The Coalescent

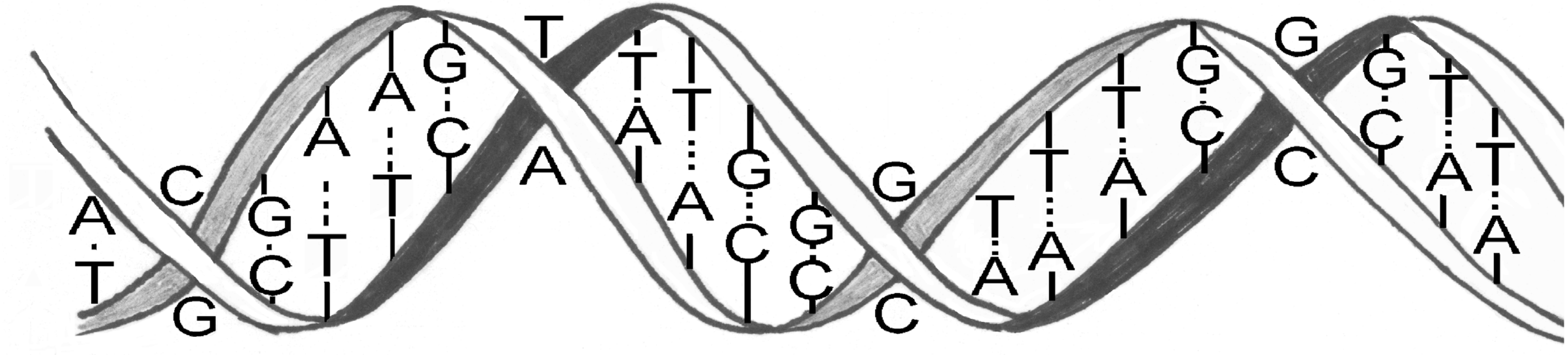
Neutral evolution backward in time

Joachim Hermisson

Mathematics and Biosciences Group
Mathematics & Max Perutz Labs,
University of Vienna

Introduction to the Coalescent

data, data, data, …

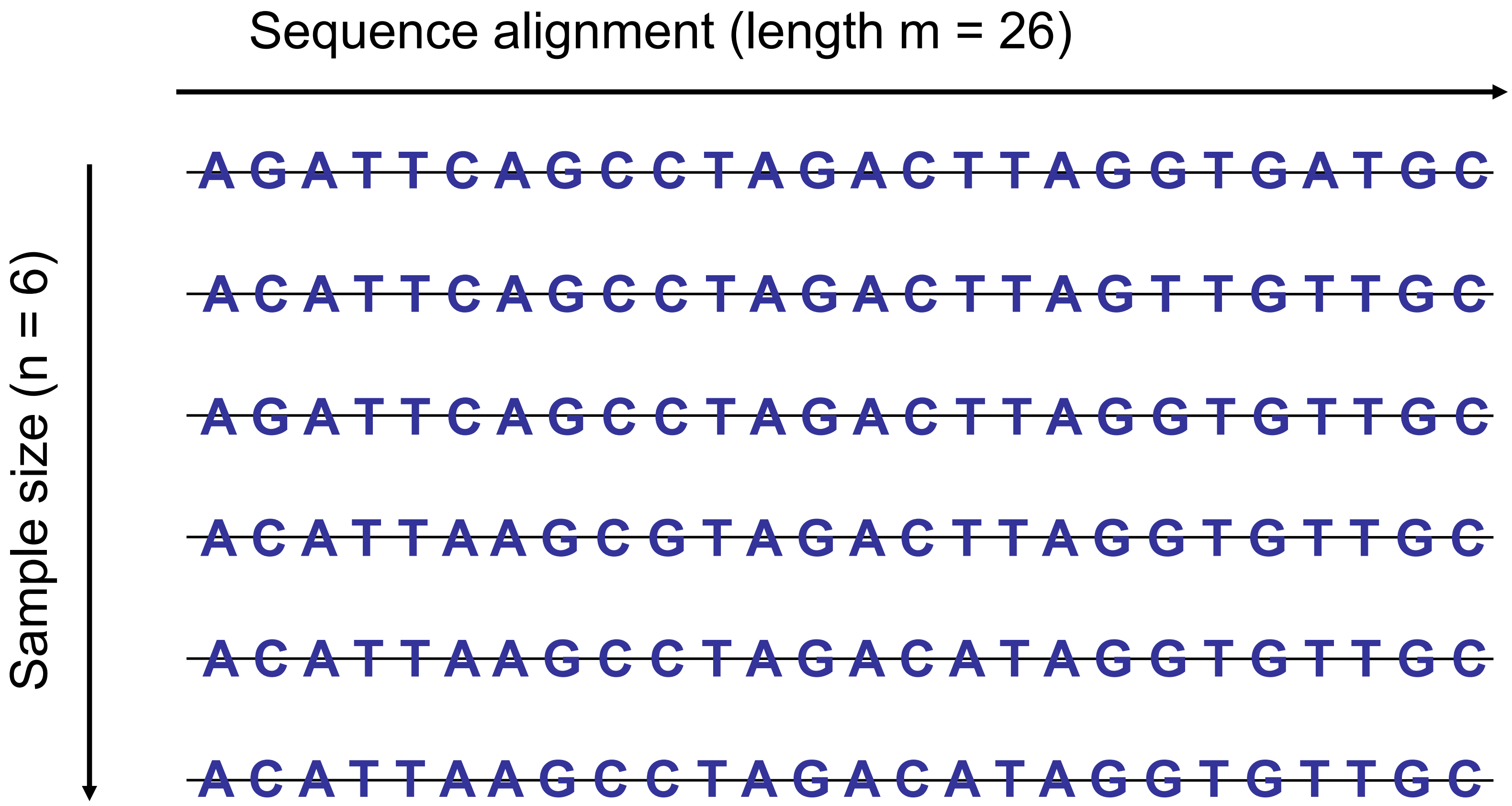


Massive accumulation of DNA sequence data

- *1980's:* Sequencing single genes (some 1000 base pairs) takes a 3-4 years PhD project
- *1990 – 2003:* Human Genome Project (~ 3 10^9 (3 billion) bases) expected: 3 billion $, final: ~ 300 Mio $
- *since 2010:* 1000 Genome Projects, first for Humans, then also for *Drosophila, Arabidopsis* …
- *today*: GWAS sample sizes > 500 000 (UK-Biobank), long reads …

Patterns of Evolution

”Summary Statistics”



Sequence alignment (length m = 26)

Sample size (n = 6)

A G A T T C A G C C T A G A C T T A G G T G A T G C

A C A T T C A G C C T A G A C T T A G T T G T T G C

A G A T T C A G C C T A G A C T T A G G T G T T G C

A C A T T A A G C G T A G A C T T A G G T G T T G C

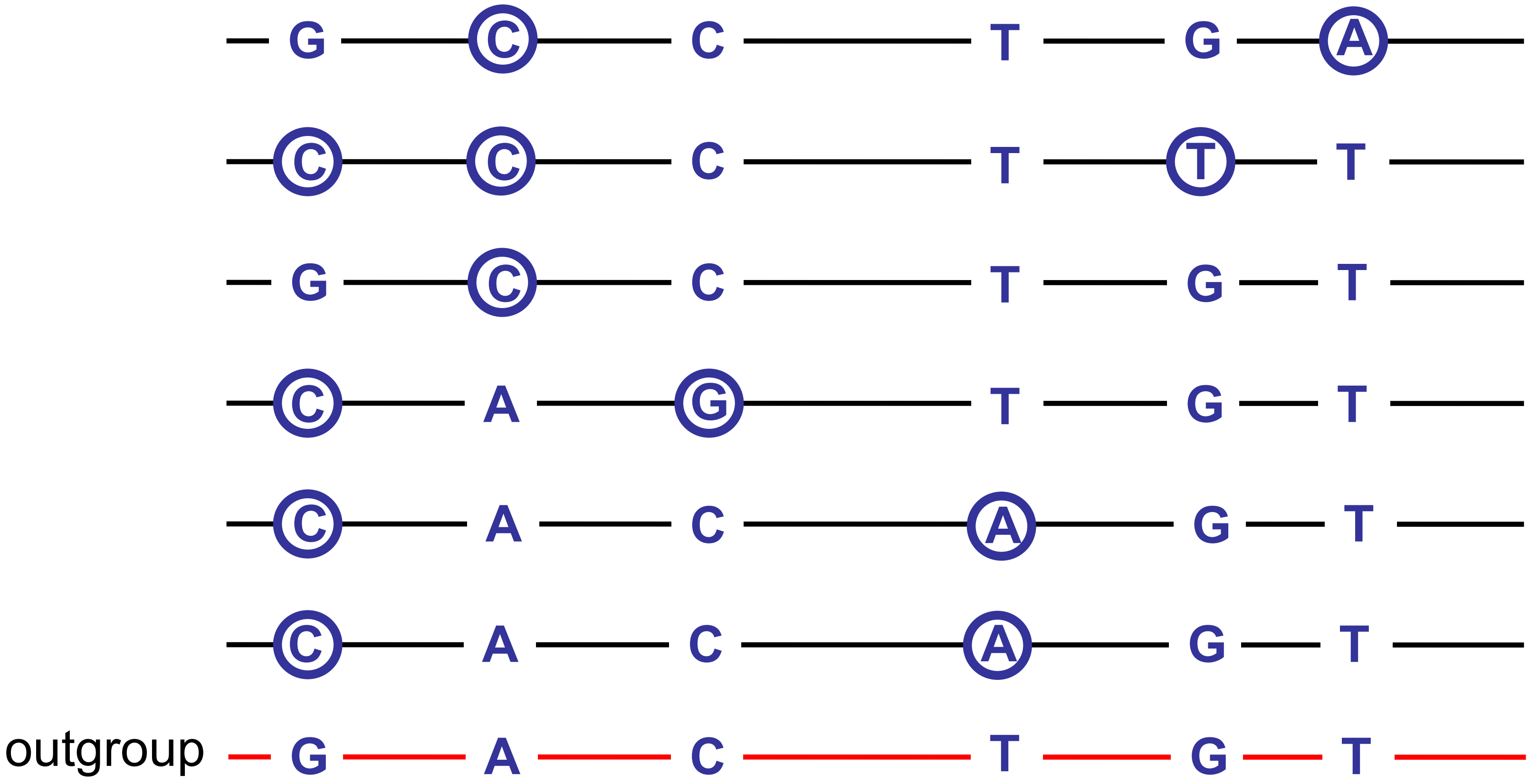
A C A T T A A G C C T A G A C A T A G G T G T T G C

A C A T T A A G C C T A G A C A T A G G T G T T G C

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

Patterns of Evolution
”Summary Statistics”

