
Patterns of Evolution

Wright-Fisher model

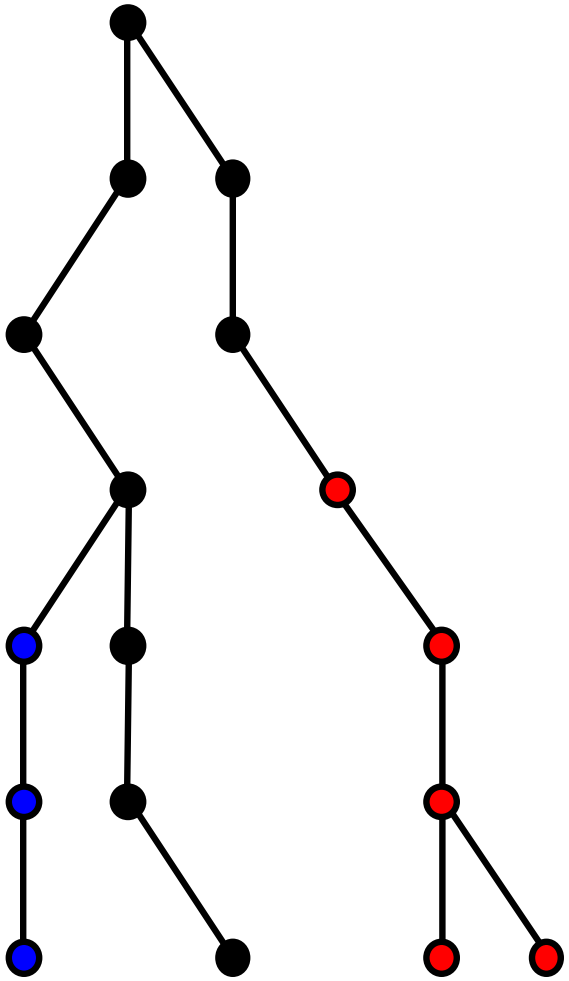


Patterns of Evolution

Wright-Fisher model

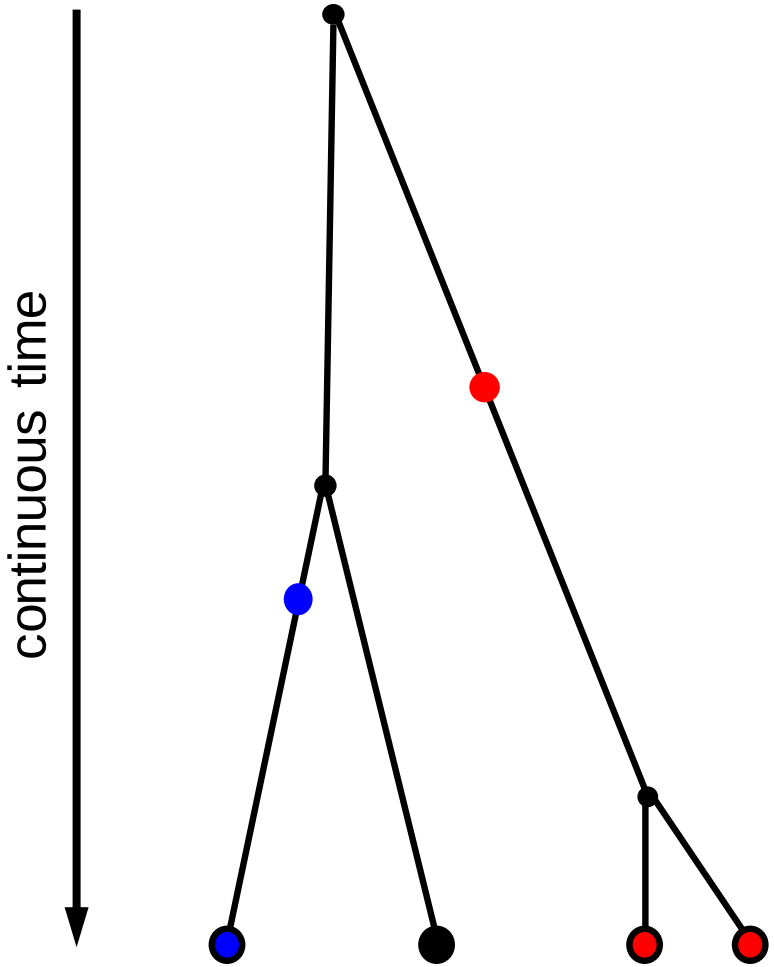


Patterns of Evolution
coalescence process



All information about the genetic variation pattern is contained in the sample genealogy.

Patterns of Evolution
coalescence process



All information about the genetic variation pattern is contained in the sample genealogy.

Construct a process to generate genealogies:

„coalescence-process“

Coalescent Theory
The standard neutral model

Haploid Wright-Fisher population of size $2N$:

- Genetic differences have no consequences on fitness
- No population subdivision

} Exchangable offspring distribution, independent of any *state label* (genotype, location, age, …)

