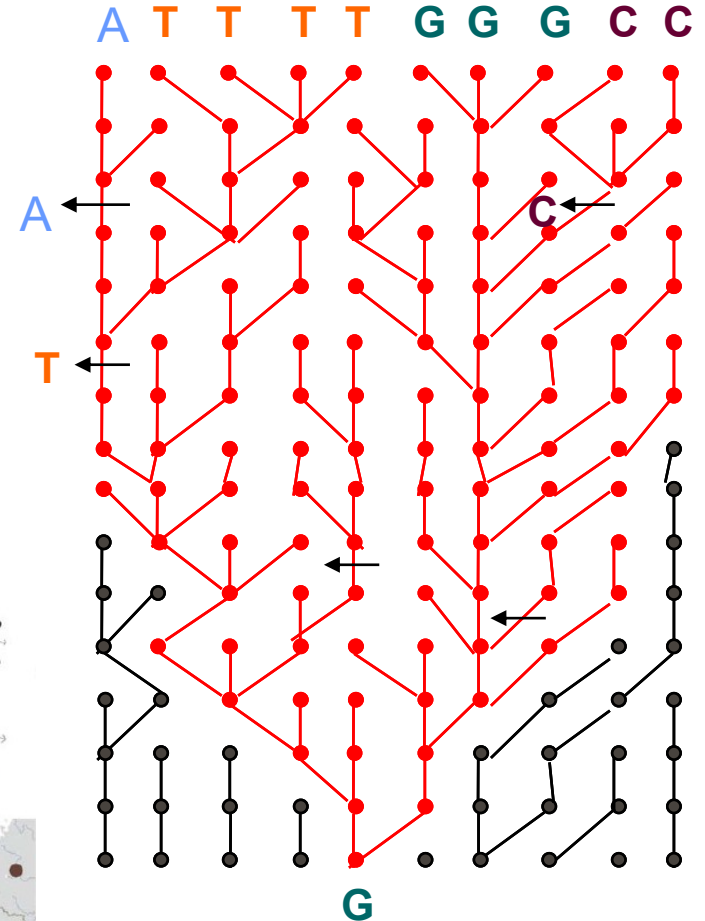
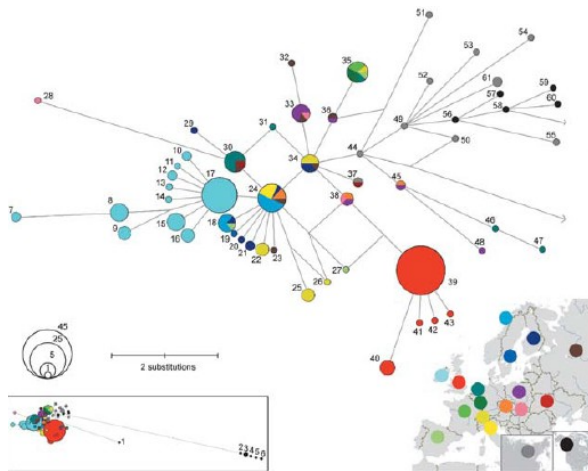
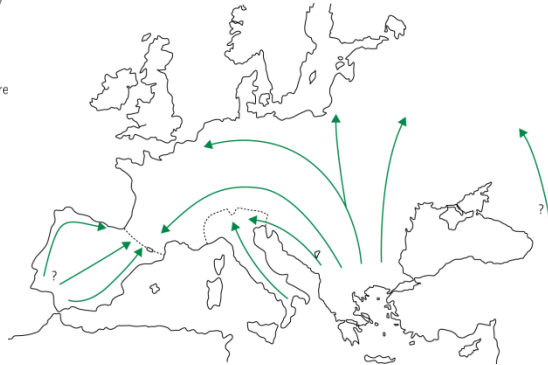
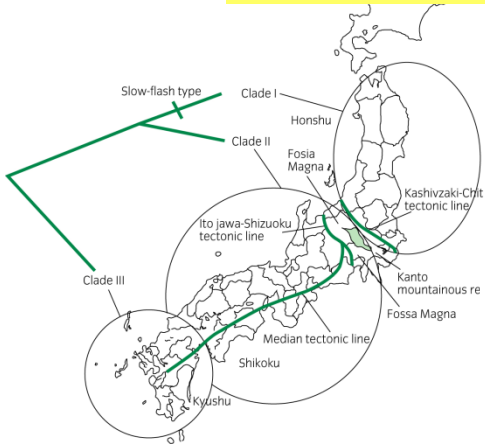
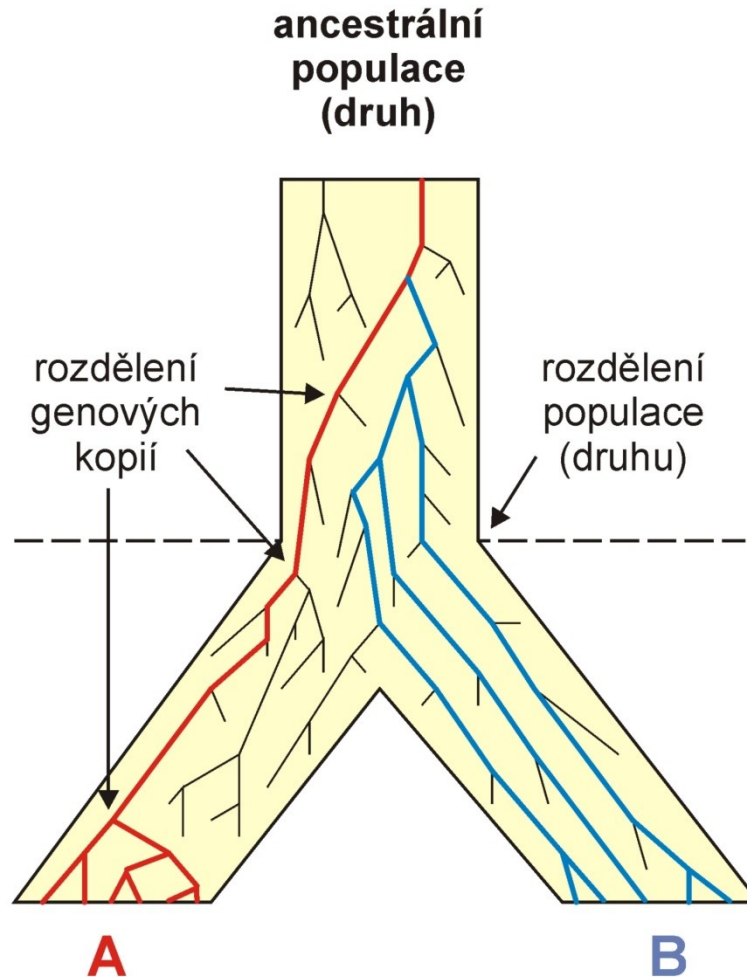


COALESCENT AND PHYLOGEOGRAPHY

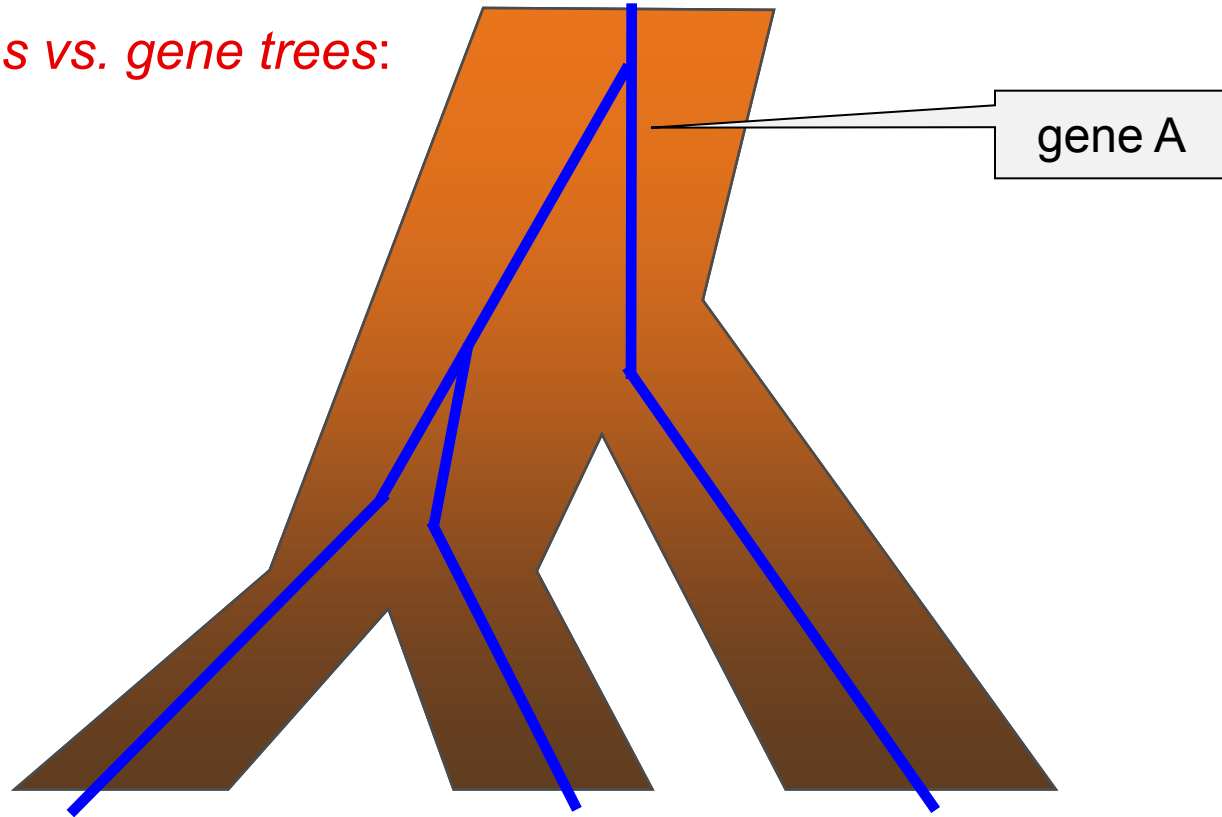


COALESCENCE

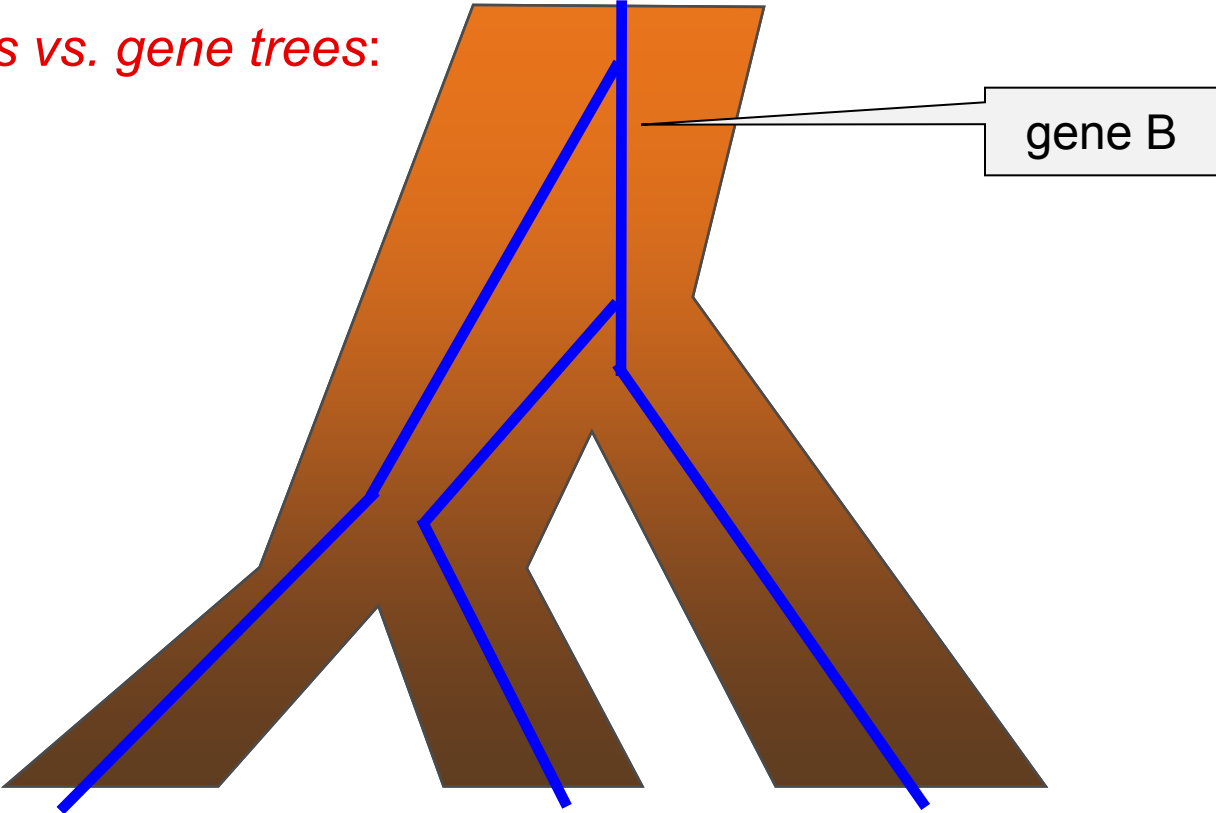
Fate of individual gene copies in the population → gene trees



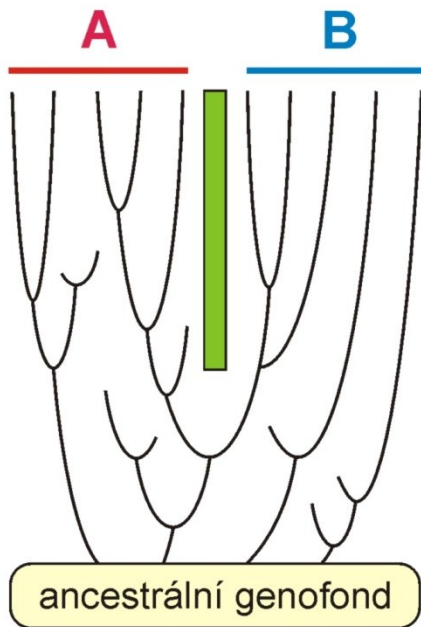
Species trees vs. gene trees:



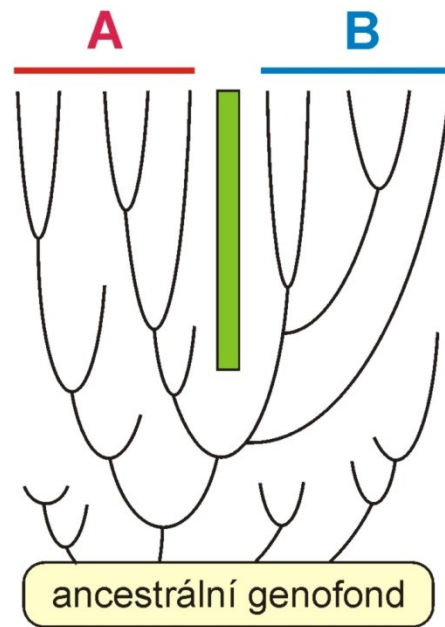
Species trees vs. gene trees:



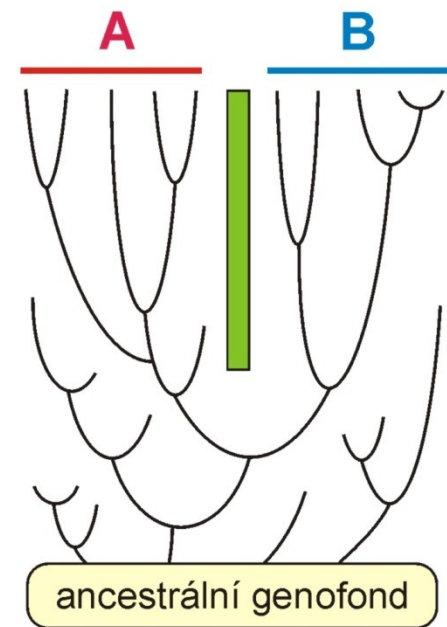
Phylogenetic relationships of two descendant populations (eg. mtDNA):



polyphyly

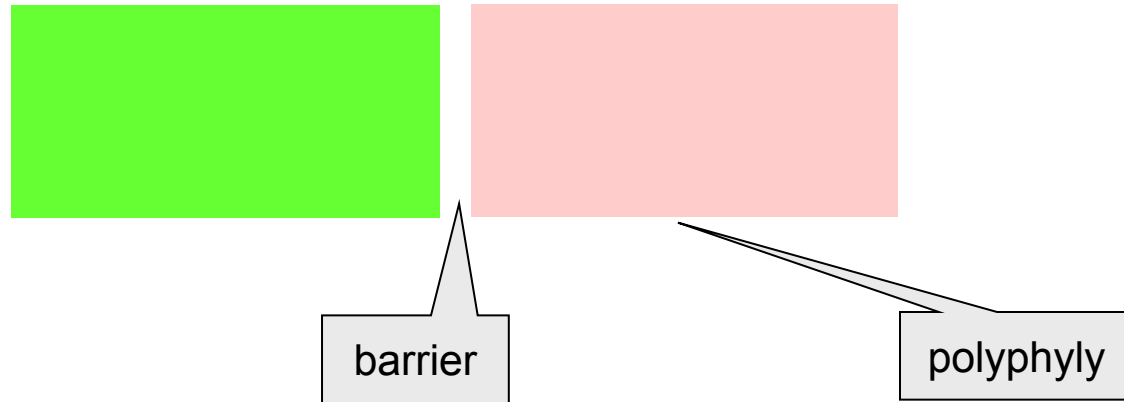


paraphyly

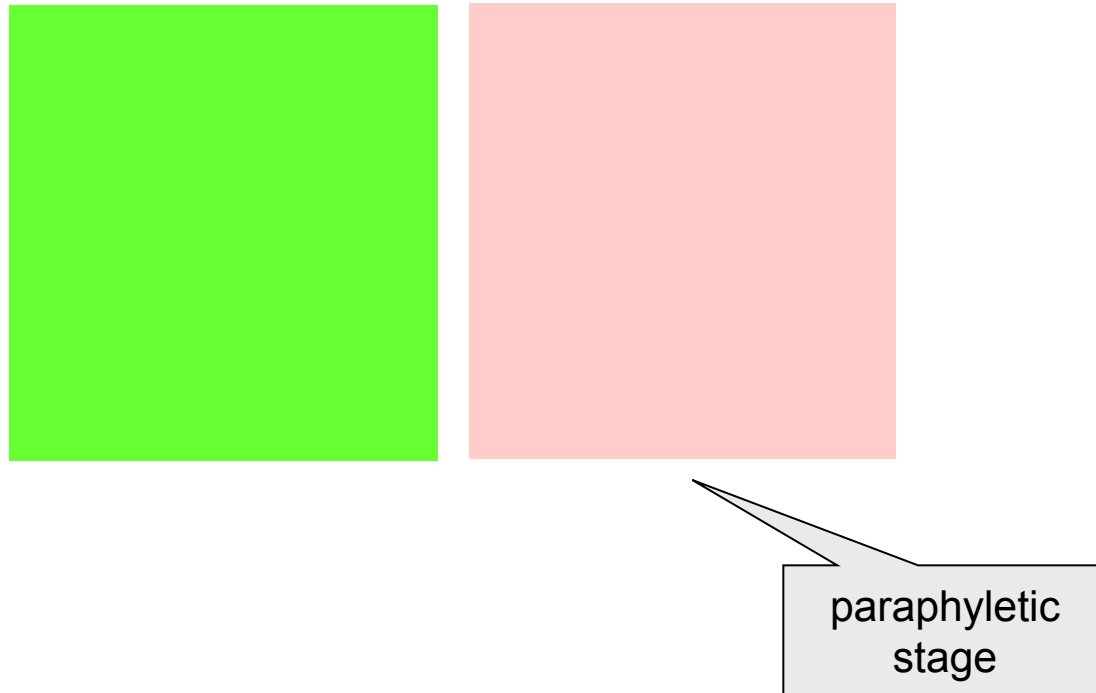


reciprocal monophyly

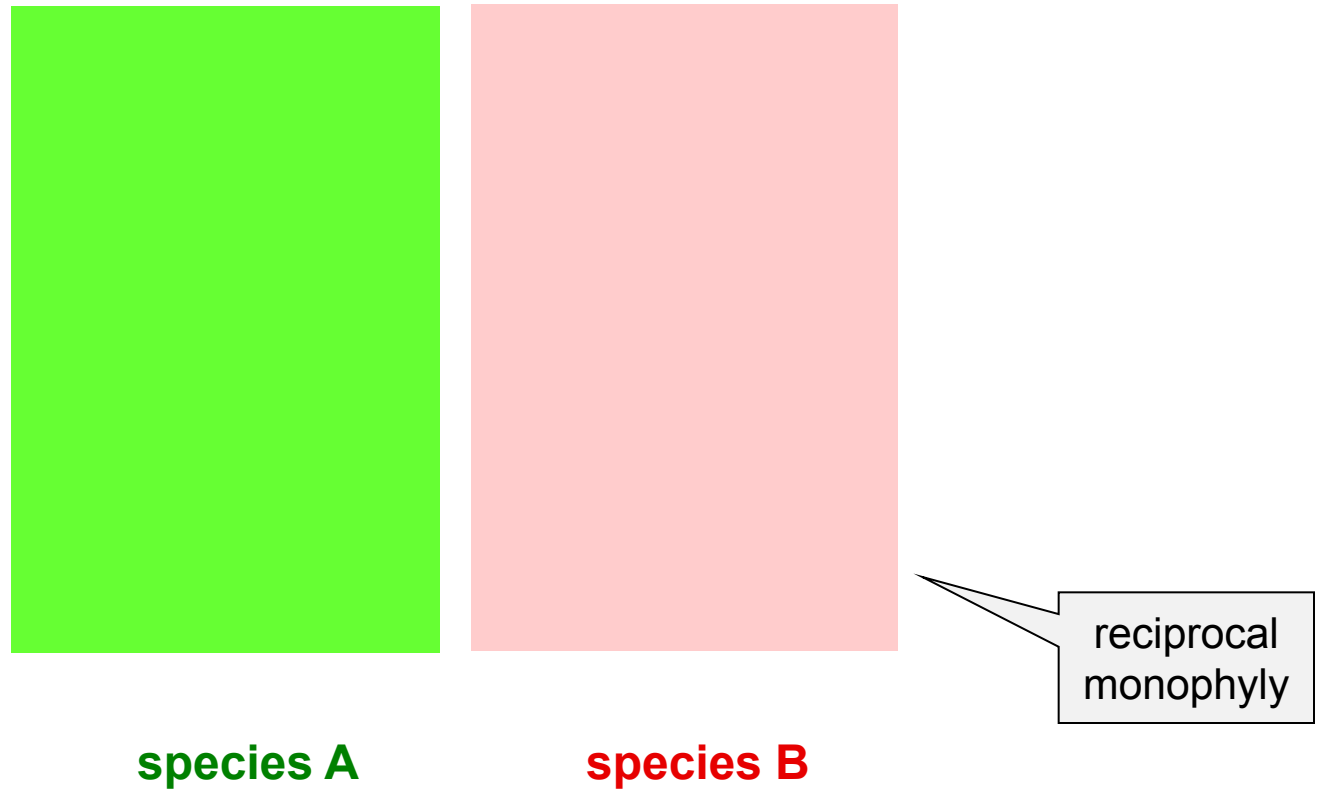
Ancestral polymorphism and lineage sorting



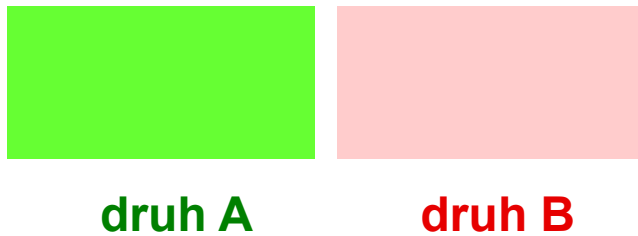
Ancestral polymorphism and lineage sorting



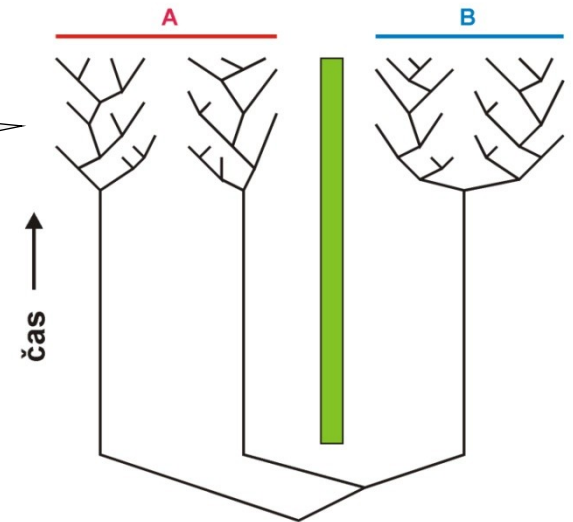
Ancestral polymorphism and lineage sorting



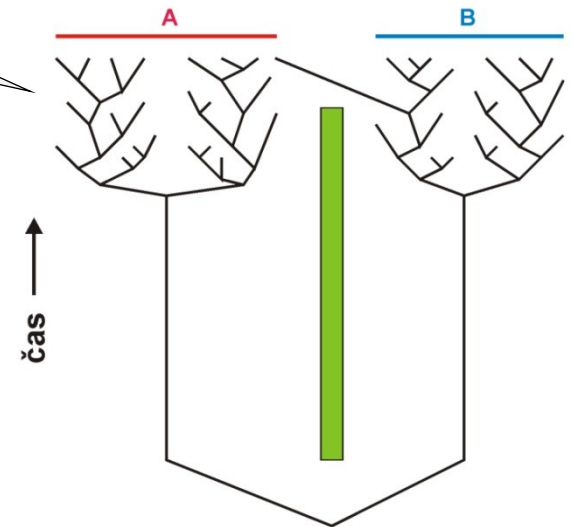
Ancestral polymorphism and lineage sorting



incomplete lineage sorting



recent gene flow



Problem: it is often difficult to distinguish between incomplete lineage sorting and consequences of gene flow

