The Neutral Model

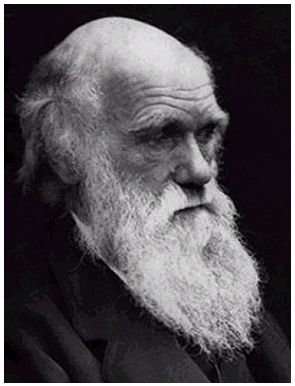
Evolution without selection

Joachim Hermisson

Mathematics and Biosciences Group

Mathematics & MFPL, University of Vienna

Evolutionary processes



Darwin

selection theory →

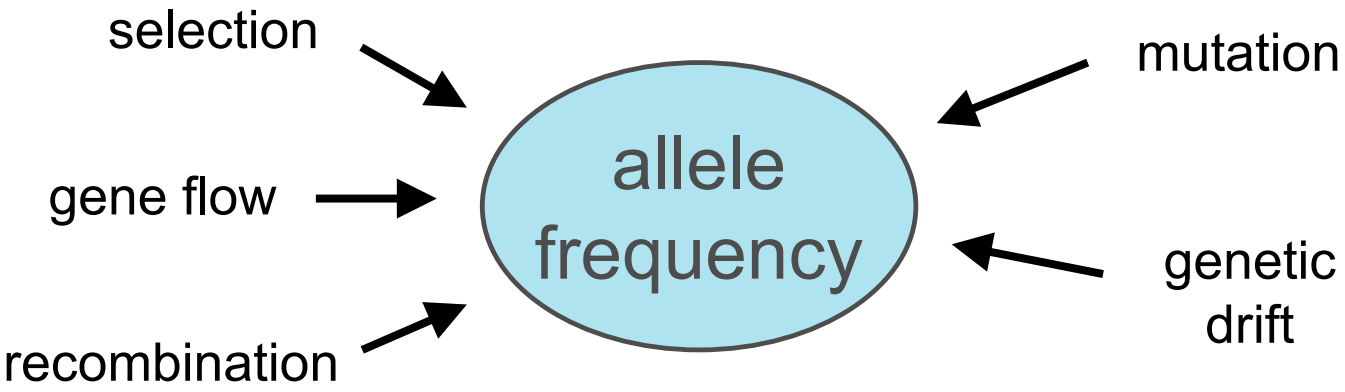
1920 – 1930 „modern synthesis“

← laws of inheritance

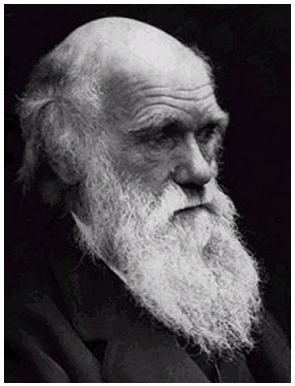
Mendel

Population Genetics

describes change of allele frequencies
under the elementary evolutionary processes:



Evolutionary processes



Darwin

selection theory →

1920 – 1930 „modern synthesis“

← laws of inheritance

Mendel

Population Genetics

describes change of allele frequencies under the elementary evolutionary processes:

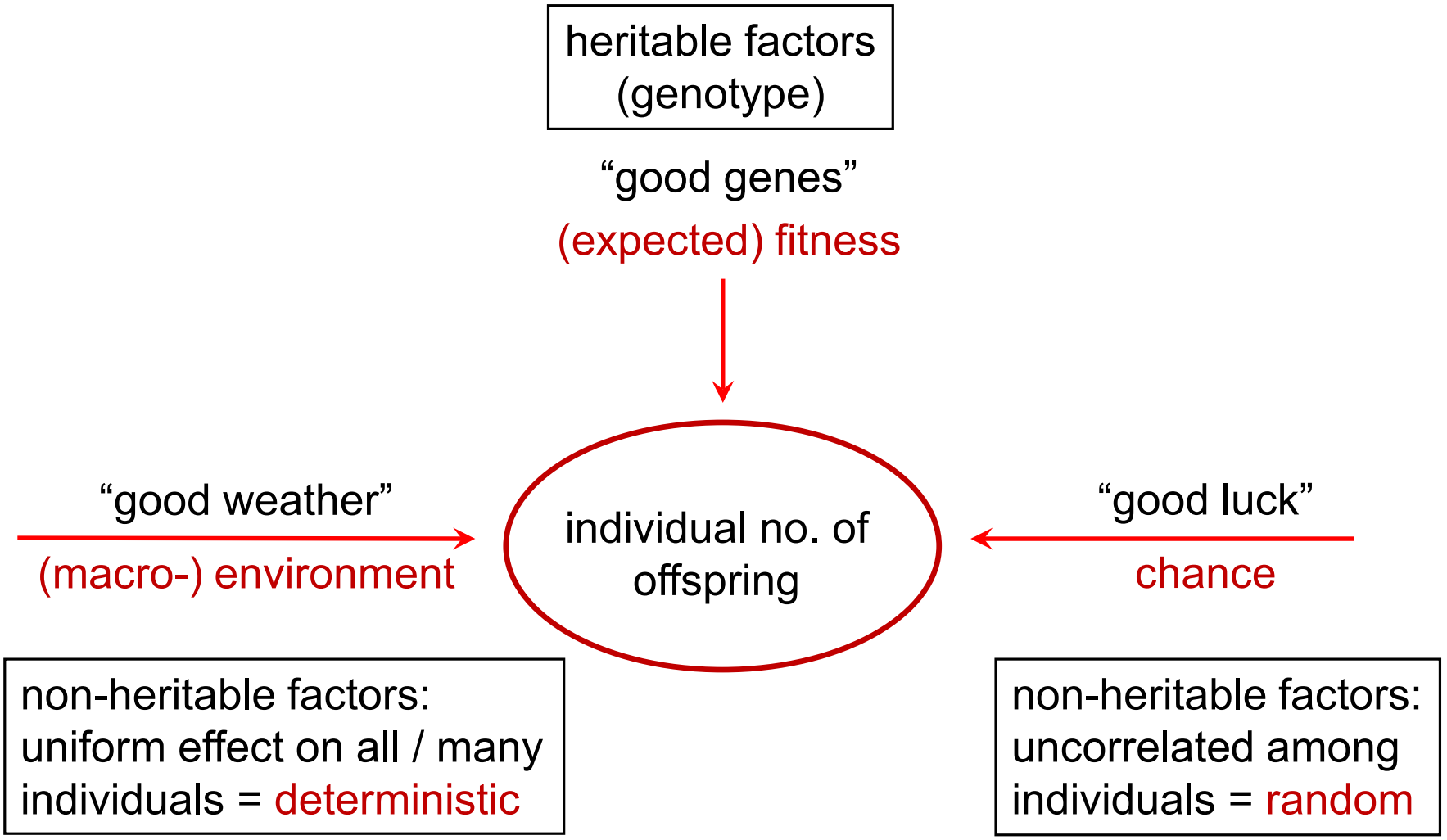
allele frequency

← mutation

← genetic drift

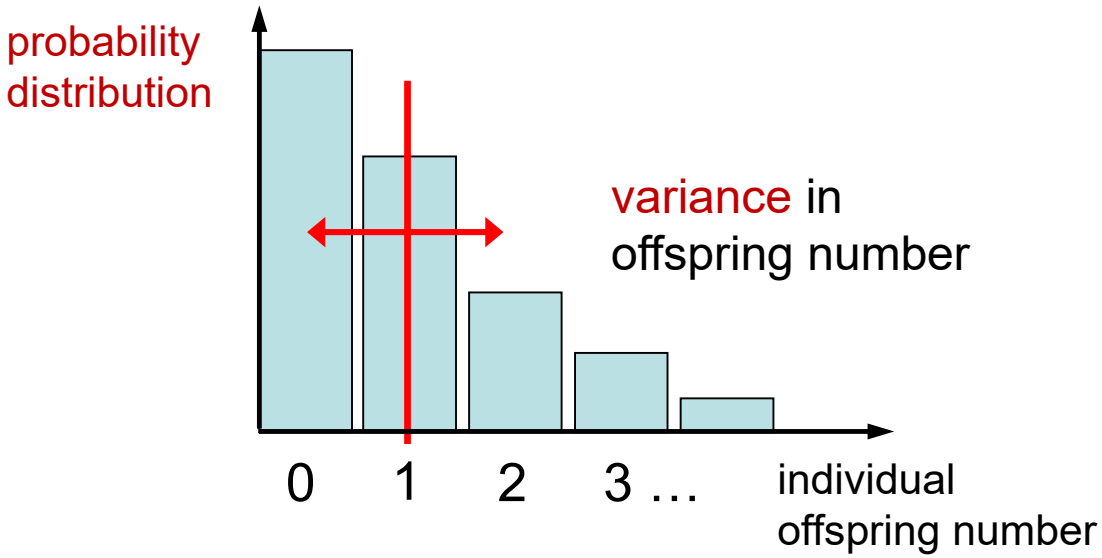
“neutral theory“

What is random genetic drift?



What is random genetic drift?

How do we define drift ?



mean offspring number
(function of genotype, environment):
= 1 for neutrality & constant population size

What is random genetic drift?

How do we define drift ?

Change in allele counts due to variation in offspring number that is

- independent of external factors
- independent of genotypes (heritable factors)

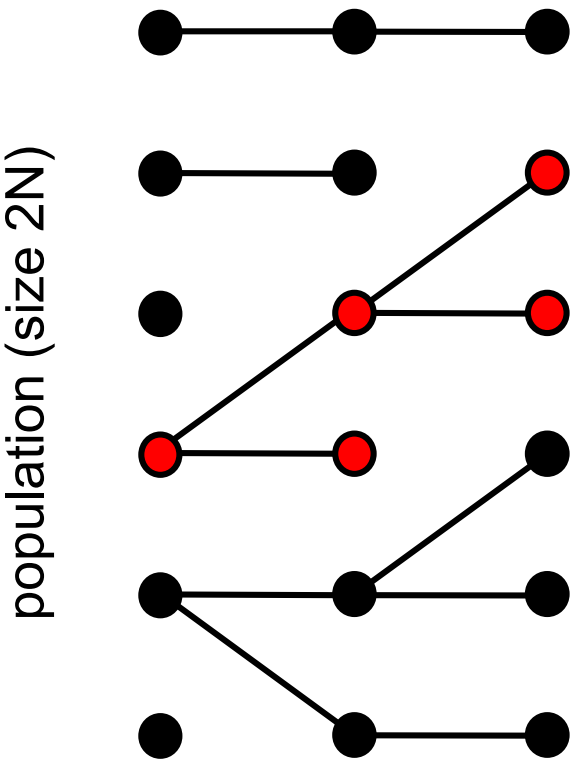
} *random*

Effects of drift in a finite population:

- allele frequencies in finite populations are not fixed, but undergo random changes from one generation to the next
- allele frequencies in sub-population diverge after population split
- alleles may get lost from a population or reach fixation

How to quantify this? – need a model

The Wright-Fisher model:
Drift as binomial sampling

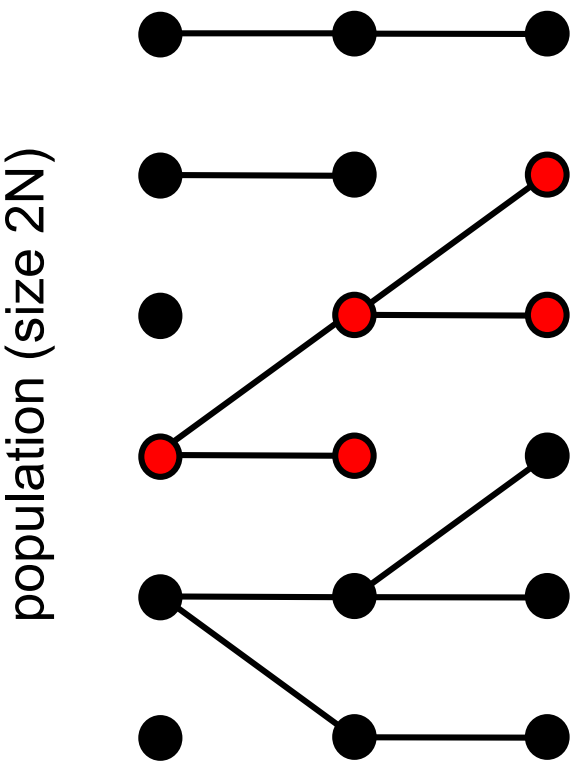


- Discrete time, non-overlapping generations
- Single haploid locus with two alleles
- Constant haploid population size 2N
- Alleles in the next generation are sampled from an *infinite gamete pool* of the current generation

➢ "Offspring randomly choose a parent" and inherit his/her genotype

➢ Sampling with replacement $\Longrightarrow$?

The Wright-Fisher model:
Drift as binomial sampling



Binomial distribution

single parent: probability for k offspring

(# trials $n = 2N$; success prob. $p = 1/2N$)

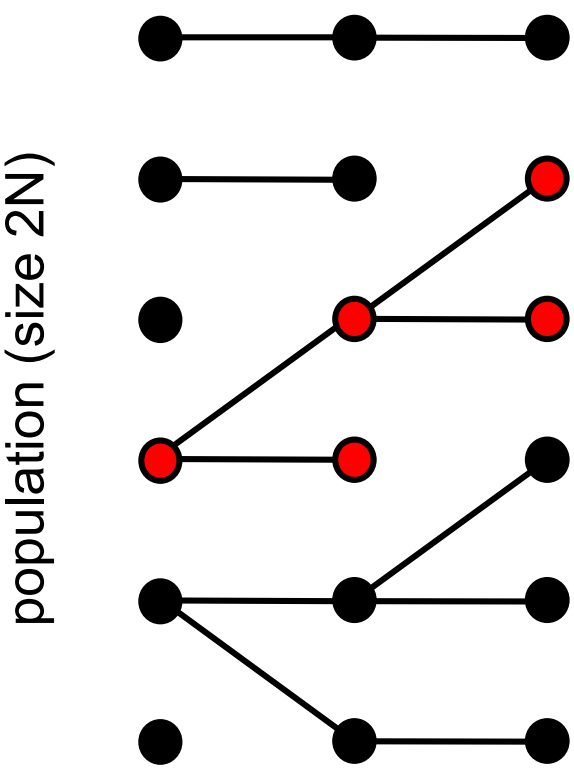
$$\Pr[k] = \binom{2N}{k}\left(\frac{1}{2N}\right)^{k}\left(1-\frac{1}{2N}\right)^{2N-k}$$

mean: $E[k] = n \cdot p = 1$

variance: $\sigma_k^2 = n \cdot p(1-p)$

$= 1 - (1/2N) \approx 1$

The Wright-Fisher model:
Drift as binomial sampling



Binomial distribution

population: allele frequency $p_t = \frac{i}{2N}$

probability for $p_{t+1} = \frac{j}{2N}$ in next generation:

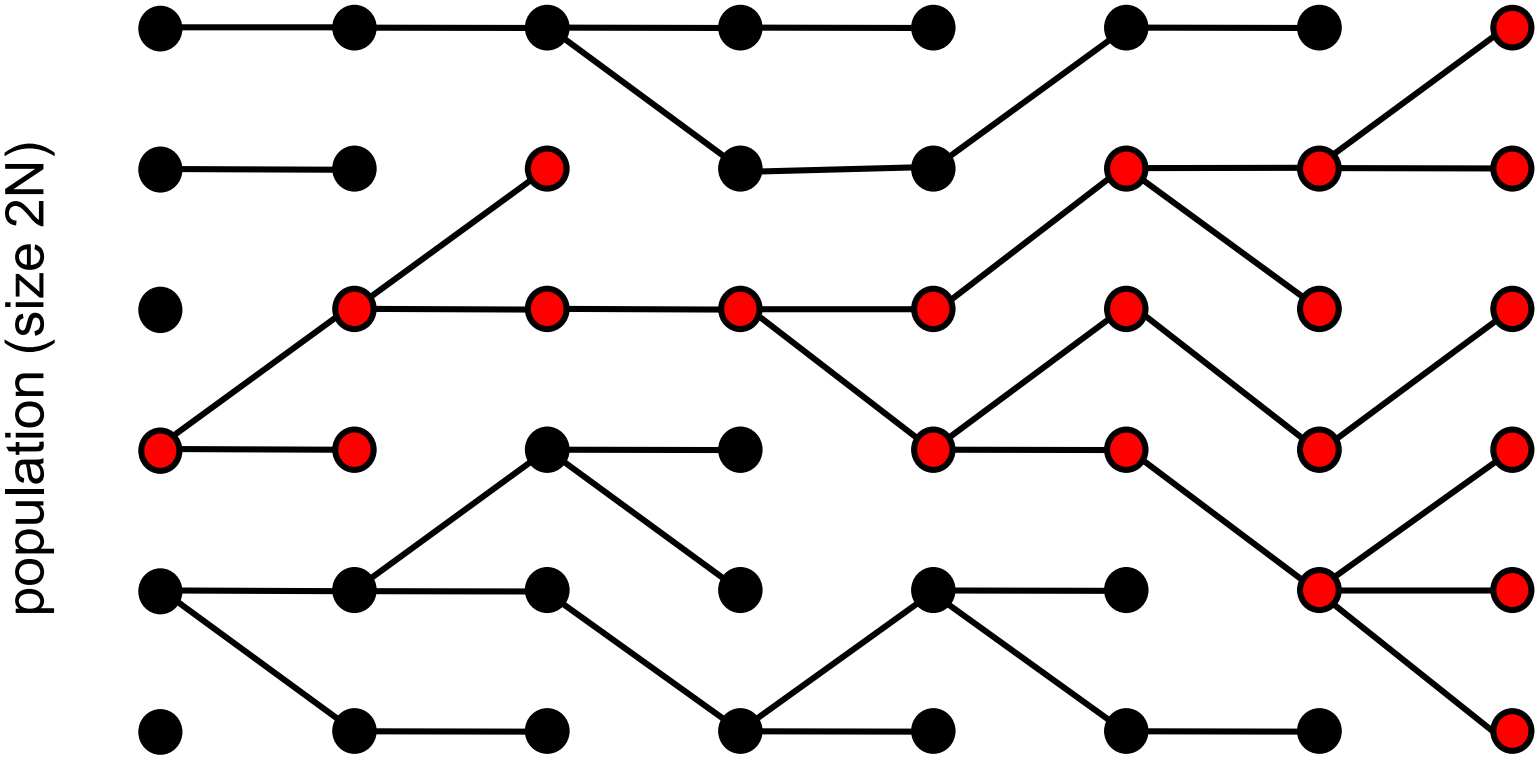
$$P_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{1}{2N}\right)^{2N-j}$$

mean: $E[p_{t+1}] = p_t$

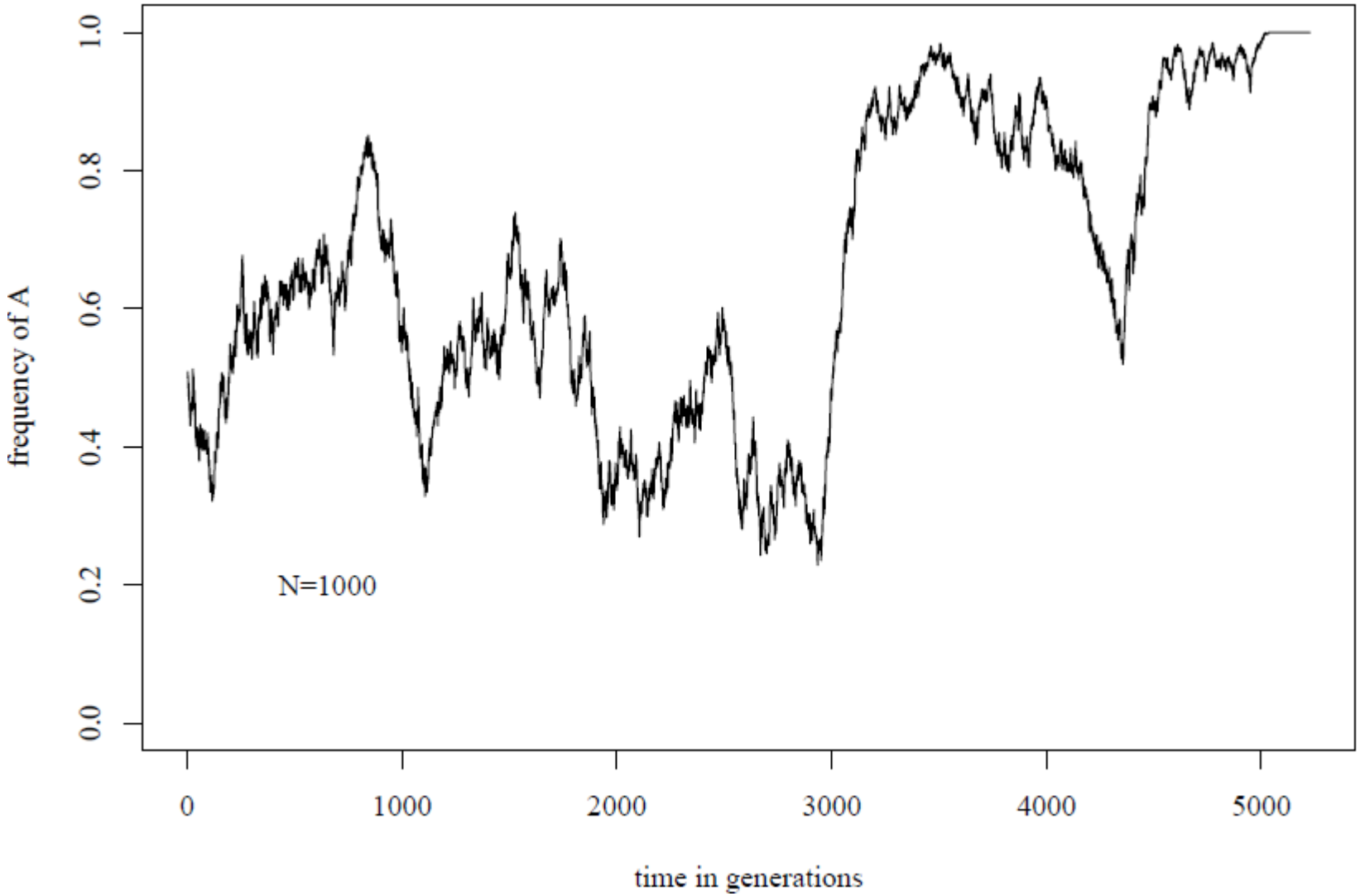
variance: $\mathrm{Var}[p_{t+1}] = \frac{p_t(1-p_t)}{2N}$

- drift stronger in small populations

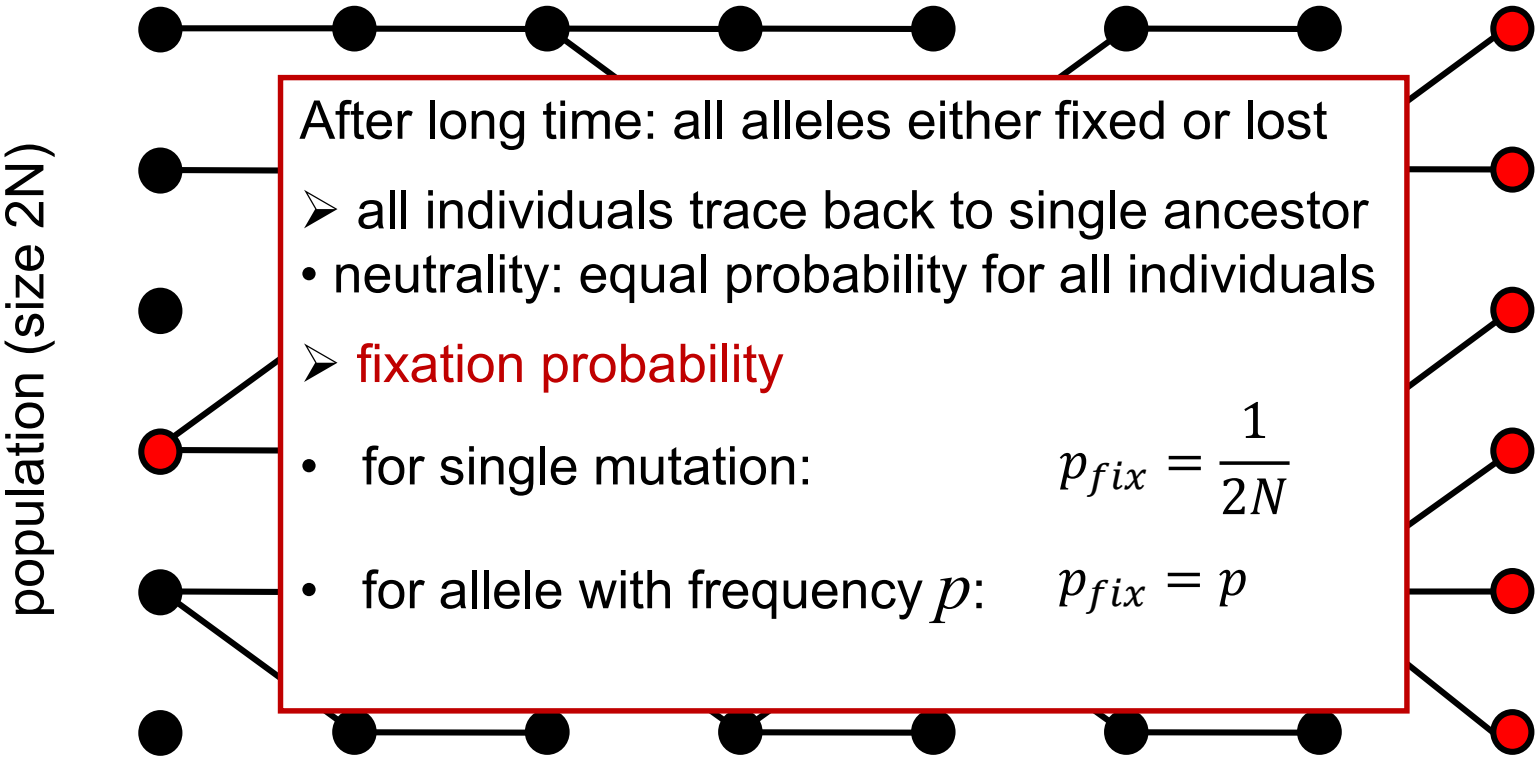
The Wright-Fisher model:
Fixation of neutral alleles



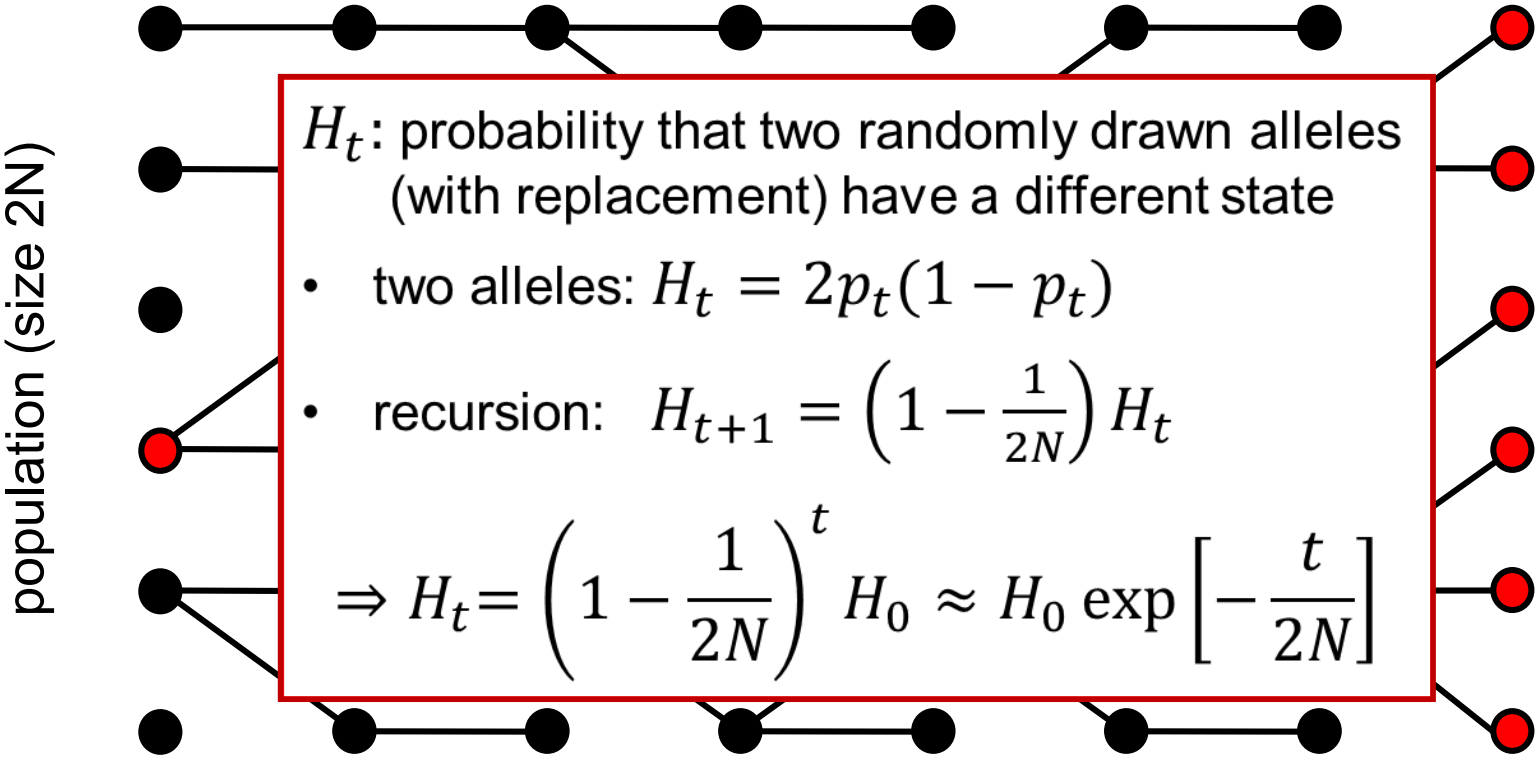
The Wright-Fisher model:
Fixation of neutral alleles



The Wright-Fisher model:
Fixation of neutral alleles



The Wright-Fisher model:
Heterozygosity



The neutral theory
Drift and Mutation

Can drift (plus mutation) explain observed patterns of diversity ?

Motoo Kimura 1960's: Neutral theory of molecular evolution

The observed natural diversity at the molecular level is largely the result of neutral evolution: patterns of genetic variation and diversity are explained by drift and mutation

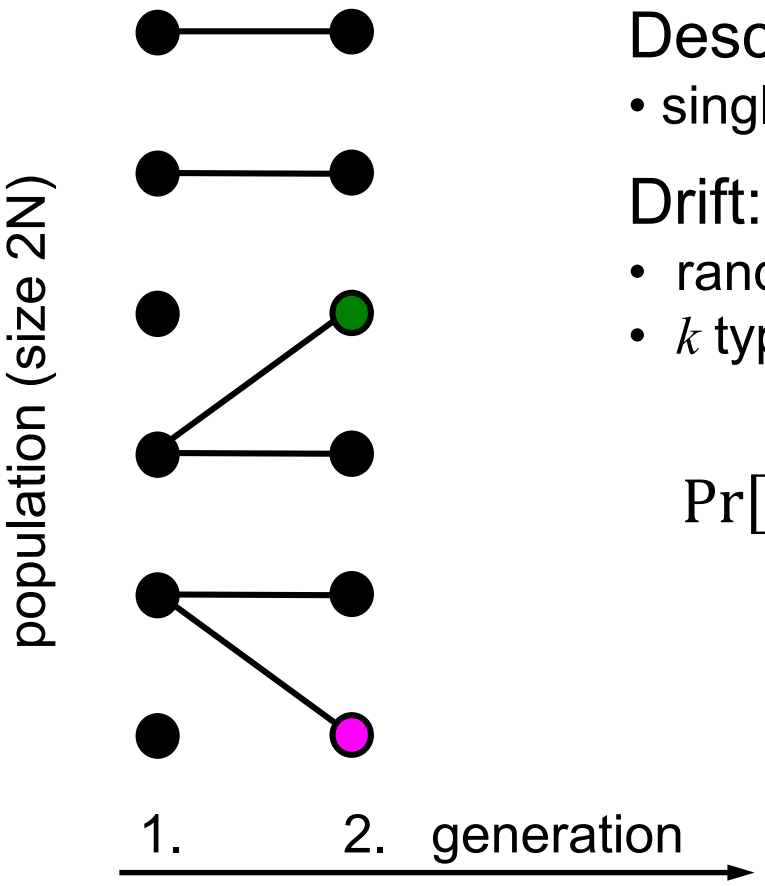
What neutral theory does not say:

"Selection is not important for evolution"

- Purifying selection responsible for conserved DNA (eg in genes)
- Adaptive evolution due to positive selection
- But: adaptive diversity tiny relative to neutral diversity

Which patterns do we expect? – need a model

The neutral theory
Drift and Mutation

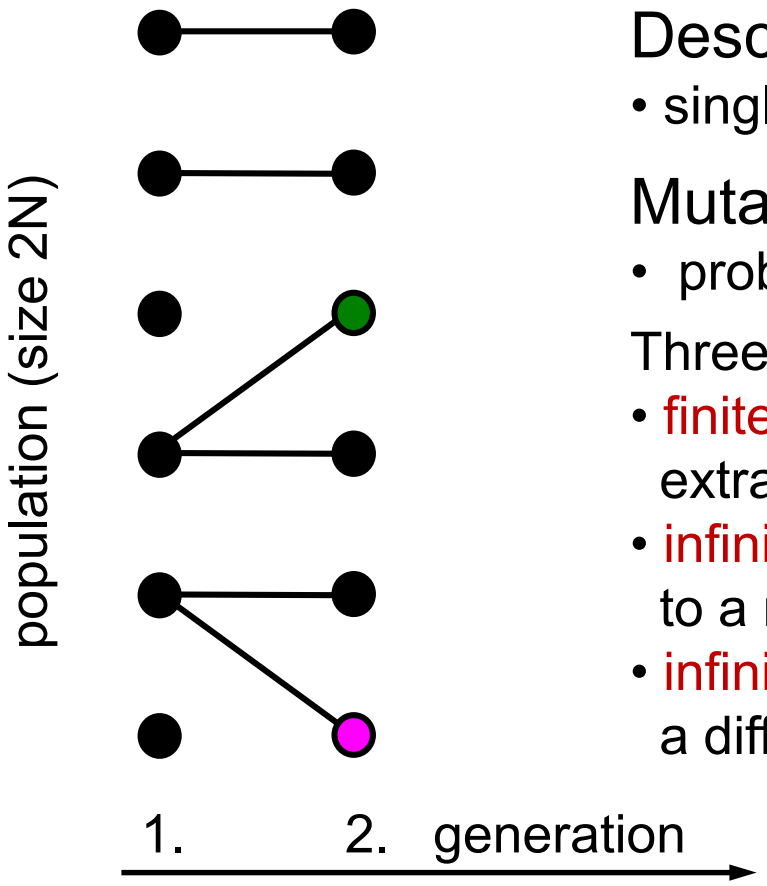


Description of neutral genetic variation
- single locus, multiple alleles

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

$$\Pr[p_1', \ldots, p_k'] = \frac{N!}{\prod_i (p_i' N)!} \prod_i {p_i}^{p_i' N}$$