⇩

- Constant population size

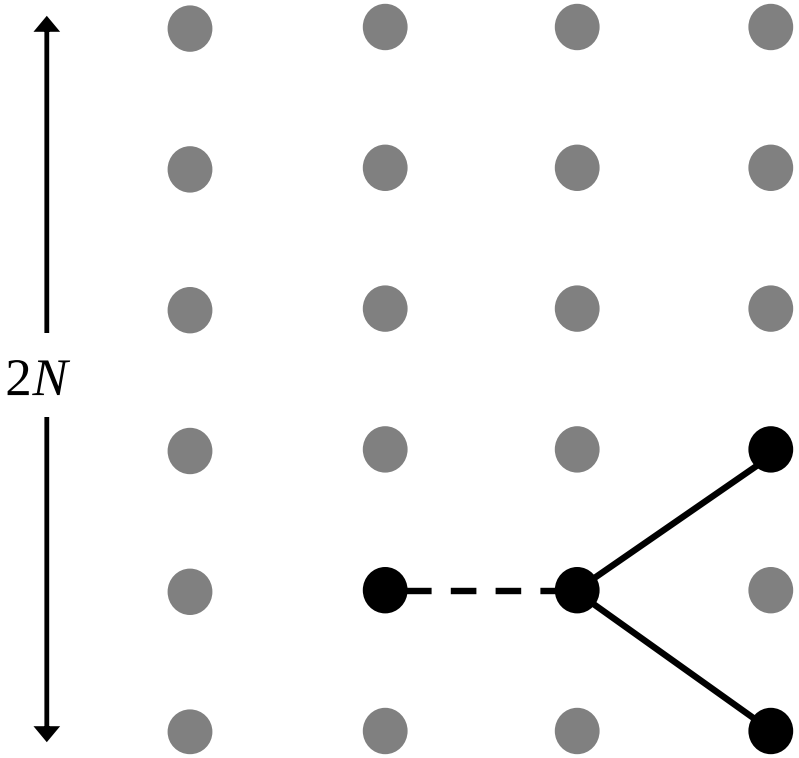
➢ Wright-Fisher: multinomial sampling

Individuals are equivalent with respect to descent

`State´ and `Descent´ are decoupled

⇨ 2 steps:
1. Construct genealogy independently of the state
2. Decide on the state only afterwards

Coalescent Theory
Construction of the Genealogy: Sample Size 2



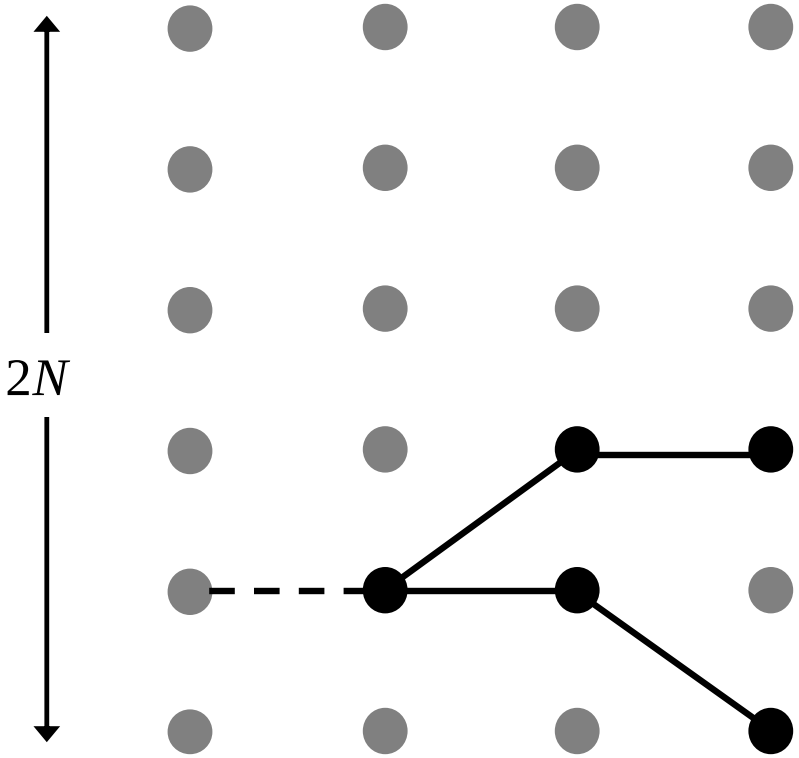
Coalescence probability

… in a single generation:

$$p_{c,1} = \frac{1}{2N}$$

Coalescent Theory

Construction of the Genealogy: Sample Size 2



Coalescence probability

… in a single generation:

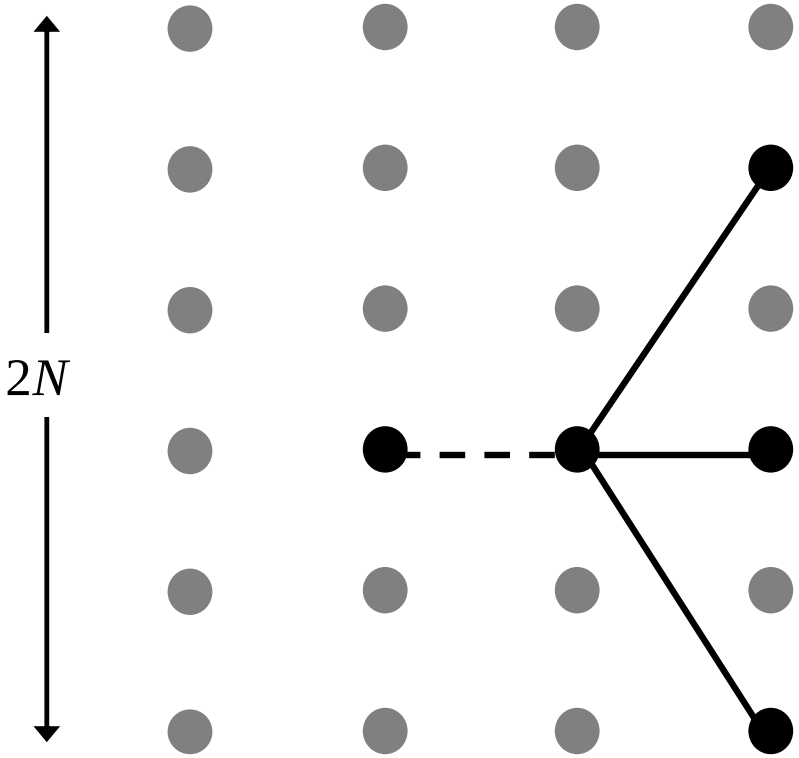
$$p_{c,1} = \frac{1}{2N}$$

… for exactly t generations:

$$p_{c,t} = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

Coalescent Theory

Construction of the Genealogy: Sample Size n

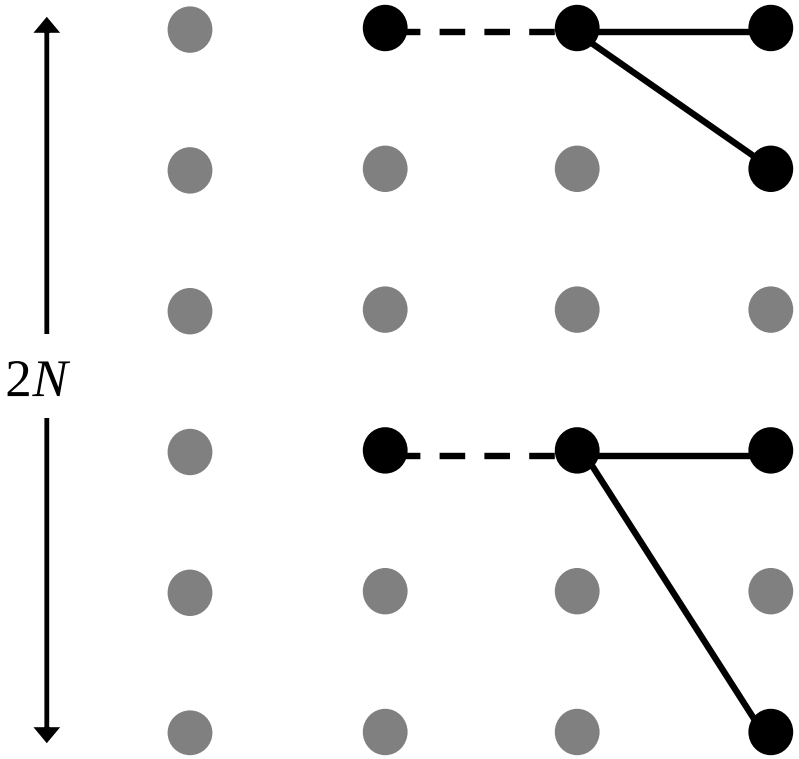


Multiple (e.g. triple) mergers:

$$p_{triple} = \frac{1}{4N^2} = \mathrm{O}\left[N^{-2}\right]$$

Coalescent Theory

Construction of the Genealogy: Sample Size n



Multiple (e.g. triple) mergers:

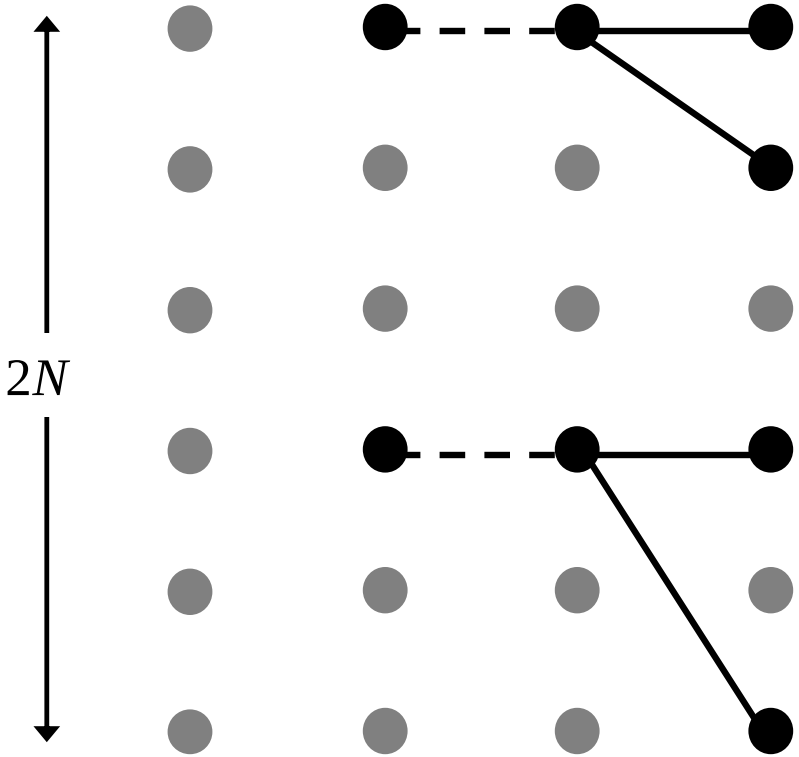
$$p_{triple} = \frac{1}{4N^2} = \mathrm{O}\left[N^{-2}\right]$$

Multiple coalescences:

$$\Pr \propto p_{c,t}^2 = \mathrm{O}\left[N^{-2}\right]$$

Coalescent Theory

Construction of the Genealogy: Sample Size n



Multiple (e.g. triple) mergers:

$$p_{triple} = \frac{1}{4N^2} = \mathrm{O}\left[N^{-2}\right]$$

Multiple coalescences:

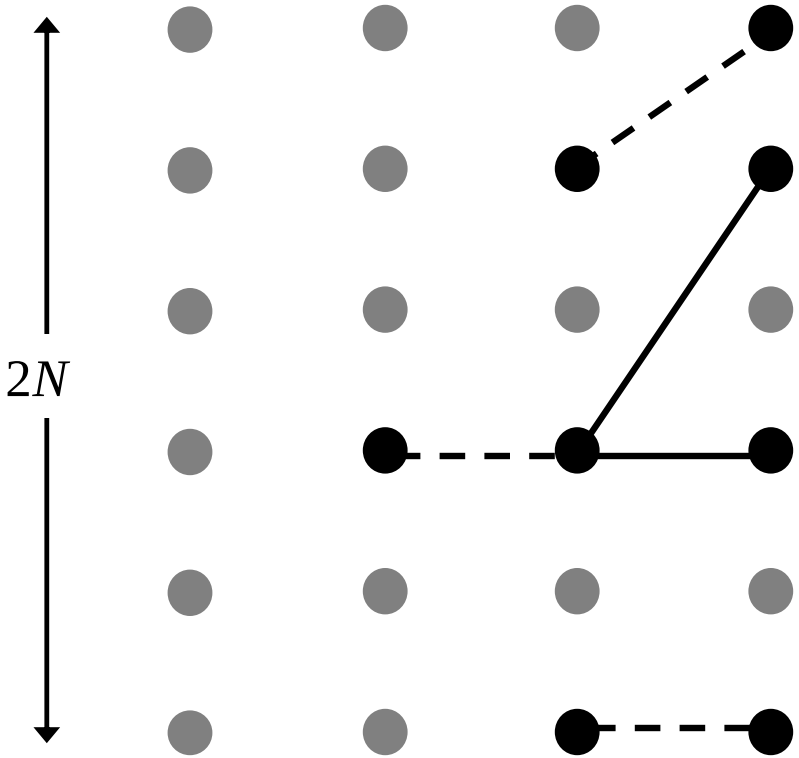
$$\Pr \propto p_{c,t}^2 = \mathrm{O}\left[N^{-2}\right]$$

can be ignored if $N >> n$:
only binary mergers for $N \to \infty$

"Kingman coalescent"