P



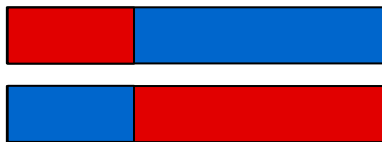
X



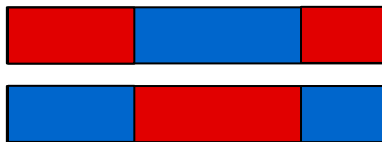
F1



F2

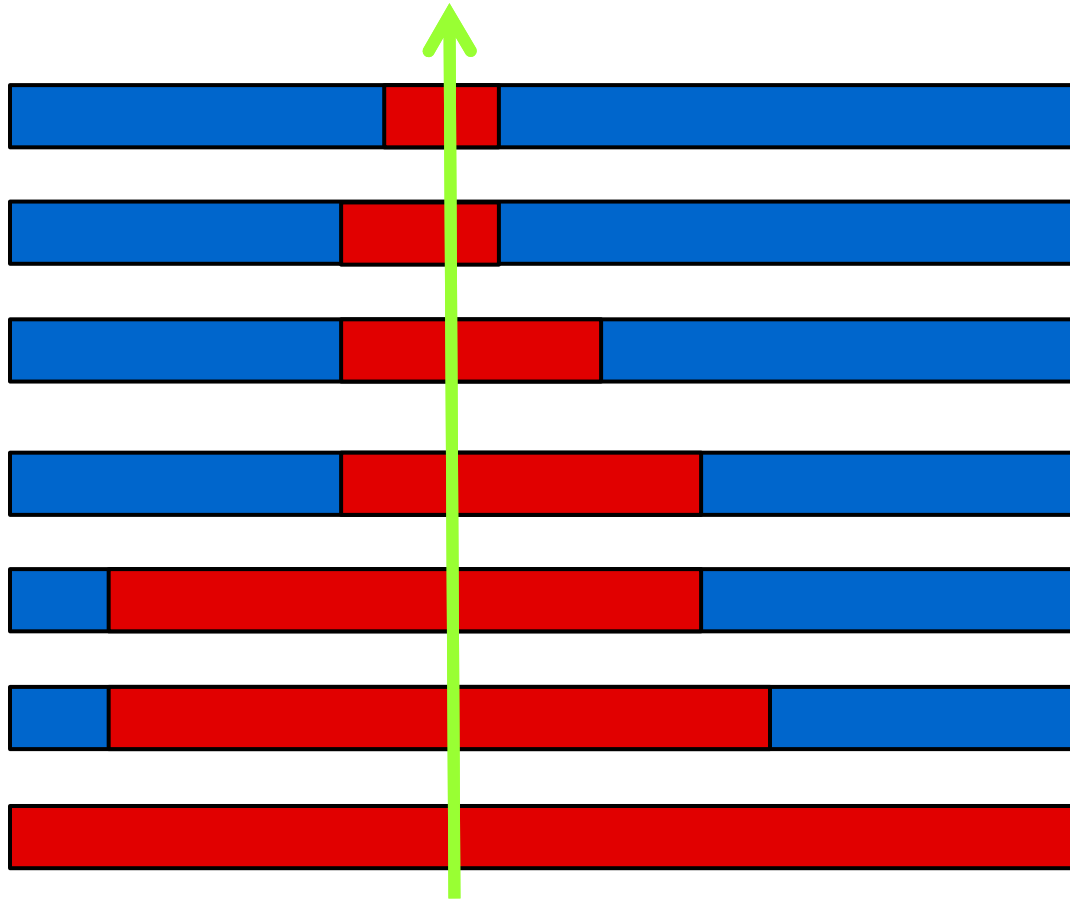


F3



⋮

Hybridization makes a cascade of blocks



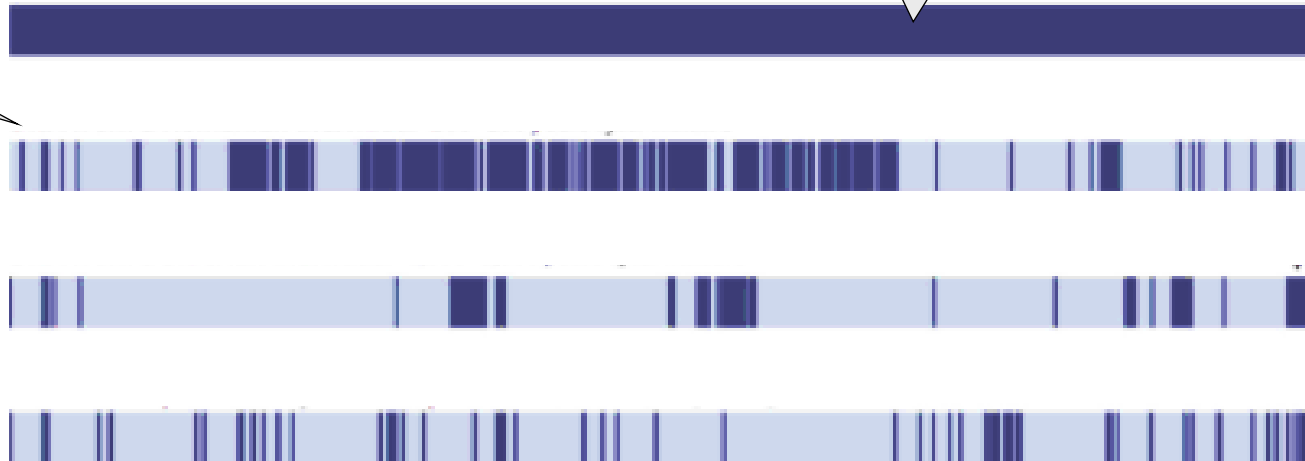
advancing introgression into 'blue' genome

Romania, ~40 kya,
mating before
200–100 years

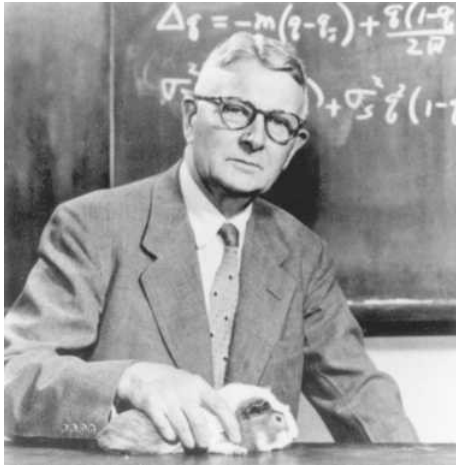
chromosome 12
of Neanderthal

Siberia, ~45 kya,
mating before
8000–5000

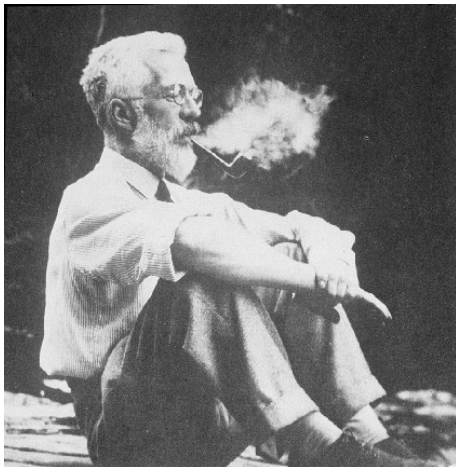
contemporary China,
54–49 kya



Wright-Fisher model:



Sewall Wright



Ronald A. Fisher

W-F population:

haploid or diploid-hermaphrodite

finite size, no fluctuations of N

random mating

complete isolation (no gene flow)

discrete generations

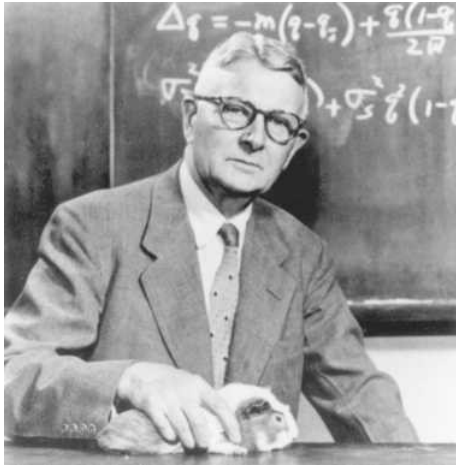
no age structure

no selection

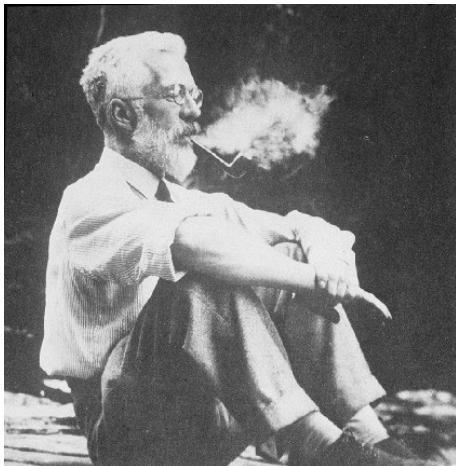
variance of gamete sampling

→ Poisson distribution

Lineage sorting in W-F model:



Sewall Wright



Ronald A. Fisher

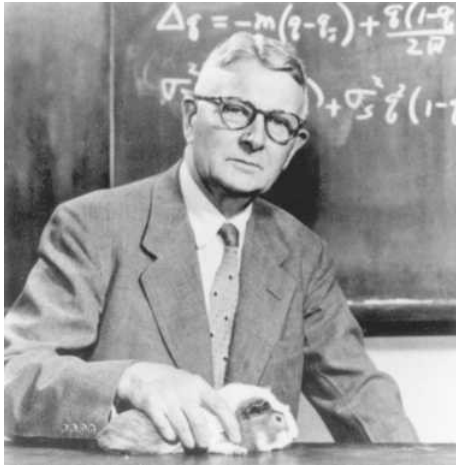
generace 1



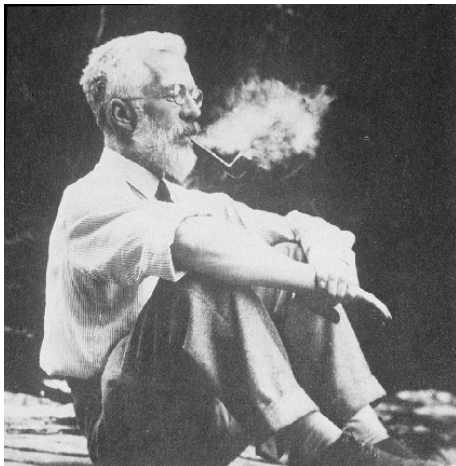
time



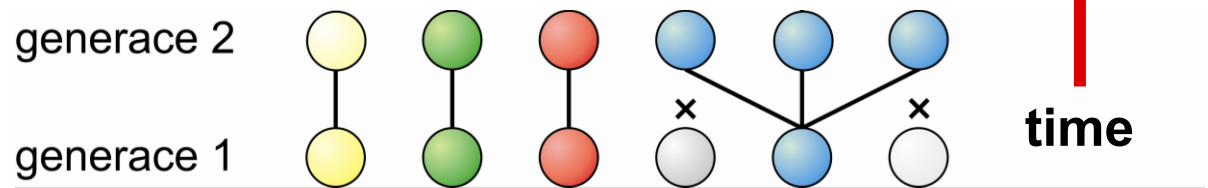
Lineage sorting in W-F model:



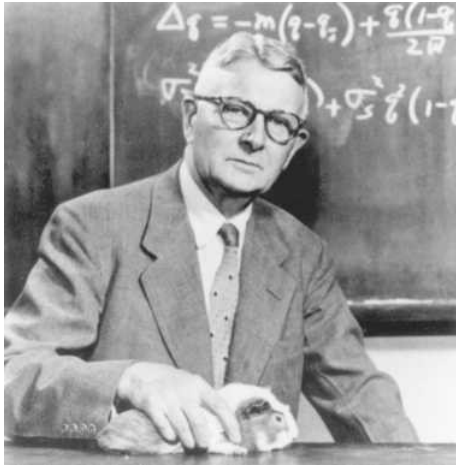
Sewall Wright



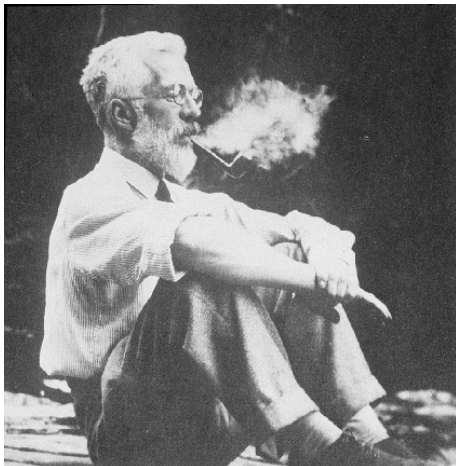
Ronald A. Fisher



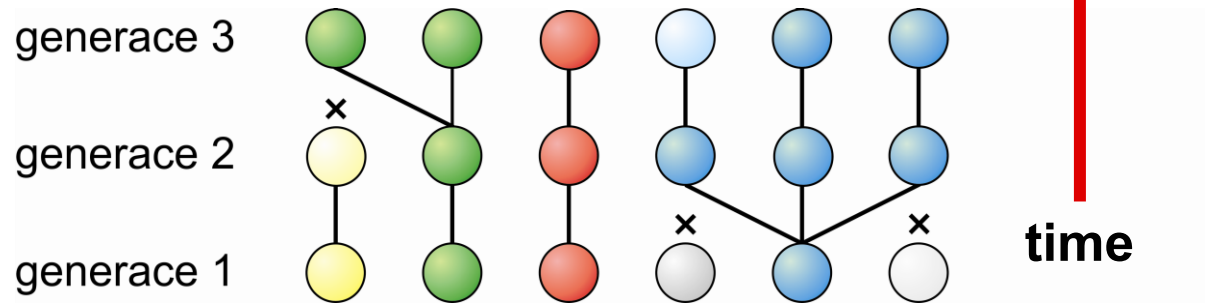
Lineage sorting in W-F model:



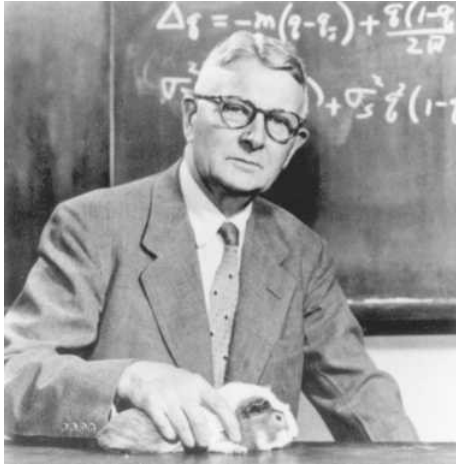
Sewall Wright



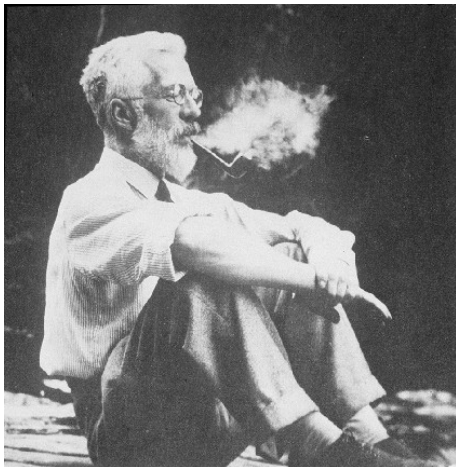
Ronald A. Fisher



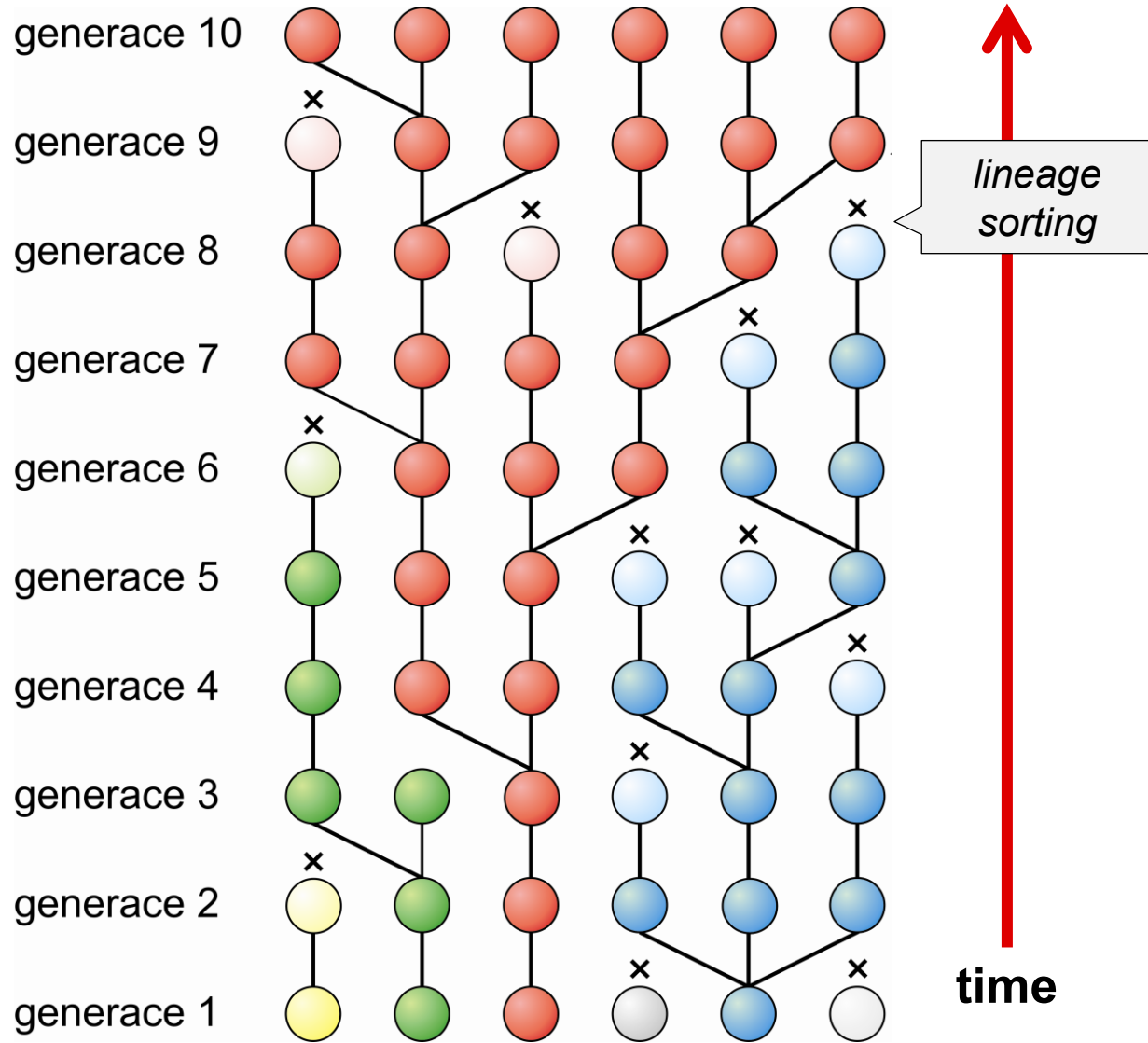
Lineage sorting in W-F model:

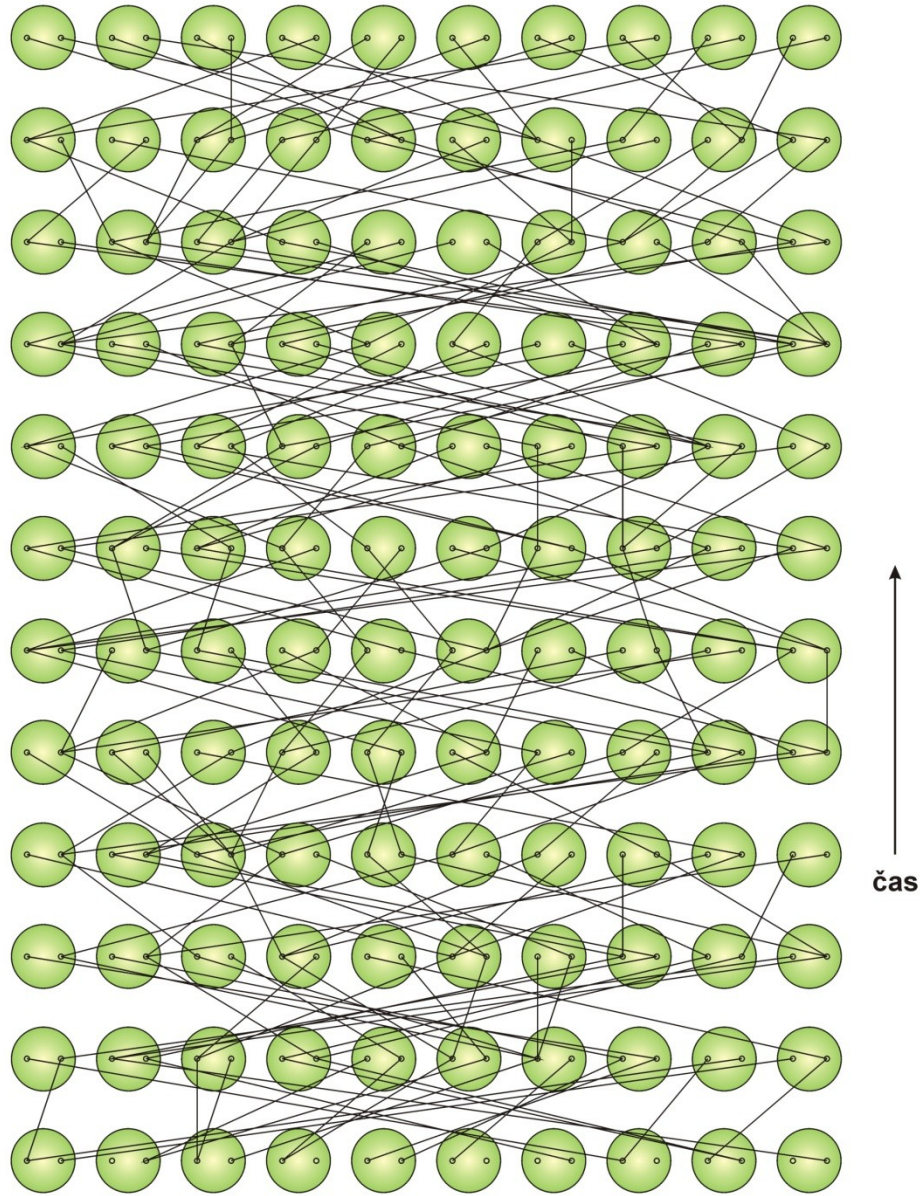


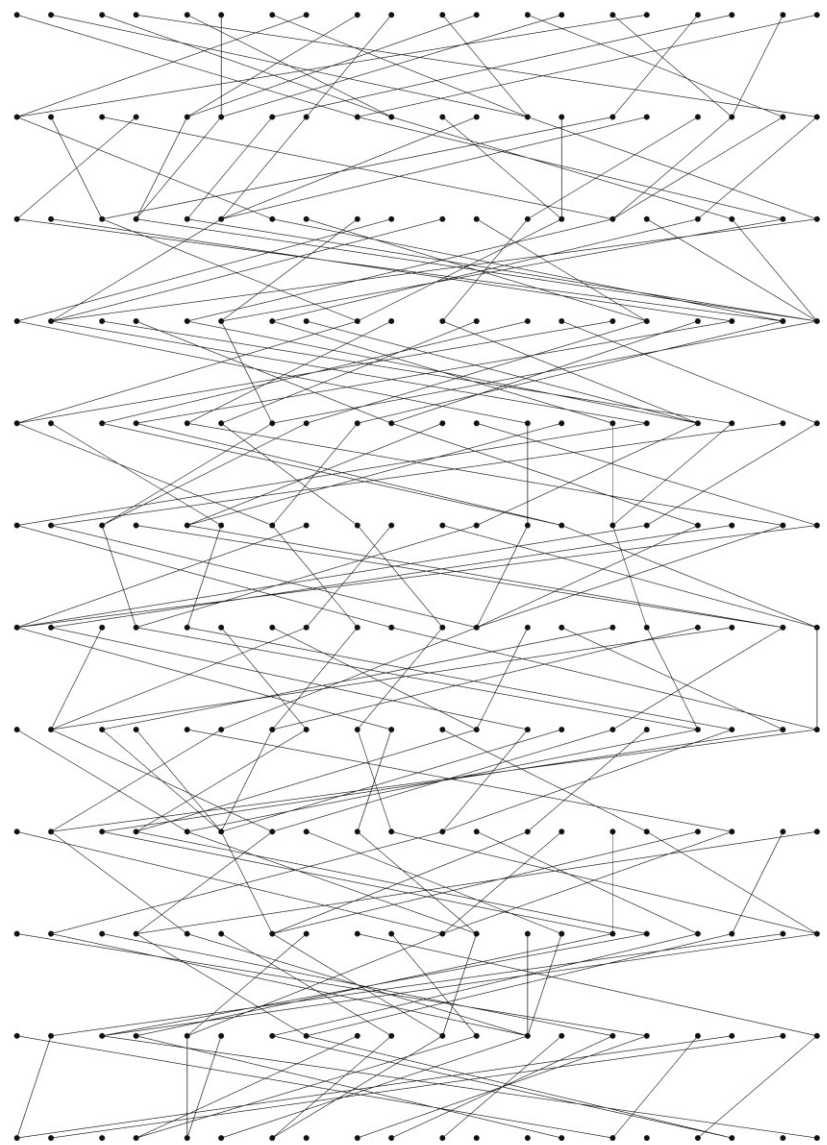
Sewall Wright



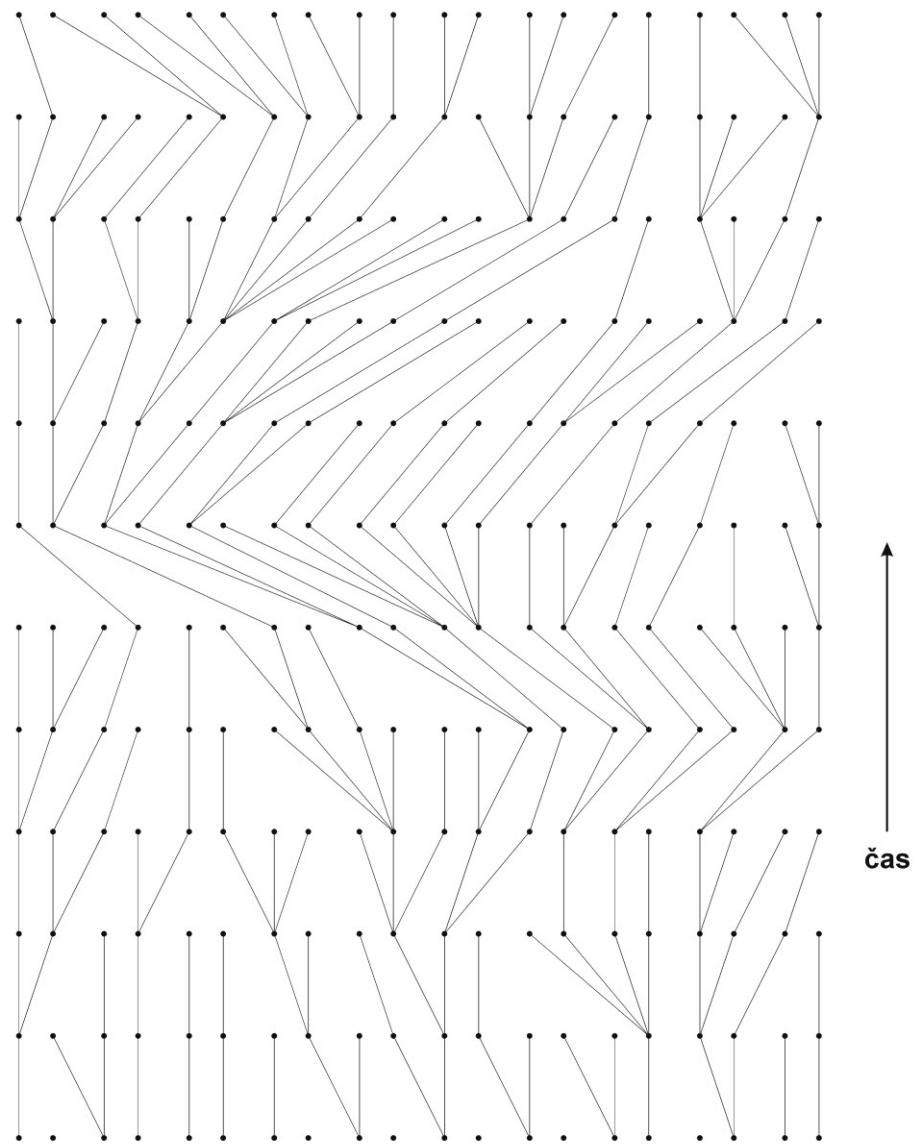
Ronald A. Fisher



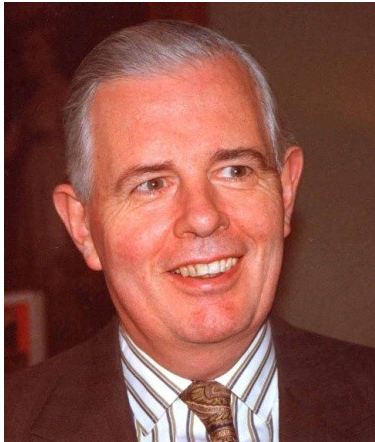




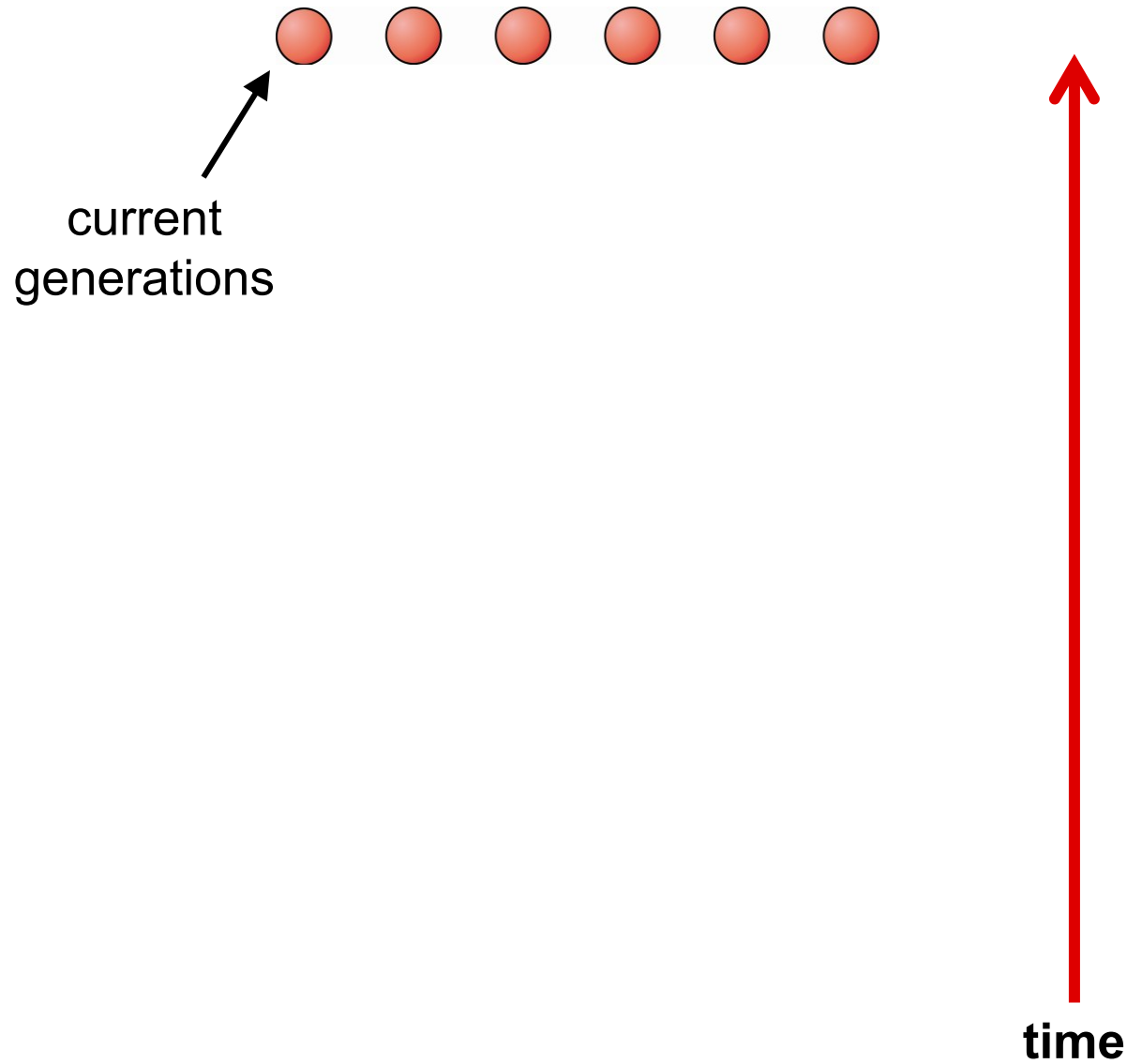
↑
čas



Coalescent:



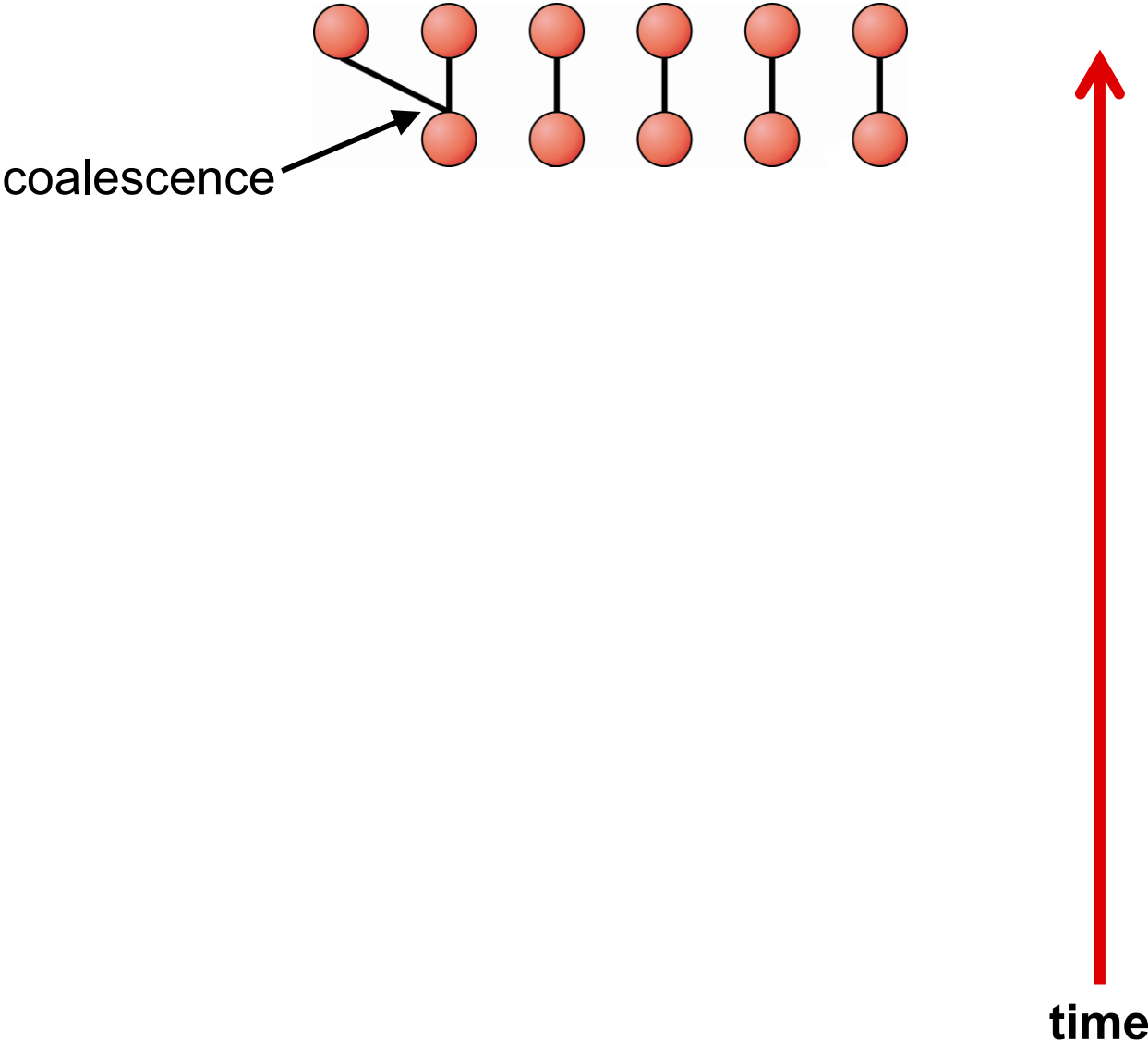
John F.C. Kingman



Coalescent:



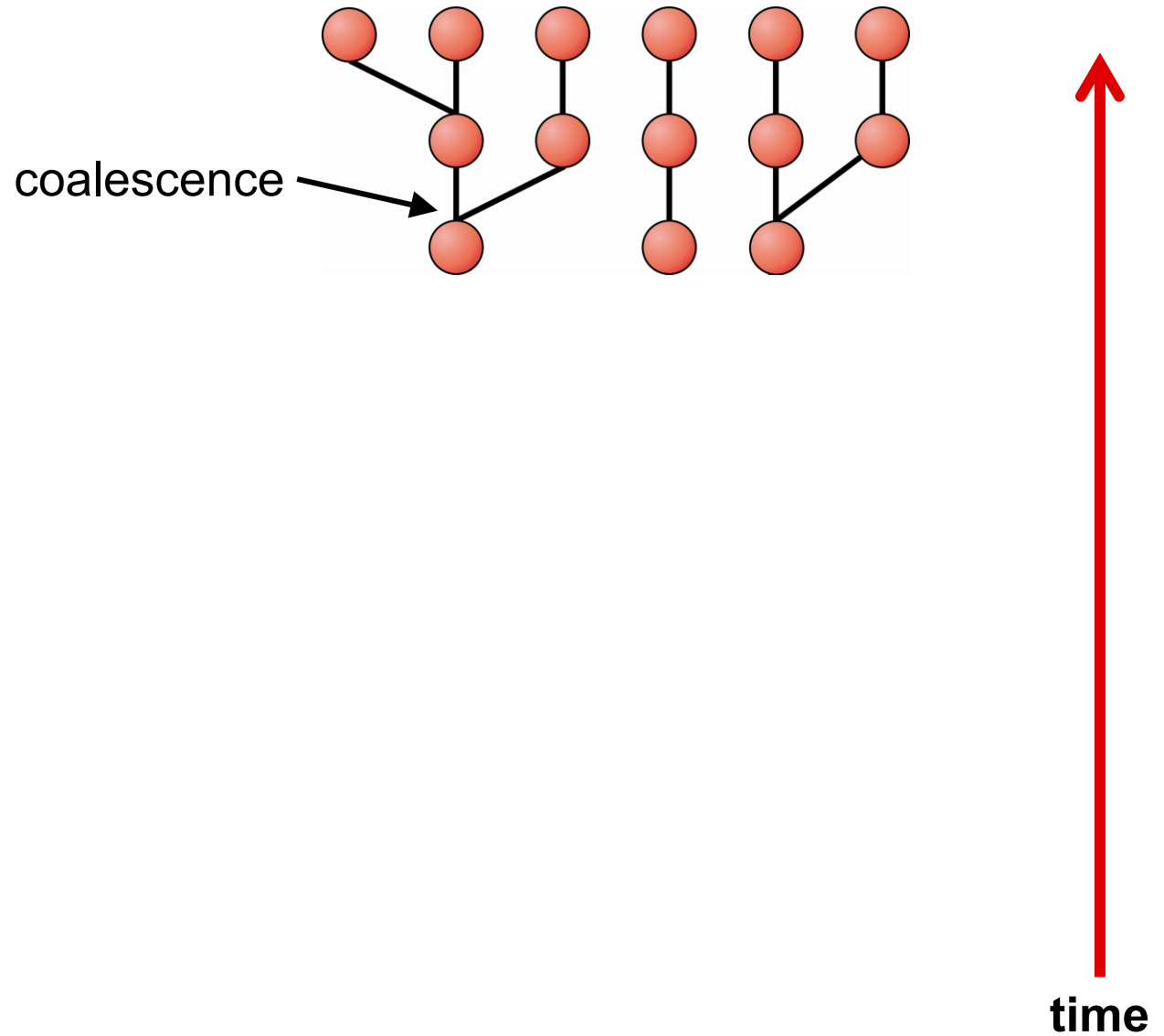
John F.C. Kingman



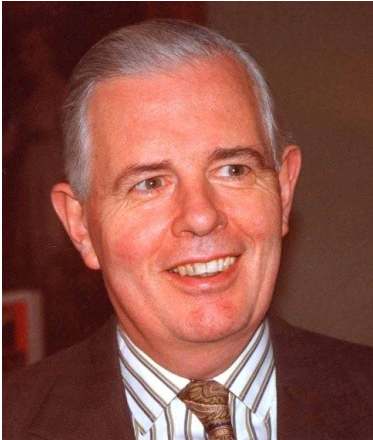
Coalescent:



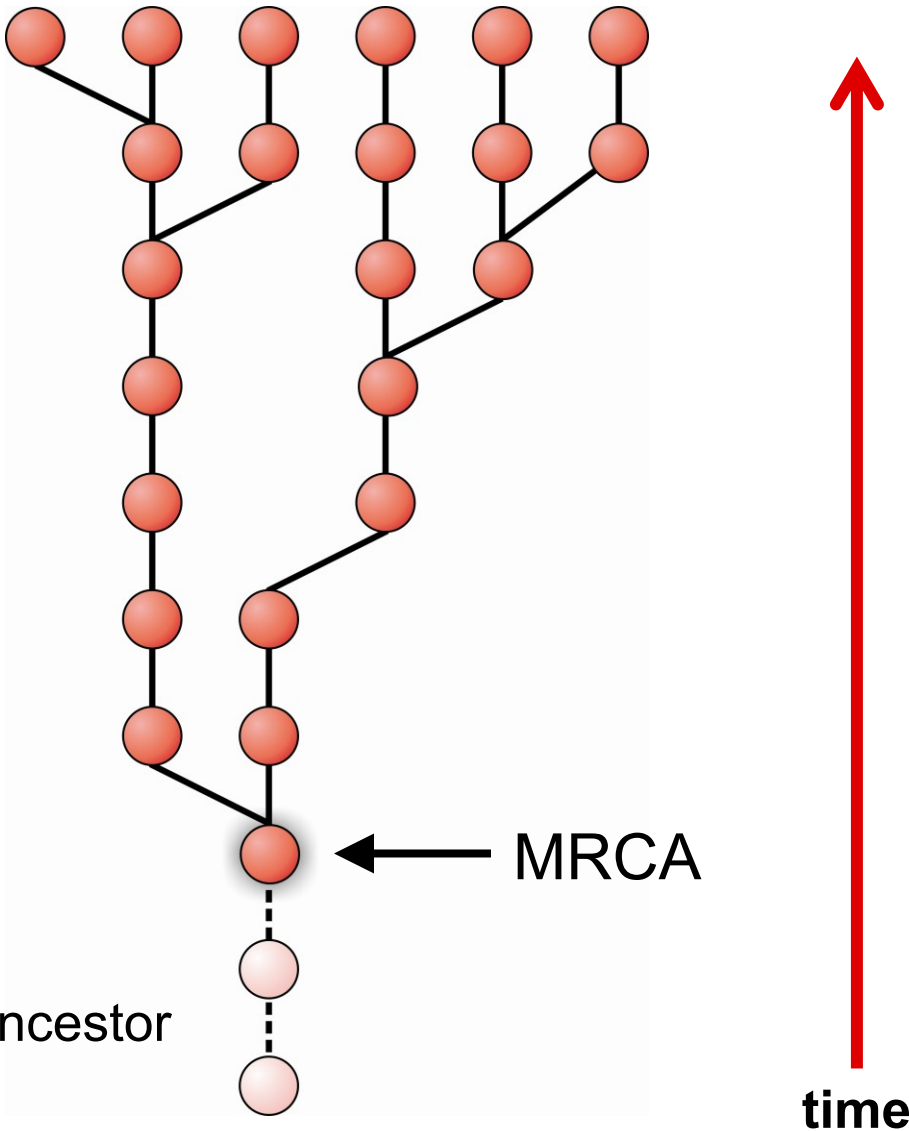
John F.C. Kingman



Coalescent:

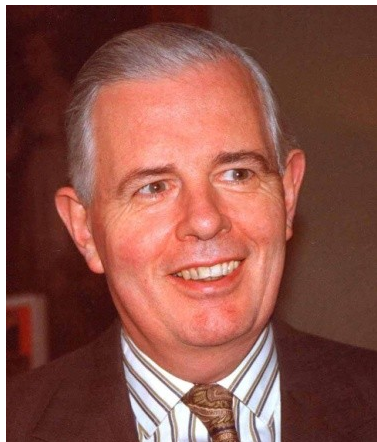


John F.C. Kingman



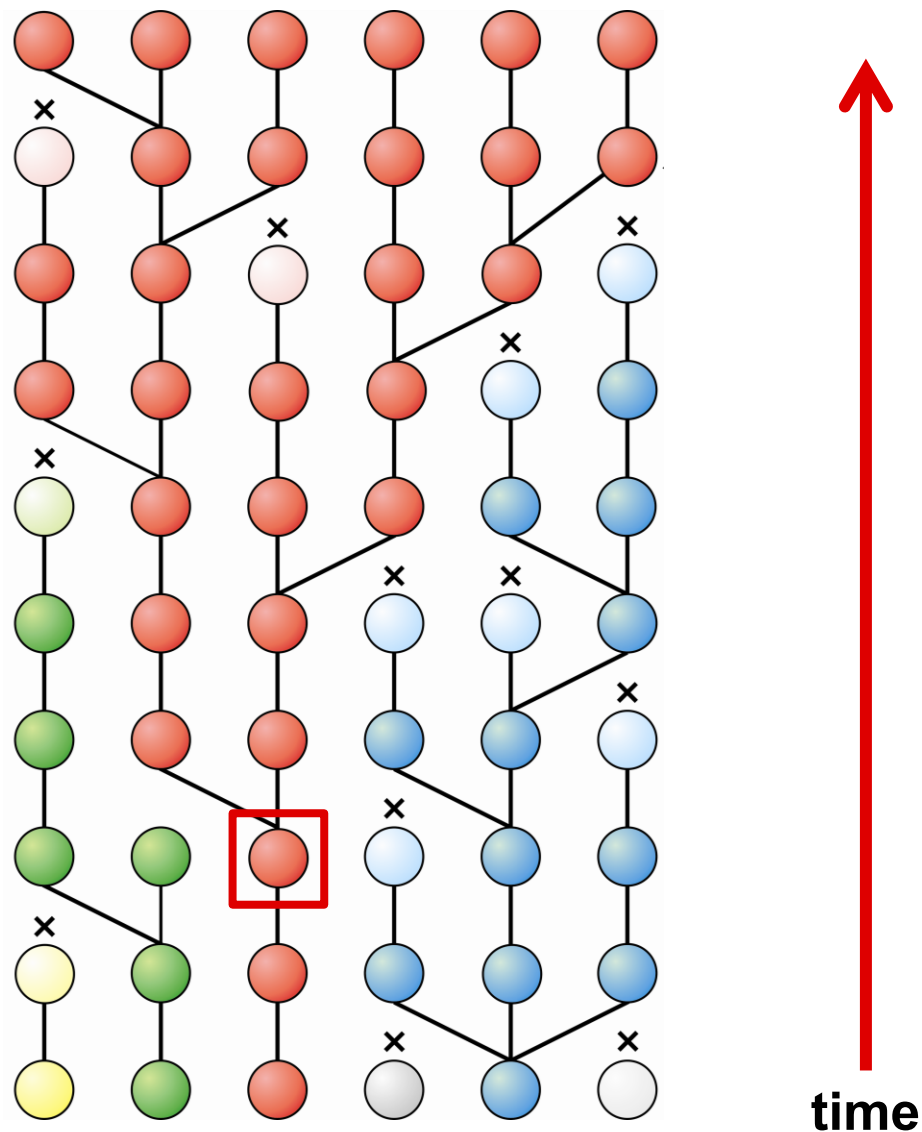
MRCA = most recent common ancestor

Coalescent:

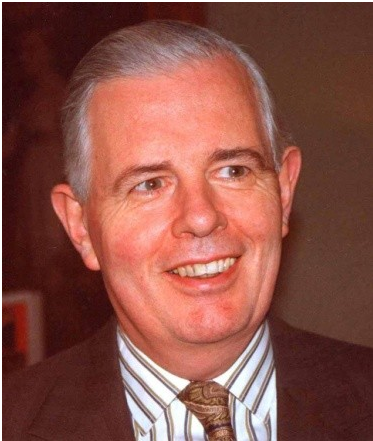


John F.C. Kingman

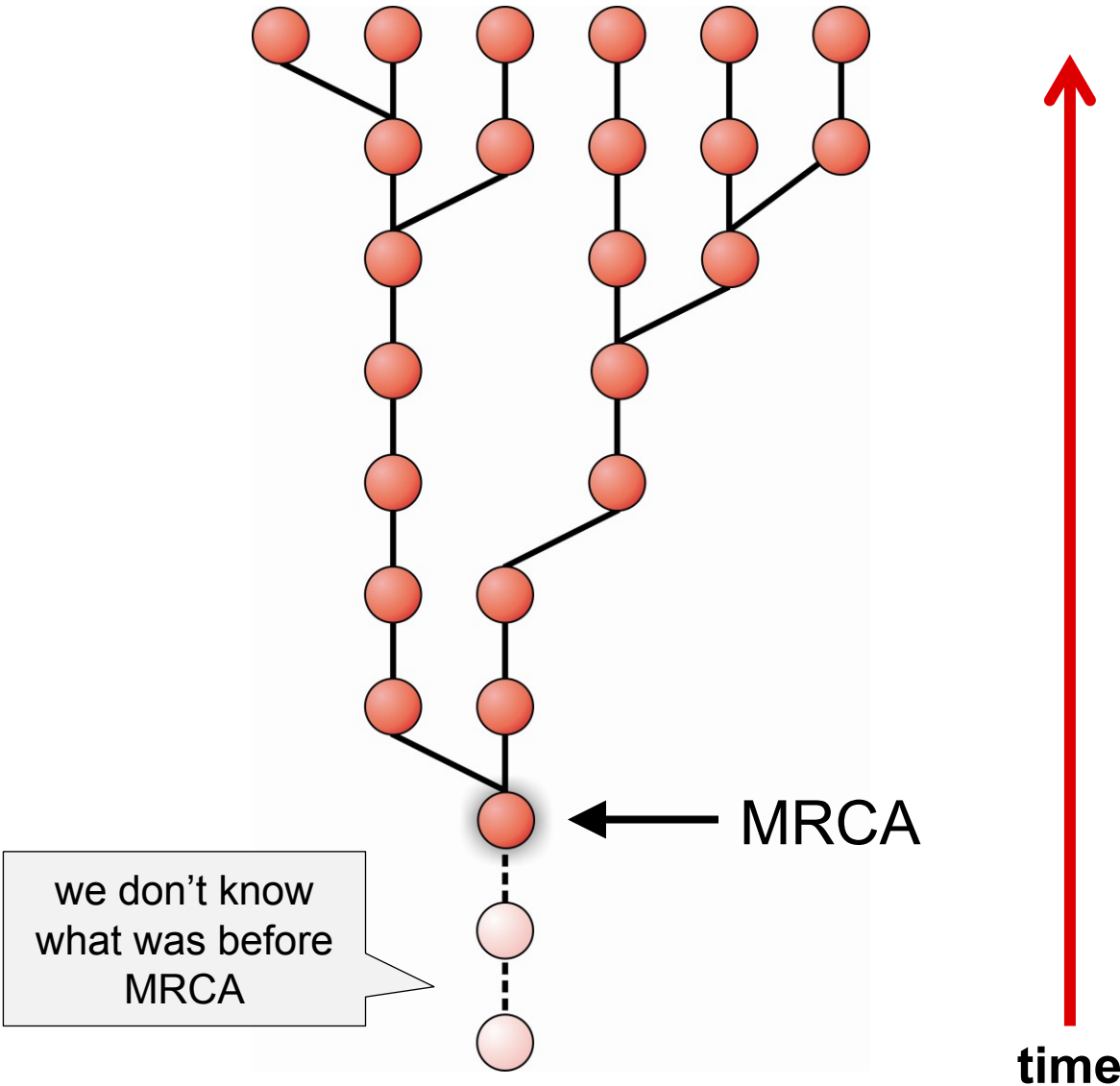
we don't know how many copies were in generation of MRCA



