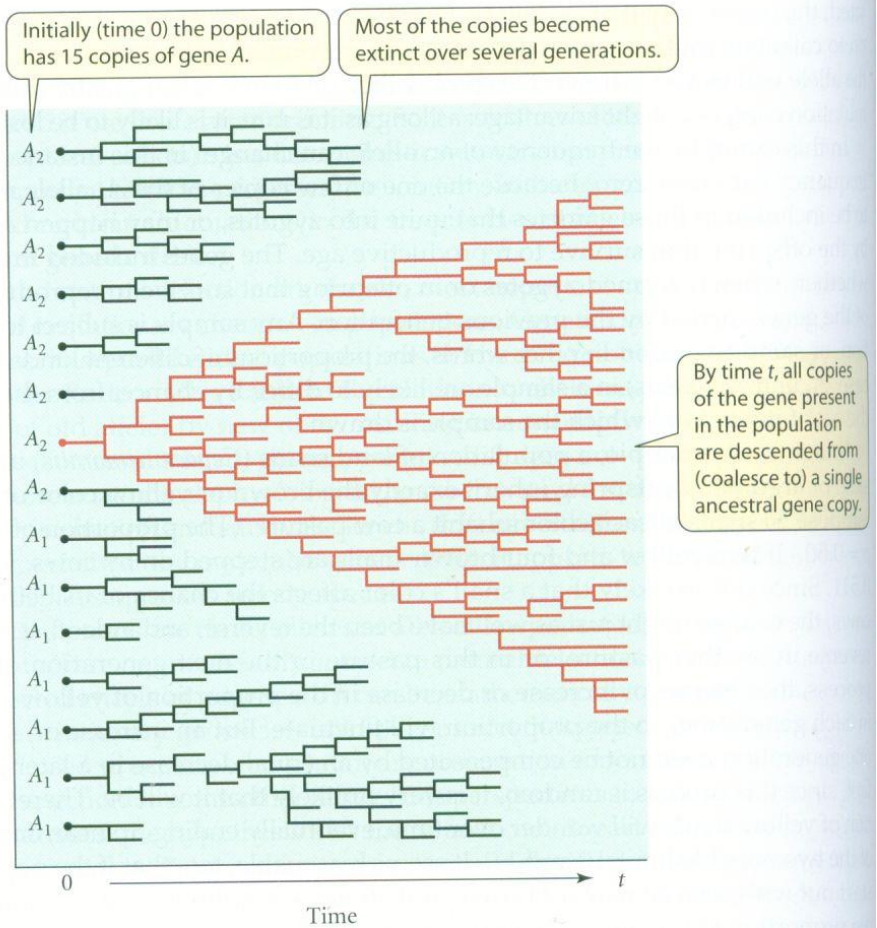
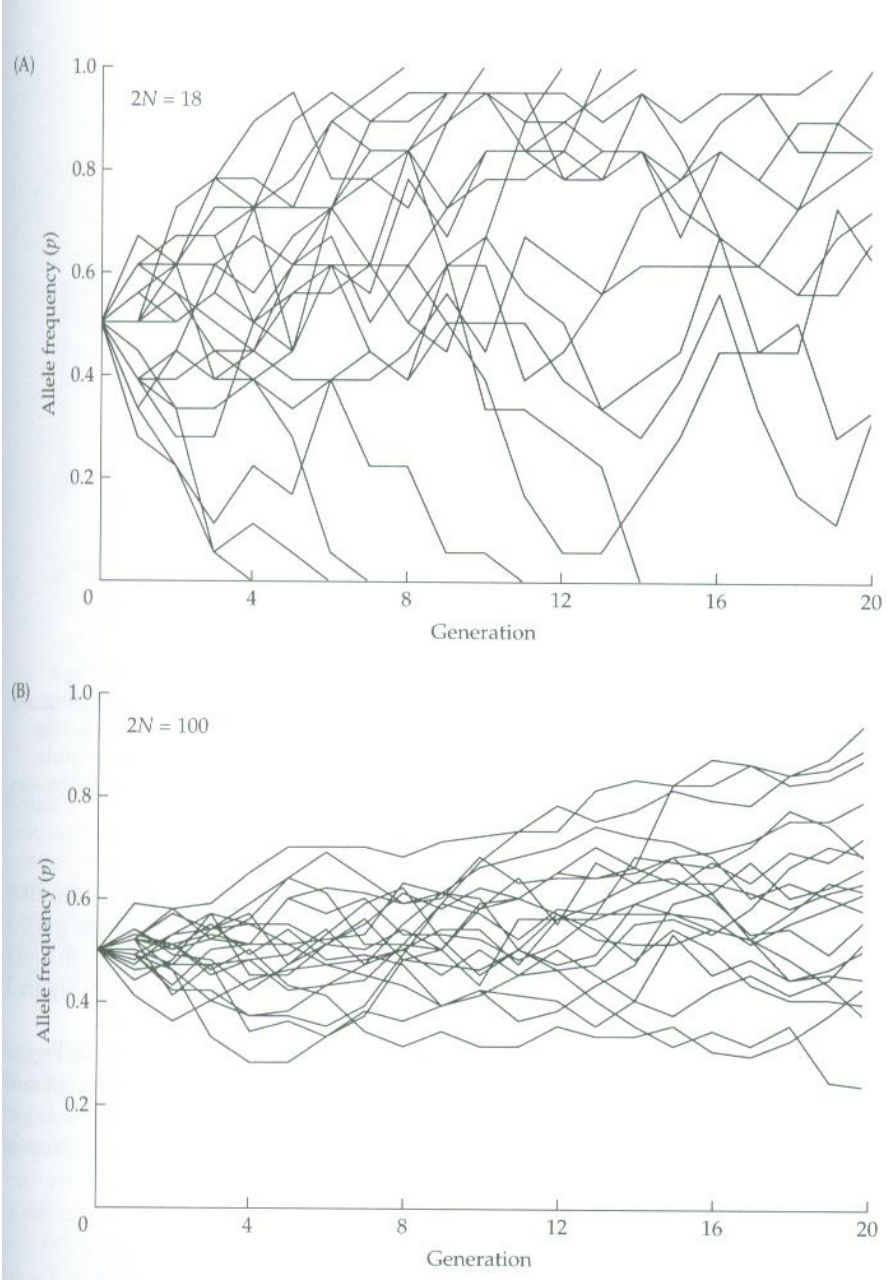
# Koalescence

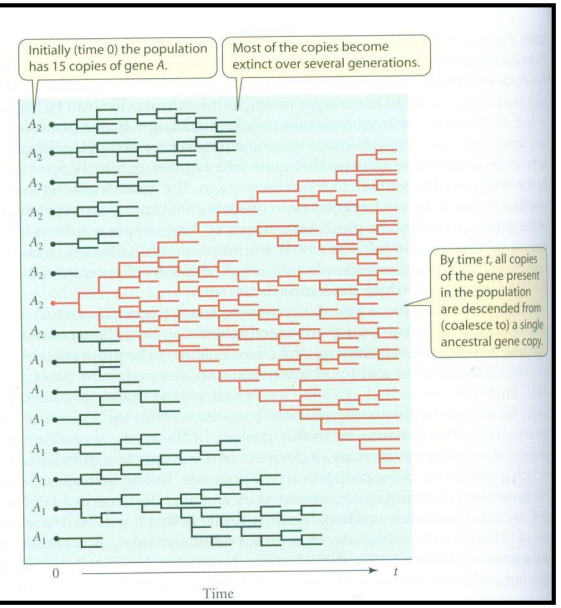

Initially (time 0) the population has 15 copies of gene A.
Most of the copies become extinct over several generations.
By time t, all copies of the gene present in the population are descended from (coalesce to) a single ancestral gene copy.
A2
A2
A2
A2
A2
A2
A2
A2
A1
A1
A1
A1
A1
A1
A1
0
t
Time

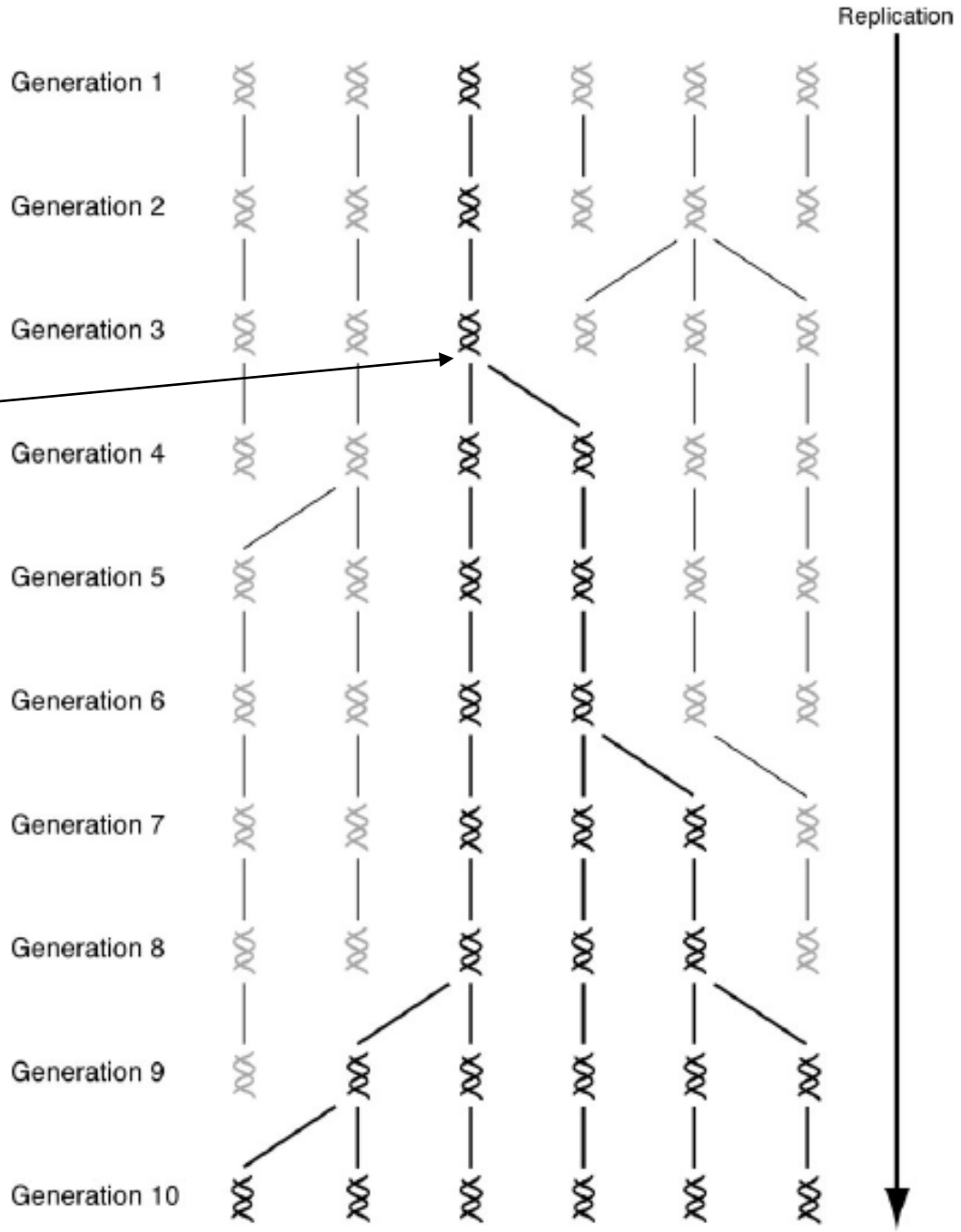

(A)
2N = 18
Allele frequency (p)
1.0
0.8
0.6
0.4
0.2
0
4
8
12
16
20
Generation
(B)
2N = 100
Allele frequency (p)
1.0
0.8
0.6
0.4
0.2
0
4
8
12
16
20
Generation

**Coalesce**
*spojit se, splynout, sloučit se*

*Didaktická simulace*
*http://www.coalescent.dk/*

**Coalescence**
= splynutí linií



# Koalescence

- Matematický model, který popisuje průběh genealogií.

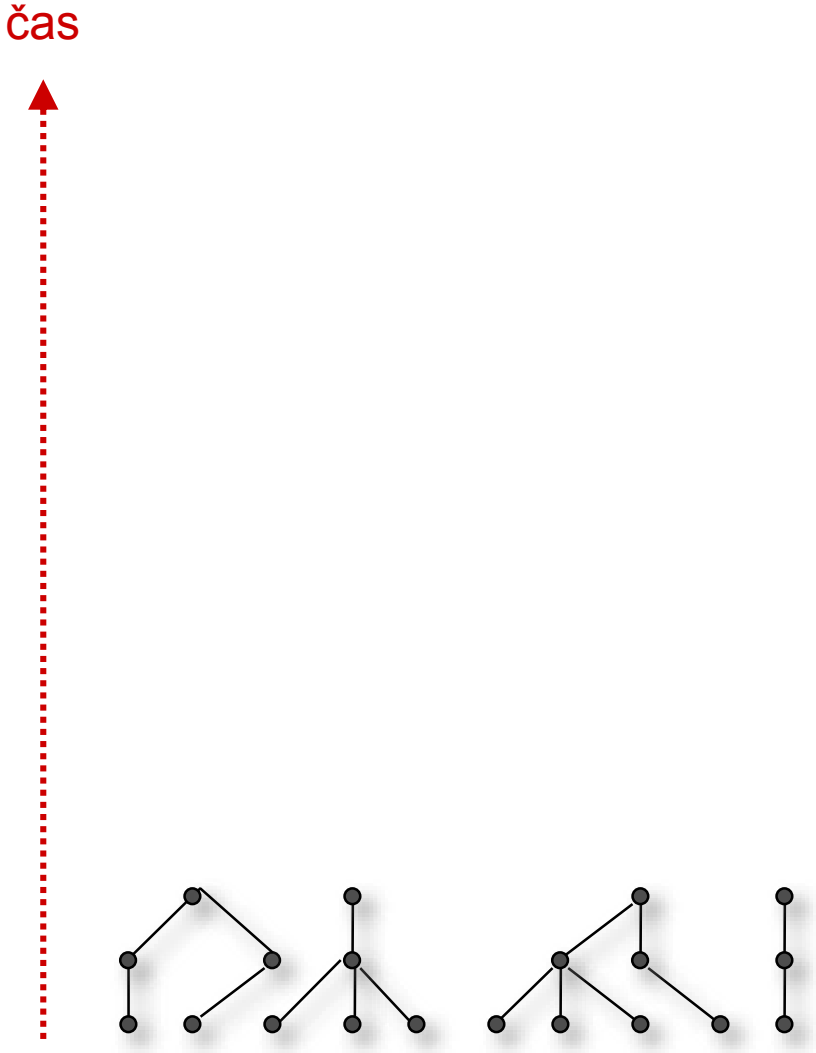
- Postupujeme opačně v čase než u klasických modelů populační genetiky (Wright-Fisher model)

