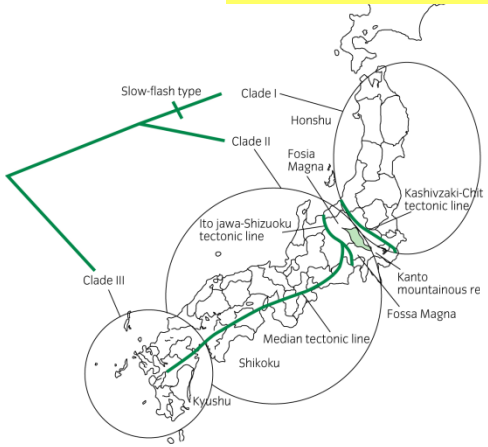
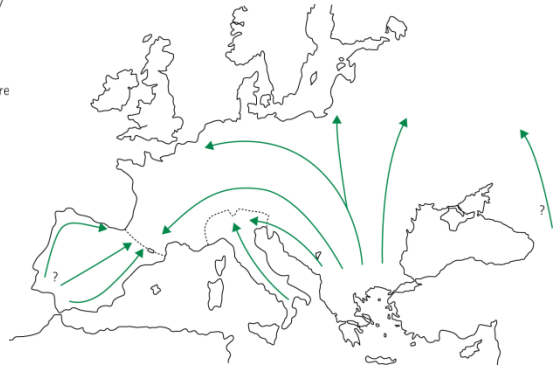
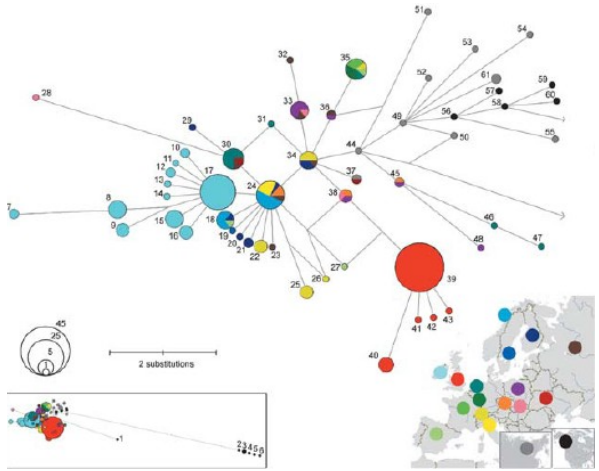
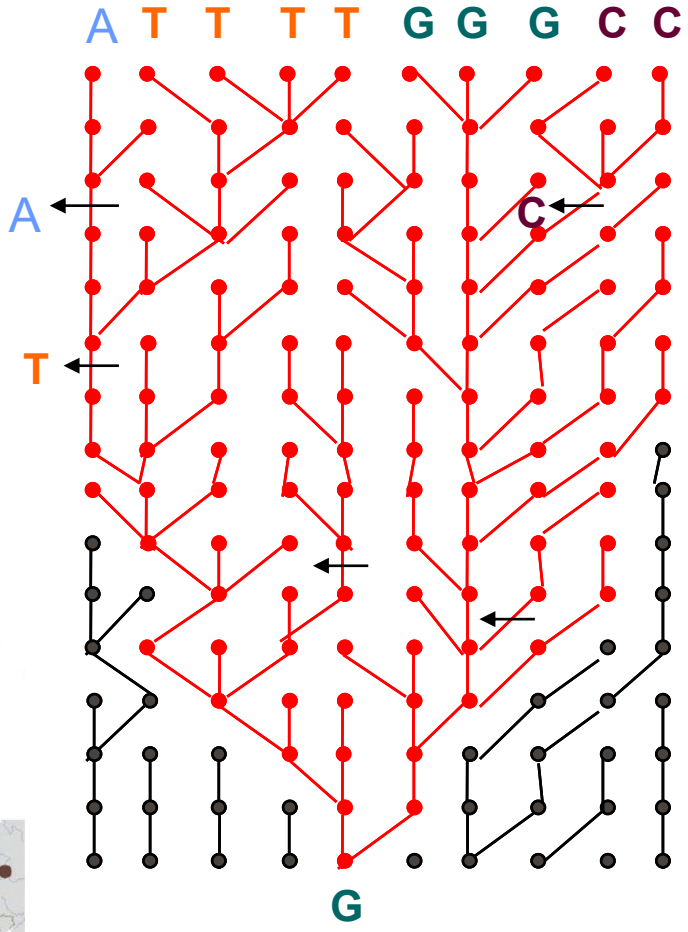
COALESCENT AND PHYLOGEOGRAPHY

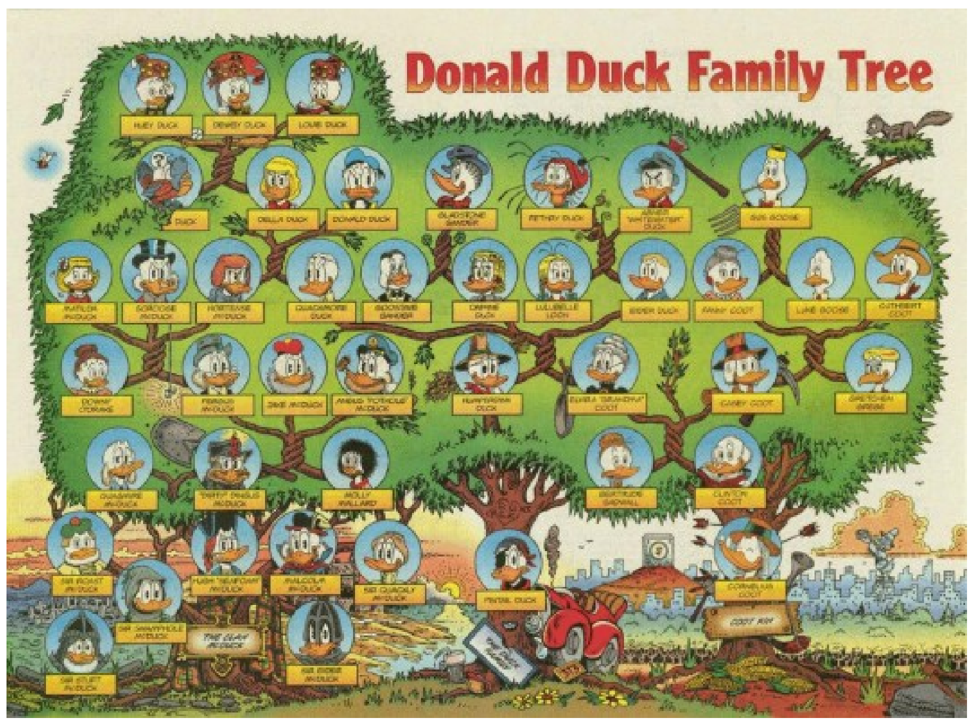








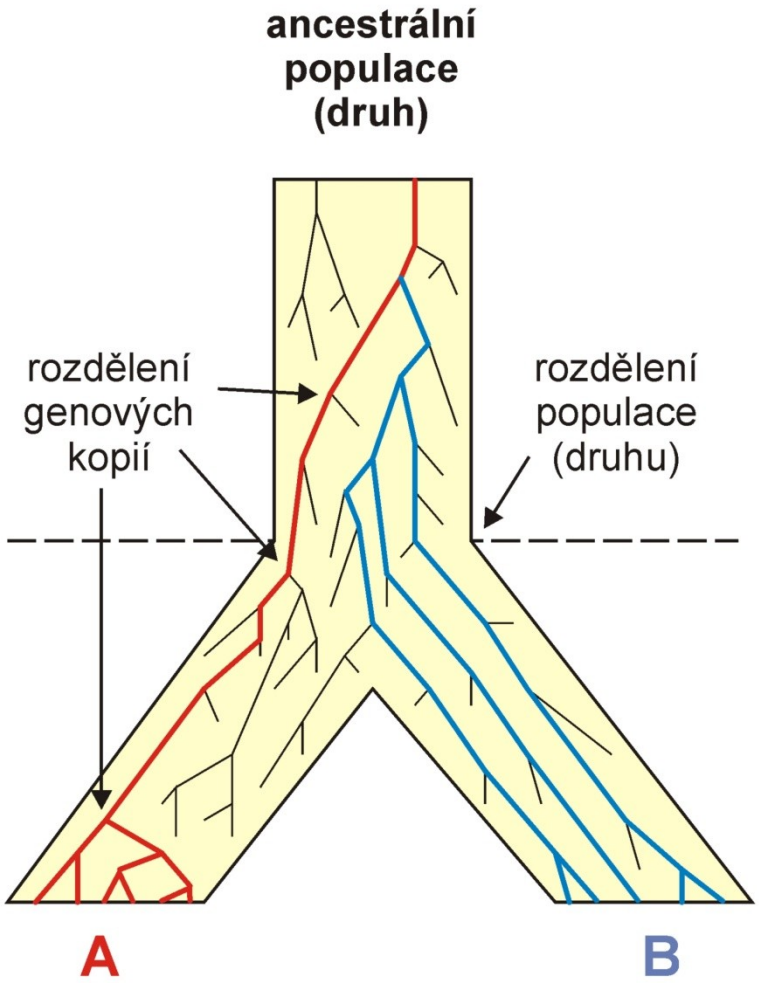
Genealogická tabule panovnického rodu Přemyslovců a jejich potomků na českém trůně
Sestaveno k 700.výročí úmrtí Václava III., posledního přemyslovského panovníka, krále českého, polského a uherského dne 4.8.1306
letopočet
generace



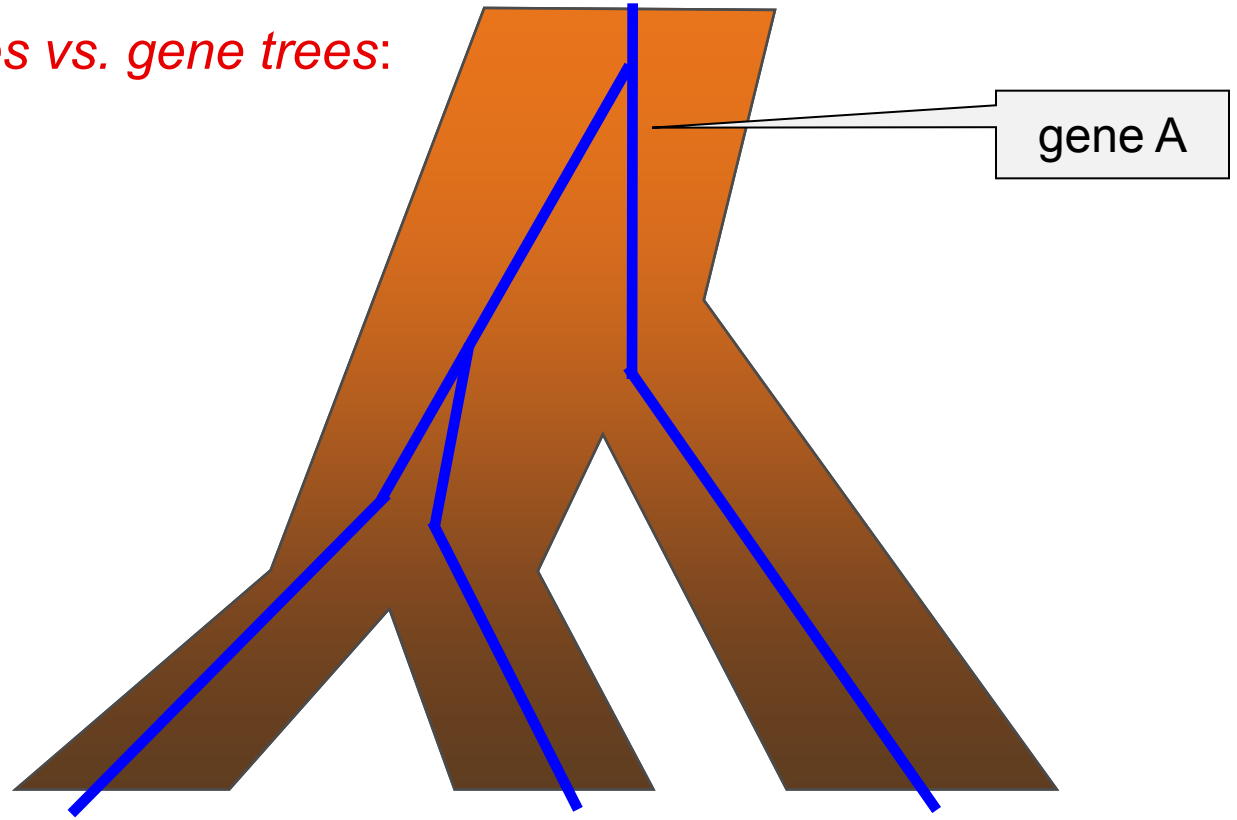
Donald Duck Family Tree
HUEY DUCK
DEWEY DUCK
LOUIE DUCK
DELLA DUCK
DONALD DUCK
GLADSTONE GANDER
FETHRY DUCK
GUS GOOSE
SCROOGE McDUCK
HORTENSE McDUCK
QUACKMORE DUCK
LULUBELLE LOON
EIDER DUCK
FANNY COOT
LUKE GOOSE
CUTHBERT COOT
DOWNY O'DRAKE
FERGUS McDUCK
JAKE McDUCK
CASEY COOT
GRETCHEN GREBE
MOLLY MALLARD
PINTAIL DUCK
CORNELIUS COOT

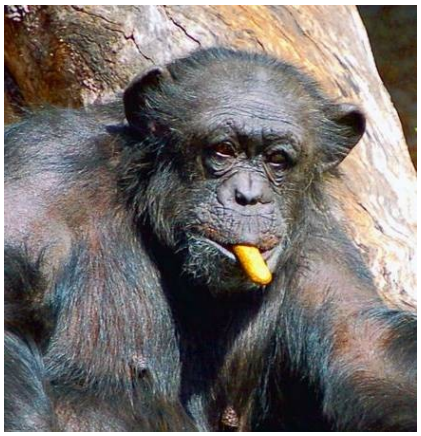
COALESCENCE

Fate of individual gene copies in the population → gene trees

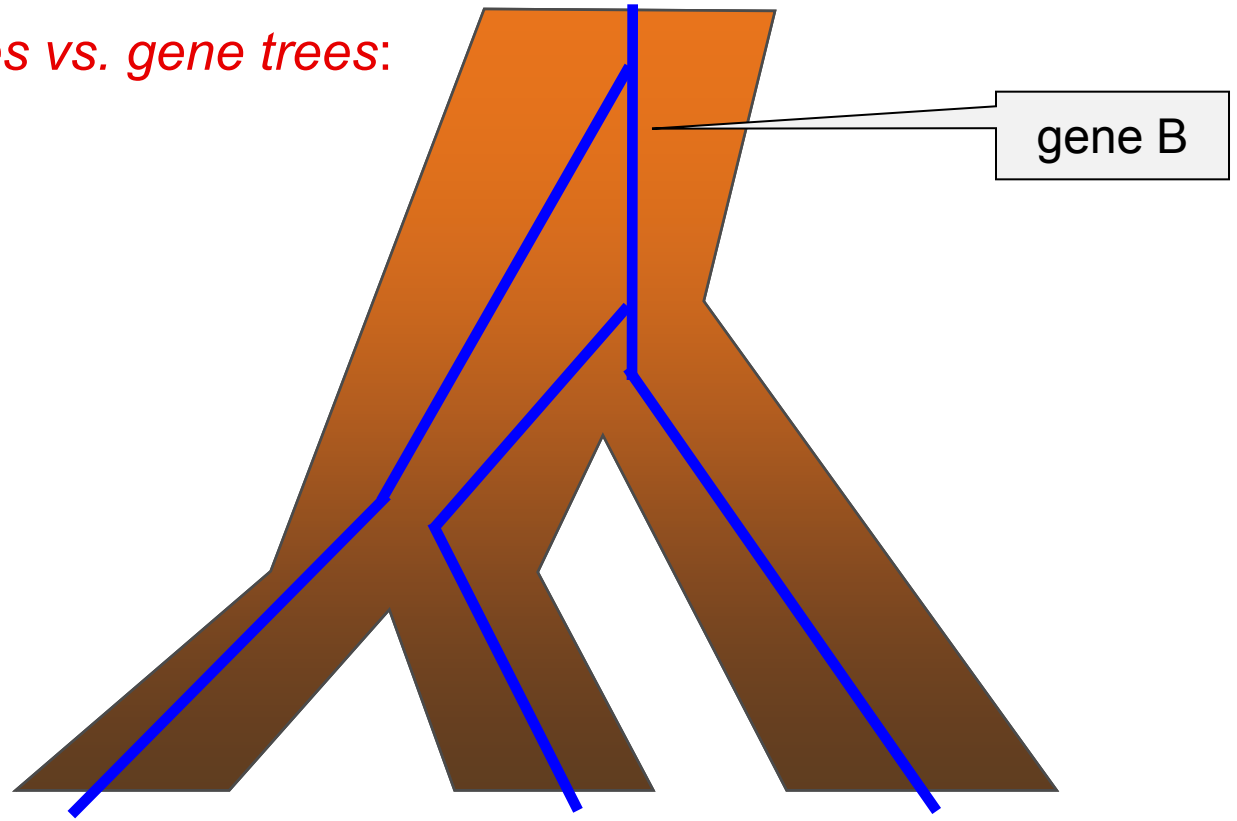


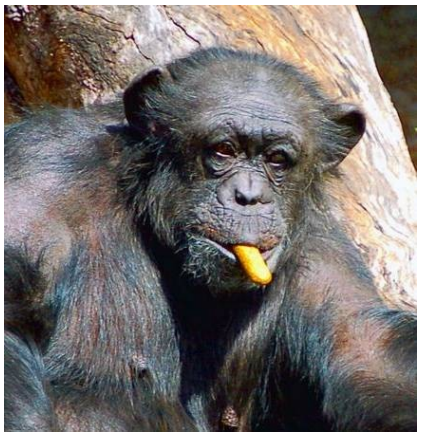
Species trees vs. gene trees:



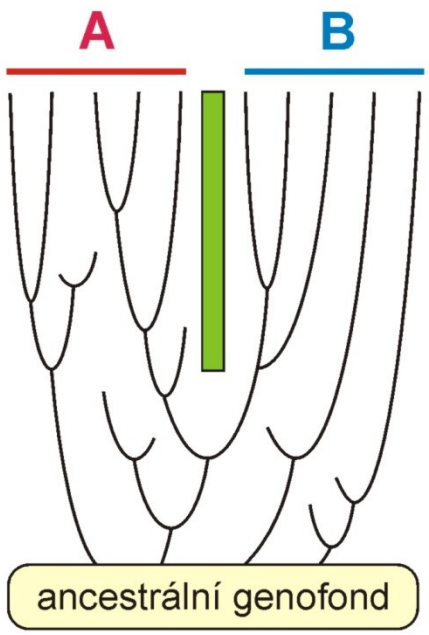


Species trees vs. gene trees:

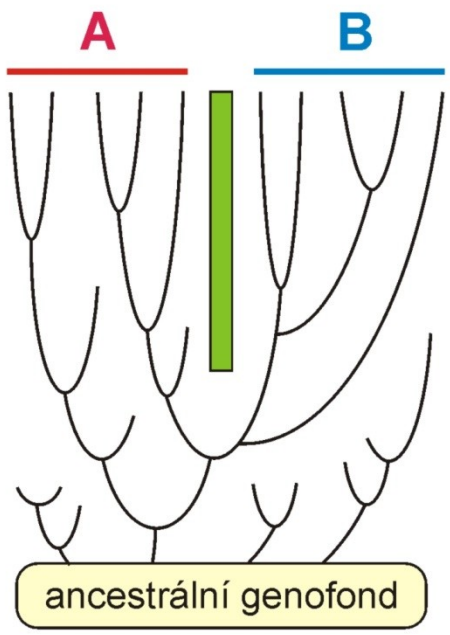




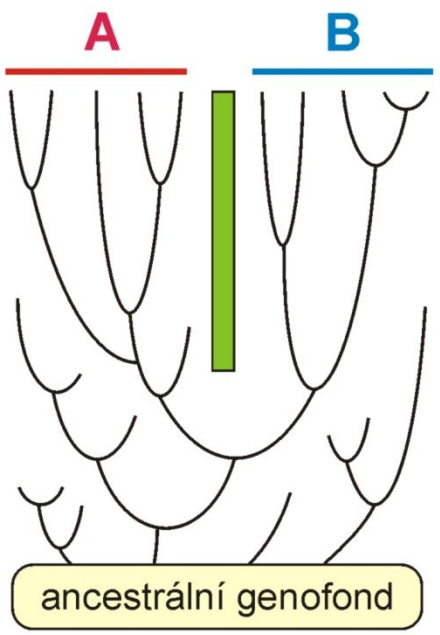
Phylogenetic relationships of 2 descendant populations (eg. mtDNA):



polyphyly

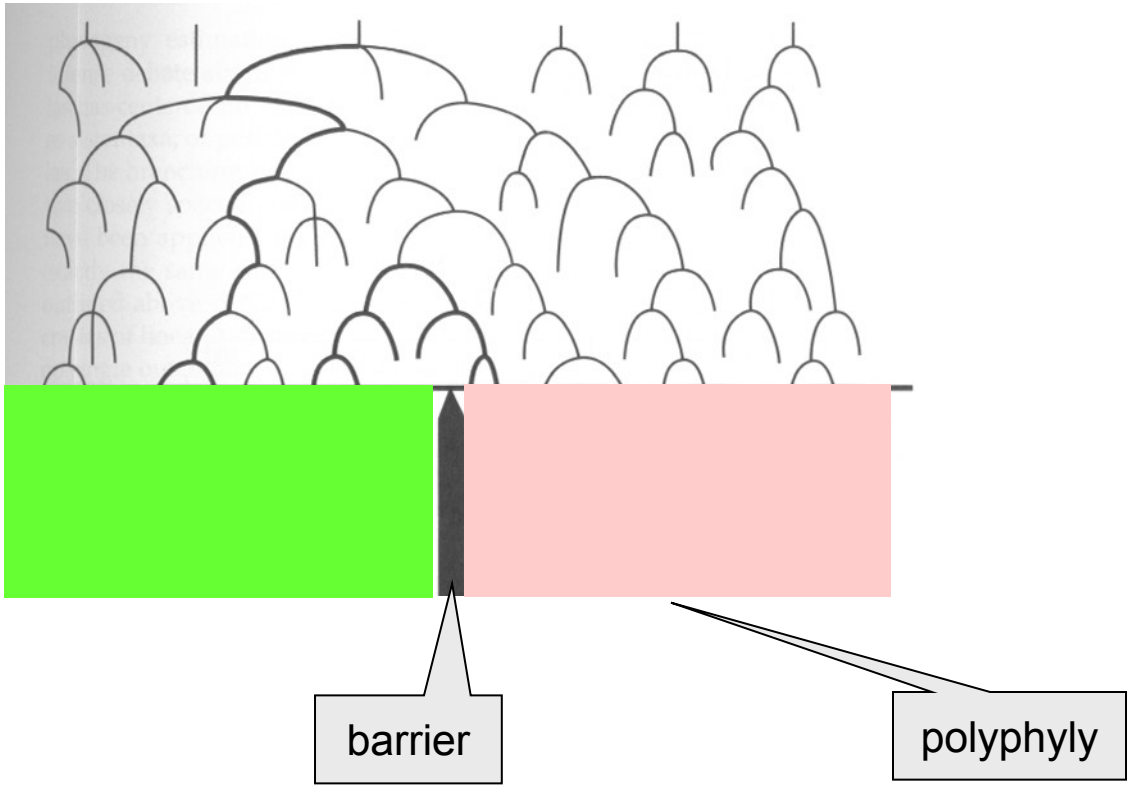


paraphyly

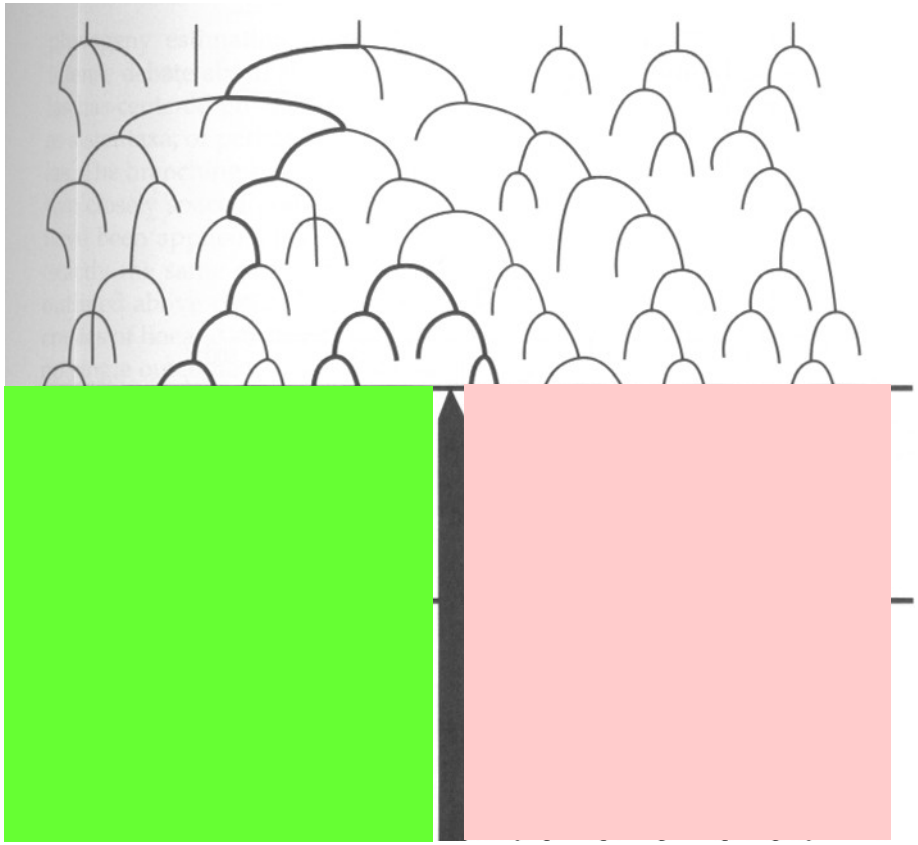


reciprocal monophyly

Ancestral polymorphism and lineage sorting

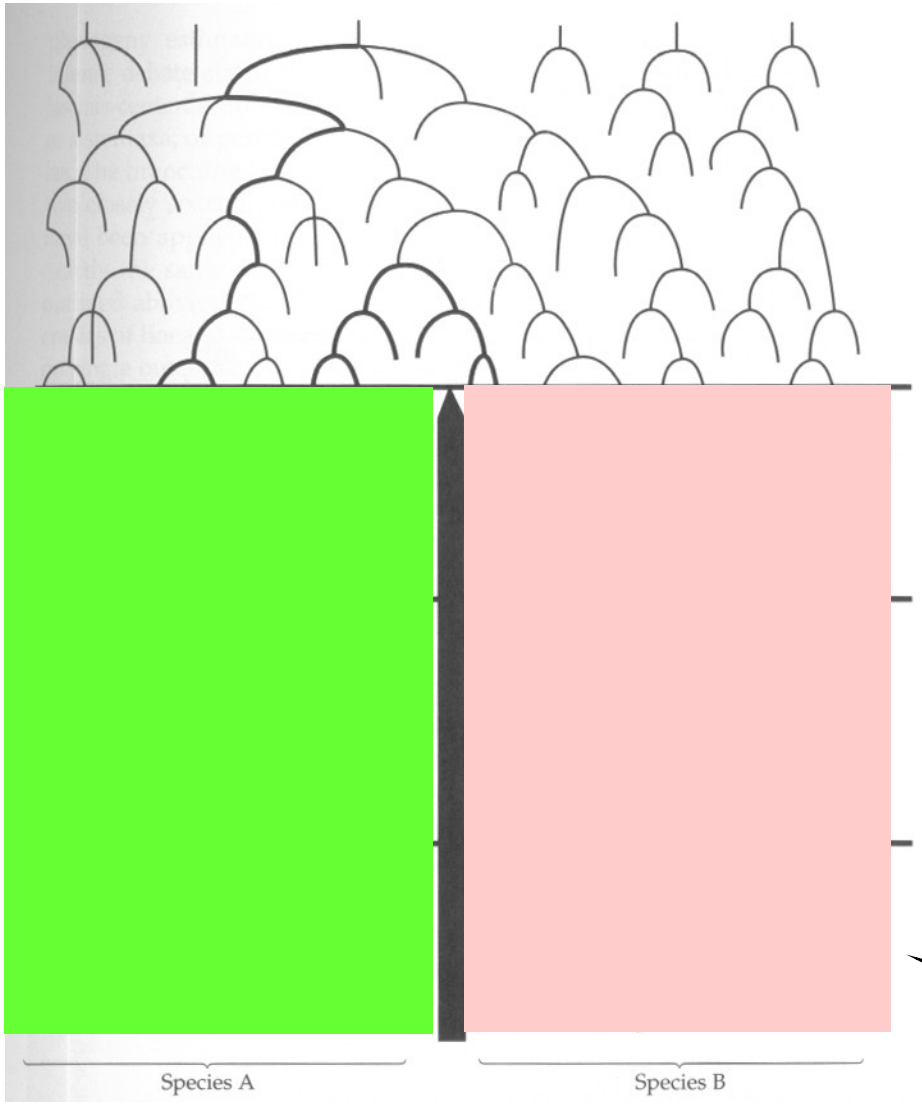


Ancestral polymorphism and lineage sorting



paraphyletic stage

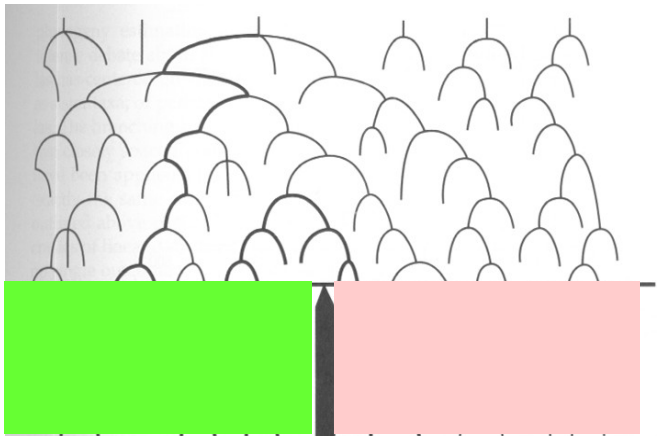
Ancestral polymorphism and lineage sorting