Coalescent:



John F.C. Kingman



we don't know what was before MRCA

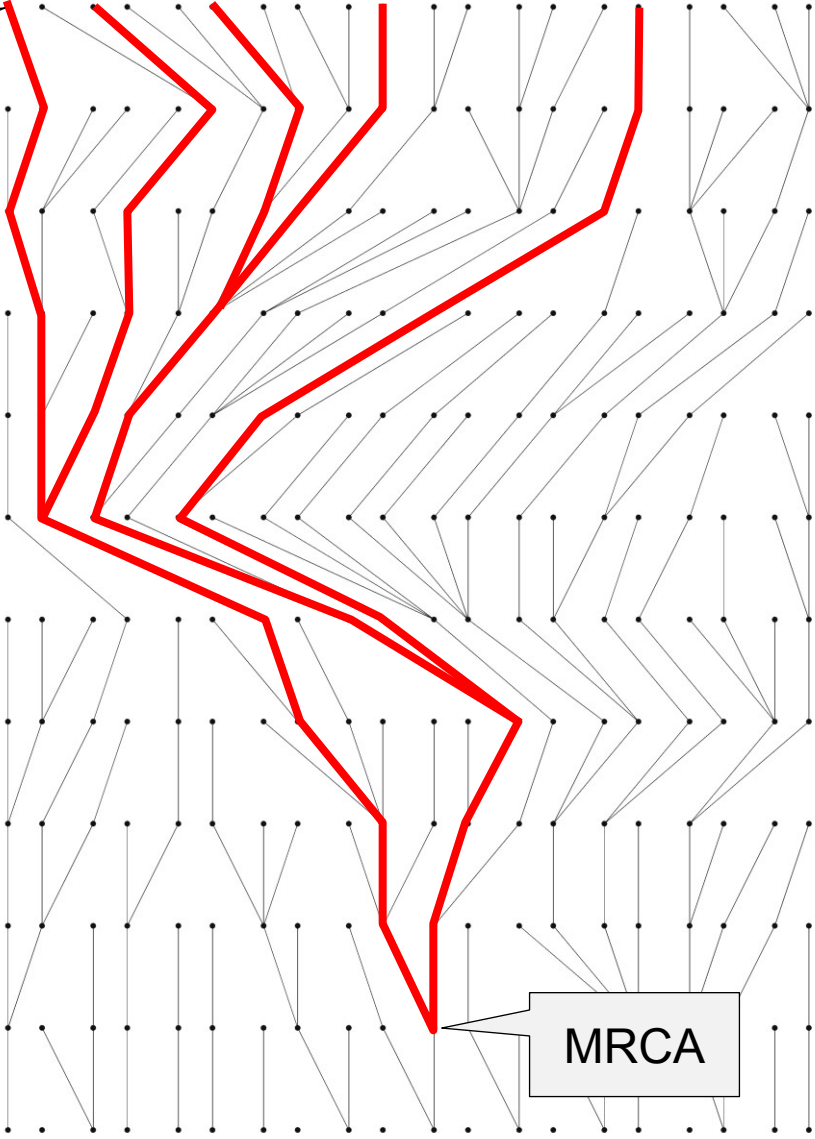
MRCA

time

$n = 5$ copies
in sample

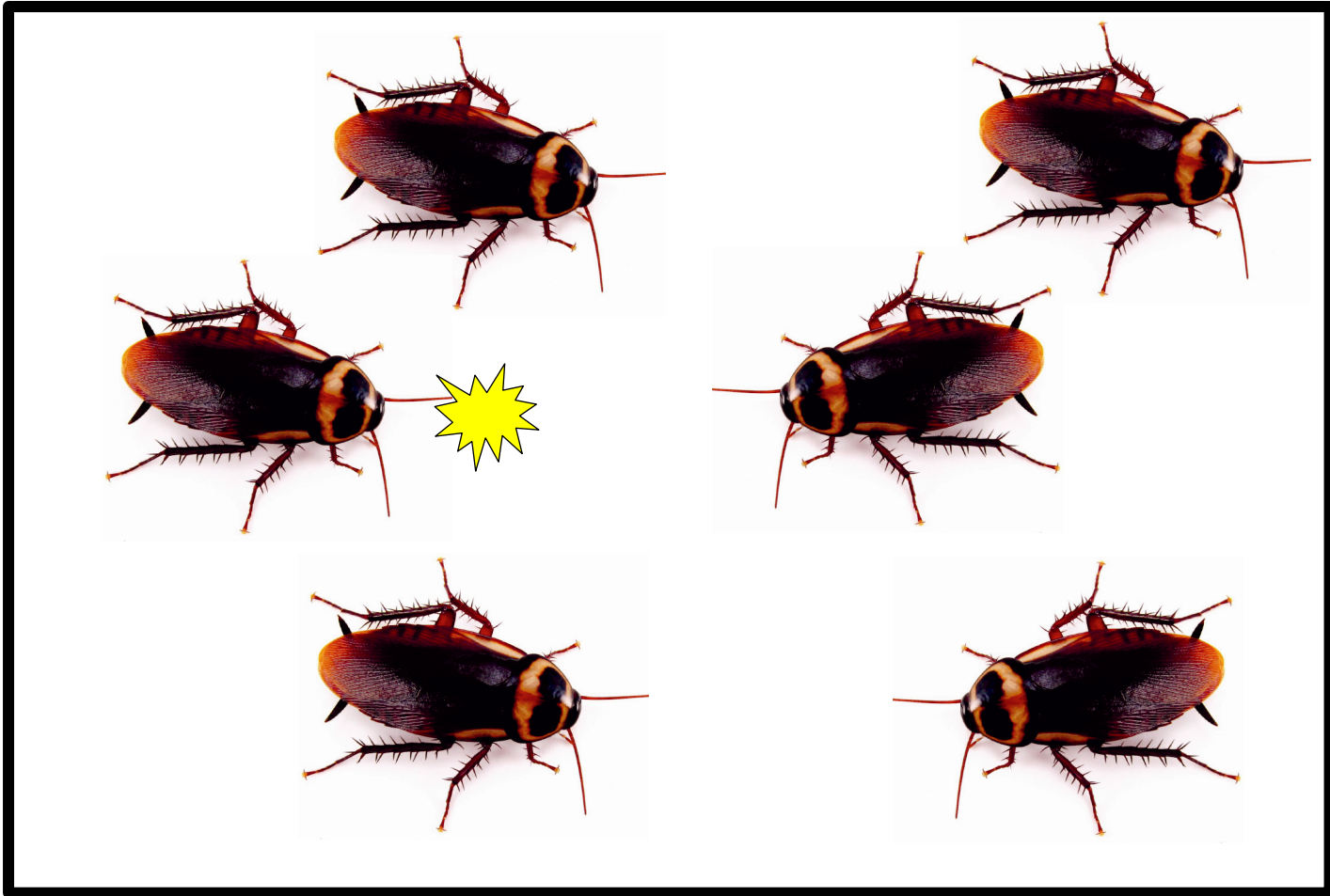
usually
 $n \ll N$

$N = 20$
copies in
population

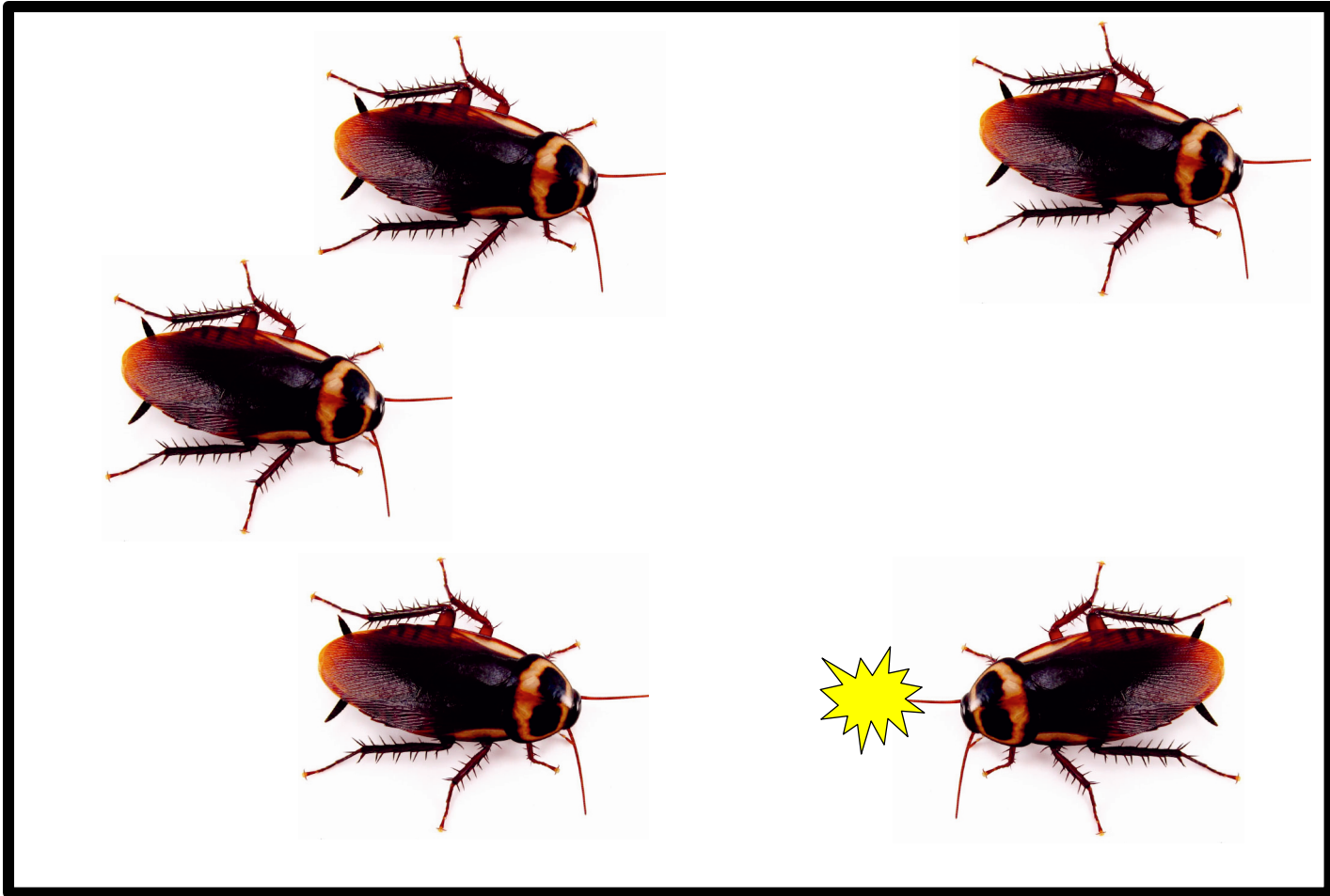


MRCA

čas



Probability of encounter of 2 cockroaches is $n(n - 1)/4N$, where n = number of cockroaches in box, N = number of „places“ in box



after coalescence, number of cockroaches (copies)
is reduced by 1 ...

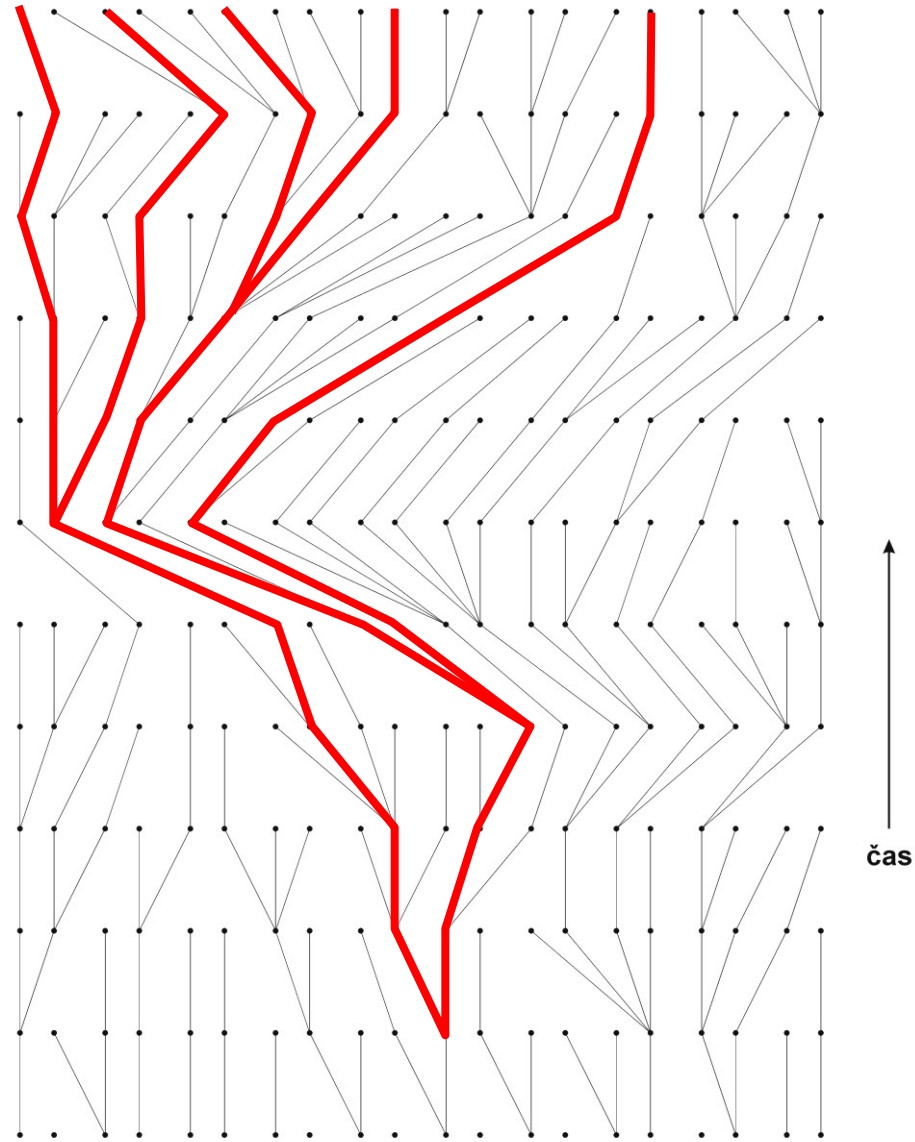
with decreasing number of cockroaches (n), time to next contact (coalescence) increases

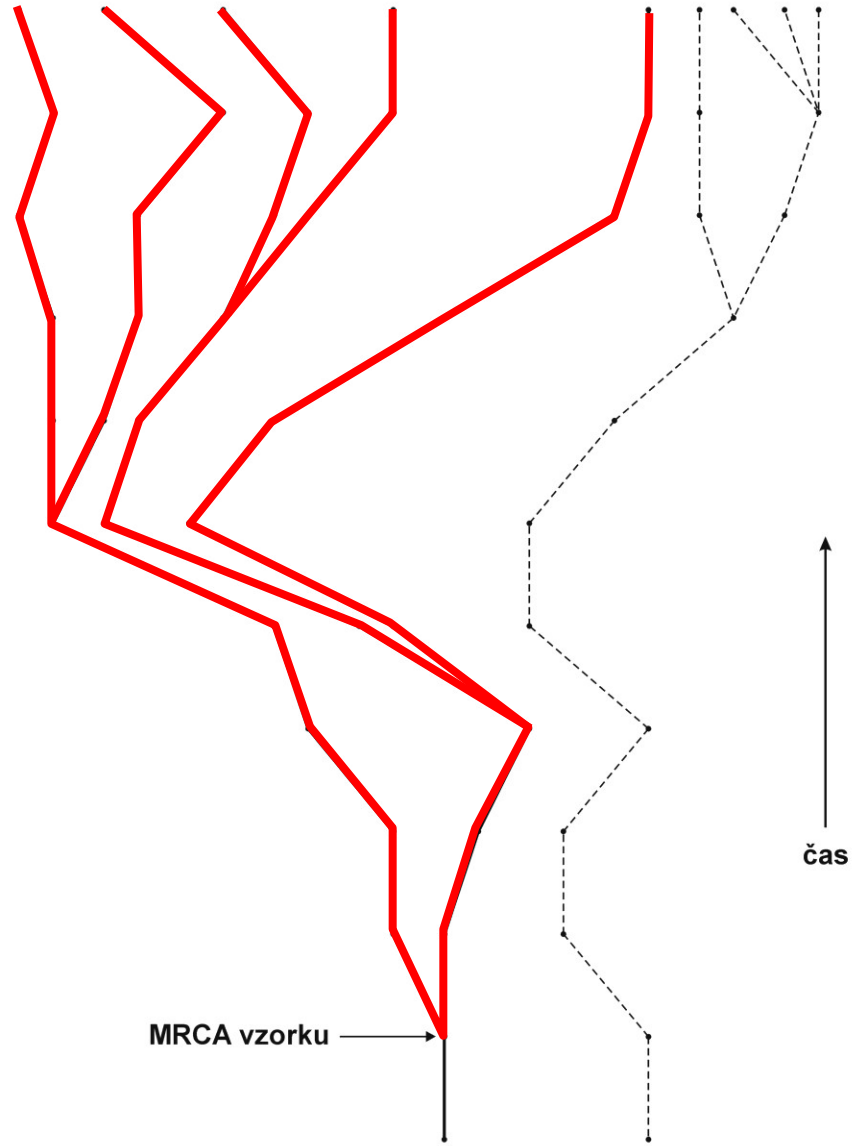


after coalescence, number of cockroaches (copies) is reduced by 1 ...



... to finish with just 1 copy





Kingman's coalescent:

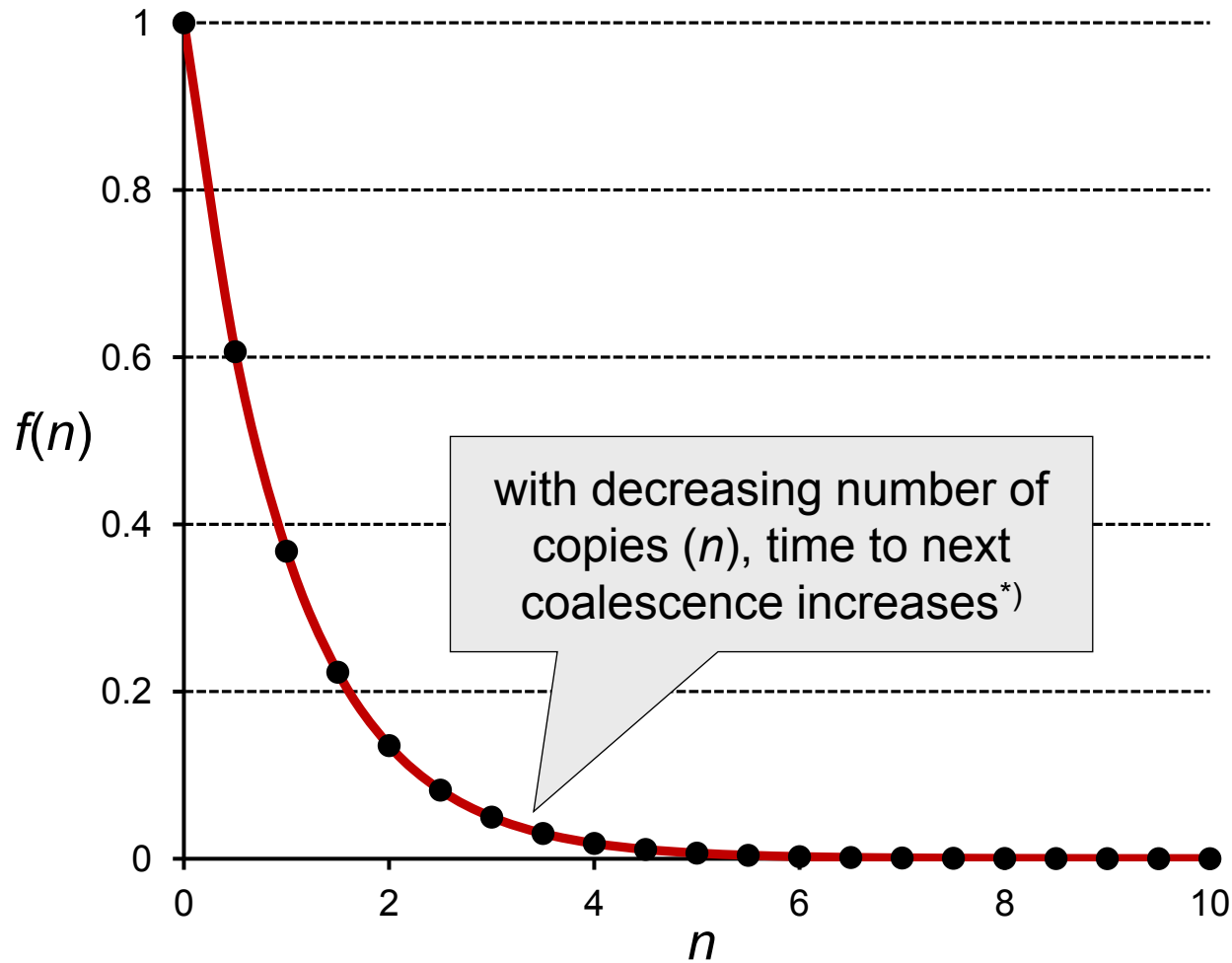
the process of coalescence gets slower with decreasing number of remaining copies, (for large $n \sim 4N$, for 2 copies $\sim 2N$)

coalescence of the last k copies takes $(1 - 1/n)/(1 - 1/k)$

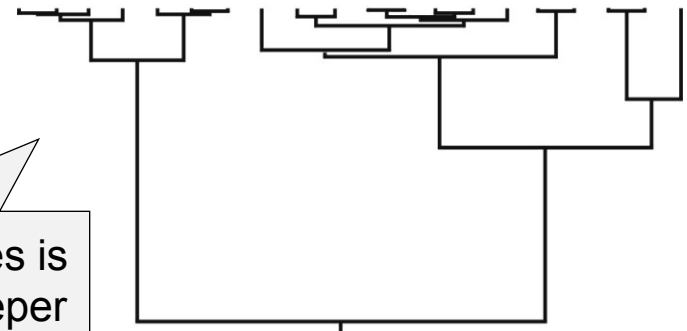
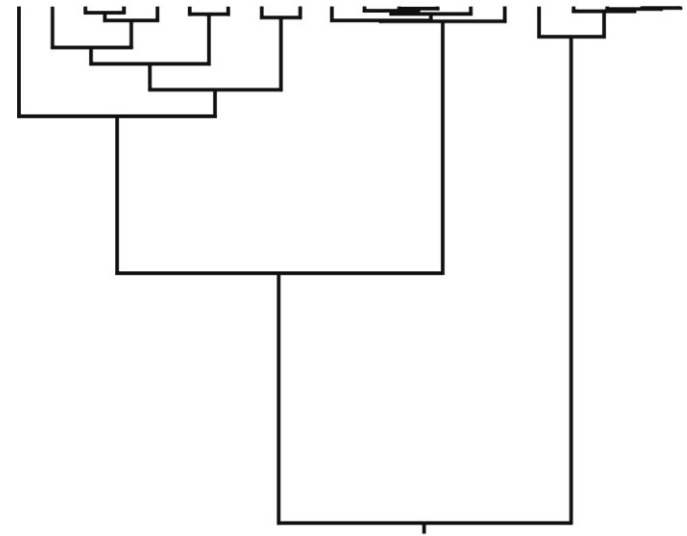
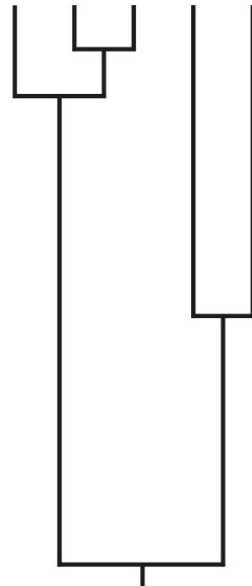
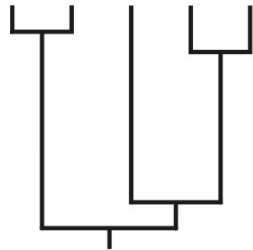
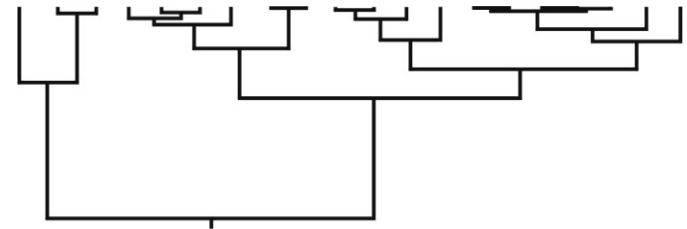
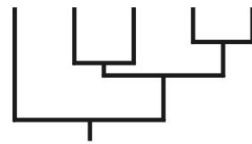
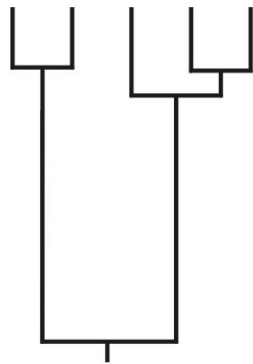
\Rightarrow first 90% copies coalesce during 9% of total time, remaining 91% of time, we wait for coalescence of the last 10% copies!