The neutral theory
Drift and Mutation

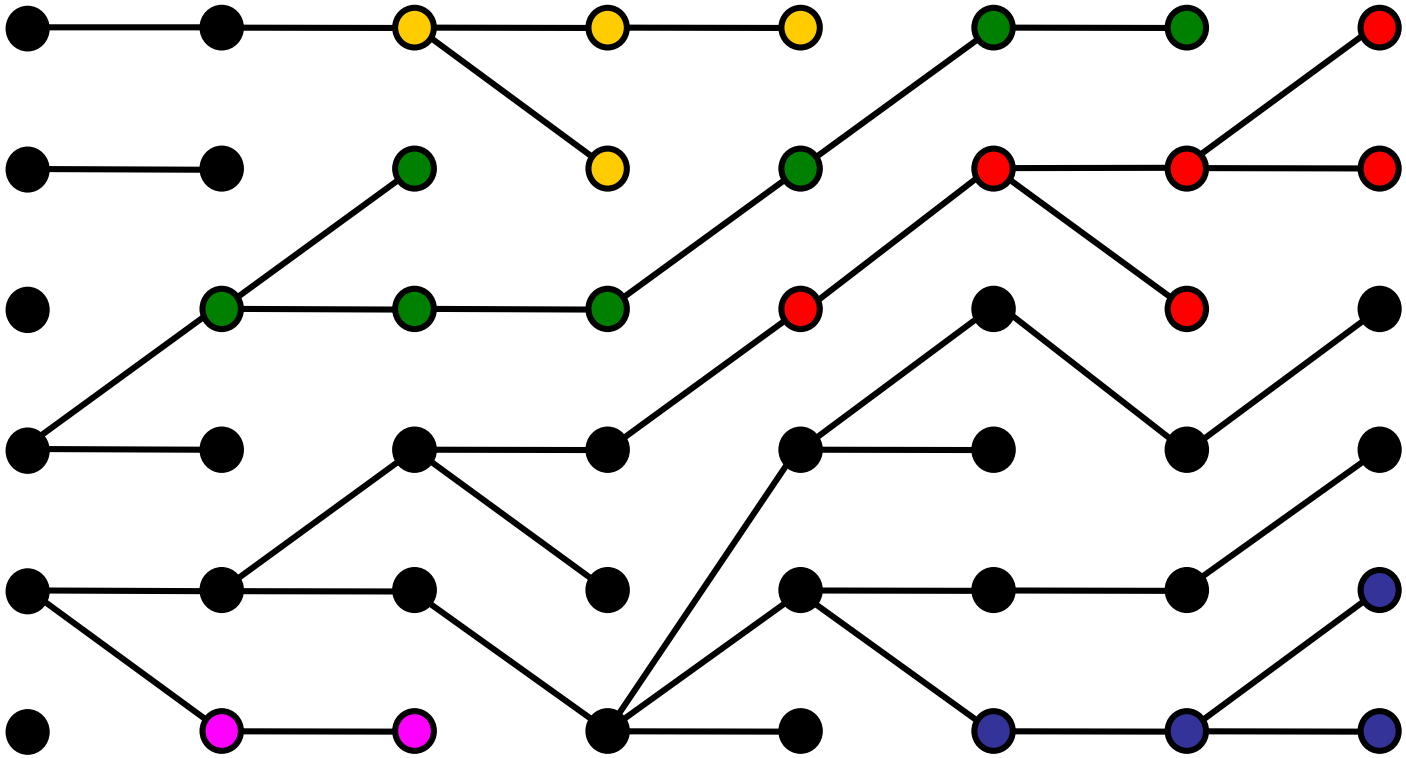


Description of neutral genetic variation
- single locus, multiple alleles

Mutation:
- probability u for each offspring

Three different mutation schemes:
- finite alleles model: like deterministic, include extra step in *infinite gamete pool*
- infinite alleles model: every mutation leads to a new allele (“new color”)
- infinite sites model: every mutation occurs at a different site and thus *remains visible*

The neutral theory
Drift and Mutation



The neutral theory
Drift and Mutation

1. Neutral variation: expected heterozygosity H / nucleotide diversity E[π]

- What is the probability that two randomly sampled alleles from a Wright-Fisher population have a different type?

  - change in heterozygosity from generation $t \rightarrow t+1$

$$H_{t+1} = 2u + (1 - 2u)\left(1 - \frac{1}{2N}\right)H_t$$

  - in equilibrium ($H = H_{t+1} = H_t$): $H \approx \frac{\theta}{\theta + 1}$, $\theta = 4Nu$

  - nucleotide level (θ << 1): $E[\pi] = H_{\text{nucleotide}} = \frac{\theta}{\theta + 1} \approx \theta$

The neutral theory
Drift and Mutation

1. Neutral variation: expected heterozygosity H / nucleotide diversity E[π]

$H \approx \theta = 4Nu$ should increase with mutation rate u and with pop. size N

- Although an increase is observed, there are strong deviations from the prediction of neutral theory:

E Coli	$H \approx 0.16$	$u \approx 10^{-10}$	$\rightarrow N \approx 10^{8}$?	$> 10^{10}$ in each human !
Drosophila	$H \approx 0.01$	$u \approx 3 \cdot 10^{-9}$	$\rightarrow N \approx 10^{6}$?	$> 10^{15}$ (?)
Homo	$H \approx 0.001$	$u \approx 3 \cdot 10^{-8}$	$\rightarrow N \approx 10^{4}$?	$\sim 10^{9}$

Reasons?

- population bottlenecks
- selection

The neutral theory
Drift and Mutation

1. Neutral divergence: substitution rates

➢ At which rate are neutral alleles substituted in a population?

- new mutational input per generation: $2Nu$
- fixation probability for each new mutant: $p_{fix} = \frac{1}{2N}$

➢ neutral substitution rate: $2Nu \cdot \frac{1}{2N} = u$

- independent of population size !
- basis for “molecular clock” estimates

The neutral theory
Drift and Mutation

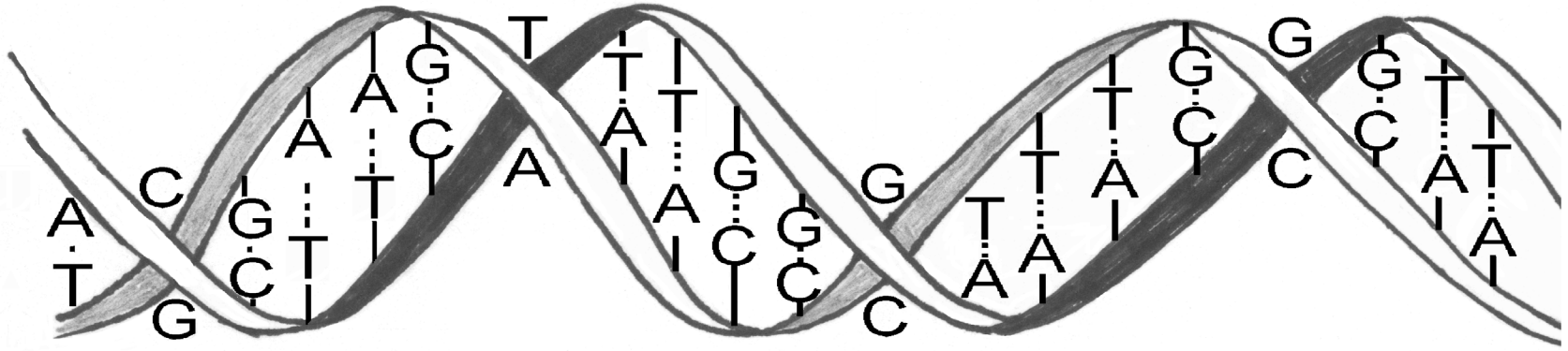
Can drift (plus mutation) explain observed patterns of diversity ?

- Sparked the fierce adaptionist / neutralist debate
- Today: selection seems to be very important even at the molecular level:
  - New mutations: many non-coding parts of the genome under selection (regulatory elements, etc)
  - Substitutions: large fractions seem to be driven by positive selection (> 50% in *Drosophila*)
- But: neutral theory generally accepted null model of molecular evolution

➢ Foundation of statistical genetics as research field

Introduction to the Coalescent

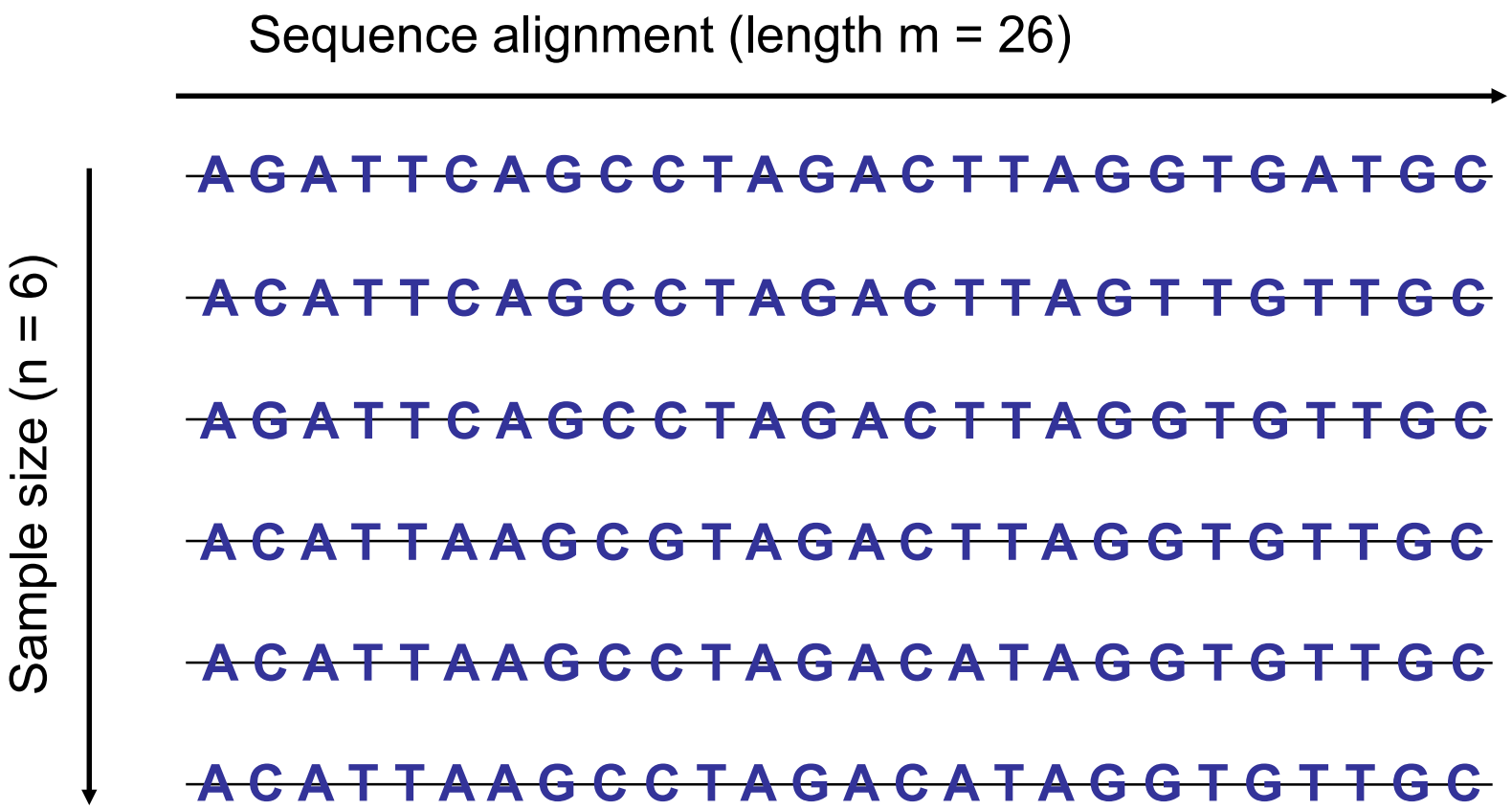
data, data, data, …



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 Projects* also for *Drosophila, Arabidopsis* …
- *today*: GWAS sample sizes > 500 000 (UK-Biobank), long reads …

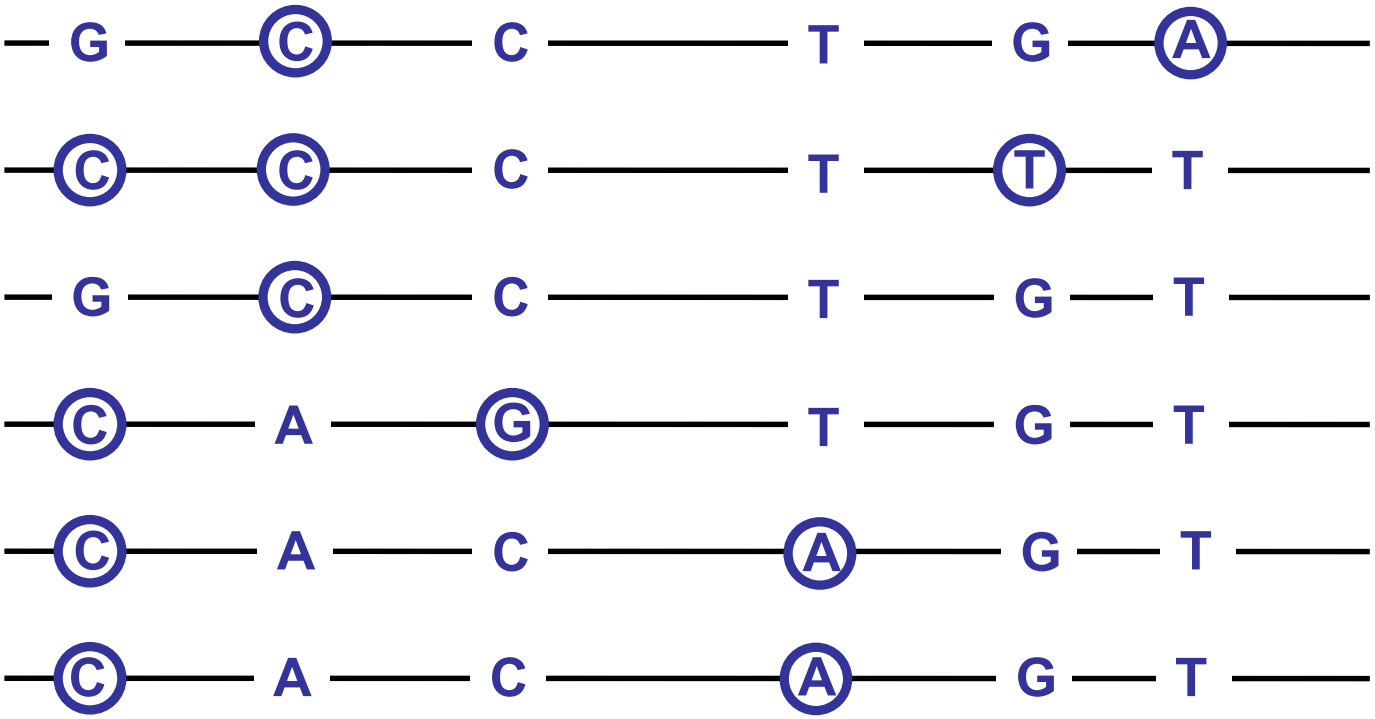
Patterns of Evolution
”Summary Statistics”