compare with outgroup …



G C C T G A

C C C T T T

G C C T G T

C A G T G T

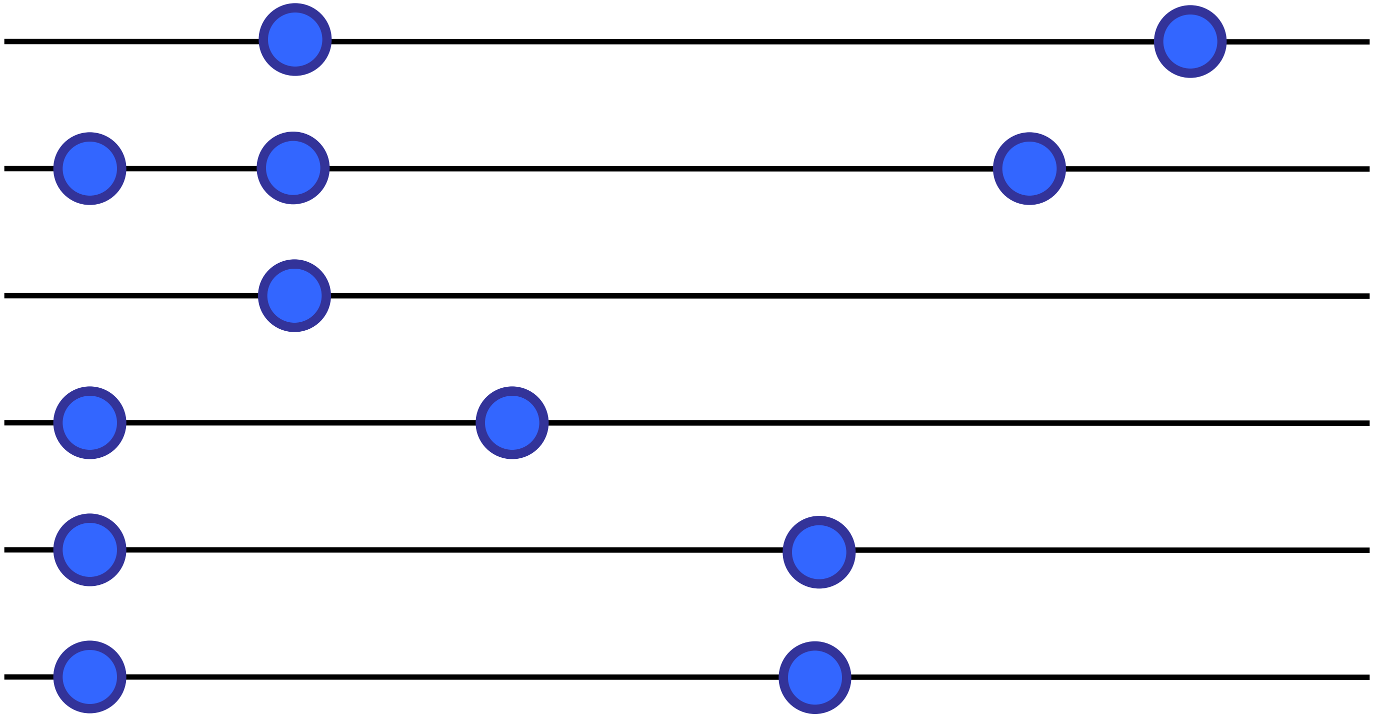
C A C A G T

C A C A G T

outgroup G A C T G T

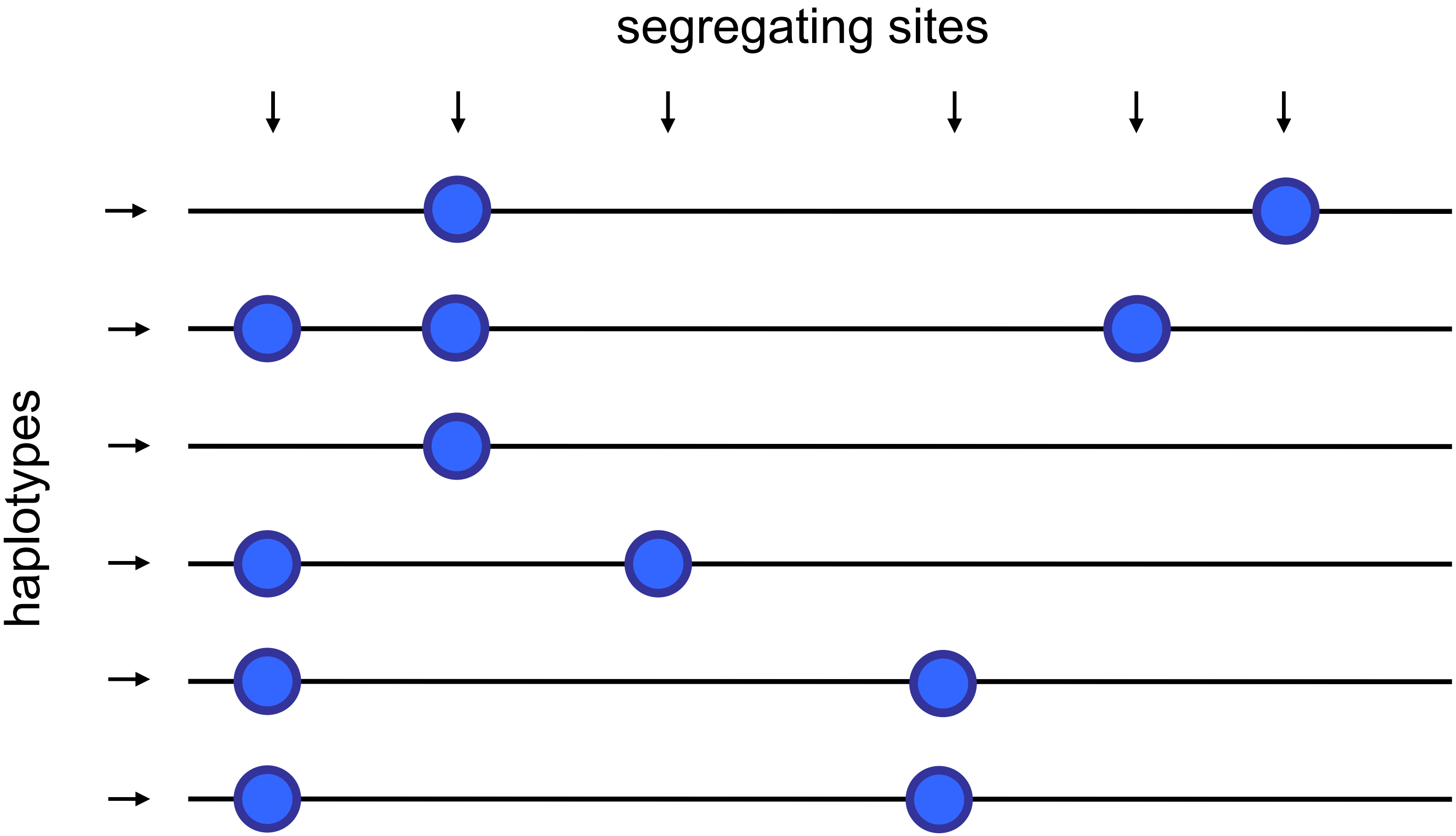
Patterns of Evolution
”Summary Statistics”

forget about molecular state …



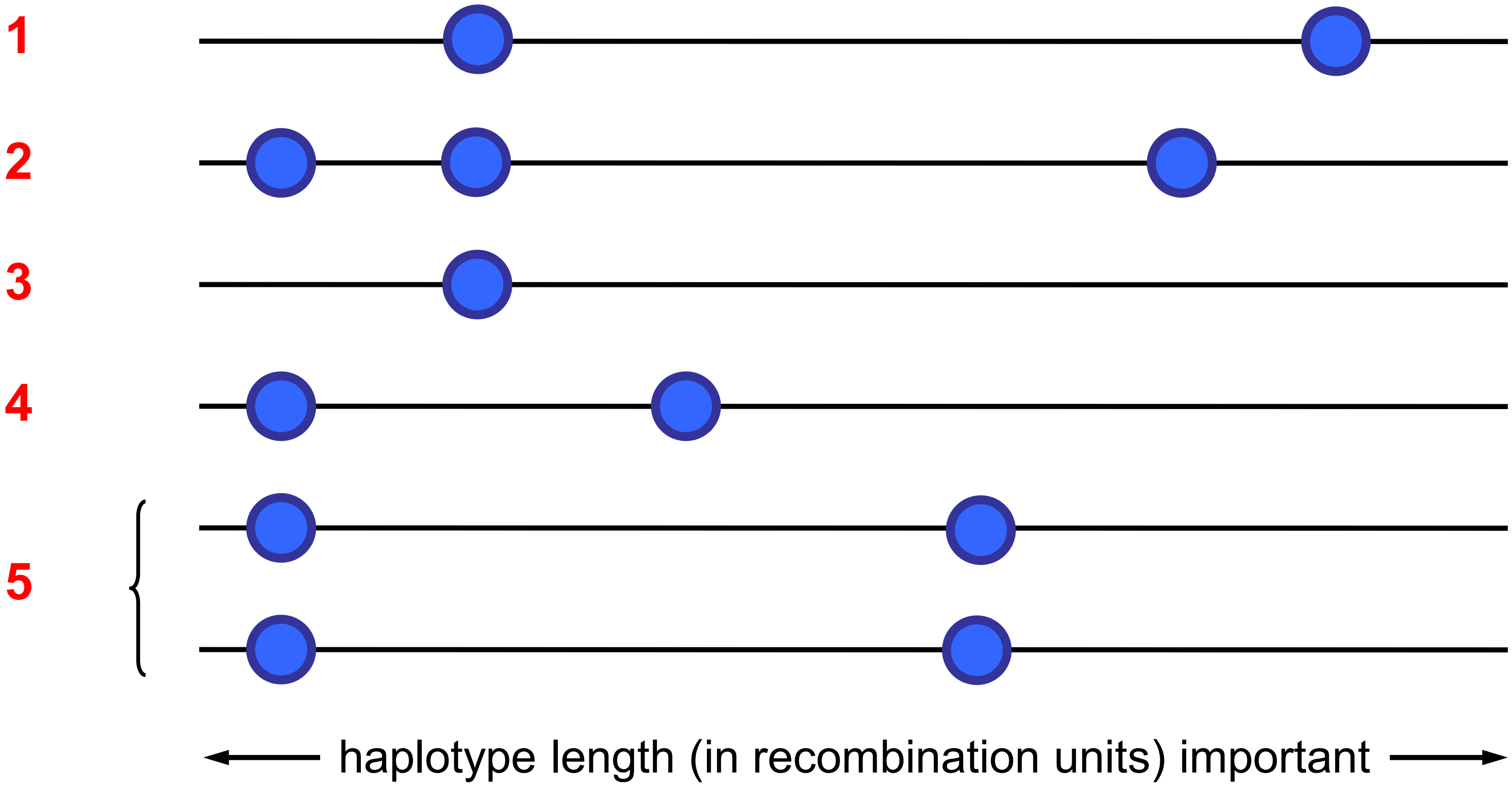
(assumes *infinite sites mutation* model)

Patterns of Evolution
”Summary Statistics”



Patterns of Evolution
Summary statistics based on haplotypes or LD

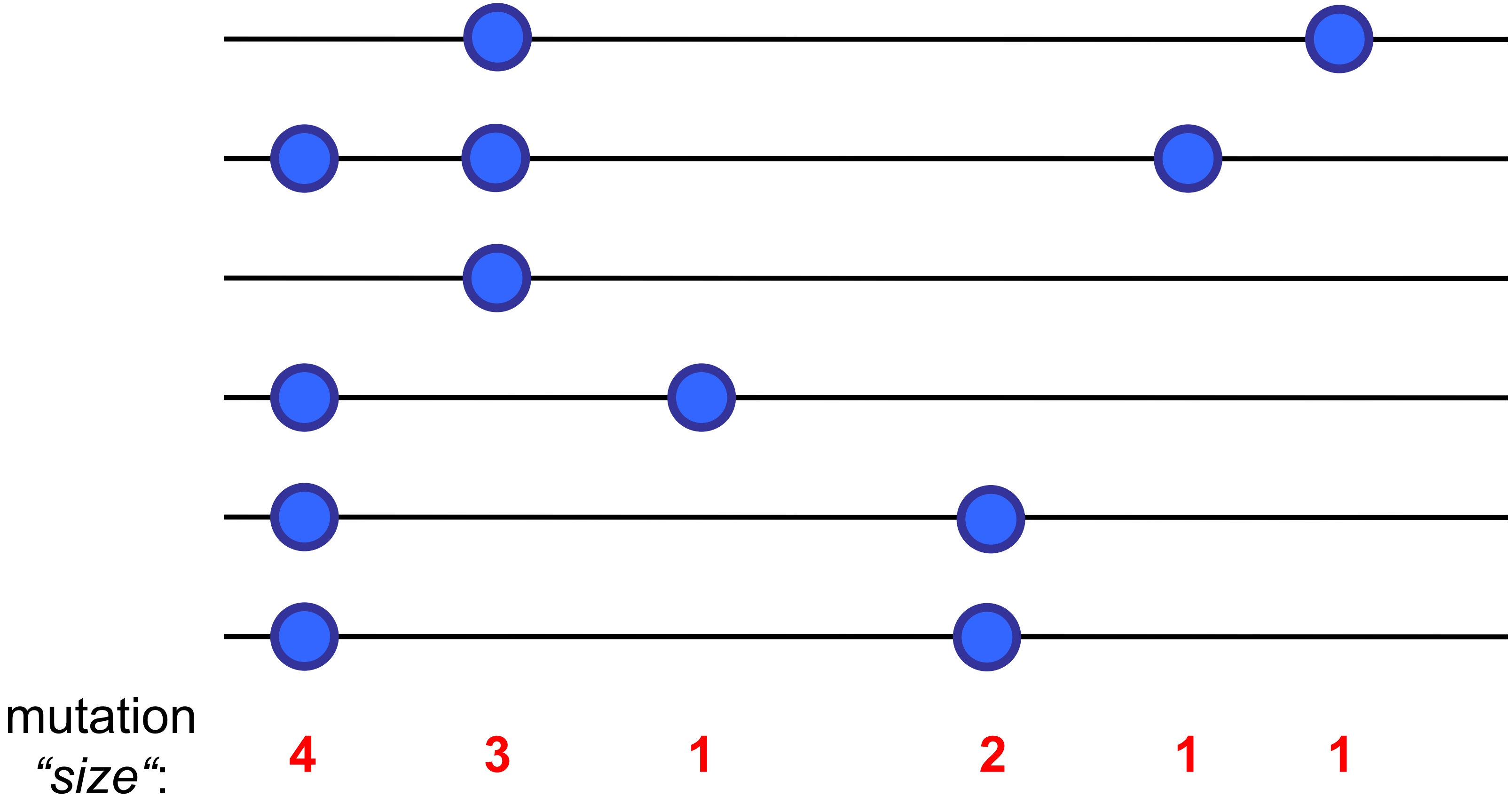
- number or frequency distribution of haplotypes
- or any other measure of linkage disequilibrium (r^2, D, …)