if there are 100 lineages, probability that the 101st lineage adds a deeper root is only 0,02% \Rightarrow including additional gene copies is unlikely to result in a deeper (older) MRCA

distribution of time between coalescences is approximately exponential:



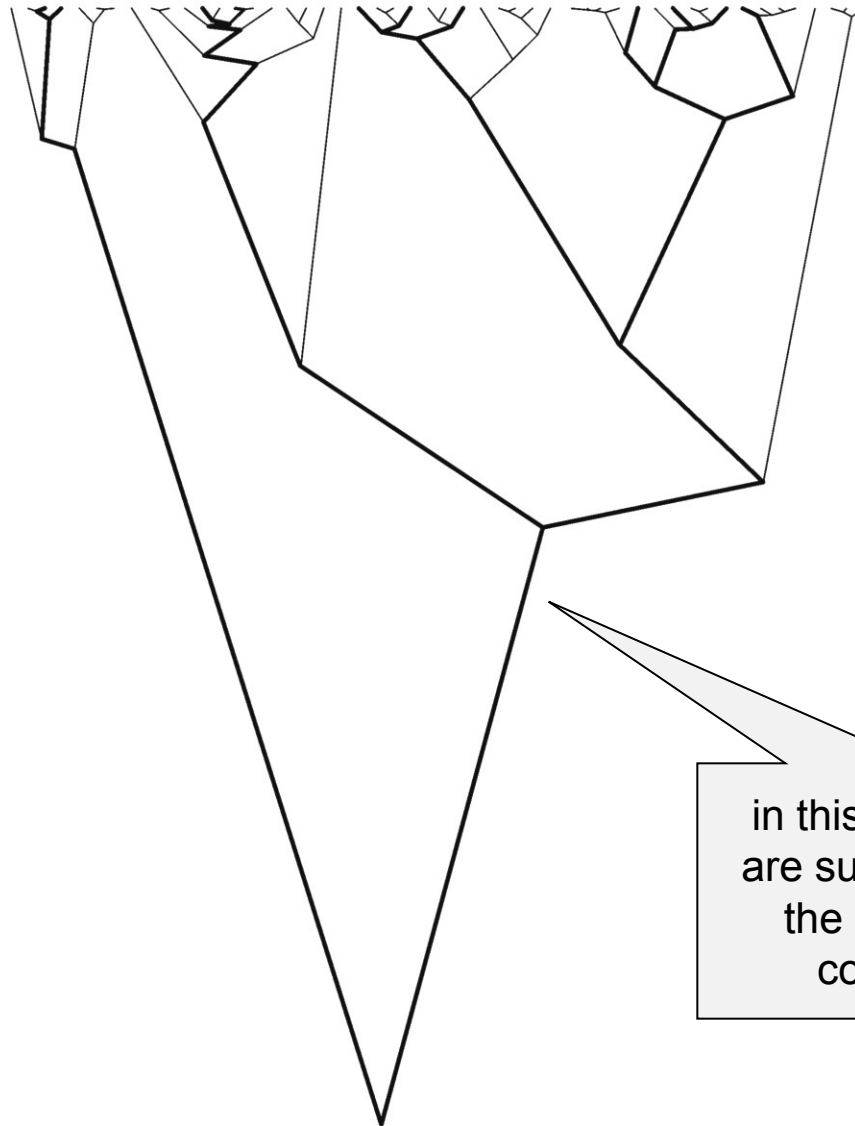
*) see number of cockroaches in the box



with decreasing number of free copies the process slows down

adding other sequences is unlikely to result in deeper coalescence

50 gene copies, 10 randomly chosen:



If we are interested in „old“ coalescences, we don't need large samples

eg. only 2 copies render, on average, 50% of coalescent time for the whole population!

By contrast, if we are interested in time to first coalescence from n to $n - 1$, estimate $N_e/[n/(n - 1)]$ is sensitive to n

eg. range of mean time between first and last coalescence for 10 genes is $0,0444N_e$ to $3,60N_e$; by increasing n to 100 genes, range will be $0,0004N_e - 3,96N_e$

by increasing n 10×
range increases 100× ...

... for last coalescence
almost no difference

Therefore, for estimates of old evolutionary events, small samples are sufficient, for estimates of recent events, large samples are necessary

Coalescent is affected by various factors, eg.:

mutation

recombination

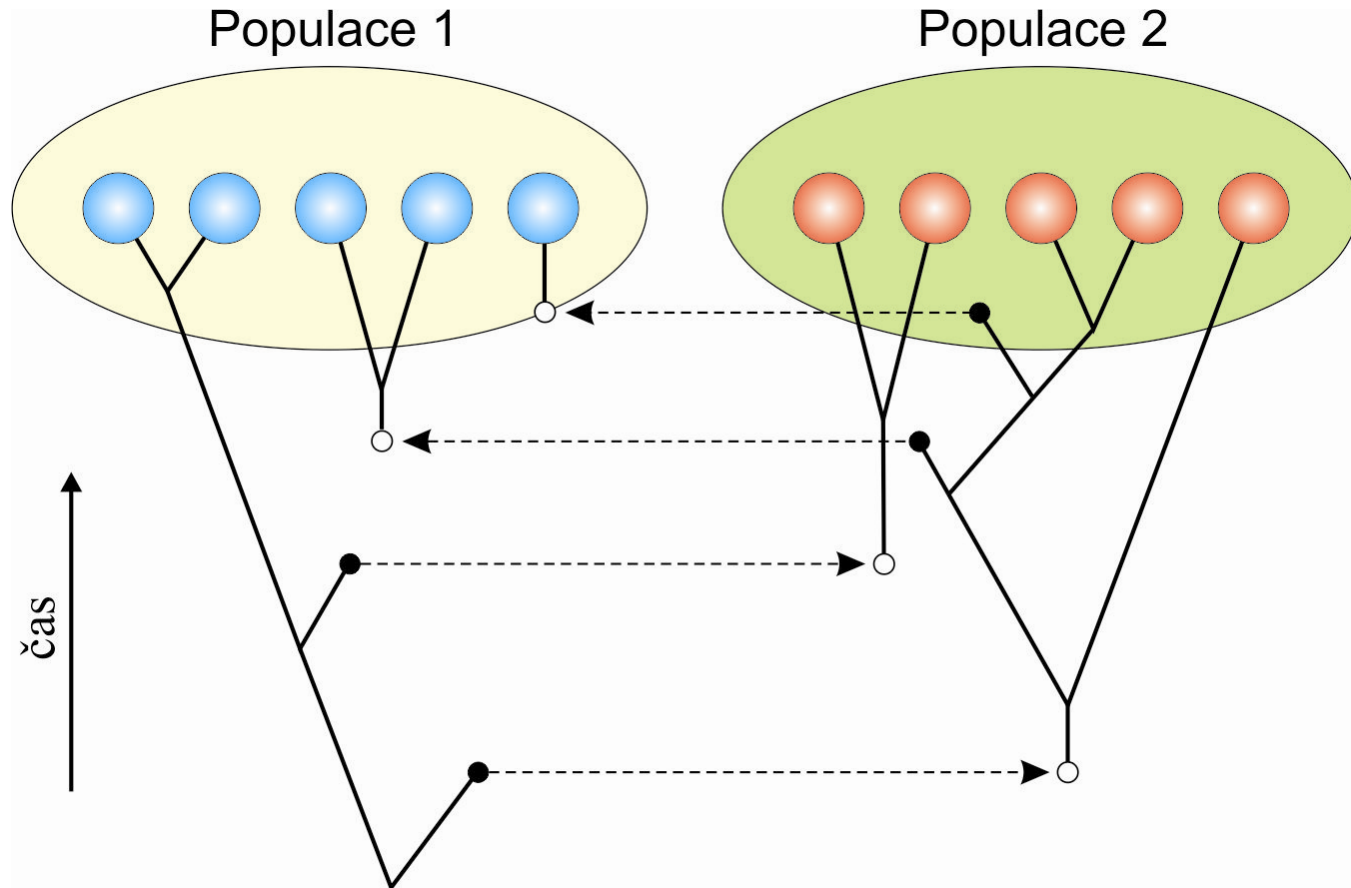
selection

changes of population size

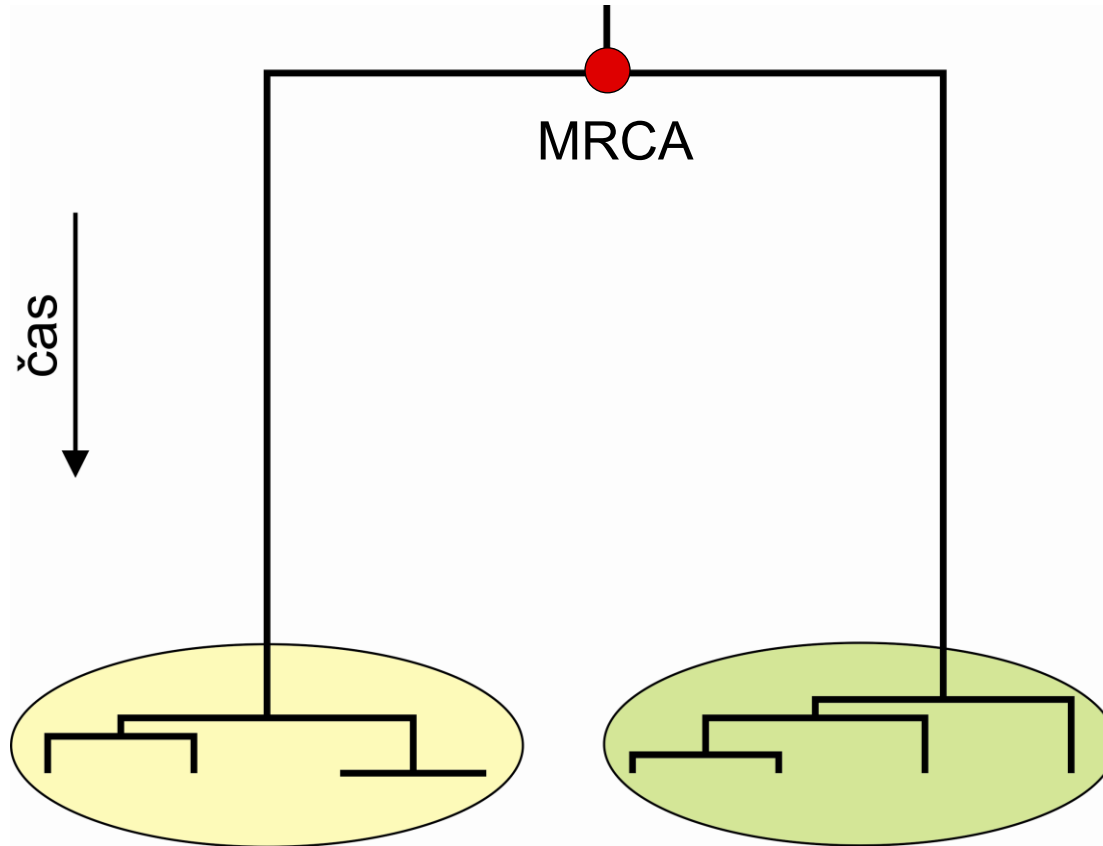
⇒ we can use coalescent theory for estimating these
parameters

Coalescent is affected by various factors, eg.:

by **migration**



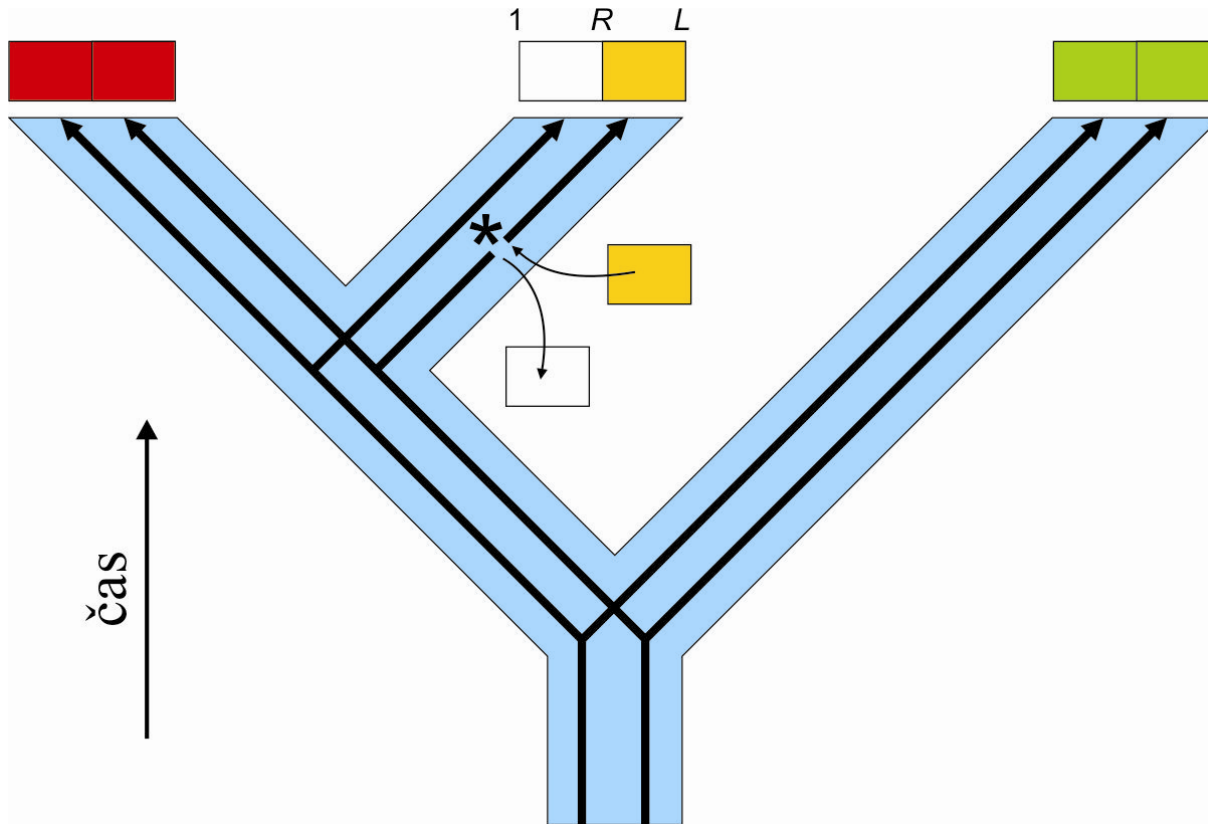
Weak migration leads to most coalescences within local populations,.....



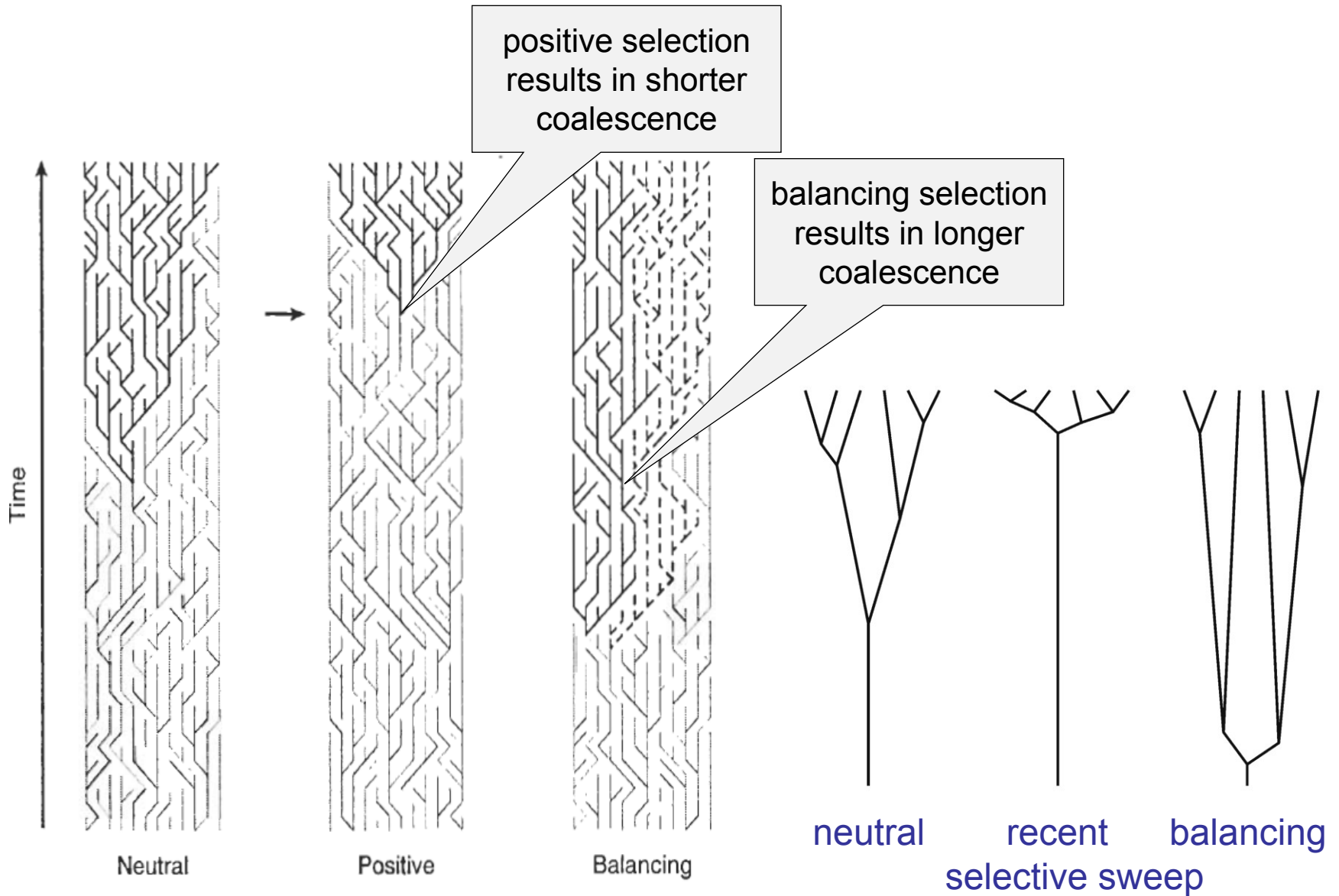
.... to increasing time to MRCA and its variance

Coalescent is affected by various factors, eg.:

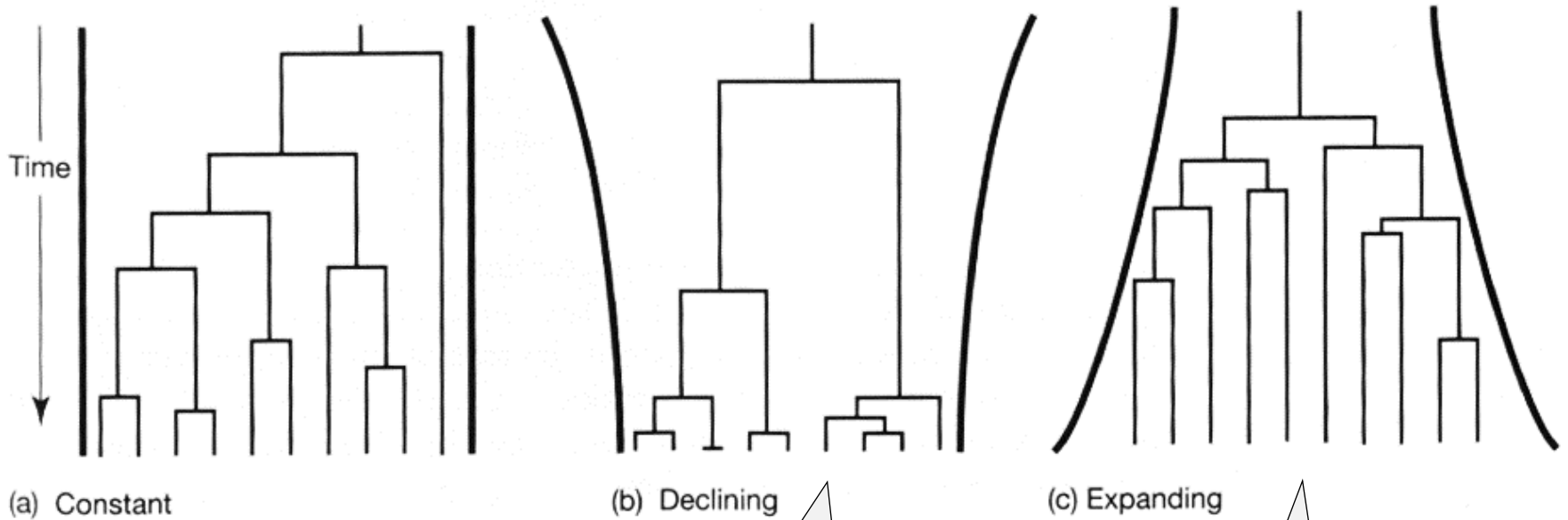
by **recombination**



Effect of selection on shape of coalescent tree



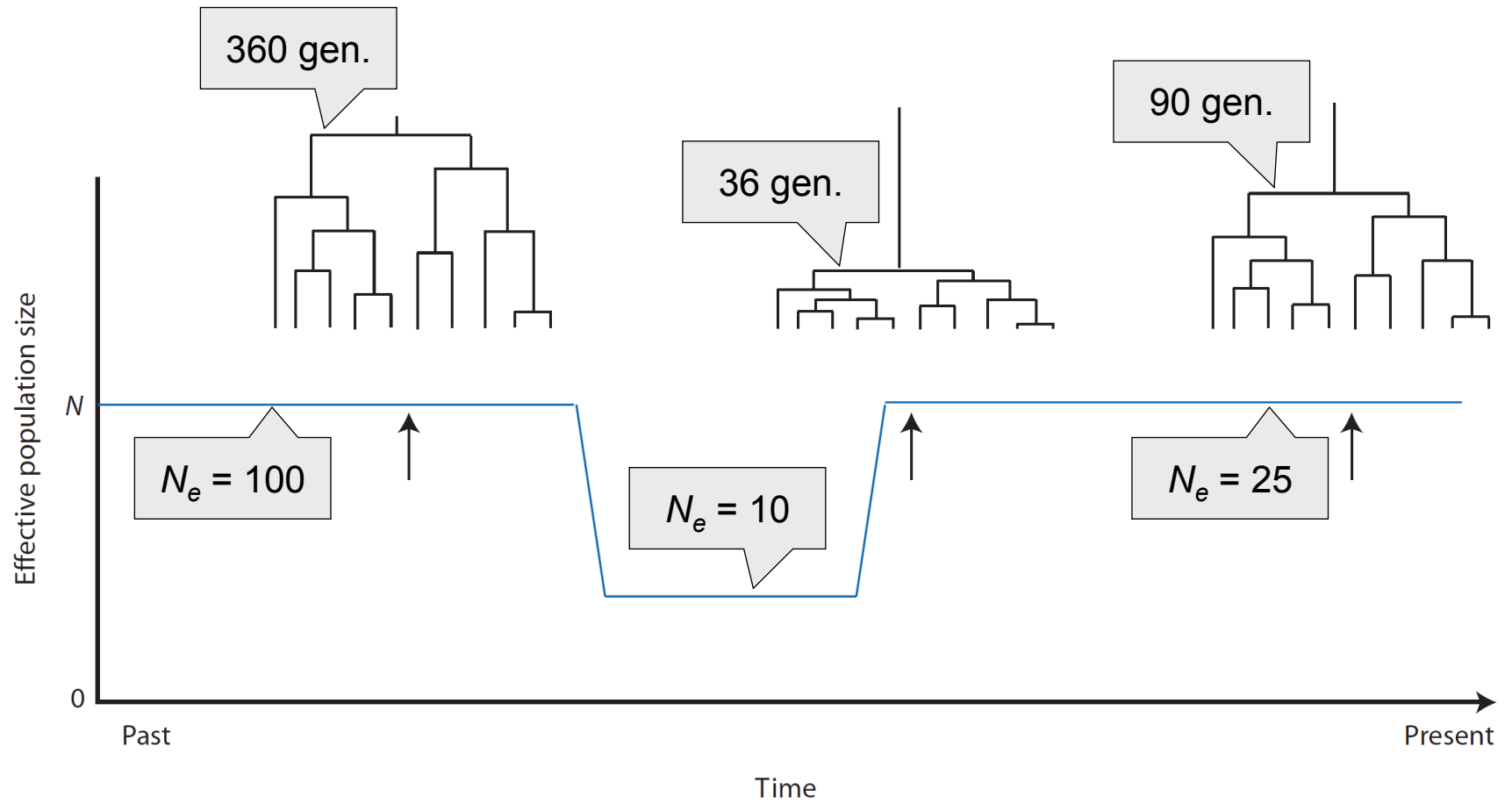
Effect of changes in population size on shape of coalescent tree



declining population:
coalescent rate
increases

growing population:
coalescent rate
decreases

$n = 10$



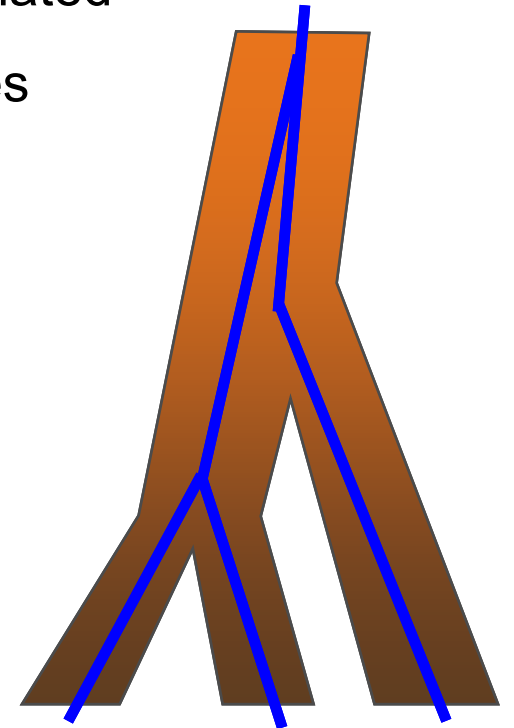
Gene vs. species trees once more:

long intervals between speciation events → gene and species trees are identical

short intervals between speciation events → gene and species trees can differ (hemiplasy)

since we assess divergence among sequences and not between species, our estimates are necessarily overestimated

discrepancies between gene trees and species trees can be minimized by using markers with low N_e , eg. mtDNA or Y chromosome



