# The Coalescent II
## Beyond the Standard Model

Joachim Hermisson

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

# Coalescent Theory
# Beyond the Standard Neutral Model

Standard neutral model:

- 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

Process → Pattern

Neutral Null Model

Distributions e.g. SFS

Reject Null ← Data

*Problem: Nature is not a toy model*

# Coalescent Theory
## Beyond the Standard Neutral Model

Standard neutral model:

- 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~~

What happens for a different offspring distribution (mean 1, variance $\sigma^2$)?

Coaleasence probability per sequence pair if parent $i$ has $k_i$ offspring:

$$p_{c,1} = E\left[\sum_{i=1}^{2N} \frac{k_i(k_i-1)}{2N(2N-1)}\right] = \frac{E[k_1^2 - k_1]}{2N-1} = \frac{\sigma^2}{2N-1}$$

# Coalescent Theory
## Beyond the Standard Neutral Model

Concept of the effective population size $N_e$:

Coaleasence probability per sequence pair:

$$p_{c,1} = \frac{\sigma^2}{2N-1} \equiv \frac{1}{2N_e} \quad ; \quad N_e = \frac{N-1/2}{\sigma^2} \approx \frac{N}{\sigma^2}$$

$N_e$ : *“coalescence effective population size“*
population size of a standard Wright-Fisher model with the
*same coalescence rate* = *same amount of drift*
as the original (non-standard) model

# Coalescent Theory
## Beyond the Standard Neutral Model

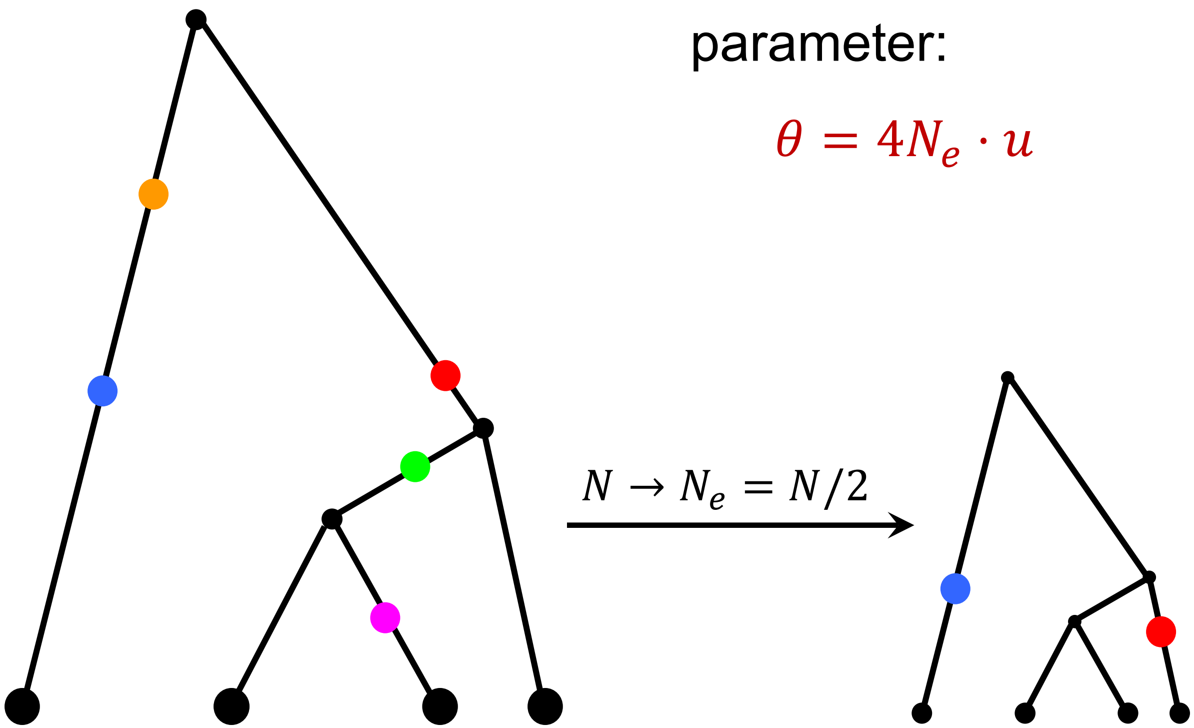
Concept of the effective population size $N_e$:

Corresponds to a rescaling of coalescent times and tree sizes:

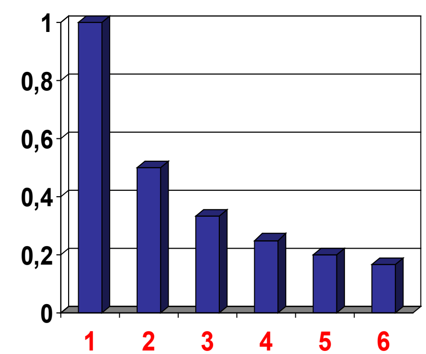
- rescaled mutation parameter:

$$\theta = 4N_e \cdot u$$

- unchanged shape of the site frequency spectrum:




$E[\xi_k]/\theta$



# Coalescent Theory
## Beyond the Standard Neutral Model

Sex dependance of offspring variance:

Consider: apes in zoo, 20♀, 20♂, but only one ♂ breeds $N_e = ?$

In general: breeding females and males: $N_f, N_m$

$$p_c = \frac{1}{4}\frac{1}{2N_f} + \frac{1}{4}\frac{1}{2N_m} = \frac{1}{8}\left(\frac{1}{N_f} + \frac{1}{N_m}\right) \stackrel{\text{def}}{=} \frac{1}{2N_e}$$

$\frac{1}{4}$: both from ♀

$\frac{1}{2N_f}$: both from same gene in ♀

$$\Longrightarrow N_e = \frac{4N_f N_m}{N_f + N_m} \leq N_f + N_m$$

E.g., apes in zoo:

$$N_e = \frac{4 \cdot 20 \cdot 1}{20 + 1} \approx 3.8$$

# Coalescent Theory
## Beyond the Standard Neutral Model

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

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

- ~~Constant population size~~ → Variable population size $N(t)$

➢ time-dependent coalescence rates $p_{c,1}(t) = \frac{1}{2N(t)}$

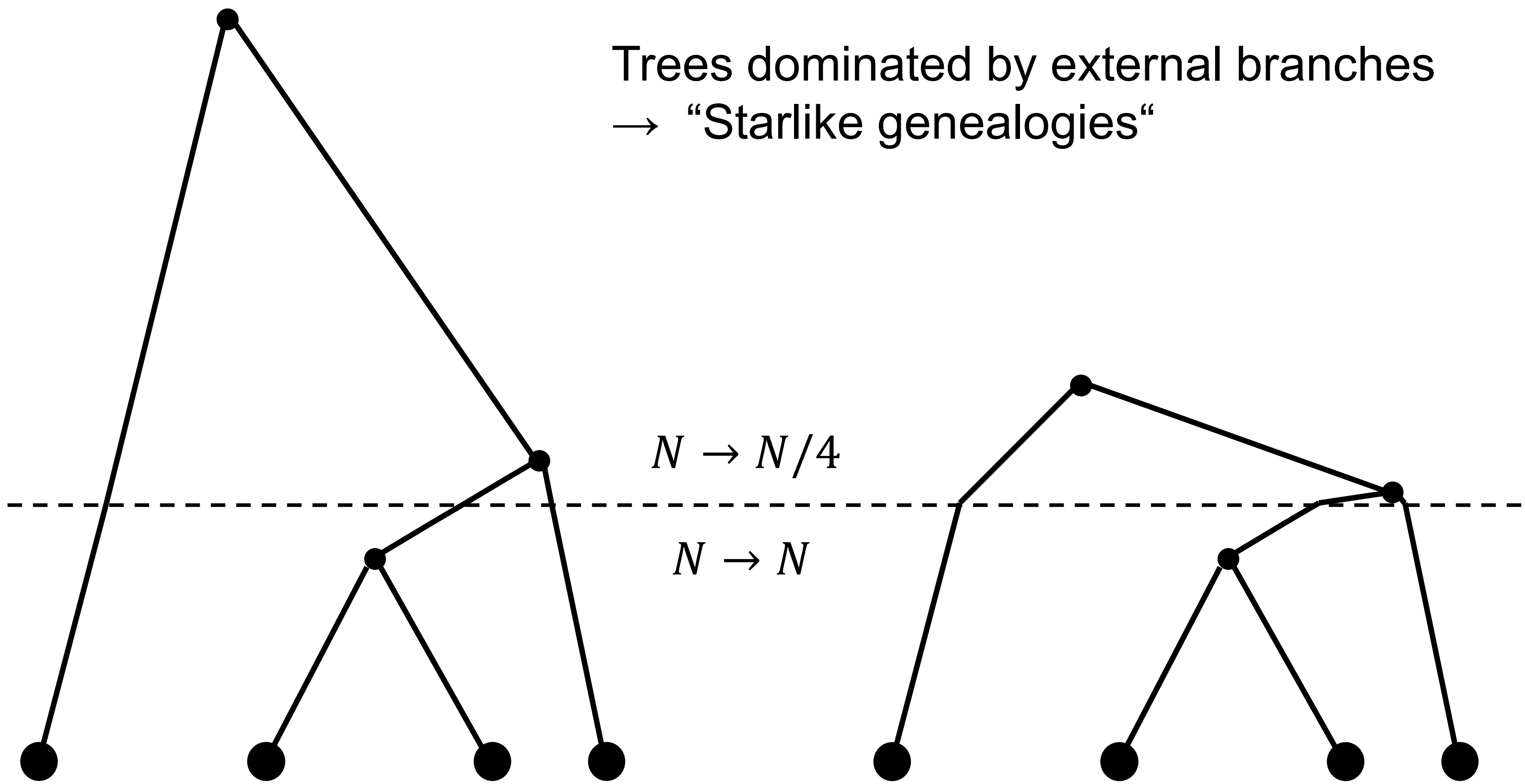
➢ can be accounted for by time-dependent rescaling of coalescent times and branch lengths