# Teorie koalescence

koalescence = splynutí linií

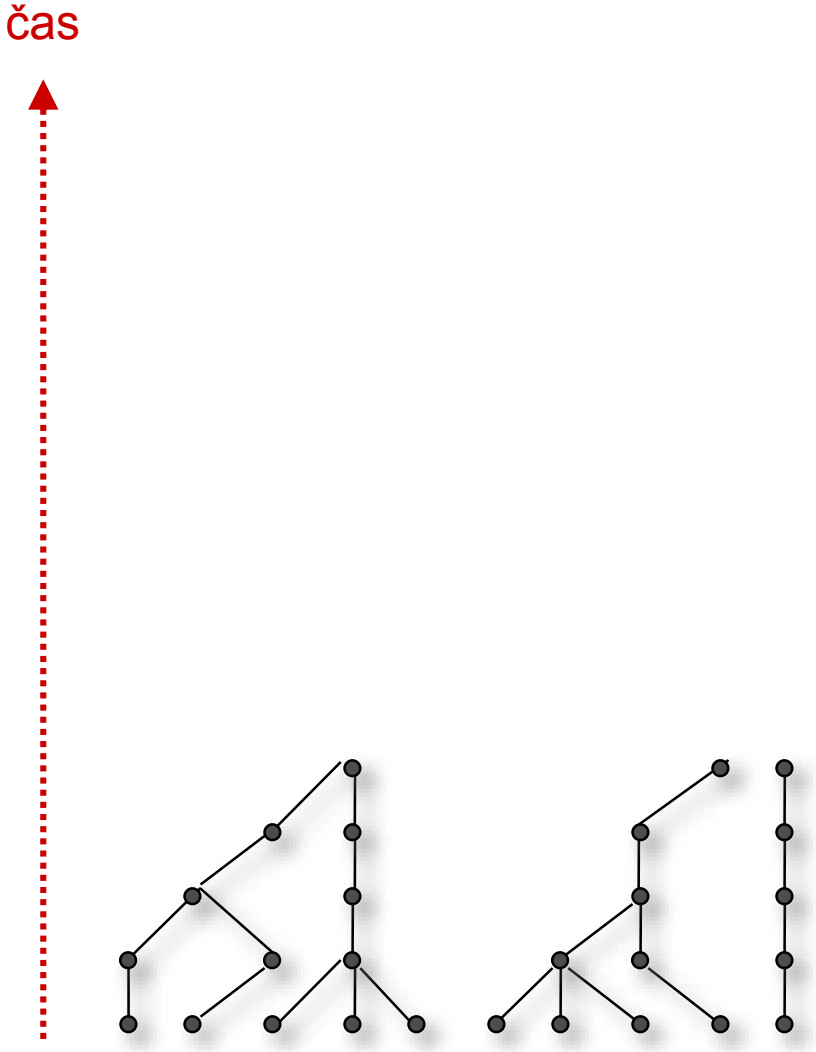
# Teorie koalescence

koalescence = splynutí linií



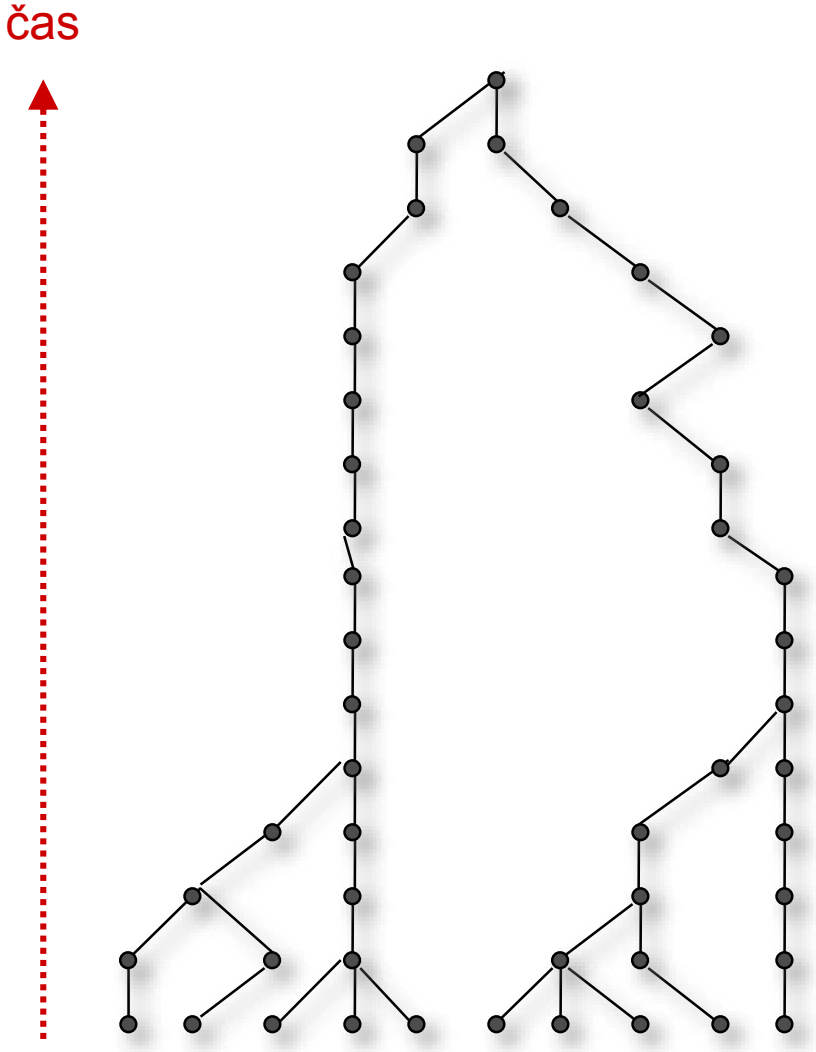
# Teorie koalescence

koalescence = splynutí linií

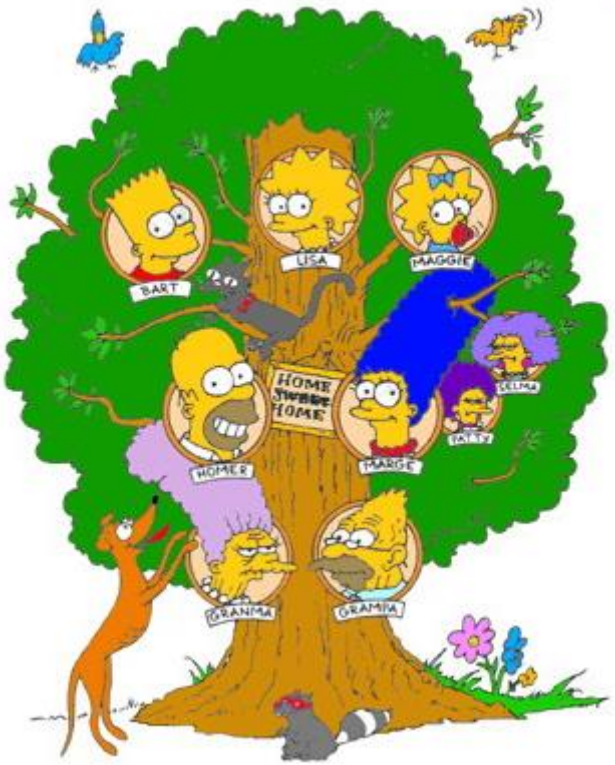


# Teorie koalescence

koalescence = splynutí linií

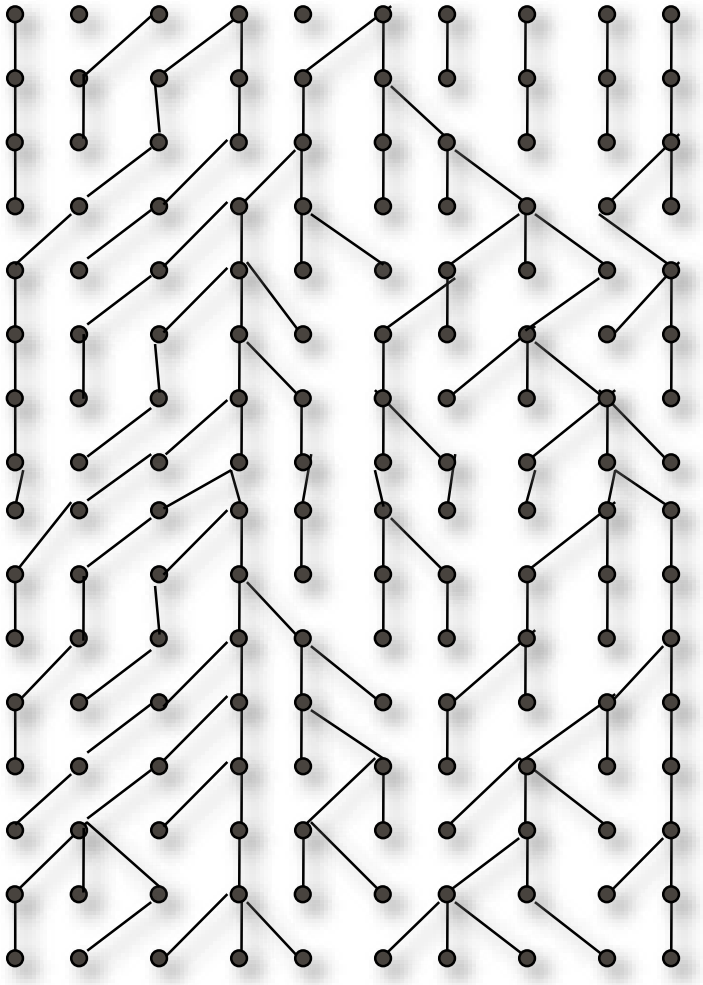


- **Bořivoj** (? - 894) = Ludmila
  - **Spytihněv** (894 - 905)
  - **Vratislav** (905 - 921) = Drahomíra
    - **Václav** (921 - 935)
    - **Boleslav I.** (935 - 967)
      - Algedina = **Boleslav II.** (967 - 999) = Emma
        - **Boleslav III.** (999 - 1003)
        - **Jaromír** (1003/4-12, 1034)
        - **Oldřich** (1012 - 1034) = Božena
          - **Břetislav** (1034 - 1055) = Jitka
            - **Spytihněv II.** (1055 - 1061) = Hidda
              - Svatobor
            - Adléta = **Vratislav II.** (1061 - 1092) (1068 - král) = Svatava
              - **Břetislav II.** (1092 - 1100)
            - **Konrád I. Brň.** (1092) = Wilpirk
              - **Oldřich Brň.** (1100)
                - Vratislav
              - Litold Znoj.
                - Konrád II. Znoj.
                - **Konrád III. Ota** (1189 - 1191)
                - Helena
            - **Ota I. Olom.** = Eufemie
              - **Svatopluk** (1107 - 1109)
                - Václav
              - Ota II. = Žofie
                - Ota III. Dětleb
            - **Jaromír** (biskup)
            - **Bořivoj II.** (1101-07, 1117-20)
            - **Vladislav I.** (1109 - 1125) = Ridenza
              - Gertruda = **Vladislav II.** (1140 - 1173) (1158 - král) = Judita
                - Svatopluk
                - Anežka
                - **Bedřich** (1172 - 1173) (1178 - 1189)
                - Vojtěch (biskup)
                - **Přemysl I.** (1192-93, 1197-30) (1204 - král)
                - **Otakar**
                - **Jindřich** (1197)
                - **Vladislav**
                - Richsa
              - Děpolt
              - **Jindřich** = Markéta
                - **Jindřich Břetislav** (1193 - 1197) (biskup, kníže)
            - **Soběslav I.** (1125 - 1140) = Adléta
              - **Soběslav II.** (1173 - 1178)
              - **Oldřich** (1173)
              - **Václav II.** (1191)
              - Vladislav
              - Marie


BART
LISA
MAGGIE
HOME SWEET HOME
SELMA
PATTY
HOMER
MARGE
GRANMA
GRAMPA

# Genové genealogie (gene trees)

Popisují cestu jakou se jednotlivé geny dědily z generace na generaci
Geny, které nejsou ve vazbě, mají jiné genealogie



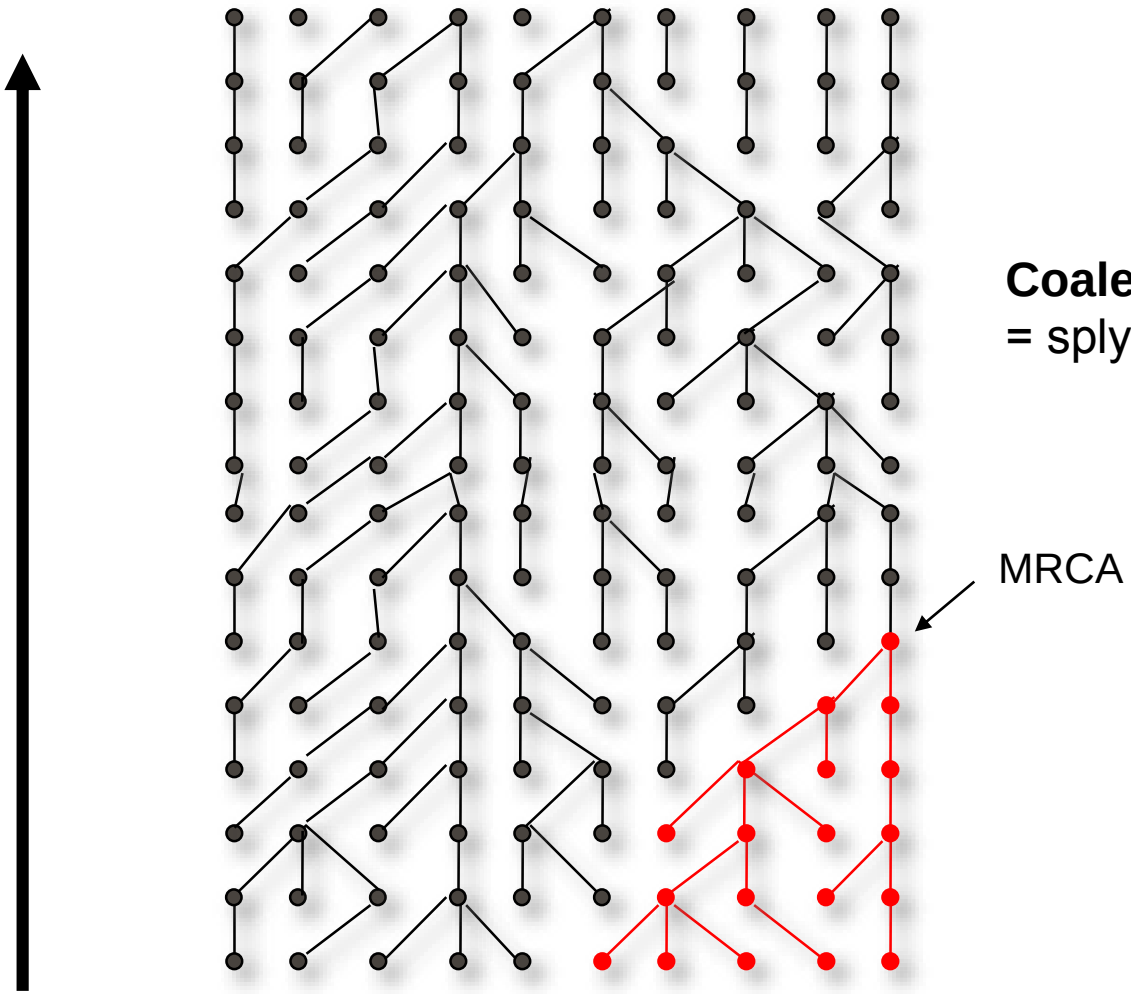
genový strom

# Genové genealogie

Geny, které mají společného předka „Identical by descent“

Nejbližší společný předek vybraných linií „Most Recent Common Ancestor“ (MRCA)

Zpět v čase



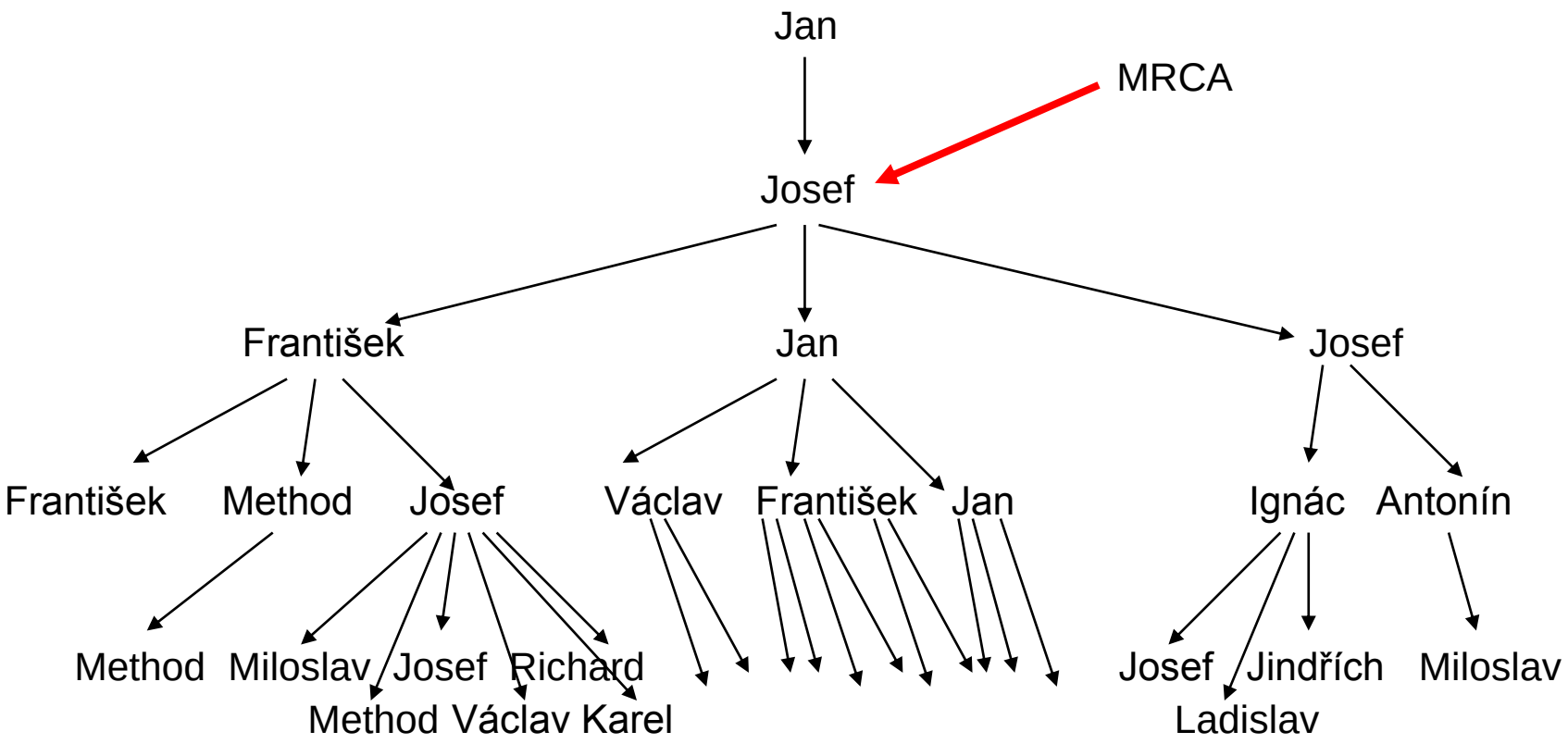

**Coalescence**
= splynutí linií v MRCA

MRCA - příklad

**Stolař z Rawengisburgu Jan Montclinger**

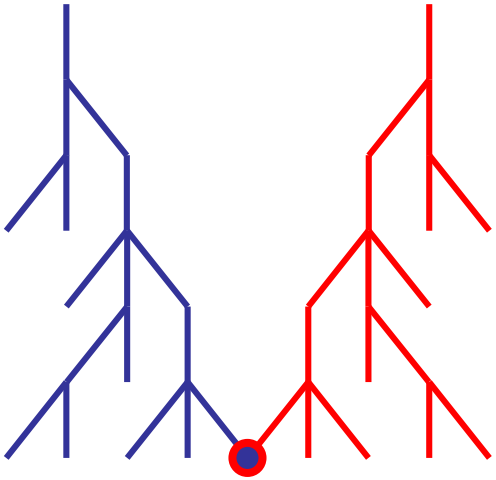
Anna Skobrtalin (vdova Škobrtalová)

* 11.2.1793 Josef Munclinger


Jan
MRCA
Josef
František
Jan
Josef
František
Method
Josef
Václav
František
Jan
Ignác
Antonín
Method
Miloslav
Josef
Richard
Method
Václav
Karel
Josef
Jindřích
Miloslav
Ladislav

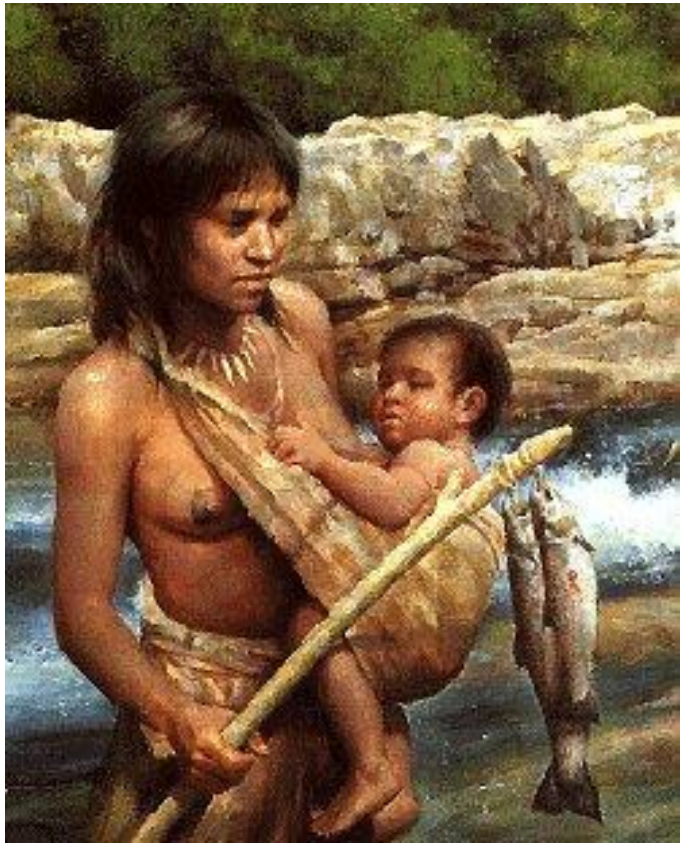
# Meč a přeslice

- Rodokmen odpovídá historii chromosomu Y (pokud ovšem ženy nebyly nevěrné)
- Rodokmen přes matky (a pro mtDNA) by byl úplně jiný!
- A jiný bude pro jiné geny



Mitochondriální Eva

Adam pro chromosom Y

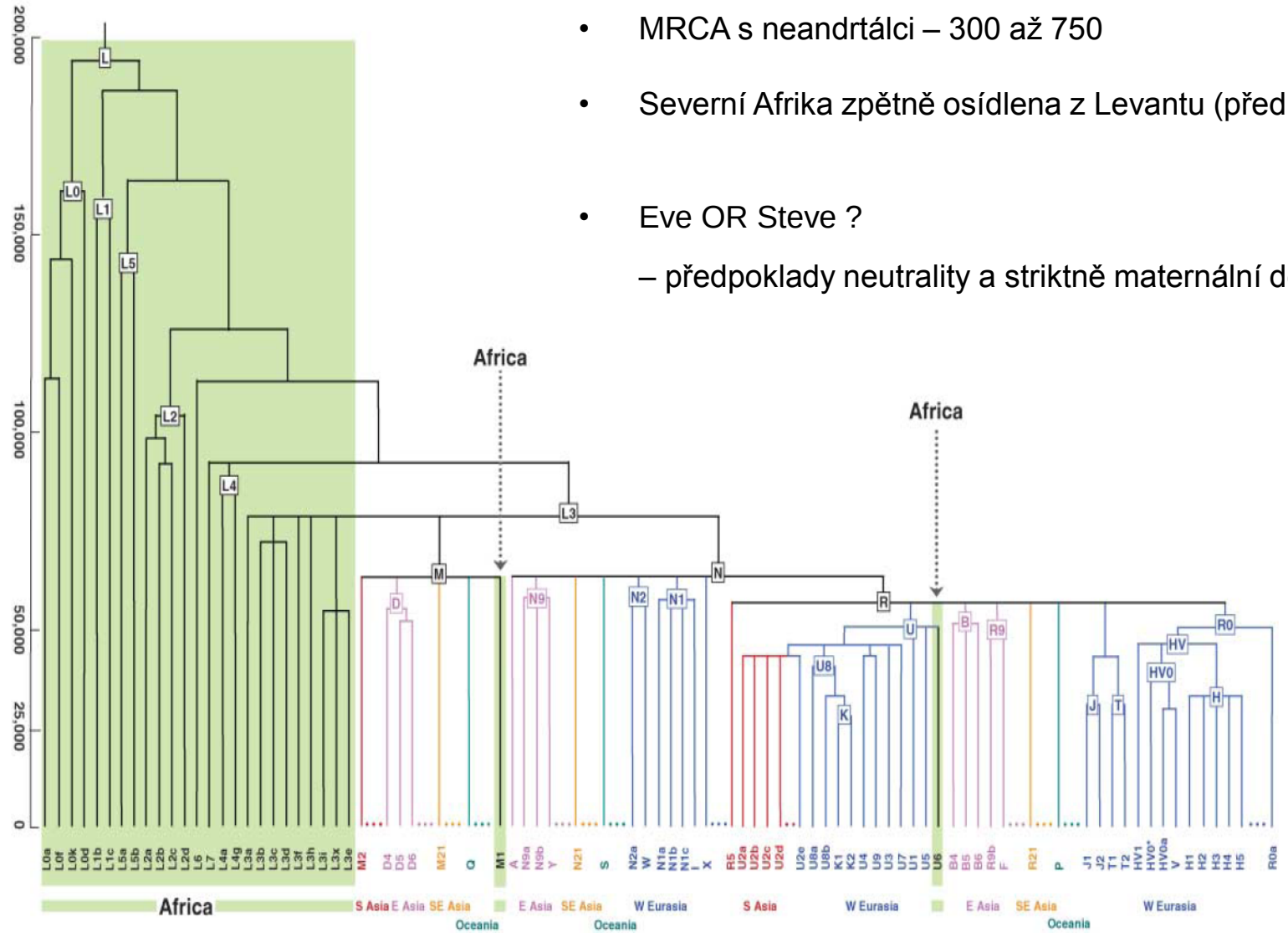


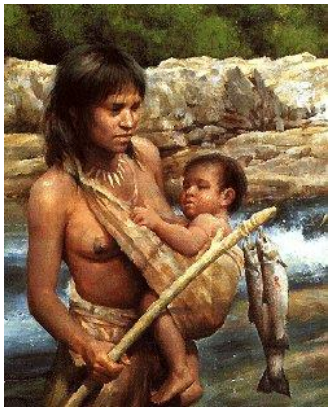
# mtDNA

## a Mitochondriální Eva

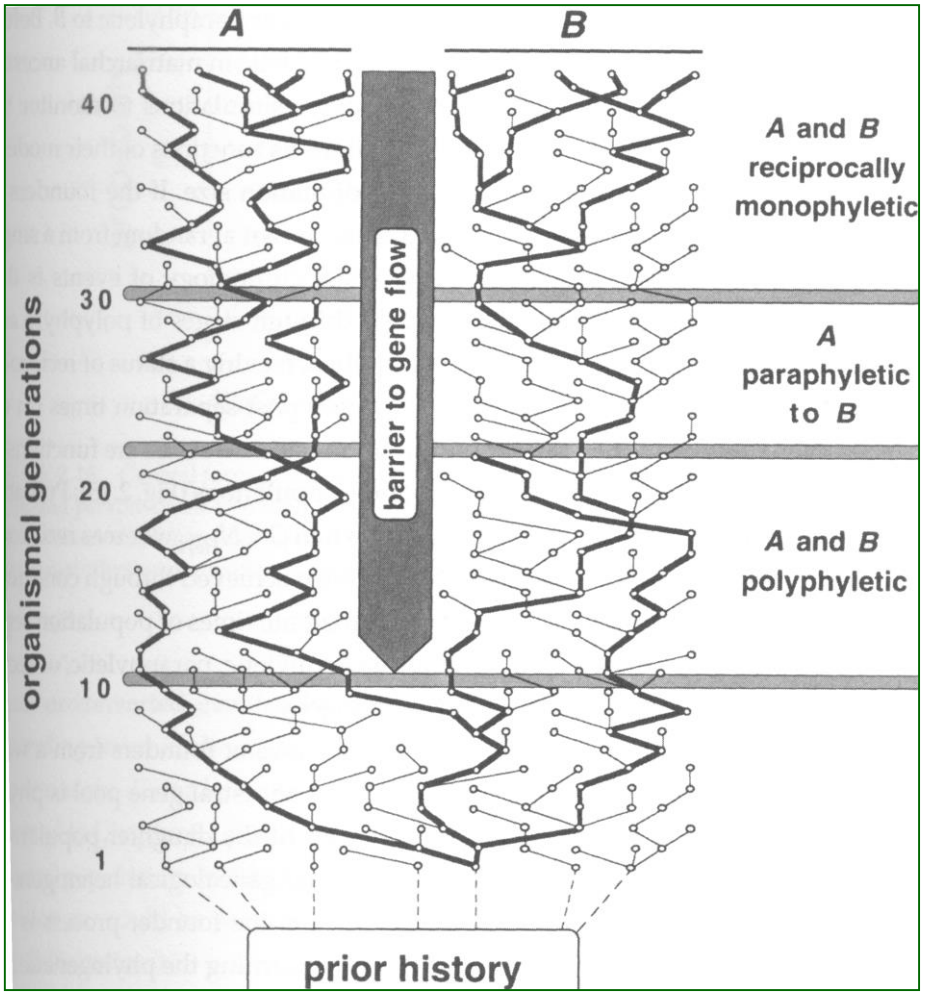
- Větve stromu
  - Afrika
  - jedna z afrických větví - zbytek světa
- MRCA pro mtDNA - 100 až 250 tisíc let
- MRCA s neandrtálci – 300 až 750
- Severní Afrika zpětně osídlena z Levantu (před 40-45 tisíci let)
- Eve OR Steve ?