reciprocal monophyly

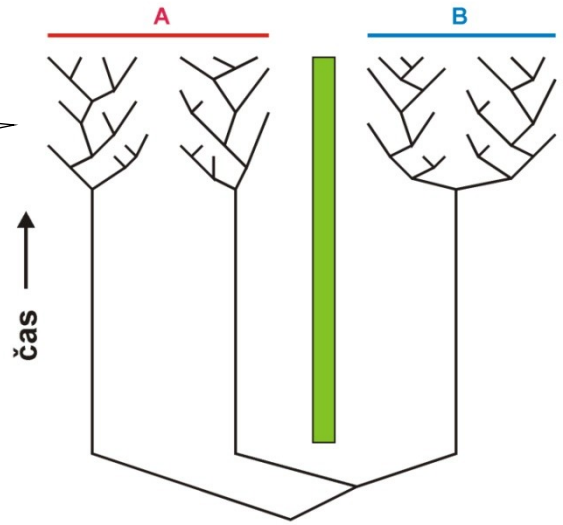
species A **species B**

Ancestral polymorphism and lineage sorting

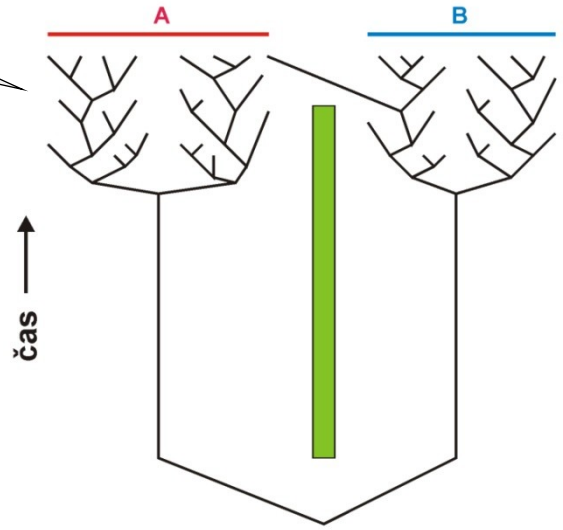


druh A druh B

incomplete lineage sorting

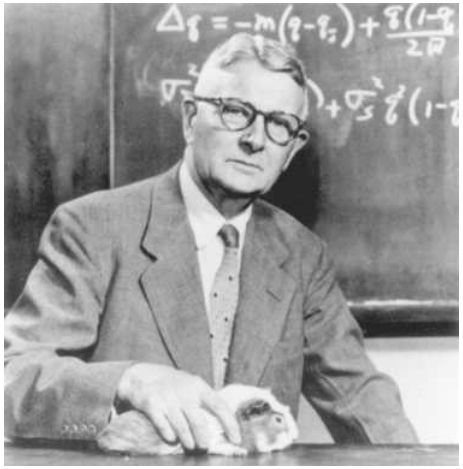


recent gene flow

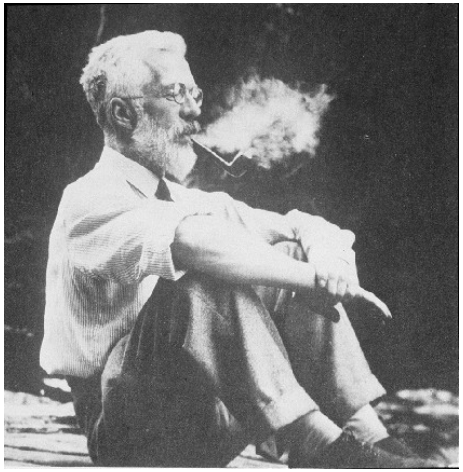


Problem: it is often difficult to distinguish between incomplete lineage sorting and consequences of gene flow

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

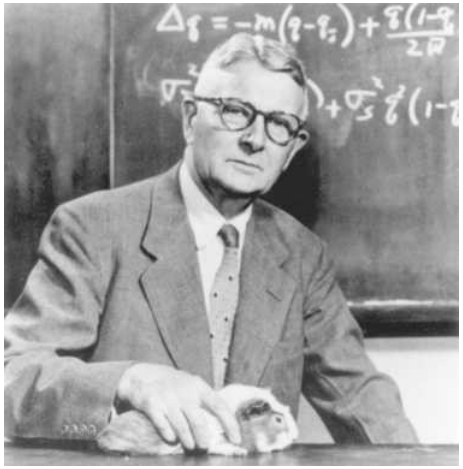
no age structure

no selection

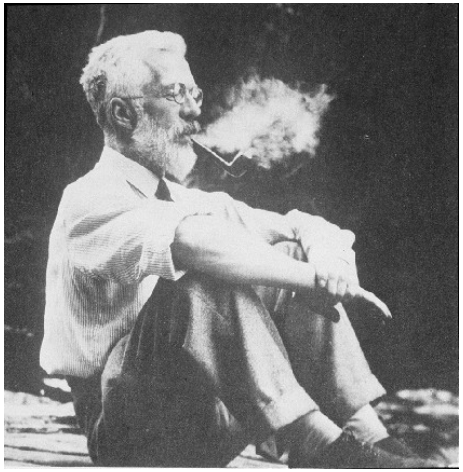
variance of gamete sampling

$\rightarrow$ Poisson distribution

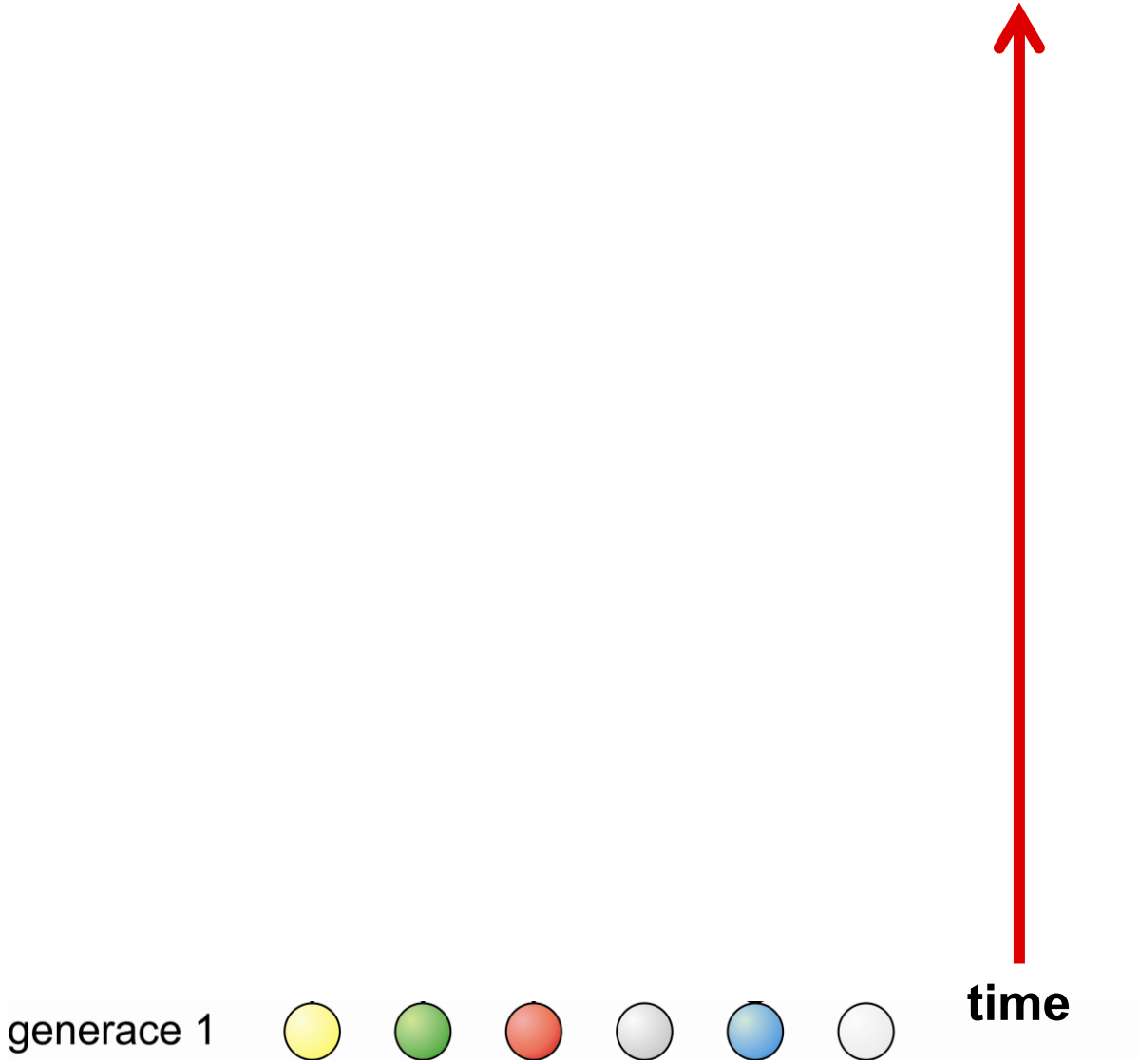
Lineage sorting in W-F model:



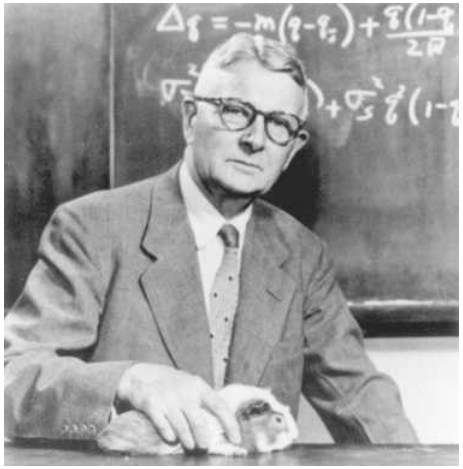
Sewall Wright



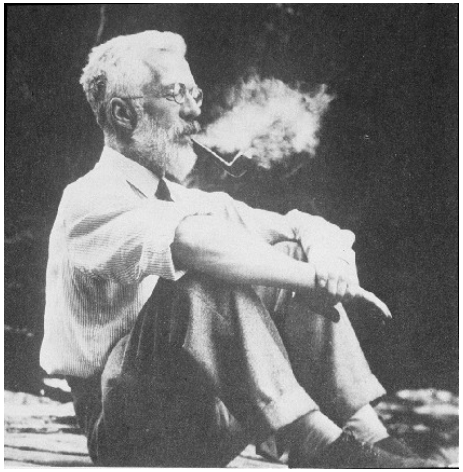
Ronald A. Fisher



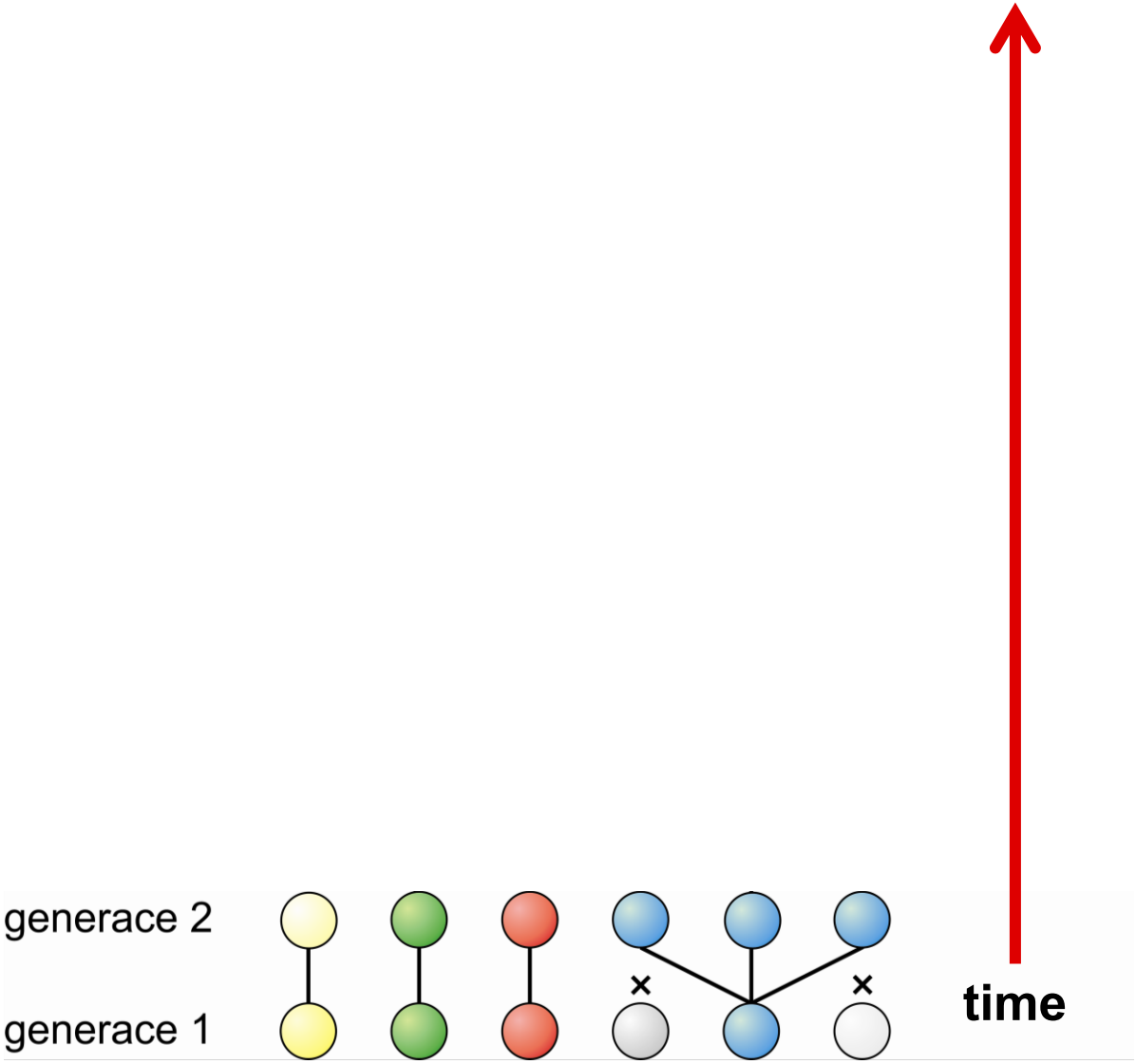
Lineage sorting in W-F model:



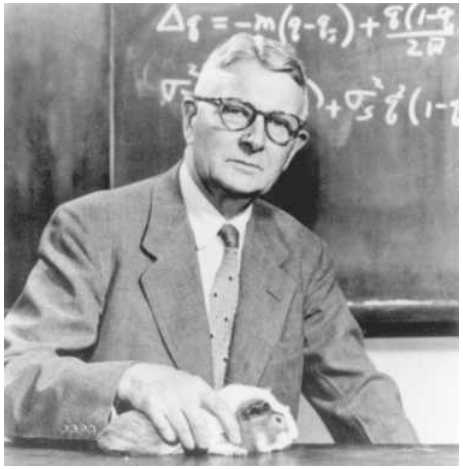
Sewall Wright



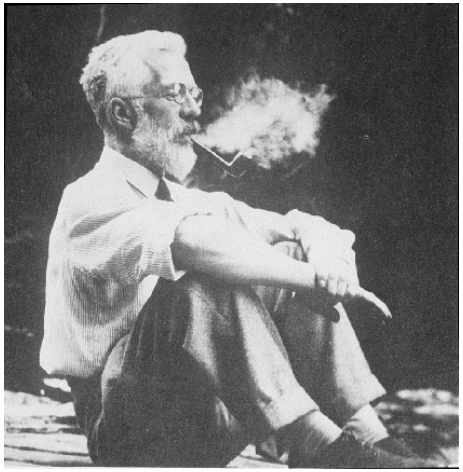
Ronald A. Fisher



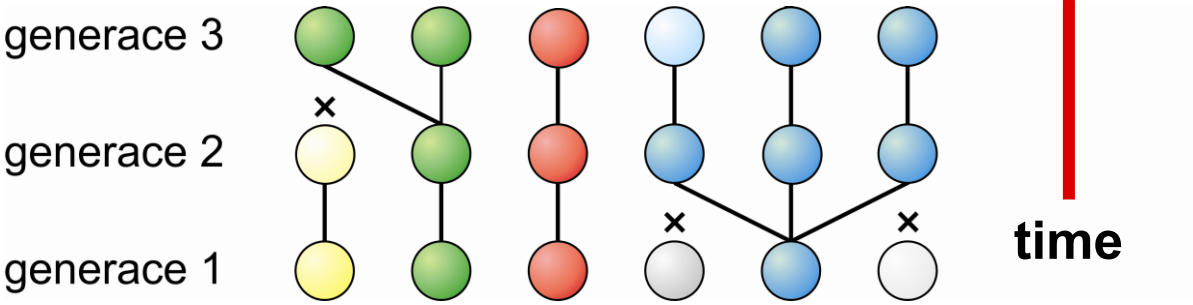
Lineage sorting in W-F model:



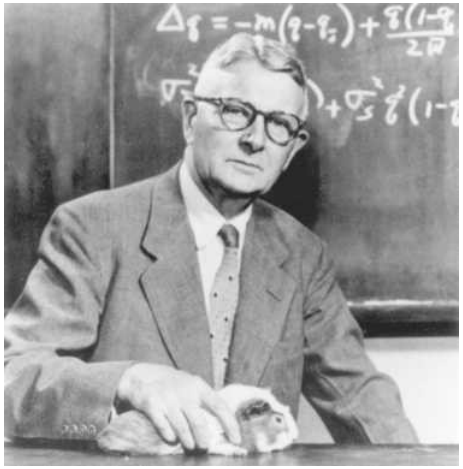
Sewall Wright



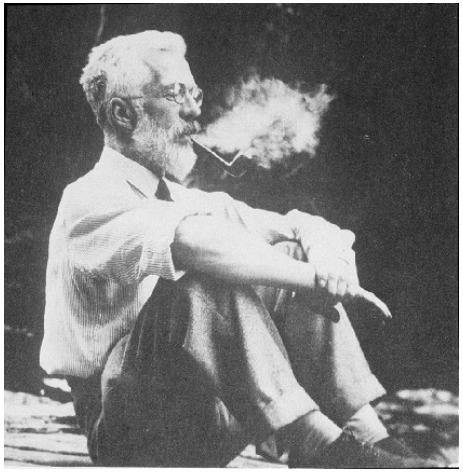
Ronald A. Fisher



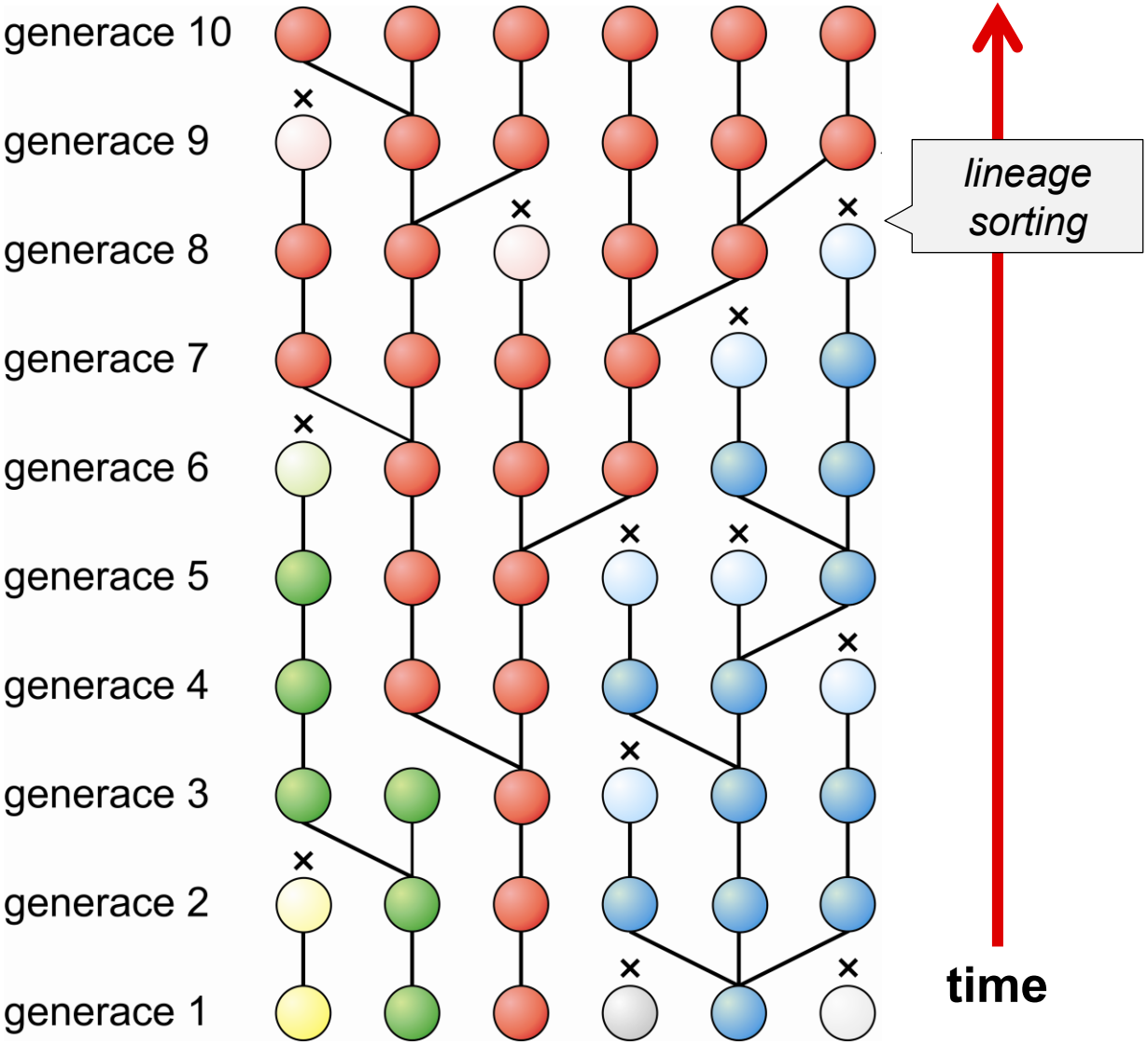
Lineage sorting in W-F model:

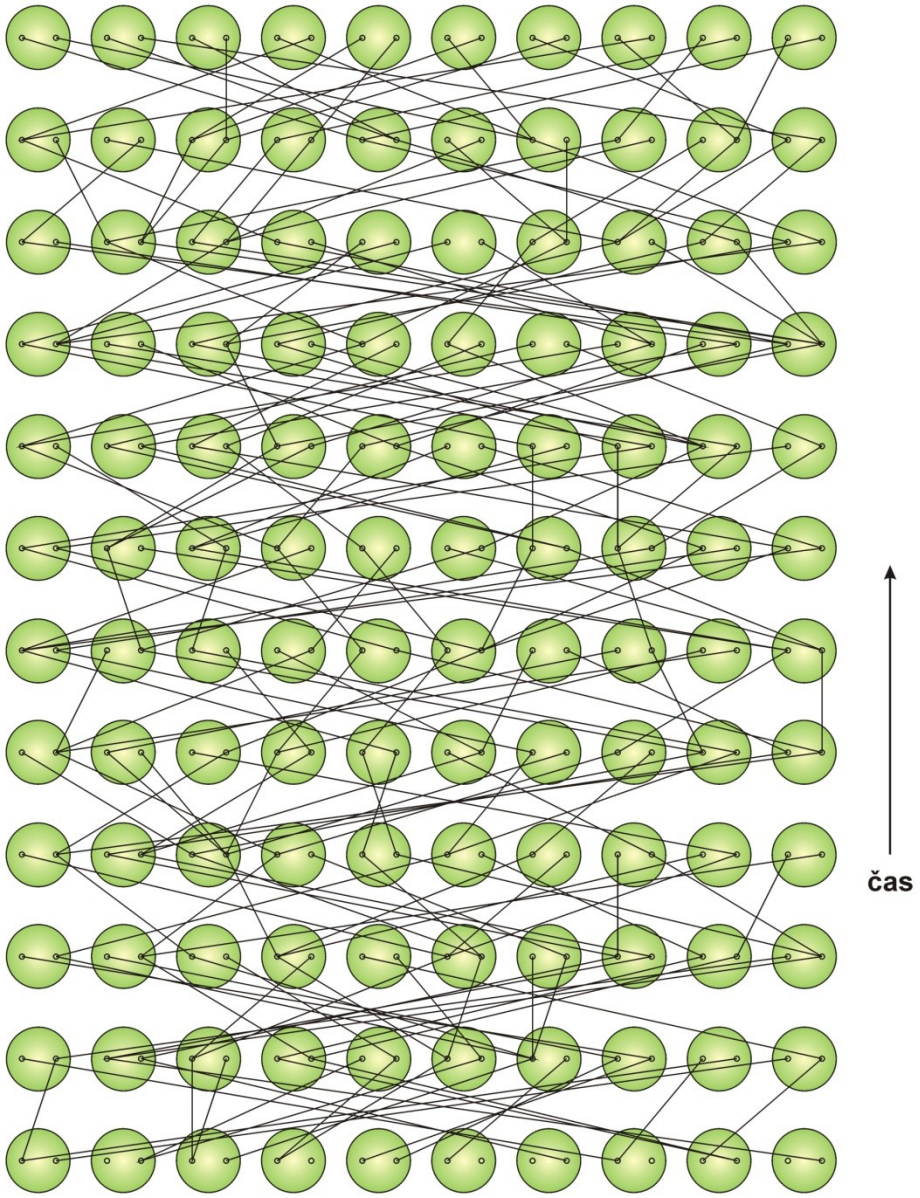


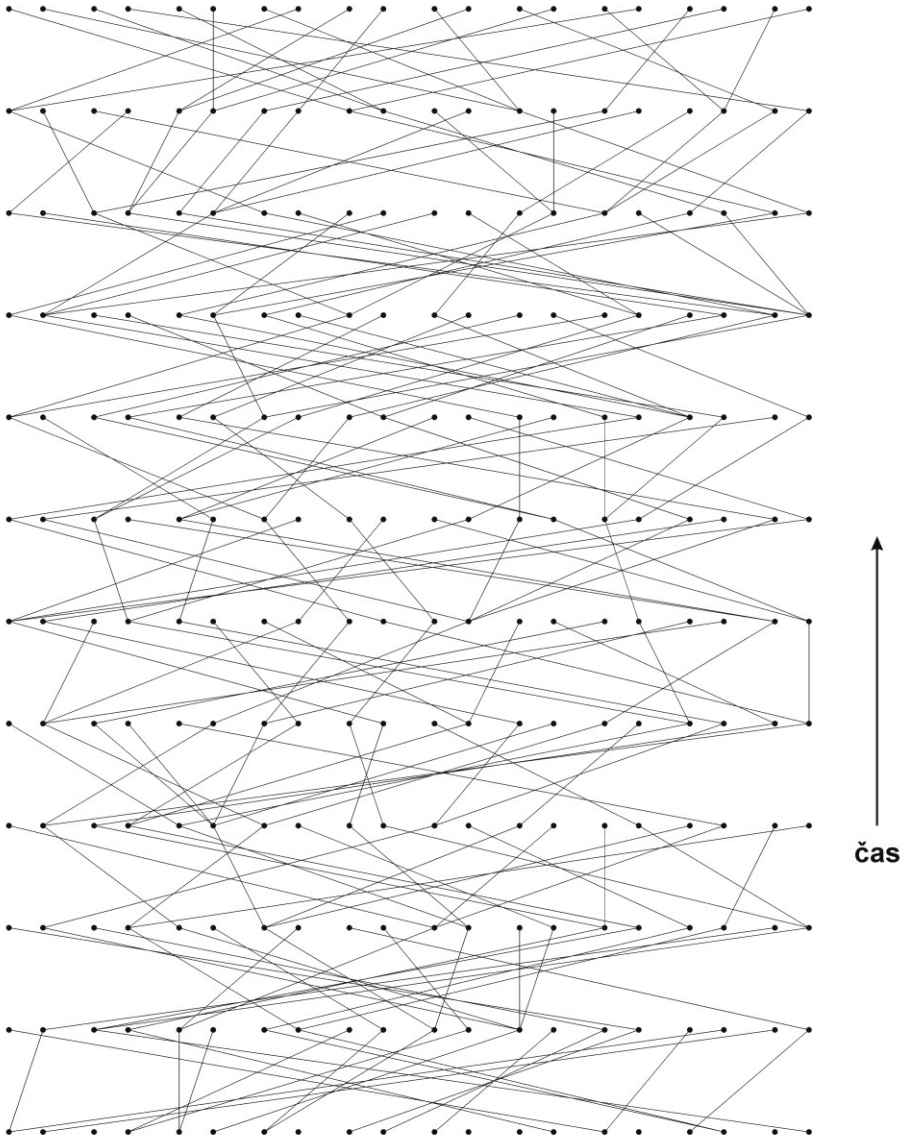
Sewall Wright

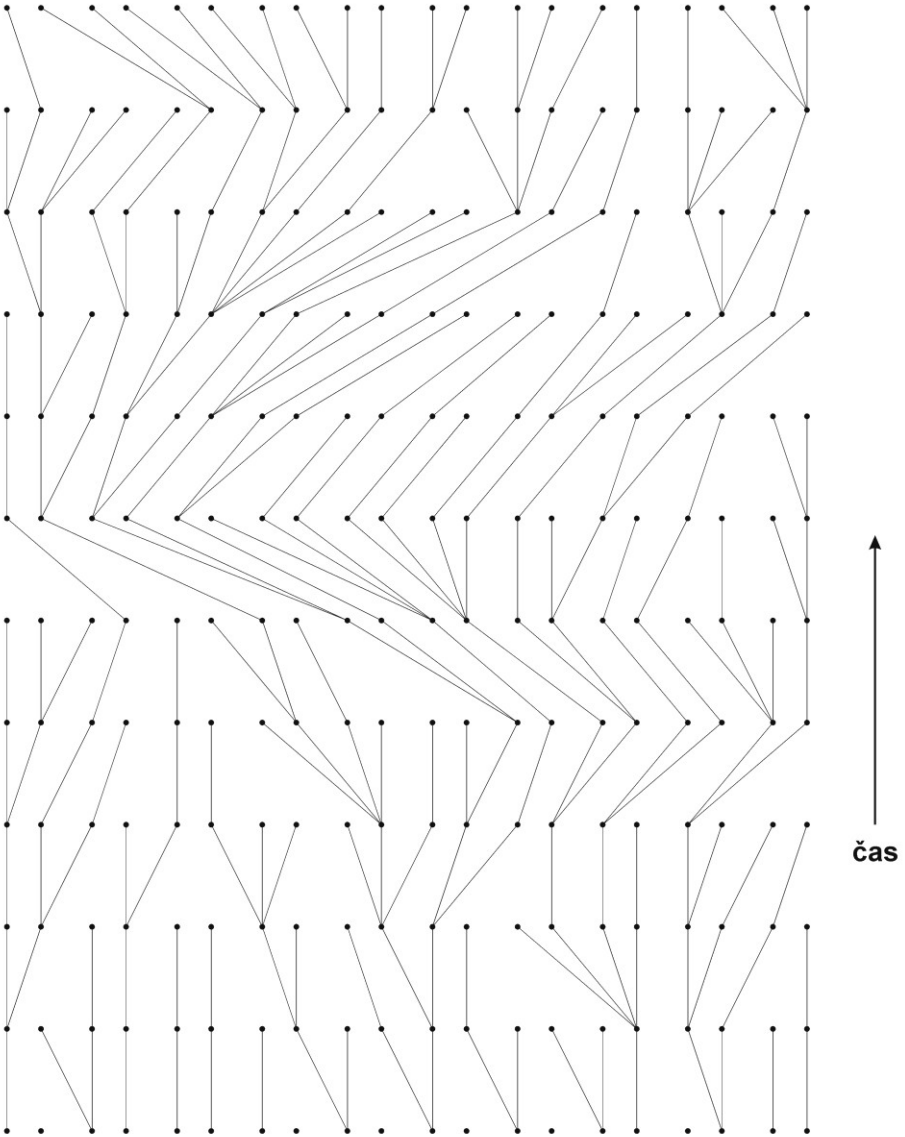


Ronald A. Fisher

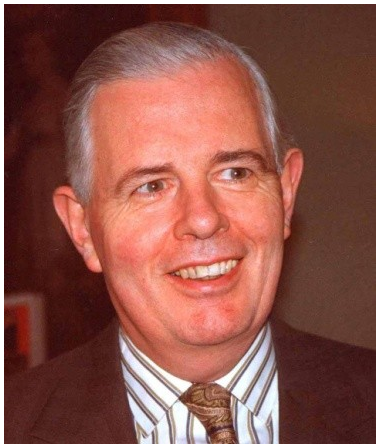



čas

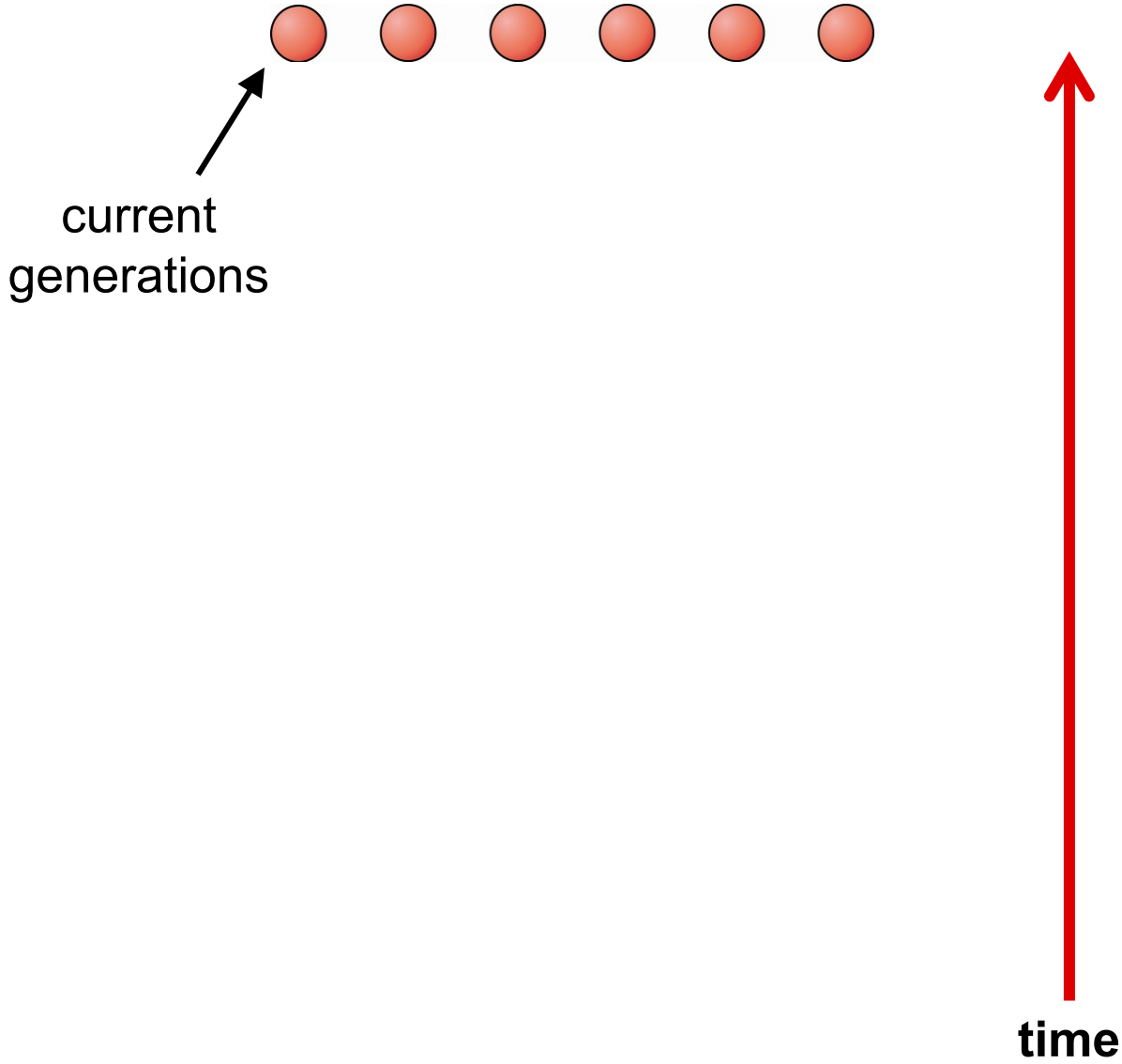

čas


čas

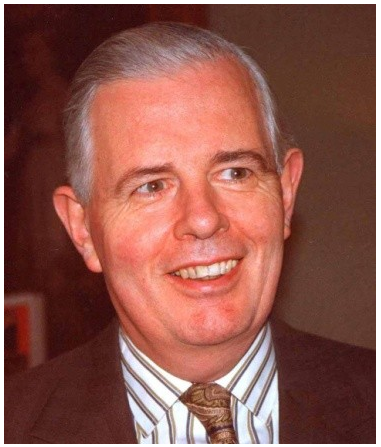
Coalescent:



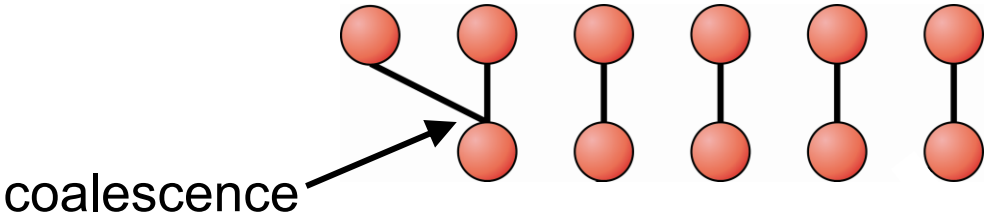
John F.C. Kingman



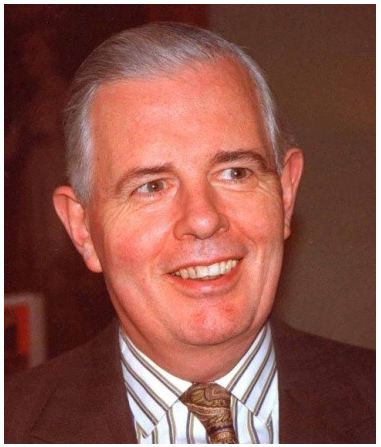
Coalescent:



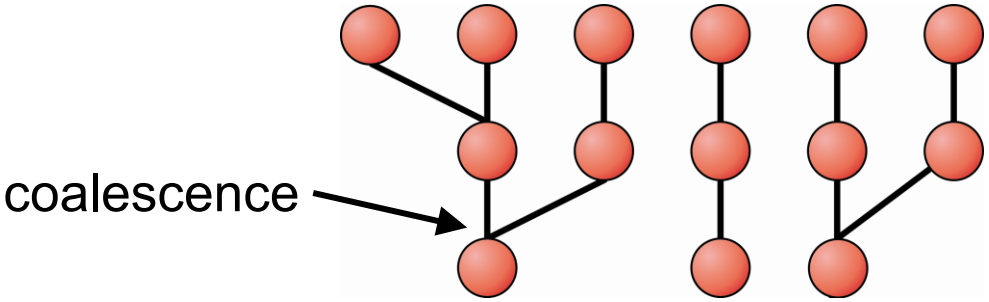
John F.C. Kingman



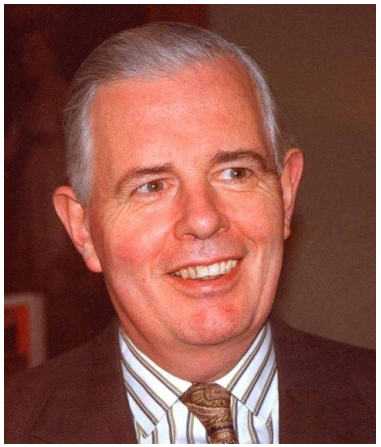
Coalescent:



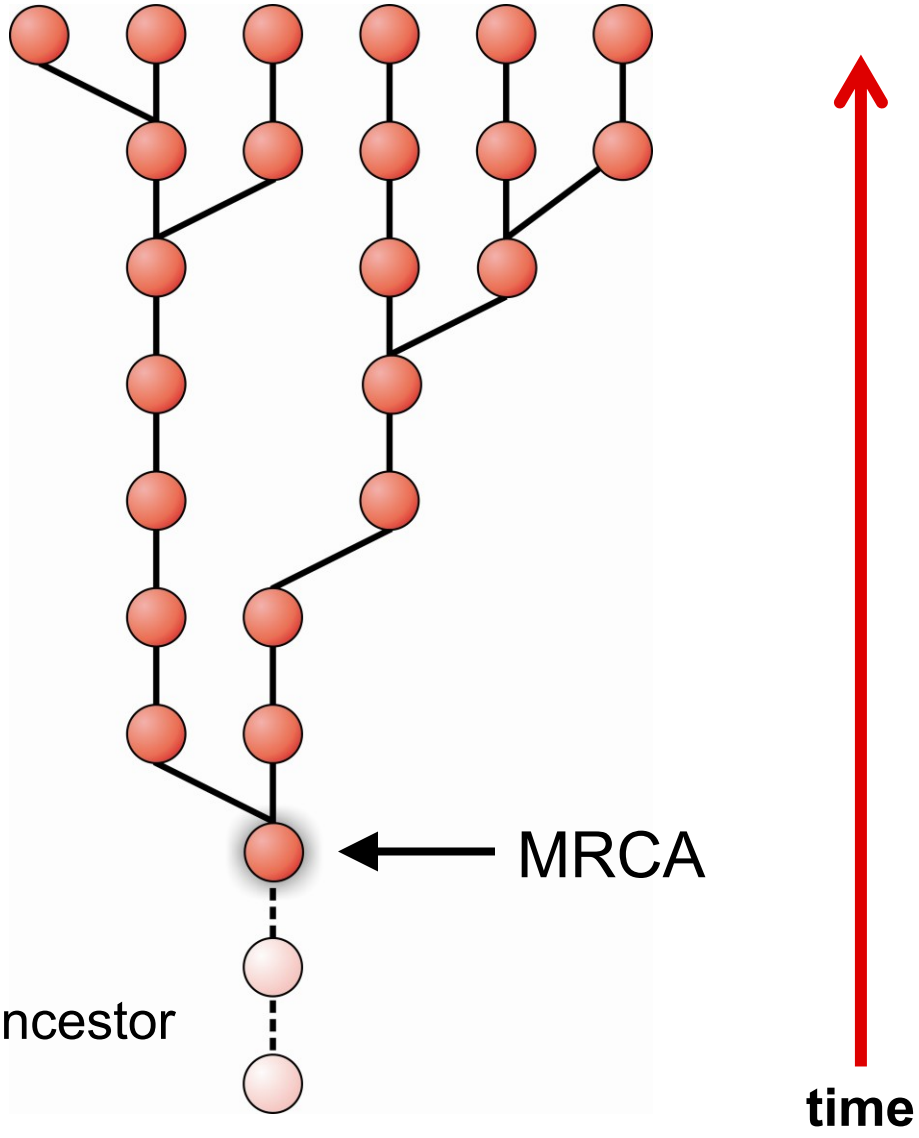
John F.C. Kingman



Coalescent:

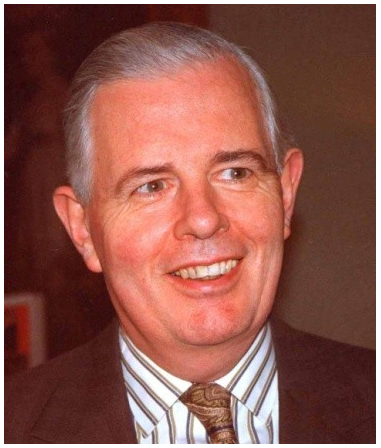


John F.C. Kingman



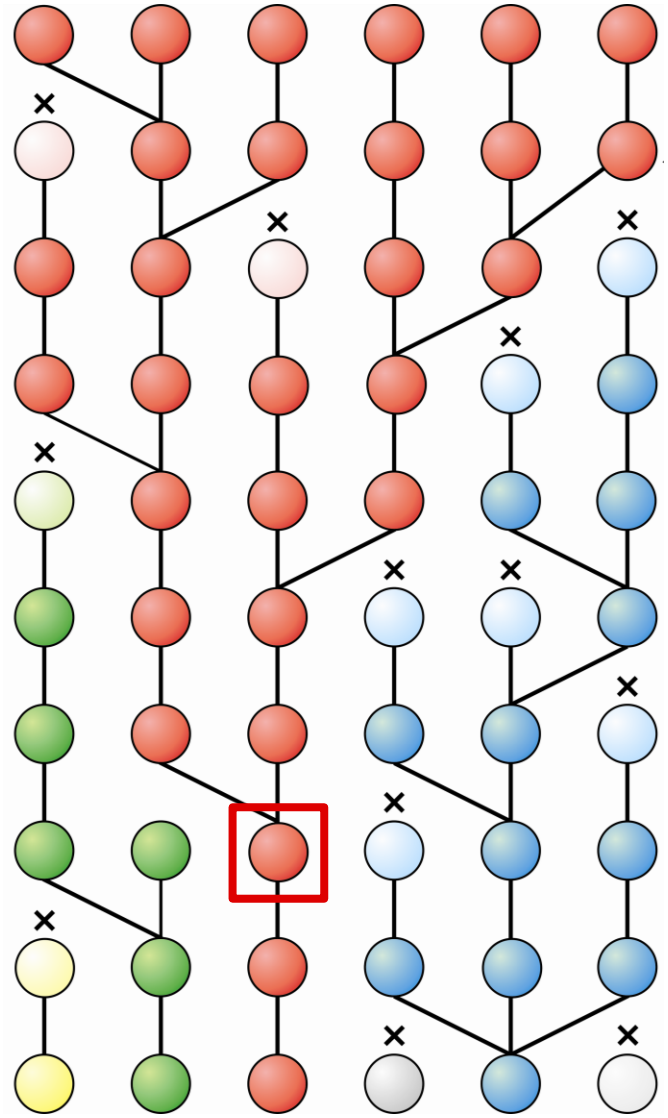
MRCA = most recent common ancestor

Coalescent:



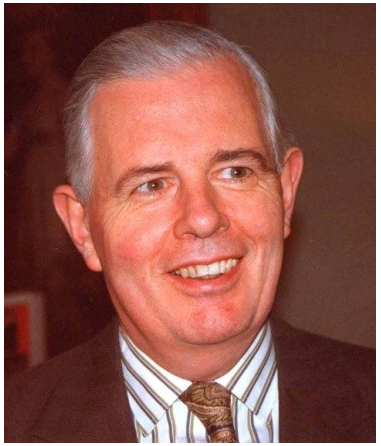
John F.C. Kingman

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

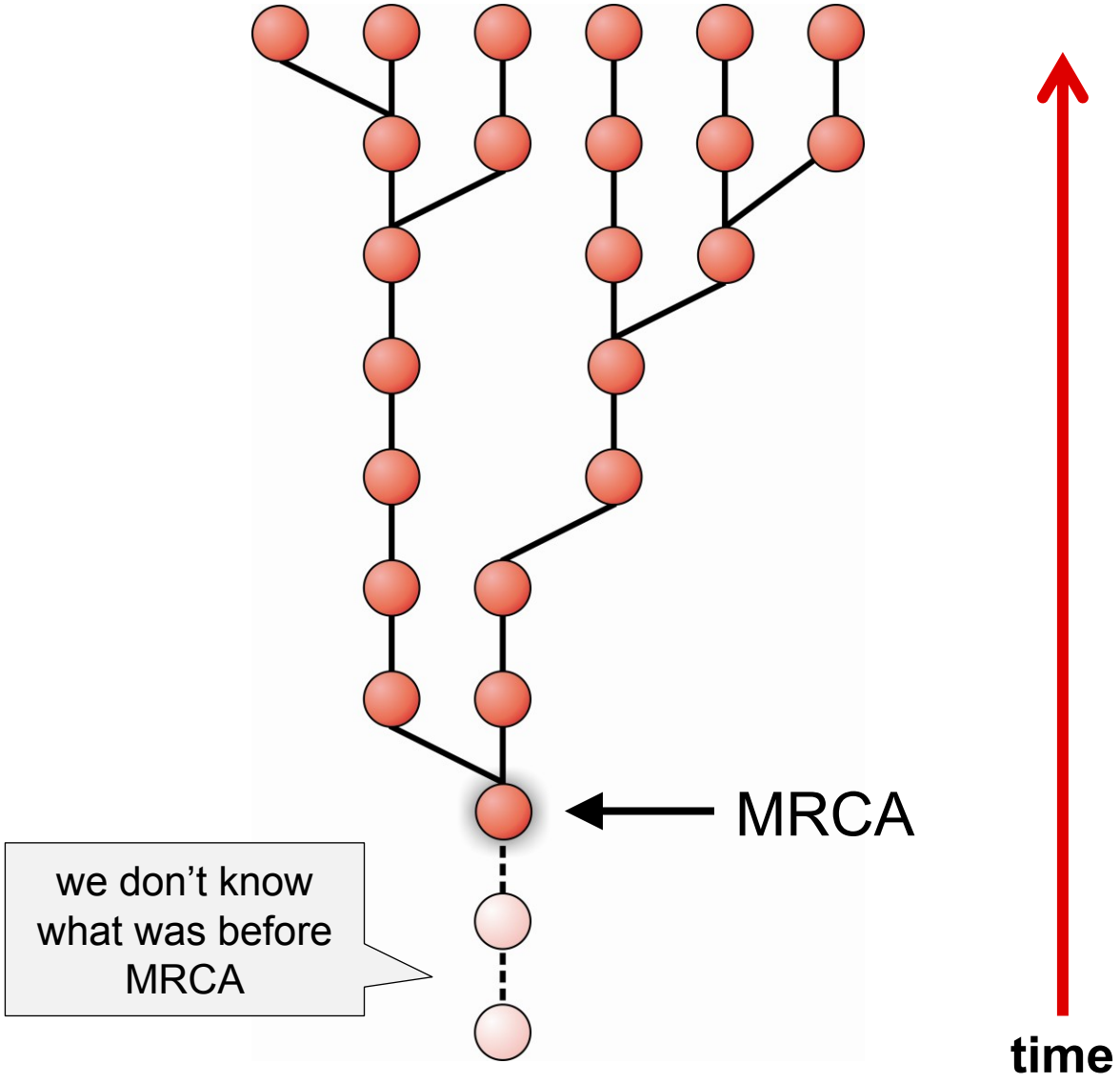


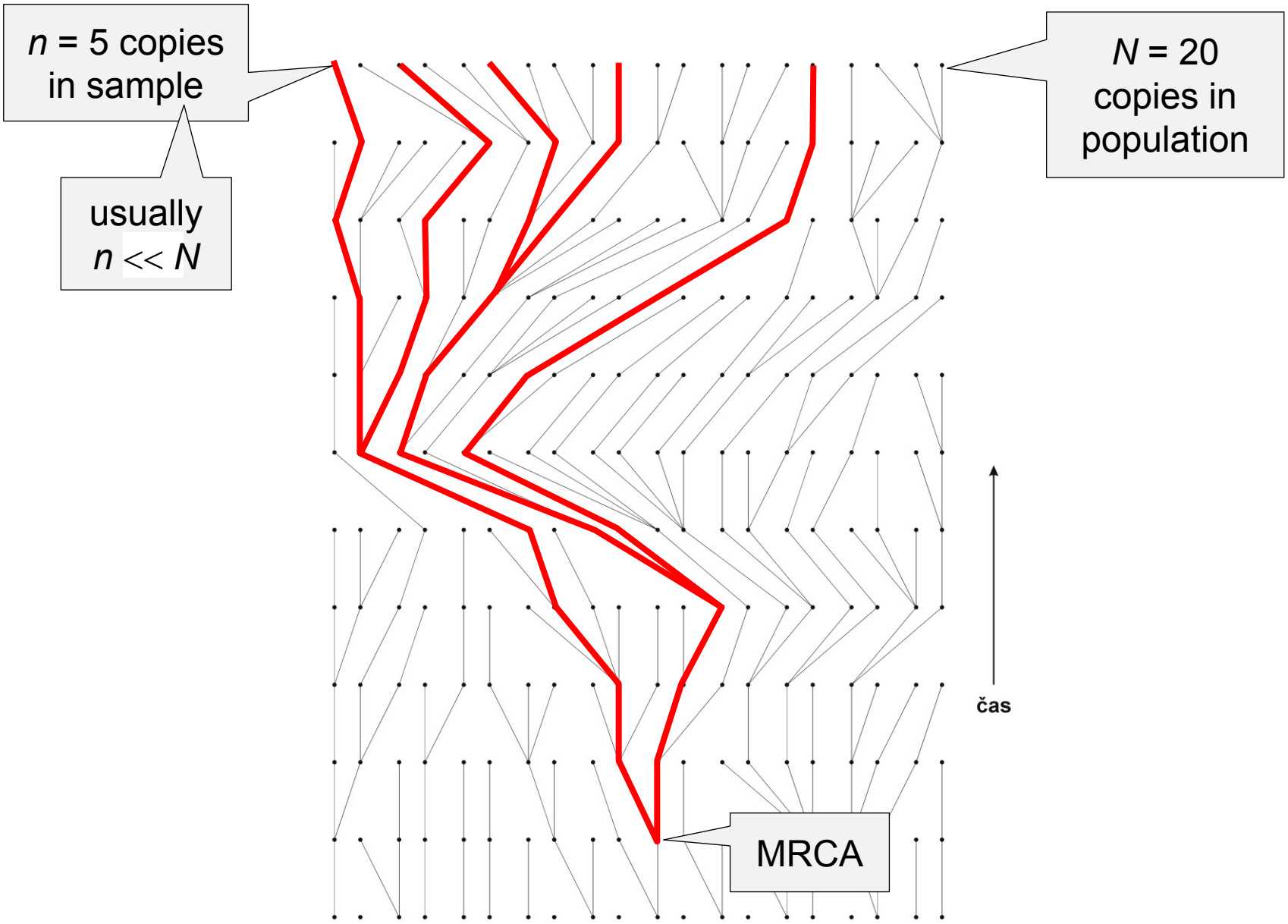
time

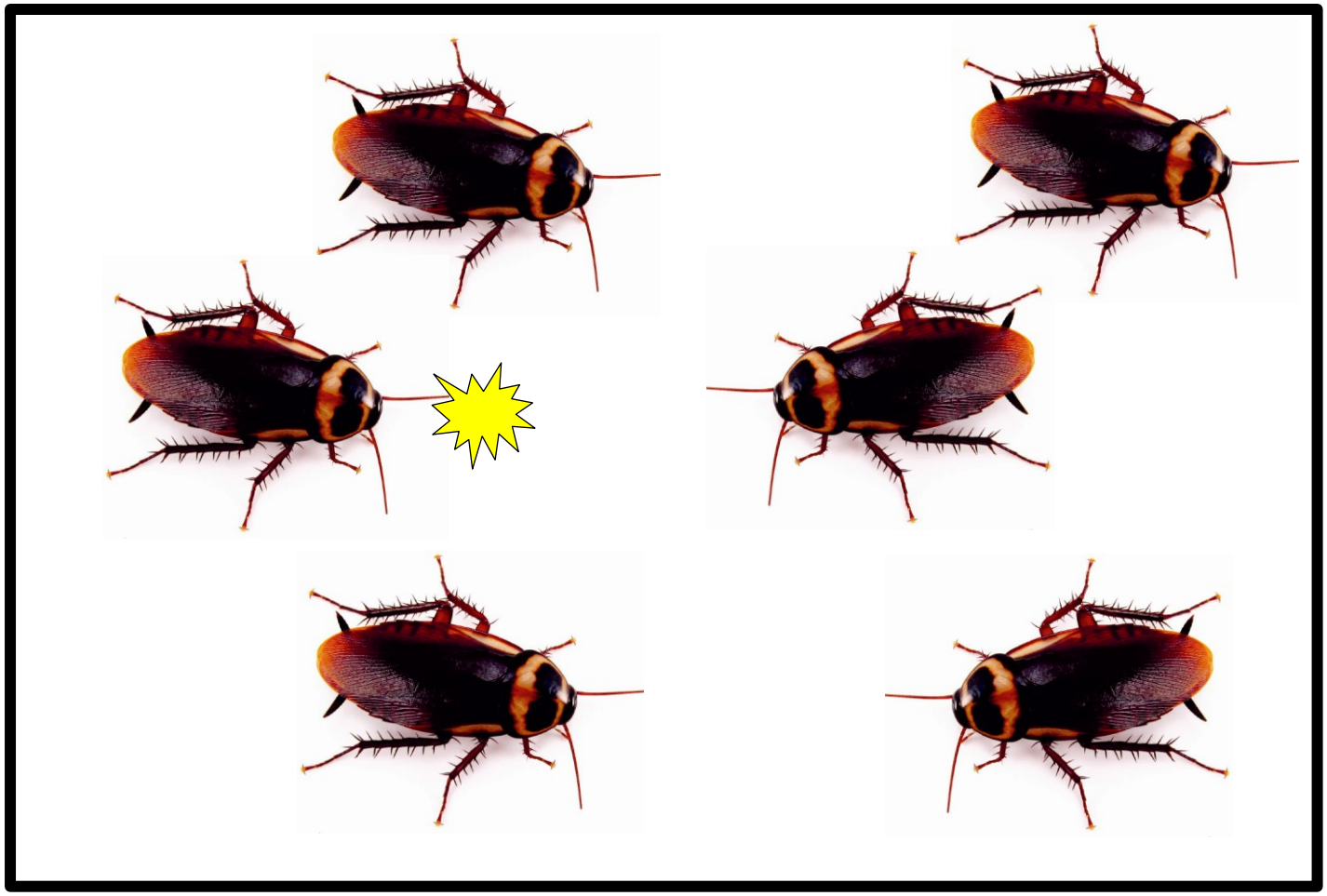
Coalescent:



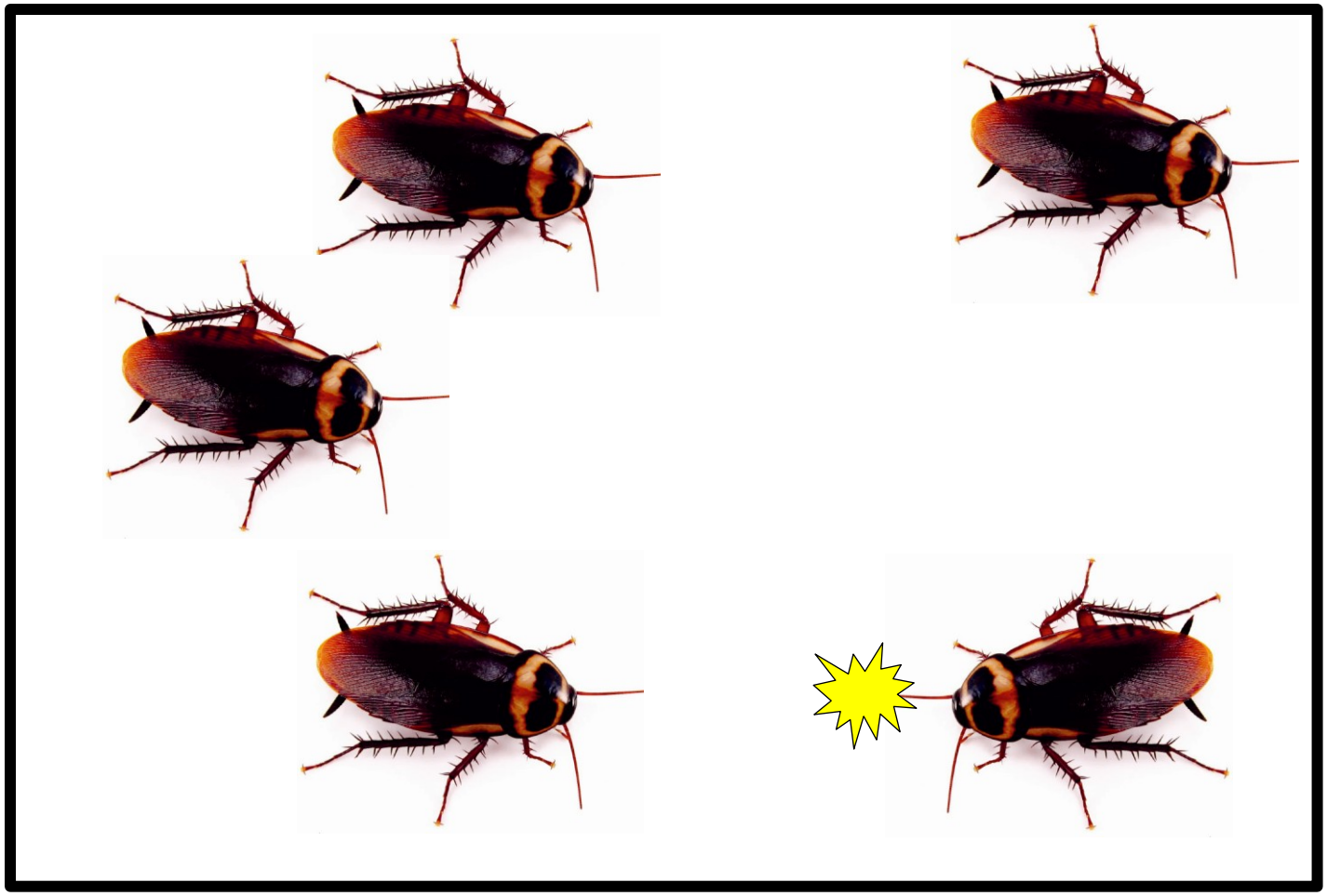
John F.C. Kingman




n = 5 copies in sample
usually n << N
N = 20 copies in population
čas
MRCA



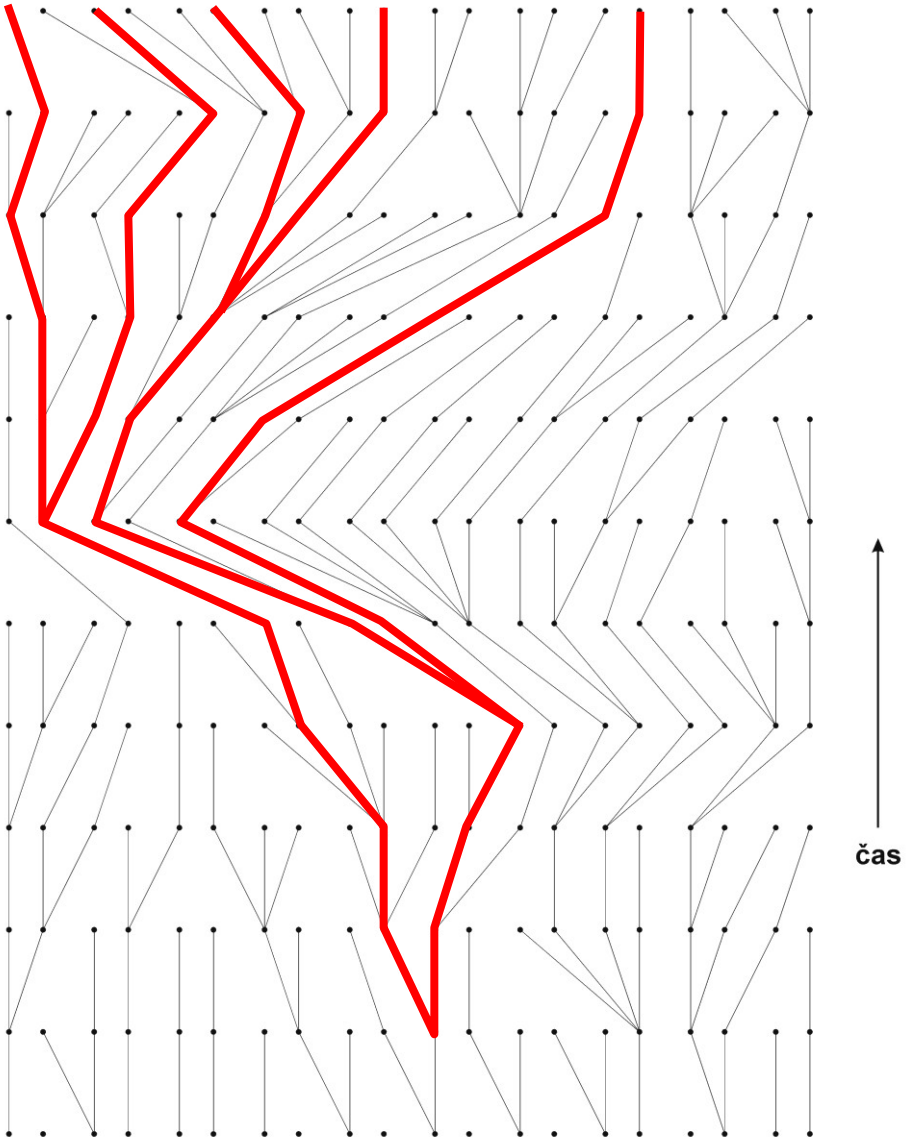
Probability of encounter of 2 cockroaches is $n(n-1)/4N$, where n = number of cockroaches in box, N = number of „places“ in box

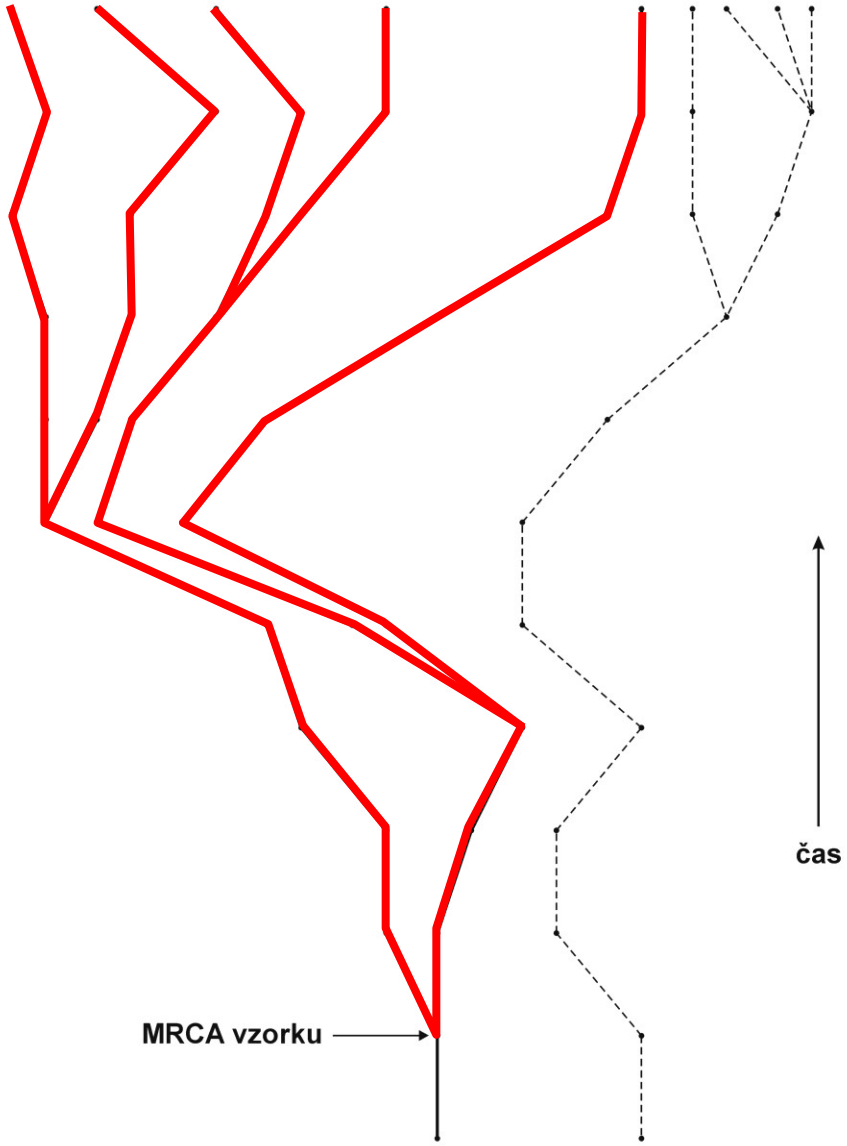


after coalescence, number of cockroaches (copies) is reduced by 1 ...

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

... to finish with just 1 copy


čas


MRCA vzorku
čas

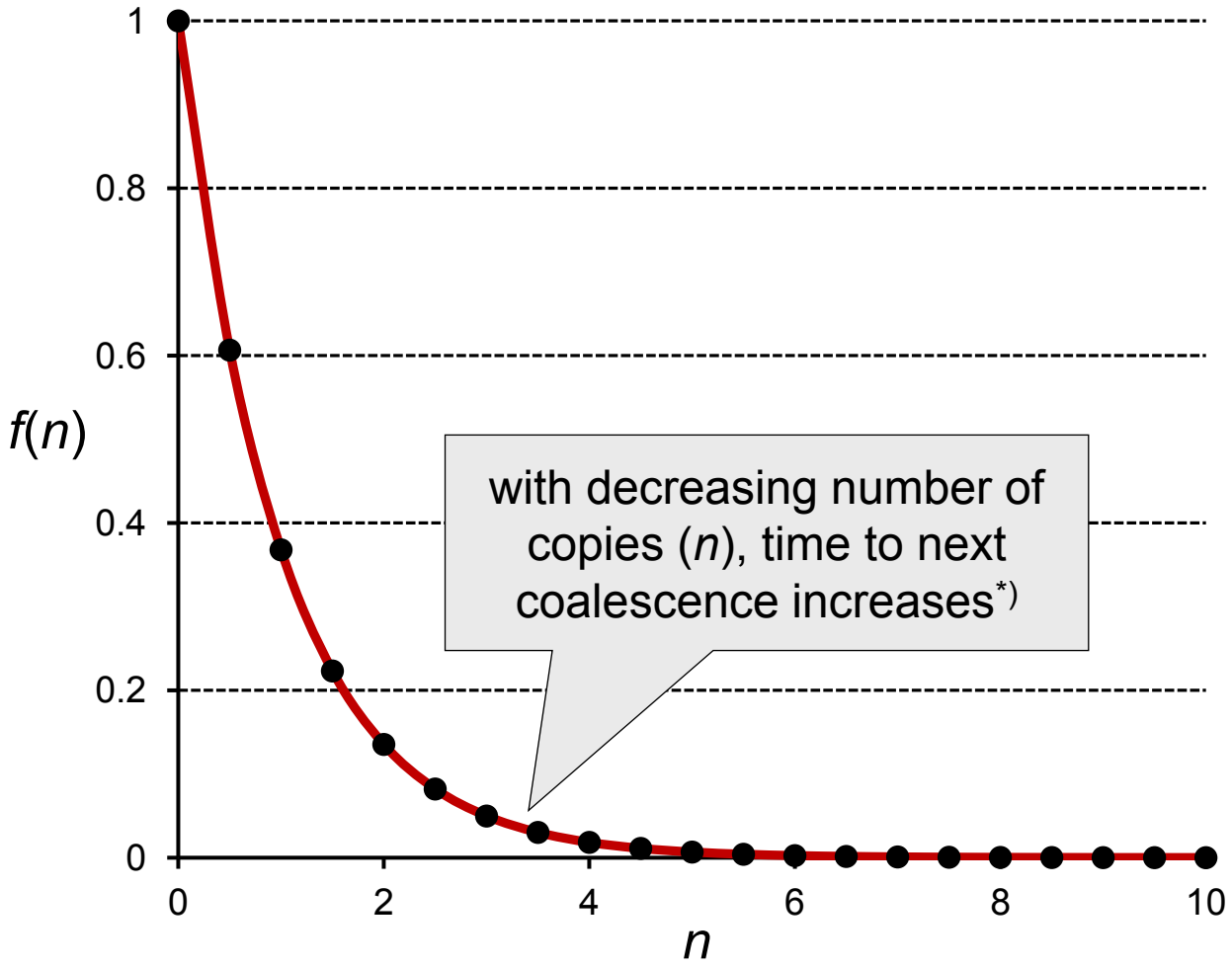
Kingman's coalescent:

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

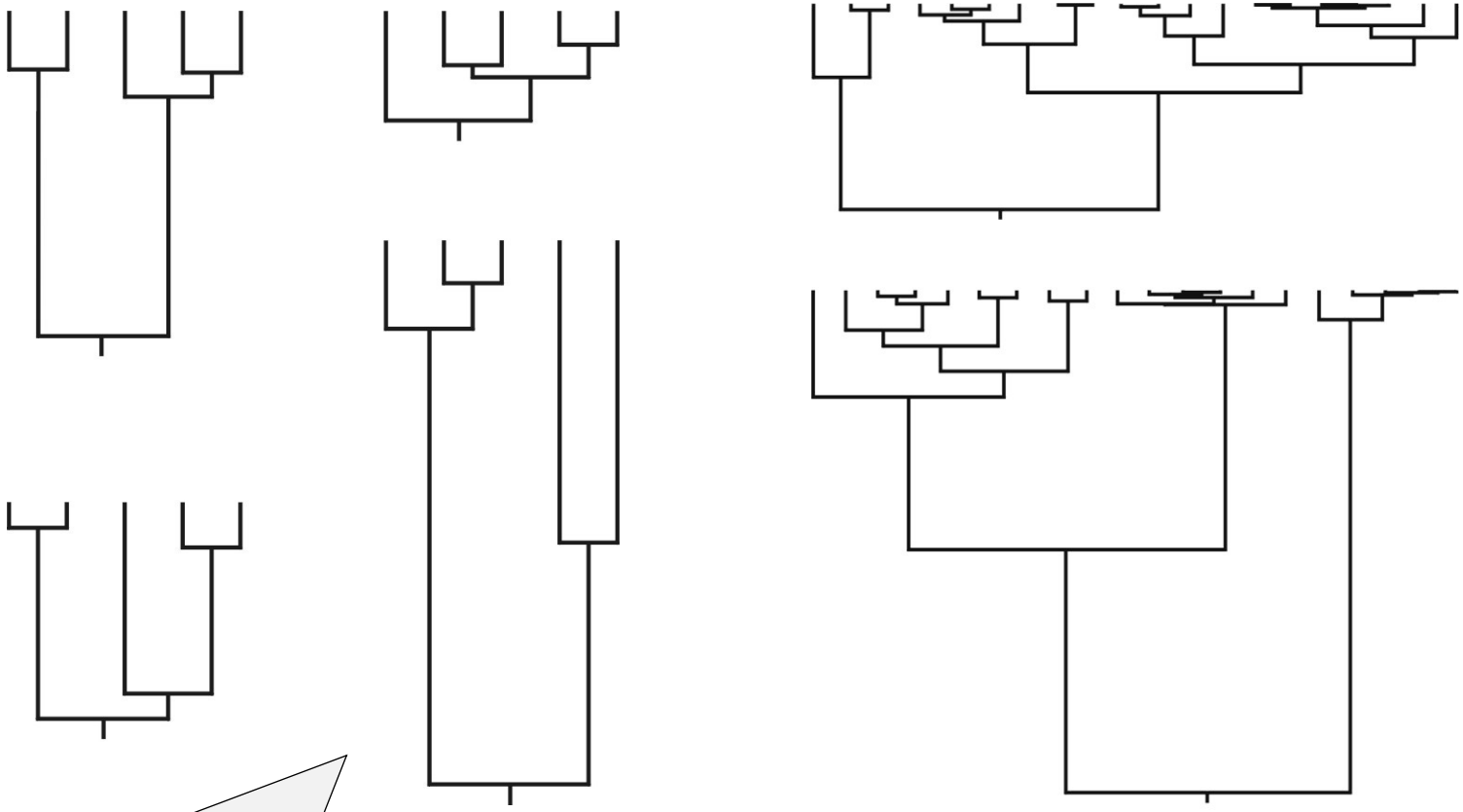
coalescence of last k copiies takes $(1 - 1/n)/(1 - 1/k)$
$\Rightarrow$ first 90% copies coalesce during 9% of total time, remaining 91% of time we wait for coalescence of last 10% copies!