… while the distribution of topologies remains unchanged

# Coalescent Theory
## Beyond the Standard Neutral Model

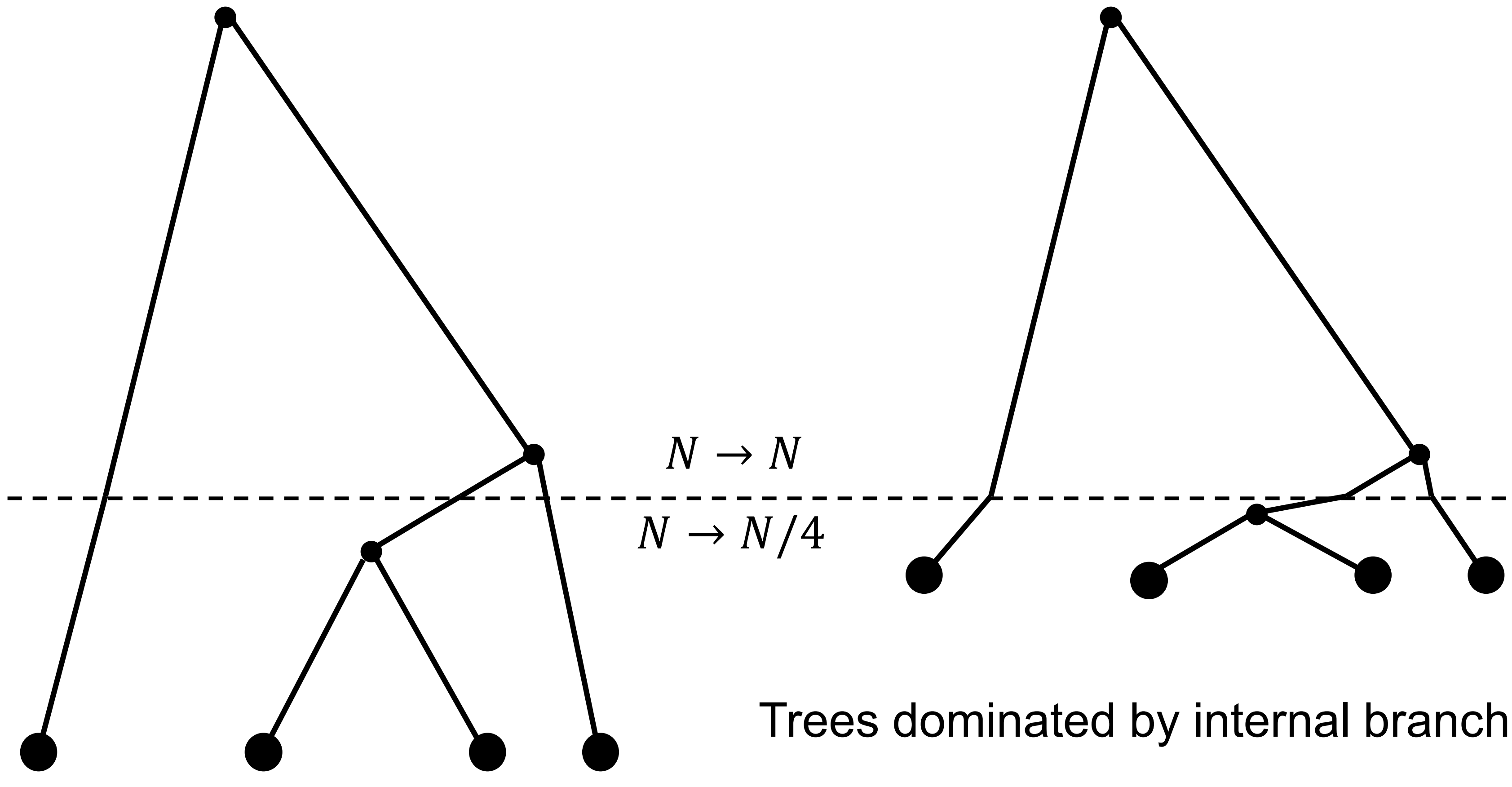
Growing populations (shrinking backward in time) :

Trees dominated by external branches
$\rightarrow$ “Starlike genealogies“



# Coalescent Theory
## Beyond the Standard Neutral Model

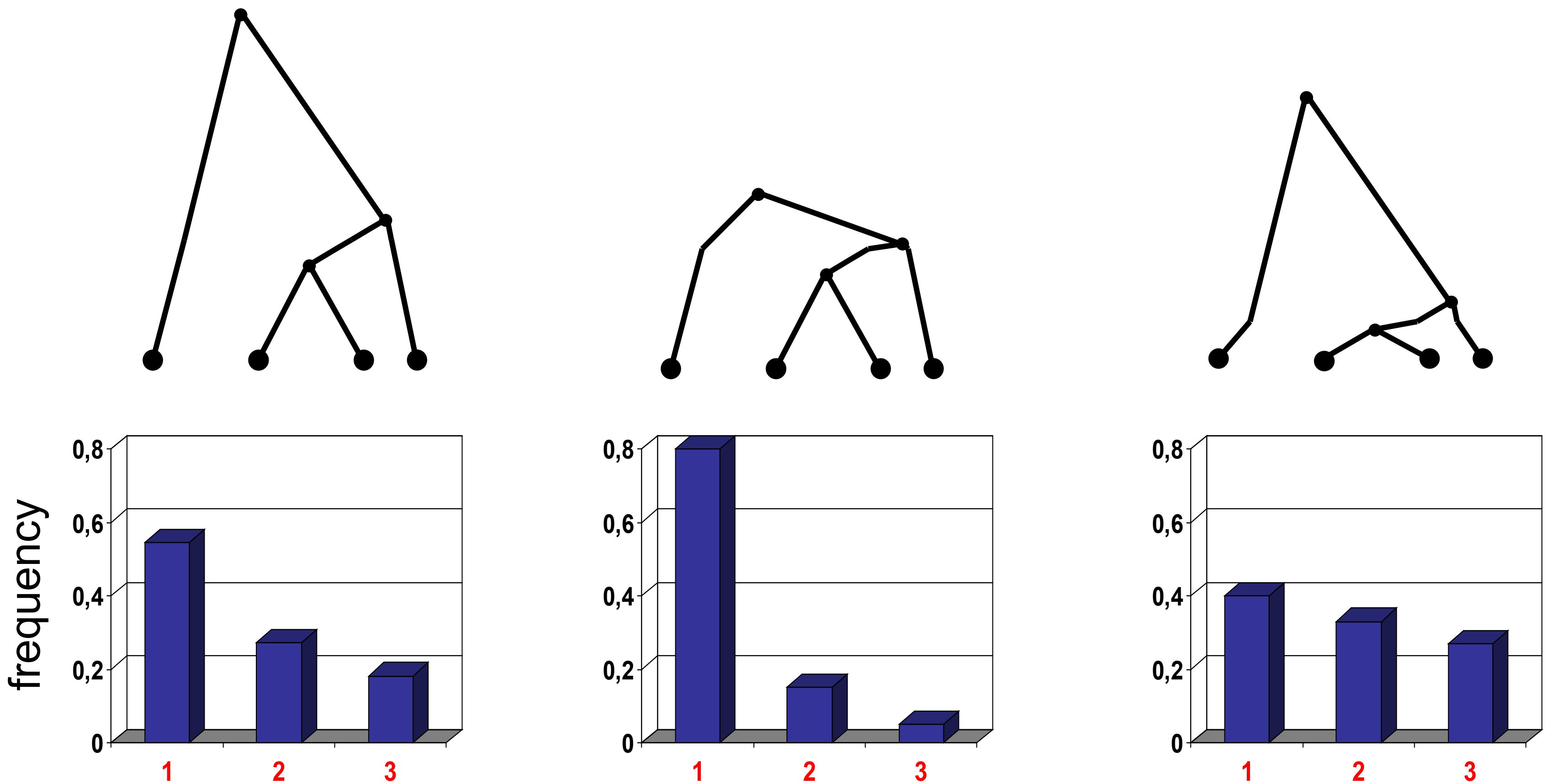
Shrinking populations (growing backward in time):