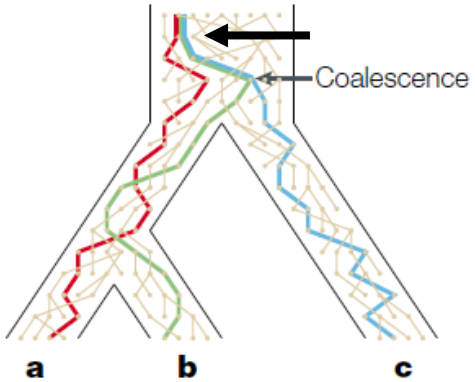
  – předpoklady neutrality a striktně maternální dědičnosti

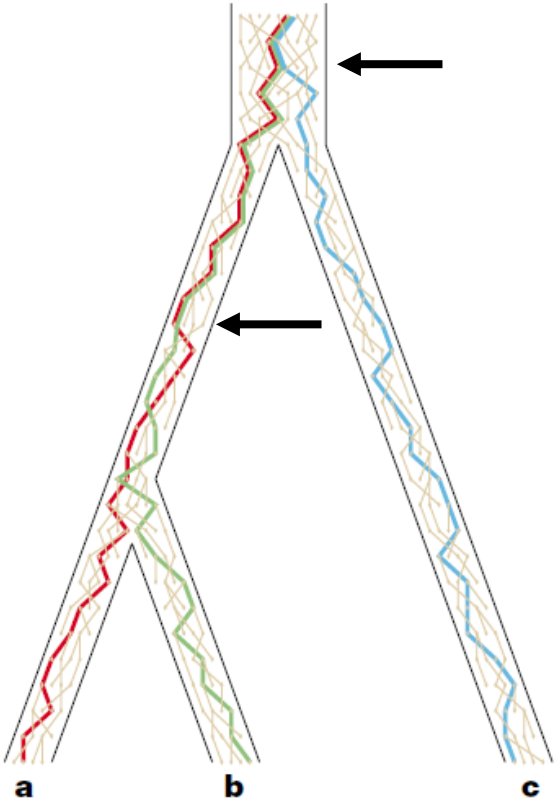


# Přetrvání ancestrálního polymorfismu
## (species tree ≠ gene tree)

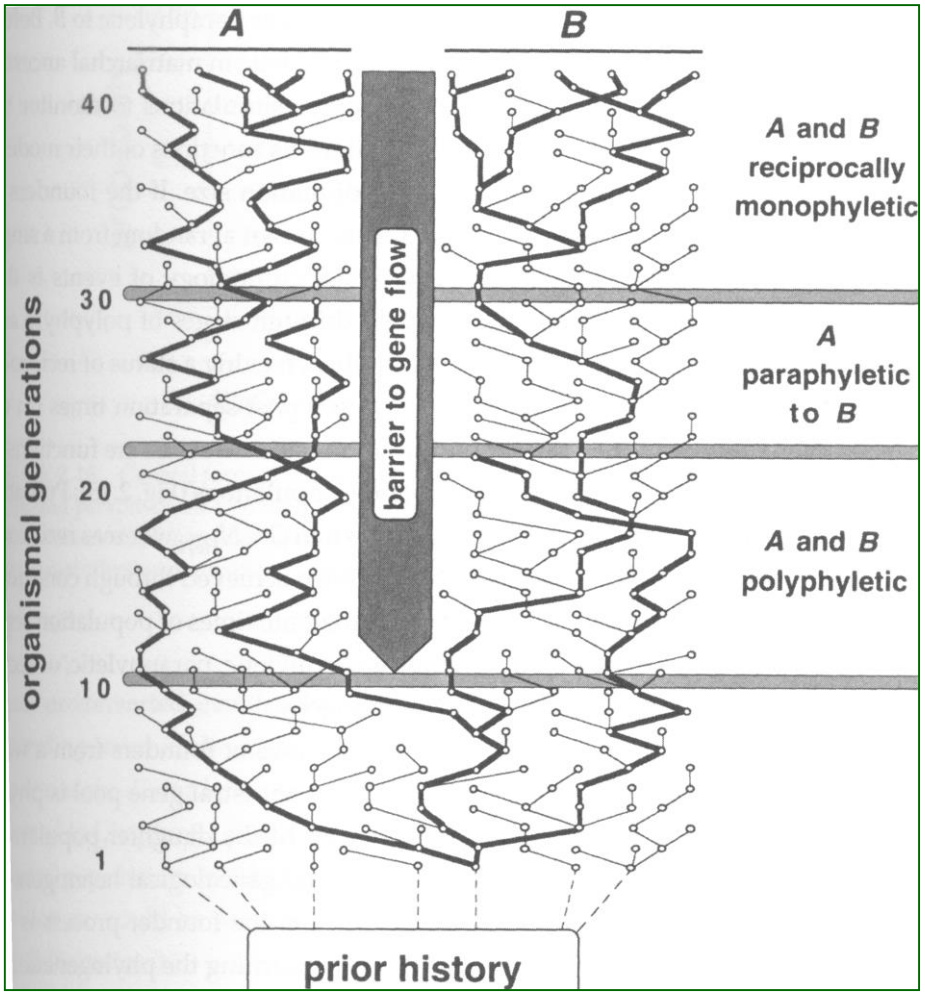


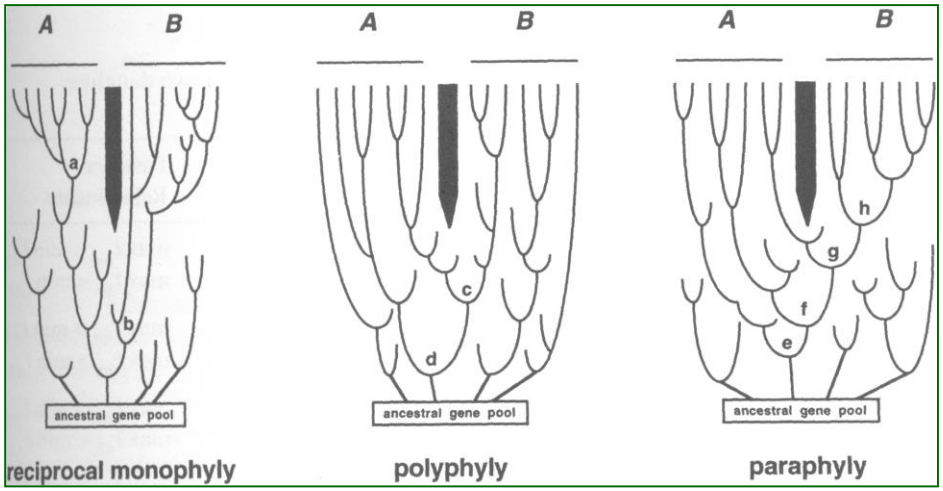



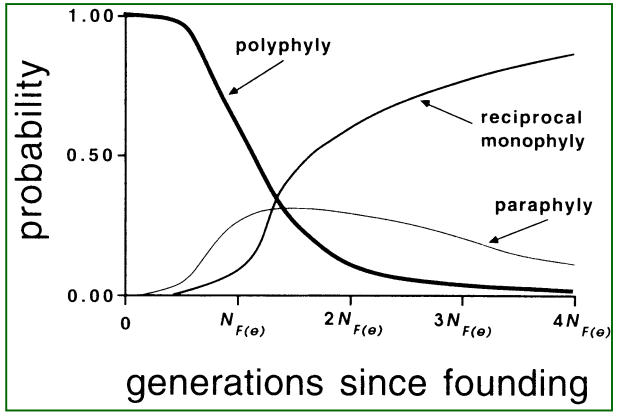



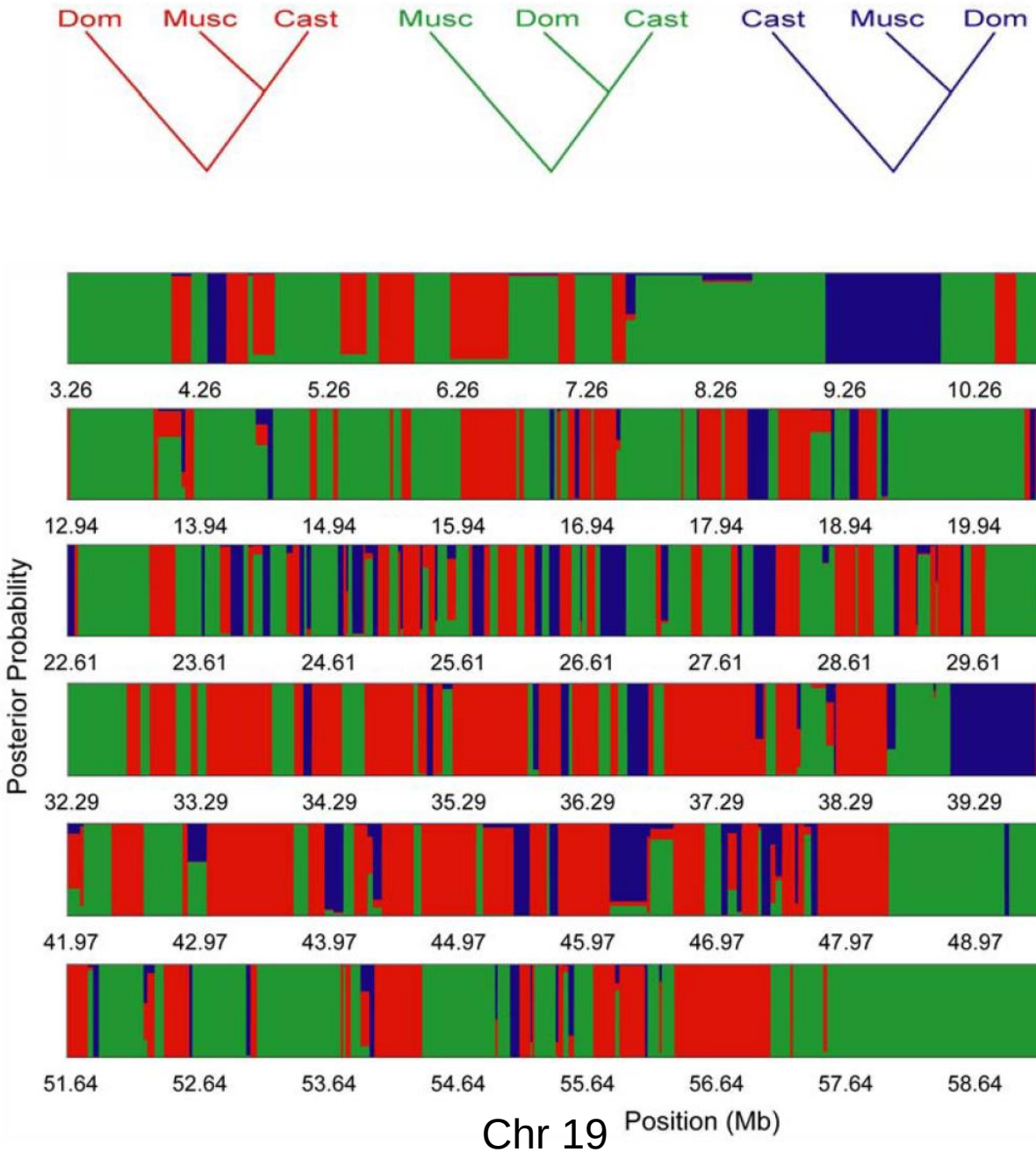
# Přetrvání ancestrálního polymorfismu

## (species tree ≠ gene tree)



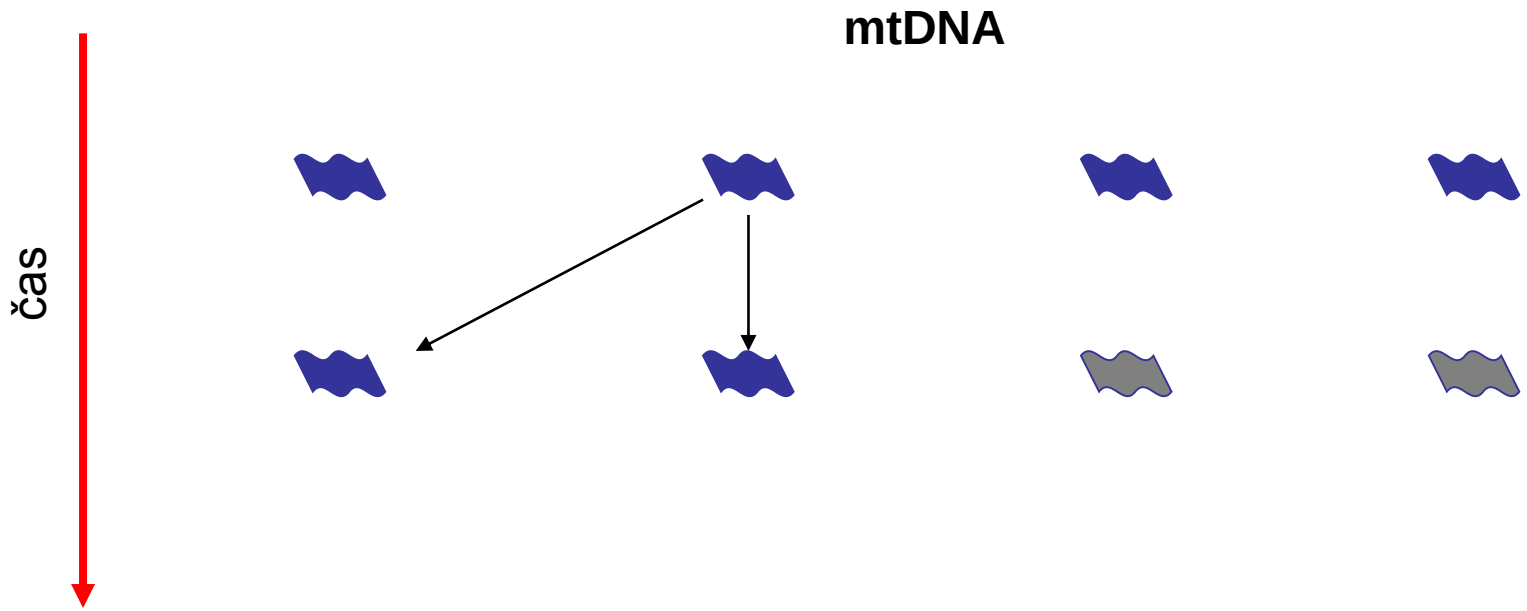




Dom
Musc
Cast
Musc
Dom
Cast
Cast
Musc
Dom
3.26
4.26
5.26
6.26
7.26
8.26
9.26
10.26
11.26
12.26
12.94
13.94
14.94
15.94
16.94
17.94
18.94
19.94
20.94
21.94
22.61
23.61
24.61
25.61
26.61
27.61
28.61
29.61
30.61
31.61
32.29
33.29
34.29
35.29
36.29
37.29
38.29
39.29
40.29
41.29
41.97
42.97
43.97
44.97
45.97
46.97
47.97
48.97
49.97
50.97
51.64
52.64
53.64
54.64
55.64
56.64
57.64
58.64
59.64
60.64
Posterior Probability
Position (Mb)

Chr 19

# Pravděpodobnost koalescence v předchozí generaci 1/N

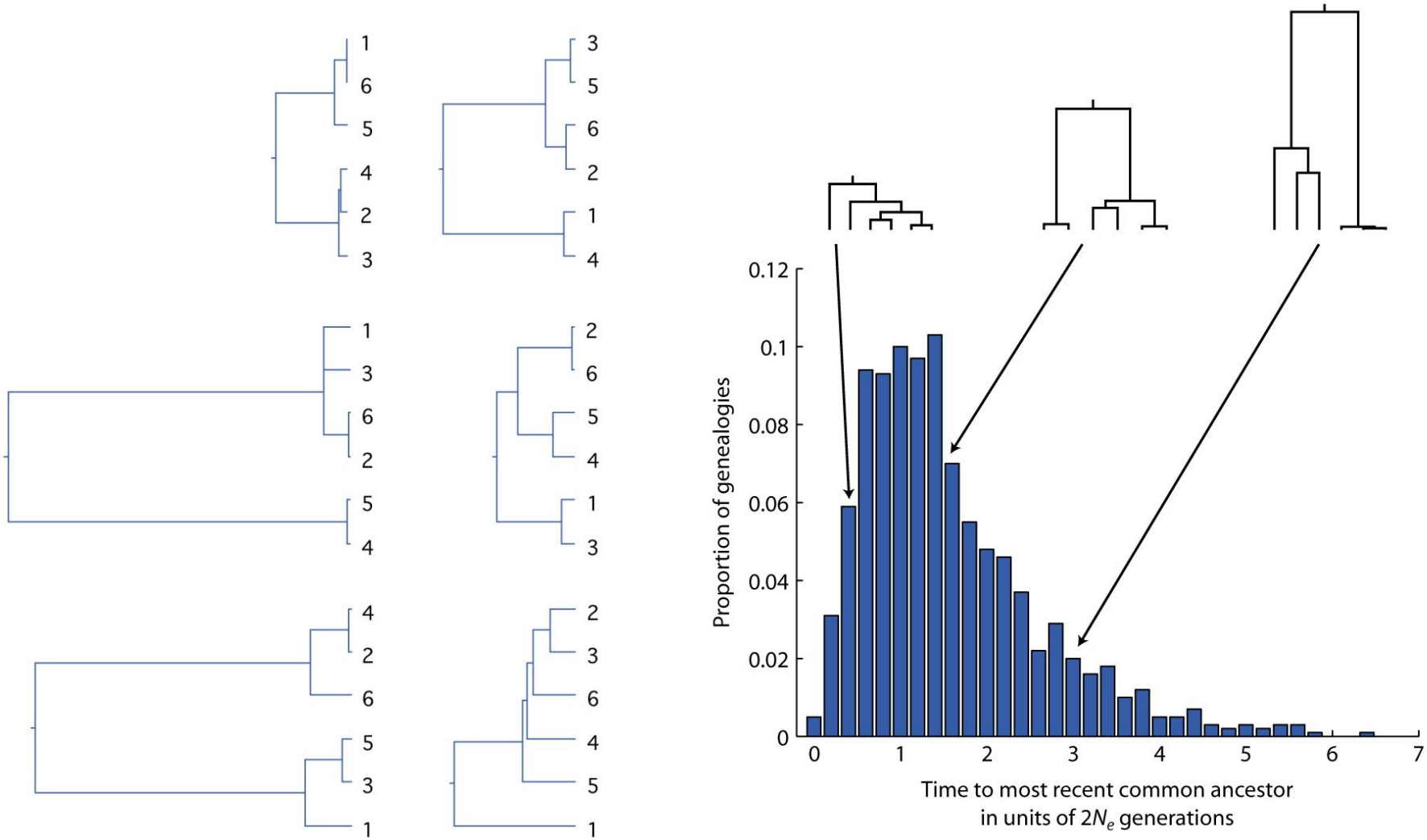



*Pozor! Často se pracuje s velikostí populace 2N jako aproximací diploidního modelu. Pravděpodobnost pak bude 1/2N*

Průměrný čas koalescence páru alel N (2N) generací
Rozptyl (variance) $(2N)^2$, s.d. 2N

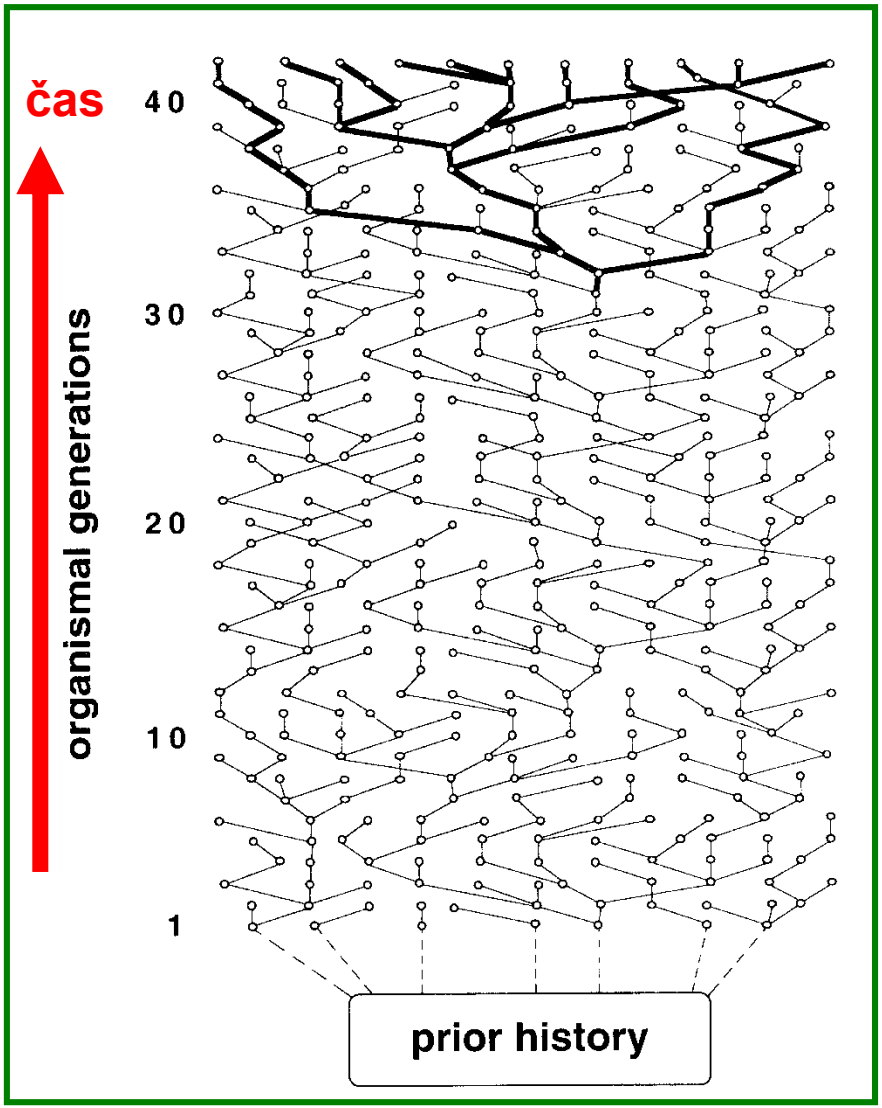
Koalescence všech alel 4N generací, opět velký rozptyl!