Coalescent Theory

Construction of the Genealogy: Sample Size n



Coalescence probability (single binary merger)

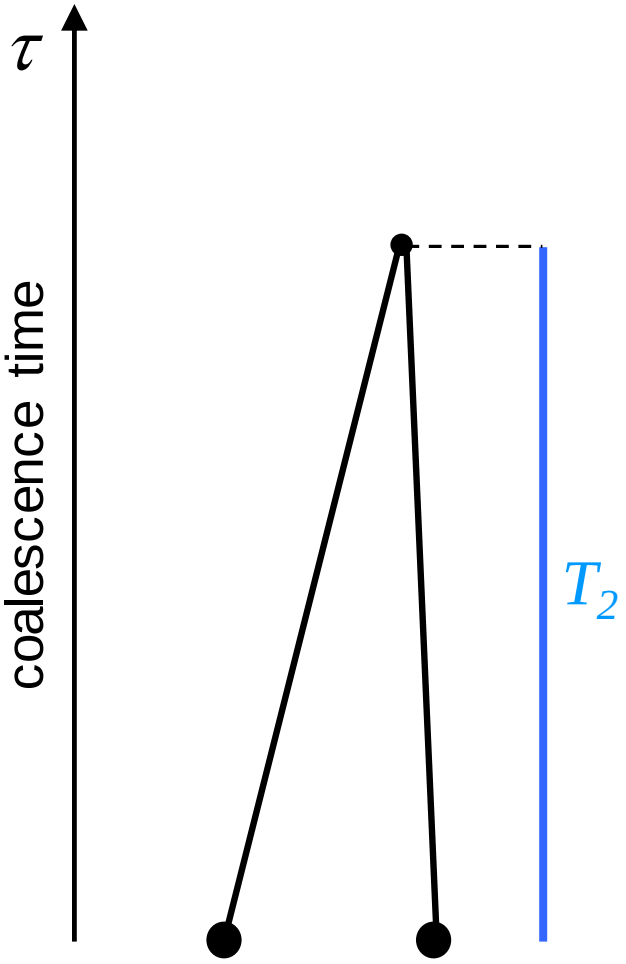
… in a single generation:

$$p_{c,1}^{(n)} = \frac{1}{2N}\binom{n}{2} = \frac{n(n-1)}{4N}$$

… for exactly t generations:

$$p_{c,t}^{(n)} = \left(1 - \frac{n(n-1)}{4N}\right)^{t-1} \frac{n(n-1)}{4N}$$

Coalescent Theory
Distribution of Coalescence Times



Define coalescence time scale:

$$\tau = \frac{t}{2N}$$

Coalescence time T_2 for sample size 2:

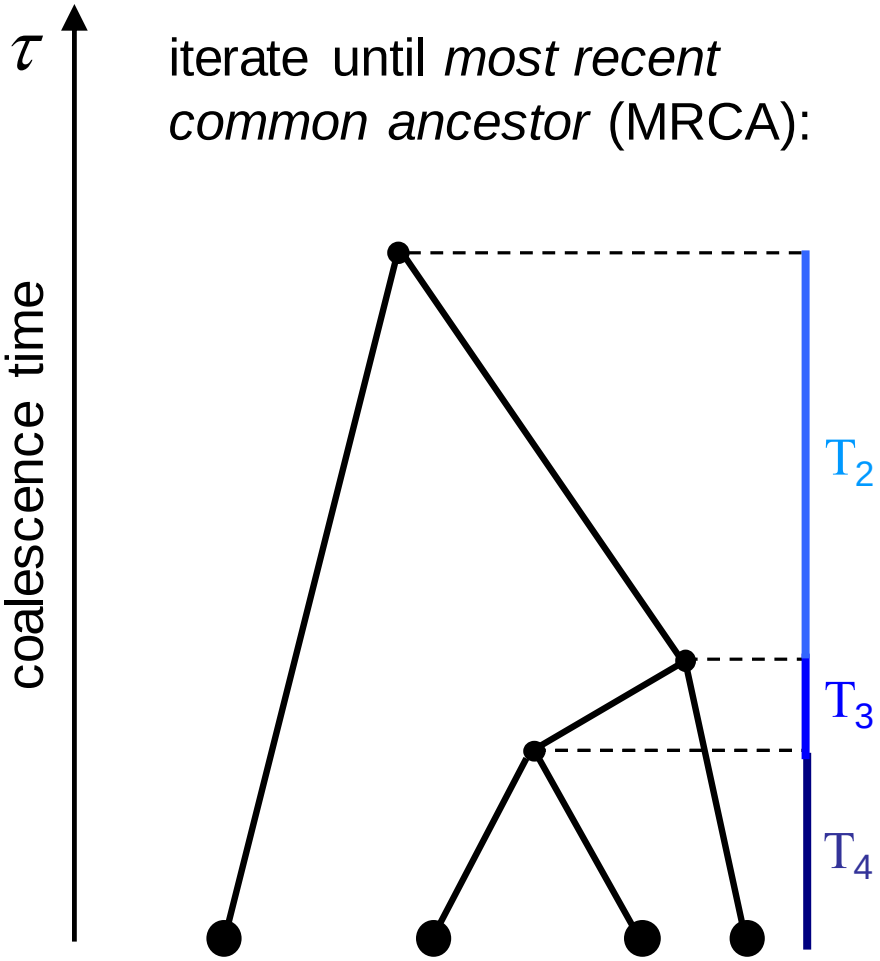
$$\Pr[T_2 > \tau] = \left(1 - \frac{1}{2N}\right)^{2N\tau}$$

$$\xrightarrow{N \to \infty} \exp[-\tau]$$

Exponential distribution with parameter 1:

$$\mathrm{E}[T_2] = 1 \qquad (2N \text{ generations})$$

Coalescent Theory
Distribution of Coalescence Times



iterate until *most recent common ancestor* (MRCA):

with sample size n:

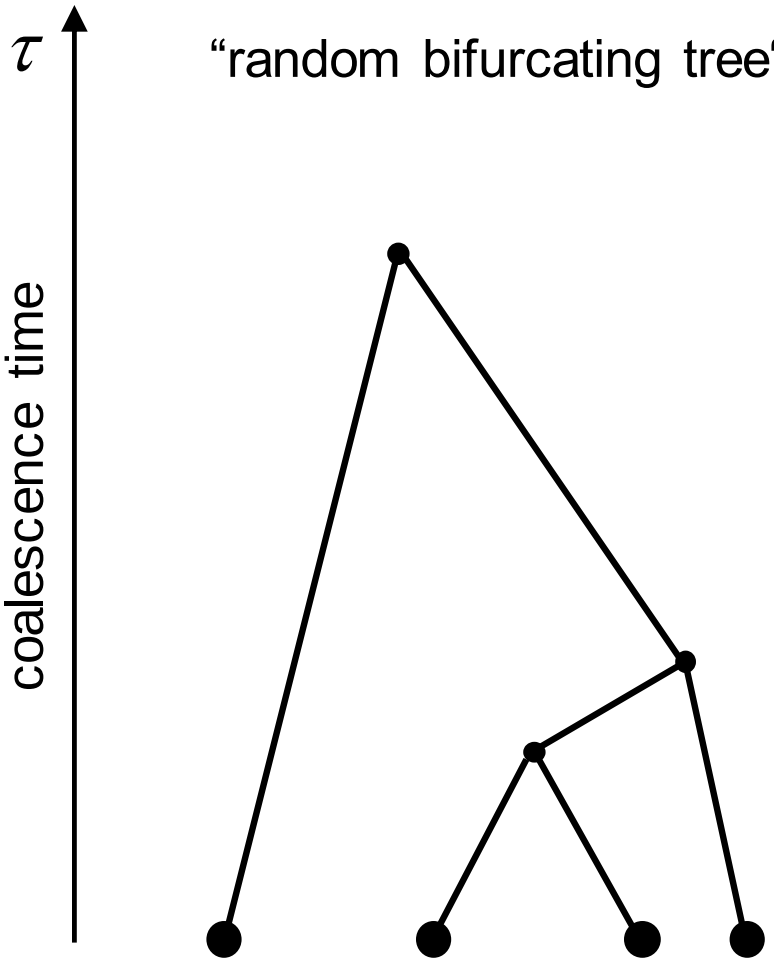
$$\Pr[T_n > \tau] = \left(1 - \frac{1}{2N}\binom{n}{2}\right)^{2N\tau}$$

$$\xrightarrow{N\to\infty} \exp\left[-\binom{n}{2}\tau\right]$$

Exponential distribution with parameter : $\binom{n}{2} = \frac{n(n-1)}{2}$

$$\mathrm{E}[T_n] = \frac{2}{n(n-1)}$$

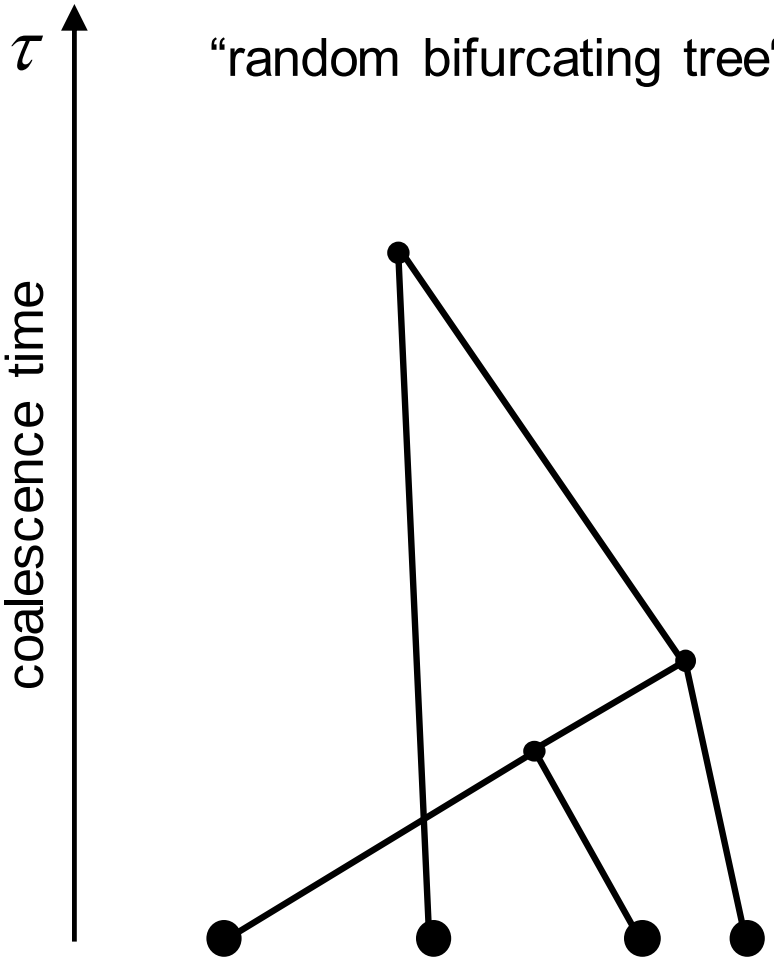
Coalescent Theory
Tree Topologies



"random bifurcating tree"

- pick two random individuals from the sample and merge
- sample size $n \rightarrow n\text{-}1$ and iterate until $n = 1$ (MRCA)
- all individuals exchangable
  - ➢ topology invariant under permutation of "leaves"