Trees dominated by internal branches

# Coalescent Theory
## Beyond the Standard Neutral Model

Expected frequency spectrum for growing/shrinking populations :



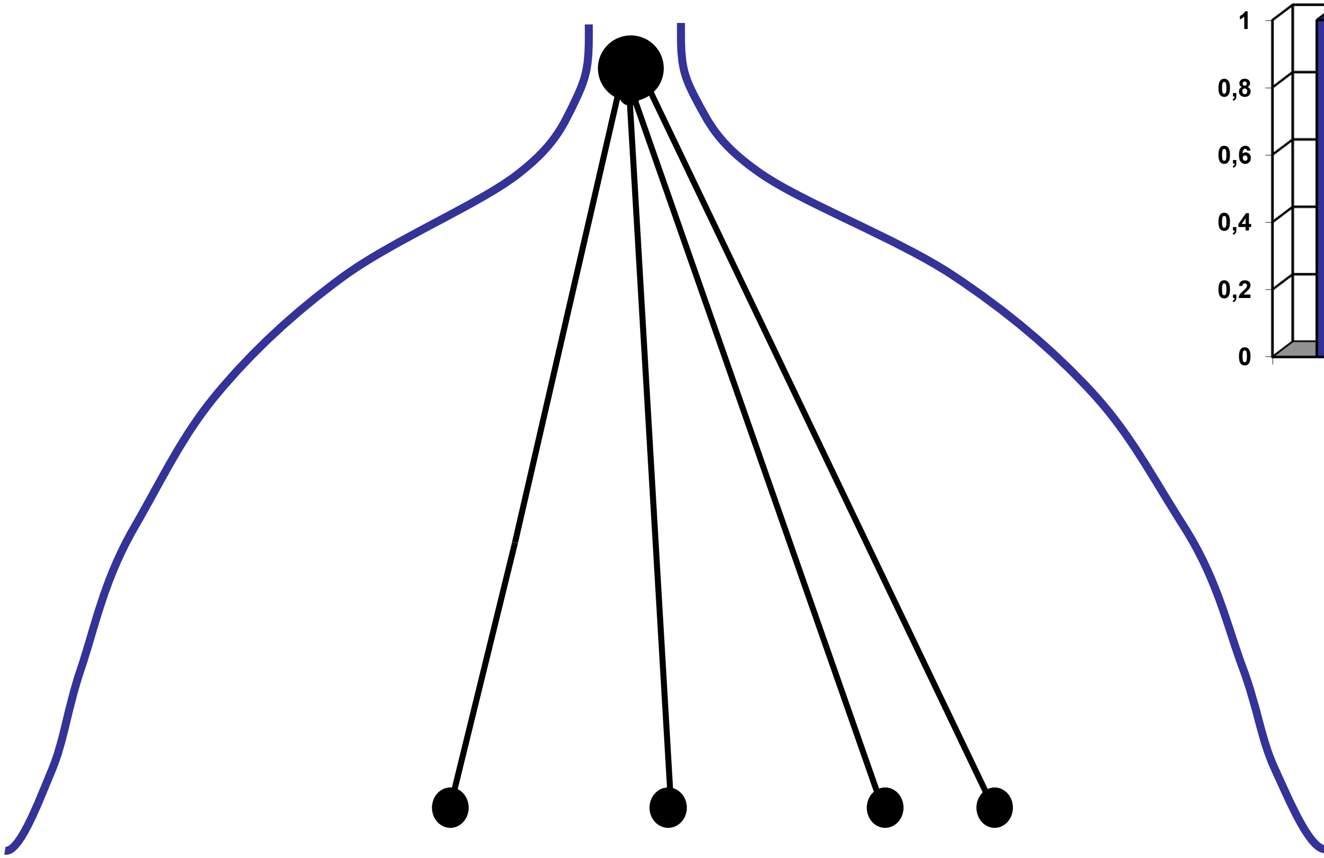

standard neutral

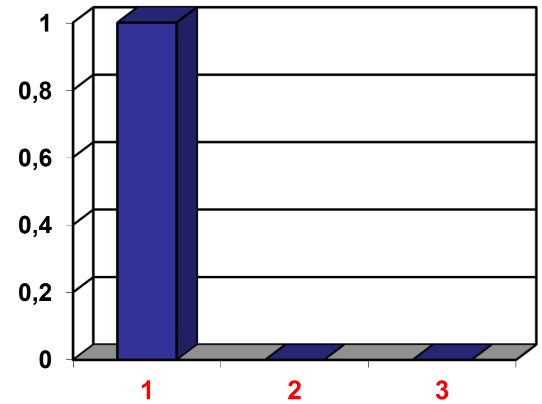
growth: excess of low frequency polymorphism

decrease: reduced low frequency polymorphism

# Coalescent Theory
## Beyond the Standard Neutral Model

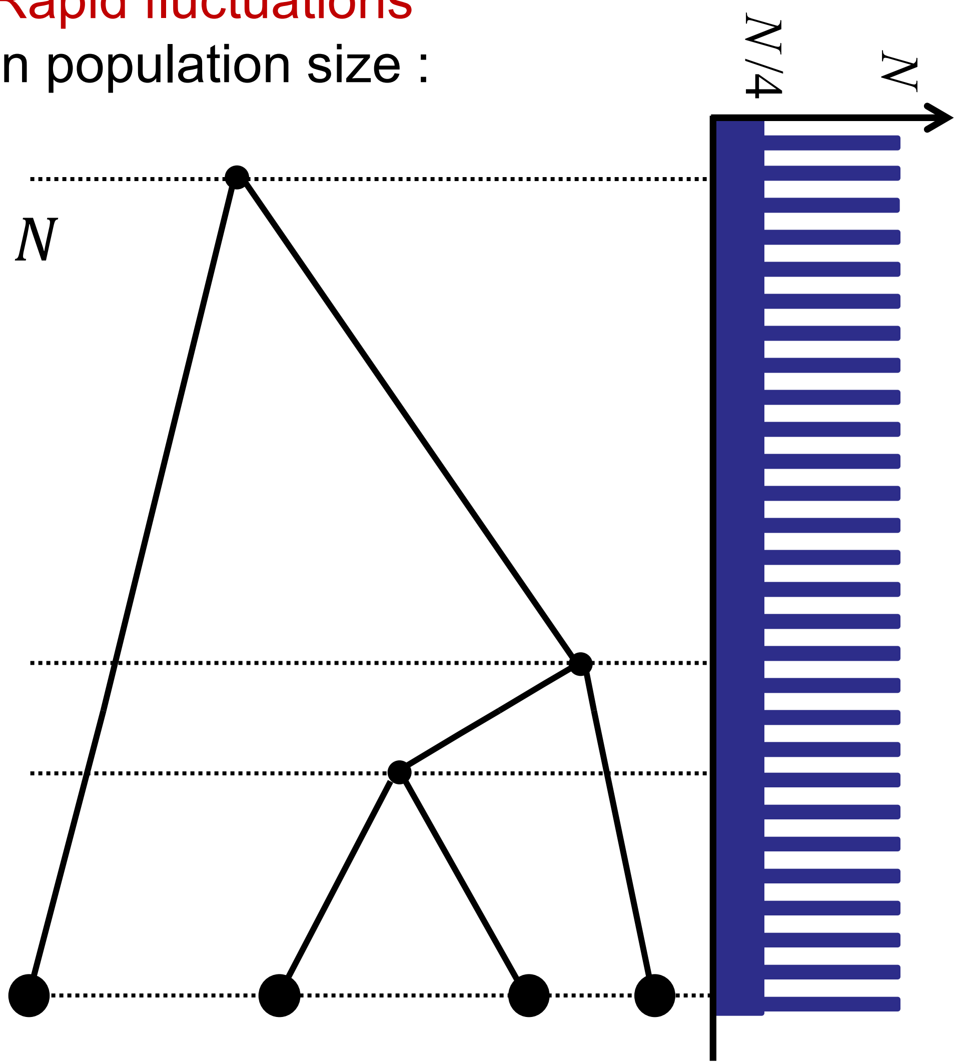
(Very) strong population growth: Starlike genealogy