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

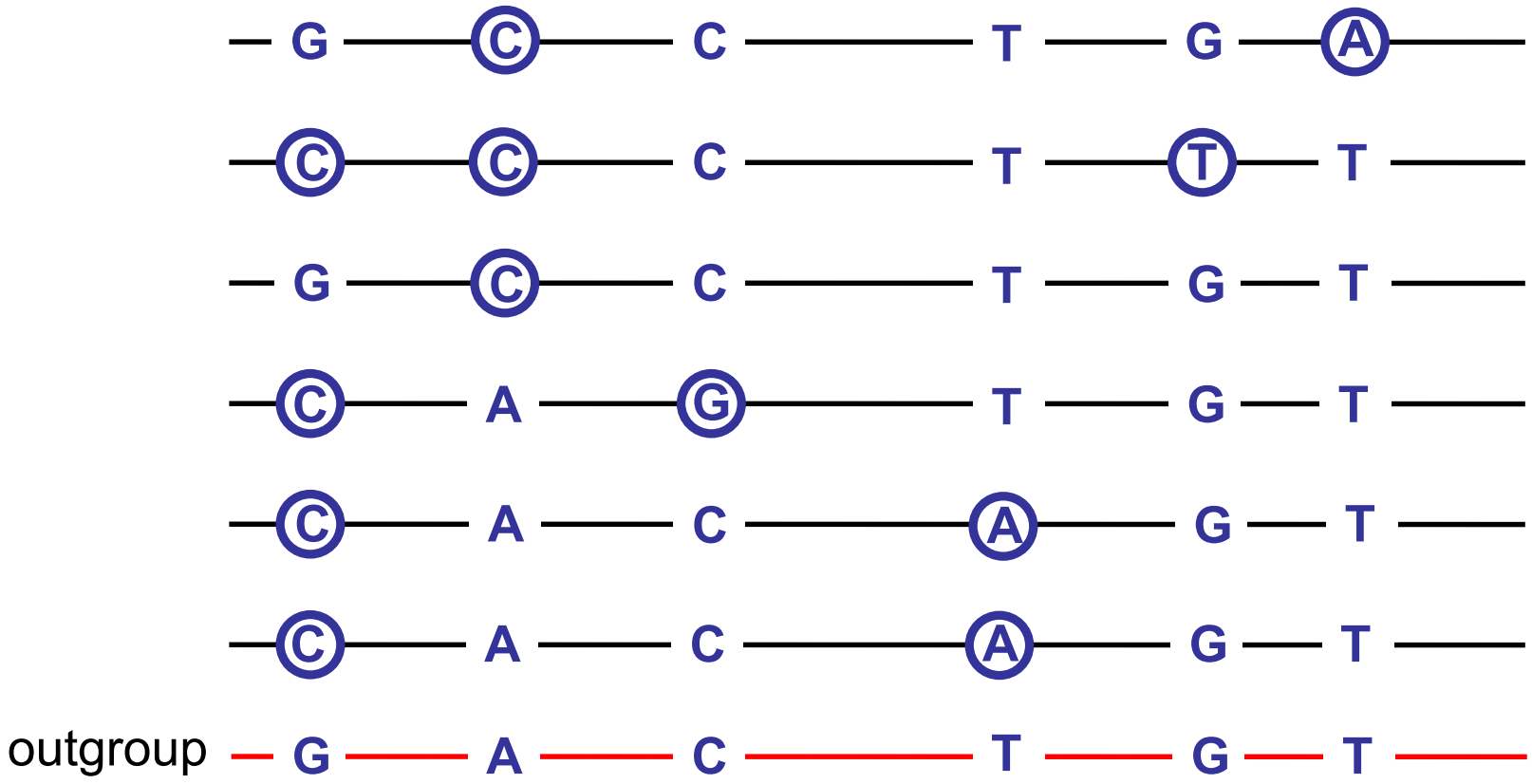
Patterns of Evolution
”Summary Statistics”

only polymorphic sites …



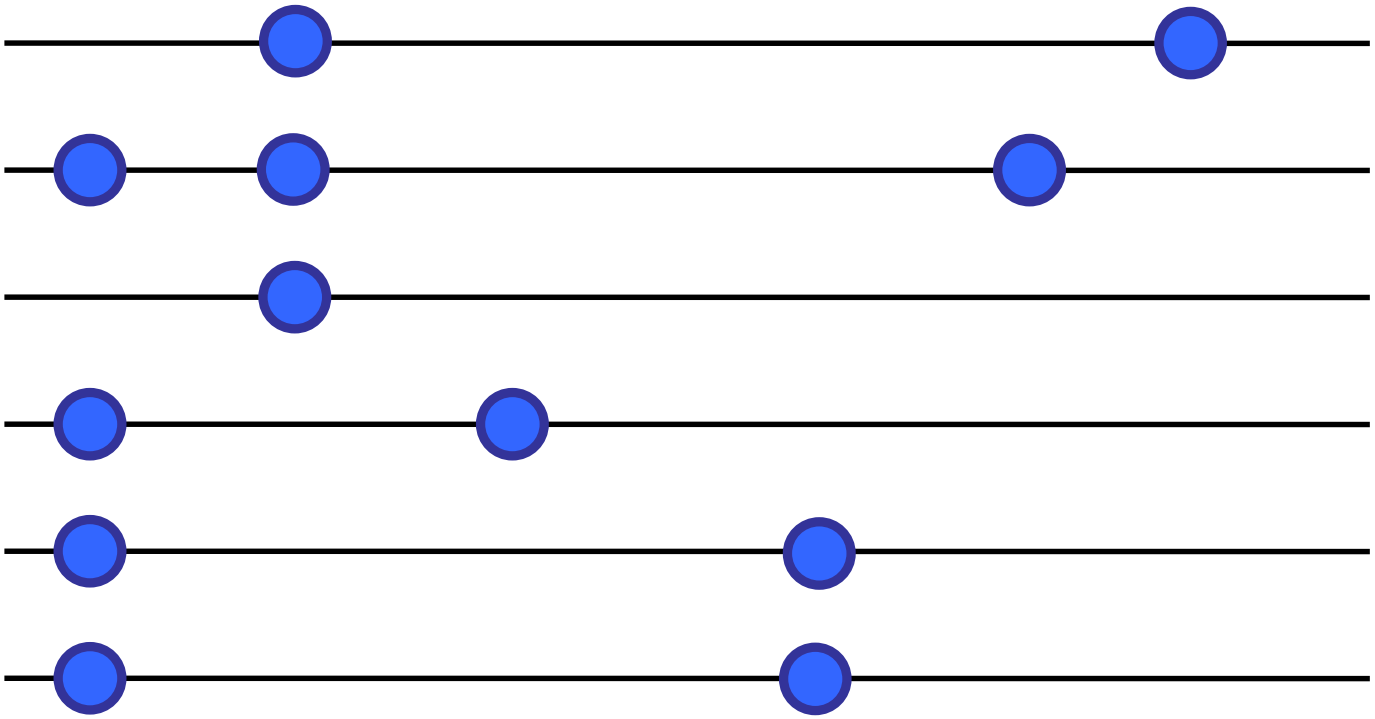
Patterns of Evolution
”Summary Statistics”

compare with outgroup …



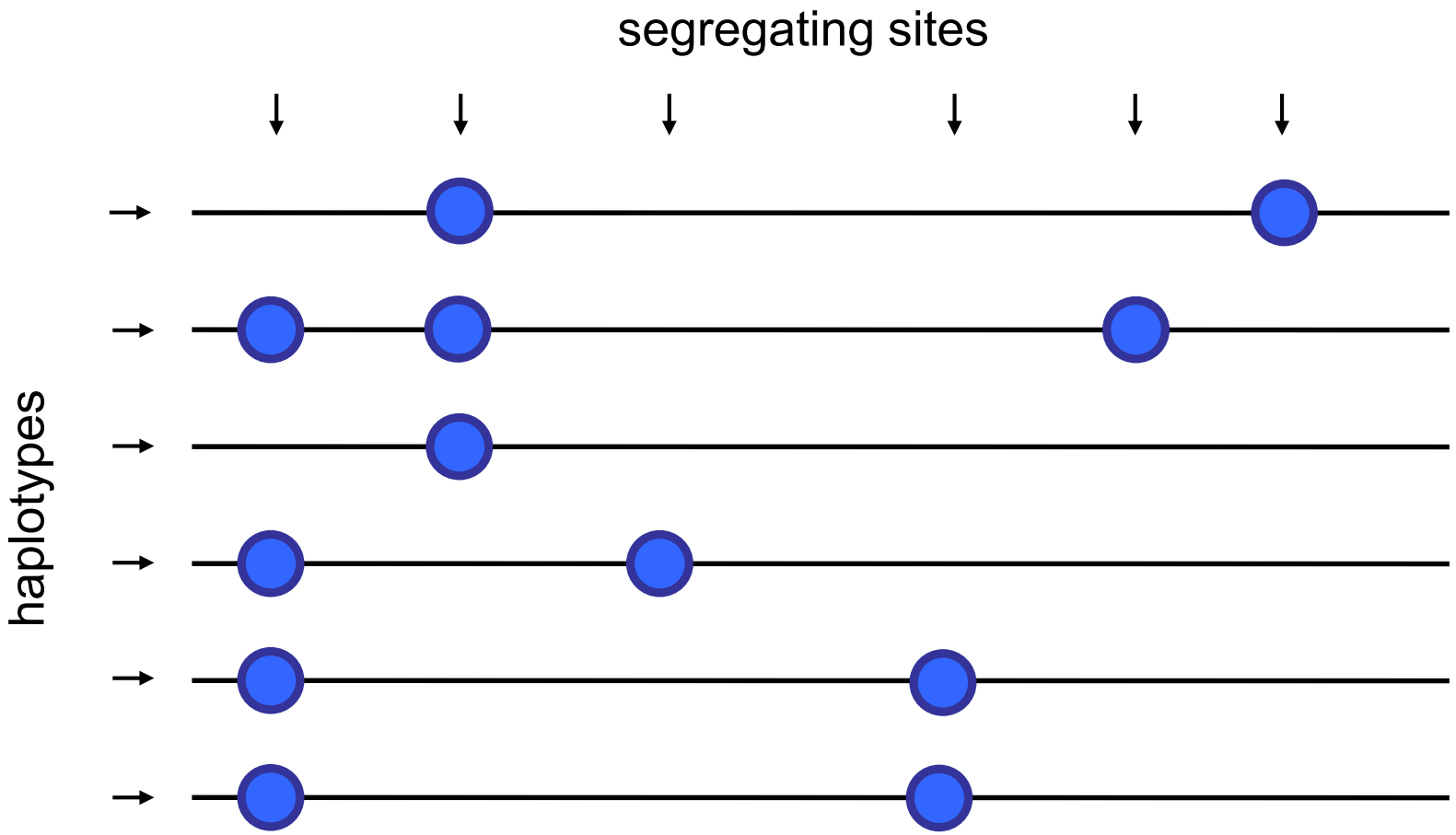
Patterns of Evolution
”Summary Statistics”

forget about molecular state …



(assumes *infinite sites mutation* model)

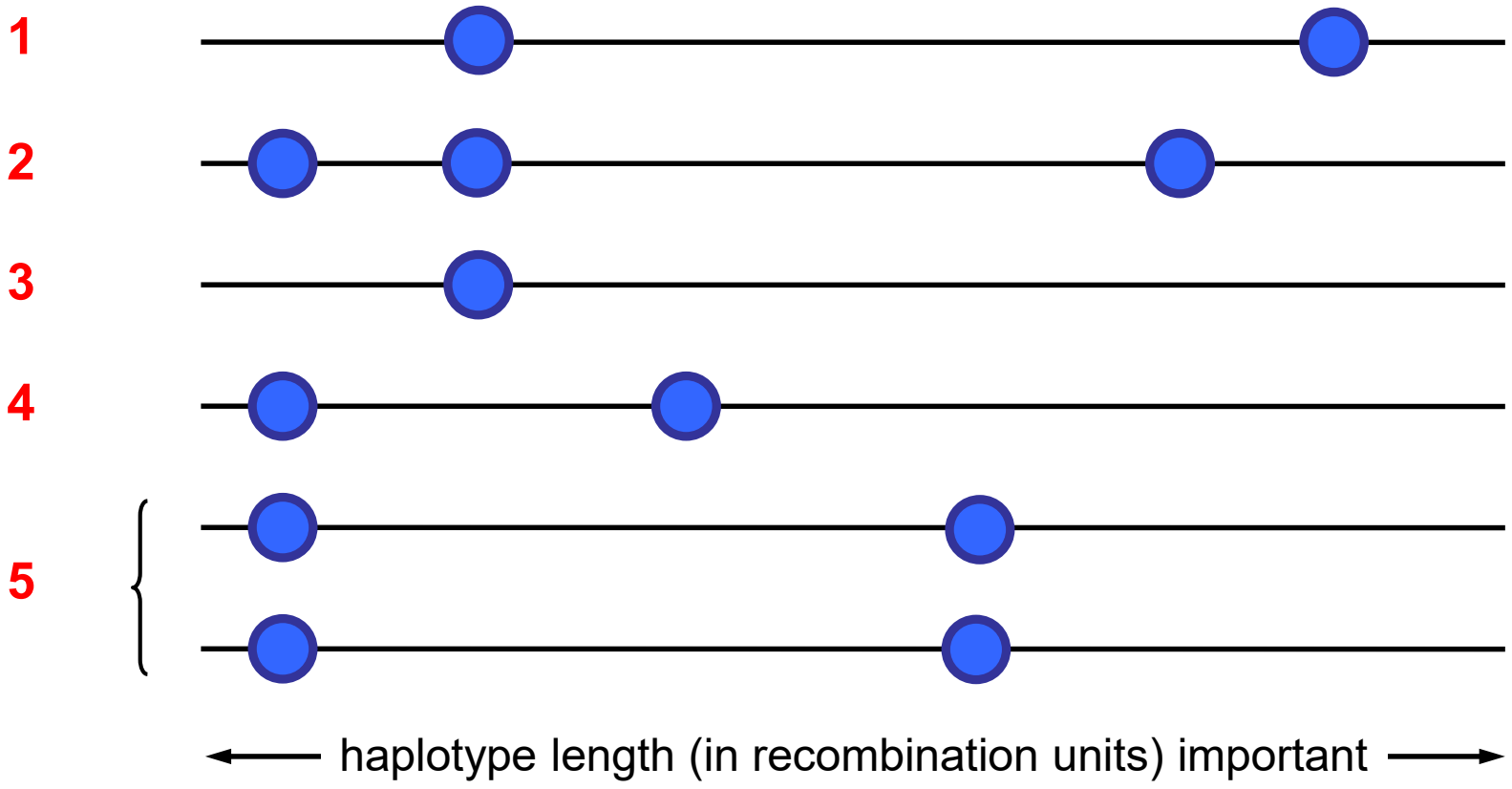
Patterns of Evolution
”Summary Statistics”



Patterns of Evolution

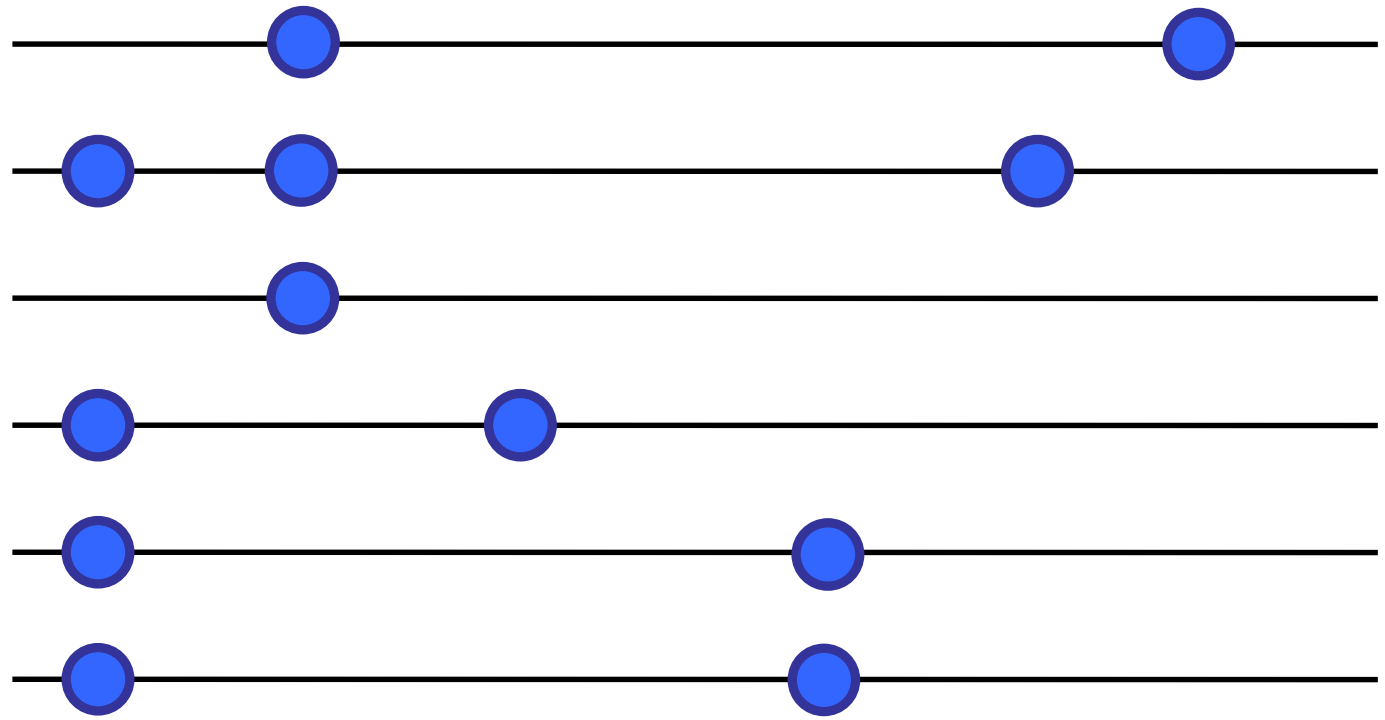
Summary statistics based on haplotypes or LD