PHYLOGEOGRAPHY

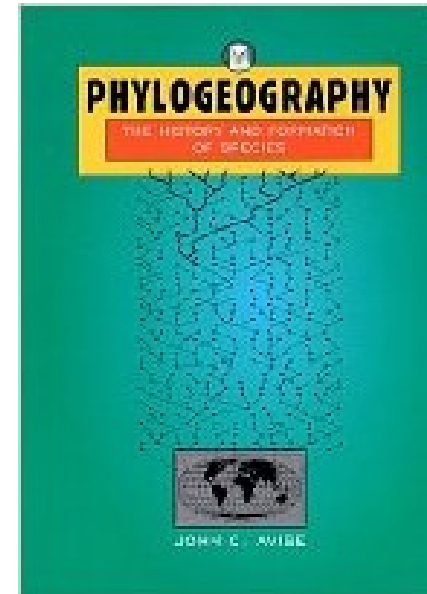
studies principles and processes affecting geographic distribution of genealogical lineages

in a way, it combines microevolutionary processes (population genetics) with macroevolution (phylogenesis)

mostly intraspecific studies or related species



John C. Avise



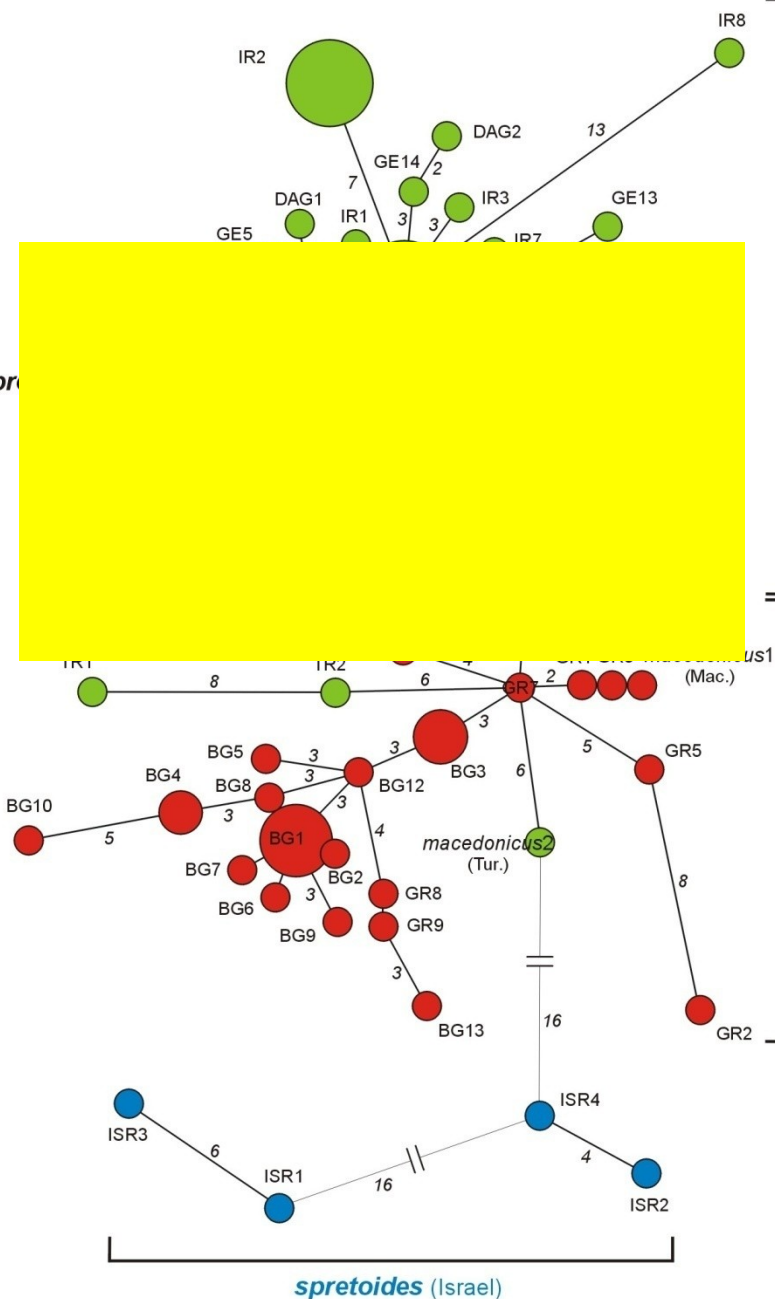


Mus macedonicus

Asia

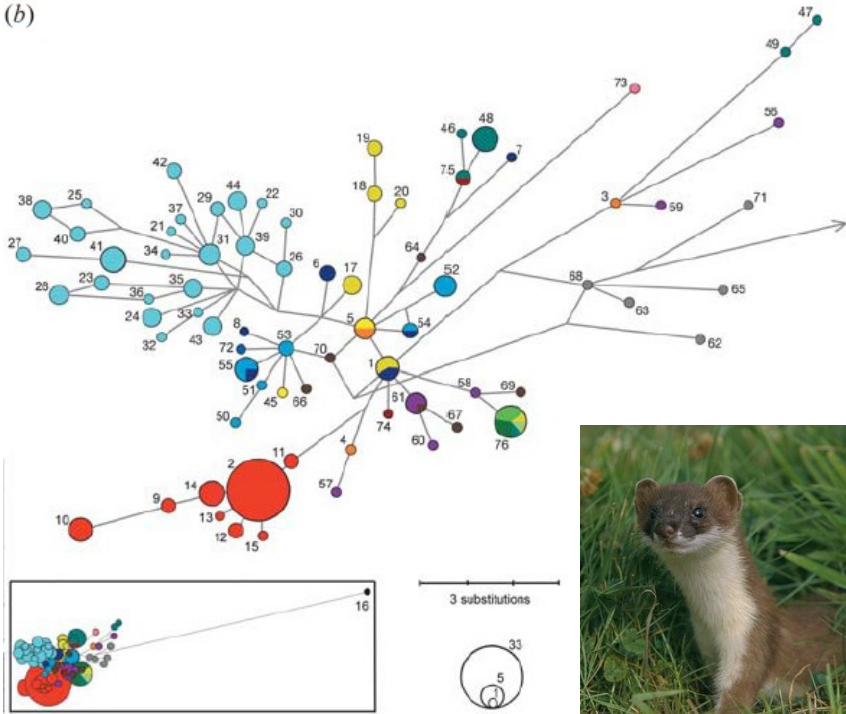
Minimum Spanning Tree (MST)
 Mimum Spanning Network (MSN)
 Median-joining network etc.

M. spr



(b)

Europe



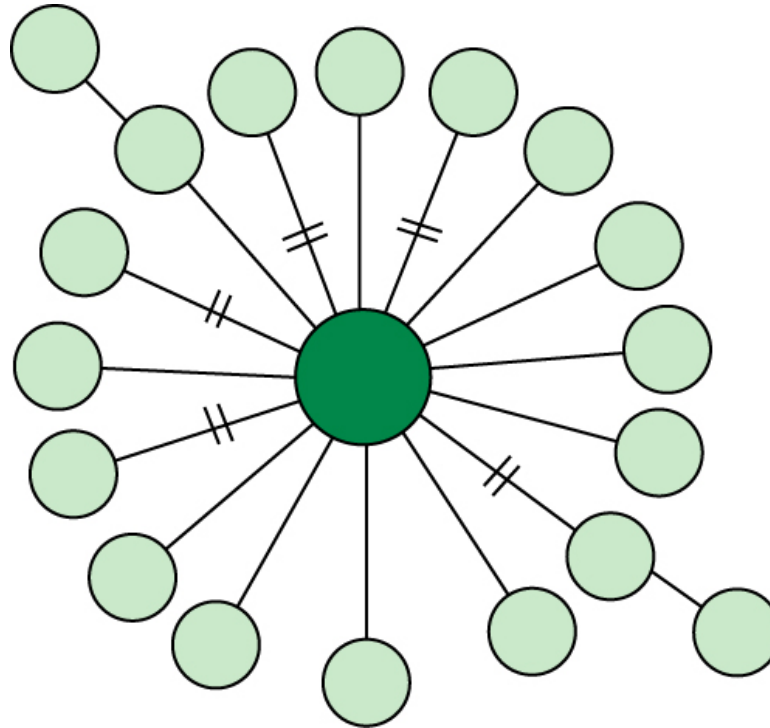
Mustela erminea

Recent expansion:

rapid expansion of a single haplotype

accumulation of low number of mutations

star structure



Changes of population size

Tajima's test (Tajima's D)

mismatch distribution (rozdělení párových neshod)

coalescent, ML or BA, MCMC

Bayesian Skyline Plot (bayesovský panoramatický graf)

1. Tajima's test

based on comparison of haplotype diversity and nucleotide diversity

primarily it is test of selective neutrality, but it can also indicate population expansion or bottleneck

Let's revisit the neutral theory:

equilibrium heterozygosity $\theta = 4N_e\mu$

if evolution is neutral, θ can be estimated in various ways,

e.g. as the mean number of pairwise differences π (or θ_π)*, or

as θ_W^{**} :

$$\theta_W = \frac{S}{\sum_{i=1}^{n-1} \left(\frac{1}{i}\right)}$$

where S = number of segregating sites
 n = number of sequences

*) nucleotide diversity

**) Watterson's theta

If NT and model of infinite sites: $\theta_\pi = \theta_W$

Fumio Tajima (1989):
$$D = \frac{\theta_\pi - \theta_W}{\sqrt{\text{Var}(\theta_\pi - \theta_W)}}$$

Eg.:

	*	*	*	*
1	ACCCG	AATTC	CAATC	CGGTT
2	AACTG	AATTC	GAATC	CGGTT
3	AACTG	AATTC	CAATC	CGGTT
4	ACCTG	AATTC	TAATC	CGGAT

pairwise comparisons:

1-2: 3 differences

1-3: 2 differences

1-4: 3 differences

2-3: 1 differences

2-4: 3 differences

3-4: 3 differences

av. $\pi = (3+2+3+1+3+3)/6 = 2,5$

S = 4 segregating sites

$\theta_W = 4/(1/1 + 1/2 + 1/3) = 4/1,83 = 2,186$

$-\theta_\pi - \theta_W = 2,5 - 2,186 = 0,314$

1. Tajima's test

very negative values indicate population expansion – prevalence of „young“ polymorphisms, when new haplotypes have arisen, but nucleotide diversity is still low

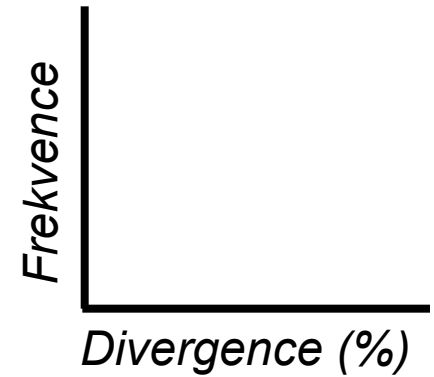
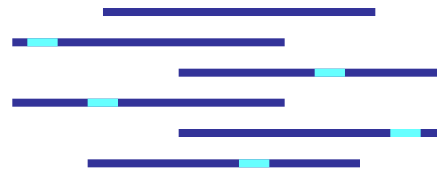
programs Arlequin, DnaSP etc.

likewise Fu's test etc.

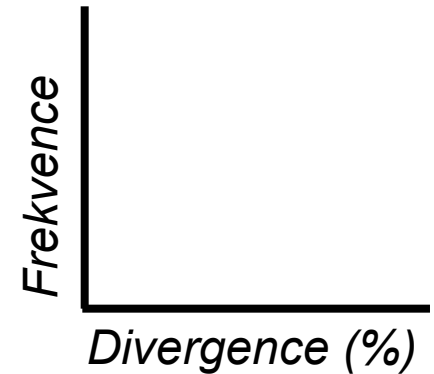
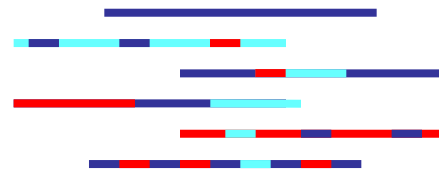
2. Mismatch distribution

pairwise comparison of all sequences → histogram

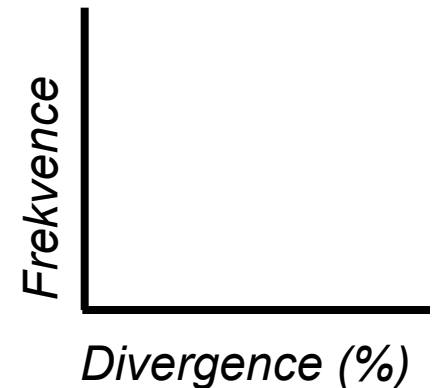
Sequences very similar

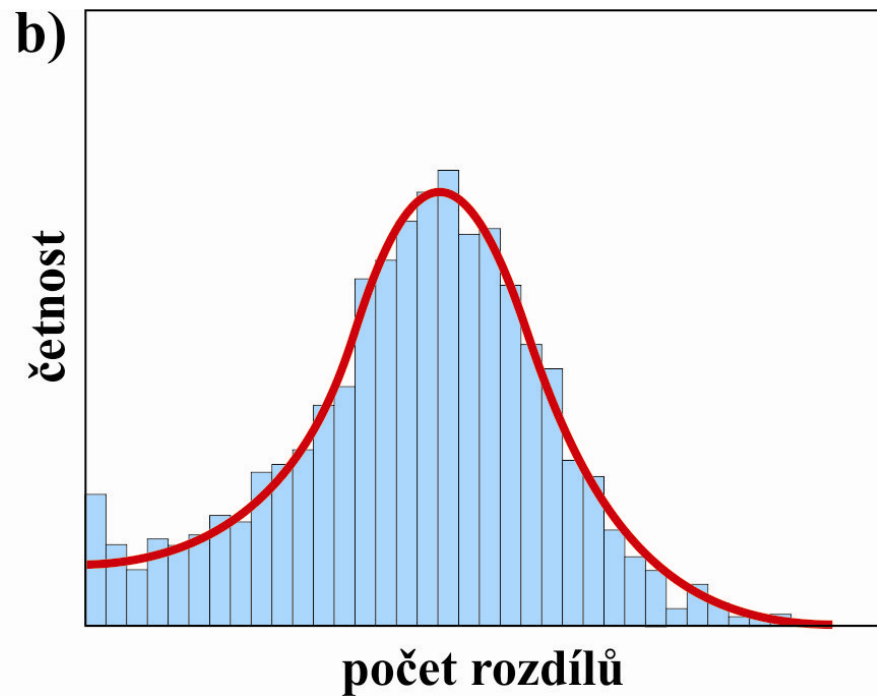
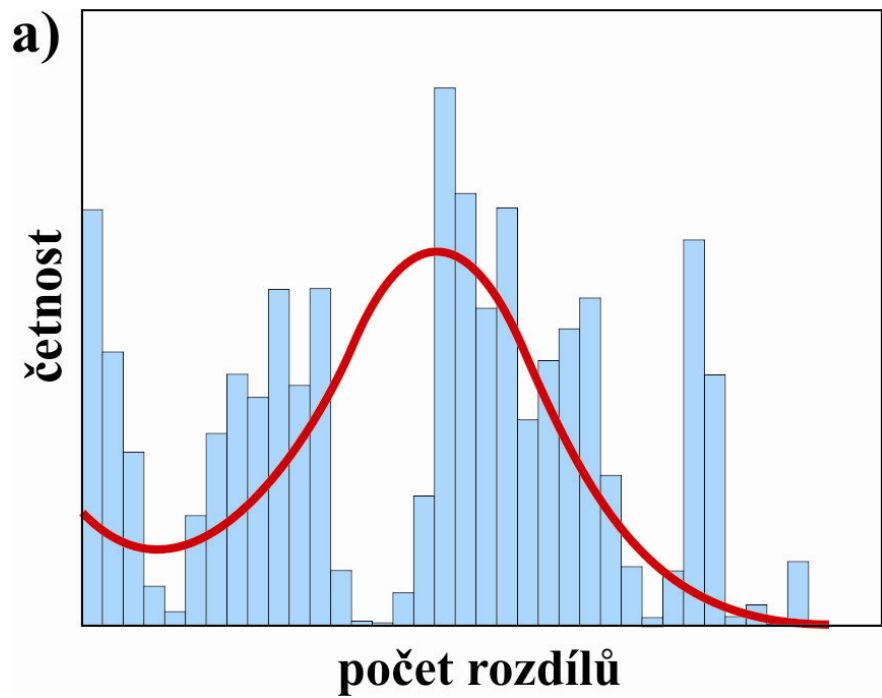


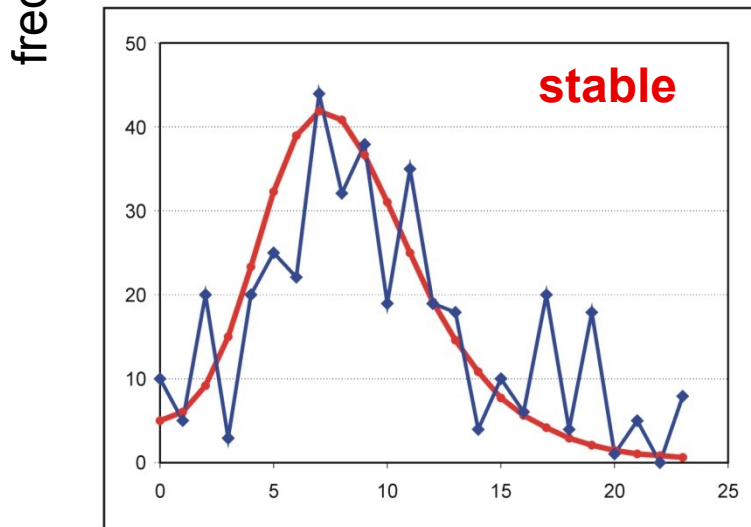
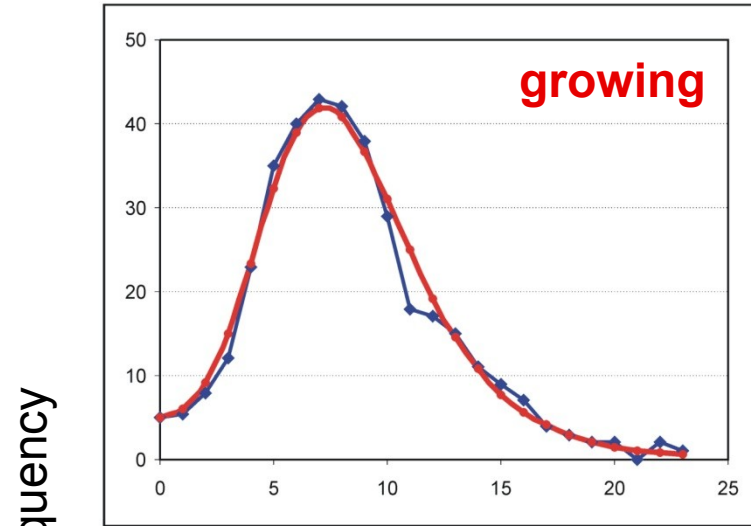
Sequences very divergent



Mixture of similar and divergent sequences







pairwise differences

test of agreement between real distribution and prediction:

Harpending's raggedness index (Harpending 1994)

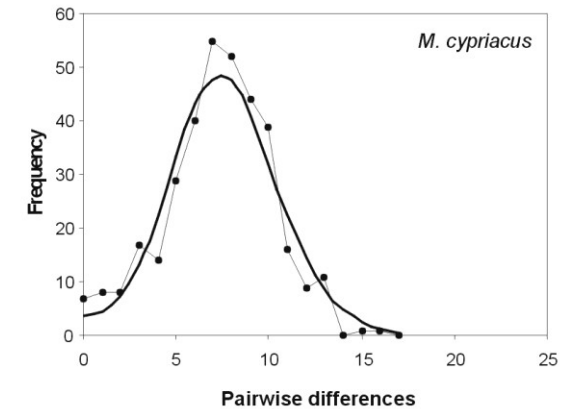
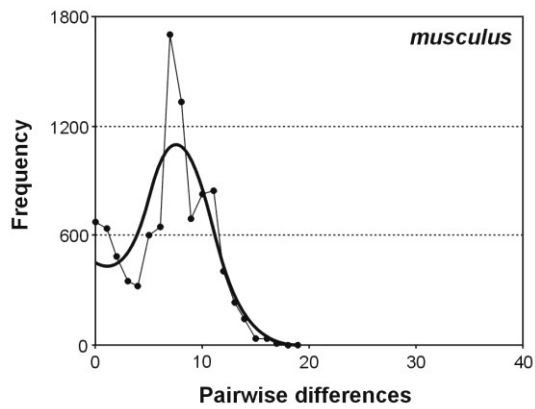
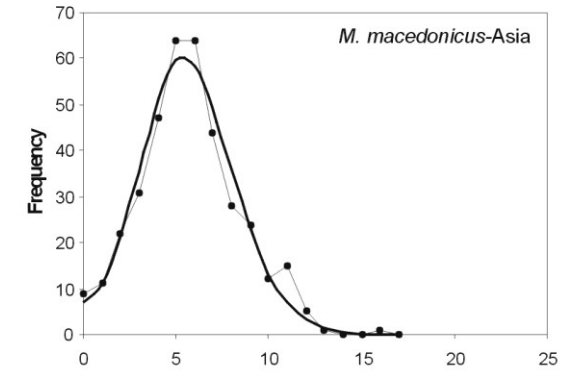
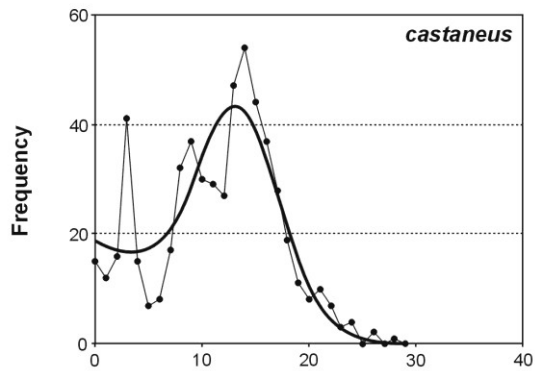
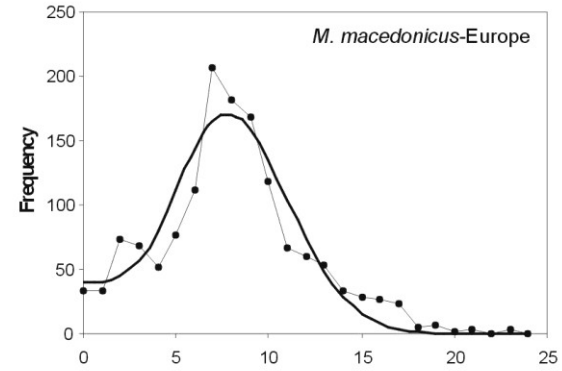
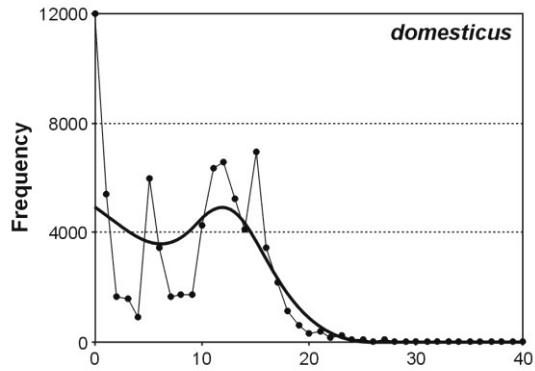
sum of squared deviations