# Coalescent Theory
## Beyond the Standard Neutral Model

Rapid fluctuations
in population size :




$$\bar{p}_{c,1} \approx \frac{1}{2}\left(\frac{1}{2N} + \frac{1}{2N/4}\right)$$

$$\Rightarrow N_e = \frac{1}{2\bar{p}_{c,1}} = \frac{2}{5}N$$

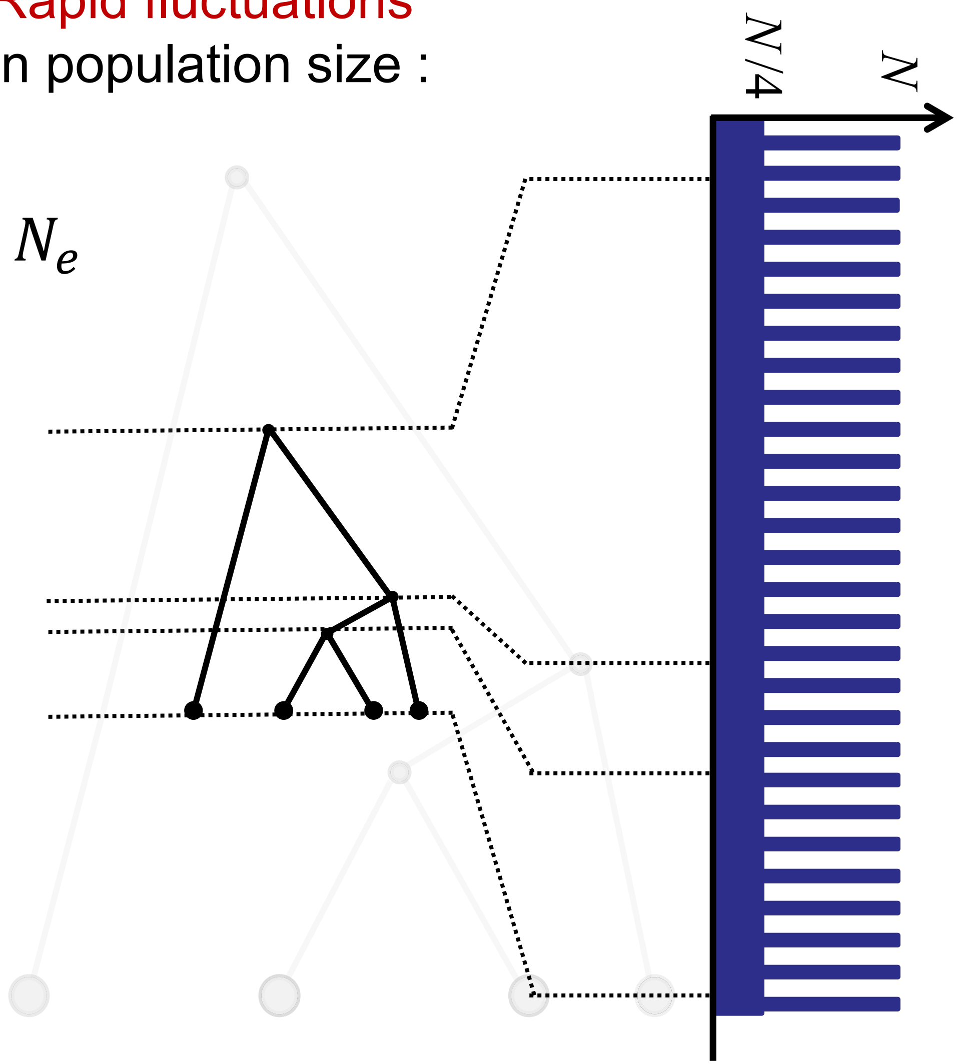
in general for period $T$ :

$$N_e = \left(\frac{1}{T}\sum_{i=1}^{T}\frac{1}{N_i}\right)^{-1}$$

harmonic mean population size

# Coalescent Theory
## Beyond the Standard Neutral Model

Rapid fluctuations
in population size :




$$\bar{p}_{c,1} \approx \frac{1}{2}\left(\frac{1}{2N} + \frac{1}{2N/4}\right)$$

$$\Rightarrow N_e = \frac{1}{2\bar{p}_{c,1}} = \frac{2}{5}N$$

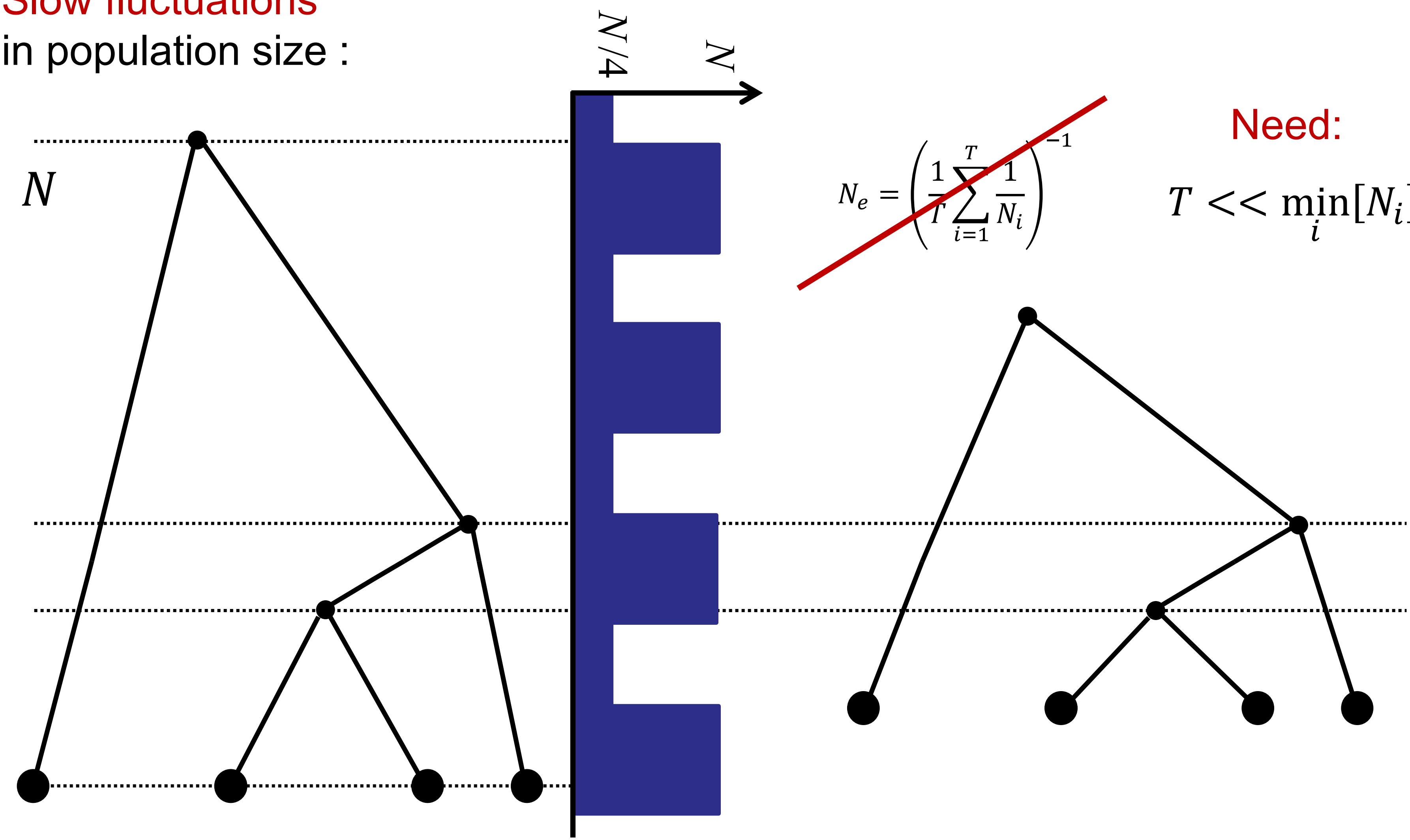
in general for period $T$ :

$$N_e = \left(\frac{1}{T}\sum_{i=1}^{T}\frac{1}{N_i}\right)^{-1}$$

harmonic mean population size

# Coalescent Theory
## Beyond the Standard Neutral Model

Slow fluctuations
in population size :




~~$N_e = \left( \frac{1}{T} \sum_{i=1}^{T} \frac{1}{N_i} \right)^{-1}$~~

Need:

$T << \min_i [N_i]$

# Coalescent Theory

## Beyond the Standard Neutral Model

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

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

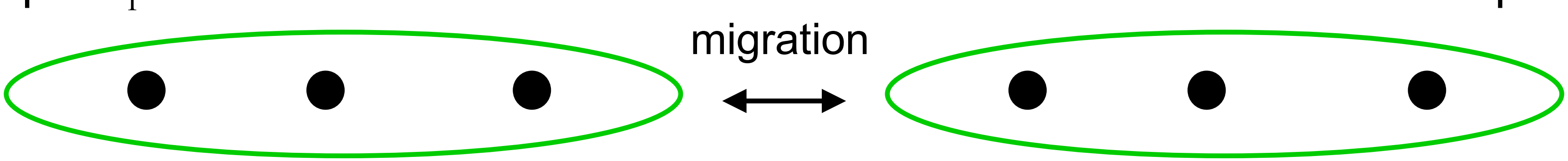
Structured population: e.g. two islands

Island 1

Island 2

size $N_1$
sample $n_1$

size $N_2$
sample $n_2$



# Coalescent Theory
## Beyond the Standard Neutral Model

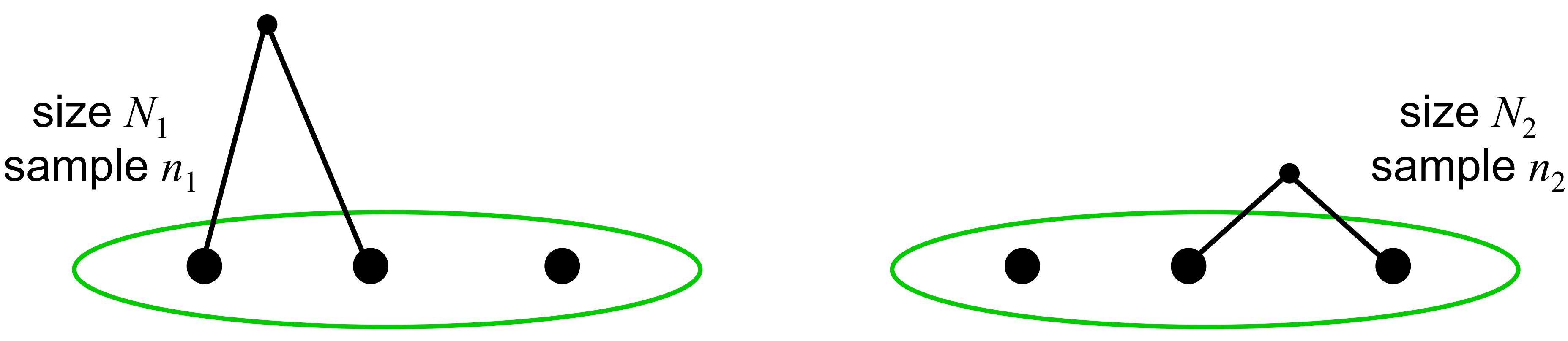
The structured coalescent: *two types of events*

coalescence on island 1

$$p_{c,1}^{(1)} = \binom{n_1}{2}\frac{1}{2N_1}$$

coalescence on island 2

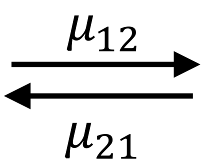
$$p_{c,1}^{(2)} = \binom{n_2}{2}\frac{1}{2N_2}$$



# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *two types of events*

**forward migration**

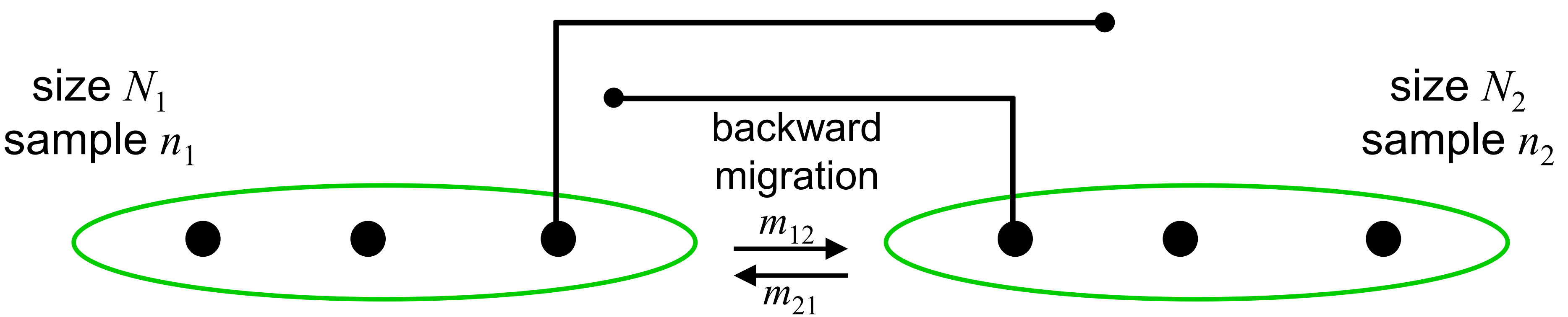



**backward migration**

$m_{ij}$: *proportion of indiv. in deme i that are new migrants from deme j*

$$m_{12} = \frac{N_2\mu_{21}}{N_2\mu_{21} + N_1(1-\mu_{12})}$$

$$m_{21} = \frac{N_1\mu_{12}}{N_1\mu_{12} + N_2(1-\mu_{21})}$$



# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *two types of events*

coalescence on island 1

$$p_{c,1}^{(1)} = \binom{n_1}{2}\frac{1}{2N_1}$$

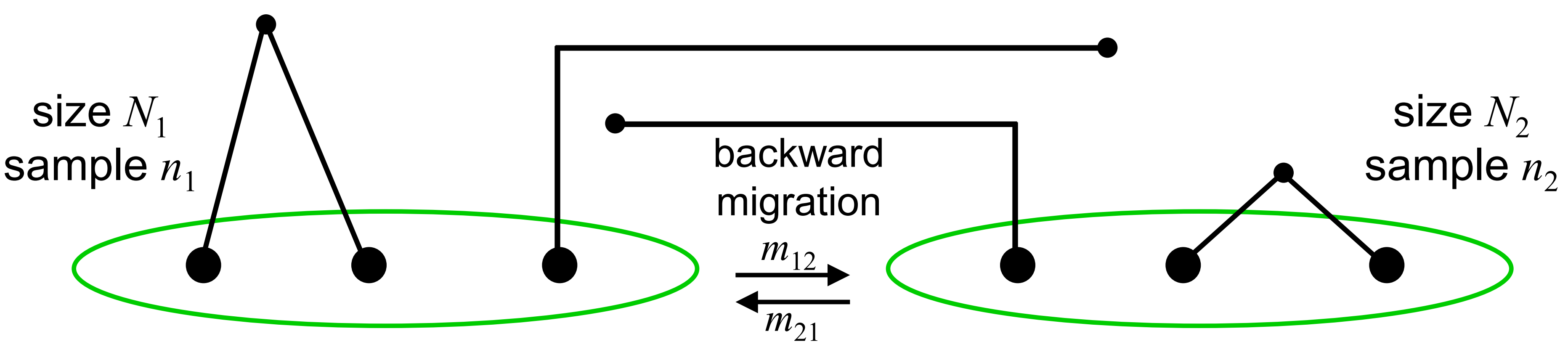
migration island 1 ↔ island 2

$$p_{m,1}^{(1\to 2)} = n_1 \cdot m_{12}$$

$$p_{m,1}^{(2\to 1)} = n_2 \cdot m_{21}$$

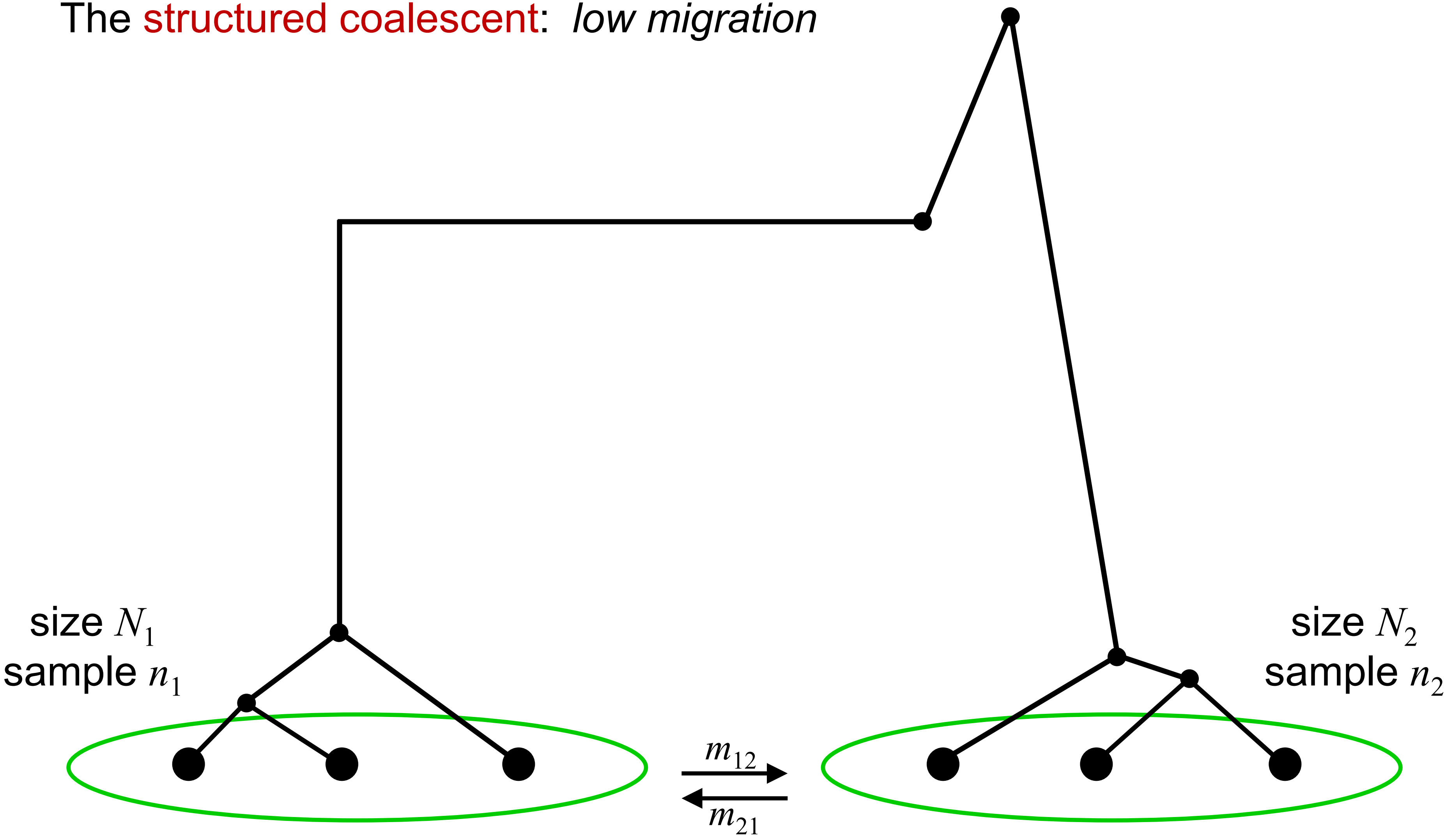
coalescence on island 2

$$p_{c,1}^{(2)} = \binom{n_2}{2}\frac{1}{2N_2}$$



# Coalescent Theory
## Beyond the Standard Neutral Model

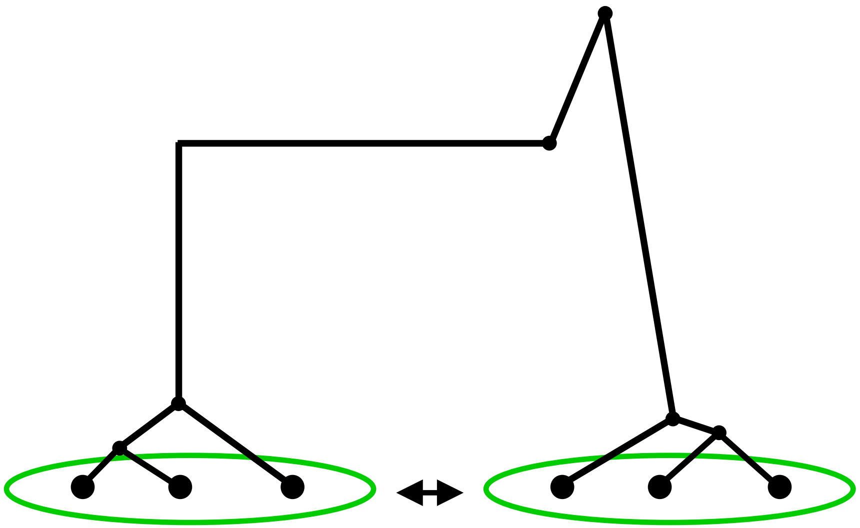
The structured coalescent: *low migration*

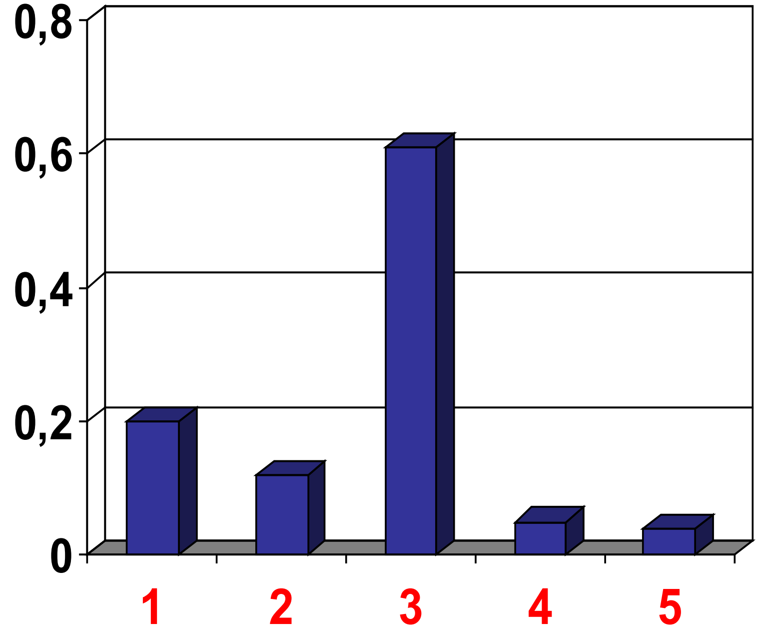


# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *low migration*

- the expected frequency spectrum



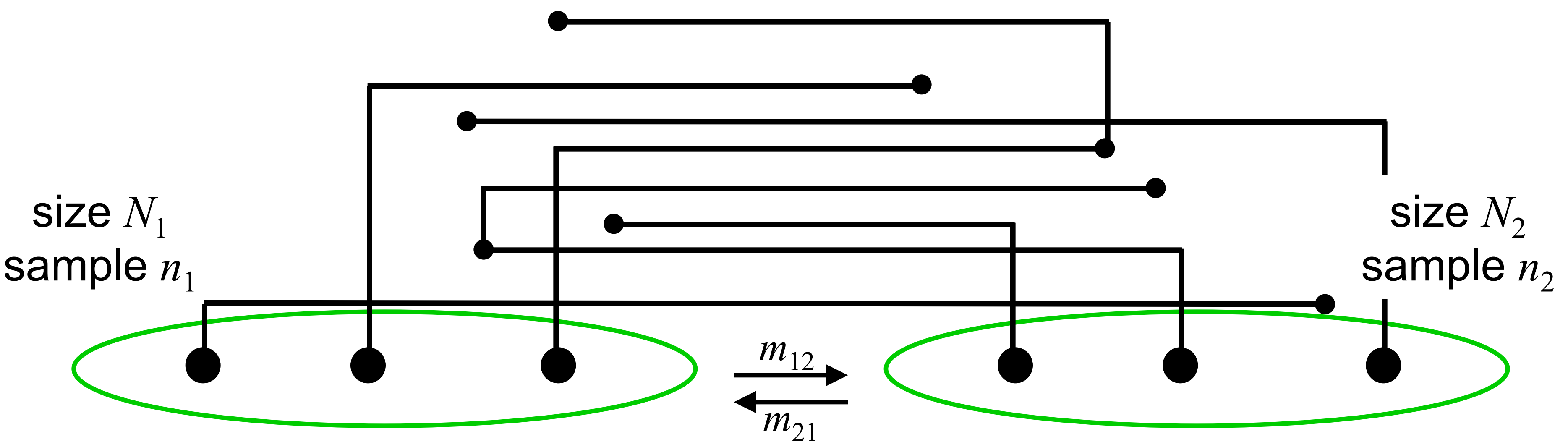


# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *strong migration* $m_{ij}N_i >> 1$

- First step: coalescent reaches migration equilibrium

$$p_1 = \frac{m_{21}}{m_{12} + m_{21}} \qquad p_2 = \frac{m_{12}}{m_{12} + m_{21}}$$