- 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

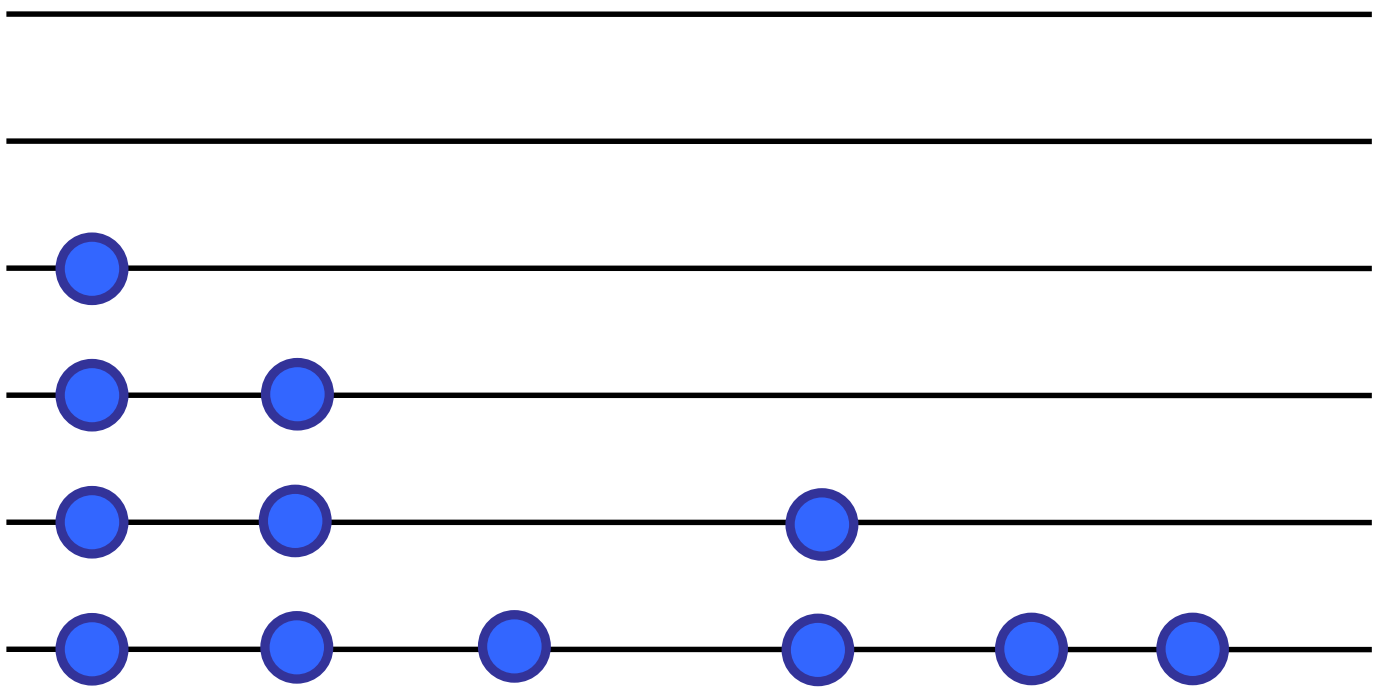


mutation *"size"*:

4 3 1 2 1 1

Patterns of Evolution
Summary statistics based on segregating sites

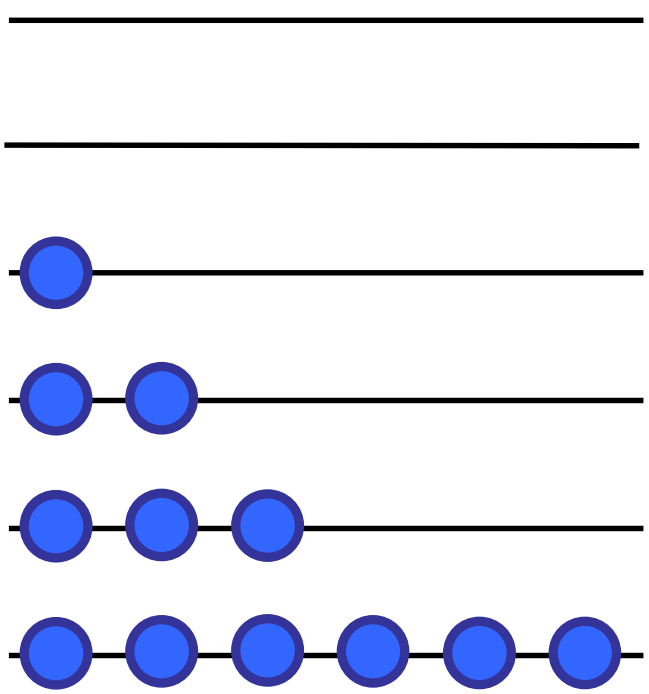
- number of segregating sites and allele frequencies
  - associations not important (“molecular bean bag“)



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“)



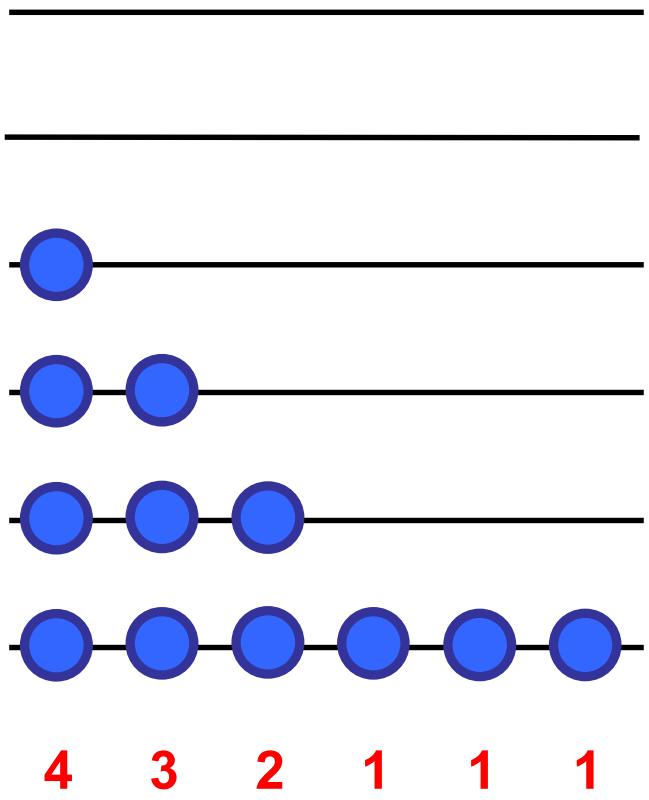
mutation “*size*“: **4 3 2 1 1 1**

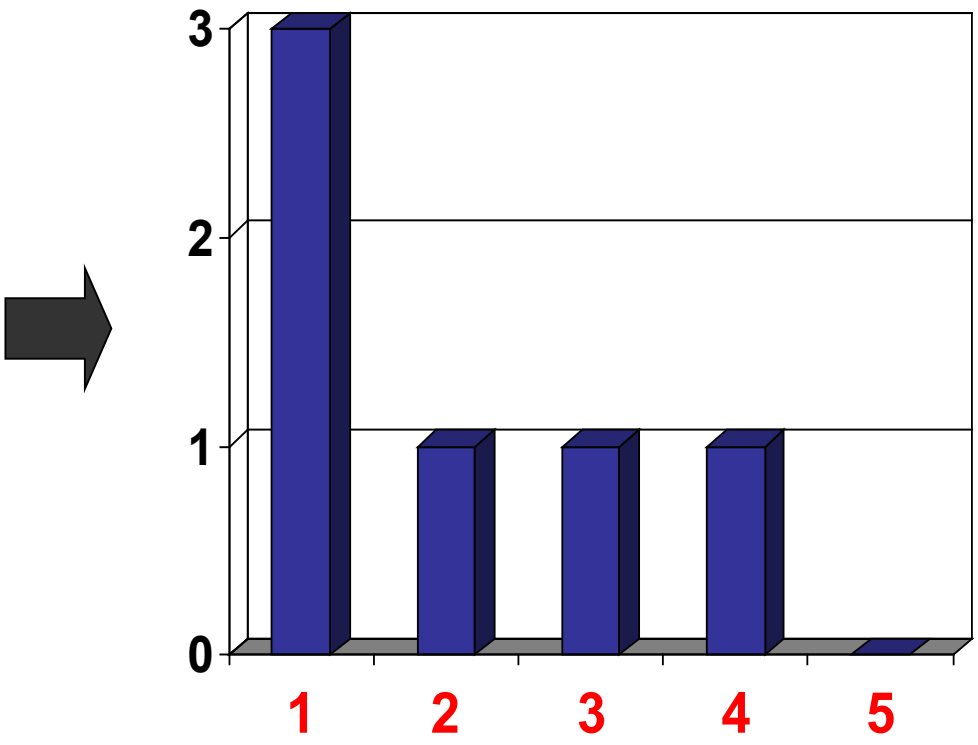
- genome position does not matter

Patterns of Evolution

Summary statistics based on segregating sites

Site Frequency Spectrum





Patterns of Evolution

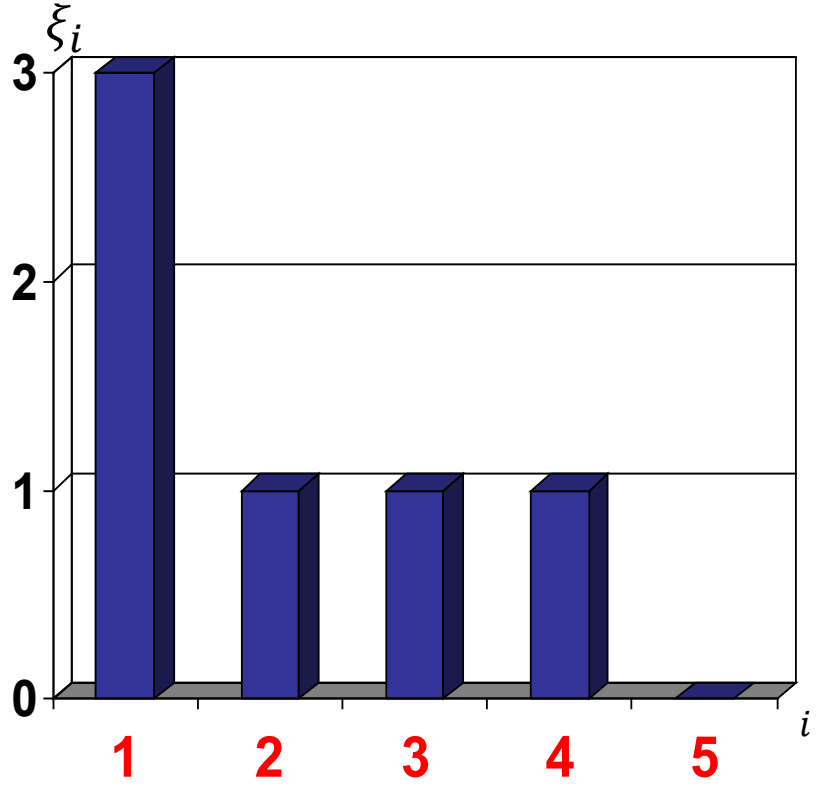
Summary statistics based on segregating sites

Site Frequency Spectrum

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

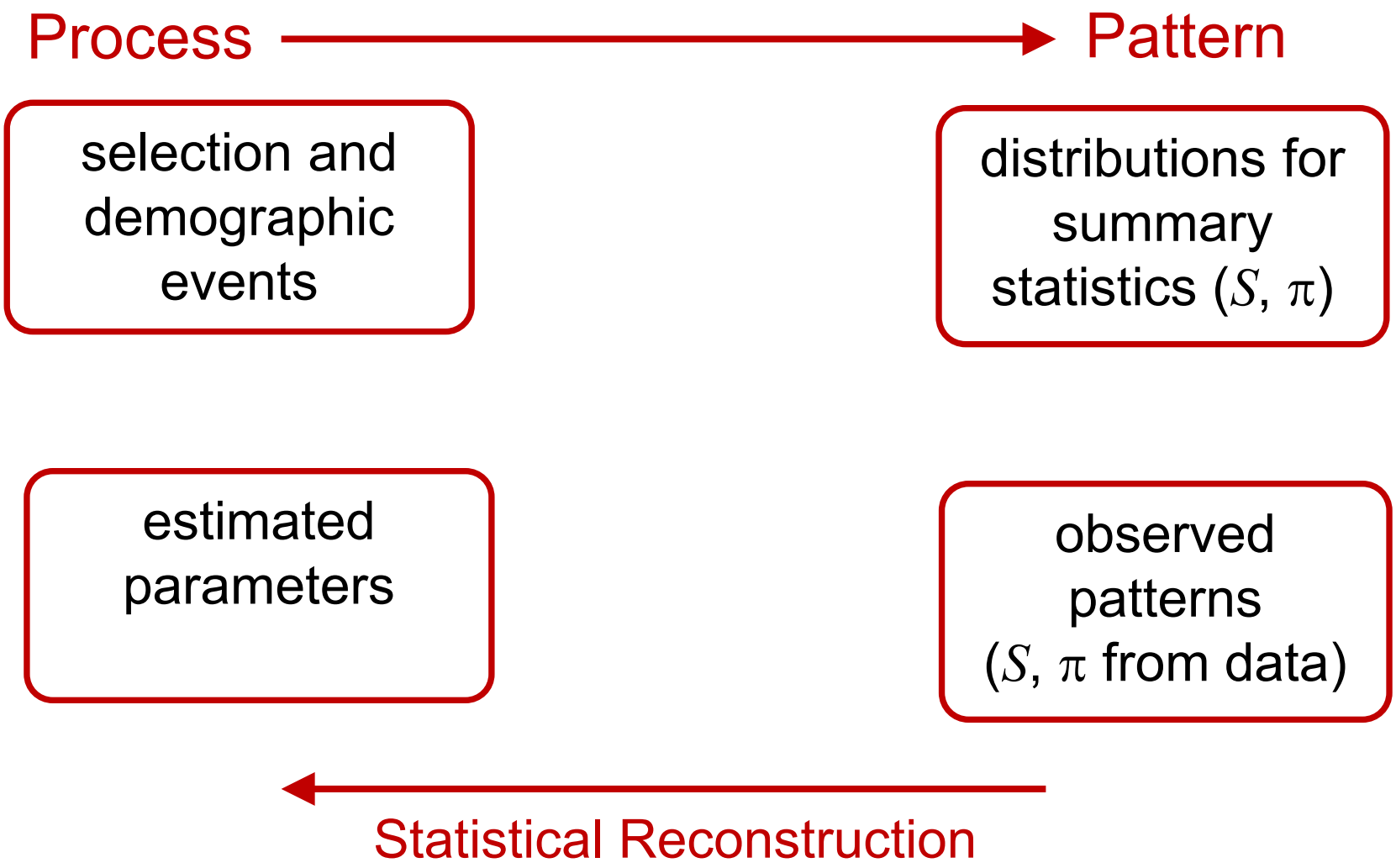
$S = \sum_{i=1}^{n-1} \xi_i$: total number of segregating sites in an sample of size n

$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i$: average number of pairwise differences



Patterns of Evolution

Reconstruction of evolutionary history



Patterns of Evolution

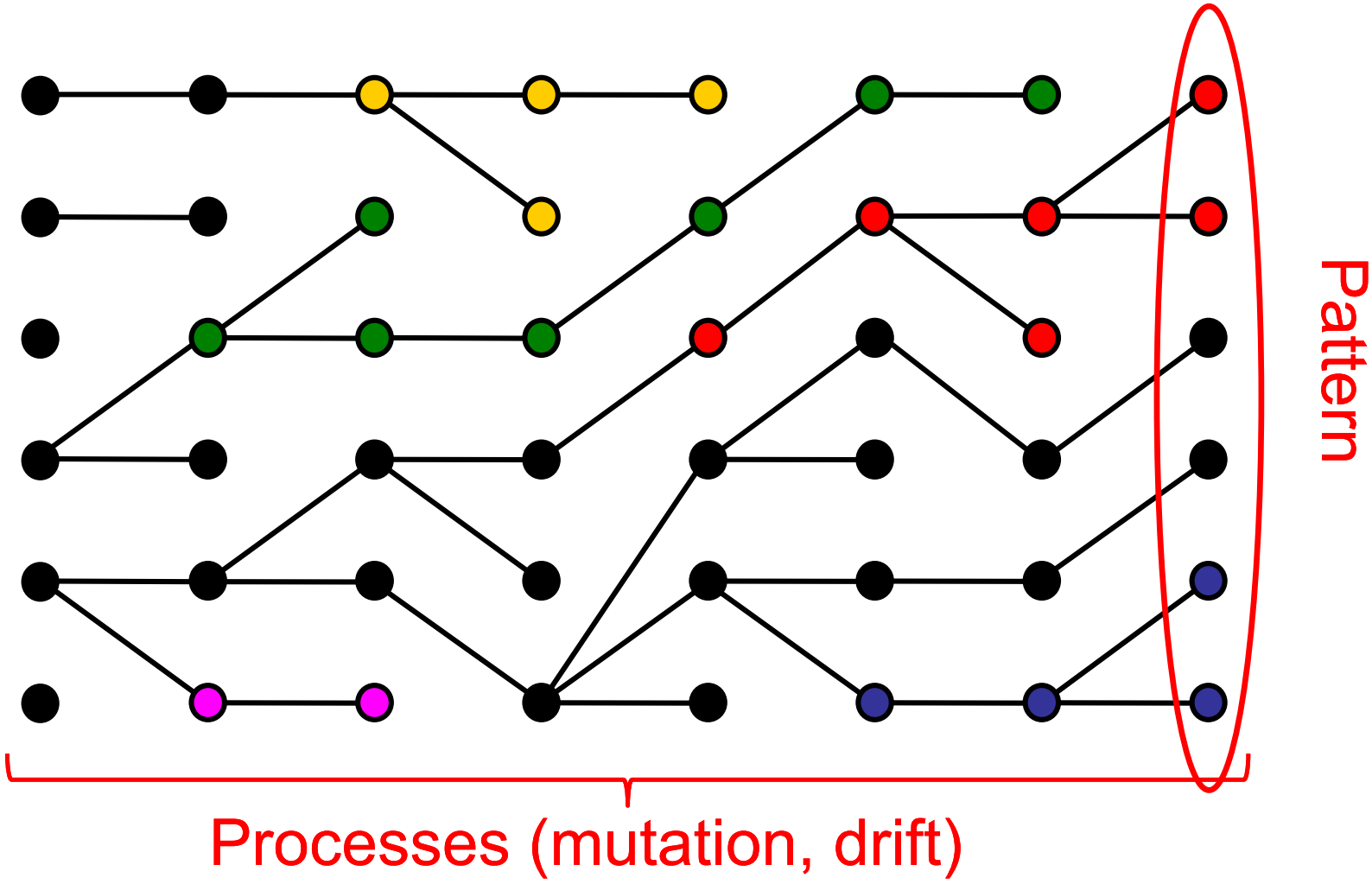
Reconstruction of evolutionary history

How does pure randomness look like ?