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

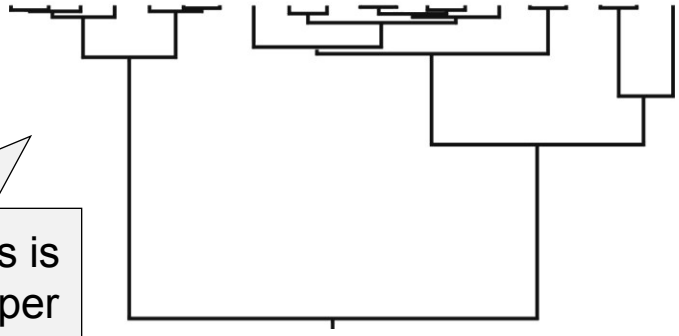
distribution of time between coalescences is approximately exponential:



*) see number of cockroaches in box

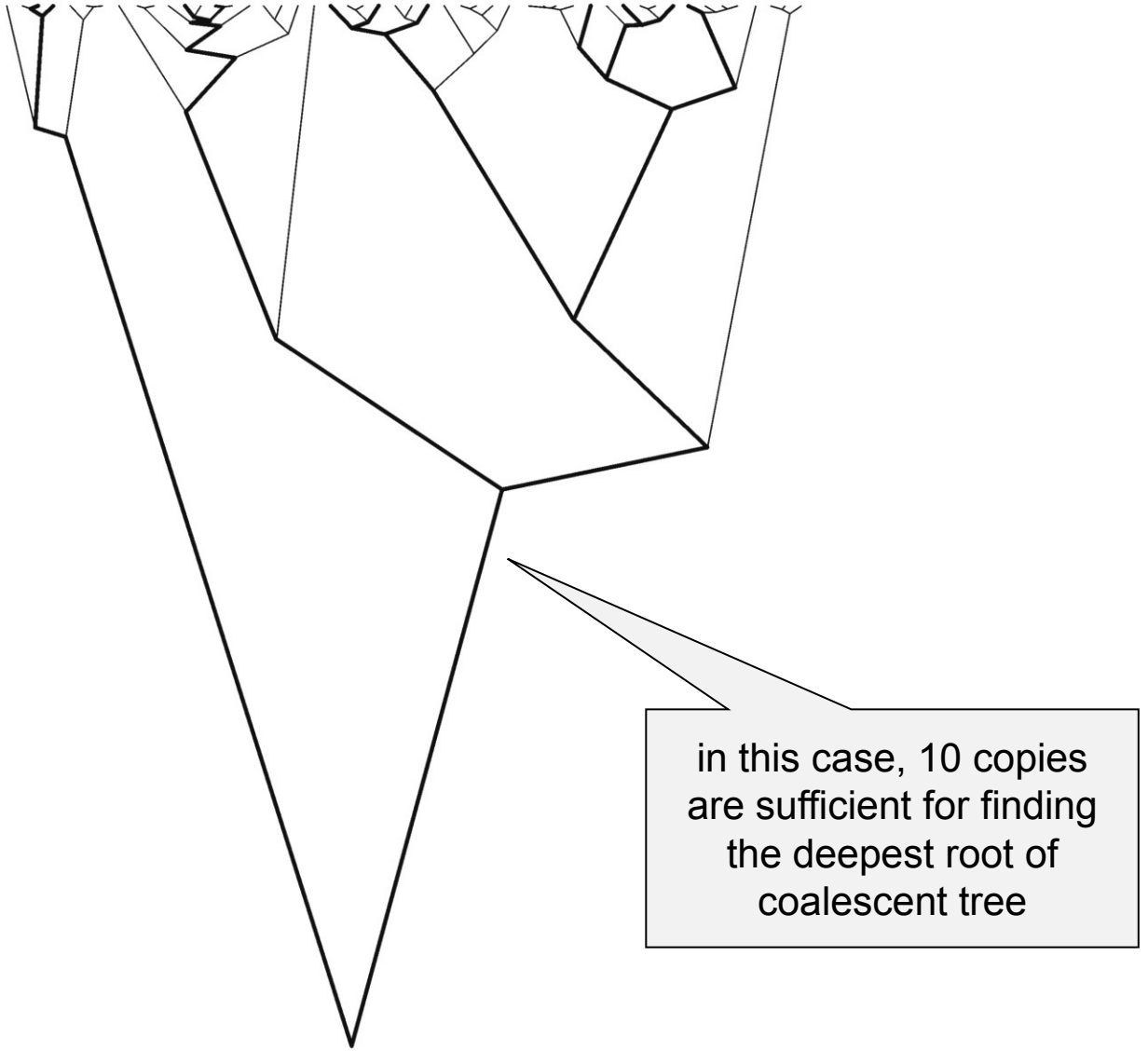


with decreasing number of free copies the process slows down



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

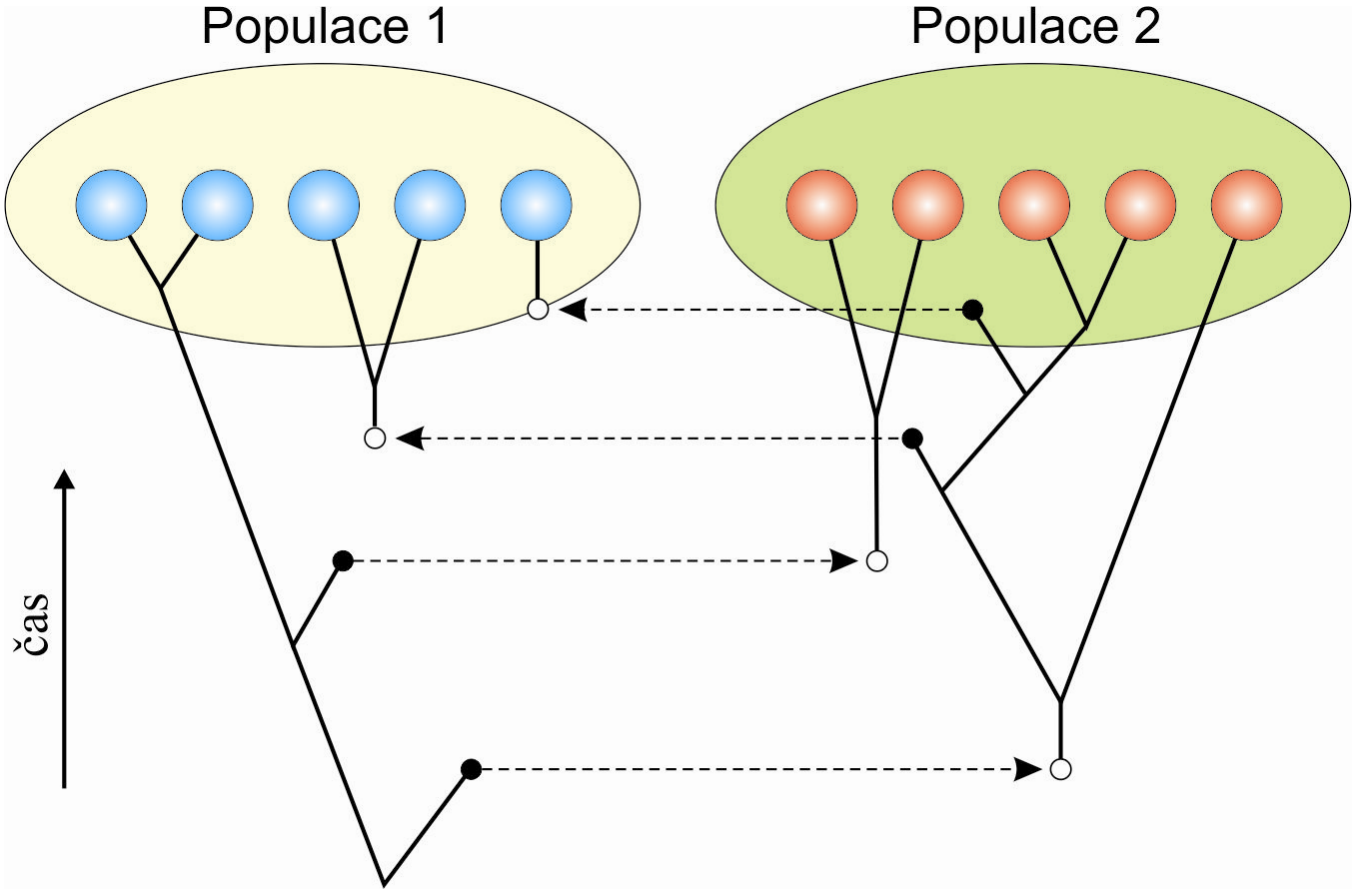
selection

changes of population size

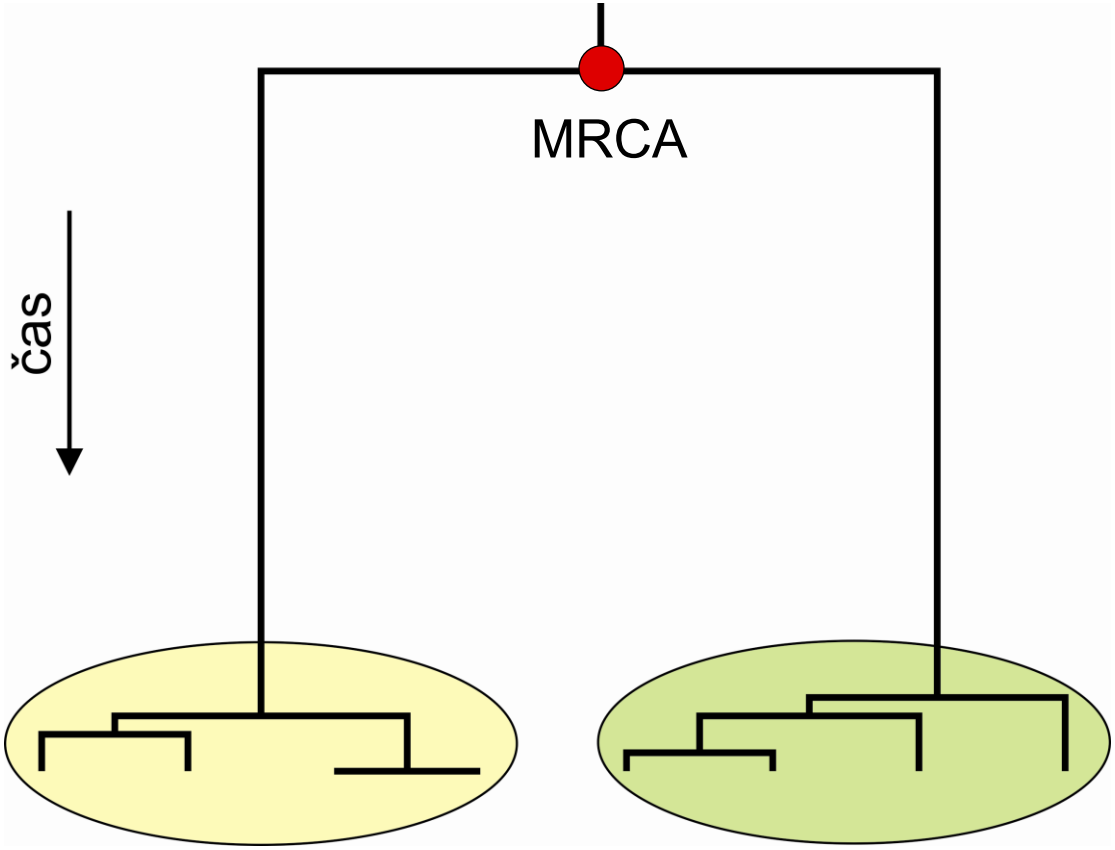
$\Rightarrow$ we can use coalescent theory for estimating these parametres

Coalescent is affected by various factors, eg.:

by migration



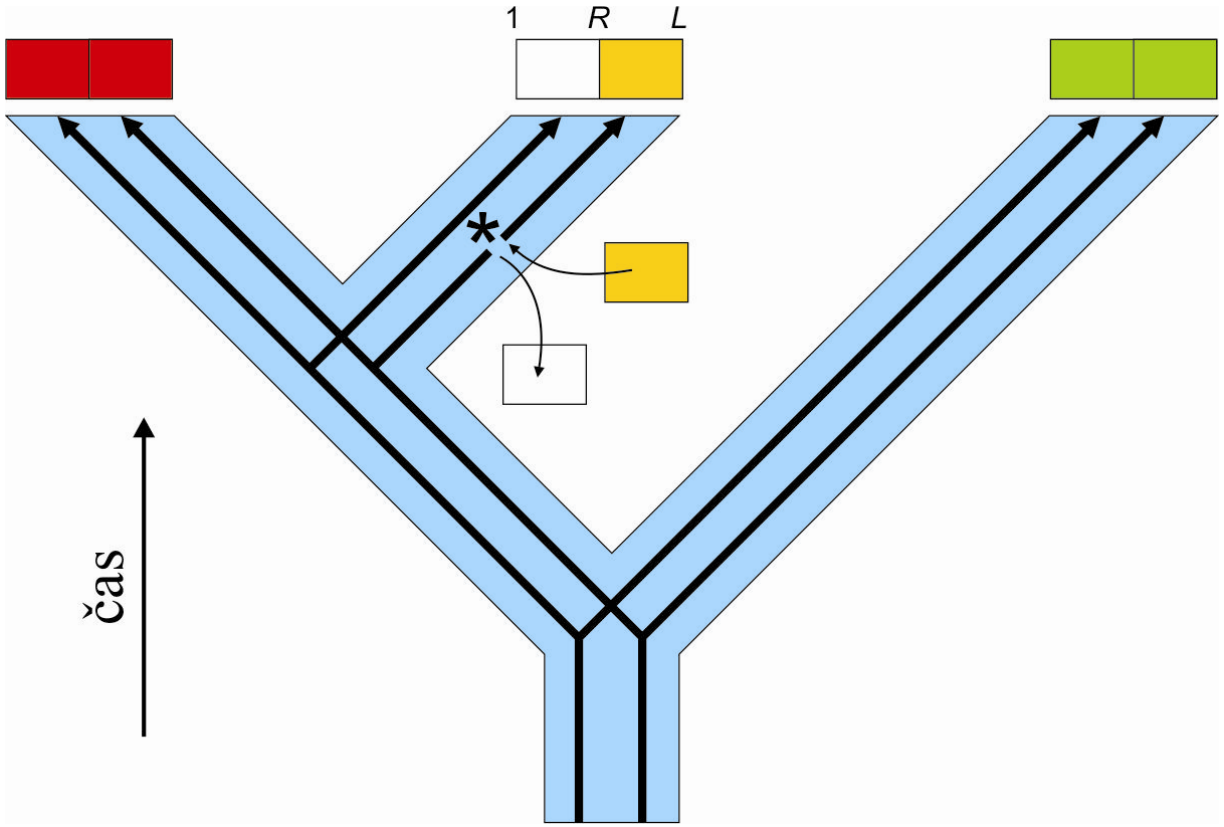
Weak migration leads to most coalescences within local populations,....



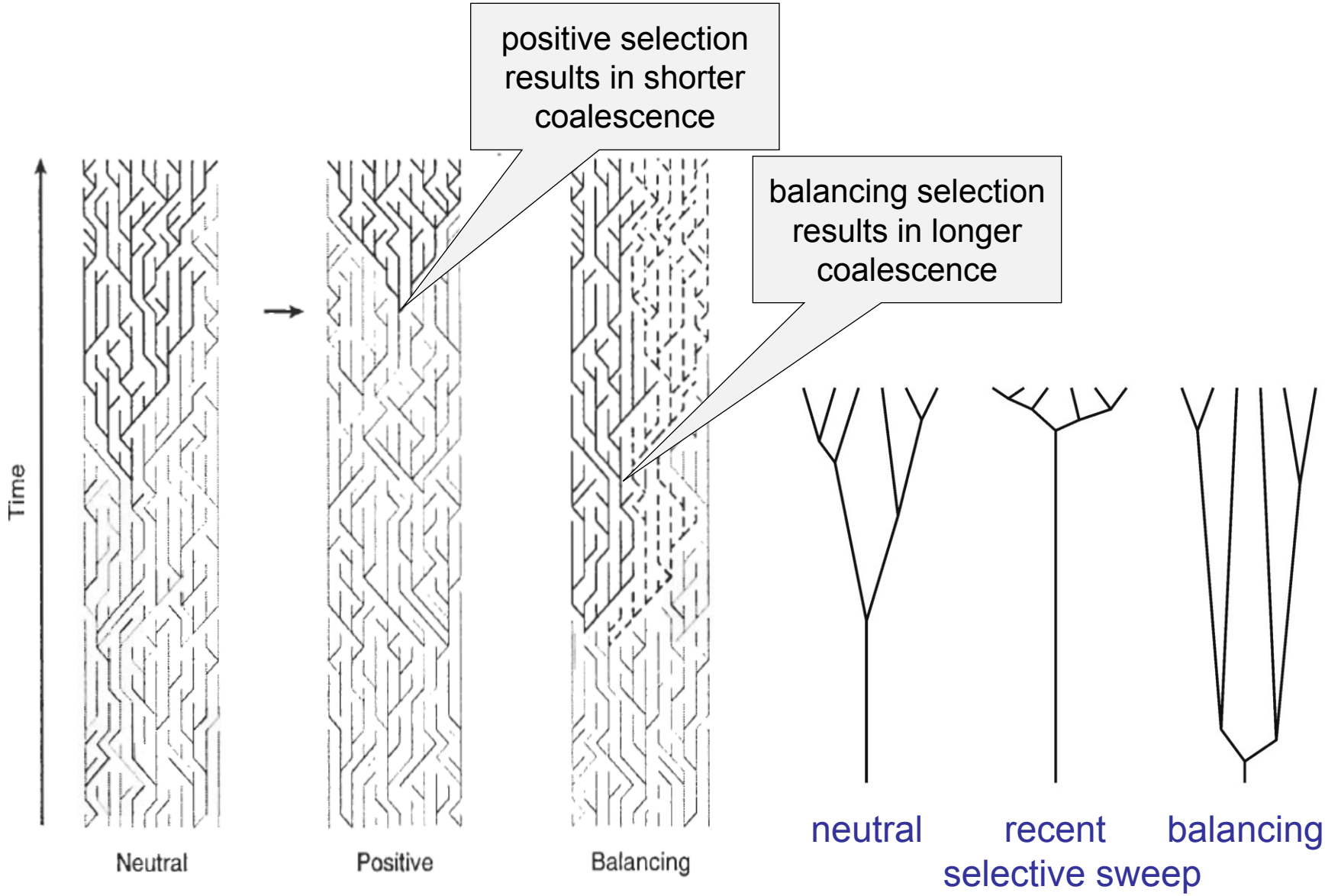
.... to increasing time to MRCA and its variance

Coalescent is affected by various factors, eg.:

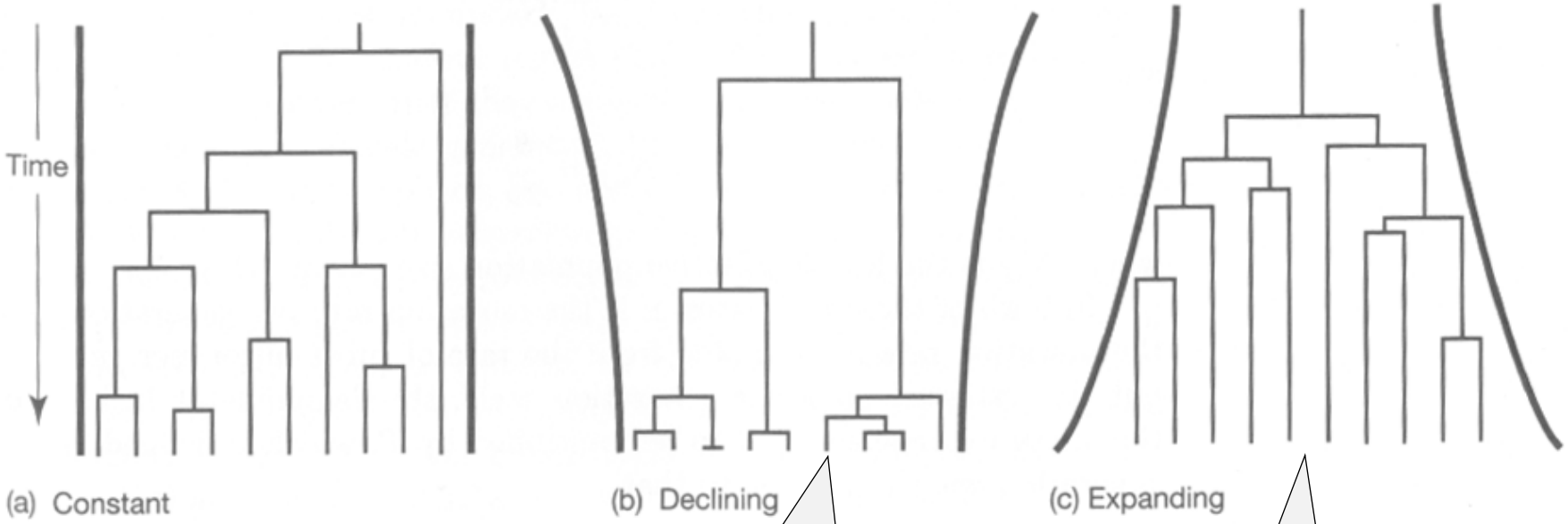
by recombination



Effect of selection on shape of coalescent tree



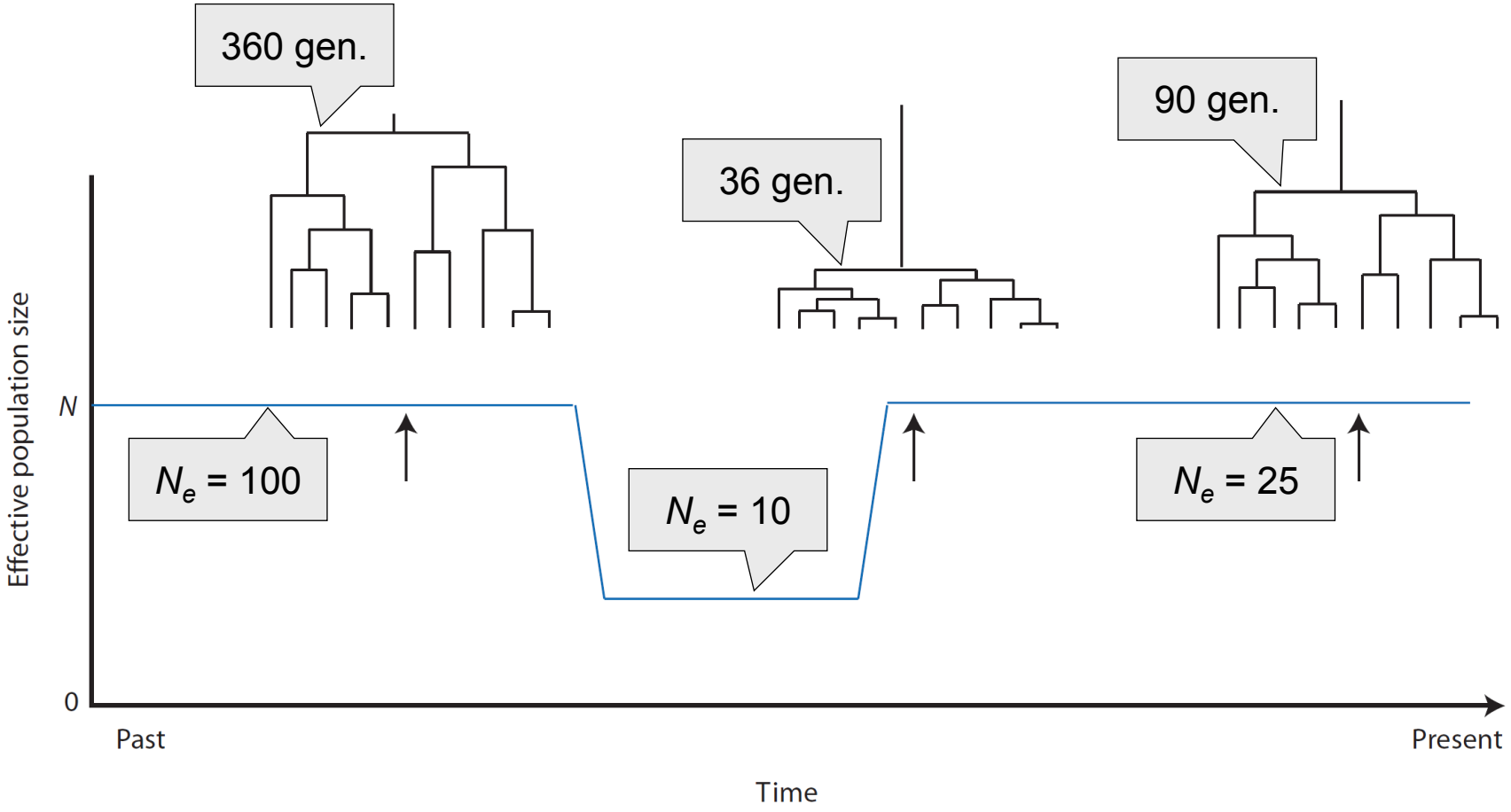
Effect of changes in population size on shape of coalescent tree



declining population: coalescent rate increases

growing population: coalescent rate decreases

$n = 10$


360 gen.
36 gen.
90 gen.
Effective population size
N
0
N_e = 100
N_e = 10
N_e = 25
Past
Present
Time