Patterns of Evolution

Summary statistics based on segregating sites

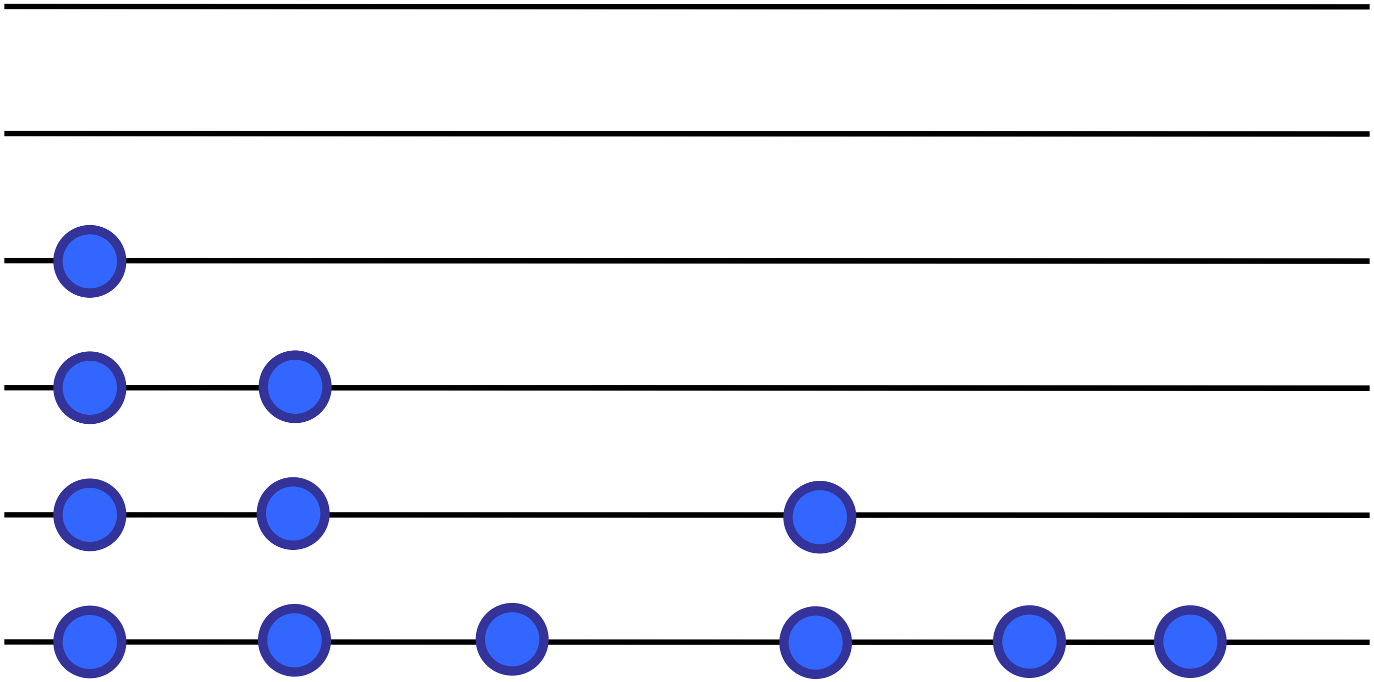
- number of segregating sites and allele frequencies



Patterns of Evolution

Summary statistics based on segregating sites

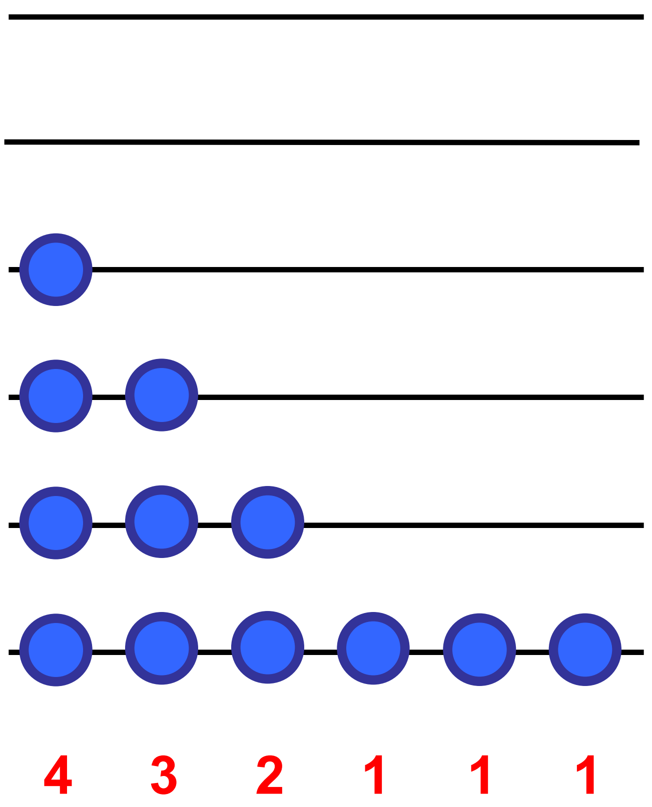
- number of segregating sites and allele frequencies
  - associations not important (“molecular bean bag“)



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“)

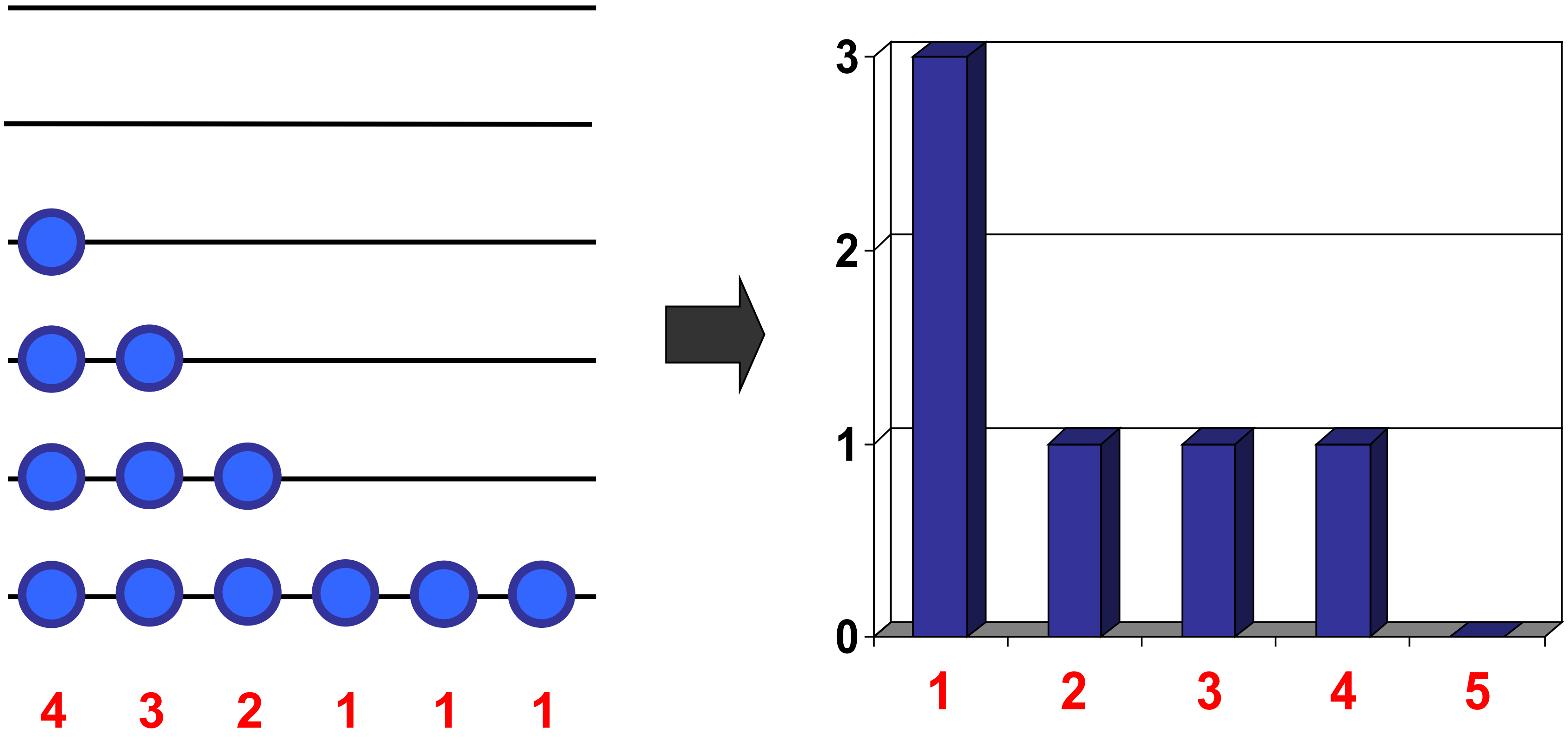


- genome position does not matter

Patterns of Evolution

Summary statistics based on segregating sites

Site Frequency Spectrum



Patterns of Evolution

Summary statistics based on segregating sites

Site Frequency Spectrum

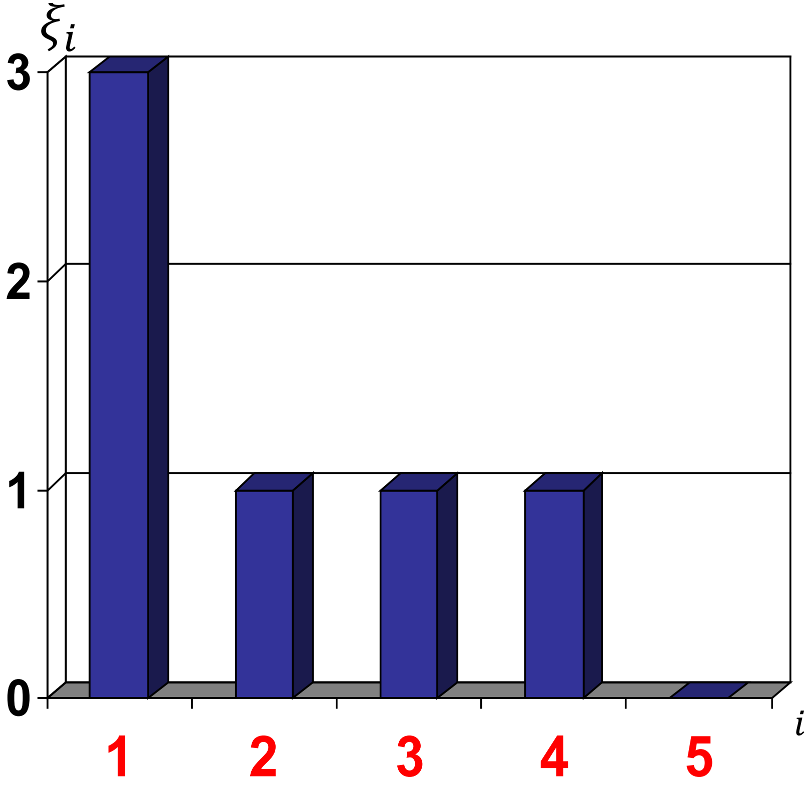
ξ_i : number of mutants that appear in i copies in the sample