- Null-model of the evolutionary theory

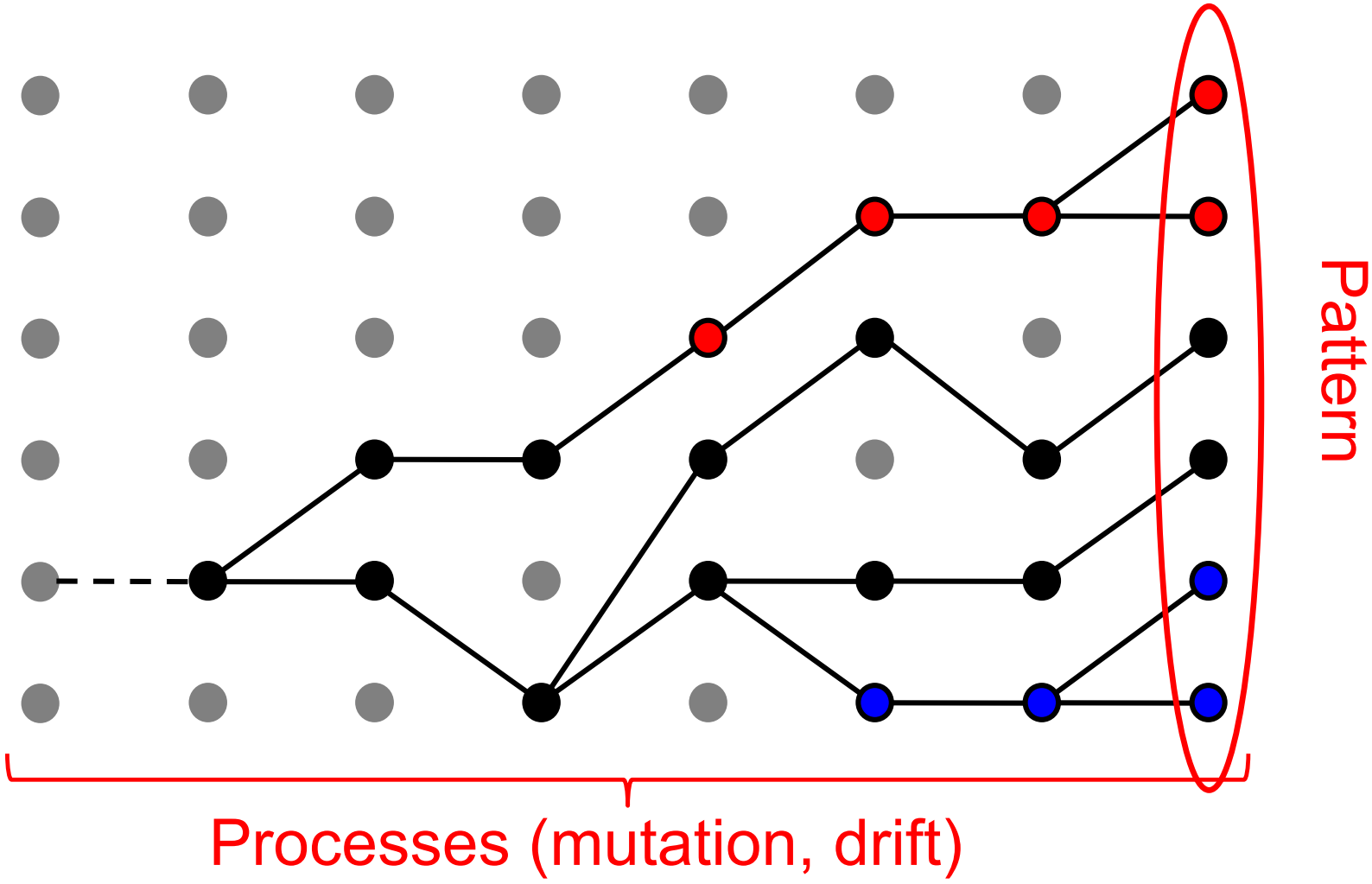
Patterns of Evolution

Wright-Fisher model



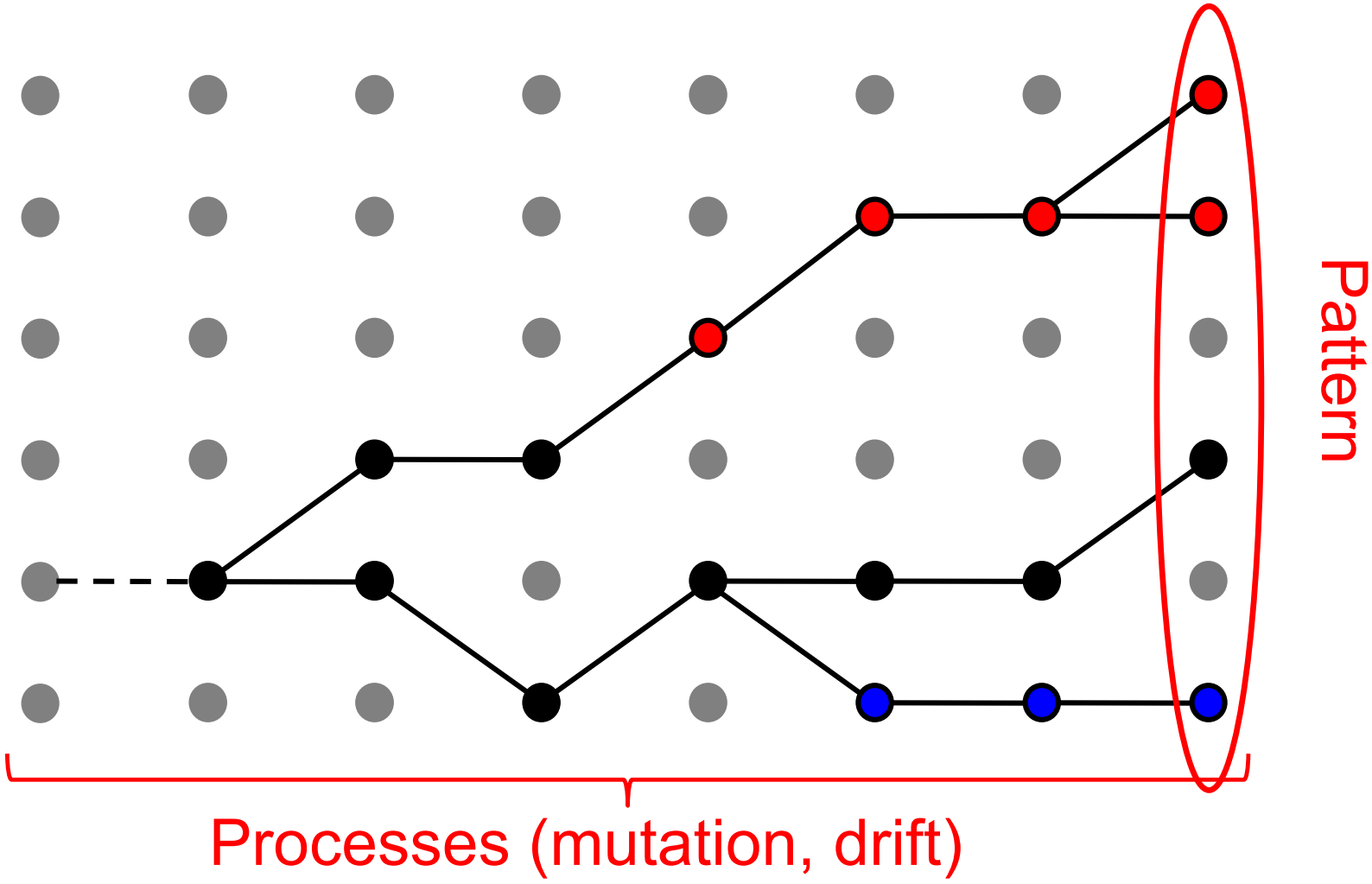
Patterns of Evolution

Wright-Fisher model

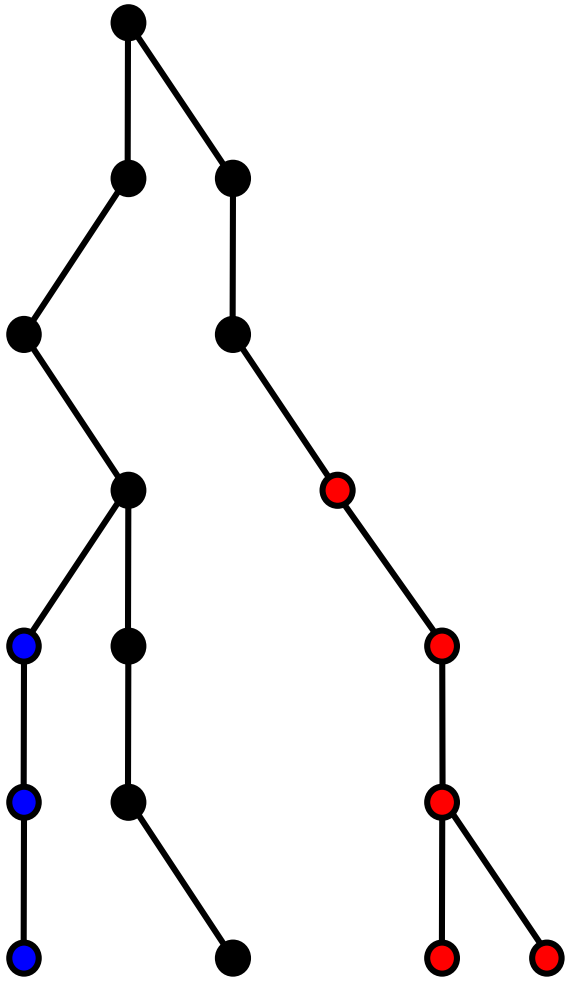


Patterns of Evolution

Wright-Fisher model

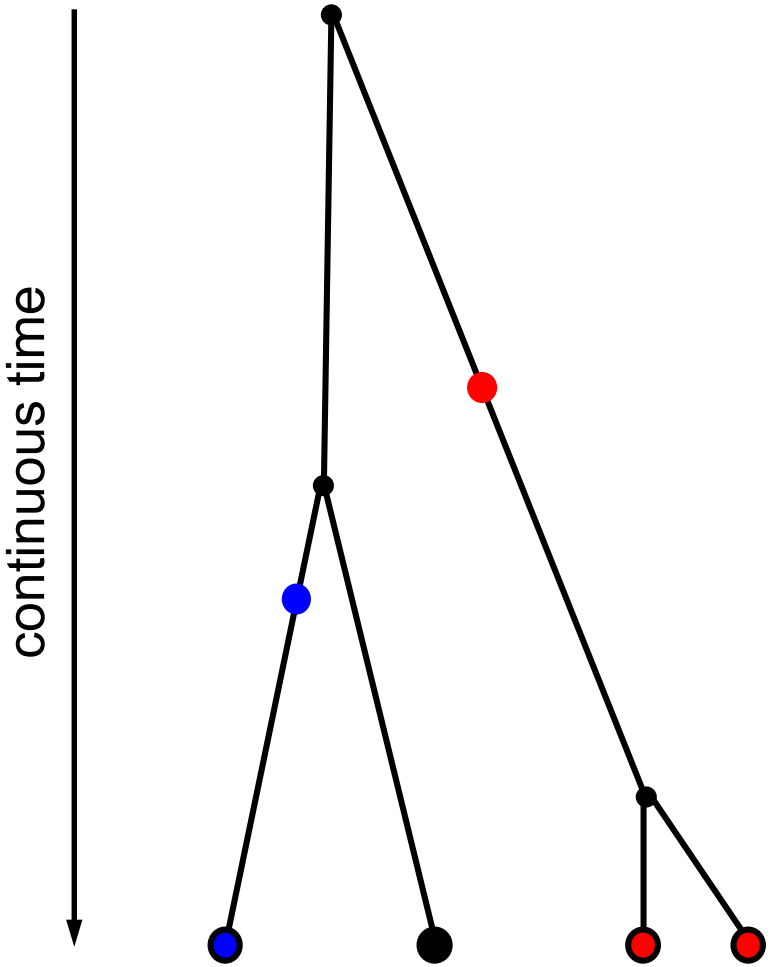


Patterns of Evolution
coalescence process



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

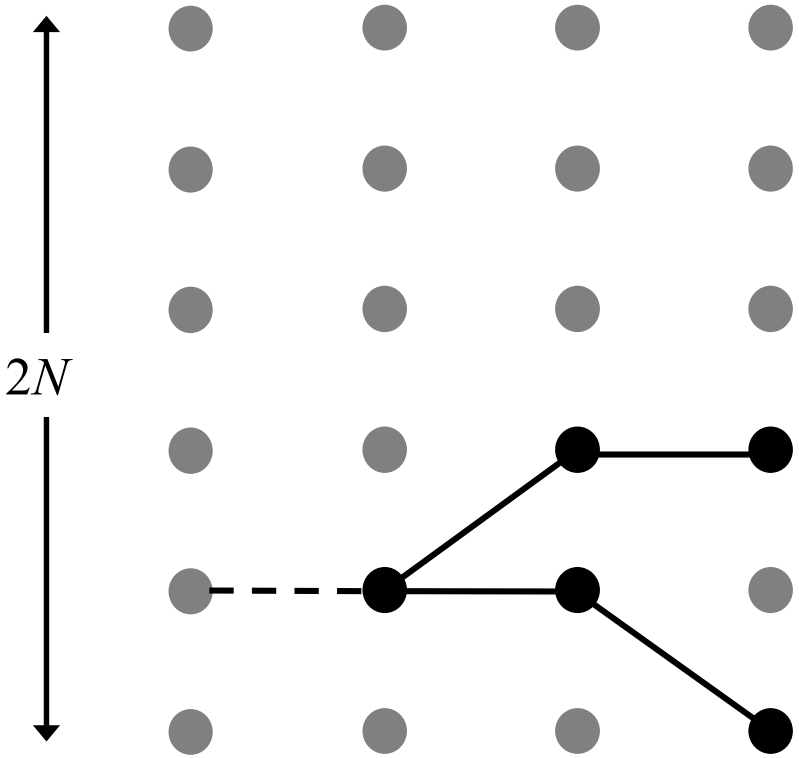
➢ Wright-Fisher: multinomial sampling

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:

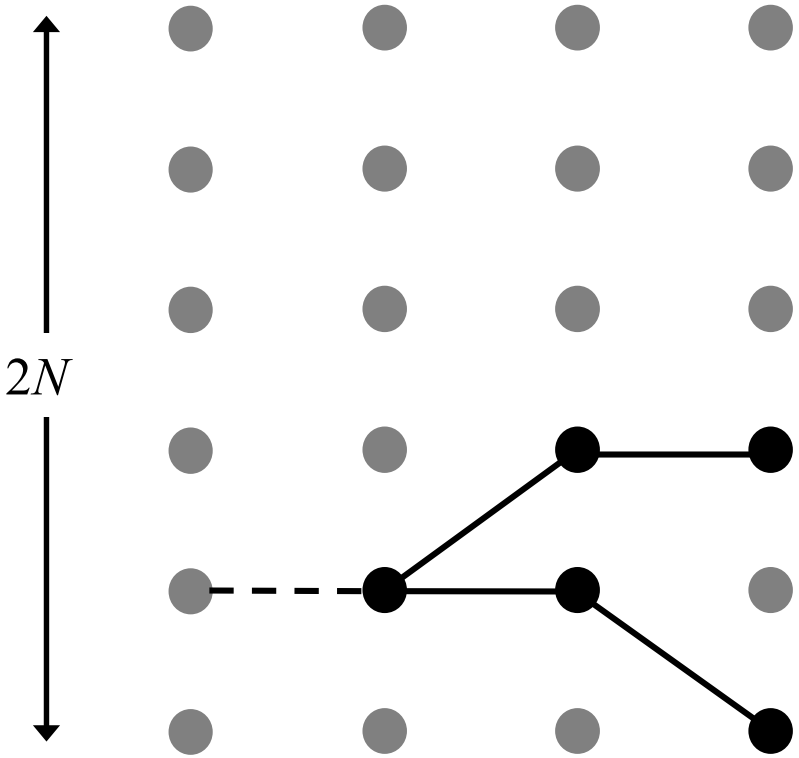
$$p_{c,1} = \frac{1}{2N}$$

… for exactly 2 generations:

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

Coalescent Theory

Construction of the Genealogy: Sample Size 2



Coalescence probability

… in a single generation:

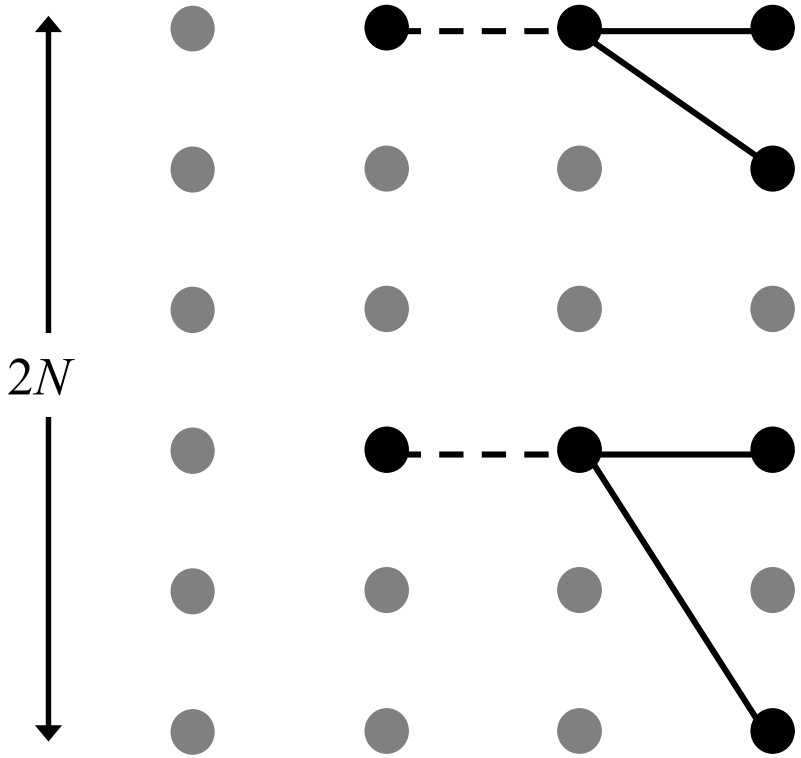
$$p_{c,1} = \frac{1}{2N}$$

… 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:

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

Multiple coalescences:

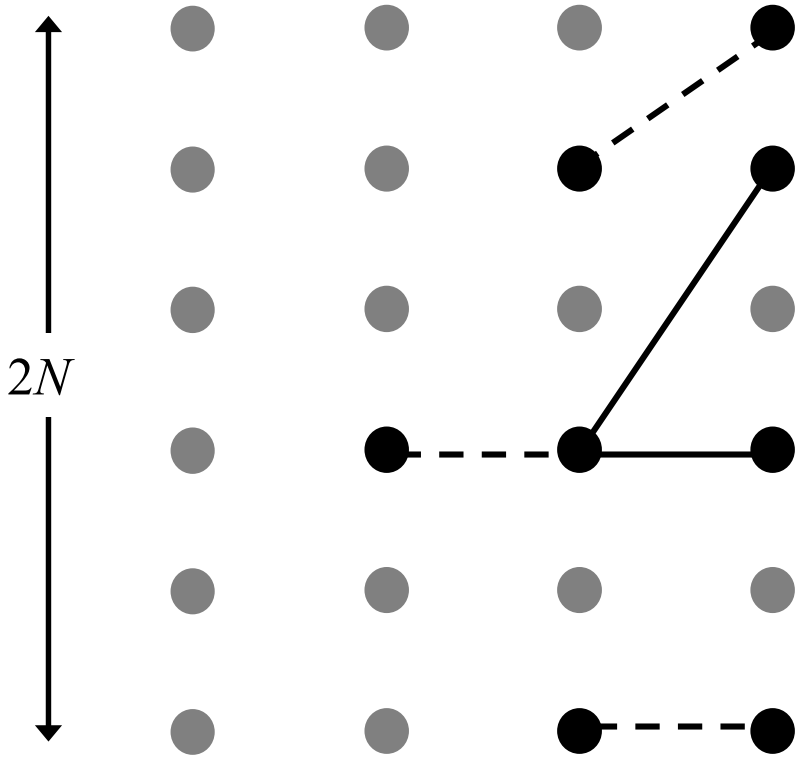
$$\Pr \propto p_{c,t}^2 = O[N^{-2}]$$

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)

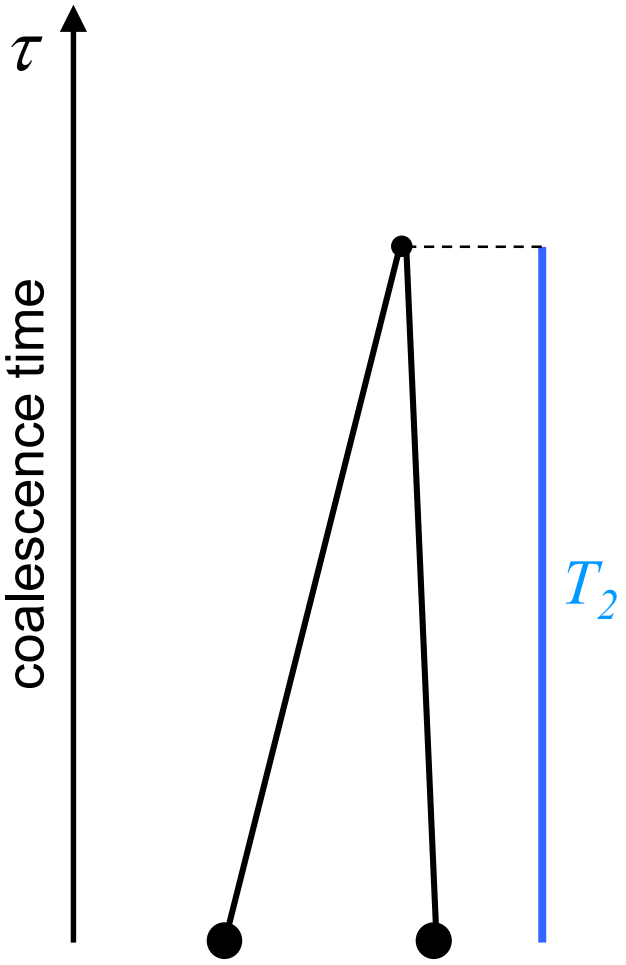
… in a single generation:

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

… 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:

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

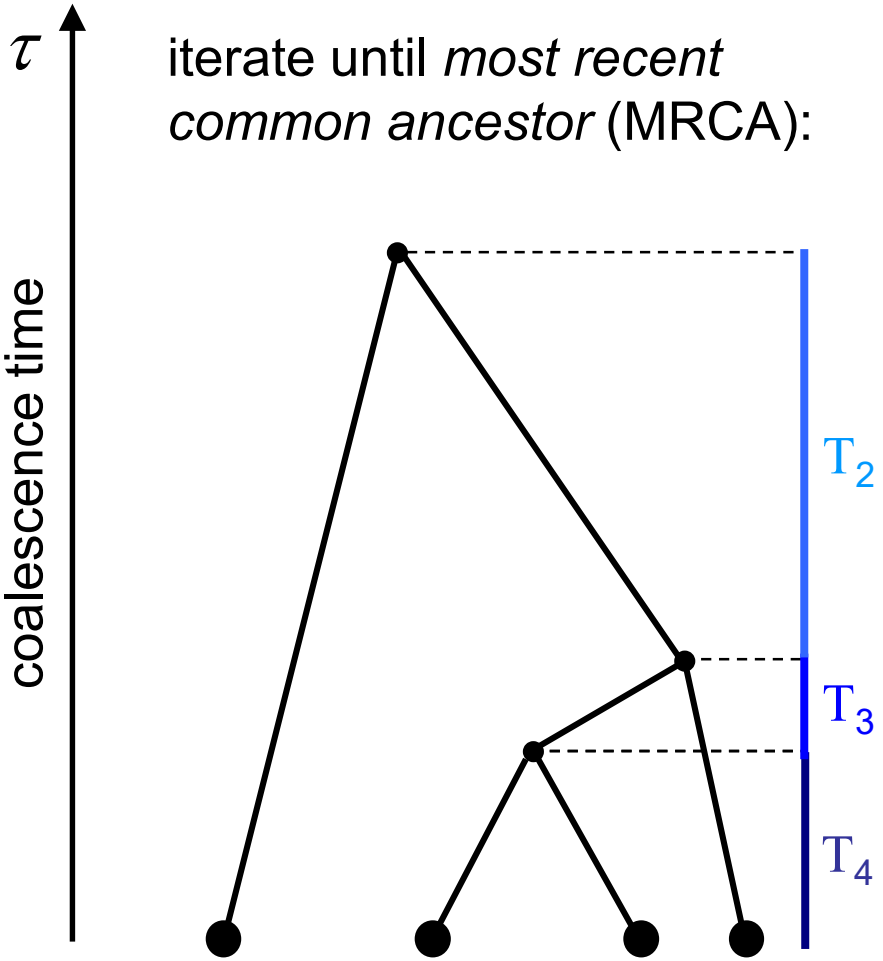
Coalescence time T_2 for sample size 2:

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

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):

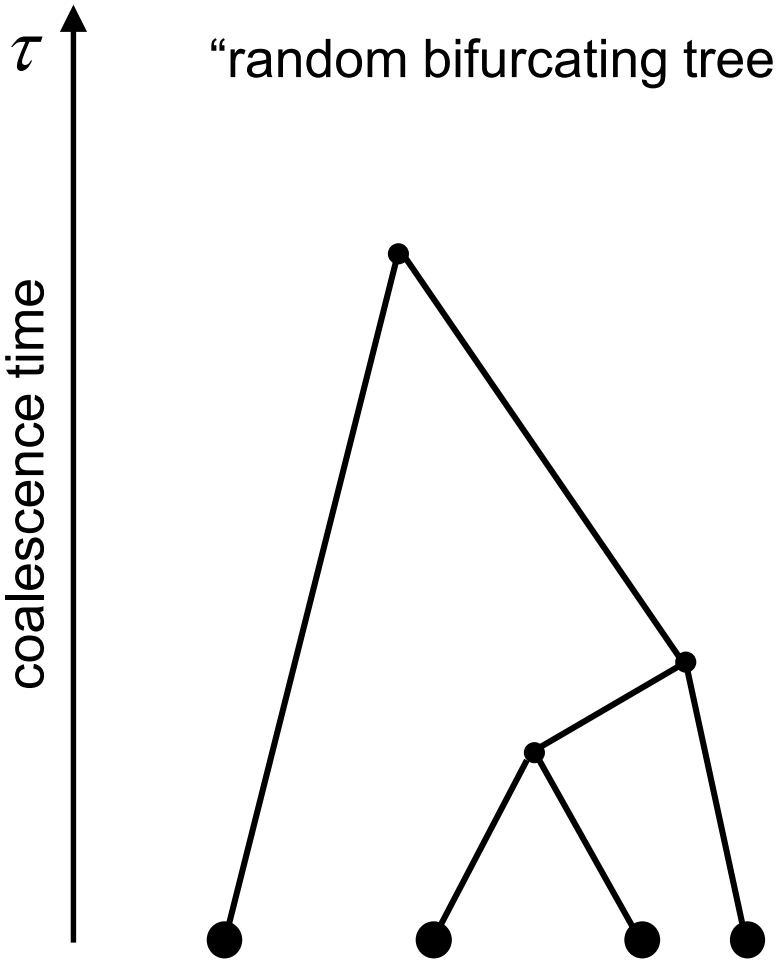
with sample size n:

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

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

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

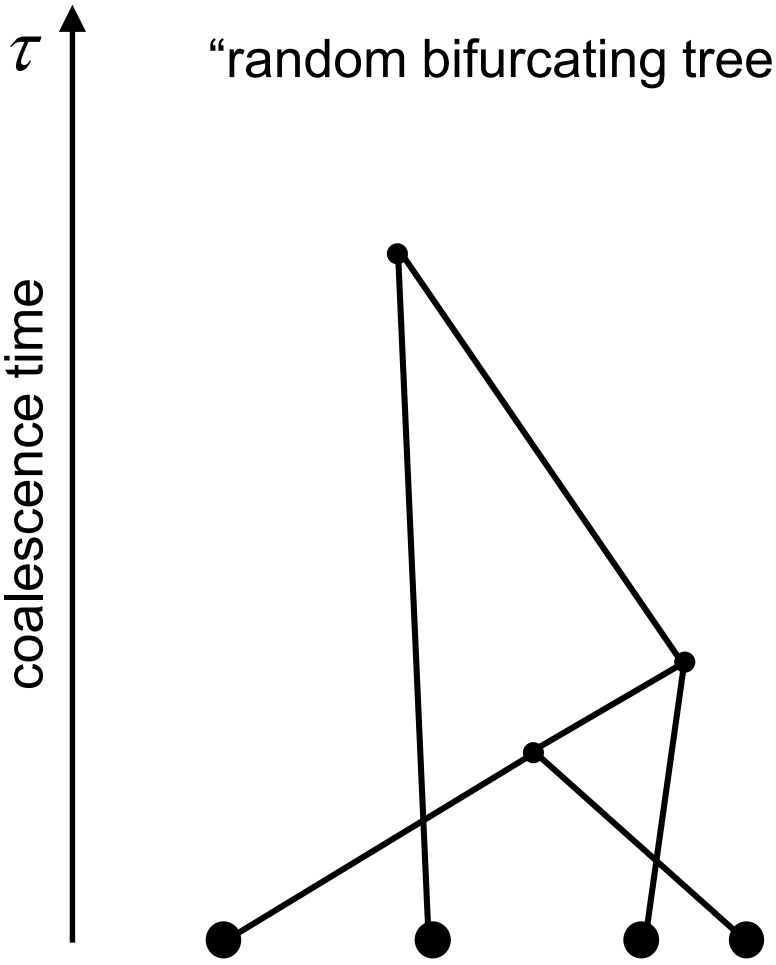
Coalescent Theory
Tree Topologies



"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



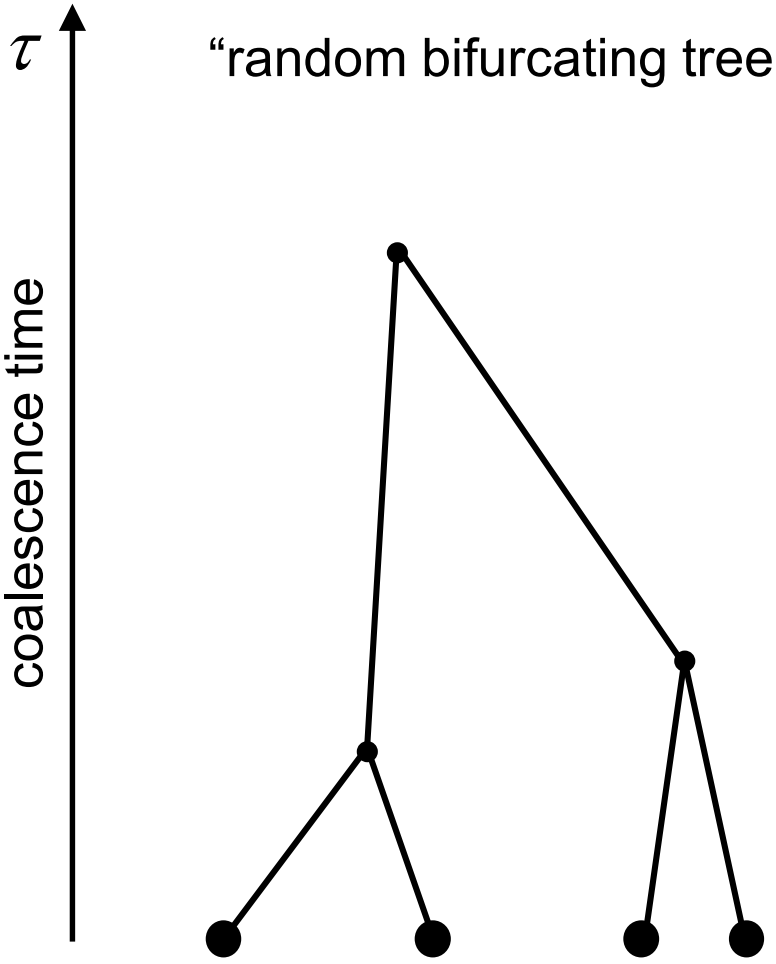
"random bifurcating tree"