time of expansion/bottleneck:

$$\tau = 1/2u,$$

where u is mutation rate for whole sequence

we can also estimate population size before and after expansion




3. ML a Bayesian inference

MCMC

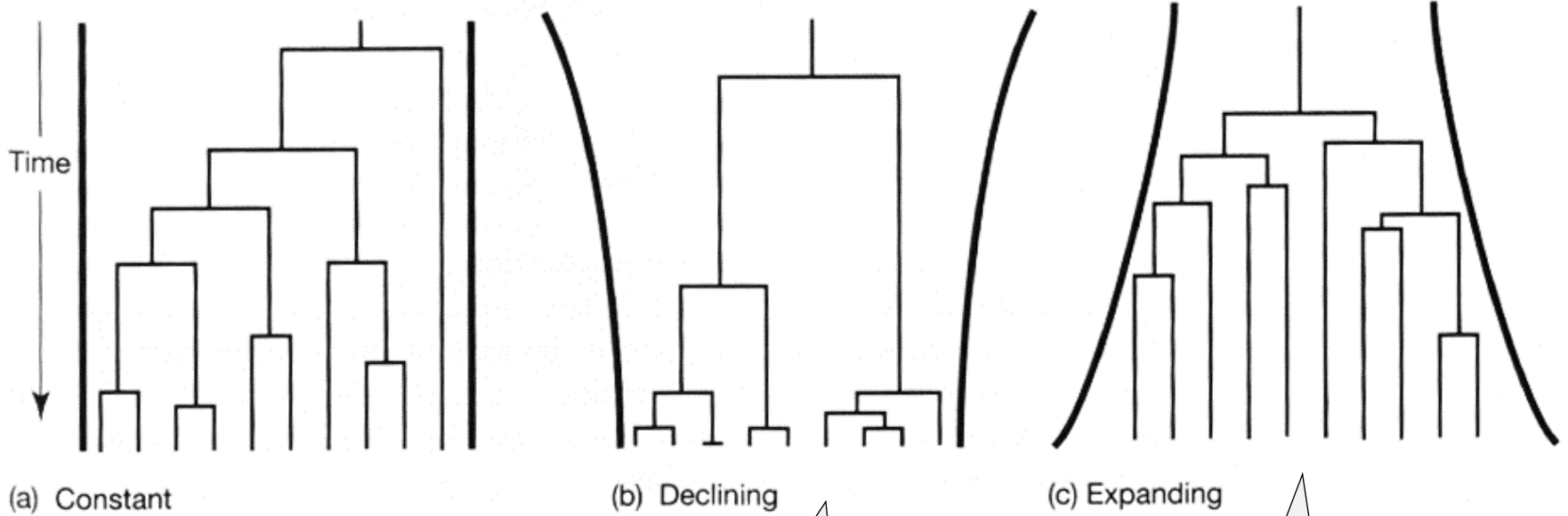
comparison of stable population model and model of exponential growth/decline using LRT with 1 degree of freedom

program Fluctuate:

growth parameter g
both ML and BA approach



4. Bayesian Skyline Plot (BSP)



declining population:
coalescent rate
increases

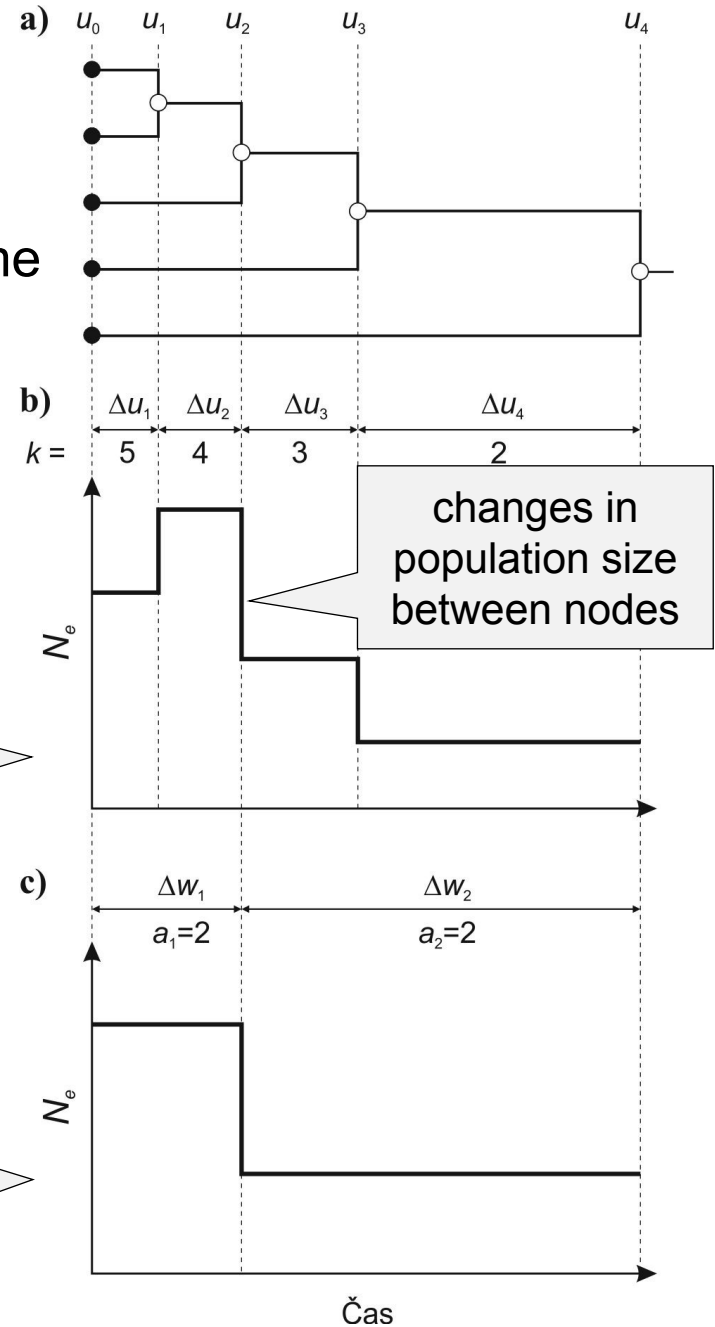
growing population:
coalescent rate
decreases

Bayesian skyline plot

distribution of genealogical lineages in time

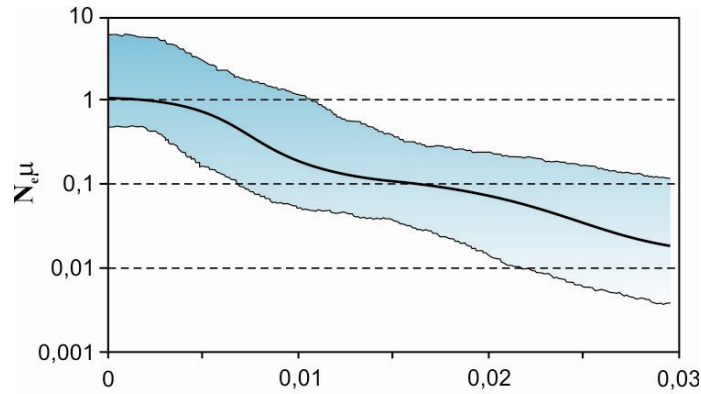
BSP is based on this approach

programs BEAST/Tracer



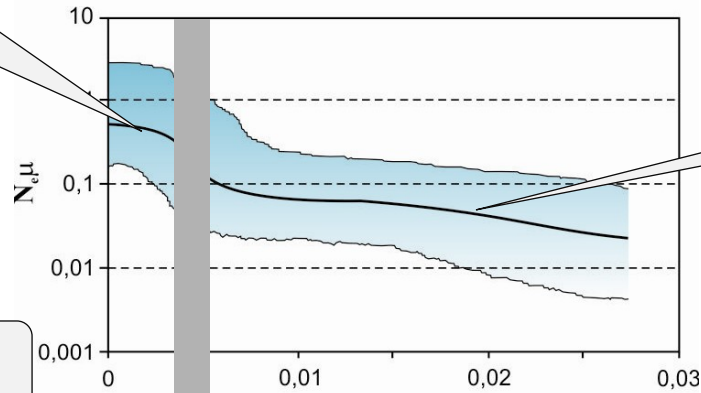
classical
BSP

generalized
BSP



domesticus

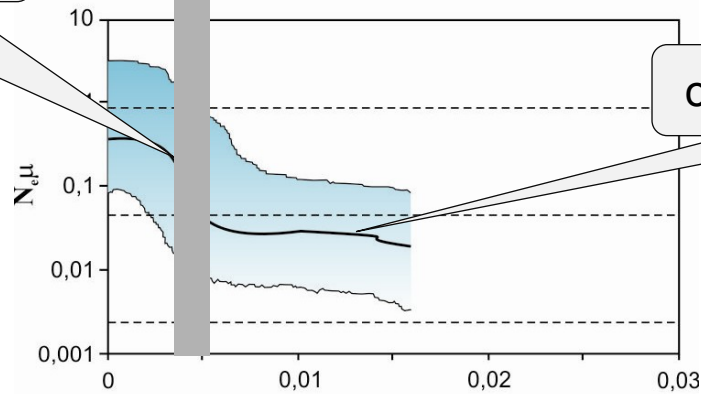
expansion to Europe



domesticus - Europe

origin outside Europe

expansion to Europe

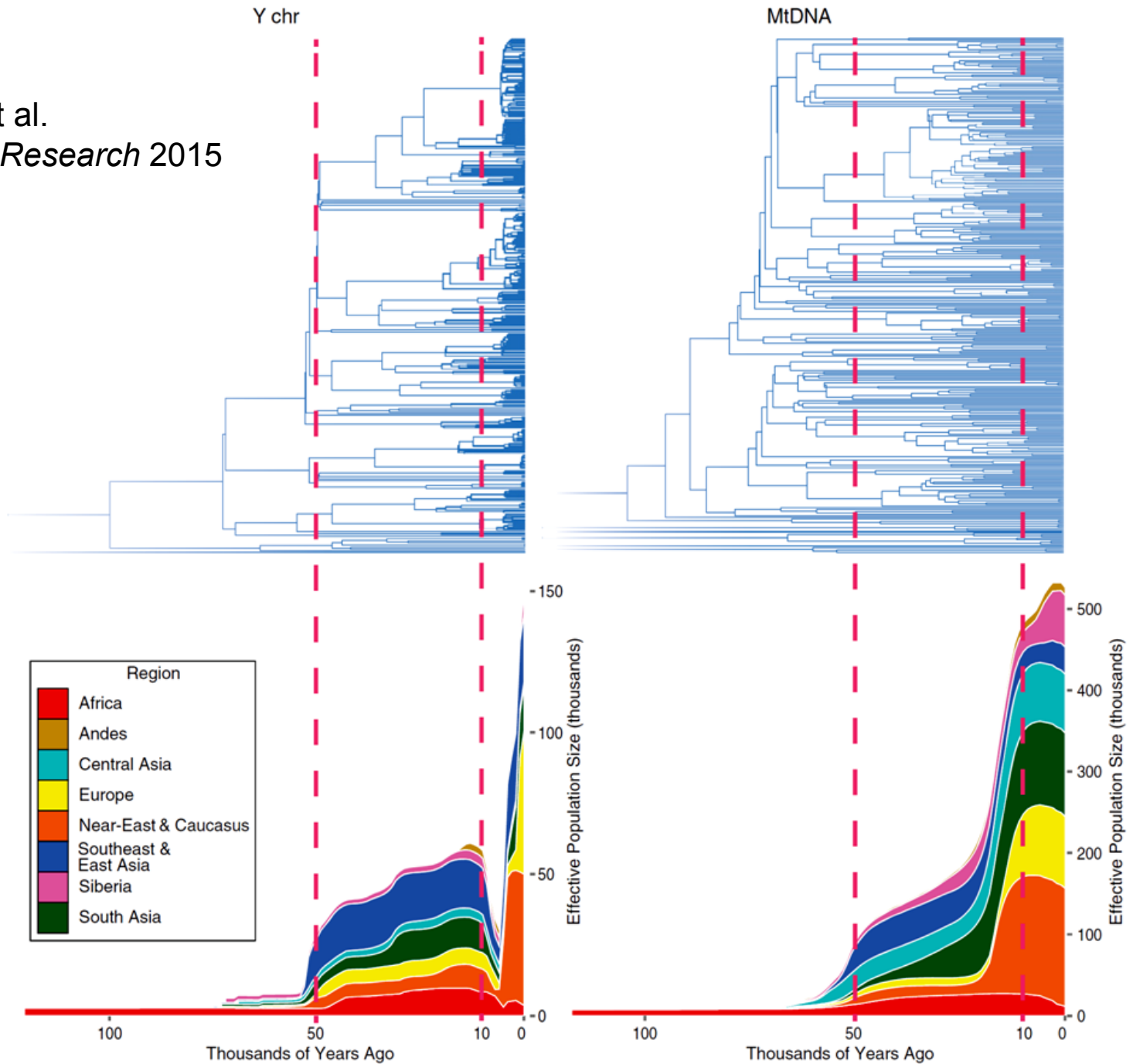


musculus - Europe

origin outside Europe

Čas (mil. let)

Karmin et al.
Genome Research 2015



Possible results of phylogeographical studies

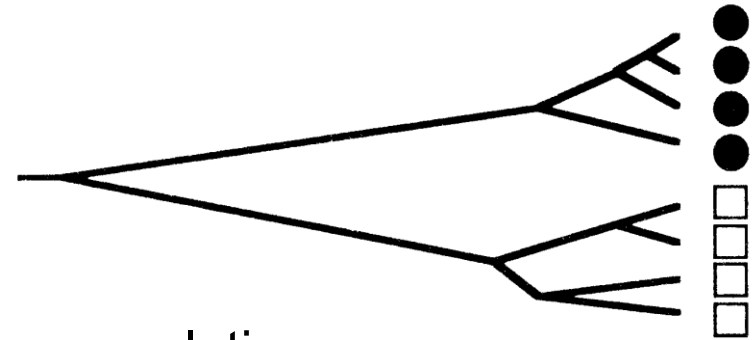
(Avice 2000)

Category I:

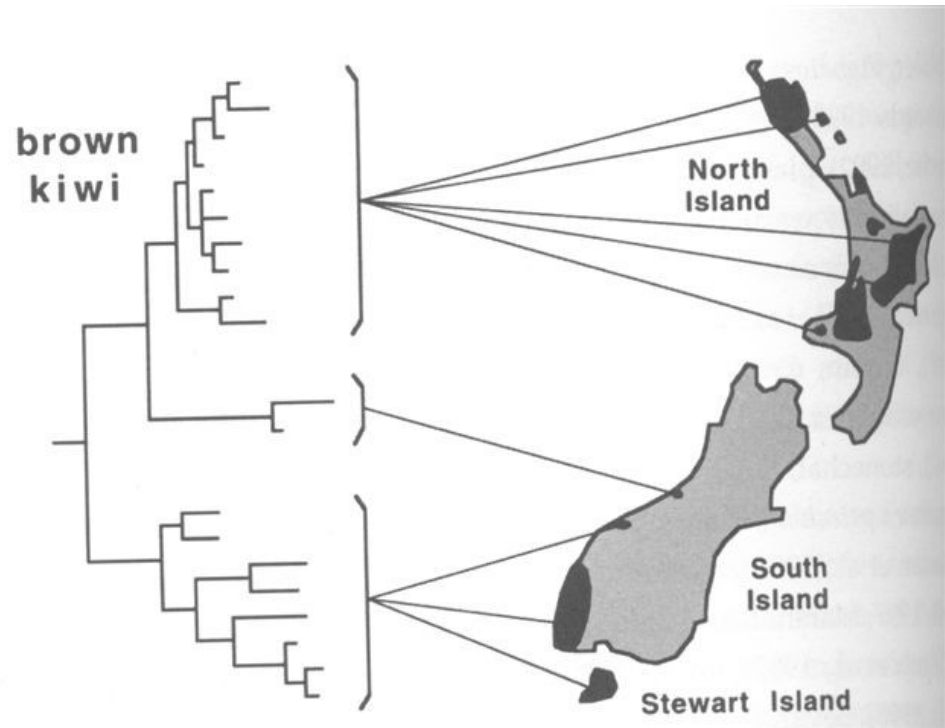
distinct allopatric lineages

barriers to gene flow or low dispersion

differences because of lineage sorting, or accumulation of new mutations

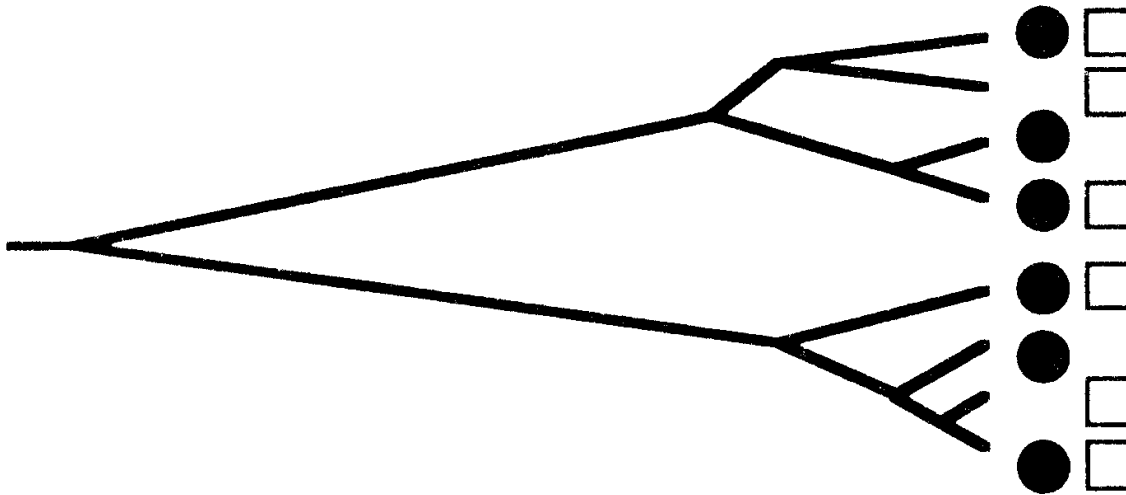


Apteryx australis



Category II:

sympatric, but deep lineages \Rightarrow secondary contact of previously separated populations



Category III:

allopatric, only slightly separated lineages

closely related, but geographically localized haplotypes

recently, populations in contact

but: gene flow sufficiently low

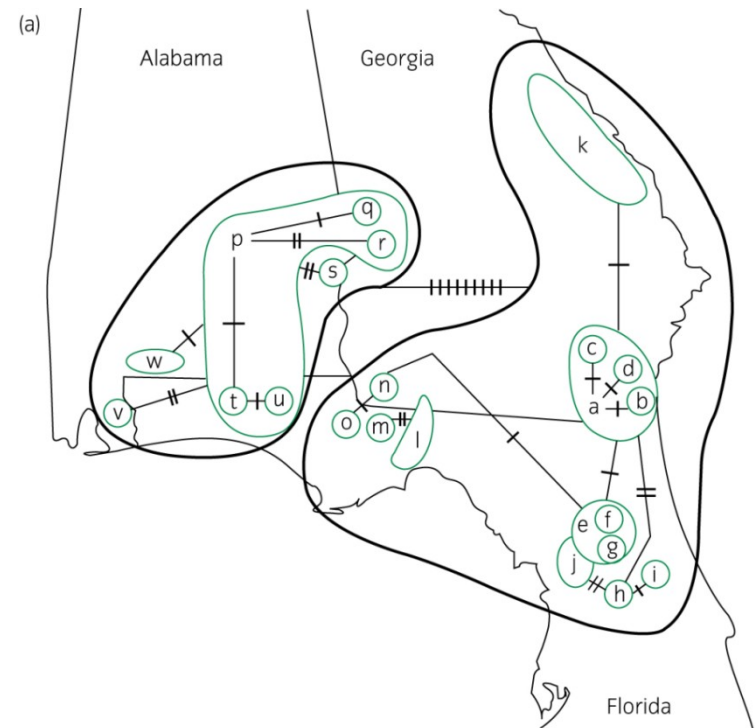
→ drift and lineage sorting → divergence of populations

often:

Category I on coarse scale

Category III on fine scale

eg.: *Geomys pinetis*



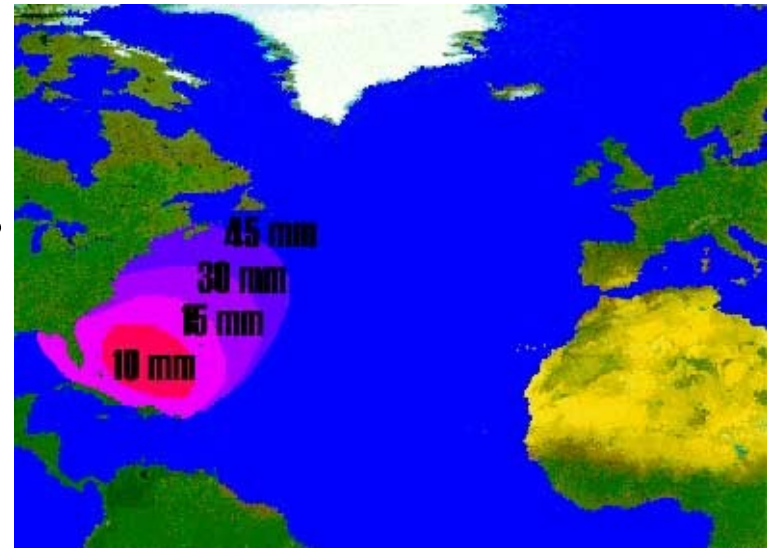
Category IV:

sympatric, only slightly separated lineages

strong gene flow

absence of geographic barriers or

recent expansion



Anguilla rostrata

Random dispersion of larvae

Panmictic aggregation
during spawning

Category V:

combination of III and IV

low divergence of lineages

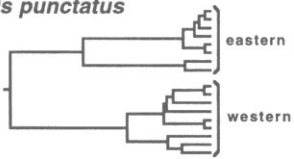
some lineages widely distributed (likely ancestral), others (new)
geographically limited

we should use private haplotypes as characters

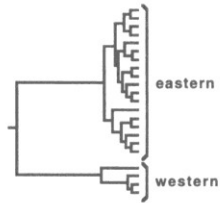
Genealogical concordance

Fishes in SE USA

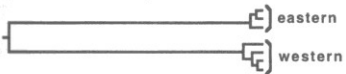
Lepomis punctatus



Gambusia affinis/ G. holbrooki



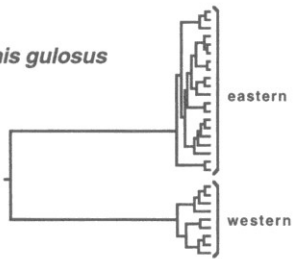
Lepomis microlophus



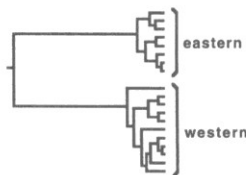
Amia calva



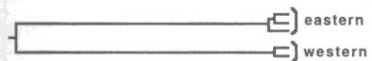
Lepomis gulosus



Micropterus salmoides



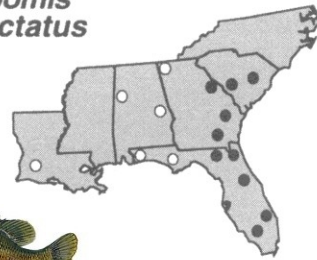
Lepomis macrochirus



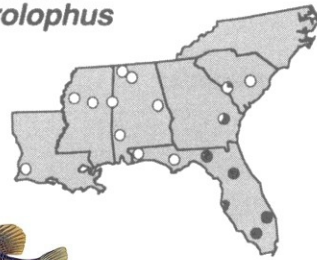
8.0 6.0 4.0 2.0 0.0
sequence divergence (%)

8.0 6.0 4.0 2.0 0.0
sequence divergence (%)

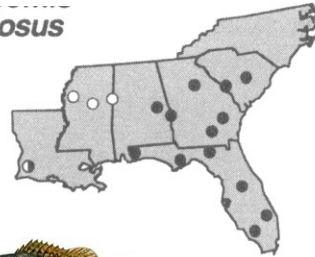
Lepomis punctatus



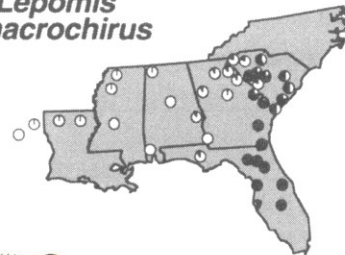
Lepomis microlophus



Lepomis gulosus



Lepomis macrochirus



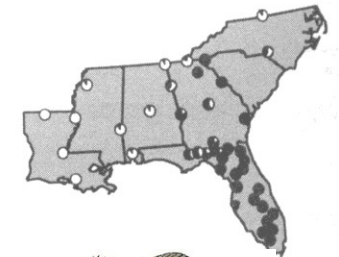
Gambusia affinis, G. holbrooki



Amia calva



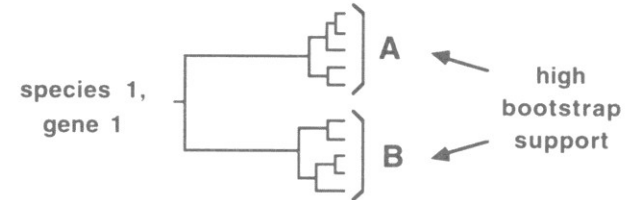
Micropterus salmoides



Genealogical concordance (congruence on different levels)