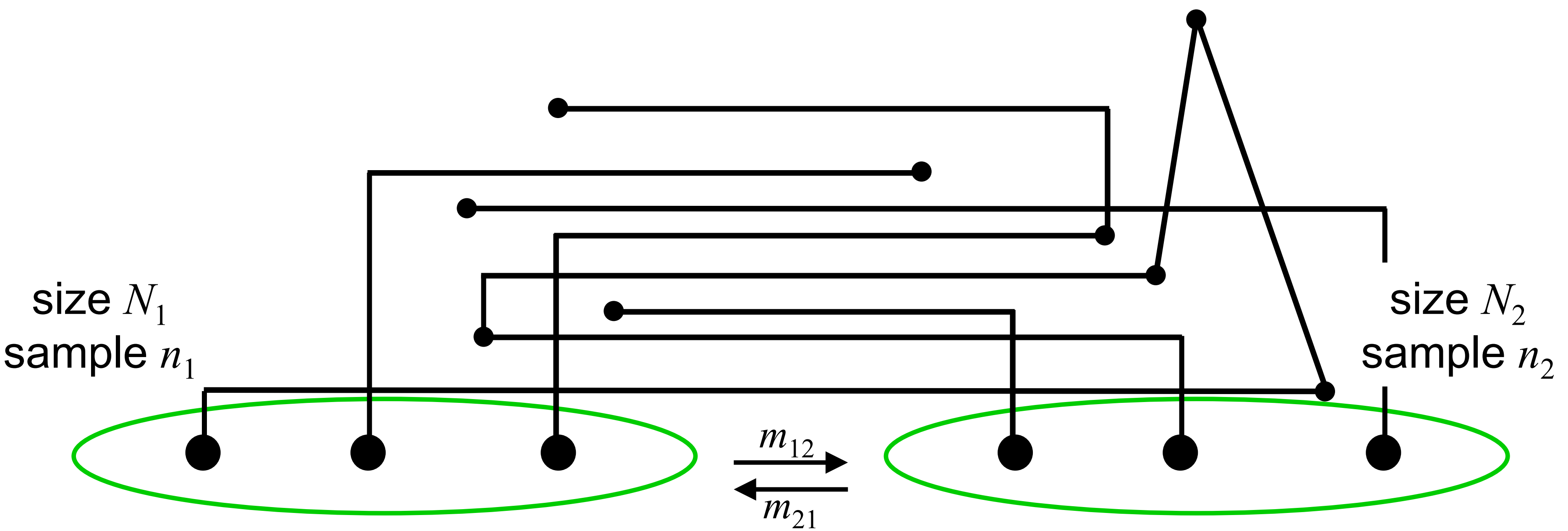


# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *strong migration* $m_{ij}N_i >> 1$

- Second step: coalescence probability in equilibrium

$$p_{c,1} = \binom{n}{2}\left(\frac{p_1^2}{2N_1} + \frac{p_2^2}{2N_2}\right)$$



# Coalescent Theory
## Beyond the Standard Neutral Model

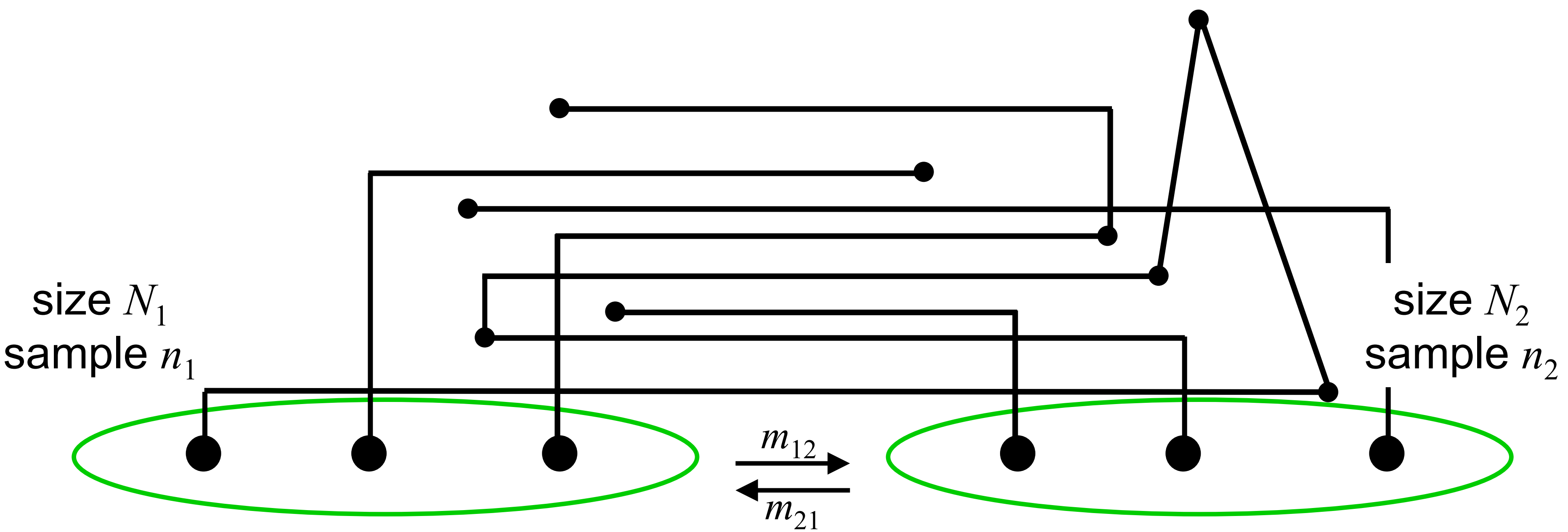
The structured coalescent: *strong migration* $m_{ij} N_i >> 1$

- In general: effective population size

$$p_{c,1} = \binom{n}{2} \cdot \sum_i \frac{p_i^2}{2N_i} \equiv \binom{n}{2} \cdot \frac{1}{2N_e}$$

(for $p_i = N_i/N$ :

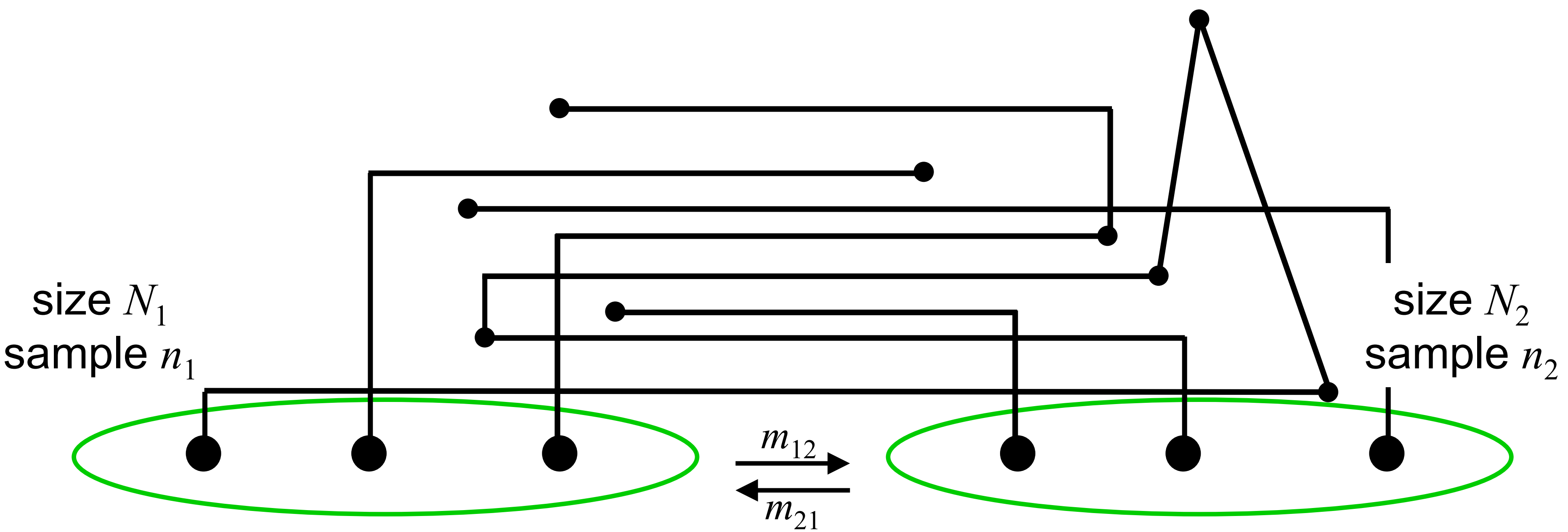
$$N_e = \sum_i N_i = N \,)$$



# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *strong migration* $m_{ij}N_i >> 1$

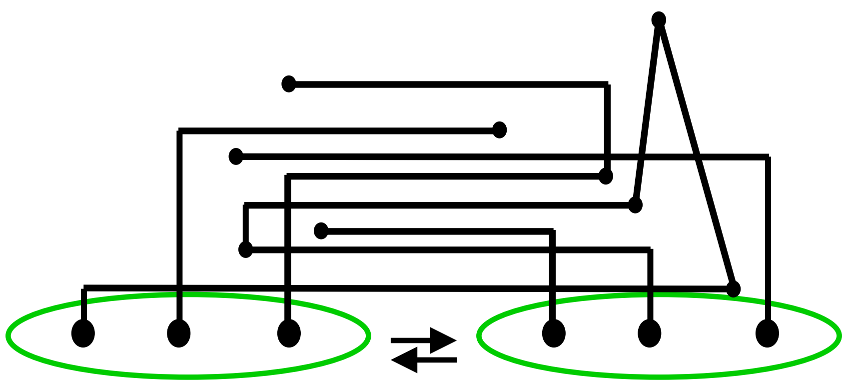
- Other kinds of structure:
  - diploid individuals (individuals = islands)
  - separate sexes (male and female “island“)
  - age structure, etc.