Vzorek n alel $\rightarrow$ 4N(1-1/n)




Waiting time – čas ke splynutí linií (coalescent event)

# Koalescence




## Lineage sorting

přežije jen jedna linie

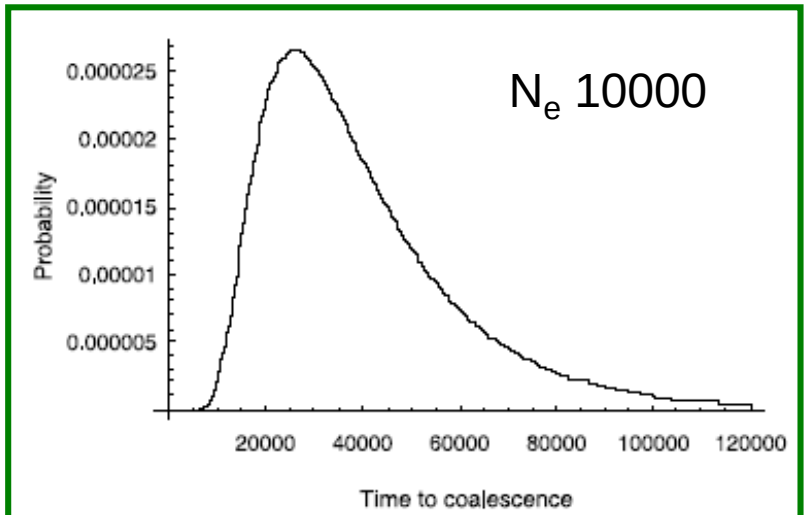
sekvence však nejsou identické díky mutacím!

## Koalescence

„splynutí“ současných linií v MRCA

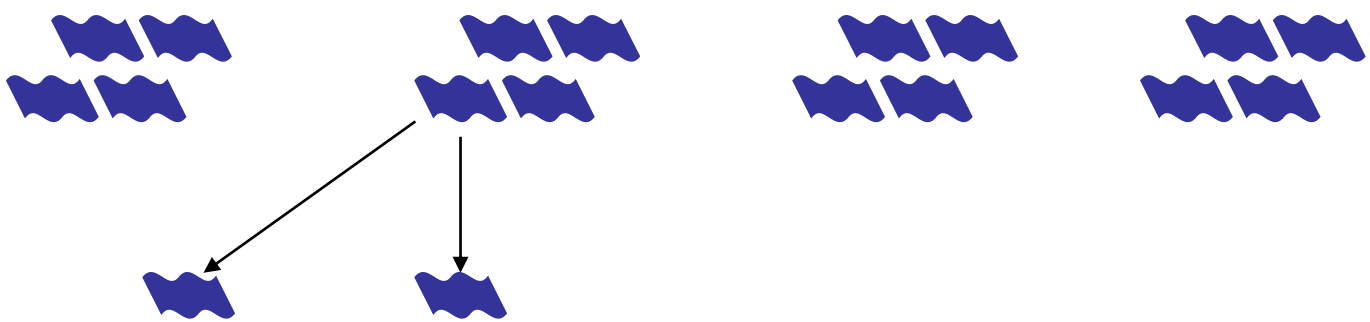
Most Recent Common Ancestor

## Čas k MCRA ~$4N_e$ generací



# N ($N_e$) může být různé pro různé geny

**mtDNA**



**jaderný lokus**

# mtDNA, Y (cpDNA) versus jaderné geny

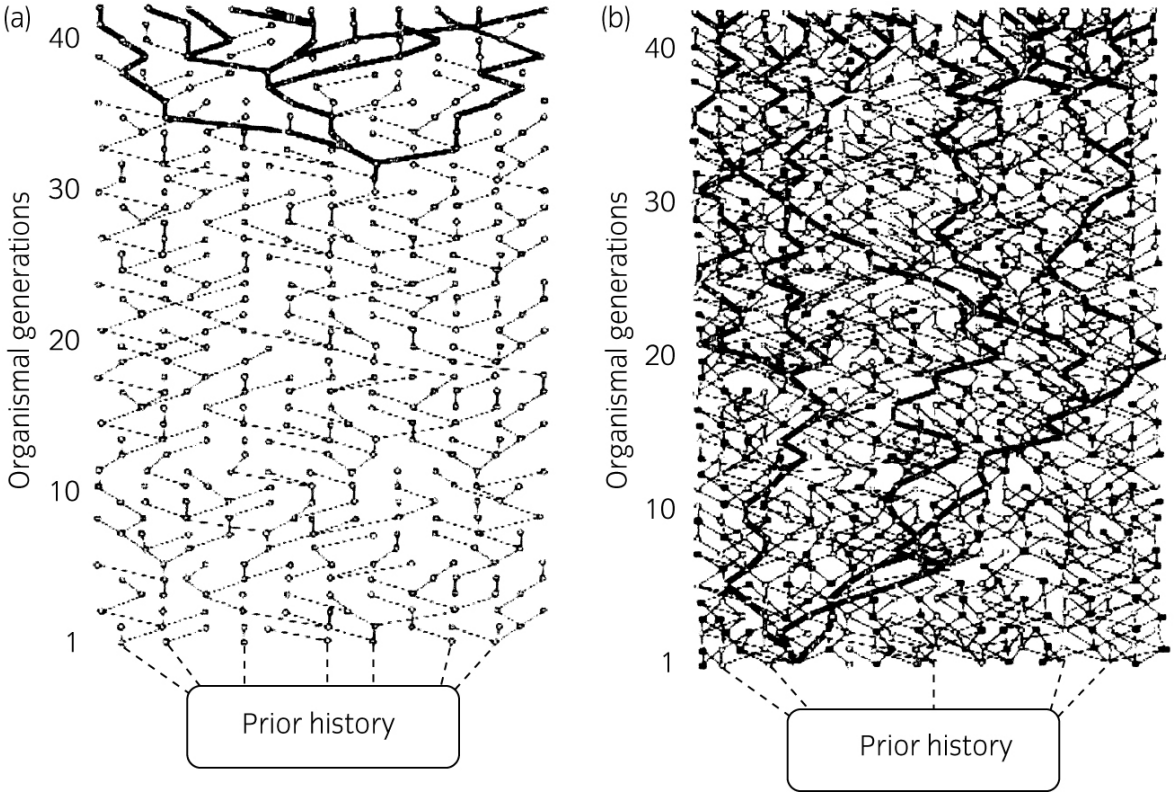
Jaderné geny:

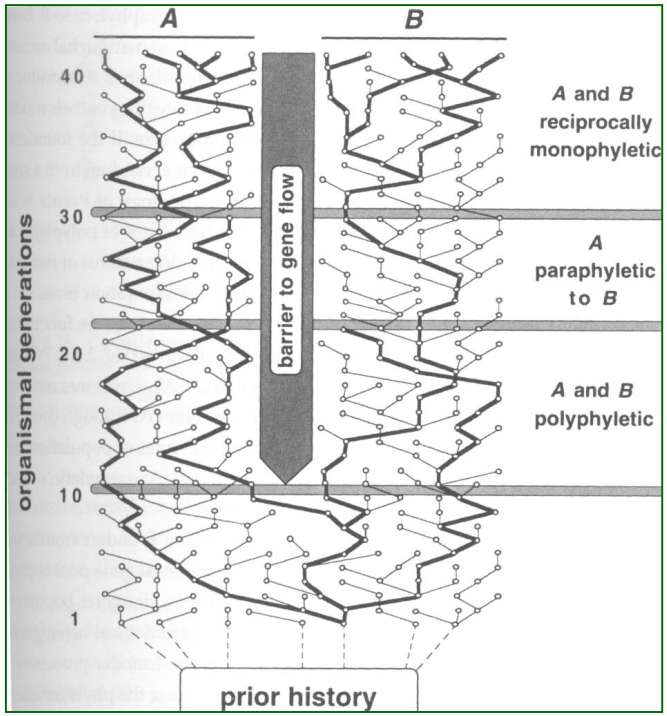
složitější osud, větší $N_e$ (4x)

→ delší doba pro lineage sorting

→ jiné výsledky u nedávno oddělených populací

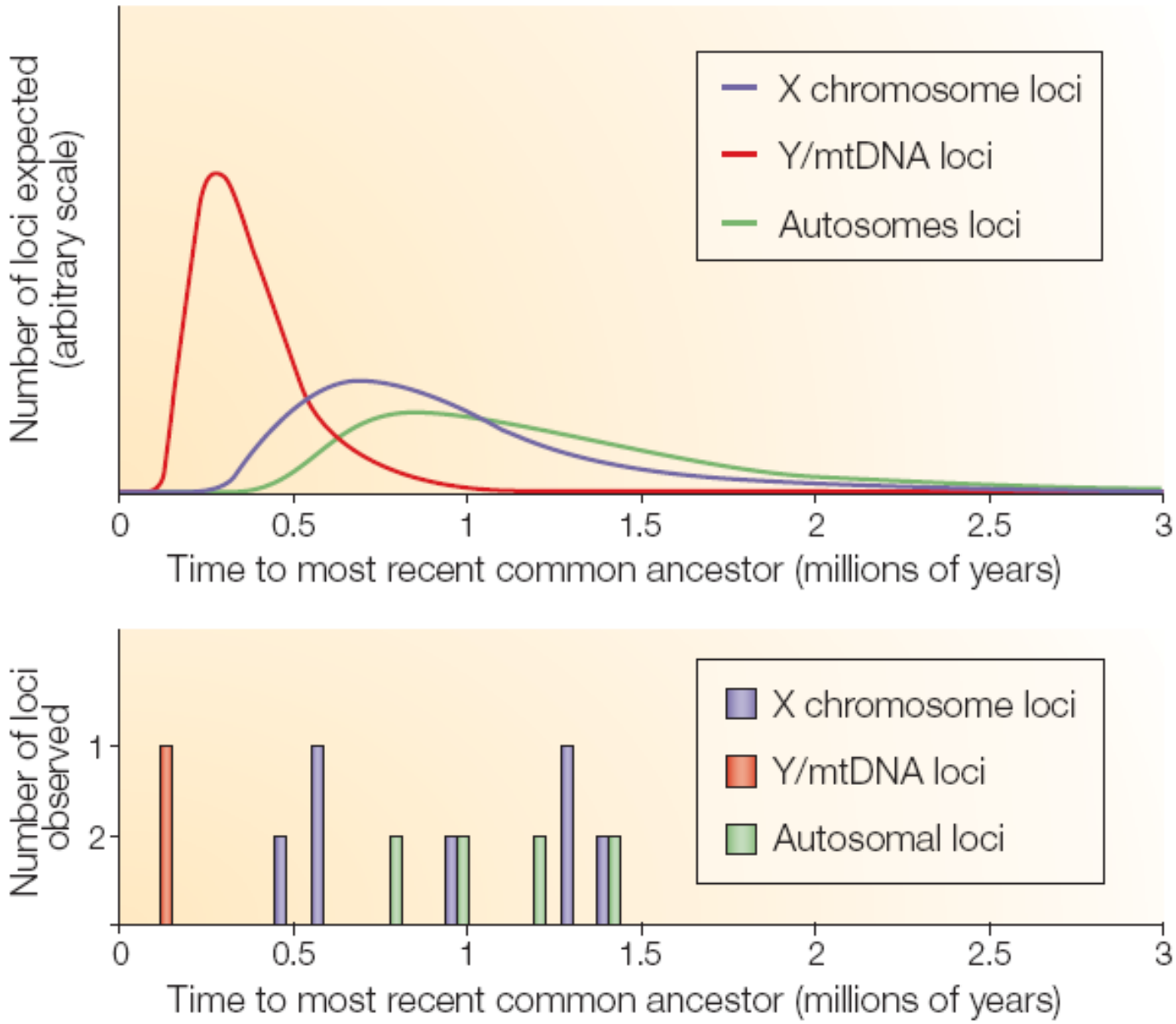
mtDNA dospěje rychleji do reciproční monofylie

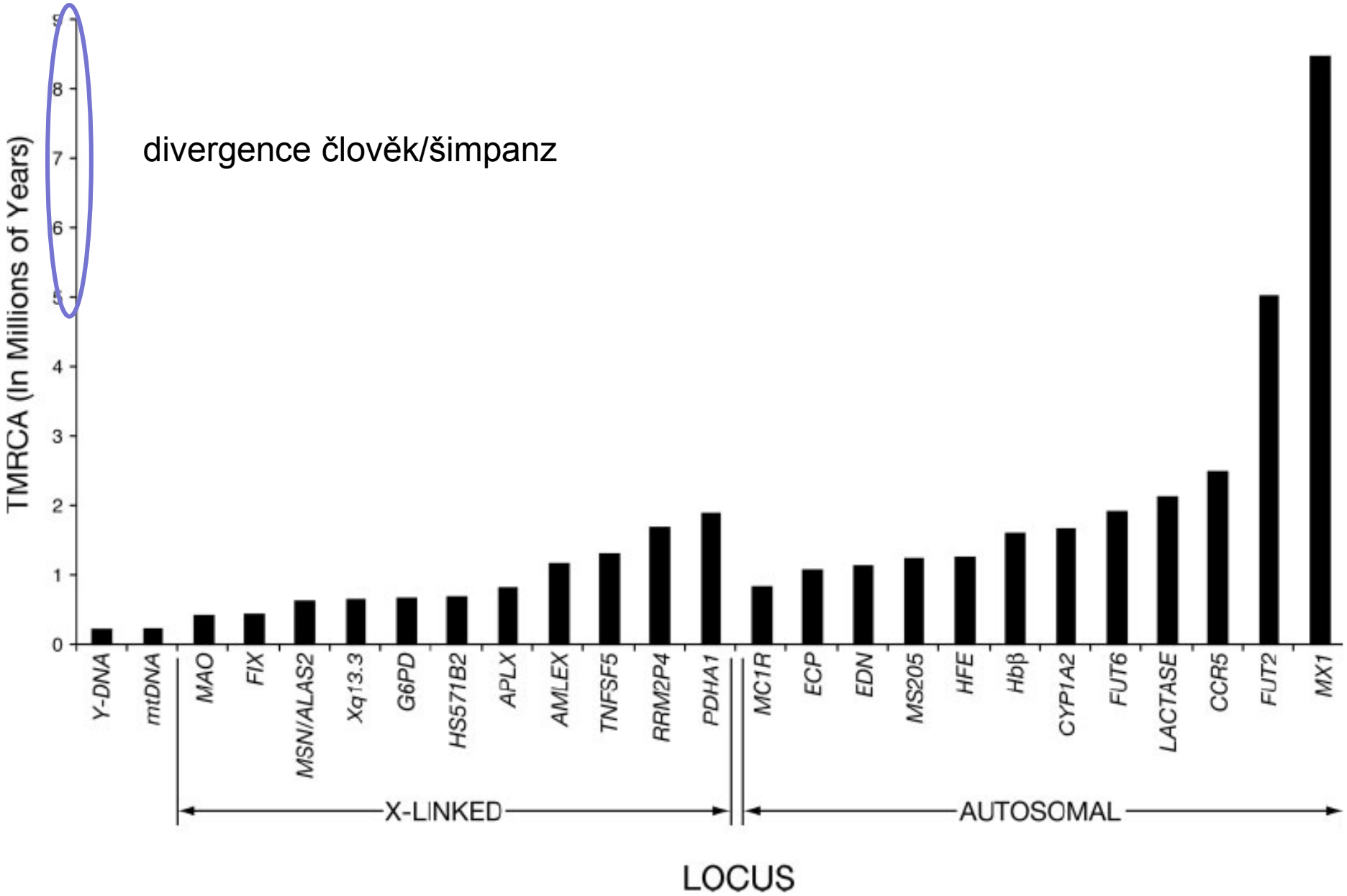


# Y a $N_e$

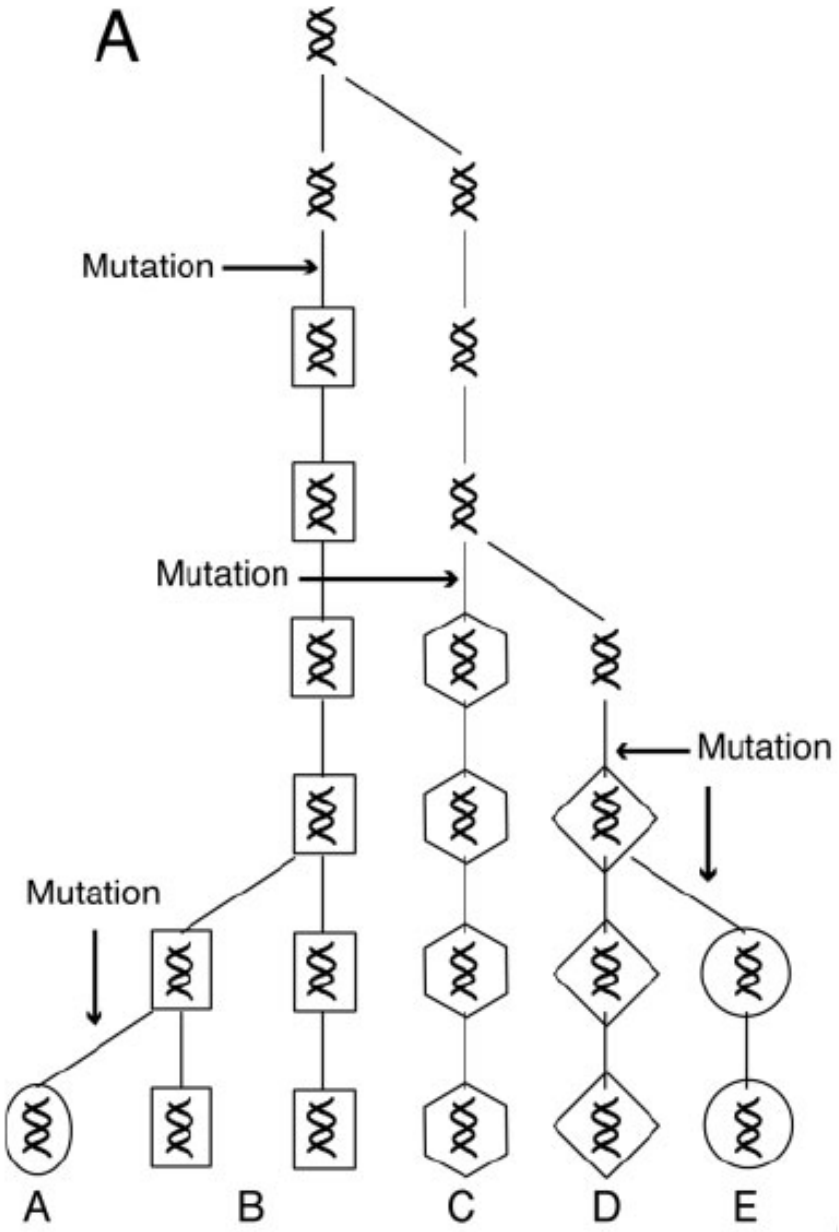
- $N_e$ je čtvrtinová vzhledem k autosomům, srovnatelná s mtDNA
- Větší náchylnost k driftu
- Rozdíl ve varianci reprodukčního úspěchu mužů (samců) a žen (samic) → nižší $N_e$, nedávné koalescence


Number of loci expected (arbitrary scale)
X chromosome loci
Y/mtDNA loci
Autosomes loci
0
0.5
1
1.5
2
2.5
3
Time to most recent common ancestor (millions of years)
Number of loci observed
1
2
X chromosome loci
Y/mtDNA loci
Autosomal loci
0
0.5
1
1.5
2
2.5
3
Time to most recent common ancestor (millions of years)


divergence člověk/šimpanz
TMRCA (In Millions of Years)
0
1
2
3
4
5
6
7
8
9
Y-DNA
mtDNA
MAO
FIX
MSN/ALAS2
Xq13.3
G6PD
HS571B2
APLX
AMLEX
TNFSF5
RRM2P4
PDHA1
MC1R
ECP
EDN
MS205
HFE
Hbβ
CYP1A2
FUT6
LACTASE
CCR5
FUT2
MX1
X-LINKED
AUTOSOMAL
LOCUS

# mutace

- Gene copies (ač mají společného předka) se liší mutacemi
- $\Theta$ scaled mutation rate, expected number of mutations separating gene copies
- $\Theta = 4N\mu$ *μ mutační rychlost*
- Závislost na *N*
- Malé populace – méně variability (kratší čas koalescence, tedy k MRCA → málo času na mutace)