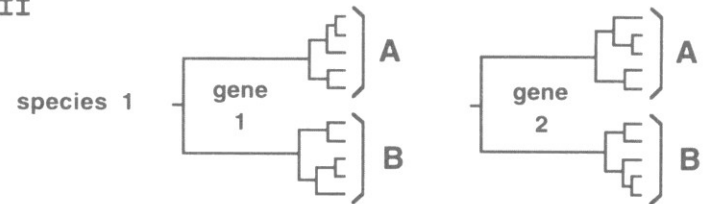
Various parts of gene sequence →

Aspect I



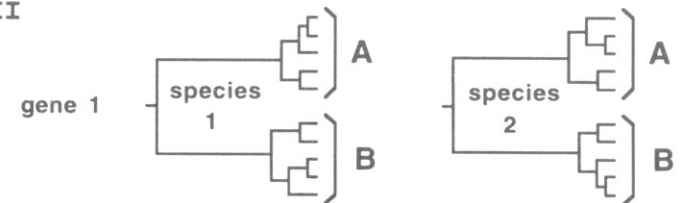
More sequences (genes) of the same species →

Aspect II



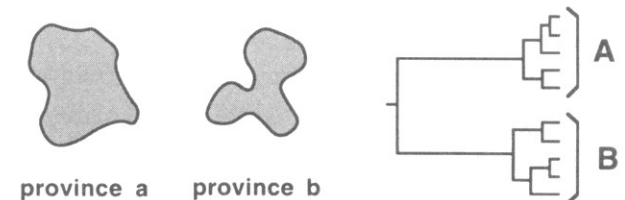
More species in the same region →

Aspect III



Support of biogeographical regions (more species, more areas) →

Aspect IV



Genetic consequences of glaciations

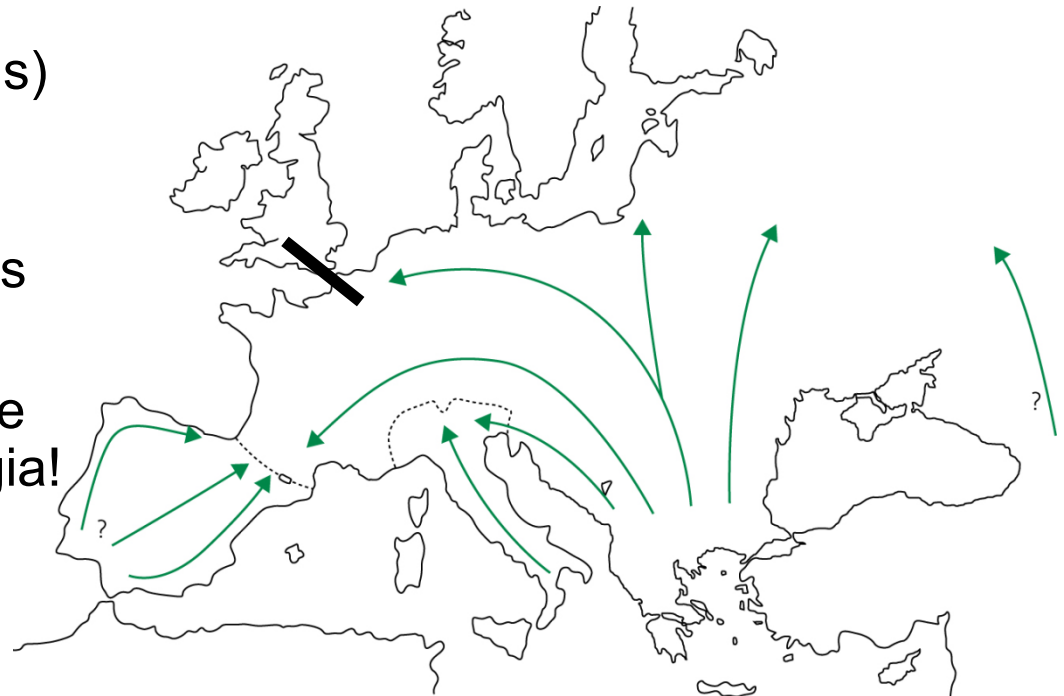
Refugia (Iberian, Apennine, Balkan peninsulas)

In refugia, small populations during relatively long time

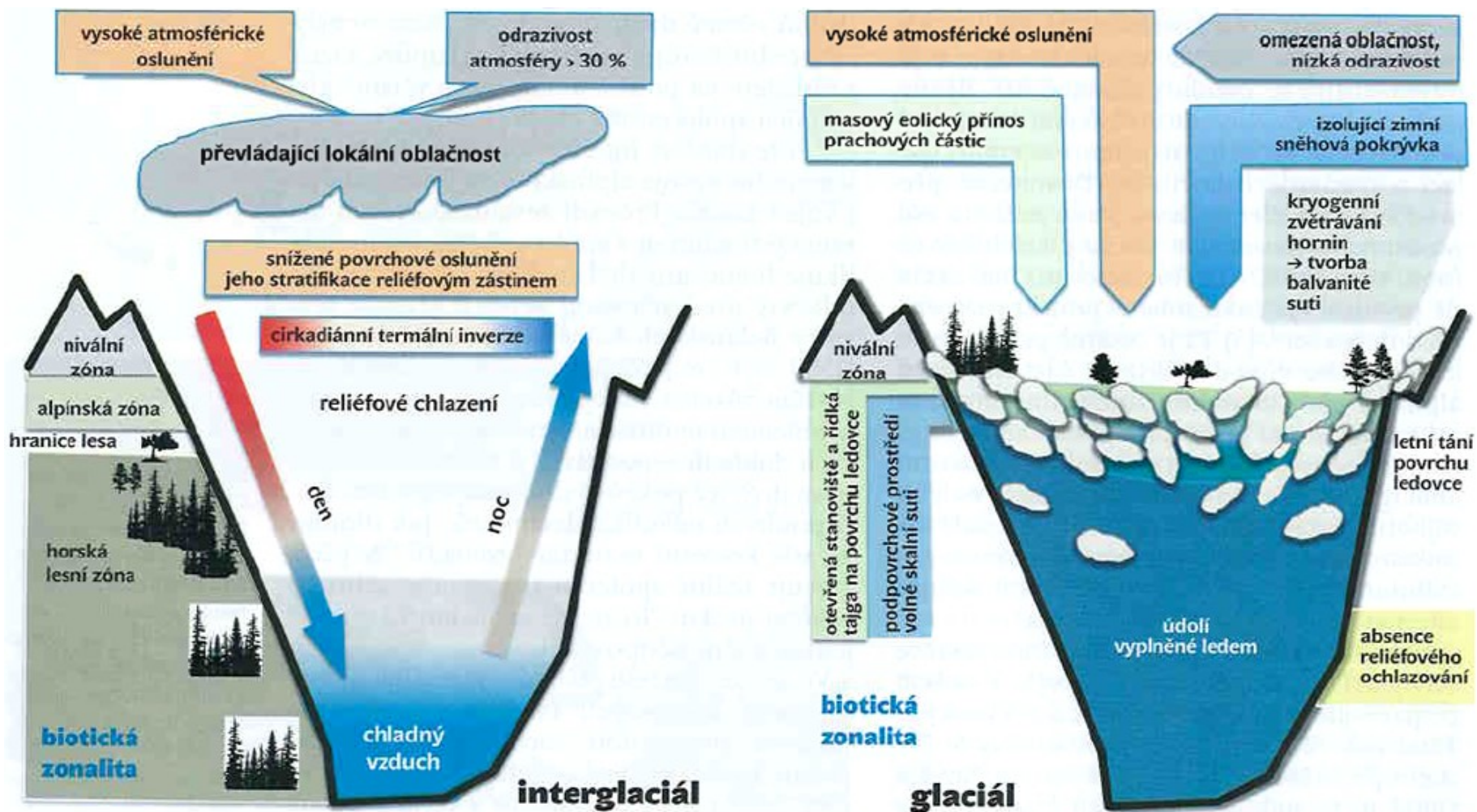
Lineage sorting (+ mutations)

Subsequent expansion →
intraspecific hybrid zones

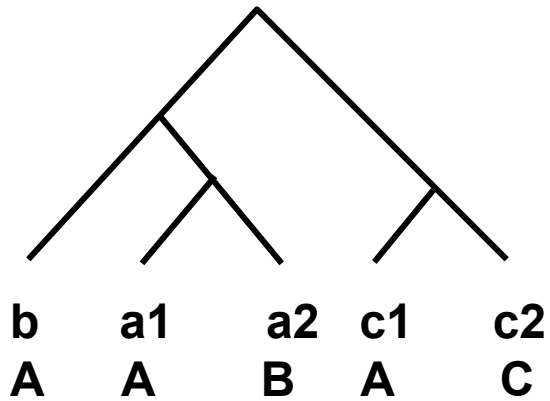
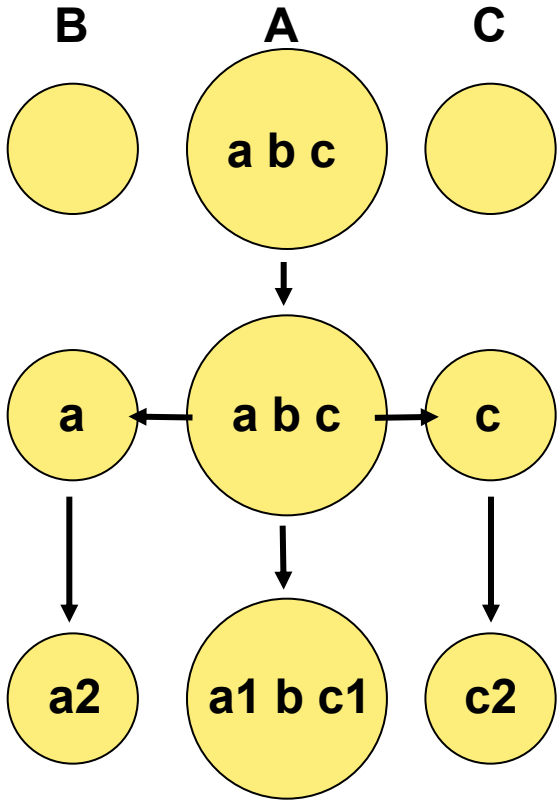
But in several species, there
were also northern refugia!



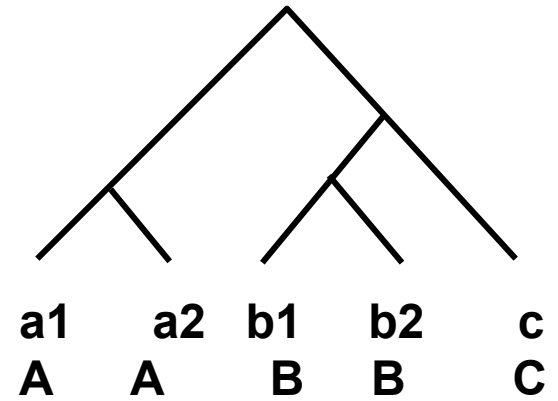
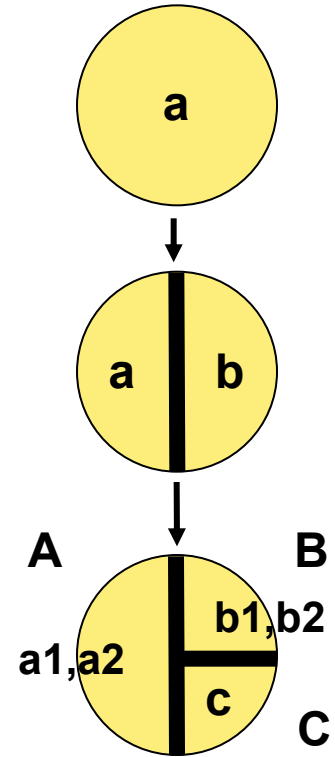
Chorthippus parallelus



dispersal



vicariance



Relationship between genetic population structure, sex-specific dispersal and gene flow regimes (Avice 2000)

female dispersal and gene flow
 low → high

male dispersal and gene flow

low
↓
 high

<p style="color: blue;">geographic structure in:</p> <p>mtDNA YES</p> <p>autosomes yes</p> <p>chr. Y yes</p>	<p style="color: blue;">geographic structure in:</p> <p>mtDNA NO</p> <p>autosomes yes</p> <p>chr. Y ***</p>
<p style="color: blue;">geographic structure in:</p> <p>mtDNA (in females) YES</p> <p>autosomes no</p> <p>chr. Y no</p>	<p style="color: blue;">geographic structure in:</p> <p>mtDNA NO</p> <p>autosomes no</p> <p>chr. Y no</p>

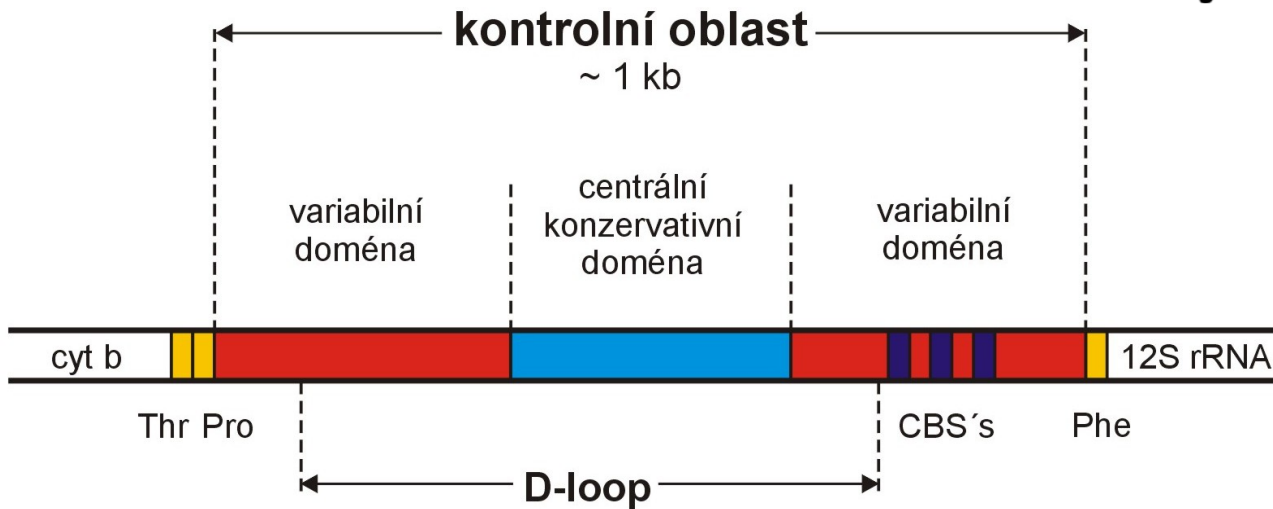
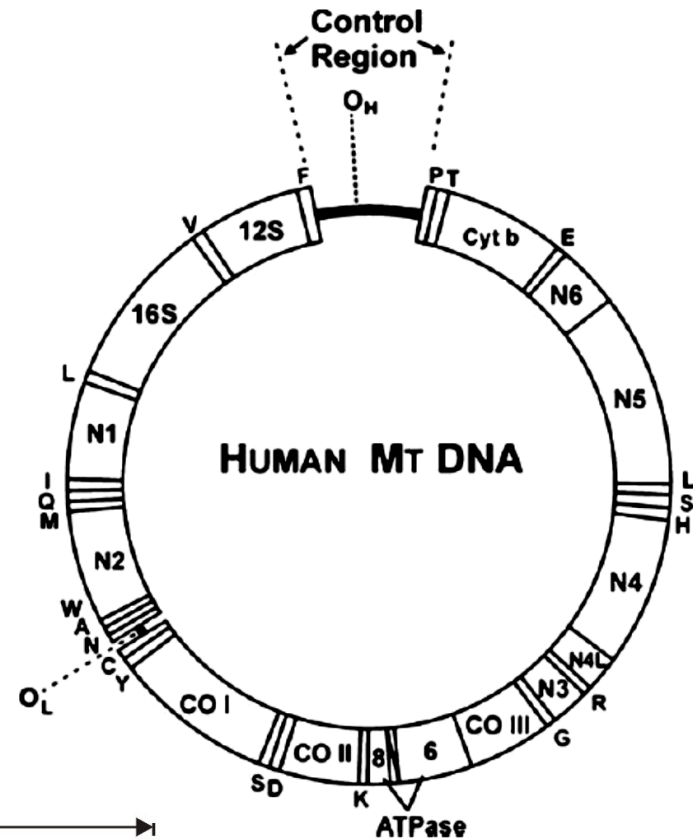
markers:

mtDNA sequences

Y chr. sequences

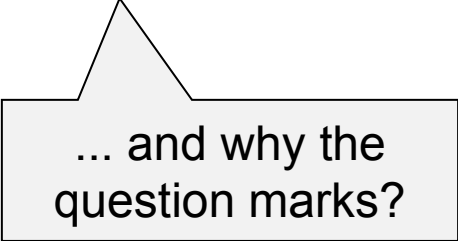
microsatellites

SNP



Why mtDNA advantageous?

- ? Small (15-20 kb), circle molecule
- ? Without introns
- ? Minimum of non-coding regions
- ? Uniparental (maternal)
- ? Non-recombining
- ? Only one type in many copies in the cell
- ? Neutrality (same fitness of different variants)



... and why the question marks?

Problems for population genetics:

Neutrality

Interspecific transmission

Nuclear pseudogenes

Biparental inheritance

Recombination

Neutrality?

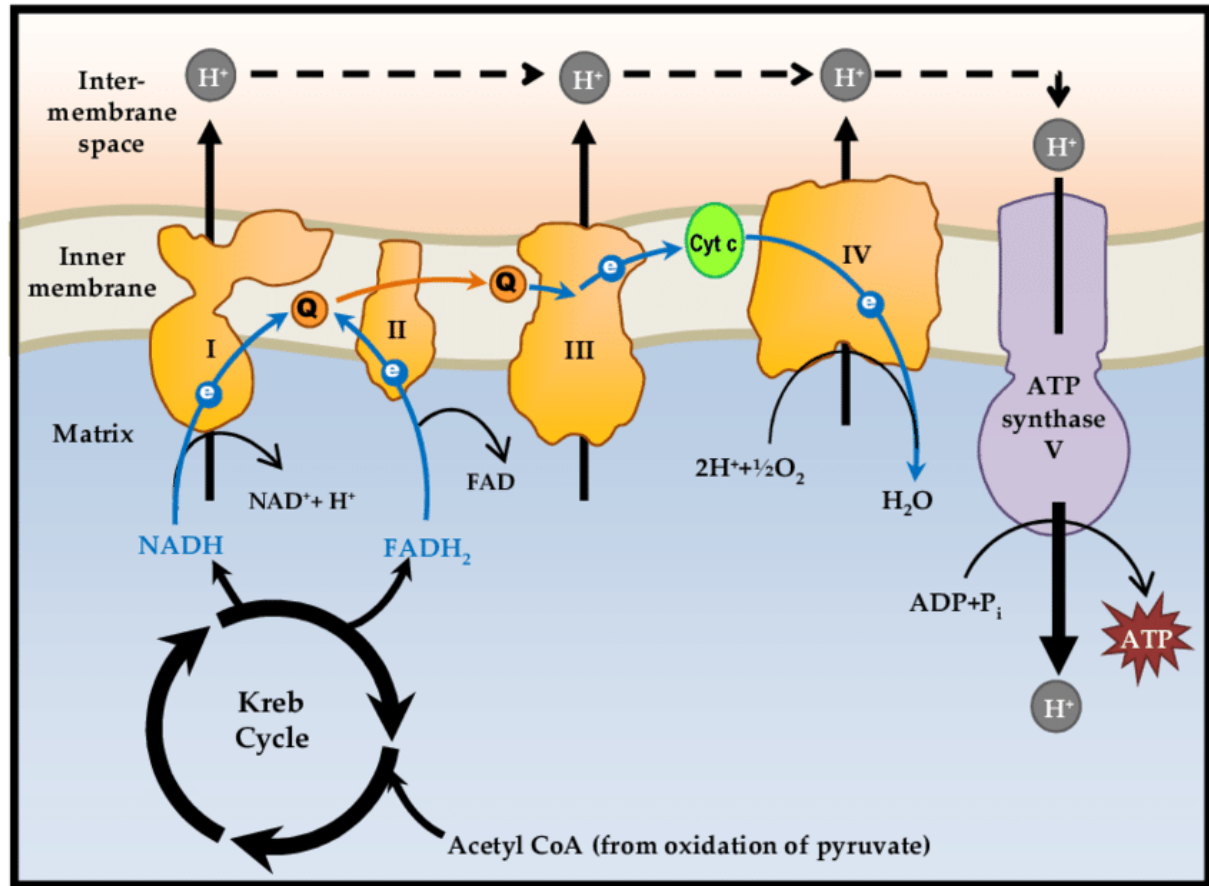
influence on fitness (experimental evidence):

mouse (*Mus*)

fruit fly (*Drosophila*)

human

OXPPOS



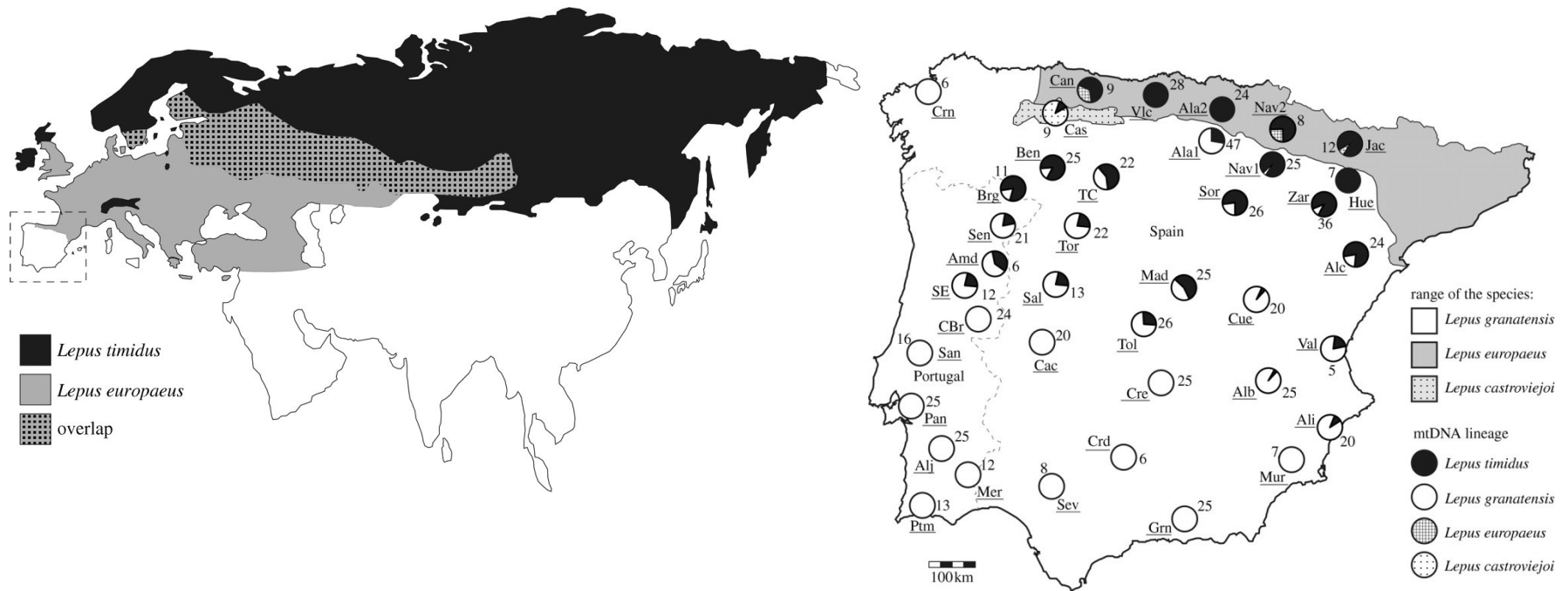
Interspecific introgression:

hairs in Spain:

presence of *Lepus timidus* mtDNA in *L. granatensis*, *L. castroviejoi* and *L. europaeus*

however, *L. timidus* disappeared at the end of the last glacial; multiple transmission of various mtDNA lineages

= mtDNA capture



Nuclear Mitochondrial DNA = NUMT:

copies of mtDNA segments integrated to nuclear DNA

loss of function

molecular fossils

similarity with original sequence → risk of amplification instead of mtDNA
⇒ problem!!

various appearance in different groups and different species within the groups

eg.: numt > 12,5 kb in 7 felid species

humans: 27 numts after split from chimpanzee lineage

What to do?

ultracentrifugation (usually fresh samples needed, or at least deep-frozen)

tissues with large number of mitochondria (eg. muscles)

long-range PCR

RT-PCR

electronic PCR (in species with known genomes)

Recombination of mtDNA:

necessary conditions:

- biparental inheritance – fusion of mitochondria

- existence of protein machinery for recombination: also in humans

biparental inheritance:

- despite myths, father's mitochondria usually transmitted to the zygote, where they are labelled and subsequently eliminated (in mammals, mitochondria are labelled by father's nuclear genes)

→ in some species paternal leakage: *Mus*, *Drosophila*, *Parus*, *Homo*

Recombination of mtDNA:

biparental inheritance:

Gyllensten et al., 1991: Paternal inheritance of mitochondrial DNA in mice. *Nature* 352: 255–257.

F1 hybrids *Mus spretus* × C57BL

frequency of paternal mtDNA relative to maternal $\approx 10^{-4}$

Maternal Inheritance of Mouse mtDNA in Interspecific Hybrids: Segregation of the Leaked Paternal mtDNA Followed by the Prevention of Subsequent Paternal Leakage

Hiroshi Shitara,^{*,†} Jun-Ichi Hayashi,^{*} Sumiyo Takahama,[†] Hideki Kaneda[†] and Hiromichi Yonekawa[†]

Shitara et al., 1998: *Genetics* 148: 851–857.

F1 hybrids *Mus spretus* × C57BL

leakage of paternal mtDNA not in all tissues

only in F1, not in subsequent generations (in backcrosses) → species-specific exclusion