Coalescent Theory
Tree Topologies



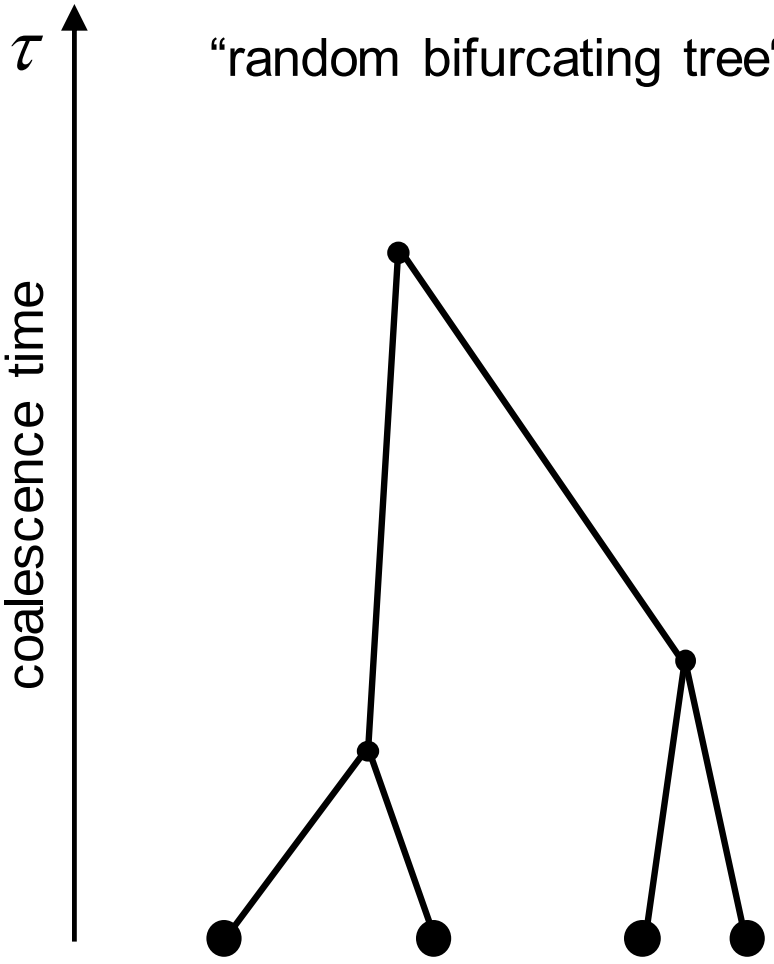
"random bifurcating tree"

- pick two random individuals from the sample and merge
- sample size $n \rightarrow n\text{-}1$ and iterate until $n = 1$ (MRCA)
- all individuals exchangable

➢ topology invariant under permutation of "leaves"

same topology

Coalescent Theory
Tree Topologies



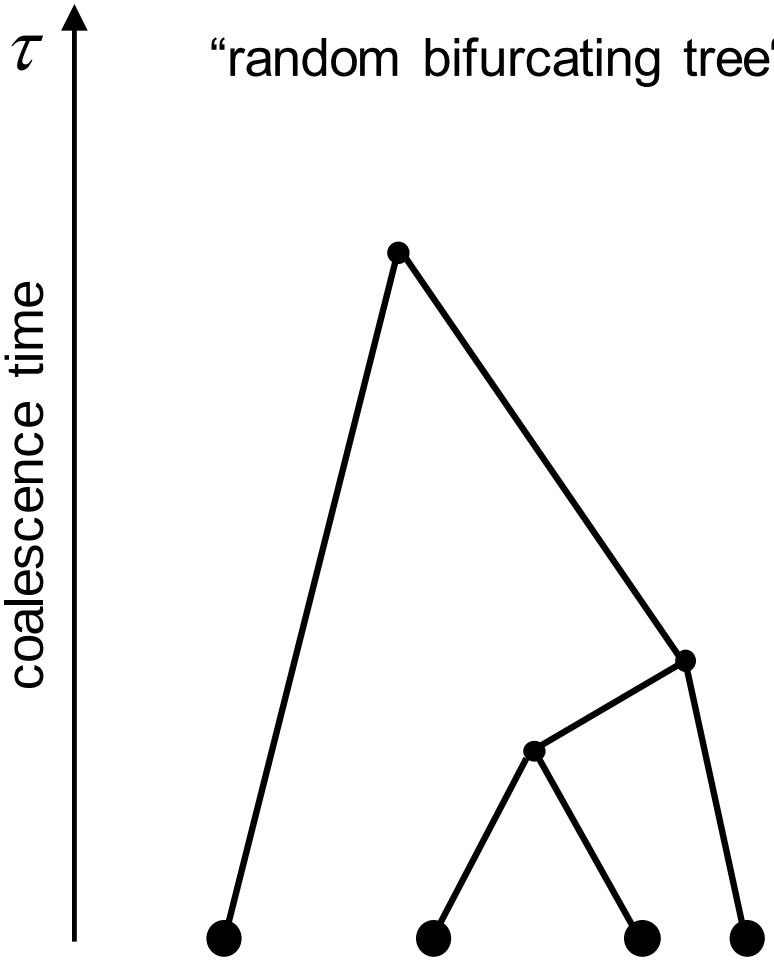
"random bifurcating tree"

- pick two random individuals from the sample and merge
- sample size $n \rightarrow n\text{-}1$ and iterate until $n = 1$ (MRCA)
- all individuals exchangable

➢ topology invariant under permutation of "leaves"

different topology

Coalescent Theory
Tree Topologies



"random bifurcating tree"

- pick two random individuals from the sample and merge
- sample size $n \rightarrow n\text{-}1$ and iterate until $n = 1$ (MRCA)
- all individuals exchangable

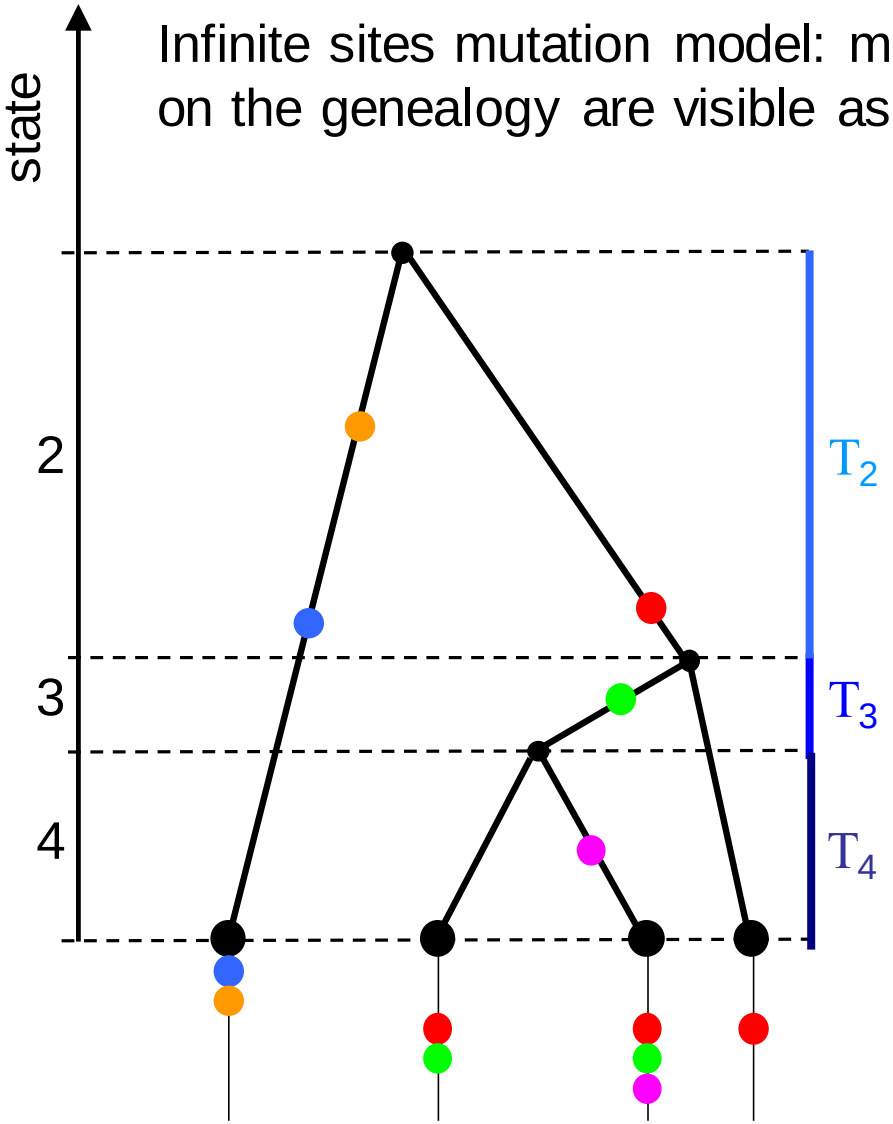
➢ topology invariant under permutation of "leaves"