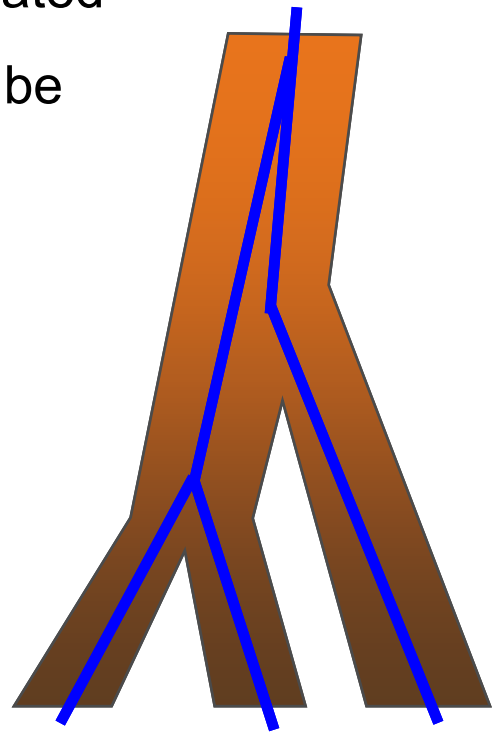
Gene vs. species trees once more:

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



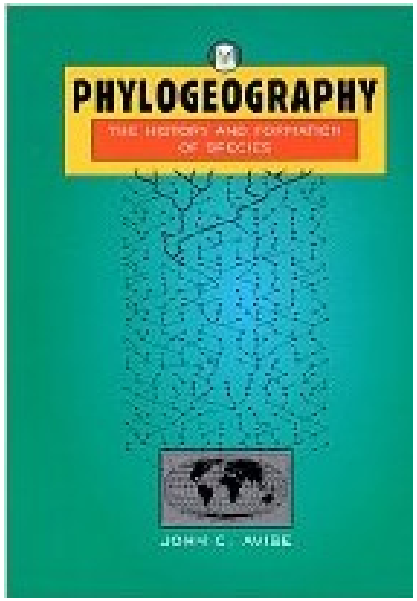
PHYLOGEOGRAPHY

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

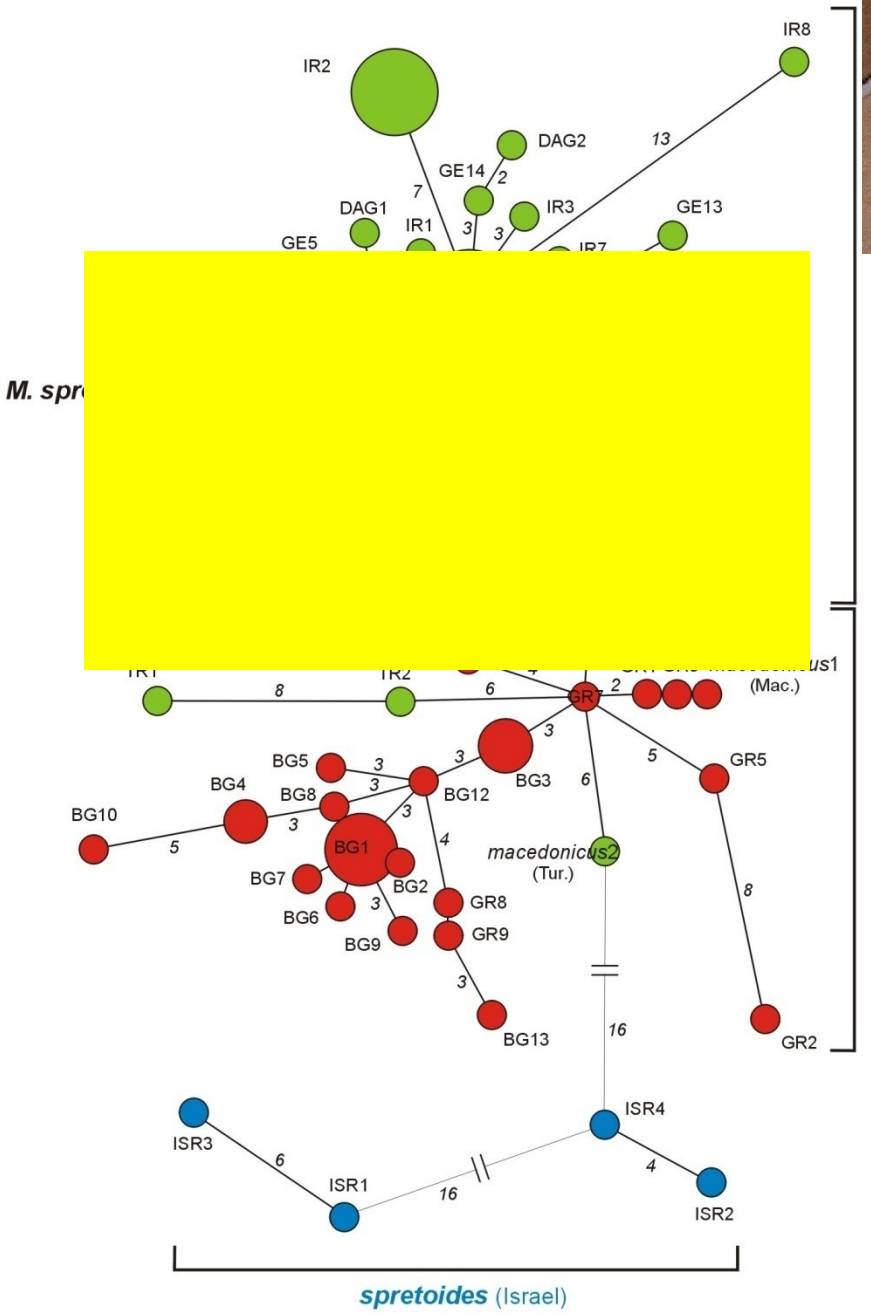


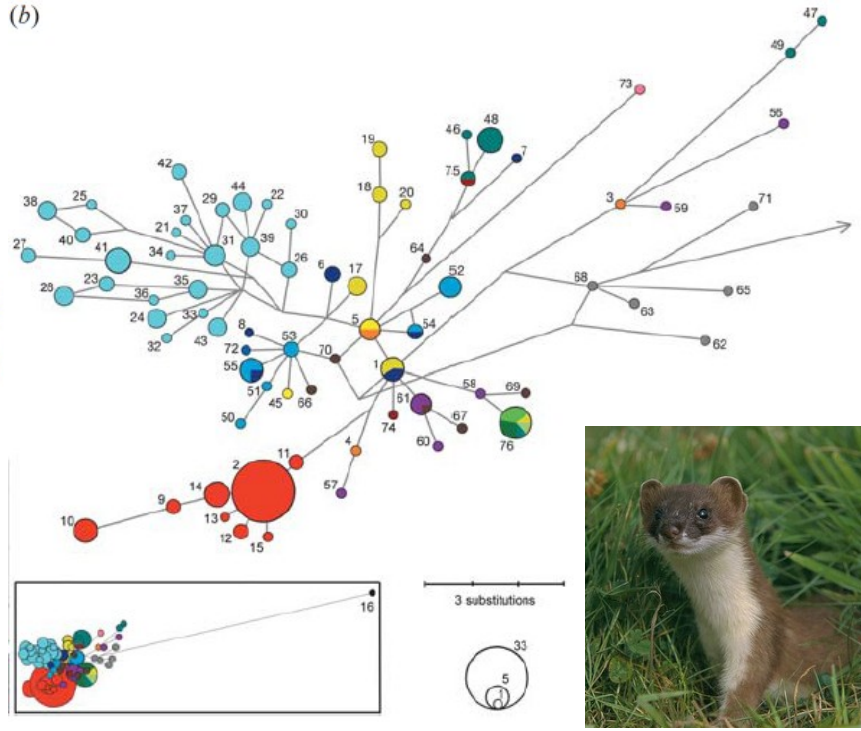
Mus macedonicus

Minimum Spanning Tree (MST)

Mimum Spanning Network (MSN)

Median-joining network etc.





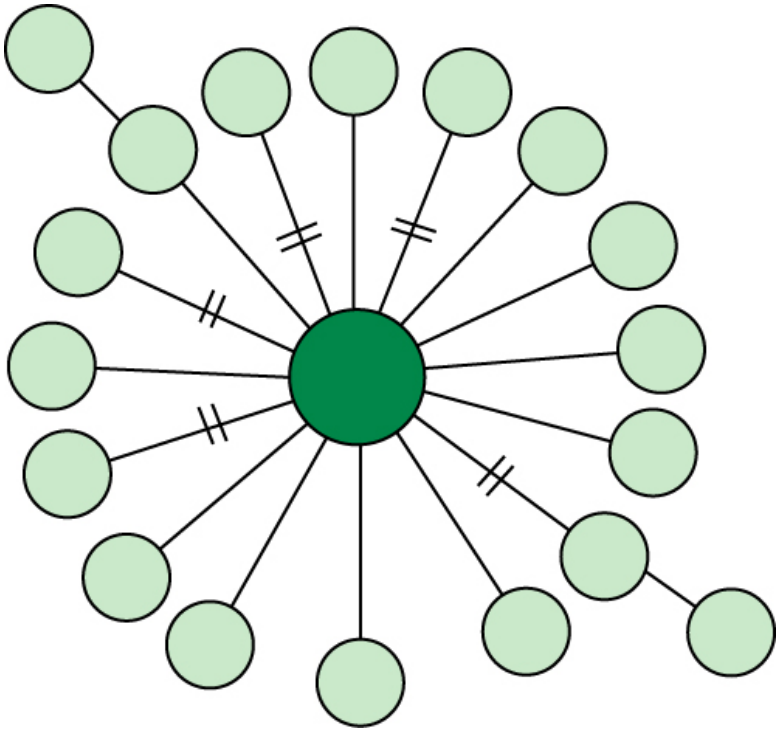
Mustela erminea

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:

equlibrium heterozygosity $\theta = 4N_e\mu$

if evolution neutral, θ can be estimated in various ways, e.g.

as 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

*) nucleotide diversity

**) Watterson's theta

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

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

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. π = (3+2+3+1+3+3)/6 = 2,5

S = 4 segregating sites

θ_W = 4/(1/1 + 1/2 + 1/3) = 4/1,83 = 2,186

- θ_π - θ_W = 2,5 – 2,186 = 0,314

1. Tajima's test

very negative values indicate population expansion – prevalence of „young" polymorphisms, when new haplotypes were arising, but nucleotide diversity still low

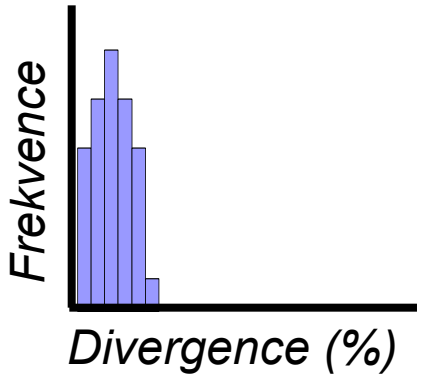
programs Arlequin, DnaSP etc.

likewise Fu's test etc.

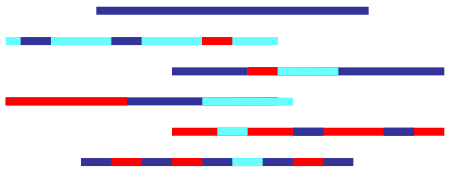
2. Mismatch distribution

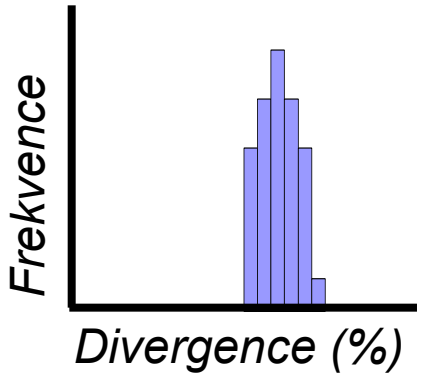
pairwise comparison of all sequences → histogram

Sequences very similar

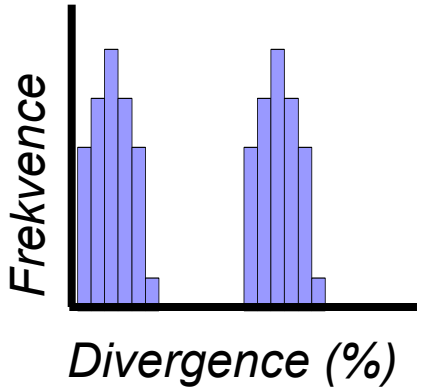


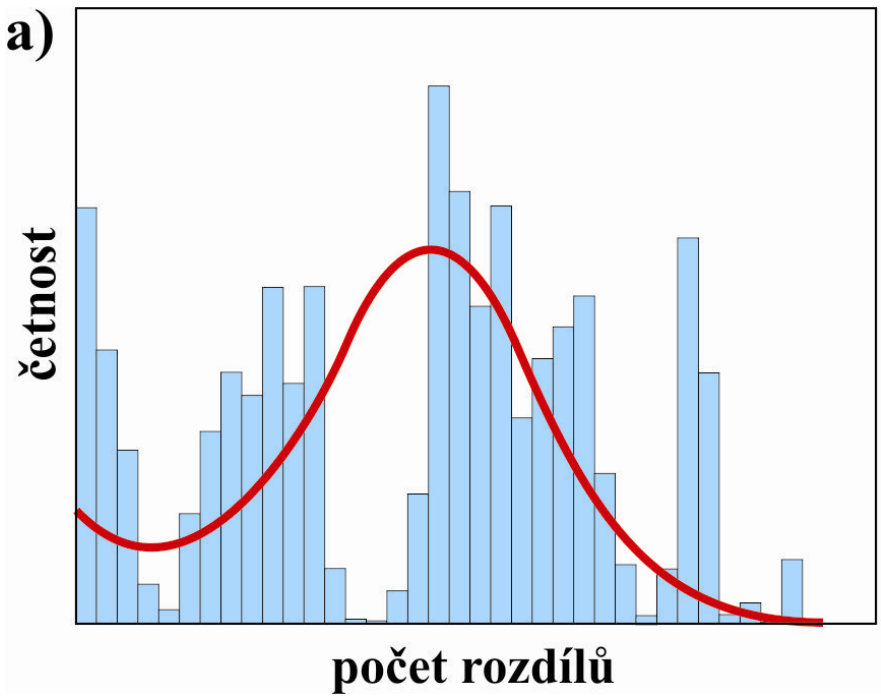
Sequences very divergent

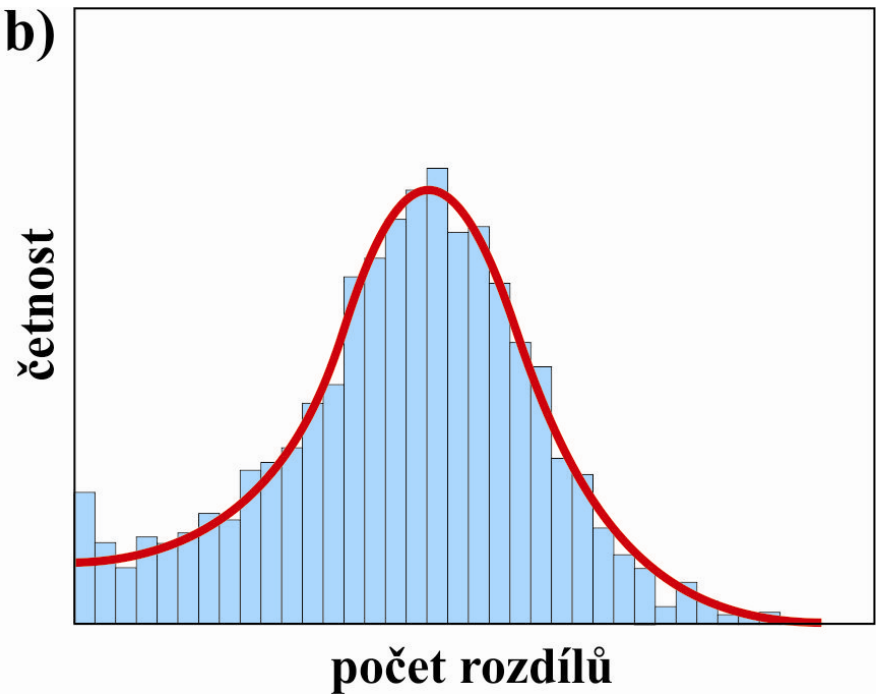


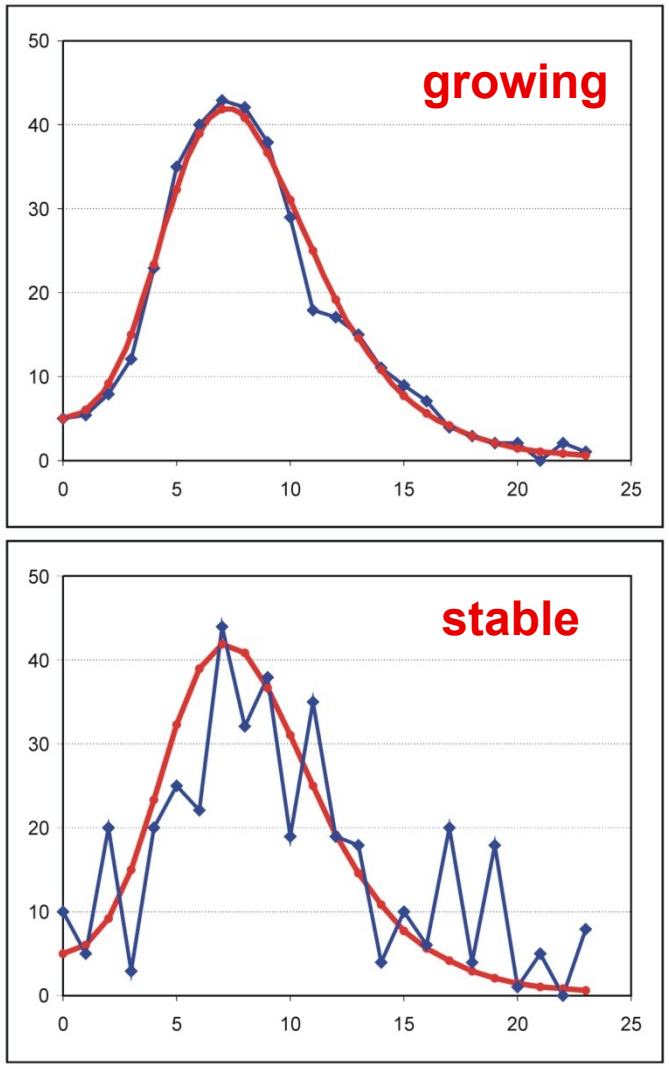


Mixture of similar and divergent sequences




a)
četnost
počet rozdílů


b)
četnost
počet rozdílů



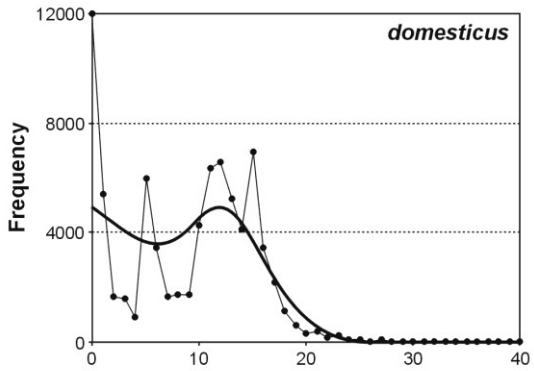
test of agreement between real distribution and prediction:

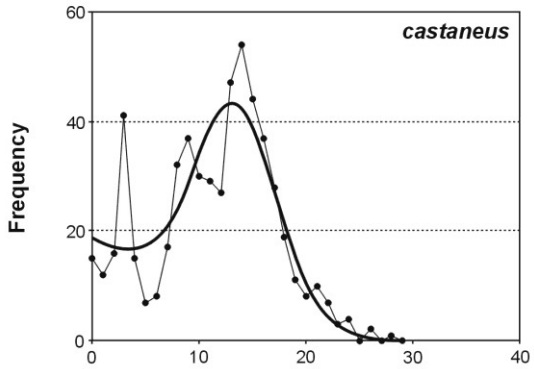
Harpending's raggedness index (Harpending 1994)

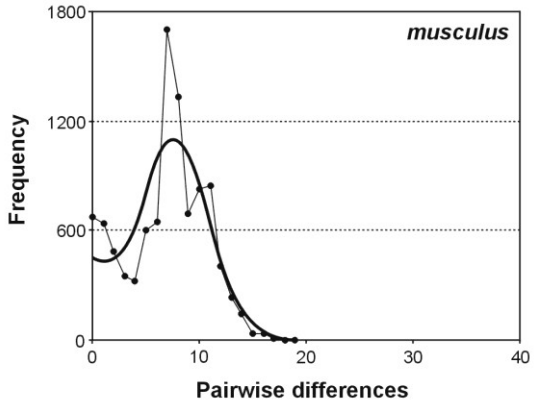
sum of squared deviations

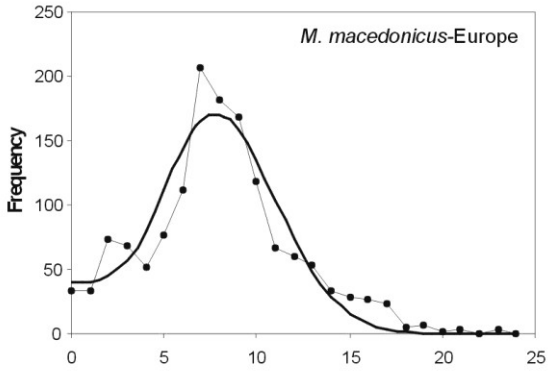
time of expansion/bottleneck:
$\tau = 1/2u$,
where u is mutation rate for whole sequence

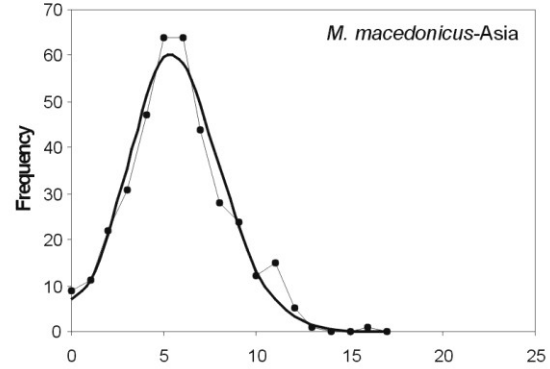
we can also estimate population size before and after expansion

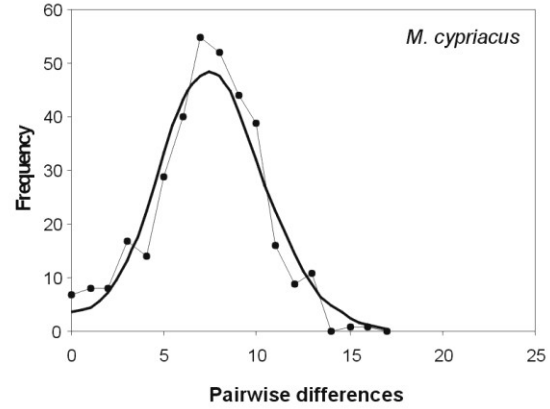

domesticus
Frequency
12000
8000
4000
0
0
10
20
30
40


castaneus
Frequency
60
40
20
0
0
10
20
30
40


musculus
Frequency
1800
1200
600
0
0
10
20
30
40
Pairwise differences


M. macedonicus-Europe
Frequency
250
200
150
100
50
0
0
5
10
15
20
25


M. macedonicus-Asia
Frequency
70
60
50
40
30
20
10
0
0
5
10
15
20
25


M. cypriacus
Frequency
60
50
40
30
20
10
0
0
5
10
15
20
25
Pairwise differences

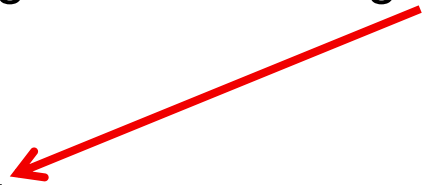
3. ML a Bayesian inference

MCMC

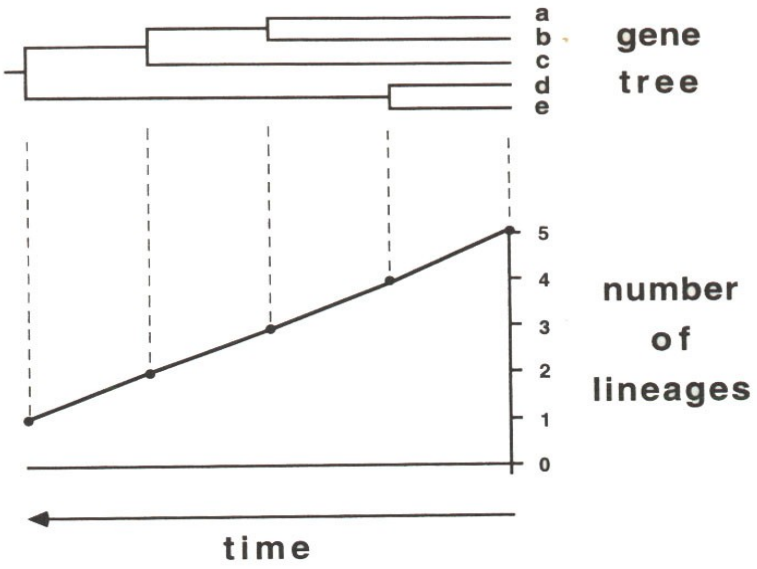
comparison of stable population model and model of exponential growth/decline using LRT with 1 degree of freedom