Lze odhadnout frekvenční spektrum mutací!
Neutrální model

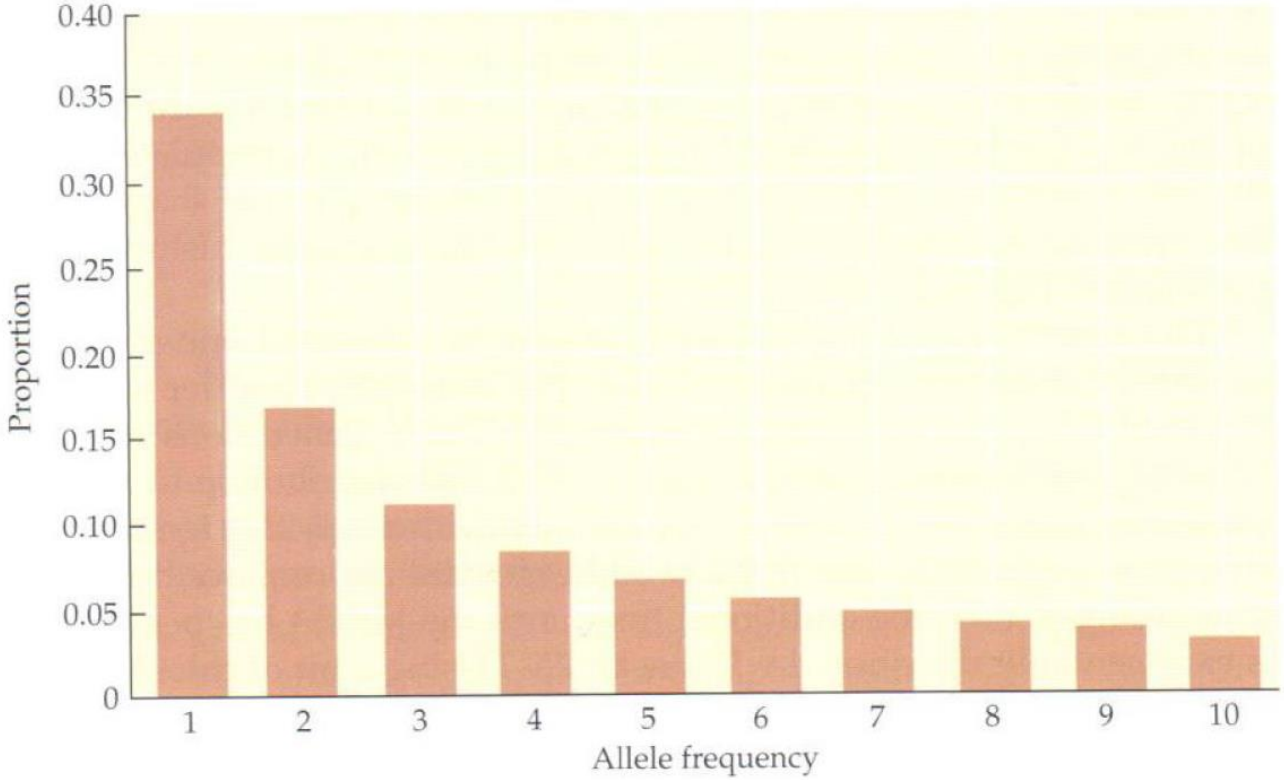
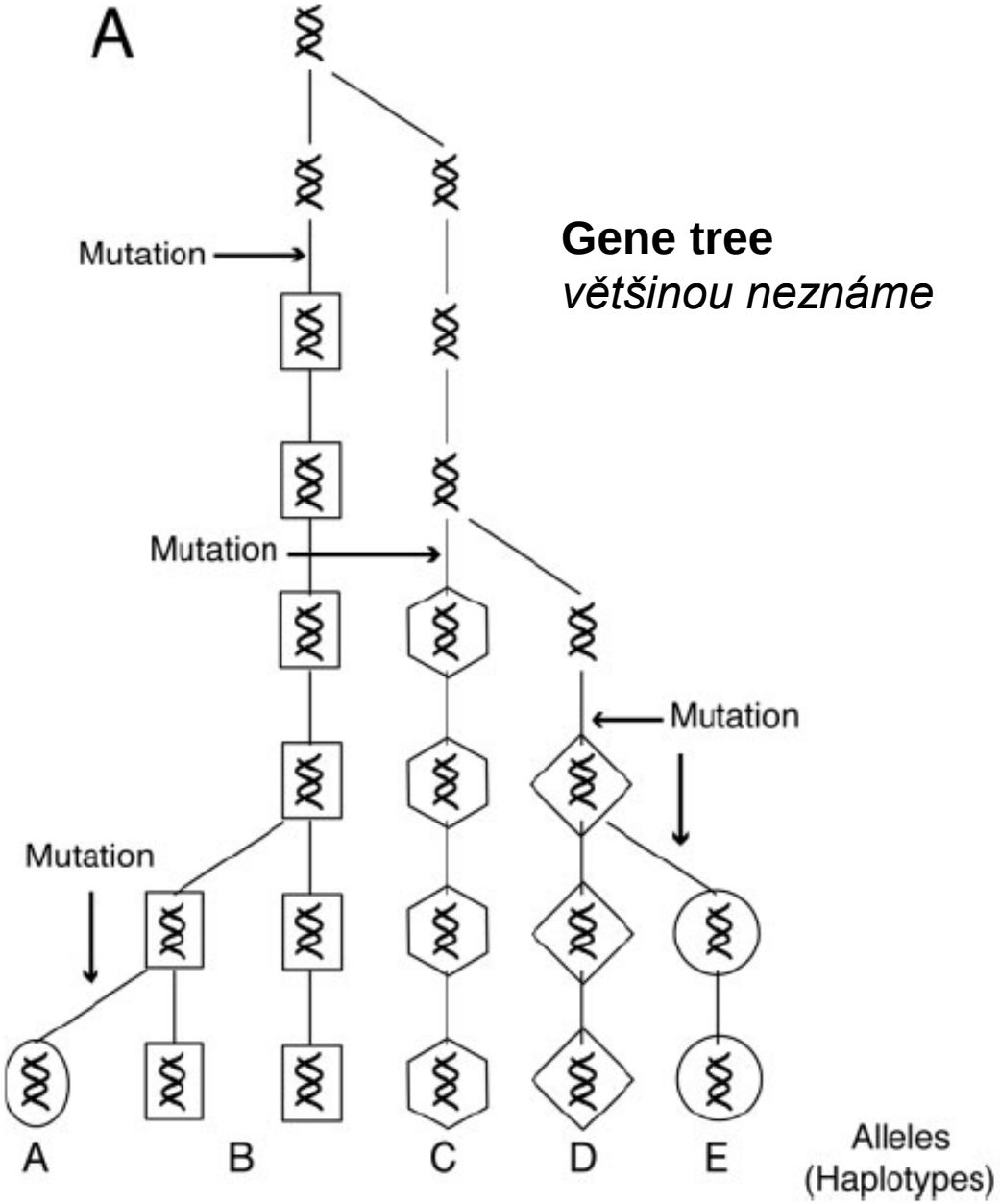
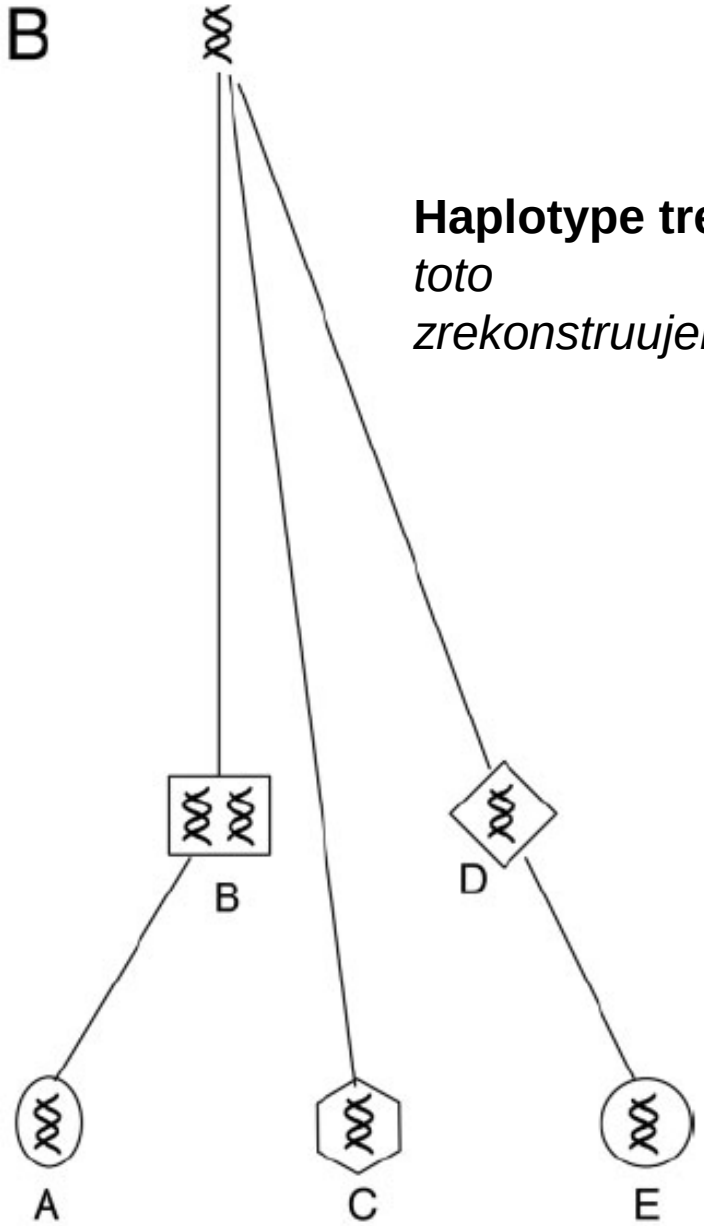



**Figure 3.9** The expected site frequency spectrum (SFS) for a sample of $n = 10$ haploid individuals under the standard neutral coalescence model with infinite sites mutation.

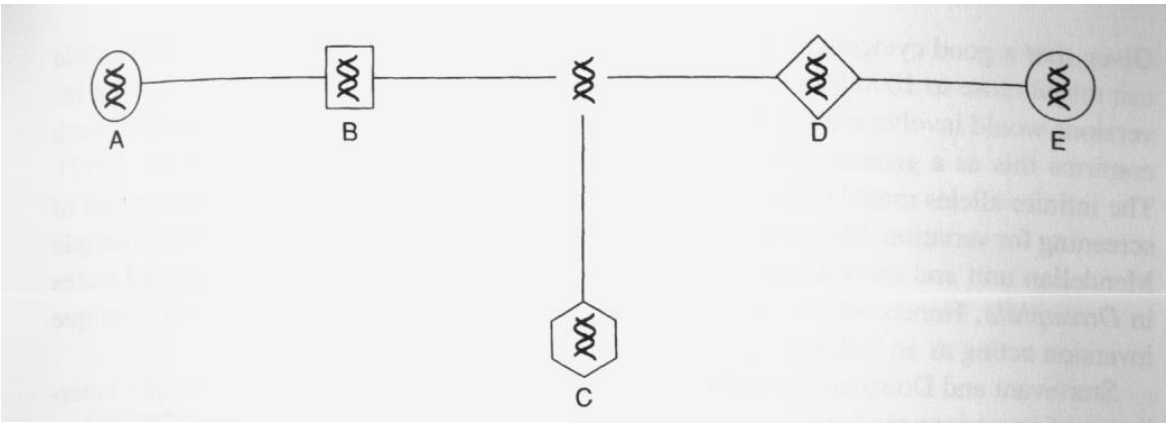


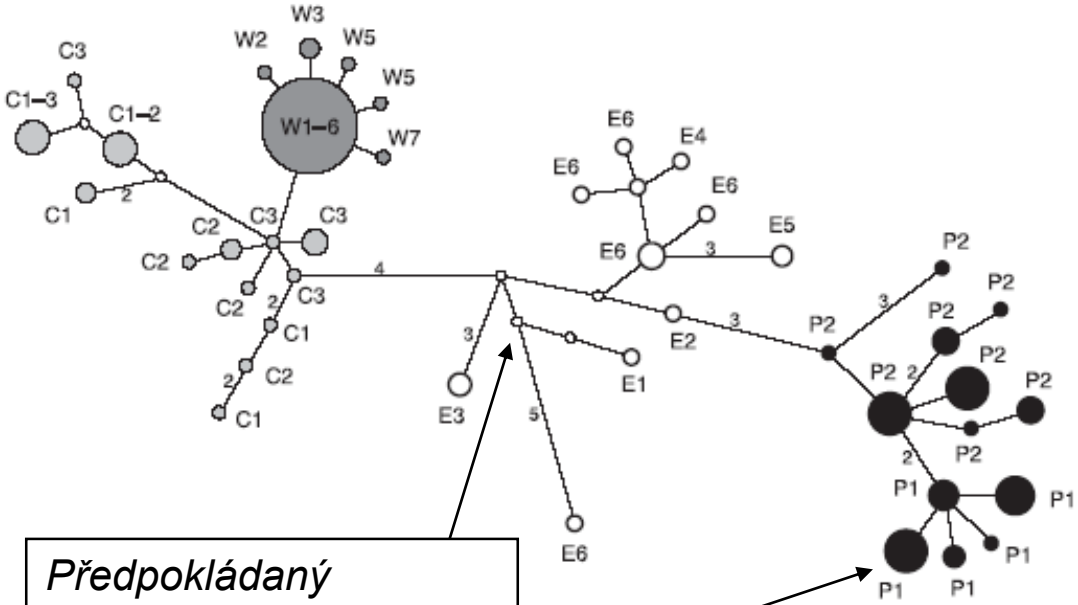

**Gene tree**
*většinou neznáme*




**Haplotype tree**
*toto zrekonstruujeme*

# Nezakořeněné stromy,
# Haplotype network, síť haplotypů (často nevíme, co bylo dřív)

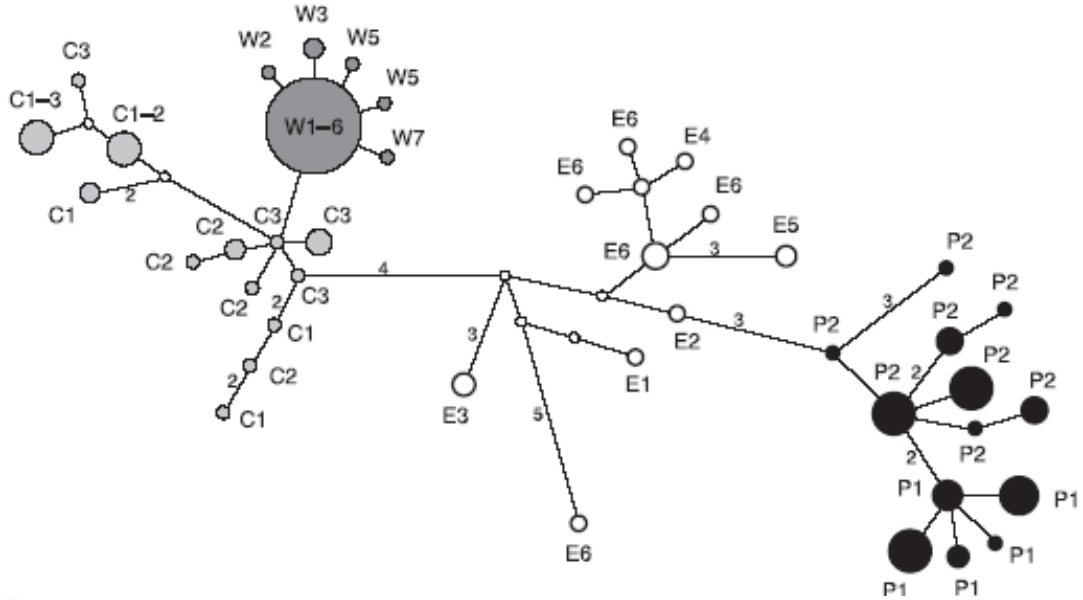


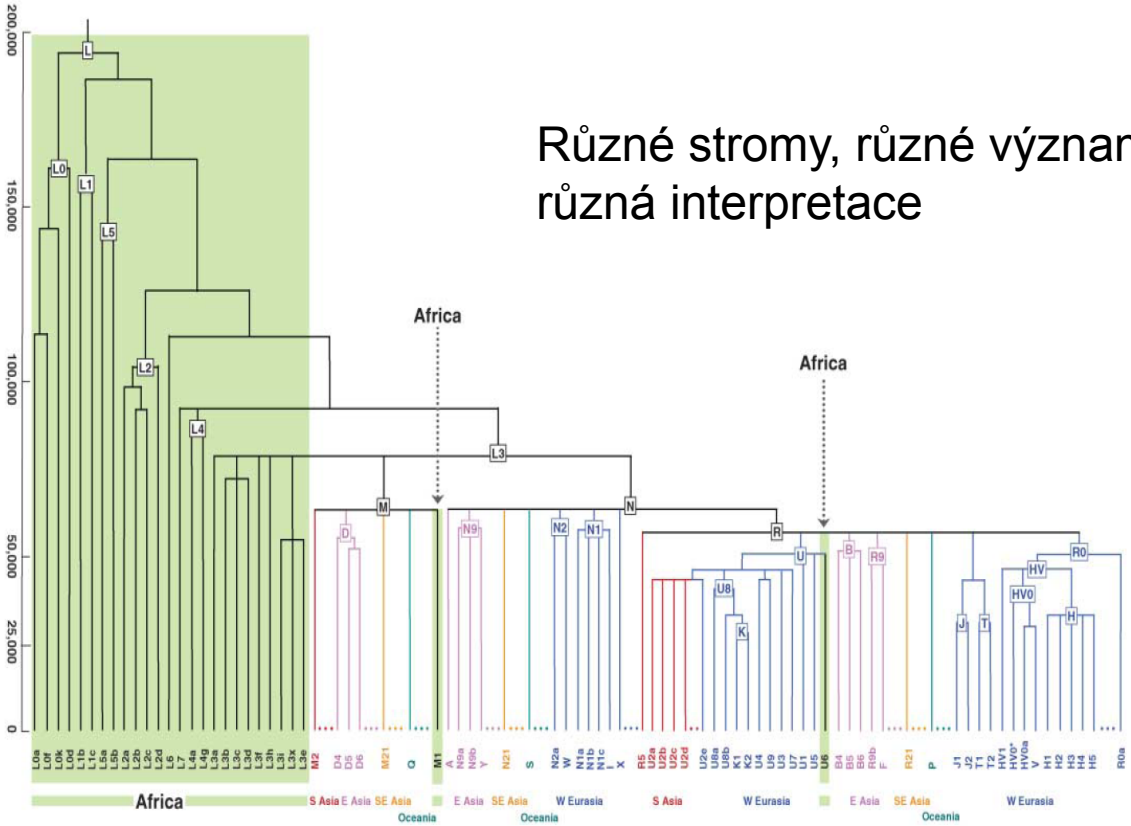




*Toto jsou sice stromy, ale bežně se jim říká sítě.*

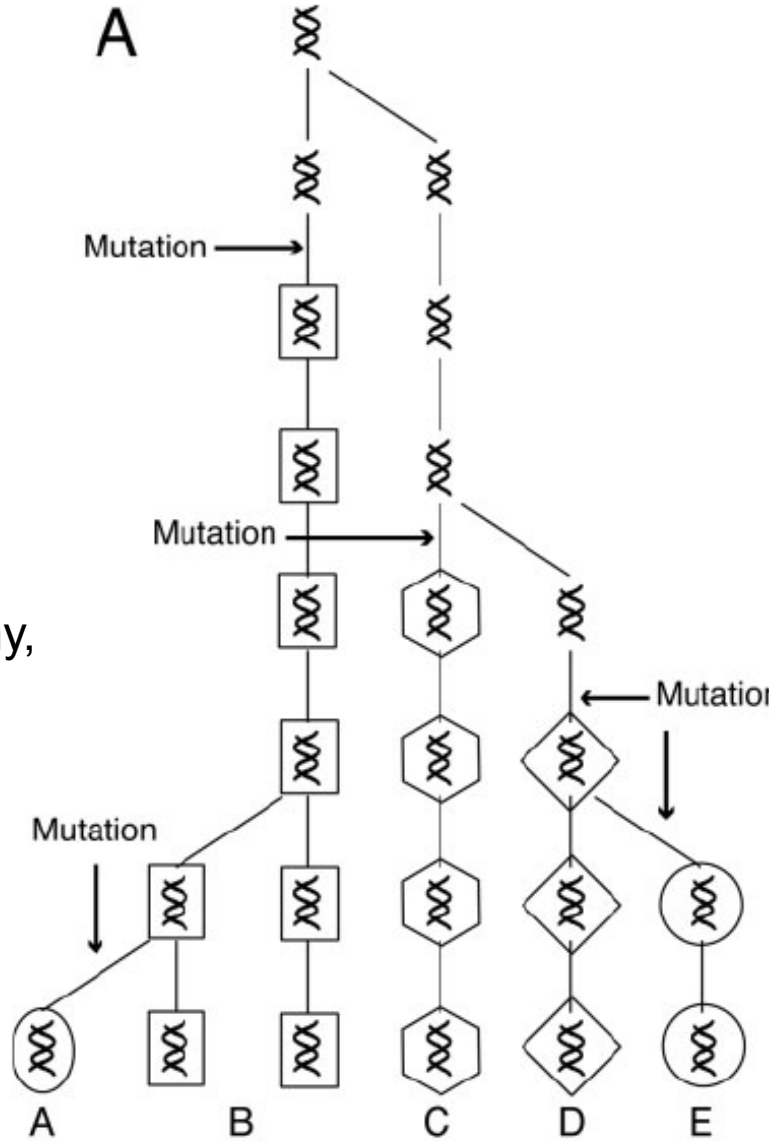
*Předpokládaný nikoliv však pozorovaný haplotyp*

*Konkrétní zjištěný haplotyp*
*Velikost kroužku = počet jedinců s tímto haplotypem*