Distribution of tree topologies

- *independent of coalescence times*
- depends only on the separation of state and descent and on the "no multiple merger" condition

Coalescent Theory
Mutation "Dropping"

Infinite sites mutation model: mutation rate u, all mutations on the genealogy are visible as polymorphisms on different sites



- only number of mutations on each branch matters
- Poisson distributed with parameter $2Nu \cdot L = \frac{\theta \cdot L}{2}$, $L = \sum_{i=j}^{k} T_i$ branch length of branch from state j through k

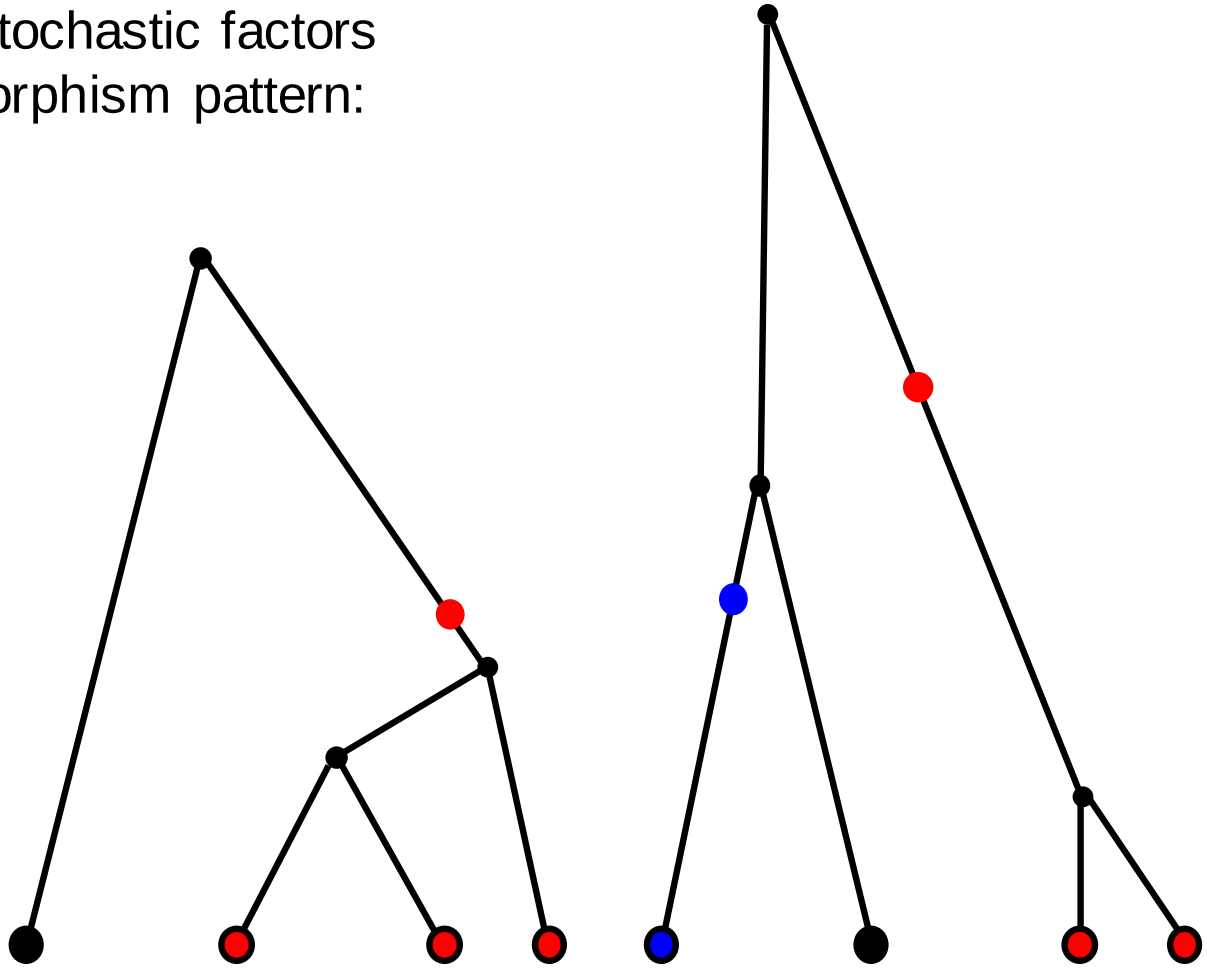
(also other mutation schemes possible)

Coalescent Theory
Basic Properties

Three independent stochastic factors determine the polymorphism pattern:

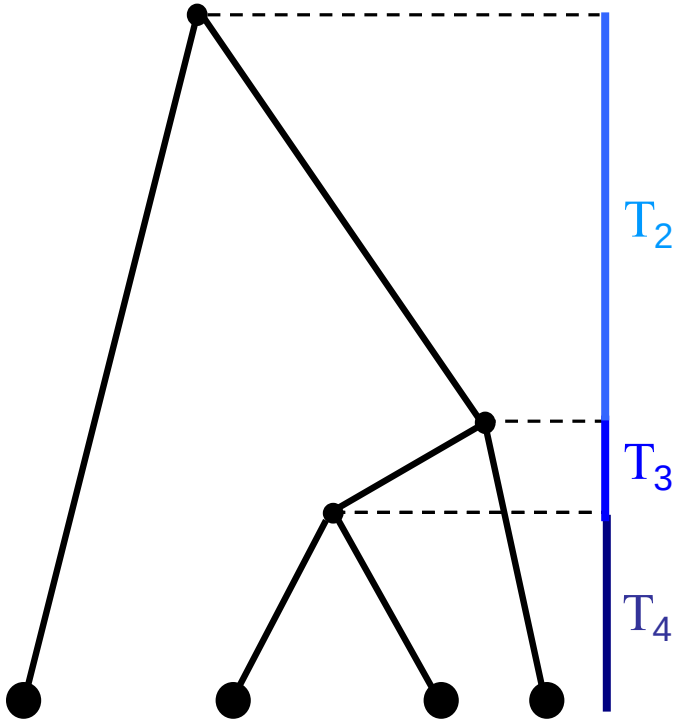
1. coalescent times
2. tree topology
3. mutation

(very easy to implement in simulations)



Coalescent Theory
Basic Properties

Time to the most recent common ancestor:



$$E[T_{MRCA}] = \sum_{k=2}^{n} E[T_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)}$$

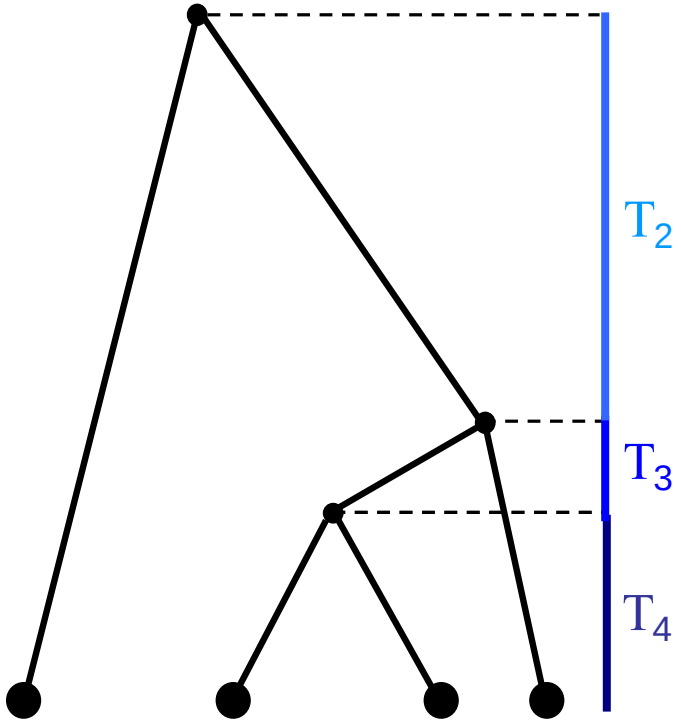
$$= 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right)$$

Compare: $E[T_2] = 1$

More than half for the last two branches!

Coalescent Theory
Basic Properties

Total length of the tree and expected number of polymorphic sites:



$$E[L_{tree}] = \sum_{k=2}^{n} kE[T_k] = 2\sum_{k=1}^{n-1} \frac{1}{k}$$

$$\Rightarrow E[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \theta \cdot a_n$$

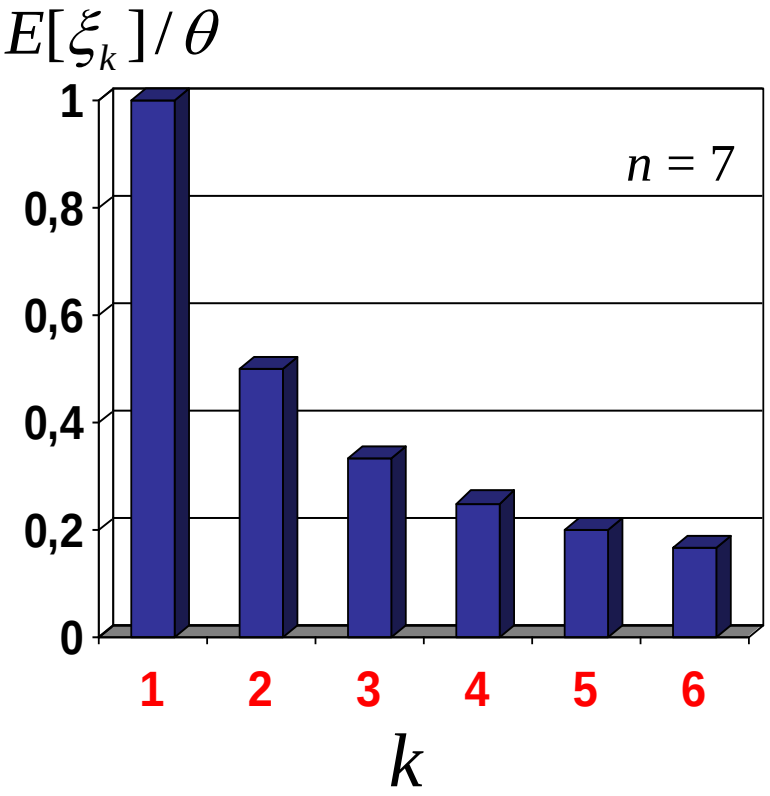
with: $a_n = \sum_{k=1}^{n-1} \frac{1}{k} \rightarrow \log n + 0.577$

(logarithmic dependence on sample size)

Coalescent Theory
Basic Properties

Expected site frequency spectrum:

ξ_k Number of mutations that appear k times in the sample (= *of size* k)



$$E[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \sum_{k=1}^{n-1} E[\xi_k]$$

indeed: $$E[\xi_k] = \frac{\theta}{k}$$

in particular: $$E[\xi_1] = \theta$$