$$S = \sum_{i=1}^{n-1} \xi_i :$$

total number of segregating sites in an sample of size n

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i :$$

average number of pairwise differences



Patterns of Evolution

Summary statistics based on segregating sites

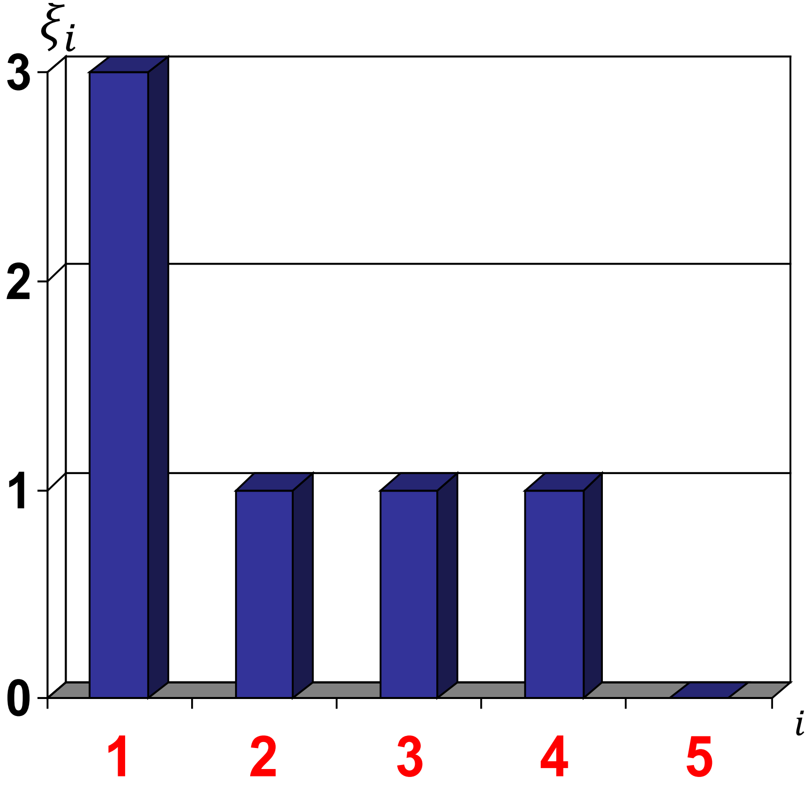
Site Frequency Spectrum

ξ_i : number of mutants that appear in i copies in the sample

$$S = \sum_{i=1}^{n-1} \xi_i :$$ total number of segregating sites in an sample of size n

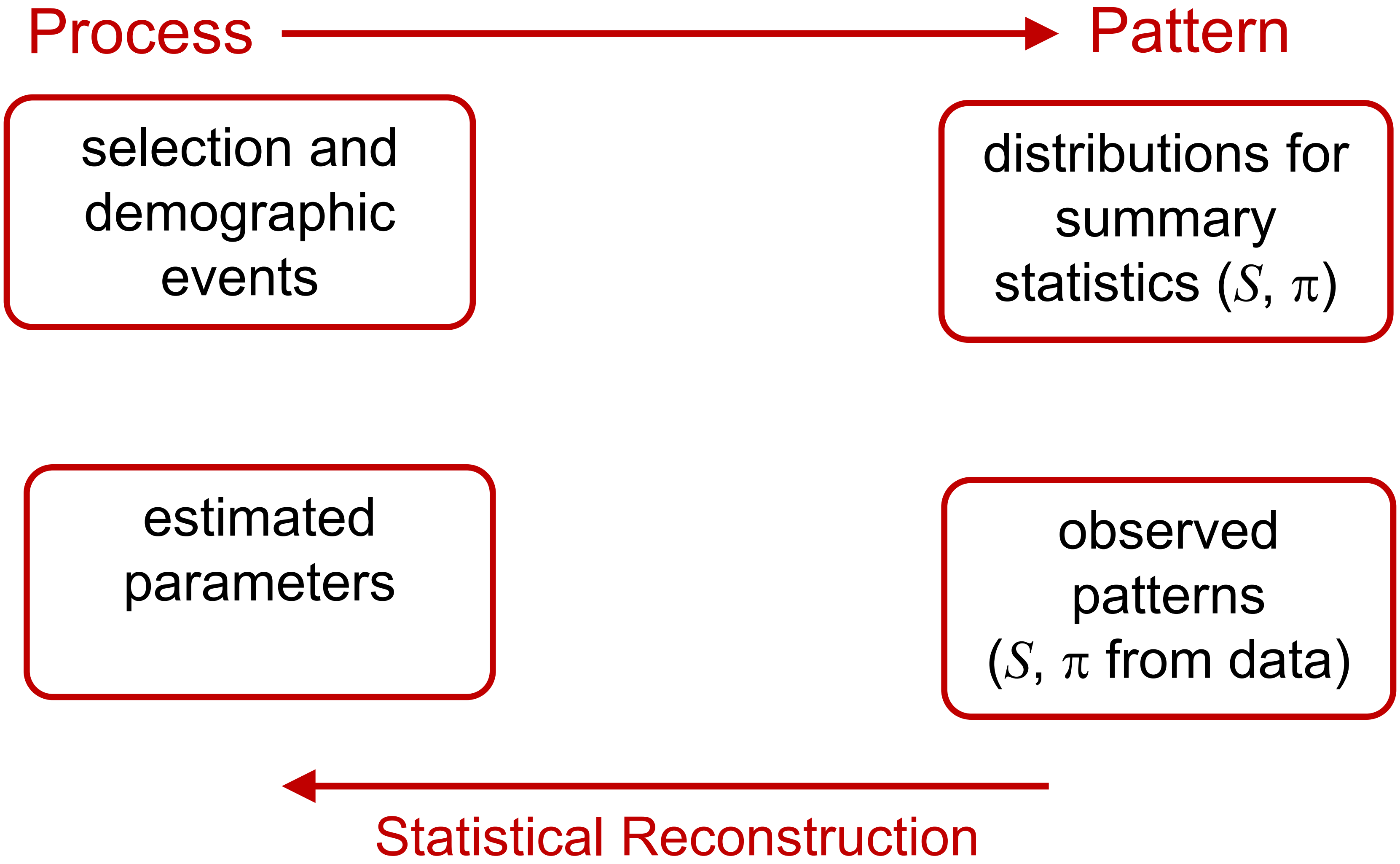
$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i :$$

each mutation of size i contributes to divergence in $i\ (n-i)$ sequence pairs



Patterns of Evolution

Reconstruction of evolutionary history



Process

Pattern

selection and demographic events

distributions for summary statistics (S, π)

estimated parameters

observed patterns (S, π from data)

Statistical Reconstruction

Patterns of Evolution

Reconstruction of evolutionary history

Process → Pattern

standard neutral model

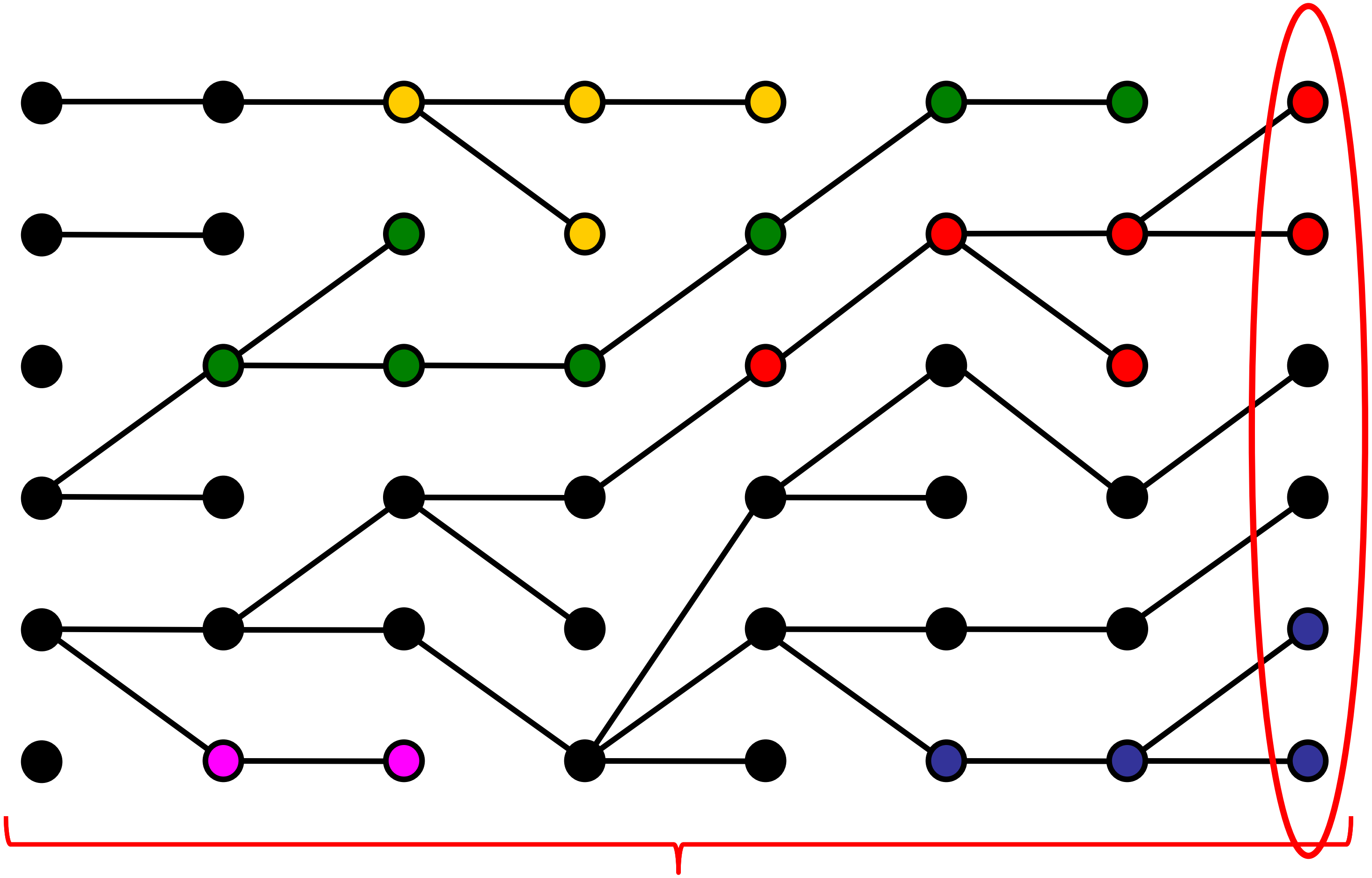
Distributions ?

How does pure randomness look like ?

- ➢ Null-model of the evolutionary theory

Patterns of Evolution

Wright-Fisher model

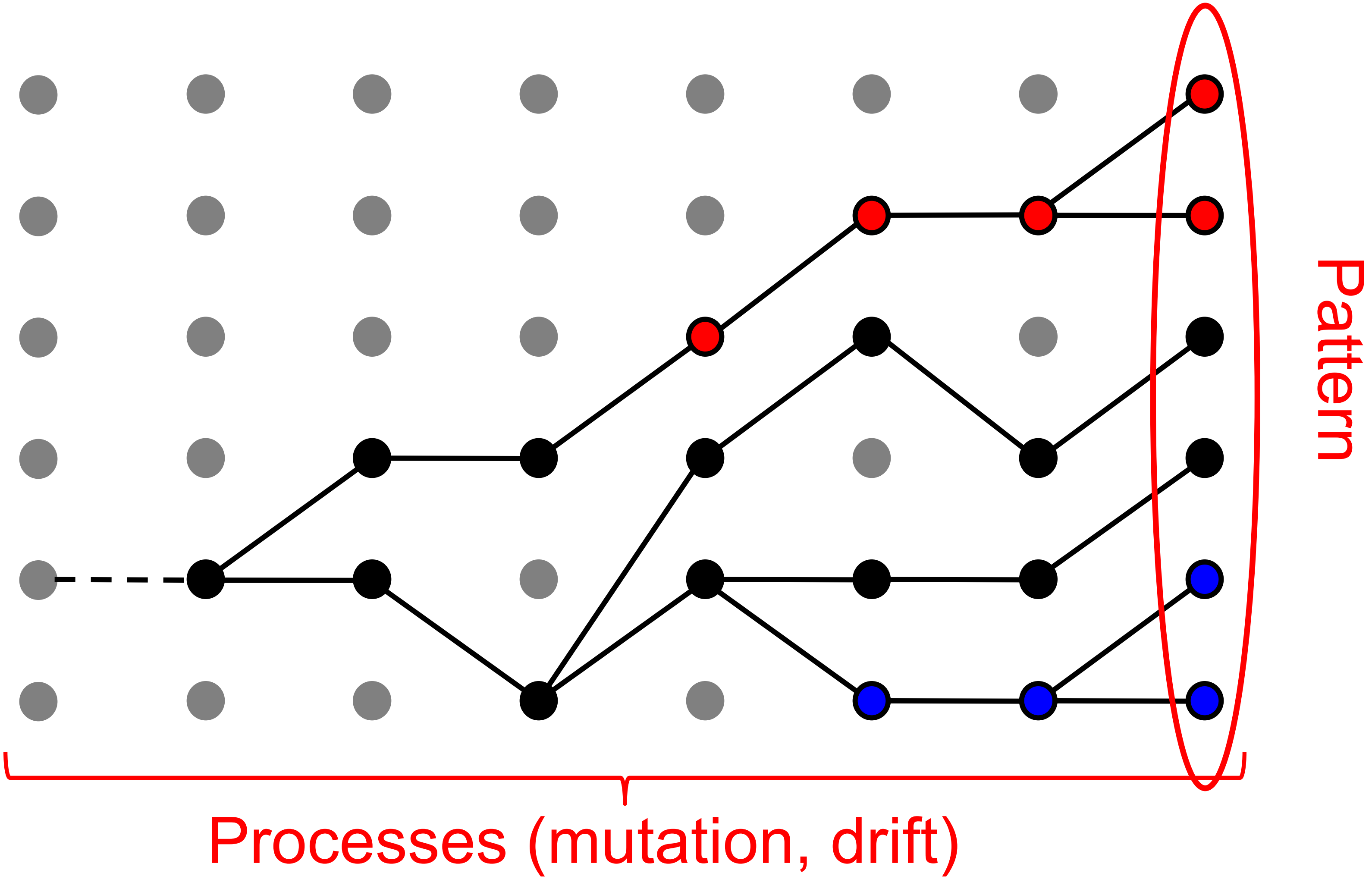


Pattern

Processes (mutation, drift)