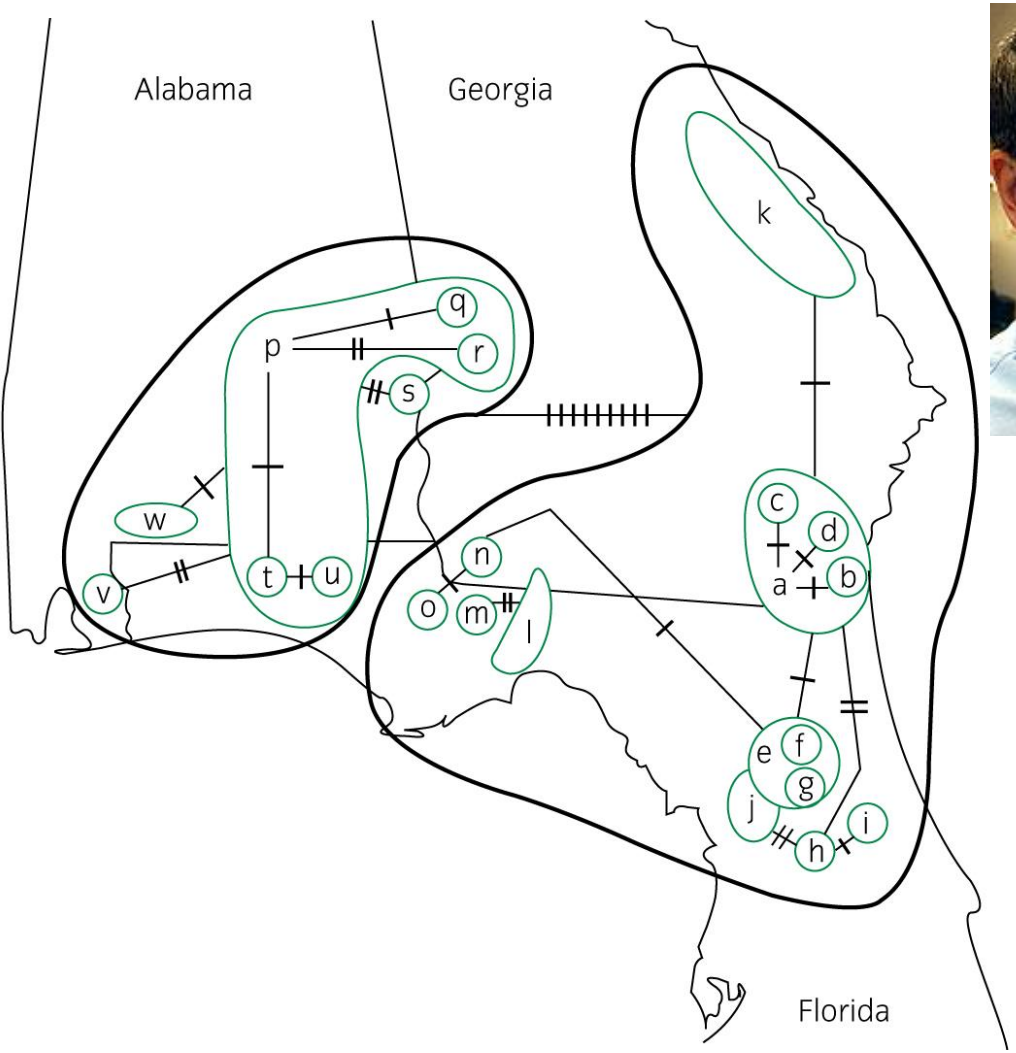

Různé stromy, různé významy, různá interpretace

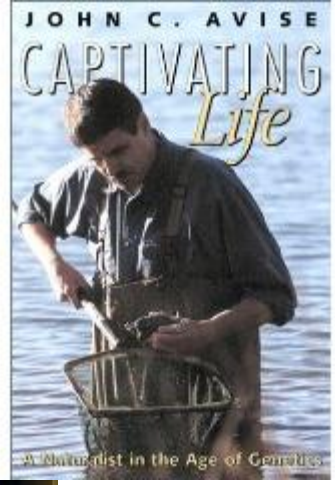


# Phylogeography

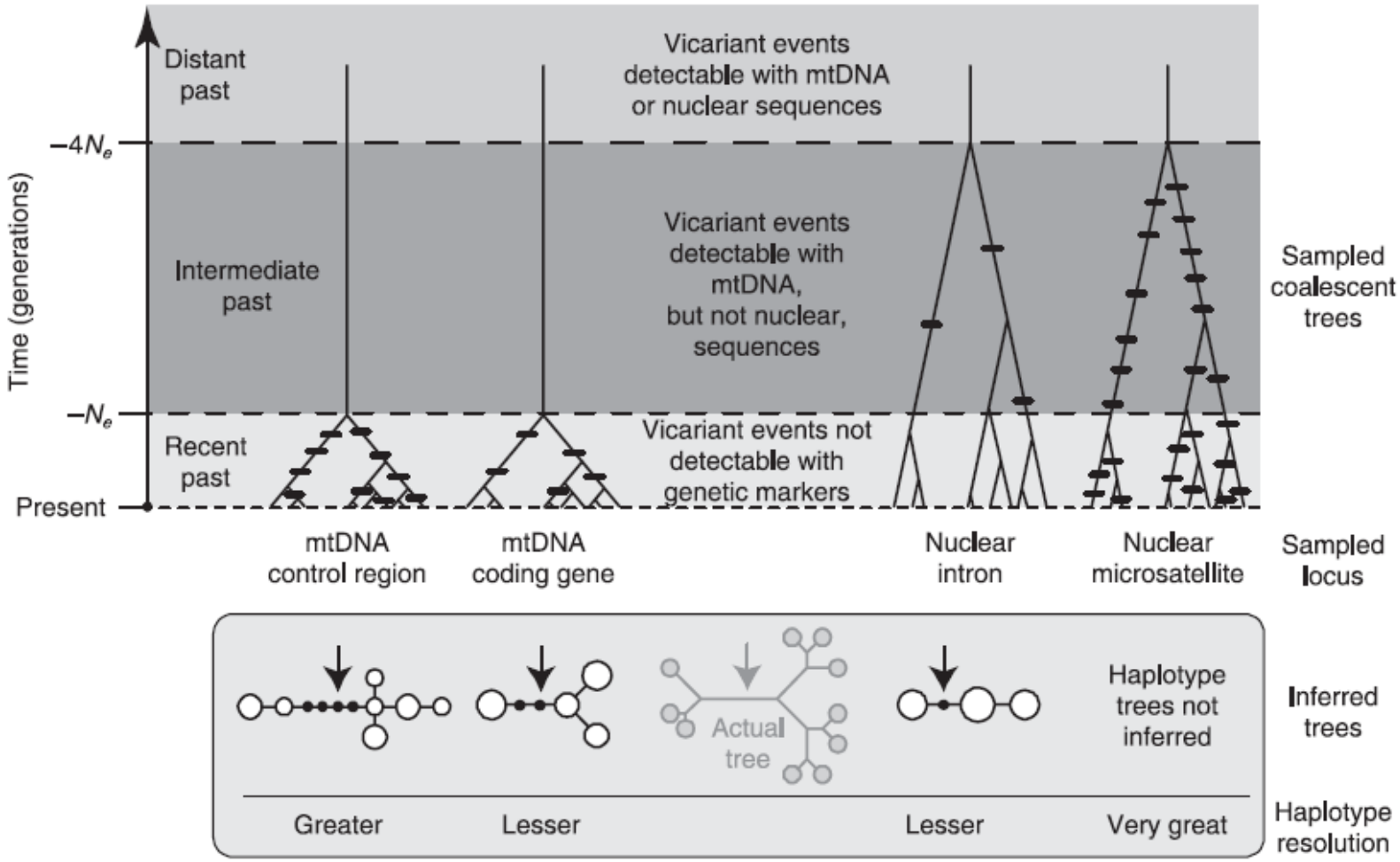
*Geomys pinetis (pytlonoš)* Avise et al. 1979

maximum parsimony network

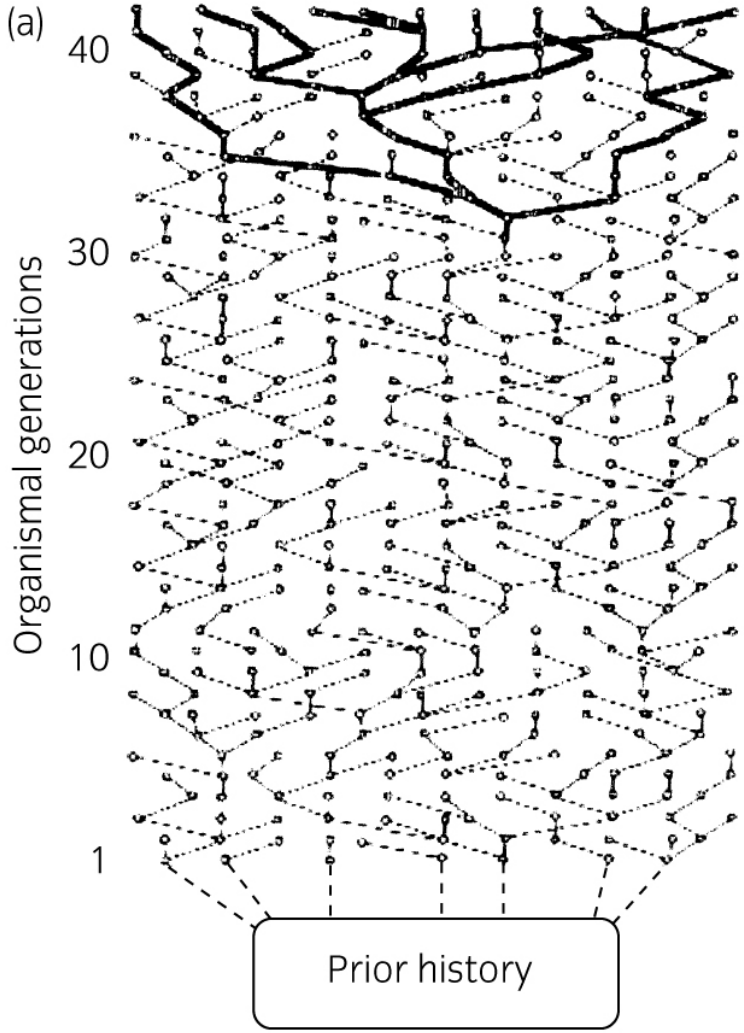


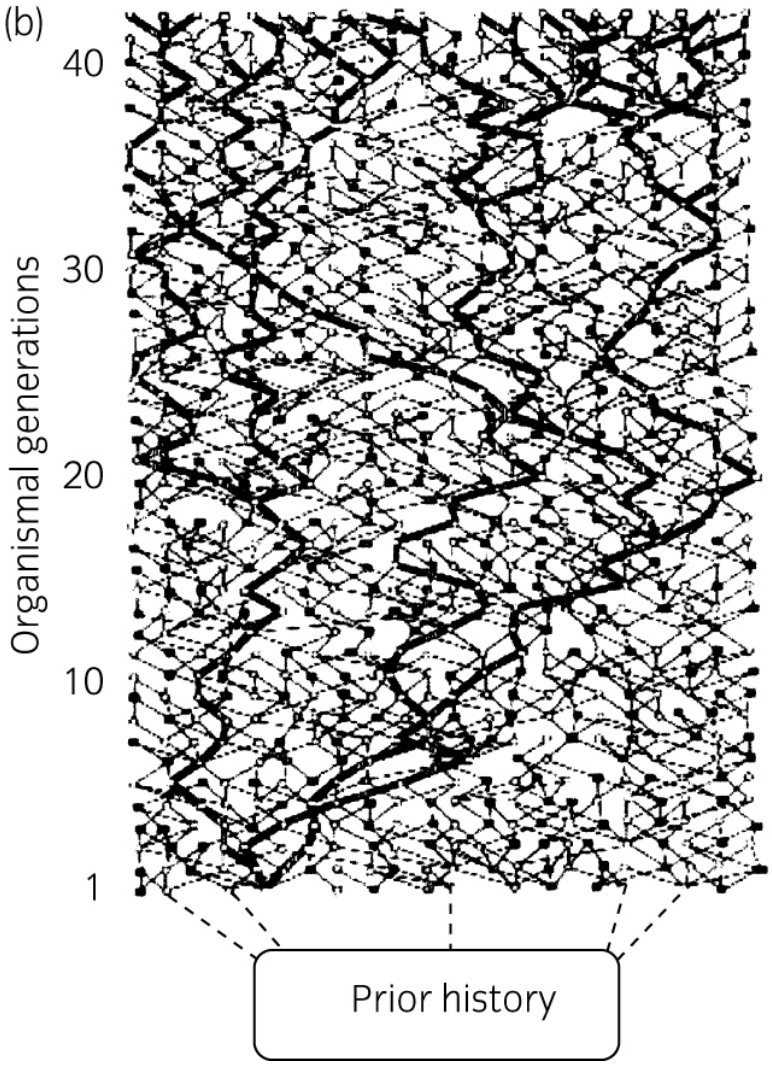


Pro výsledné haplotypové stromy (sítě) je zásadní i mutační rychlost



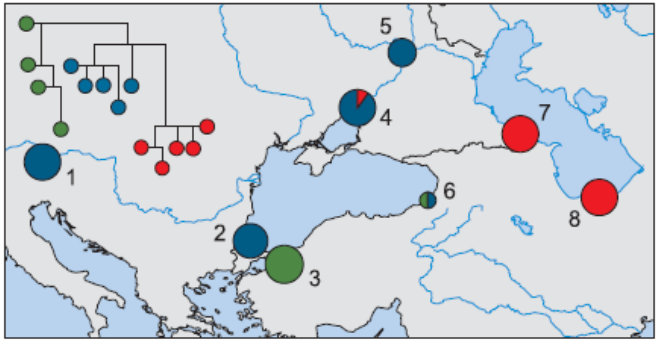
Jaderné lokusy – složitá a delší historie, často ale méně mutací



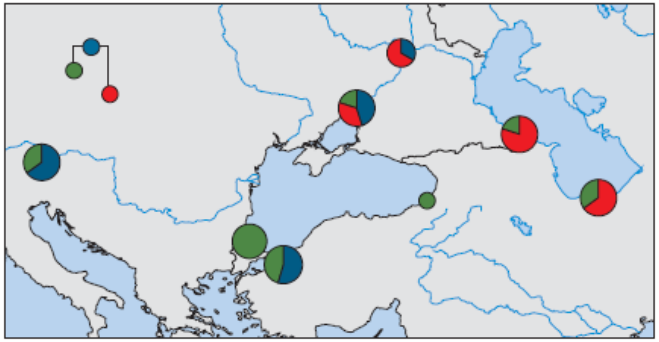



*Rutilus frisii* Plotice perleťová

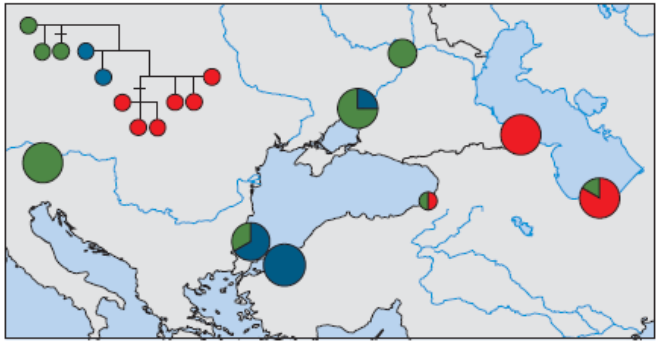
mitochondrial DNA cytochrome *b*




ATP synthase beta-subunit

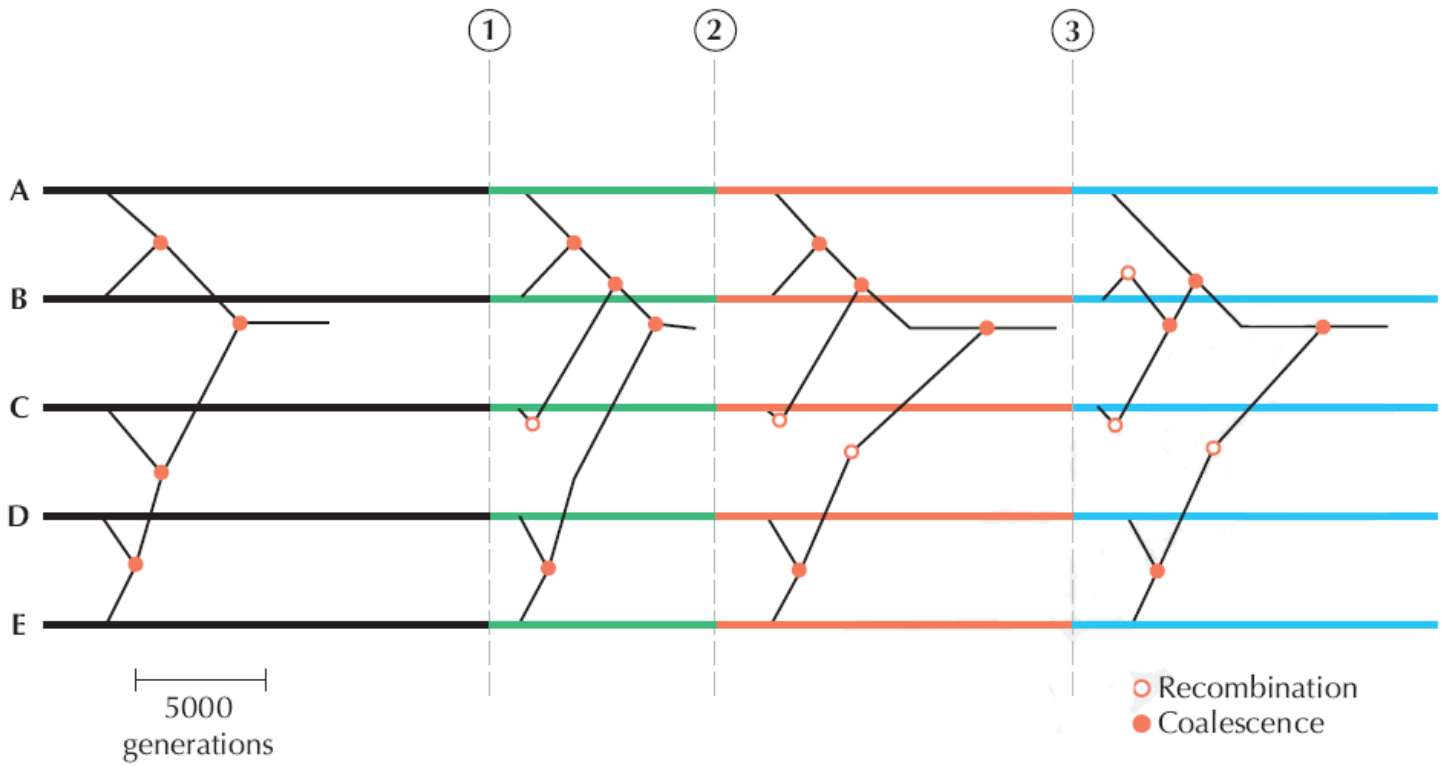


S7 ribosomal protein



# Rekombinace mění genealogie podél chromosomu

- Vystopovat společného předka pro autosomální lokusy je obtížné, protože rekombinují. Rekombinace mění genealogie.
- Na autosomech mají geny, které nejsou ve vazbě různé genealogie. Čím blíže jsou u sebe na chromosomu, tím mají podobnější genealogii.
- Jak rychle se genealogie podél genomu mění záleží na míře vazebné nerovnováhy (u druhů s velkou Ne a vysokou mírou rekombinace mohou být nerekobinující oblasti velmi krátké – několik desítek až stovek bp, tj. méně než délka genu!)



# K odhadu MRCA stačí malý vzorek

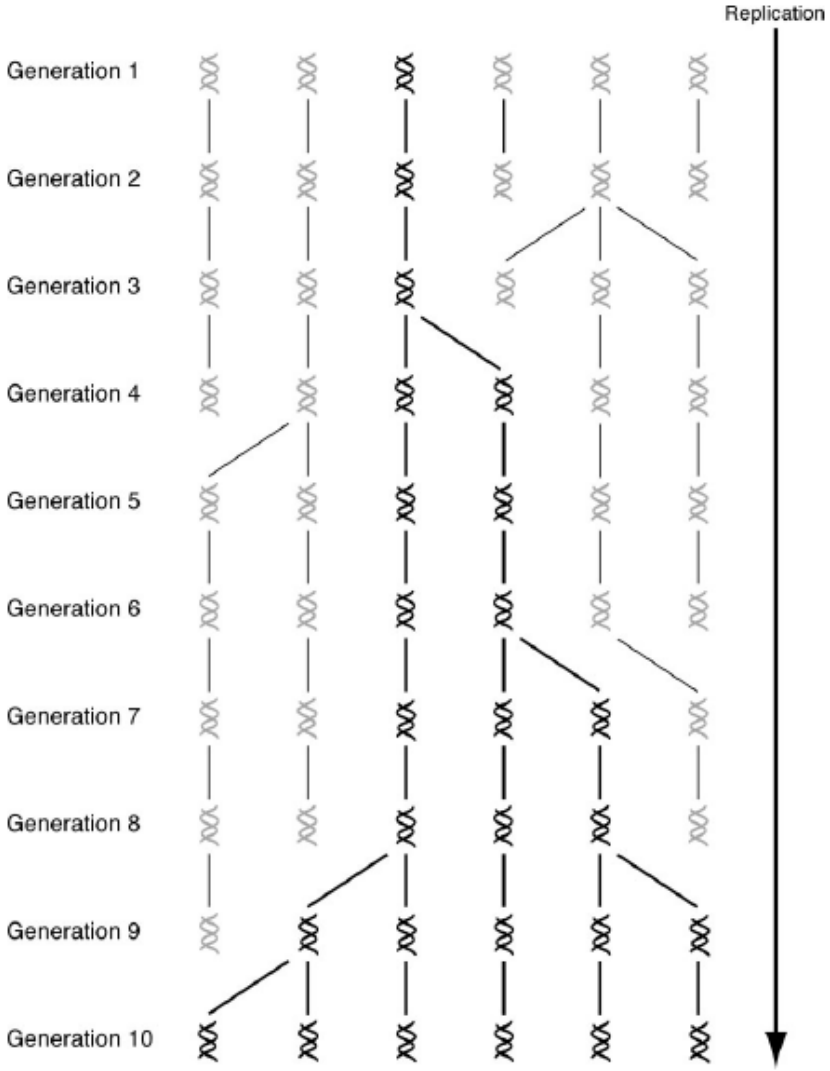
Pravděpodobnost zachycení
MRCA celé populace
vzorkem o *n* jedincích:

$(n - 1)/(n + 1)$

$n = 3 \rightarrow 0{,}5$

$n = 20 \rightarrow > 0{,}9$

Versus
Klasická populační (frekventistická) genetika
Odhad frekvence → velký vzorek



# Koalescence

## Náhodné genealogické stromy

Takto můžou vypadat genealogie vzniklé náhodným slučováním linií.

- Různé lokusy představují nezávislé replikáty společné historie. Čím víc nezávislých lokusů studujeme, tím lepší odhady o historii populace dostaneme.

# „Coalescent“
# Koalescenční teorie

- Interpretace času ke společnému předku alel pomocí evolučních procesů

- Procesy: drift, migrace, selekce, změna populační velikosti

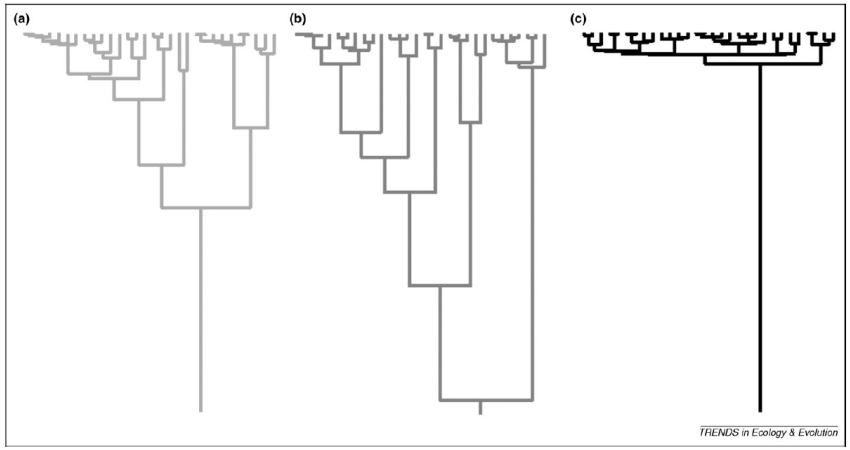



**Figure 2.** Growth signature in genealogies. Genealogies sampled from (a) constant-size, (b) exponentially shrinking and (c) exponentially growing populations.

# K čemu to teda vlastně je dobré?

- Odhad času k MRCA
- Pozorované odchylky (hodnoty parametrů)
  → biologicky zajímavé děje (demografie, selekce, migrace)
- Konkrétní štěpení stromů nás vlastně ani nezajímá
  *(rozdíl oproti fylogenetickým metodám)*

- **Detekce**:
  - Historie šíření druhů, migrace
  - Demografických událostí (expanze, bottlenecky)
  - Selekce
- **Problém**:
  - Selekce i demografické události → podobný signál

# Rozdíl mezi fylogenetickým a populačně genetickým (koalescenčním) přístupem

Data, **strom**, mutační model

Basic statistics makes the distinction between phylogenetic and coalescent approaches apparent. The fundamental equation for likelihood inference in phylogenetics is[94]

$$L = P(D|G, \mu), \qquad (1)$$

where $L$ is the likelihood (the probability of the data, given the parameters), $D$ is the data (typically DNA sequences), $G$ is the tree and $\mu$ is the collection of parameters in the mutation model. The objective of the analysis is to estimate the parameter $G$. = strom

The analogous equation in the coalescent setting is[22,42,94]

$$L = \sum_{G} P(D|G, \mu) P(G, \alpha), \qquad (2)$$

where $\alpha$ is the collection of parameters (such as population sizes and migration rates) for the population process. The objective of the analysis is typically to estimate these parameters. The tree or genealogy, $G$, is a so-called nuisance parameter, which we remove by averaging the likelihood over all possible values.