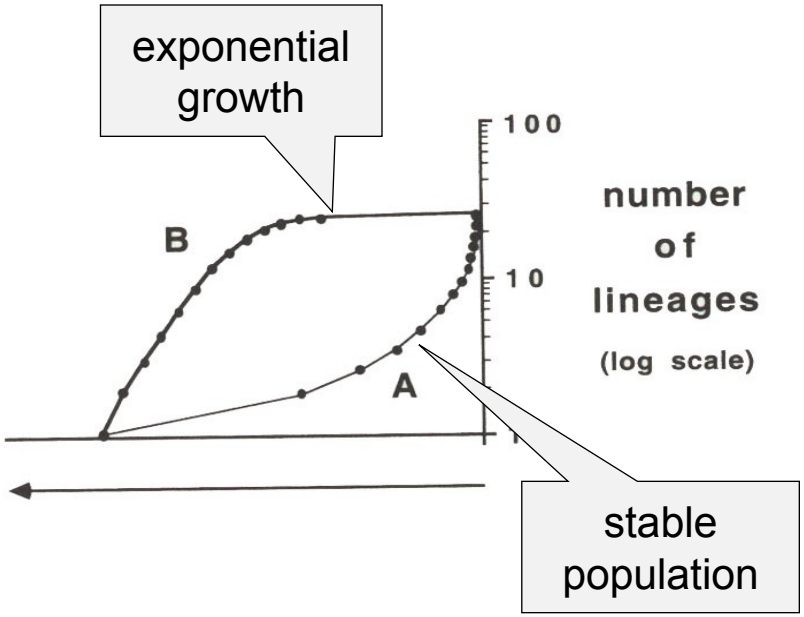
program Fluctuate:

growth parametre *g*
ML i BA approach



4. Bayesian Skyline Plot (BSP)



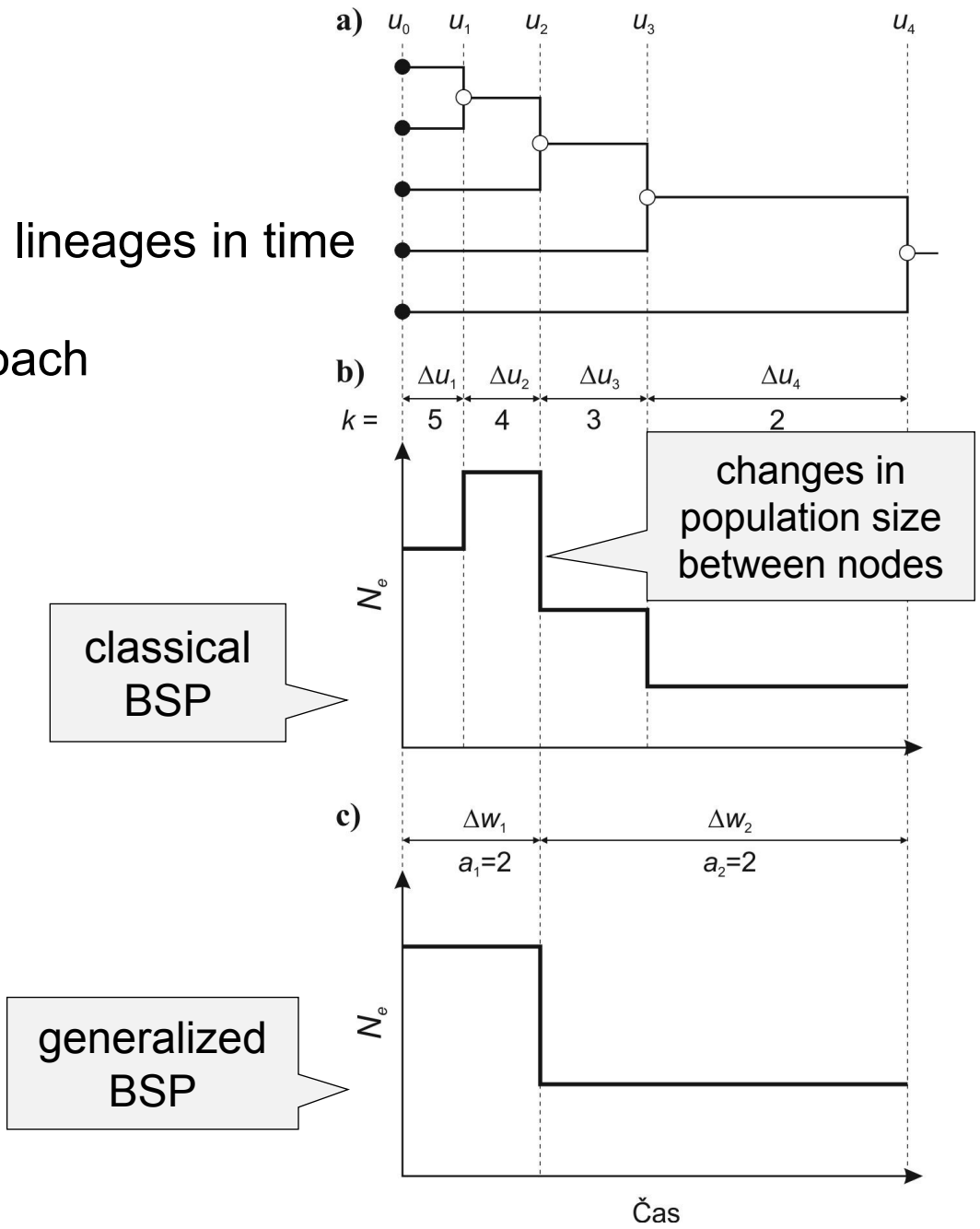


Bayesian skyline plot

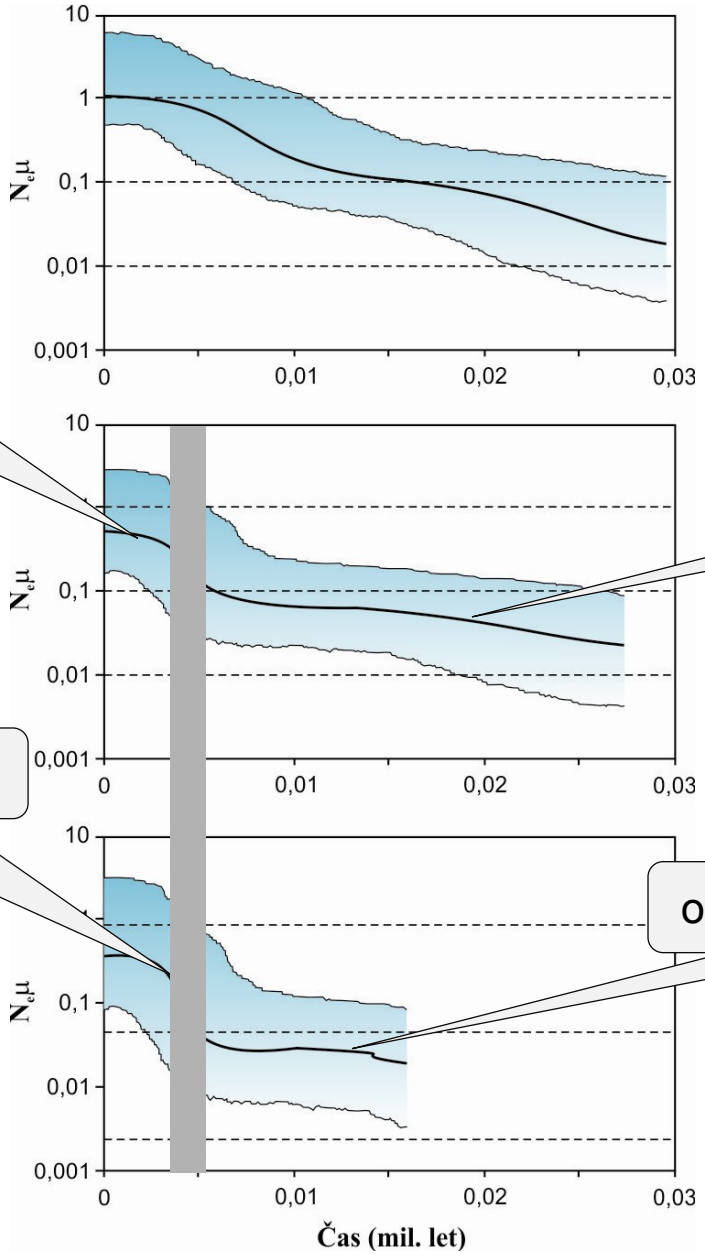
distribution of genealogical lineages in time

BSP is based on this approach

programs BEAST/Tracer



Mouse colonization of Europe



expansion to Europe

domesticus

origin outside Europe

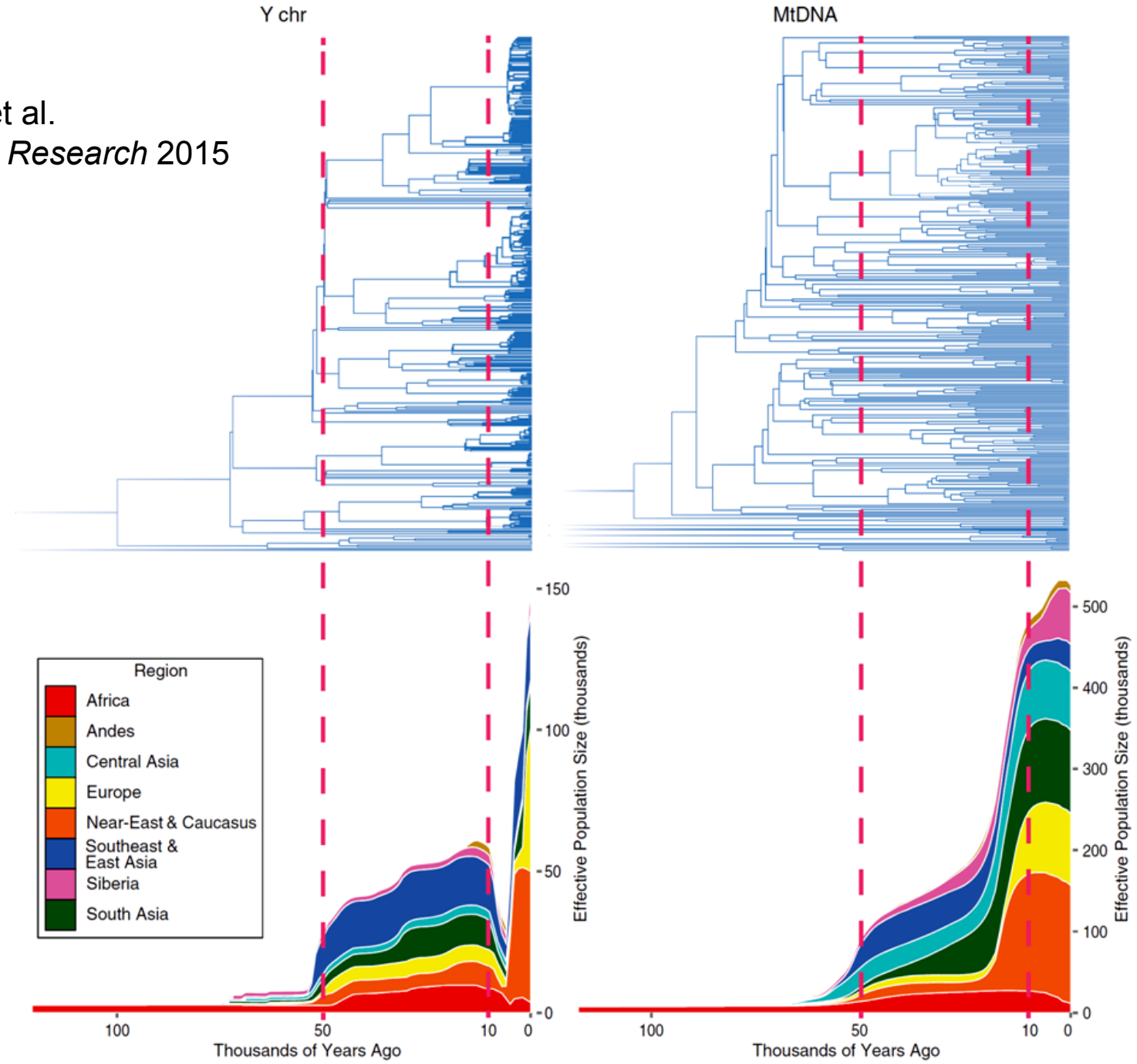
domesticus - Europe

expansion to Europe

origin outside Europe

musculus - Europe

Karmin et al.
Genome Research 2015



Possible results of phylogeografical studies

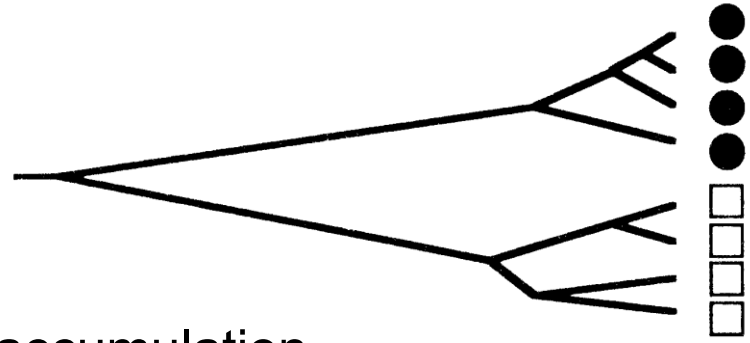
(Avise 2000)

Category I:

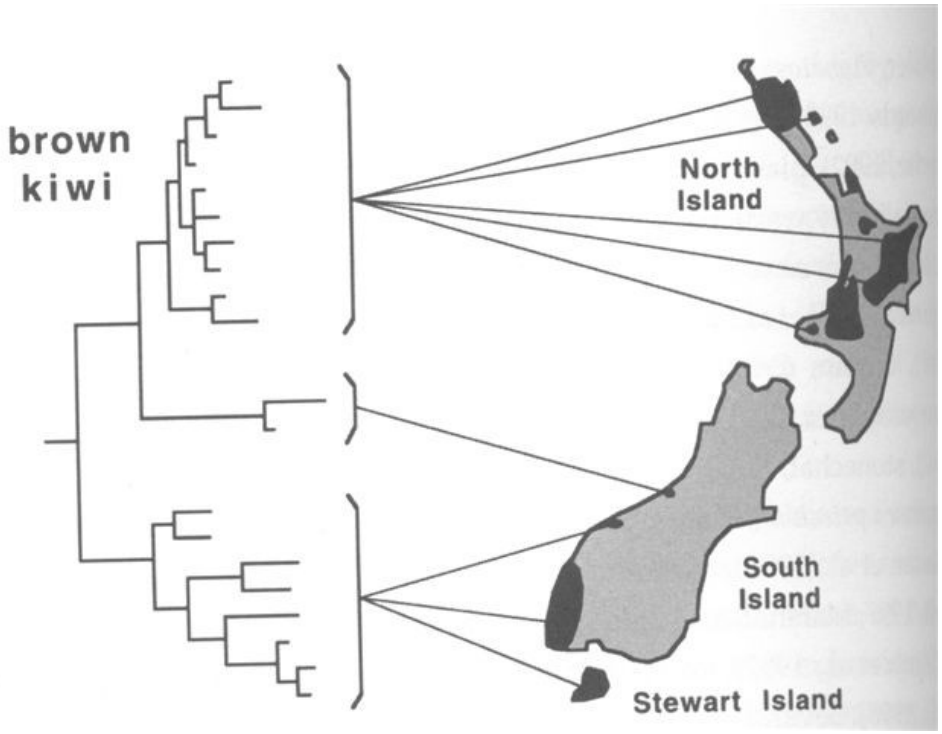
distinct allopatric lineages

barriers to gene flow or low dispersion

differences because of lineage sorting, or accumulation of new mutations

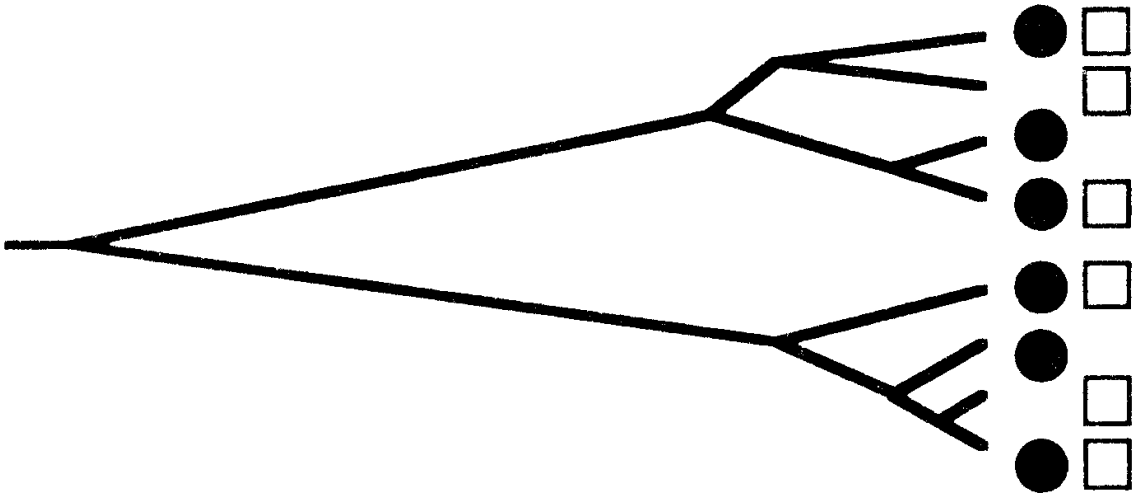


Apteryx australis



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

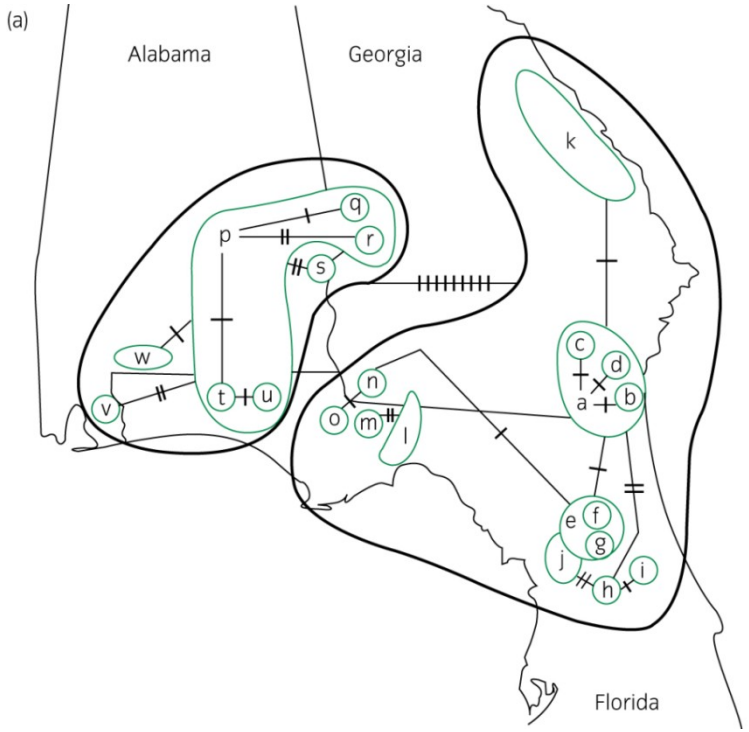
recently, populations in contact

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*



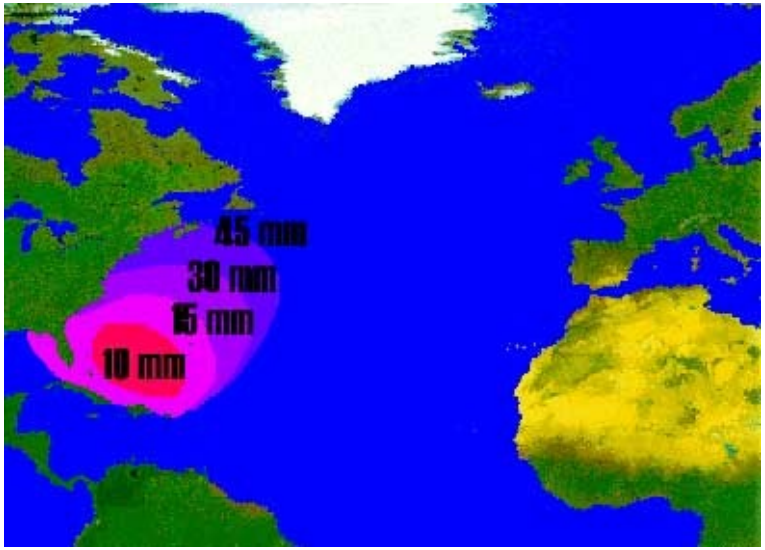
Category IV:

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:

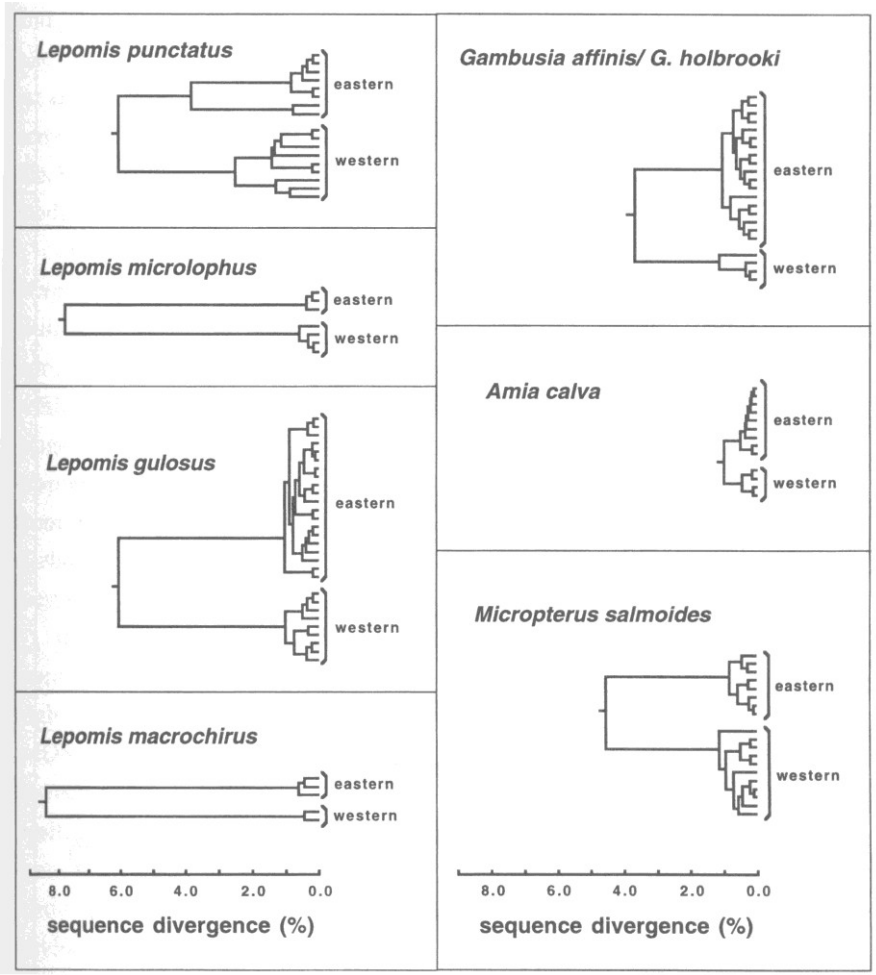
combination of III and IV

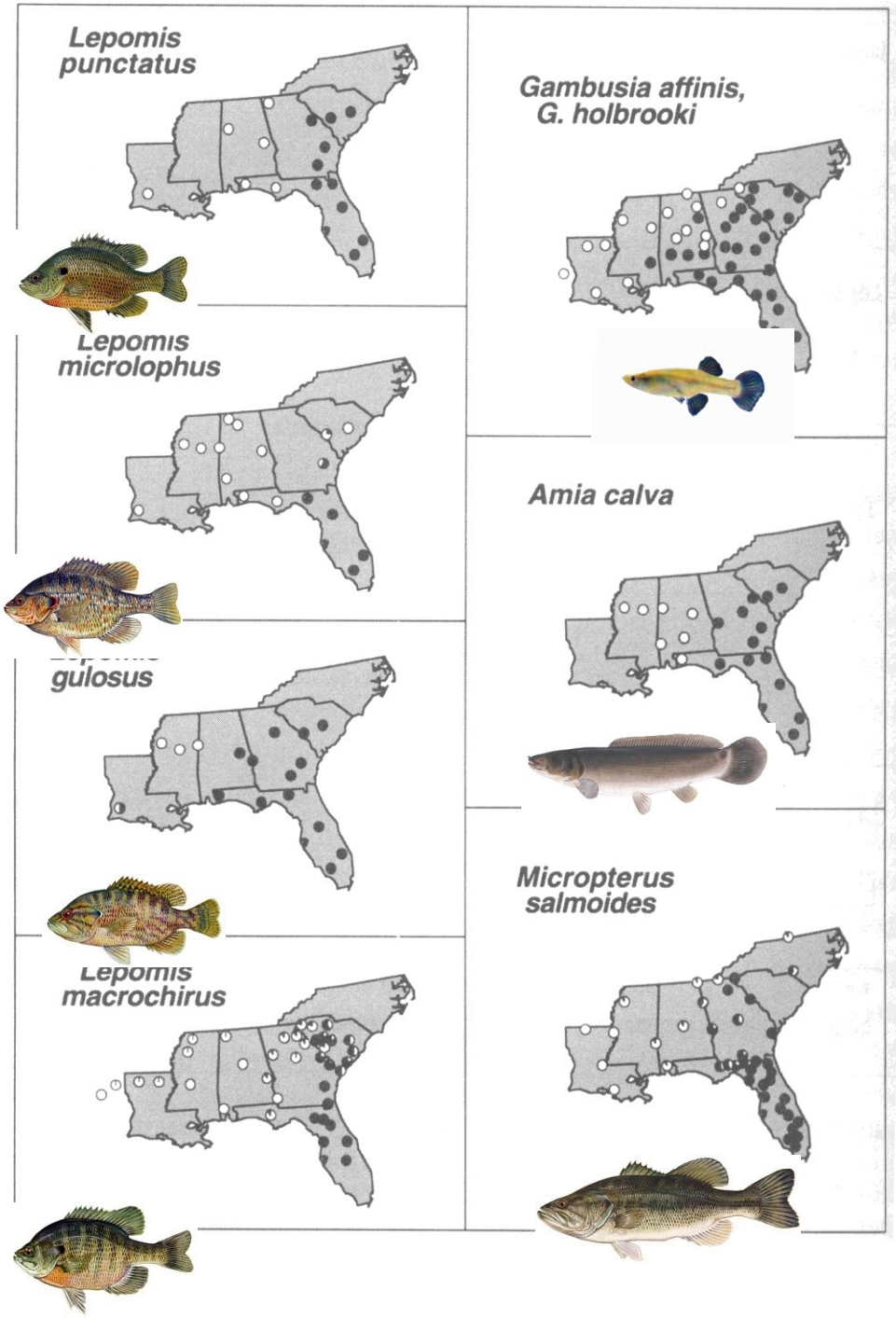
low divergence of lineages

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

we should use private haplotypes as characters

Genealogical concordance
Fishes in SE USA





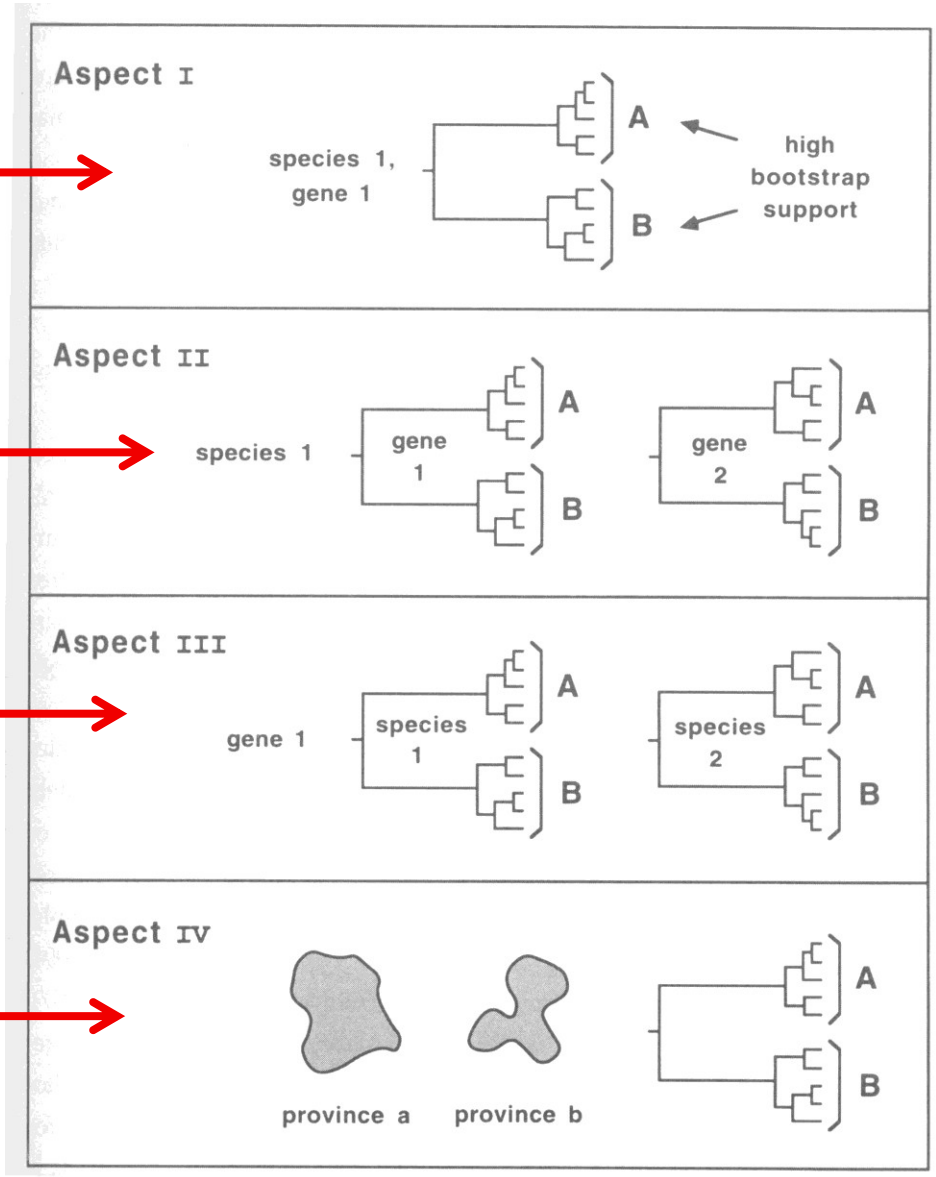
Genealogical concordance (congruence on different levels)