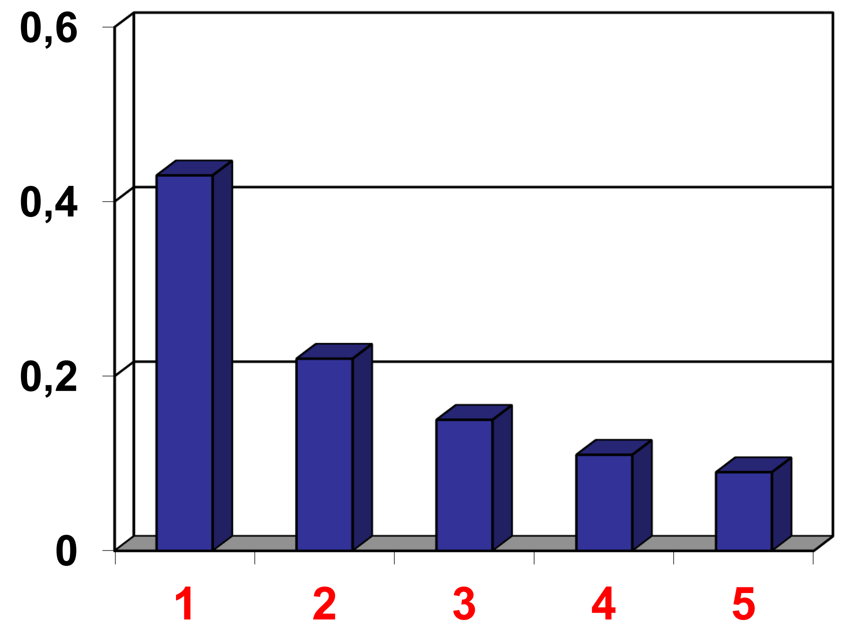


# Coalescent Theory
## Beyond the Standard Neutral Model

The structured coalescent: *strong migration* $m_{ij}N_i >> 1$

- The expected frequency spectrum






➢ Standard neutral spectrum (with effective population size)

# 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)

$\pi$-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)

# Coalescent Theory
## Test statistics

Test statistics for the deviation from neutrality:

Tajima's $D$:

$$D_T = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_W]}}$$

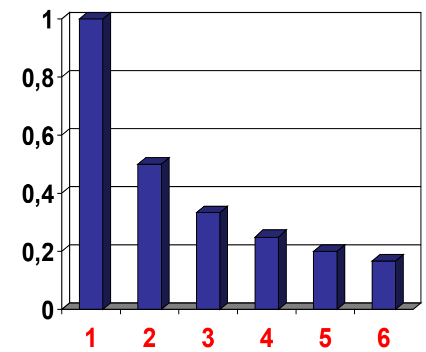
Fu and Li's $D$:

$$D_{FL} = \frac{\hat{\theta}_W - \hat{\theta}_S}{\sqrt{\mathrm{Var}[\hat{\theta}_W - \hat{\theta}_S]}}$$

Fay and Wu's $H$:

$$H_{FW} = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_H]}}$$

standard neutral evolution:




$$D_T = D_{FL} = H_{FW} = 0$$

# Coalescent Theory
## Test statistics

Test statistics for the deviation from neutrality:

Tajima's $D$:

$$D_T = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_W]}}$$

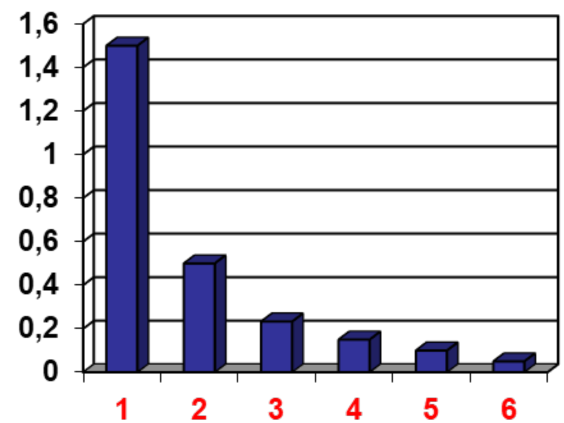
Fu and Li's $D$:

$$D_{FL} = \frac{\hat{\theta}_W - \hat{\theta}_S}{\sqrt{\mathrm{Var}[\hat{\theta}_W - \hat{\theta}_S]}}$$

Fay and Wu's $H$:

$$H_{FW} = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_H]}}$$

population growth:




$$D_T;\ D_{FL} < 0\ ;$$
$$H_{FW} > 0$$

# Coalescent Theory
## Test statistics

Test statistics for the deviation from neutrality:

Tajima‘s $D$:

$$D_T = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_W]}}$$

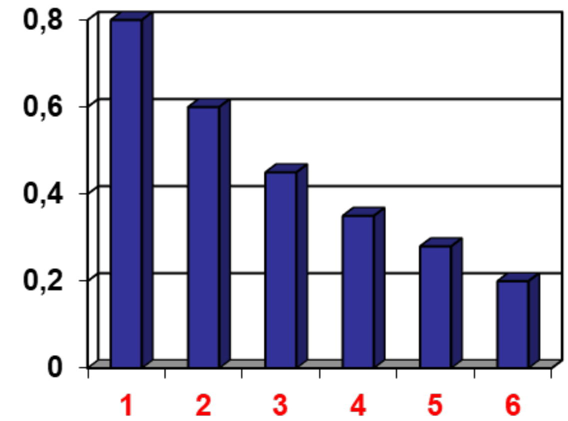
Fu and Li‘s $D$:

$$D_{FL} = \frac{\hat{\theta}_W - \hat{\theta}_S}{\sqrt{\mathrm{Var}[\hat{\theta}_W - \hat{\theta}_S]}}$$

Fay and Wu‘s $H$:

$$H_{FW} = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_H]}}$$

population decline:




$$D_T;\ D_{FL} > 0\ ;$$
$$H_{FW} < 0$$

# Coalescent Theory
## Test statistics

Test statistics for the deviation from neutrality:

Tajima's $D$:

$$D_T = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_W]}}$$

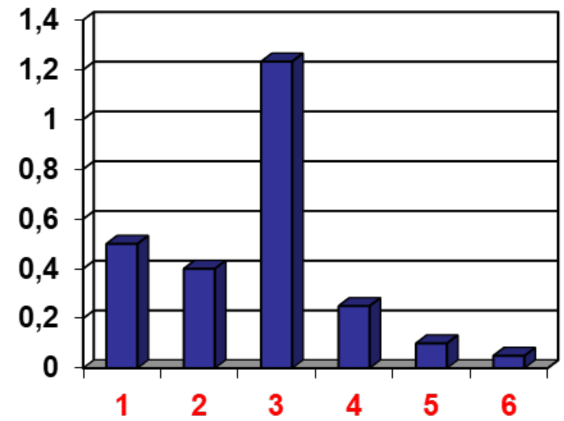
Fu and Li's $D$:

$$D_{FL} = \frac{\hat{\theta}_W - \hat{\theta}_S}{\sqrt{\mathrm{Var}[\hat{\theta}_W - \hat{\theta}_S]}}$$

Fay and Wu's $H$:

$$H_{FW} = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_H]}}$$

population structure:

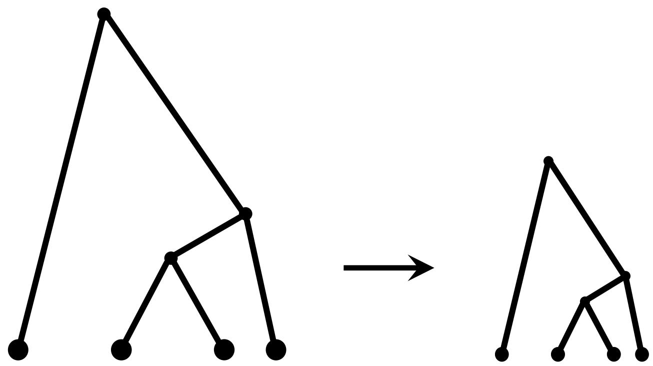



$$D_T;\ D_{FL} > 0\ ;$$
$$H_{FW} = \ ?$$

# Coalescent Theory
## Robustness of the Coalescent

Many biological details only lead to a rescaling of coalescence trees (changed effective population size):

- offspring distribution, breeding dominance
- diploids, separate sexes, …
- weak structure with rapid migration
- short-term fluctuations in population size



All these processes occur on much shorter time scales than coalescence and can therefore be averaged out. We say the coalescent is robust with respect to these details. In contrast, coalescence trees are affected by:

- “large“ demographic changes
- strong population structure
- selection

Consequence:
Only these phenomena can be detected from polymorphism data !