Data, strom, mutační model, **parametry**

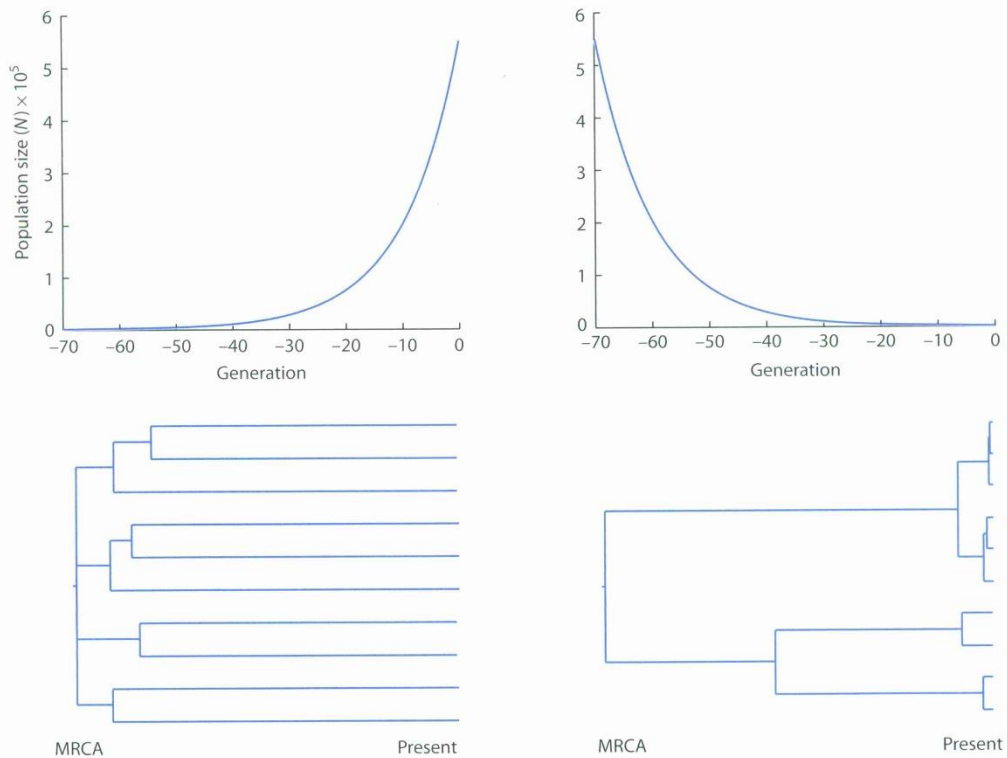



**Figure 3.29** The effects of exponential population growth or shrinkage on coalescent genealogies. The upper panels show change in population size over time with exponential growth according to $N(t) = N_0 e^{-rt}$ with $r = \pm 0.1$, yielding relatively slow exponential population growth. The two genealogies illustrate examples of waiting times that might be seen under strong exponential population growth (left) and shrinkage (right). With strong exponential population growth coalescent times are longest in the present when the population is the largest, leading to genealogies characterized by long branches near the present and very short branches in the past around the time of the MCRA. With exponential population shrinkage, coalescence times are greatest in the past near the MRCA when the population was larger and shortest near the present when the population is at its smallest size. The genealogy on the left was obtained using equation 3.91 with $r = 100$.

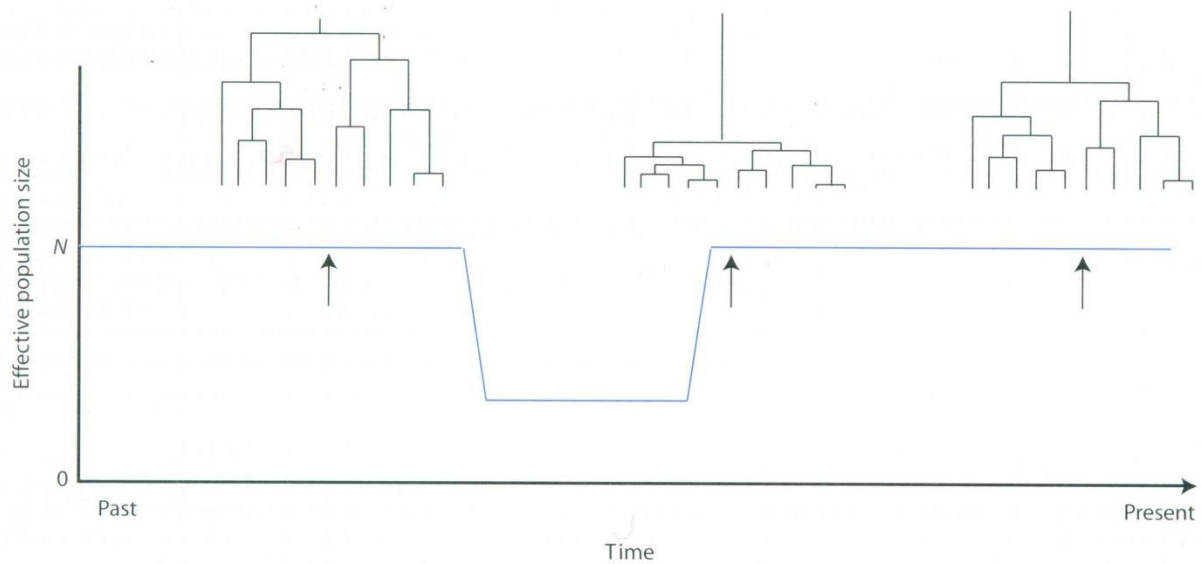



**Figure 3.28** The effects of a population bottleneck on gene genealogies. During the bottleneck the chance that two randomly sampled gene copies are derived from one copy in the previous generation $\left(\frac{1}{2N_e}\right)$ increases. This can also be thought of as a reduction in the overall height of a genealogical tree caused by the bottleneck since lineages that find their ancestors during the bottleneck lead to short branches. The overall effect of a bottleneck on coalescence among gene copies sampled in the present depends on the reduction in the effective population size and the duration. The arrows indicate the point in time when gene copies were sampled from the population.

# Mutace

- Mutace se akumulují nezávisle v jednotlivých liniích
- *μ* mutační rychlost
- *Θ (scaled mutation rate)* $4N\mu$
- Různé délky větví stromu → různé počty mutací
- *π* nucleotide site diversity
- $\mu << 1 \rightarrow \pi \approx \Theta$

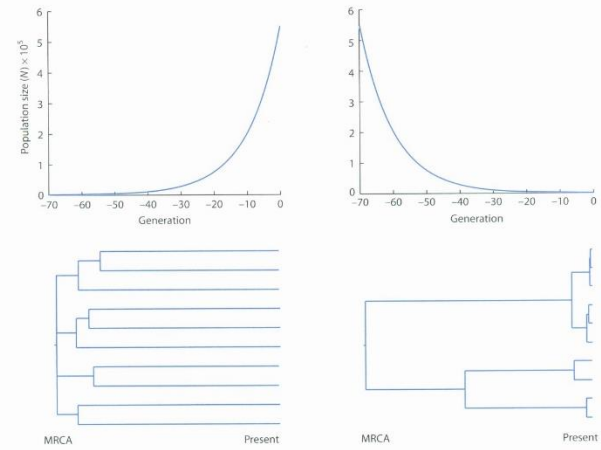



**Figure 3.29** The effects of exponential population growth or shrinkage on coalescent genealogies. The upper panels show change in population size over time with exponential growth according to $N(t) = N_0 e^{-rt}$ with $r = \pm 0.1$, yielding relatively slow exponential population growth. The two genealogies illustrate examples of waiting times that might be seen under strong exponential population growth (left) and shrinkage (right). With strong exponential population growth coalescent times are longest in the present when the population is the largest, leading to genealogies characterized by long branches near the present and very short branches in the past around the time of the MCRA. With exponential population shrinkage, coalescence times are greatest in the past near the MRCA when the population was larger and shortest near the present when the population is at its smallest size. The genealogy on the left was obtained using equation 3.91 with $r = 100$.

- *π* umíme spočítat z dat
- Pokud známe *μ*, dokážeme odhadnout *N a* dokonce odhadnout třeba celkovou délku větví či hloubku jednotlivých koalescenčních událostí

- Bayesian Skyline Plot
- Změny velikostí populací

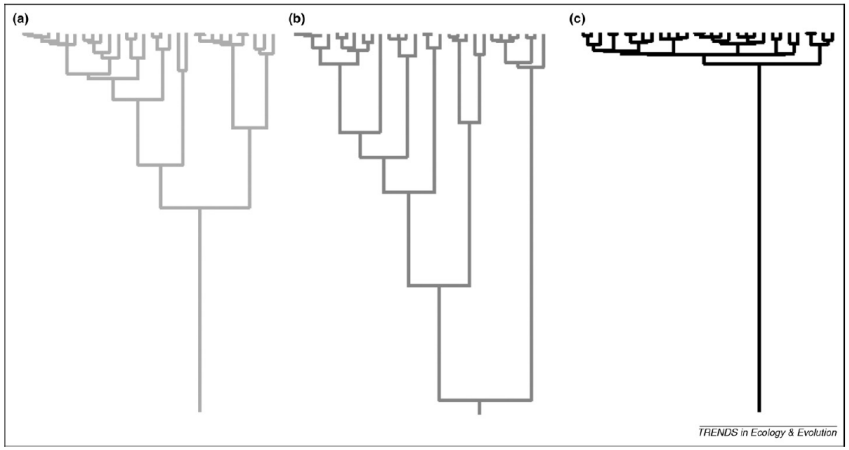
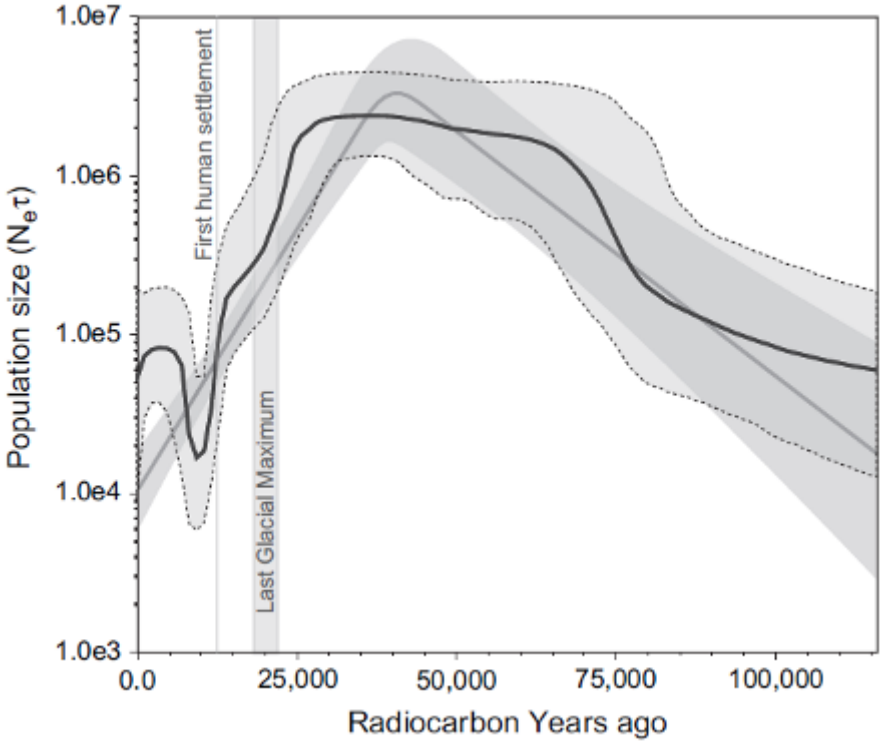



**Figure 2.** Growth signature in genealogies. Genealogies sampled from (a) constant-size, (b) exponentially shrinking and (c) exponentially growing populations.



*Bison priscus*

# Odhady migrace

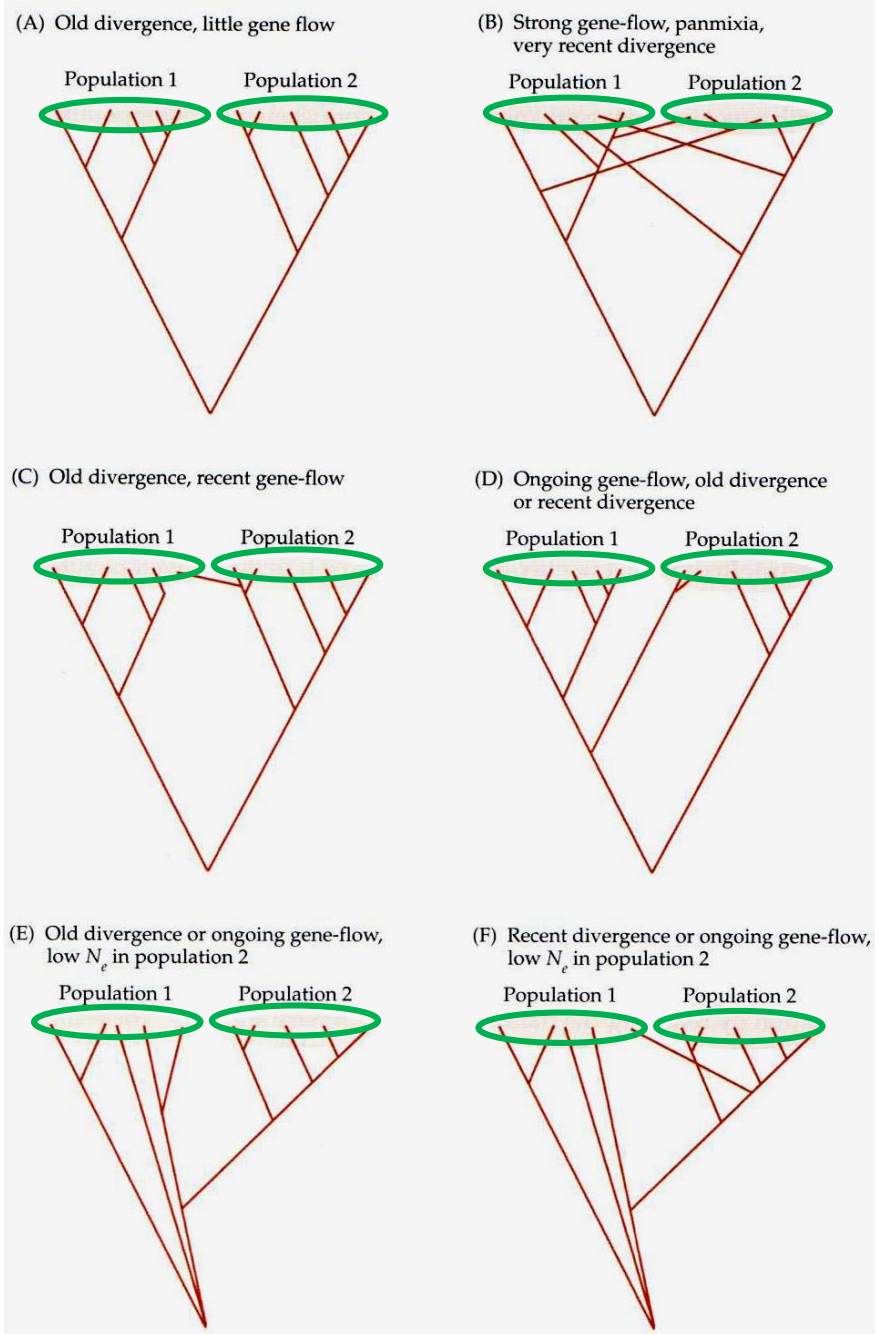
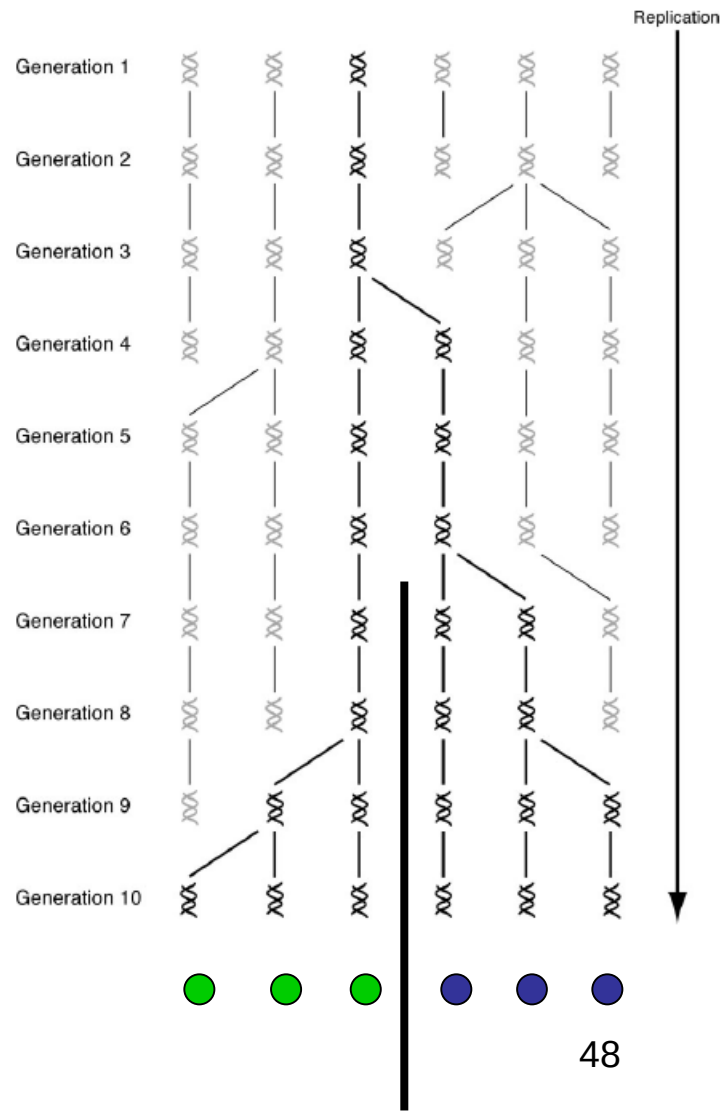
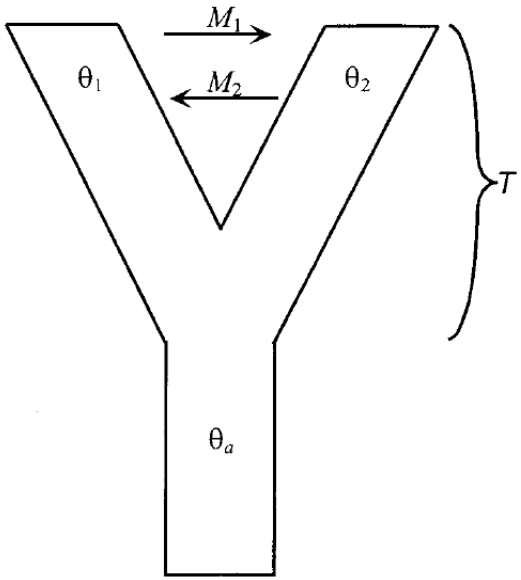


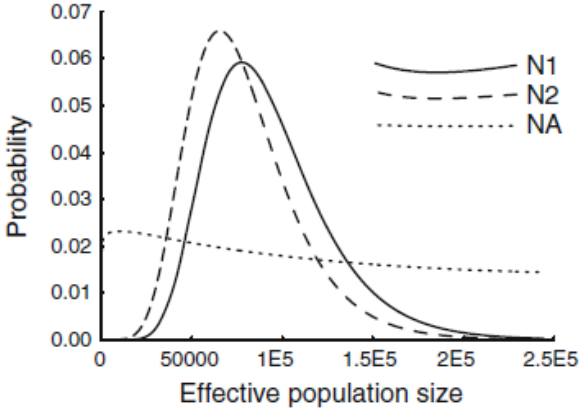

**Figure 5.8** Coalescence trees produced by different demographic and historical processes.

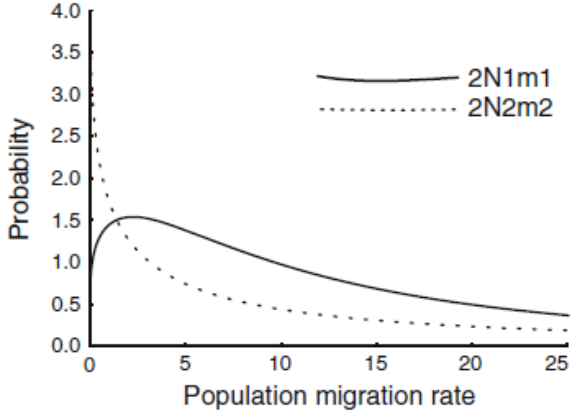


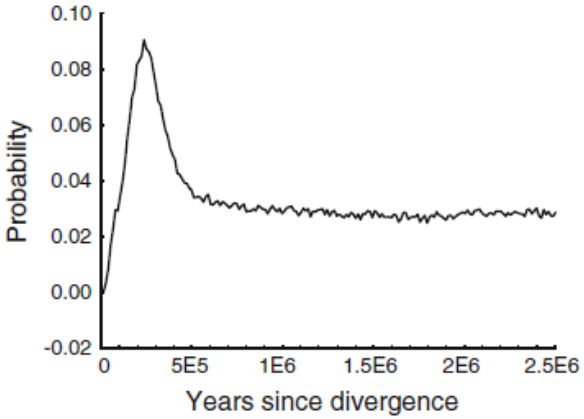
# Model „Isolation with Migration“




- Rozlišení ancestrálních polymorfismů a migrací
- Velikosti populací
- Čas divergence
- Migrace
- Programy IM, IMa, IMa2