- 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



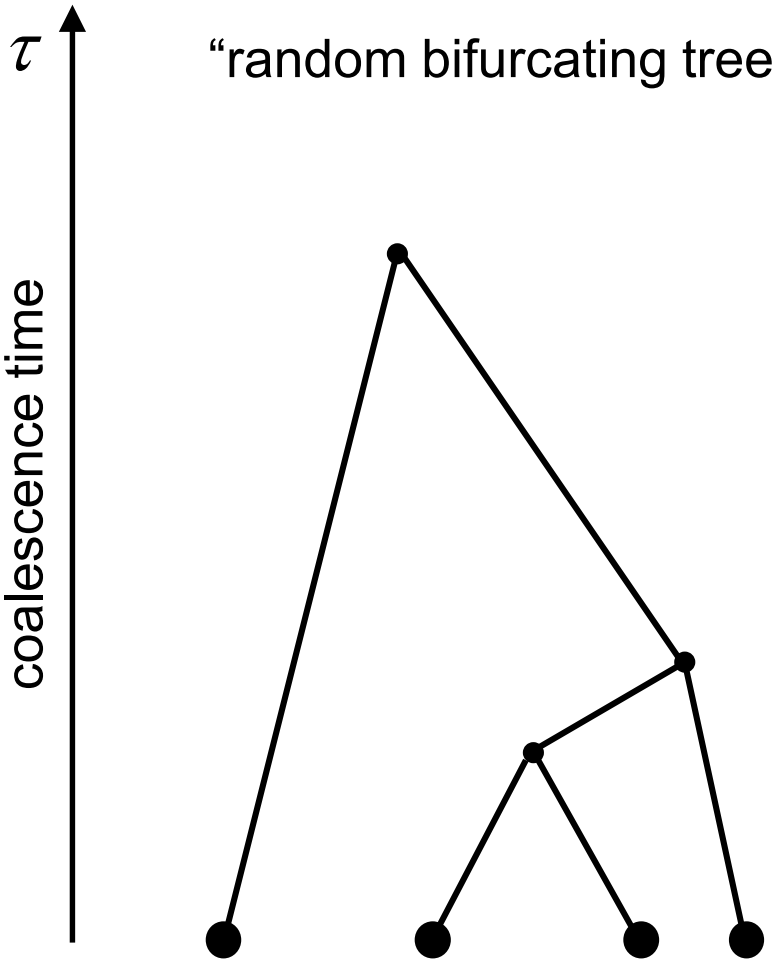
"random bifurcating tree"

- 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

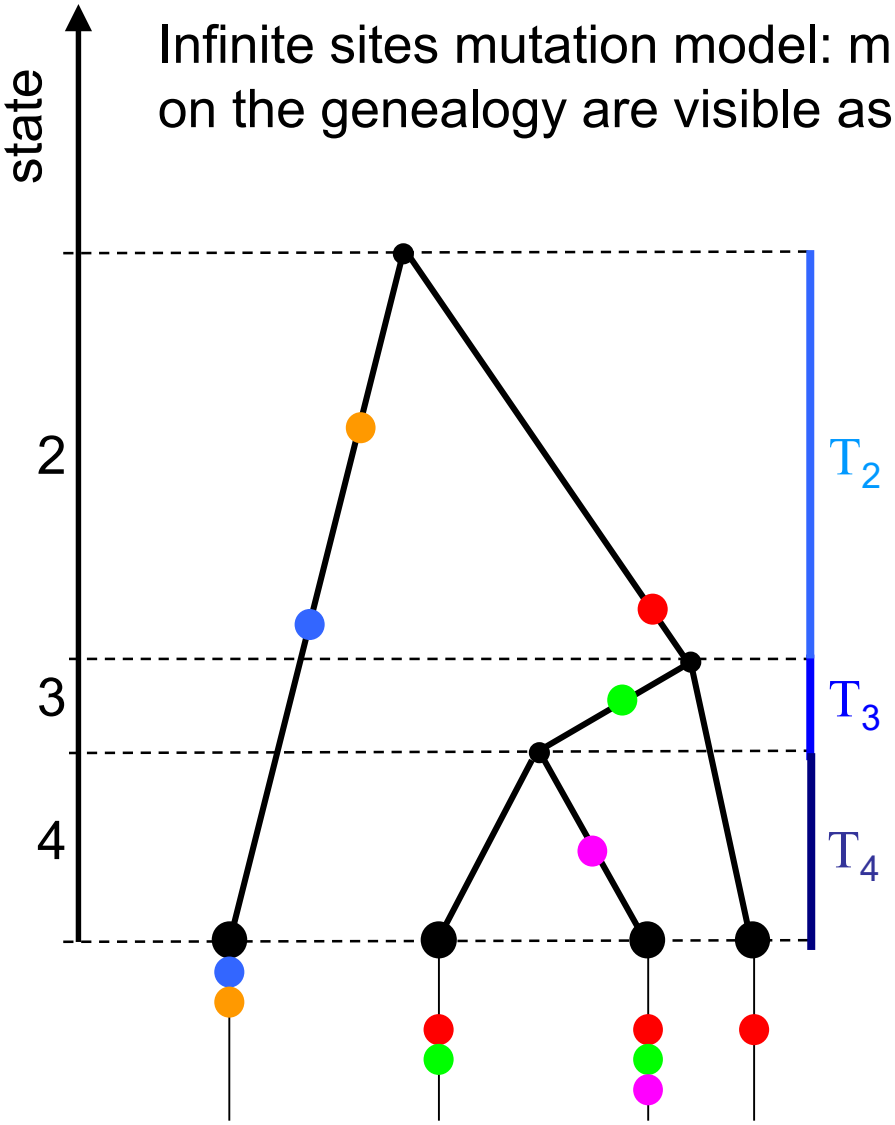
➢ topology invariant under permutation of "leaves"

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

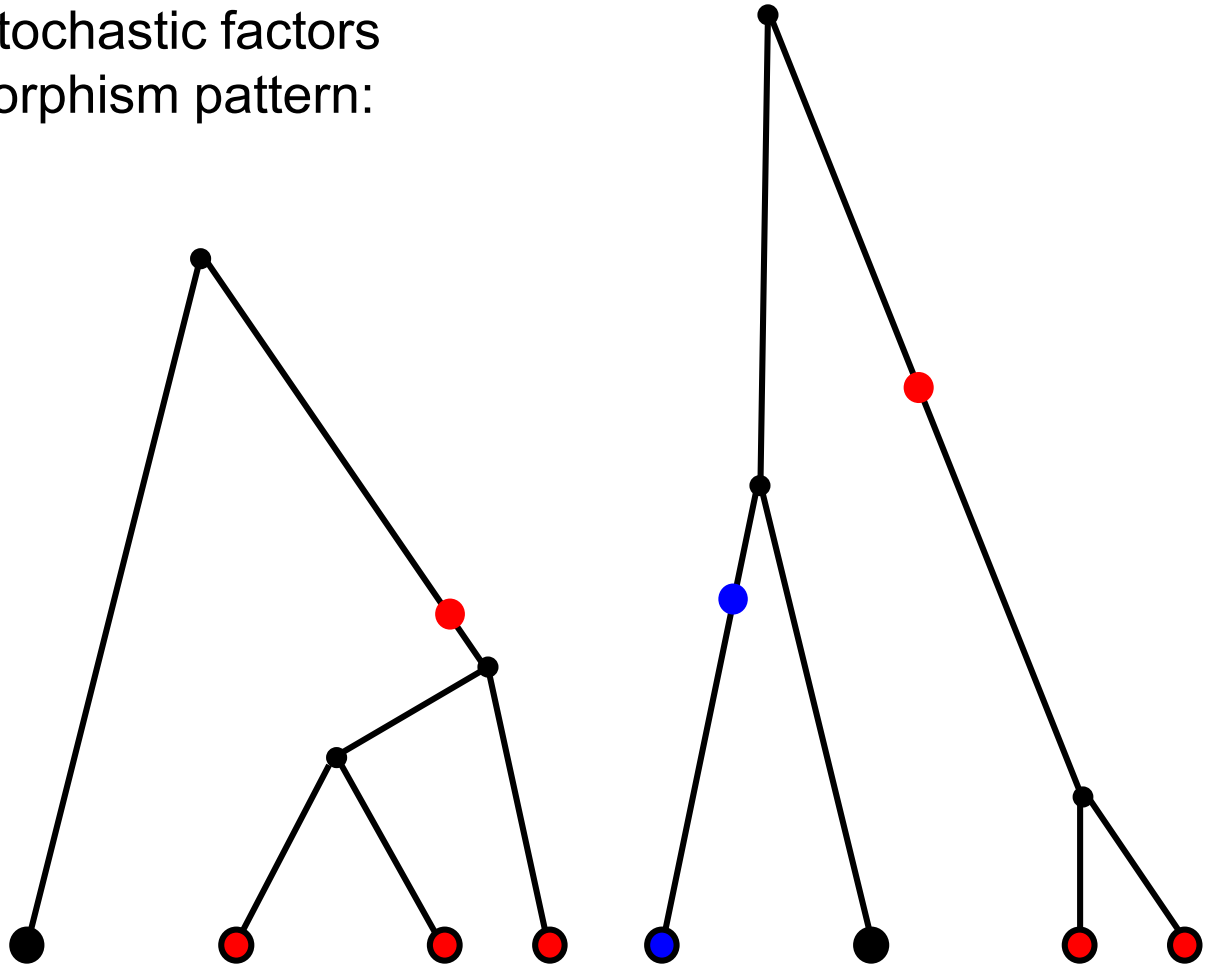
(also other mutation schemes possible)

Coalescent Theory
Basic Properties

Three independent stochastic factors determine the polymorphism pattern:

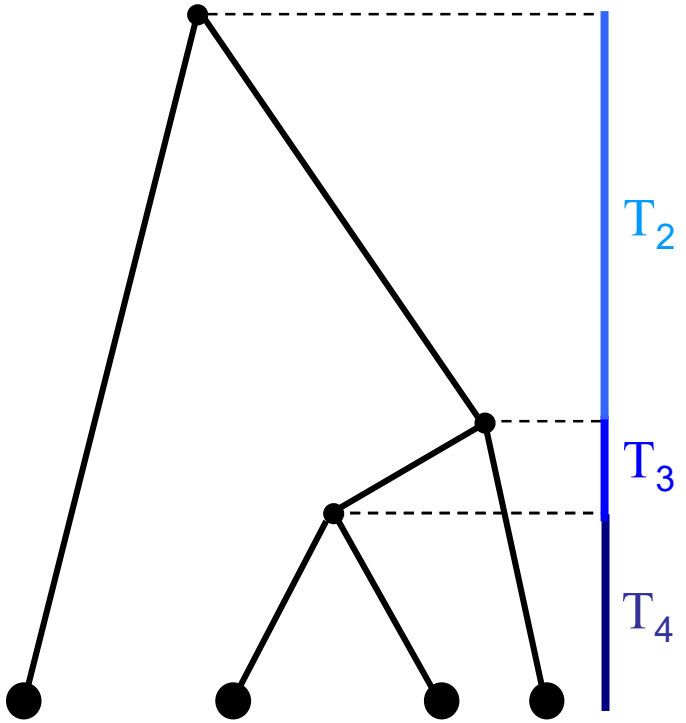
1. coalescent times
2. tree topology
3. mutation

(very easy to implement in simulations)



Coalescent Theory
Basic Properties

Time to the most recent common ancestor:



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

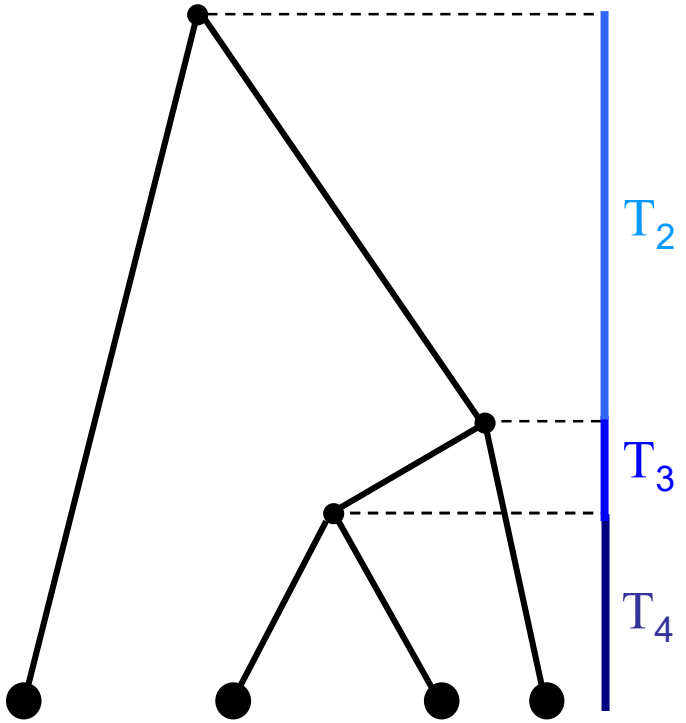
$$= 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right)$$

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:



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

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

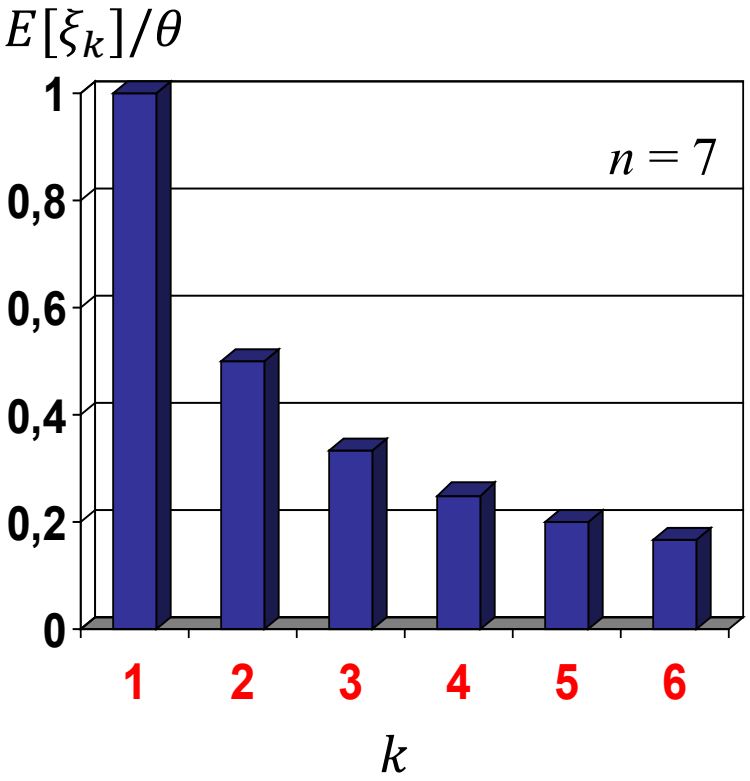
with: $a_n = \sum_{k=1}^{n-1} \frac{1}{k} \quad \rightarrow \log n + 0.577$

(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)



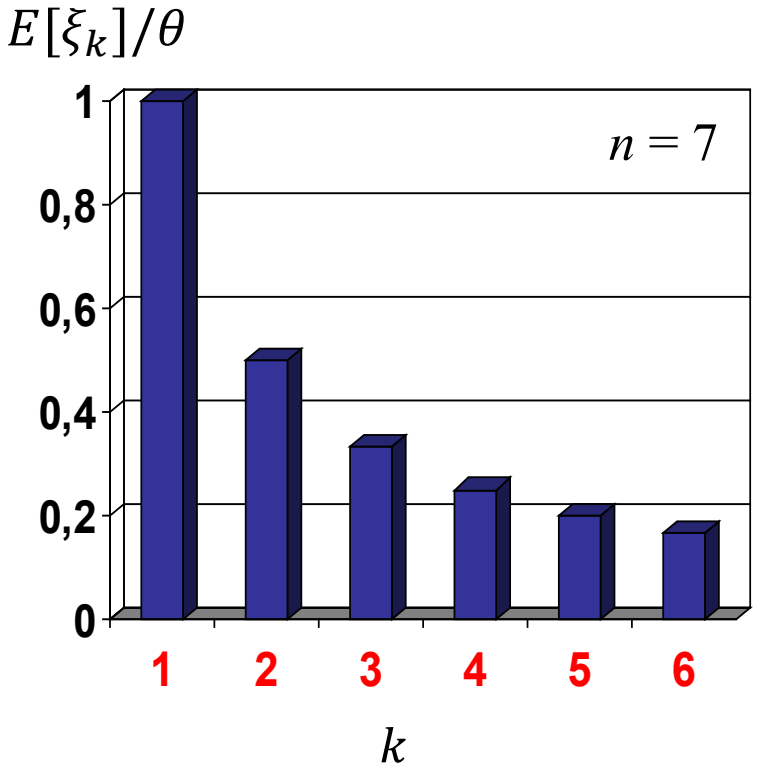
$$E[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \sum_{k=1}^{n-1} E[\xi_k]$$

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

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: $\hat{\theta}_s = \frac{n-1}{n} \underbrace{(\xi_1 + \xi_{n-1})}$ (extreme frequencies)

singletons of the folded spectrum