Various parts of gene sequence

More sequences (genes) of the same species

More species in the same region

Support of biogeographical regions (more species, more areas)



Genetic consequences of glaciations

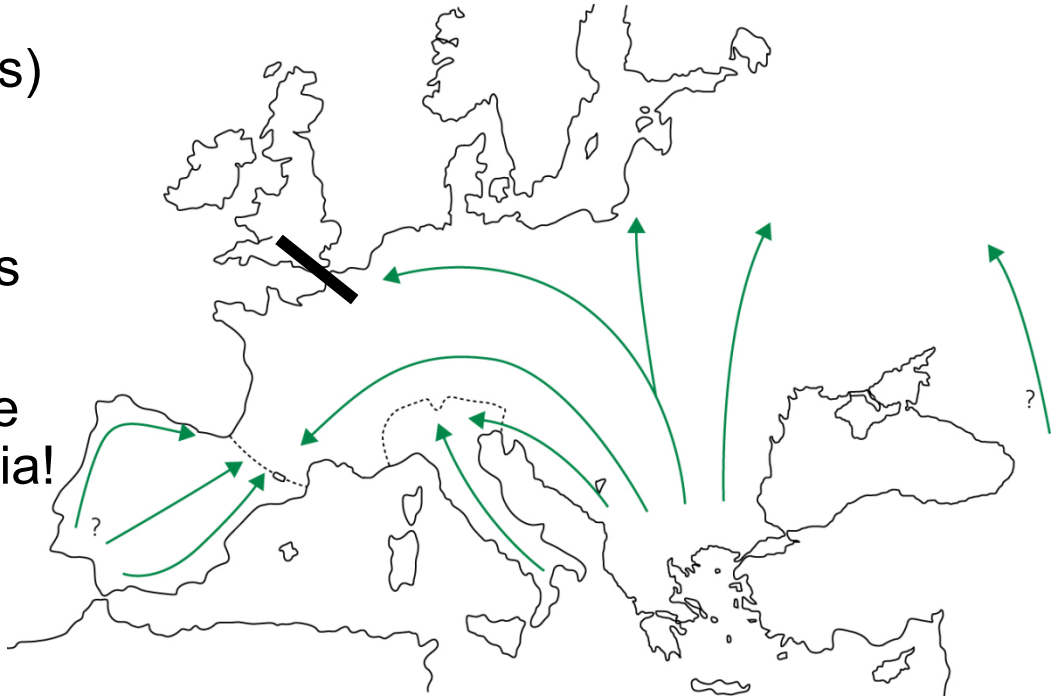
Refugia (Iberian, Apennine, Balkan peninsulas)

In refugia, small populations during relatively long time

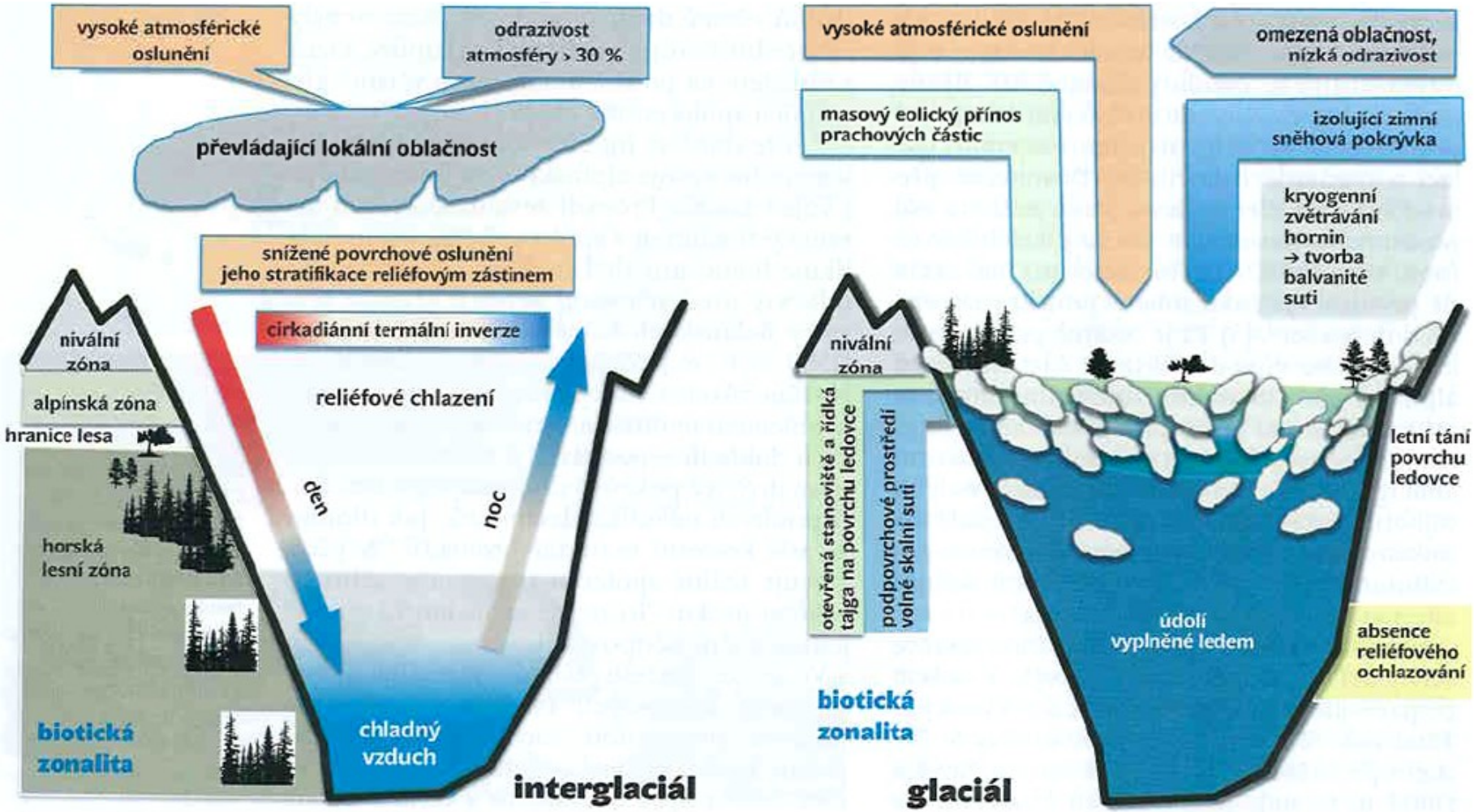
Lineage sorting (+ mutations)

Subsequent expansion → intraspecific hybrid zones

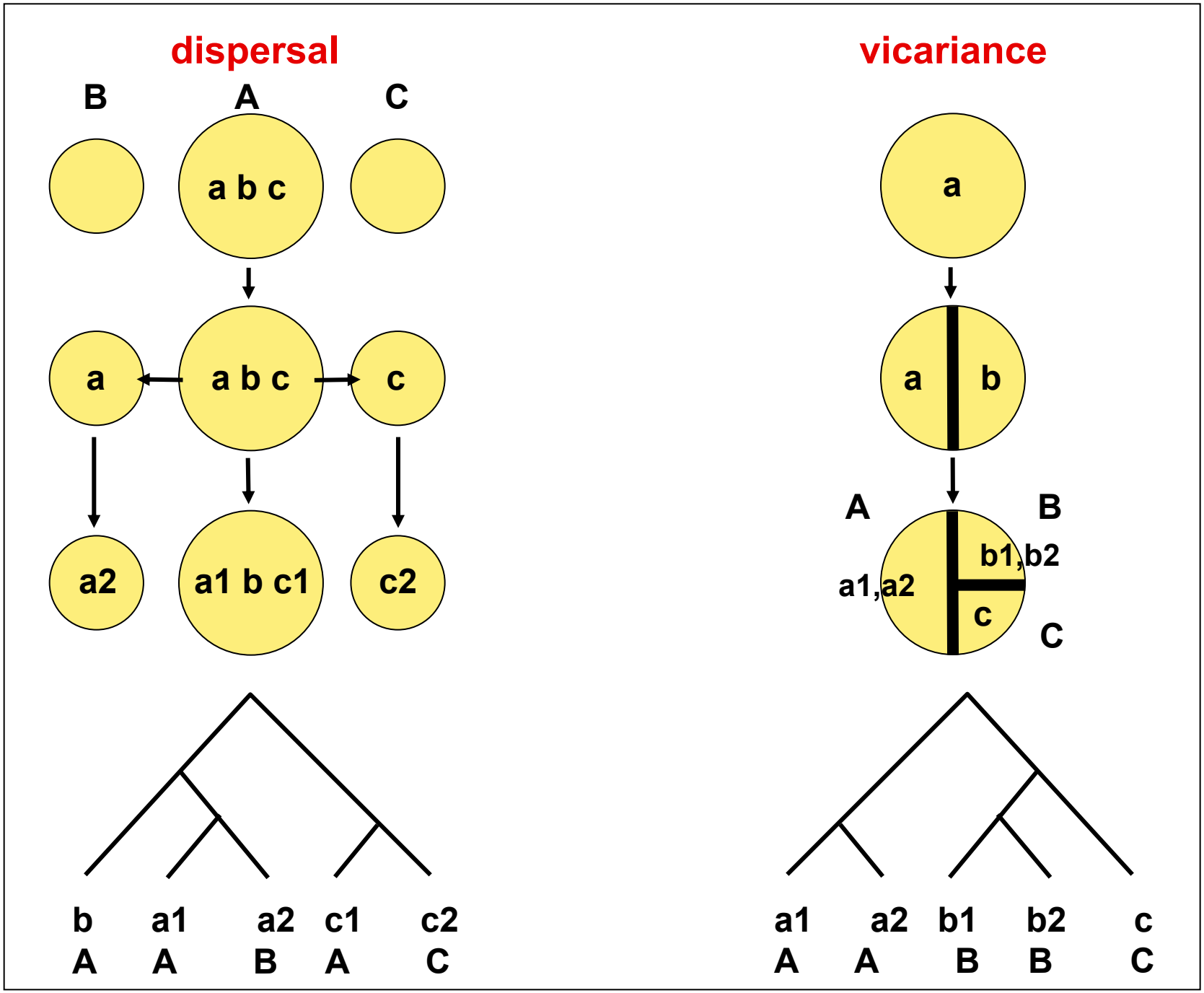
But in several species, there were also northern refugia!



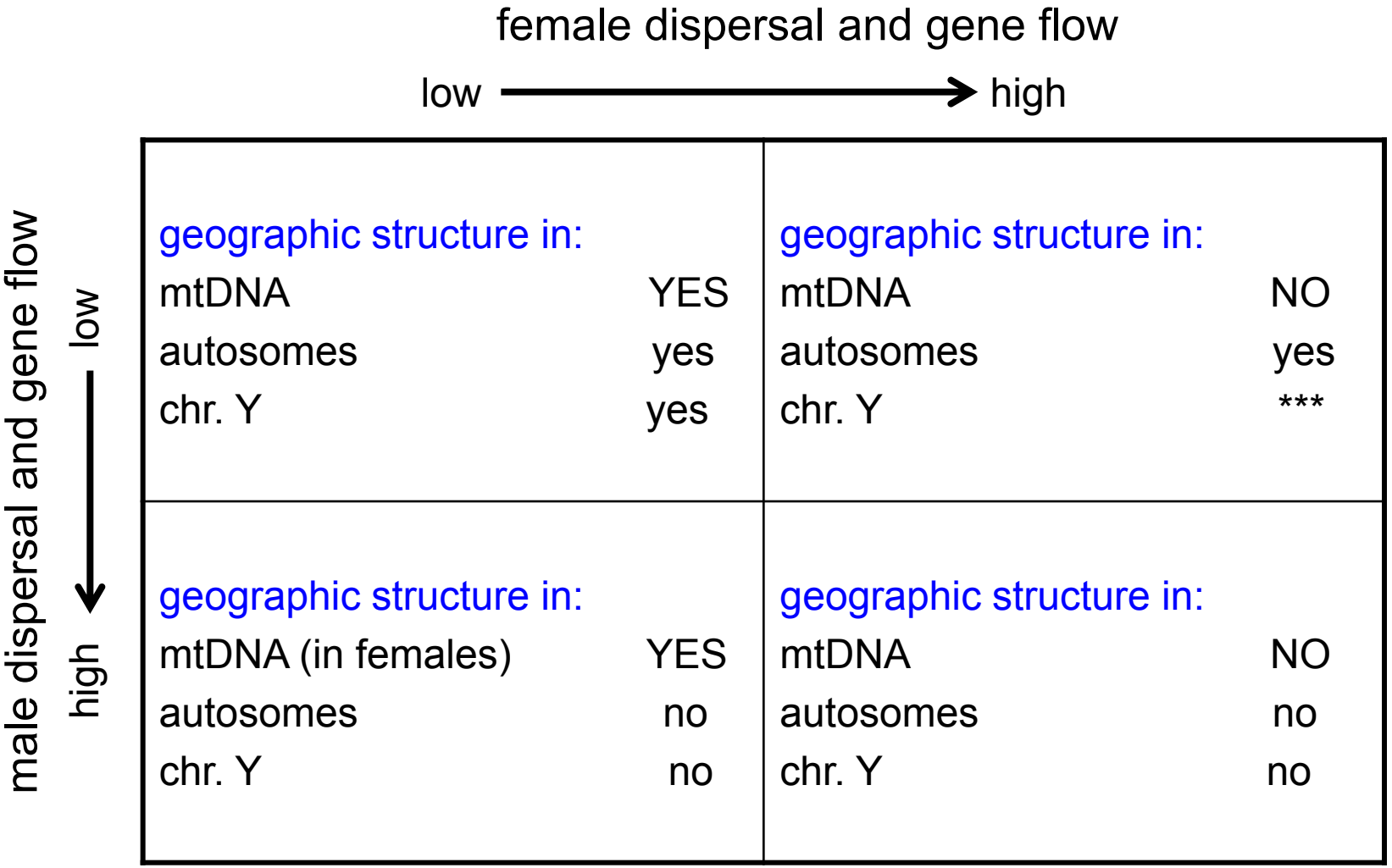
Chorthippus parallelus



Horáček, Vesmír 94 (2015)


dispersal
B
A
C
a b c
a
a b c
c
a2
a1 b c1
c2
b
A
a1
A
a2
B
c1
A
c2
C
vicariance
a
a
b
A
B
b1,b2
a1,a2
c
C
a1
A
a2
A
b1
B
b2
B
c
C

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



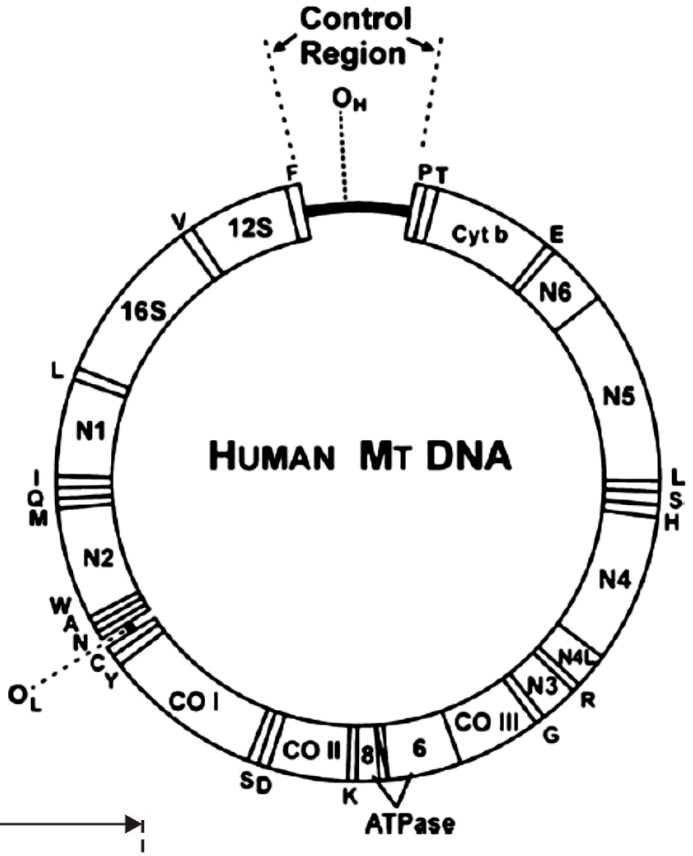
markers:

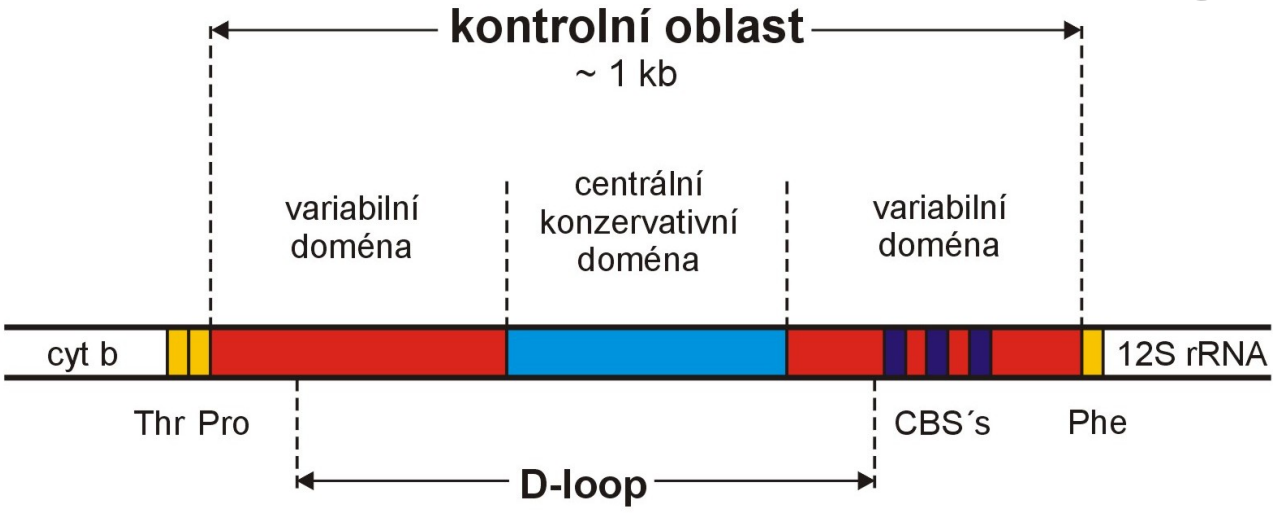
mtDNA sequences

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

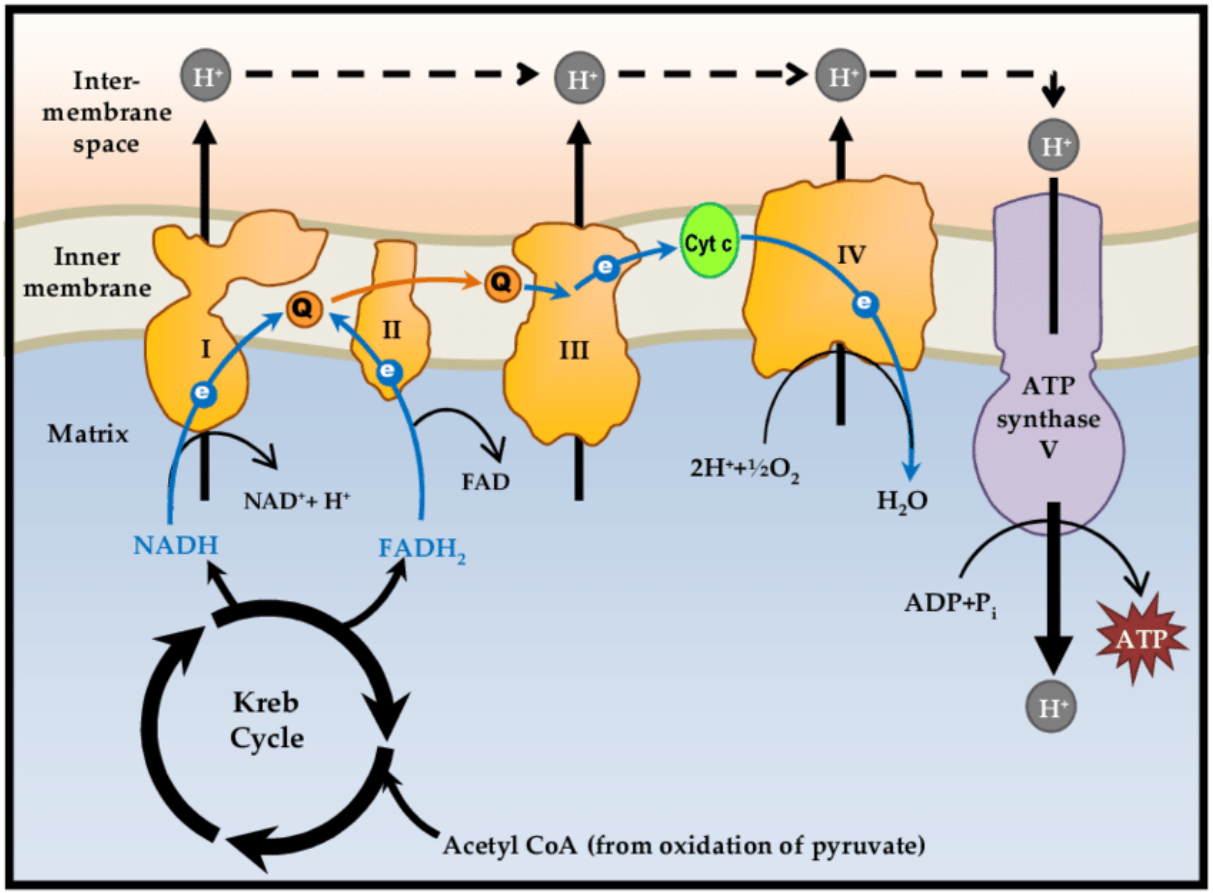
Recombination

Neutrality?

influence on fitness (experimental evidence):

- mouse (*Mus*)
- fruit fly (*Drosophila*)
- human

OXPHOS



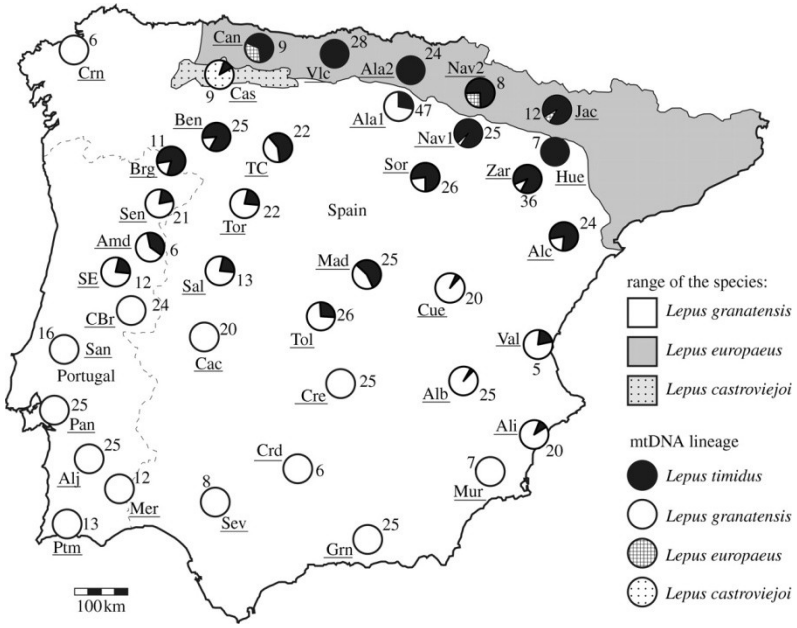
Interspecific introgression:

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)

$\rightarrow$ 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 ≈ 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