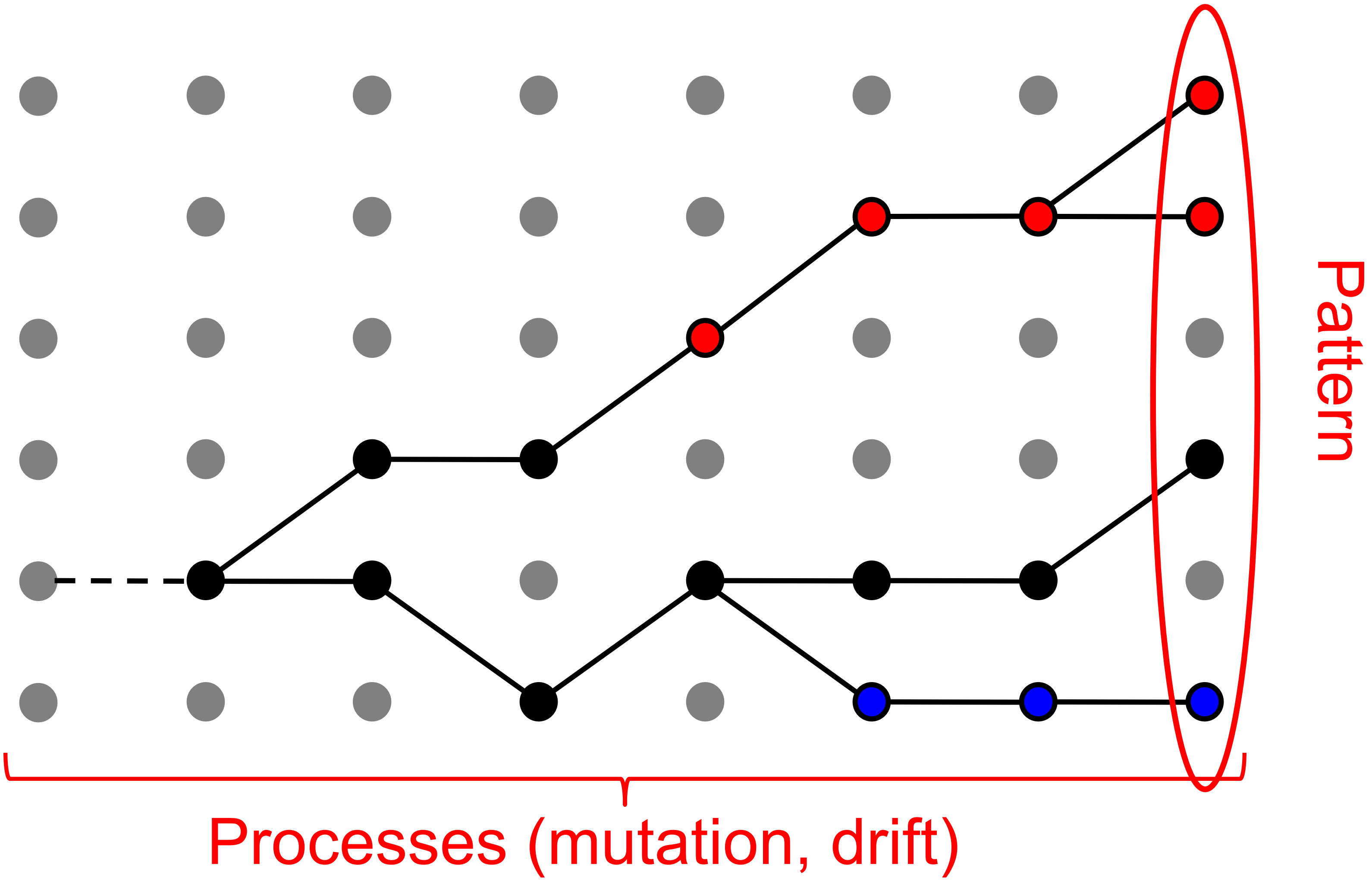
Patterns of Evolution

Wright-Fisher model

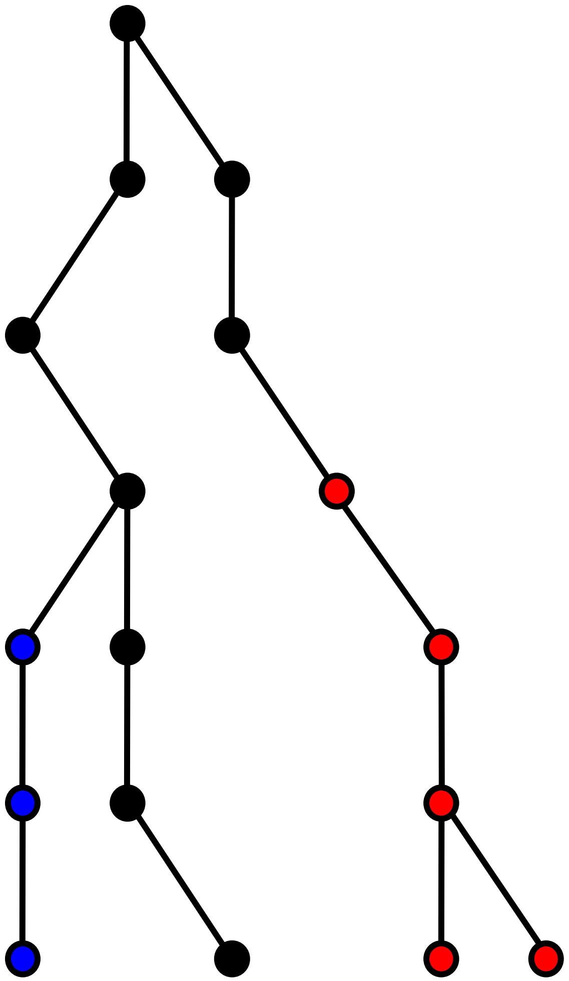


Patterns of Evolution

Wright-Fisher model



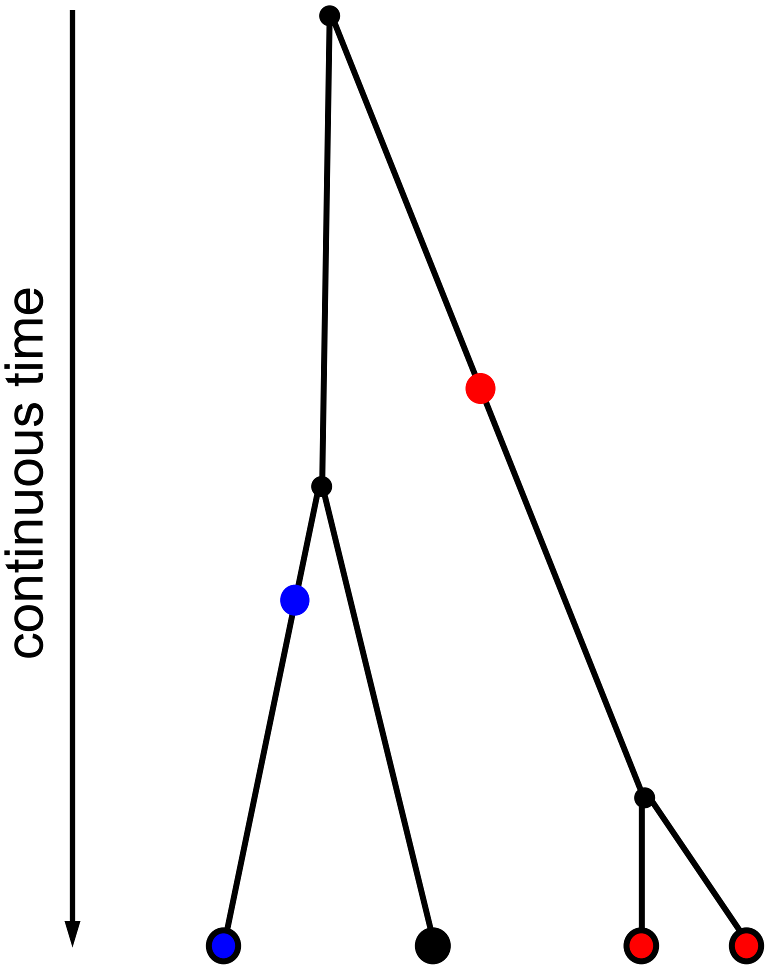
Patterns of Evolution
coalescence process



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* (genotype, location, age, …)

⇩

- Constant population size

➢ Wright-Fisher: multinomial sampling

Individuals are equivalent with respect to descent

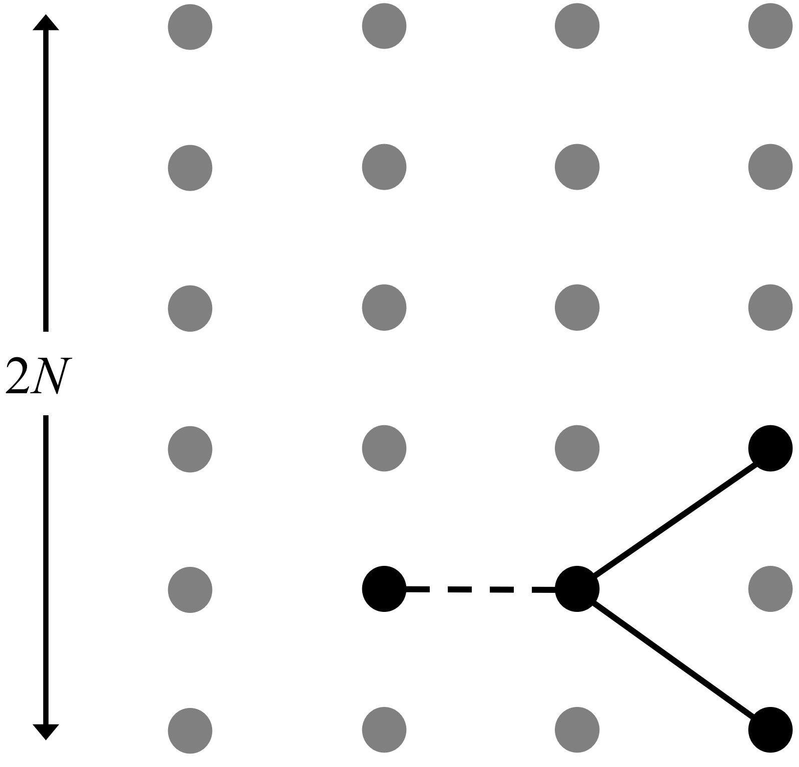
`State´ and `Descent´ are decoupled

⇨ 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



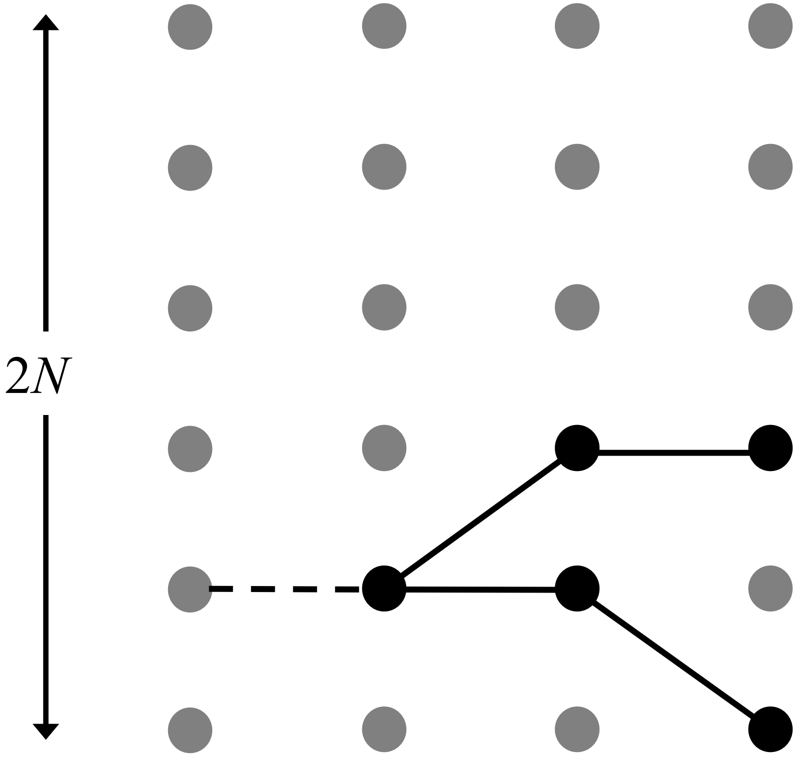
Coalescence probability

… 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:

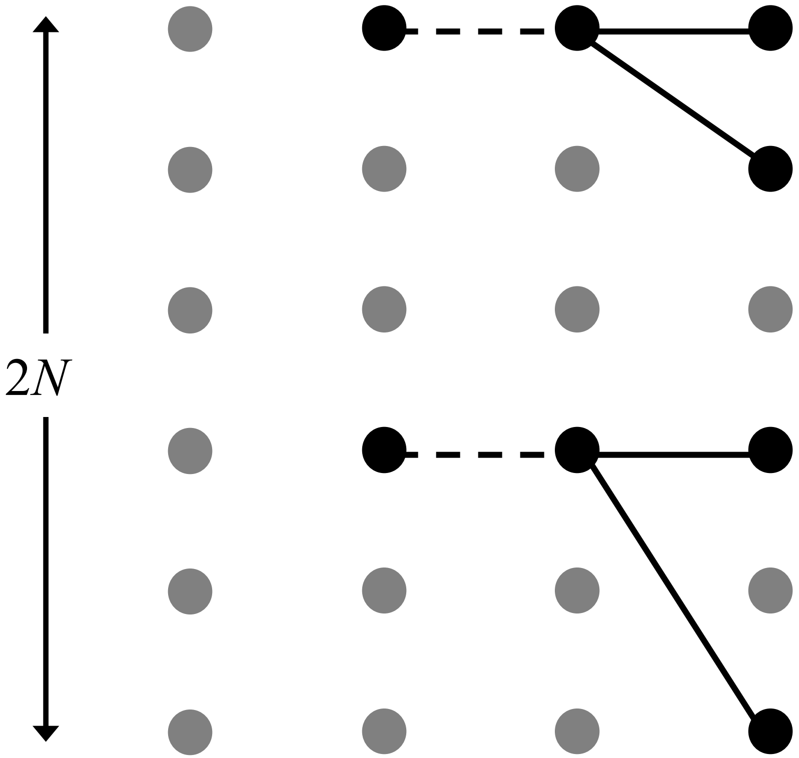
$$p_{c,1} = \frac{1}{2N}$$

… for more than t generations:

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

Coalescent Theory

Construction of the Genealogy: Sample Size n



~~Multiple (e.g. triple) mergers:~~

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

~~Multiple coalescences:~~

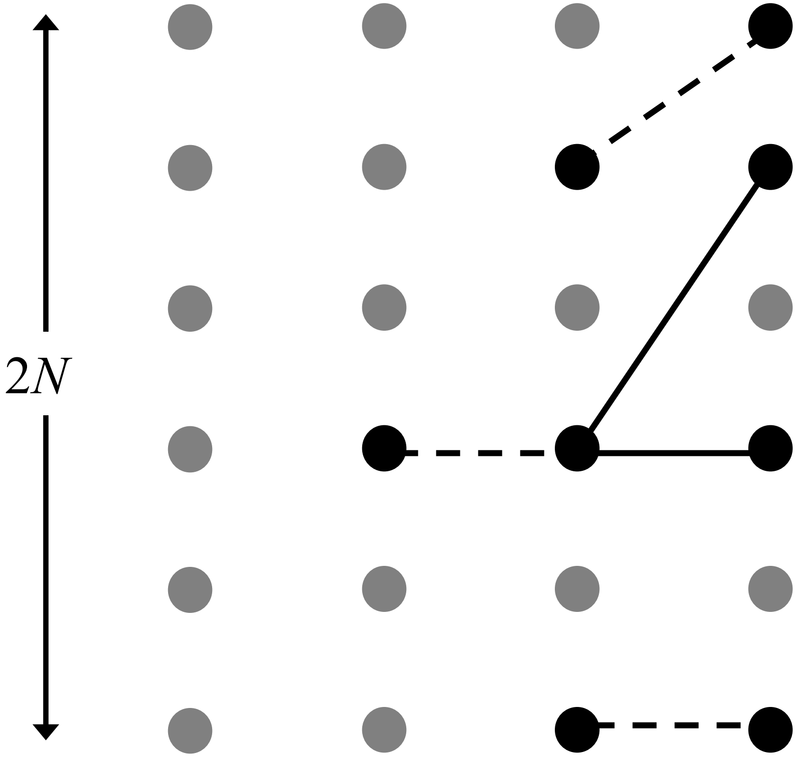
$$\Pr \propto p_{c,t}^2 = O[N^{-2}]$$

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)

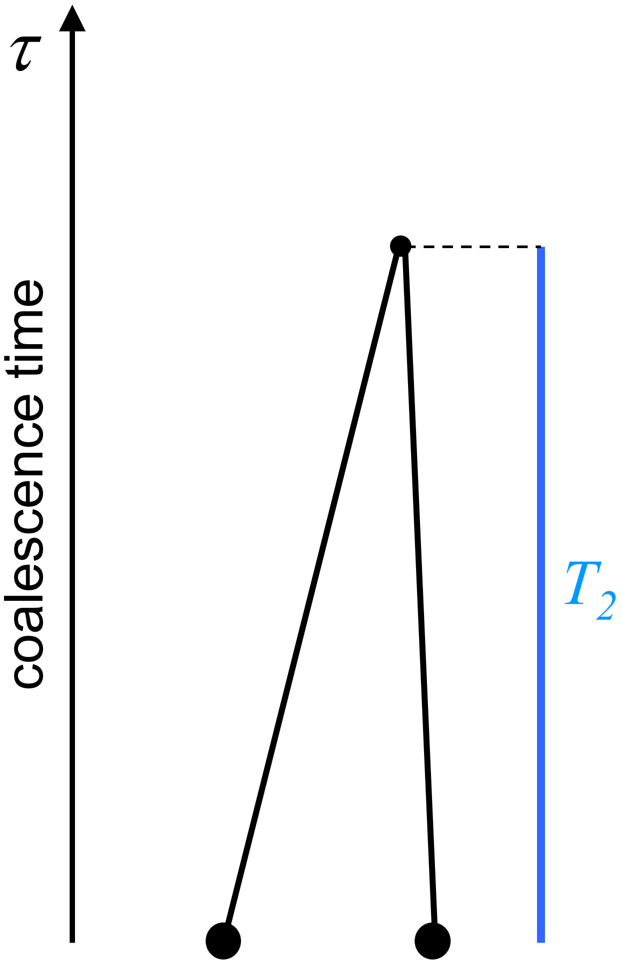
… in a single generation:

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

… for more than t generations:

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

Coalescent Theory
Distribution of Coalescence Times



Define coalescence time scale:

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

Coalescence time T_2 for sample size 2:

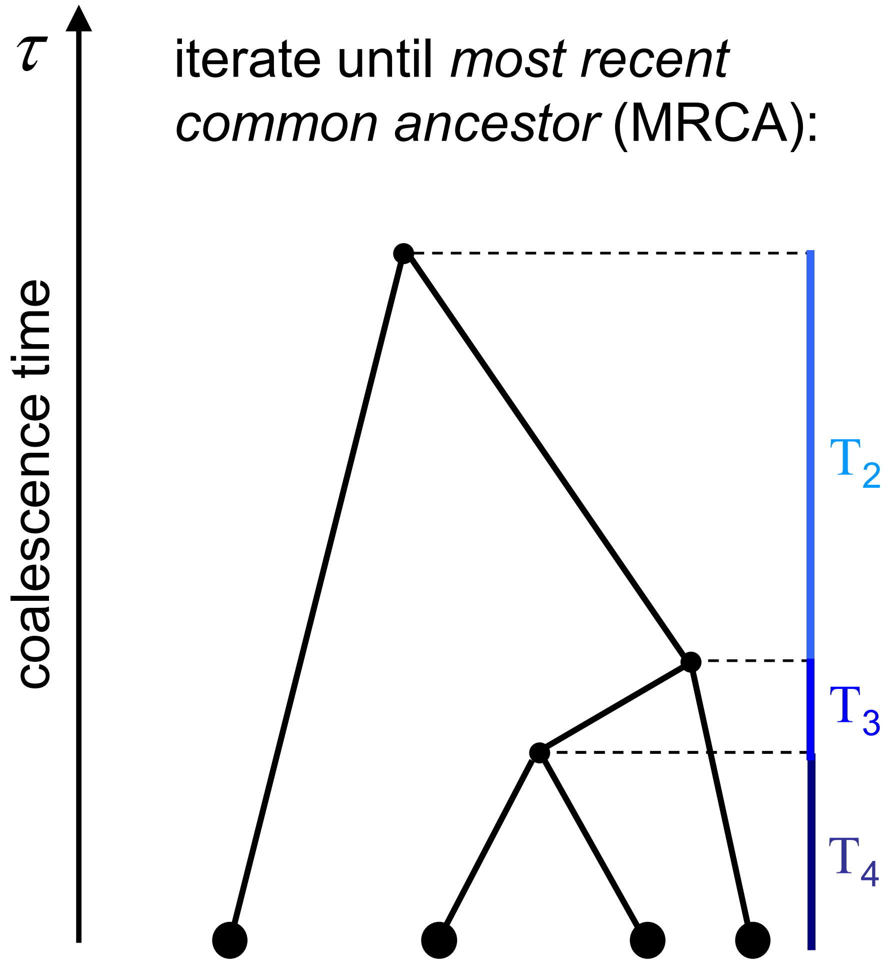
$$\Pr[T_2 > \tau] = \left(1 - \frac{1}{2N}\right)^{2N\tau}$$

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

Exponential distribution with parameter 1:

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

Coalescent Theory
Distribution of Coalescence Times



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

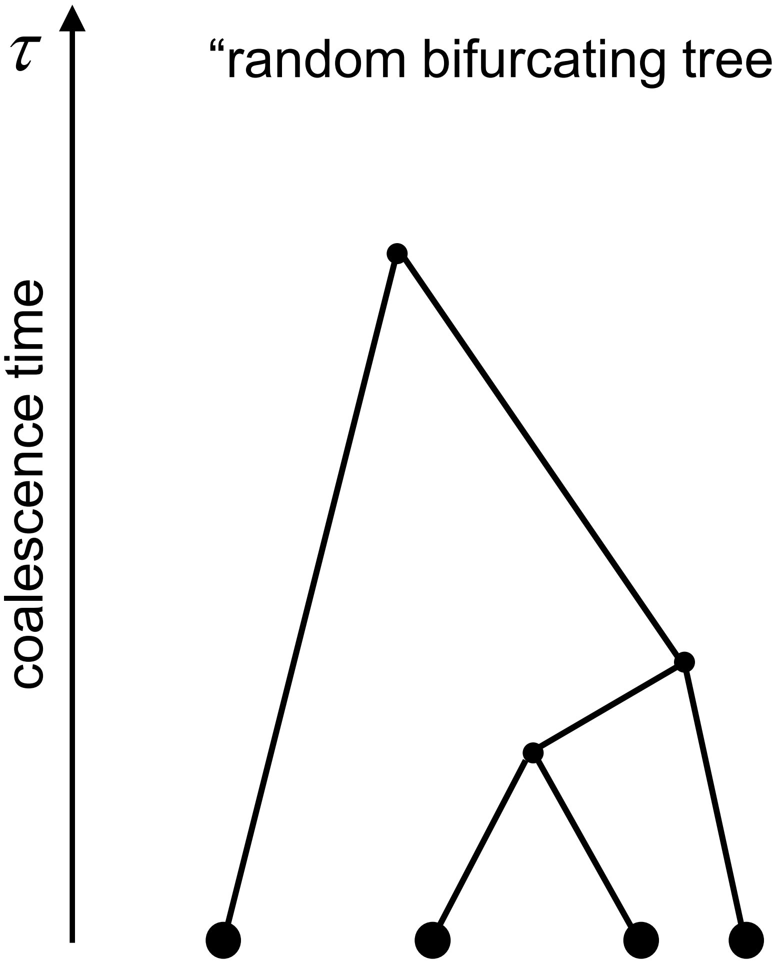
with sample size n:

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

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

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

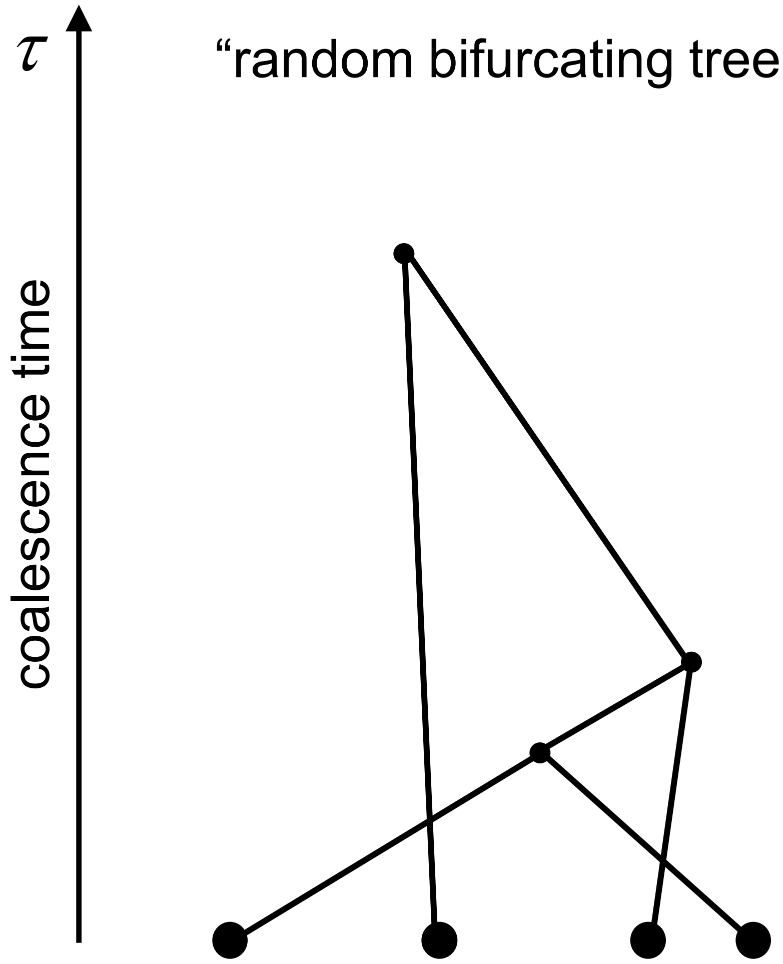
Coalescent Theory
Tree Topologies



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

➢ topology invariant under permutation of "leaves"

Coalescent Theory
Tree Topologies

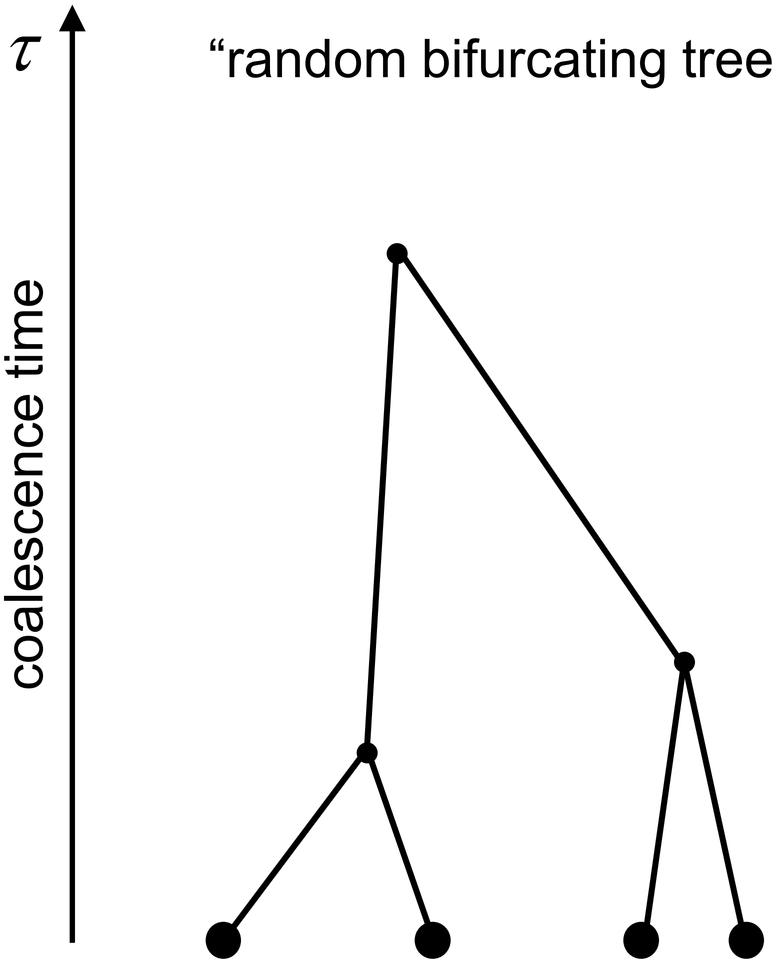


- pick two random individuals from the sample and merge
- sample size $n \to n$-1 and iterate until $n = 1$ (MRCA)
- all individuals exchangable
- ➢ topology invariant under permutation of "leaves"

same topology

Coalescent Theory
Tree Topologies

“random bifurcating tree“

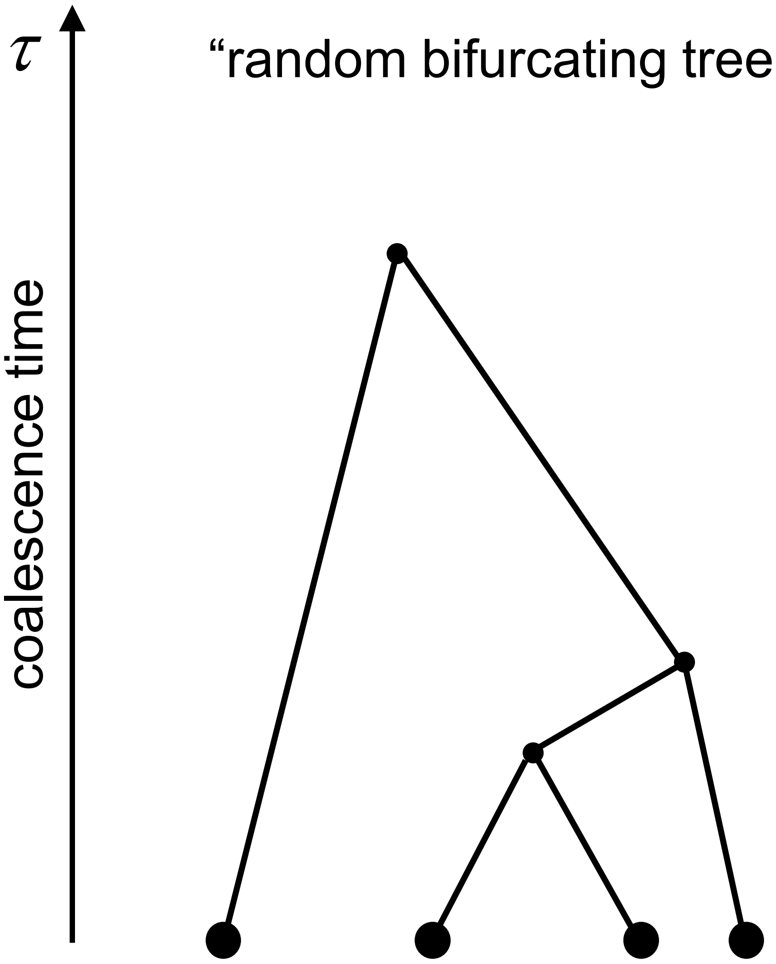


- pick two random individuals from the sample and merge
- sample size $n \rightarrow n$-1 and iterate until $n = 1$ (MRCA)
- all individuals exchangable

➢ 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 $n \rightarrow n$-1 and iterate until $n = 1$ (MRCA)
- all individuals exchangable

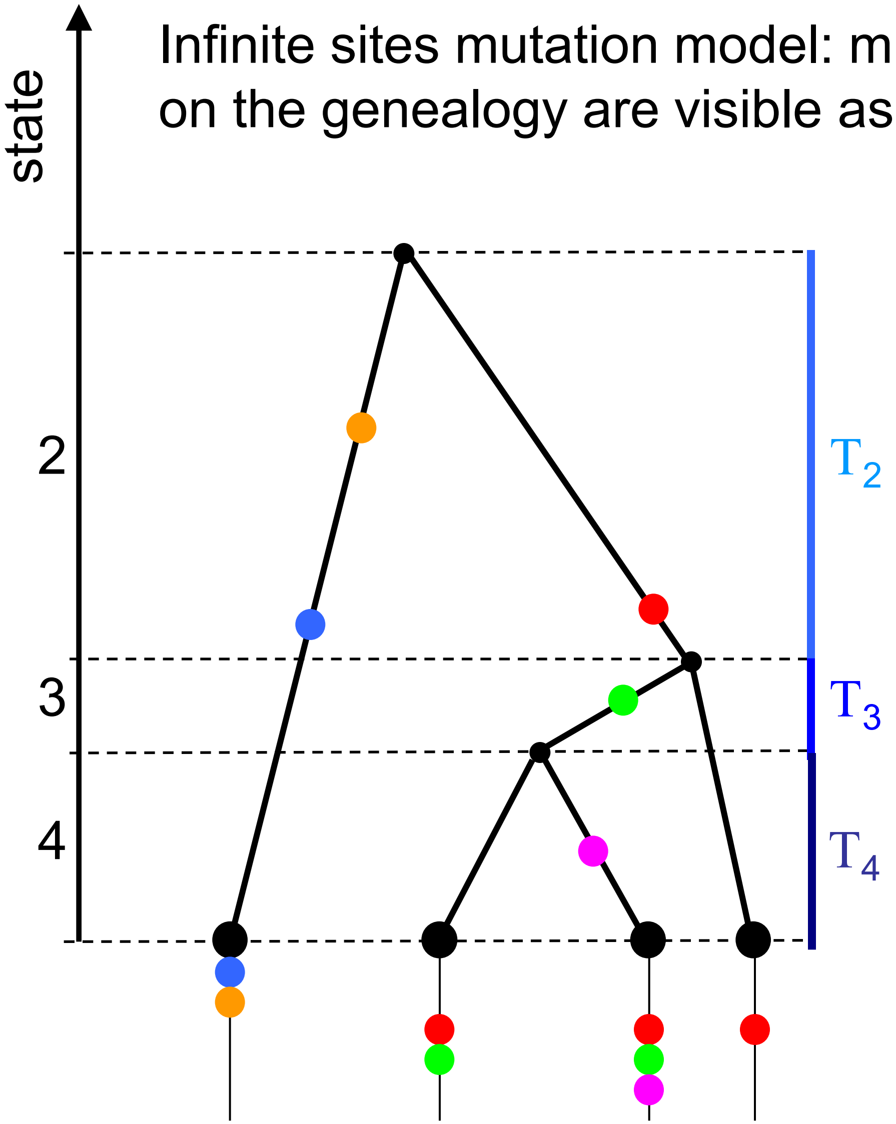
➢ topology invariant under permutation of “leaves“

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 $u \cdot 2NL = \frac{\theta \cdot L}{2}$, $L = \sum_{i=j}^{k} T_i$ branch length of branch from state j through k

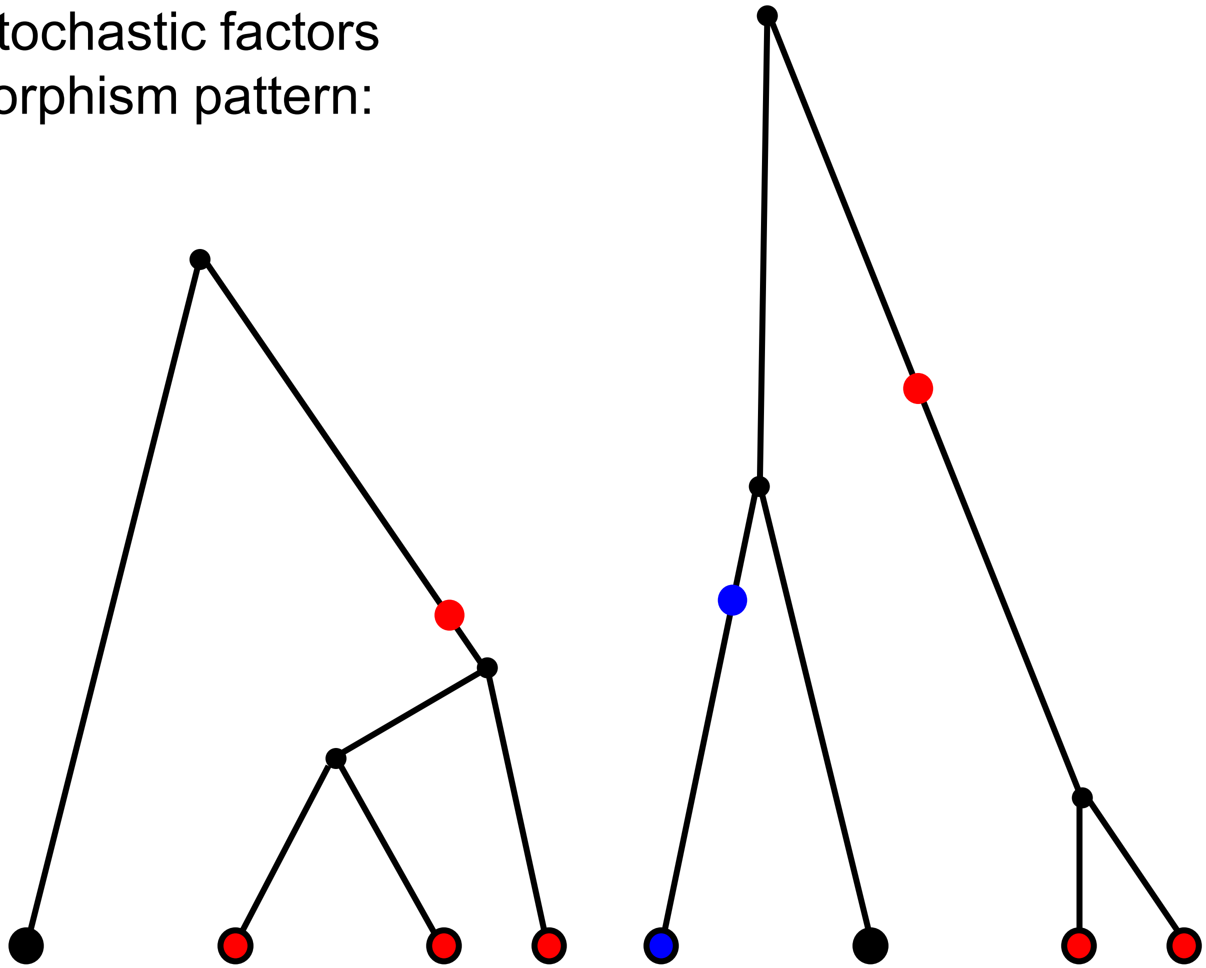
(also other mutation schemes possible)

Coalescent Theory
Basic Properties

Three independent stochastic factors determine the polymorphism pattern:

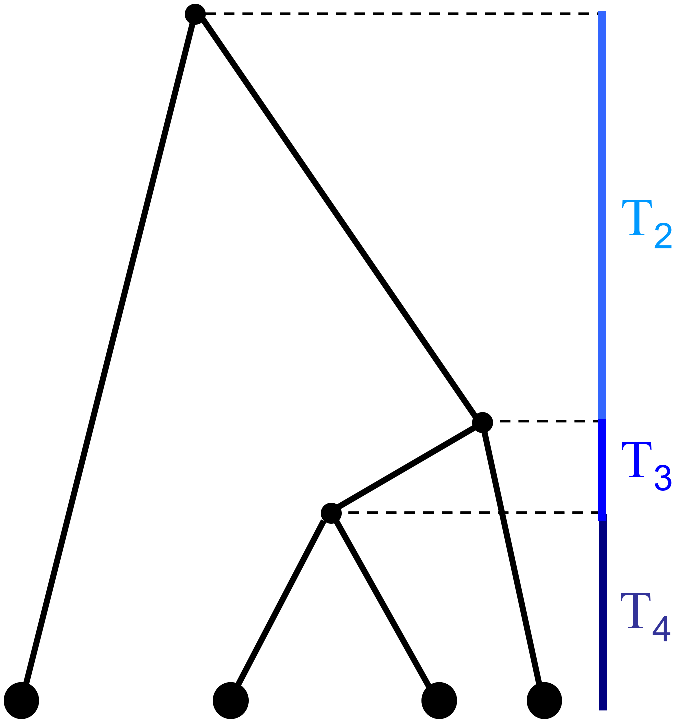
1. coalescent times
2. tree topology
3. mutation

(very easy to implement in simulations)



Coalescent Theory
Basic Properties

Time to the most recent common ancestor:



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

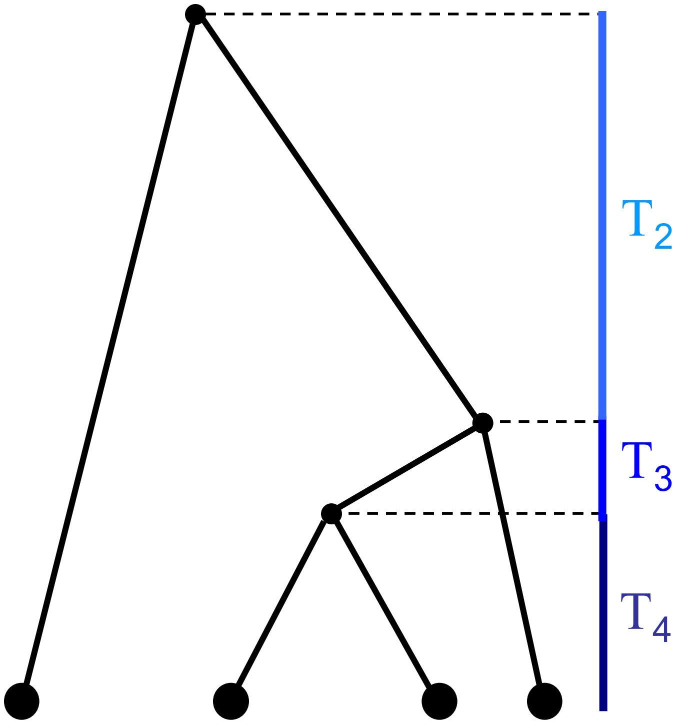
$$= 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right)$$

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:



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

$$\Rightarrow \cdots \Rightarrow E[S] = 2Nu \cdot 2 \sum_{k=1}^{n-1} \frac{1}{k} = \theta \cdot a_n$$

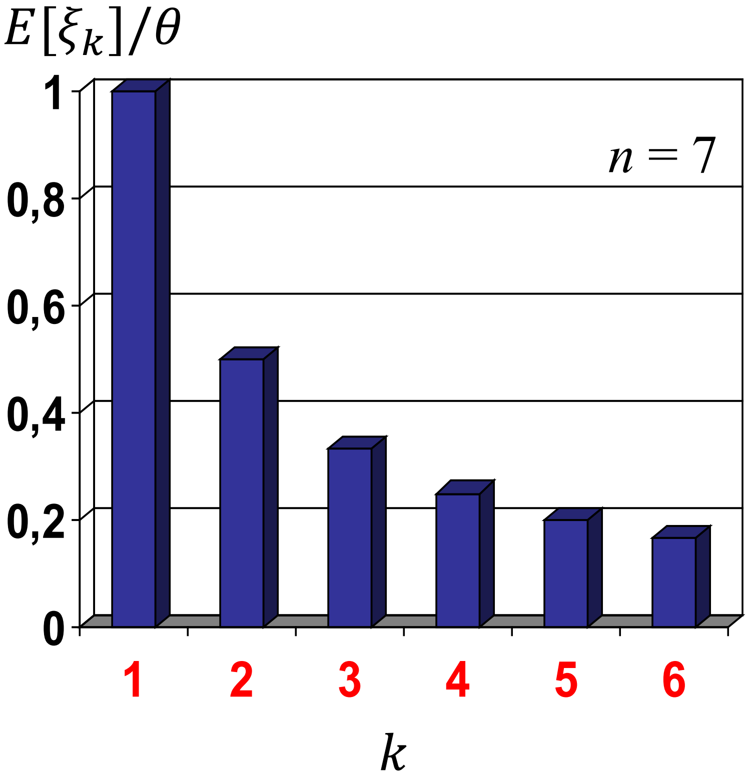
with: $a_n = \sum_{k=1}^{n-1} \frac{1}{k} \quad \rightarrow \log n + 0.577$

(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)



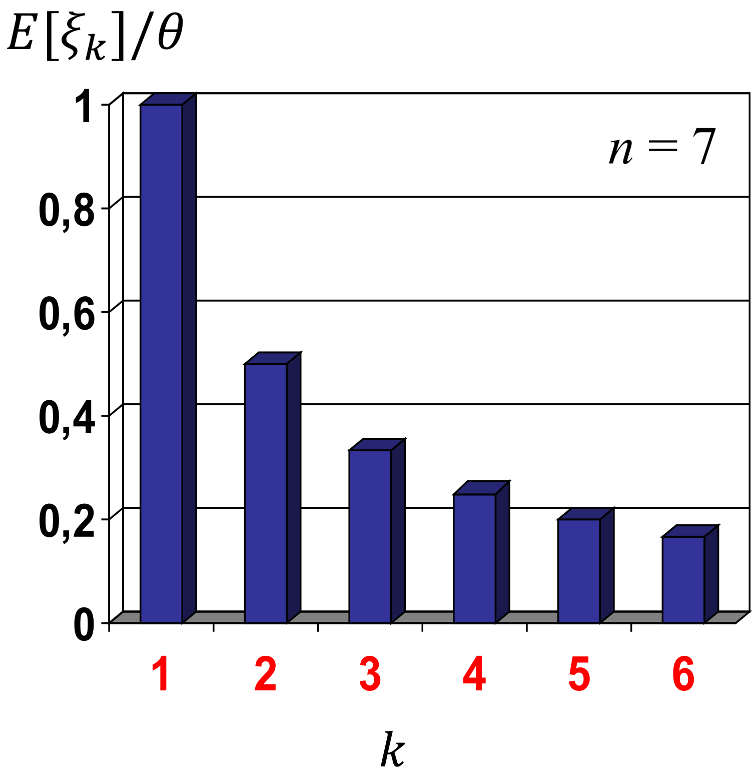
$$E[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \sum_{k=1}^{n-1} E[\xi_k]$$

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

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

Coalescent Theory
Estimators

Expected site frequency spectrum under standard neutrality:



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

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

- depends on $\theta = 4N_e u$ as only model parameter

➢ How can we estimate θ?

Coalescent Theory
Estimators

Unbiased estimators of the mutation parameter $\theta = 4Nu$:

Watterson's estimator: $\hat{\theta}_W = \frac{S}{a_n} = \sum_{k=1}^{n-1} \xi_i \Big/ \sum_{k=1}^{n-1} \frac{1}{k}$ (equal weights)

π-based estimator: $\hat{\theta}_\pi = \pi = \binom{n}{2}^{-1} \sum_{k=1}^{n-1} k(n-k)\ \xi_k$ (intermediate frequencies)

Fay and Wu's estimator: $\hat{\theta}_H = \binom{n}{2}^{-1} \sum_{k=1}^{n-1} k^2\, \xi_k$ (high frequencies)

singleton estimator: $\hat{\theta}_s = \frac{n-1}{n} \underbrace{(\xi_1 + \xi_{n-1})}_{\text{singletons of the folded spectrum}}$ (extreme frequencies)