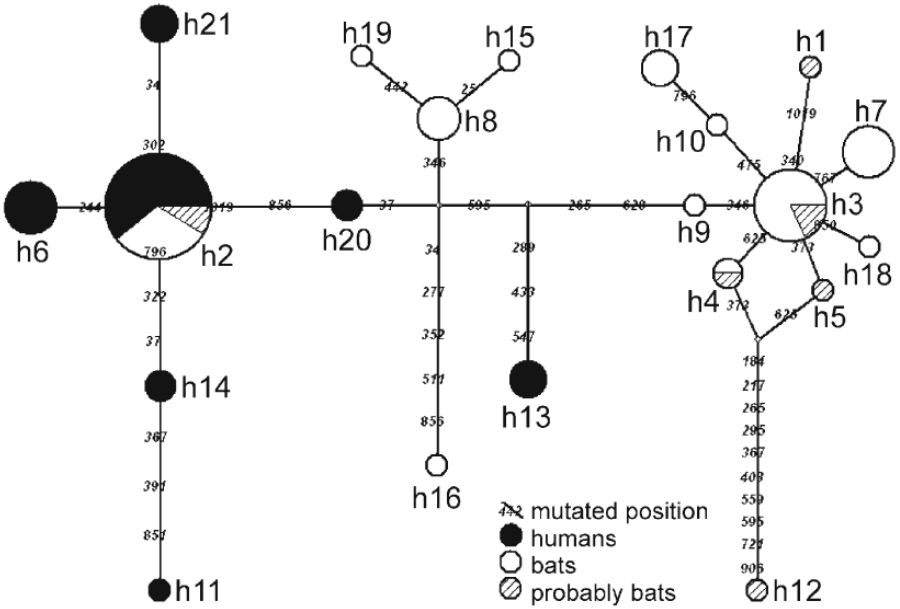
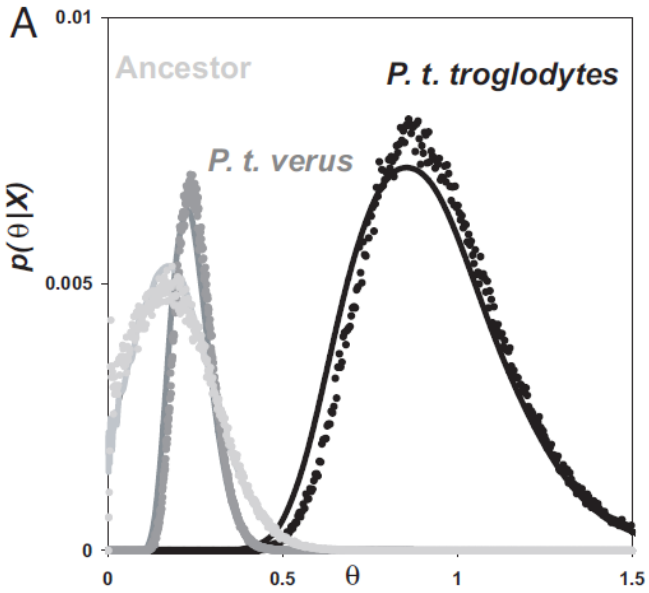
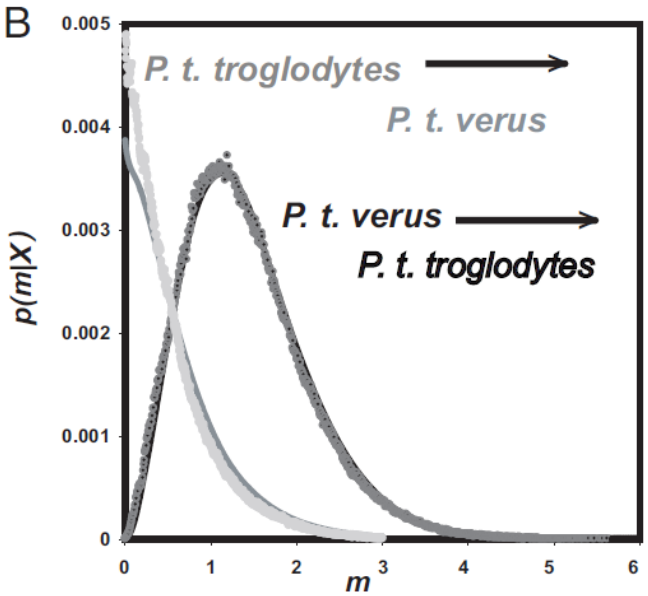




**Fig. 4** Estimated haplotype network of human- and bat-associated bedbug populations based on cytochrome oxidase subunit I and 16S rRNA gene. The populations marked as "probably from bats" with uncertain host association were shown by DFA to be originally from bats

Zbadáno 48 genů

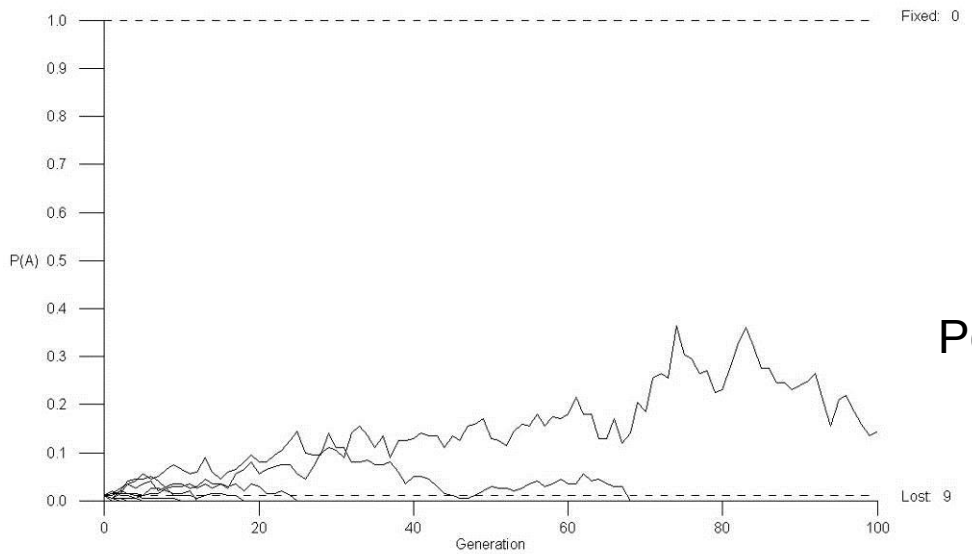



*Population size*

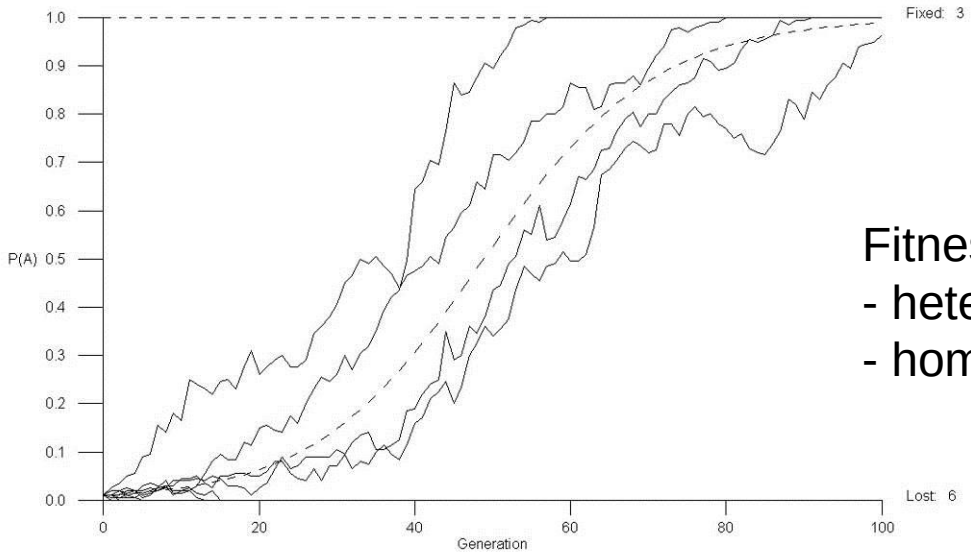
*Migration*

# Positivní selekce



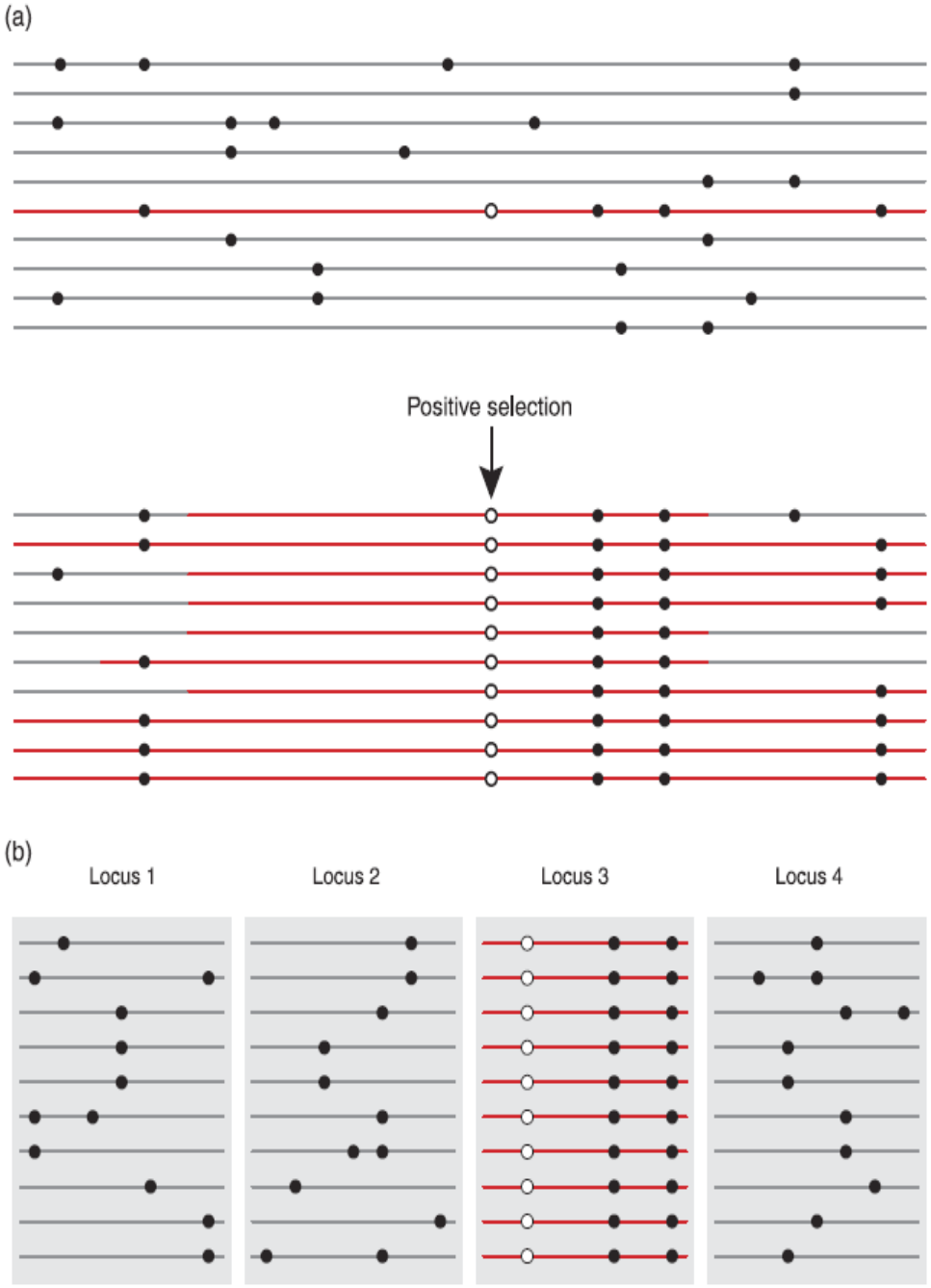

Simulováno 10 populací
Nová mutace
N je 100

Pouze drift



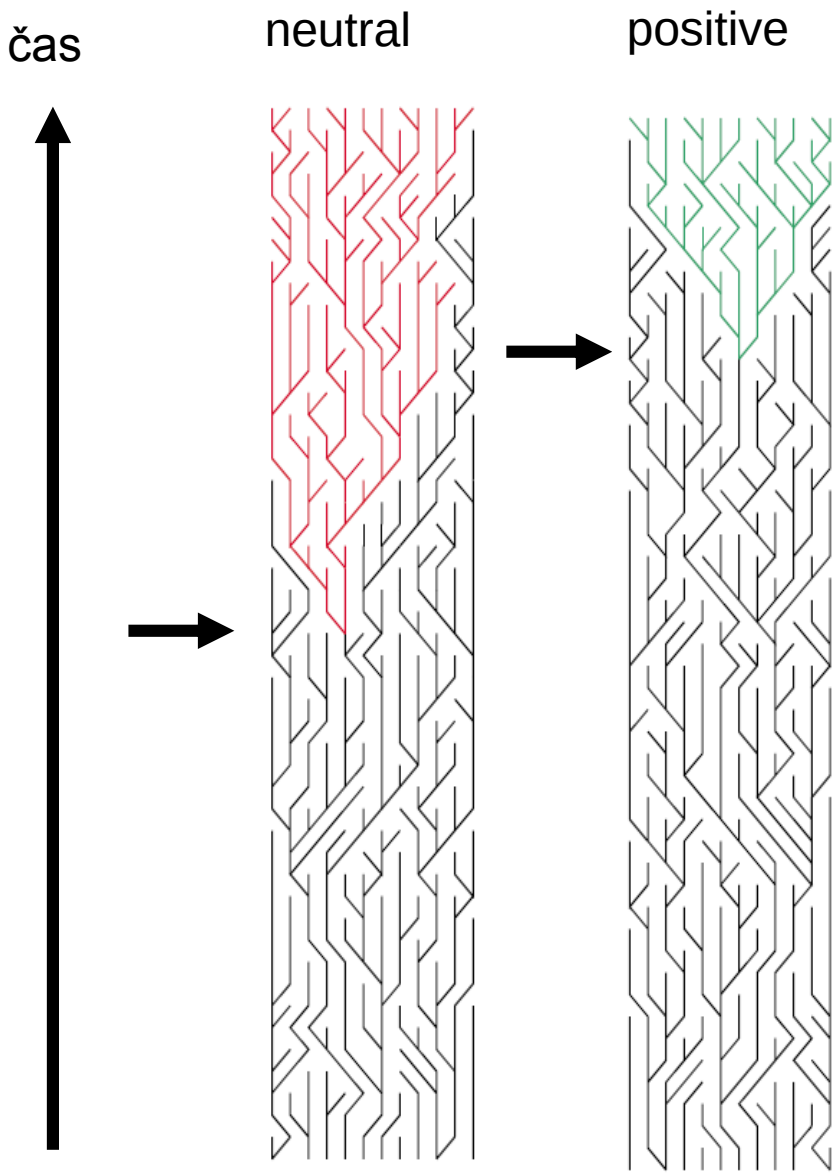

Fitness
- heterozygota pro novou alelu o 10 % vyšší
- homozygota o 20 % vyšší

# Selective sweep, hitchhiking



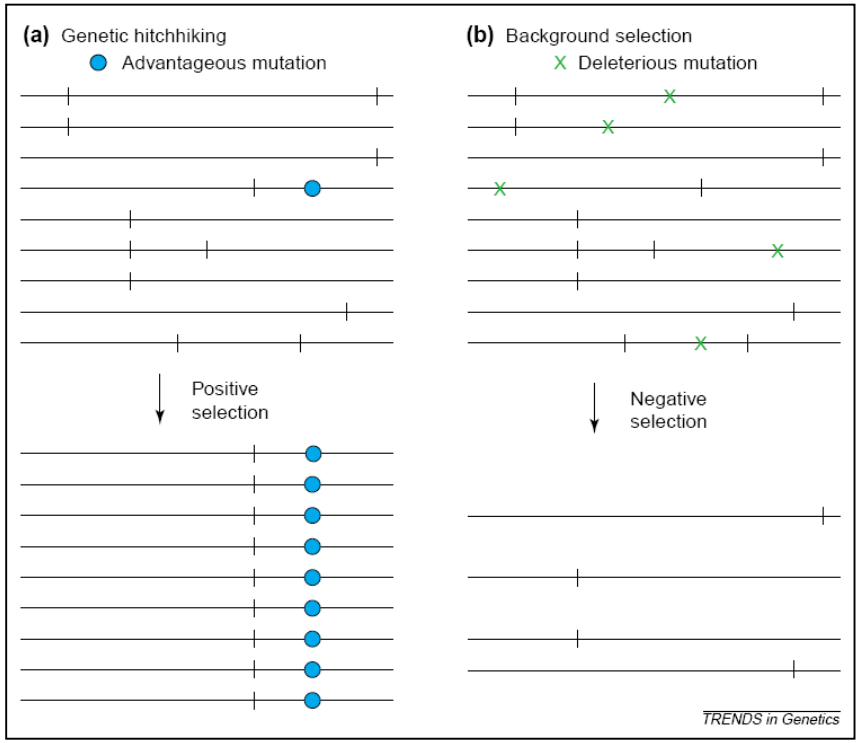
# Positive selection

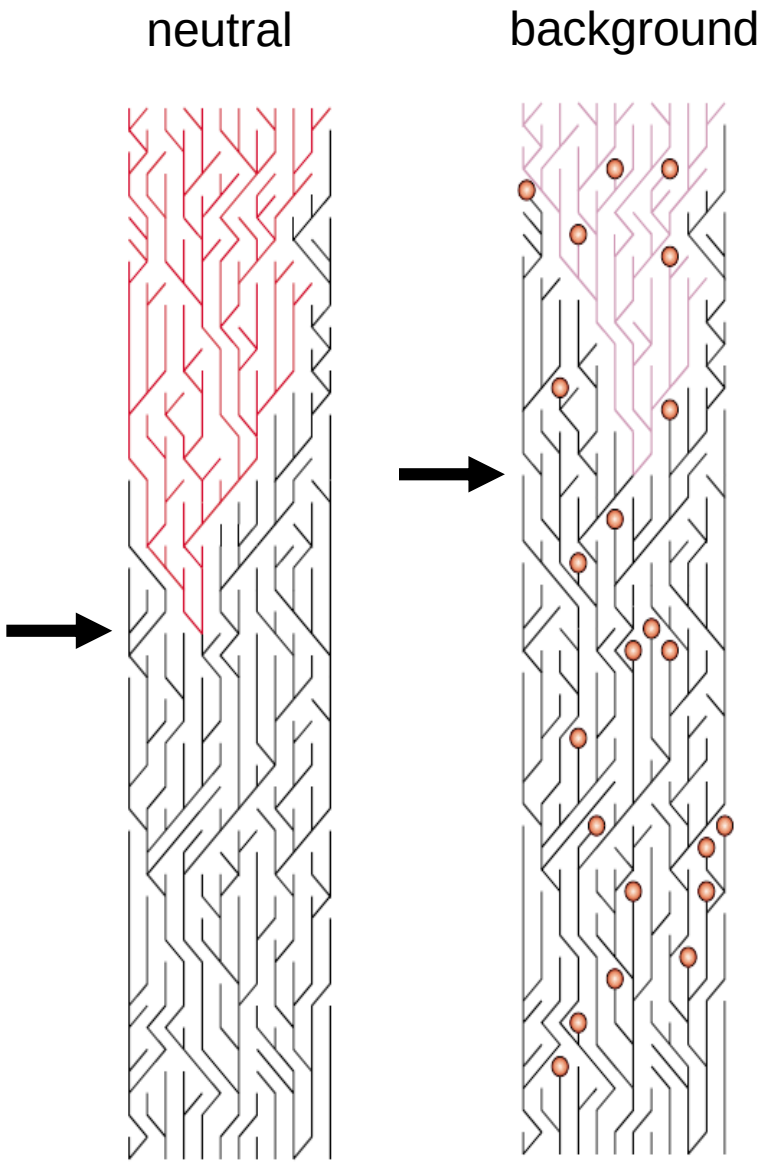
- Výhodná varianta
- Rychlé šíření
- Společný předek nedávno
- → selective sweep



# Background selection

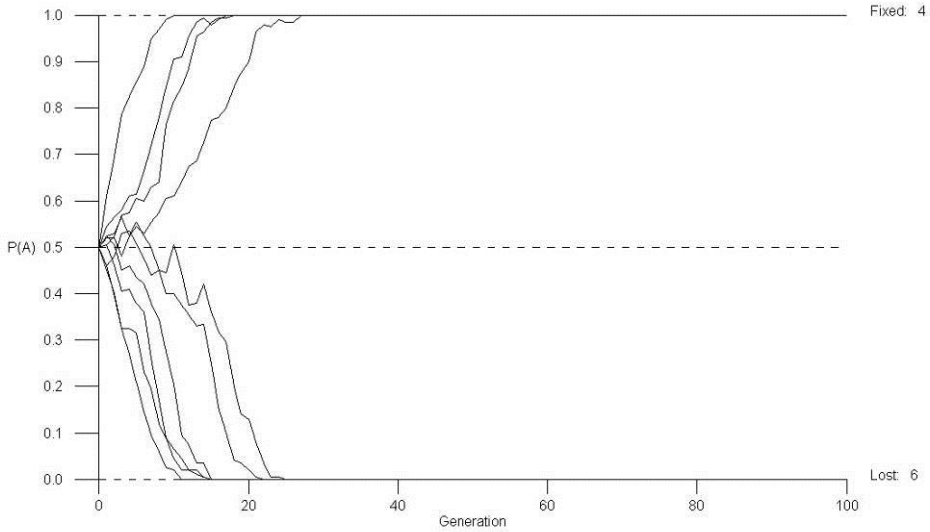
- Odstraňování škodlivých mutací
- Zánik linií (i těch, co nesou v blízkosti neutrální nebo slabě výhodné mutace)
- Snížení $N_E \rightarrow$ rychlejší fixace
- Kratší čas ke společnému předkovi



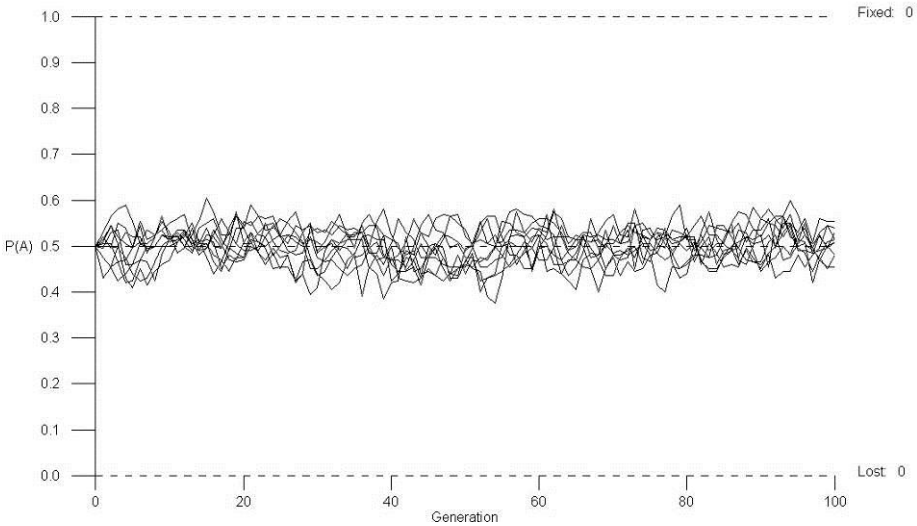



# Role heterozygotů

- Against
  - heterozygotes (heterozygote disadvantage, underdominance), nestabilní
  - homozygotes (heterozygote advantage, overdominance), stabilní
    → frekvence alel odpovídá *w* homozygotů




underdominance




overdominance

# Balancing selection

- Overdominance  je jen jedním z příkladů

  **Balancing selection** = selekce udržující polymorfismus

- Jsou i jiné příčiny

- Často – **Frequency-dependent selection**

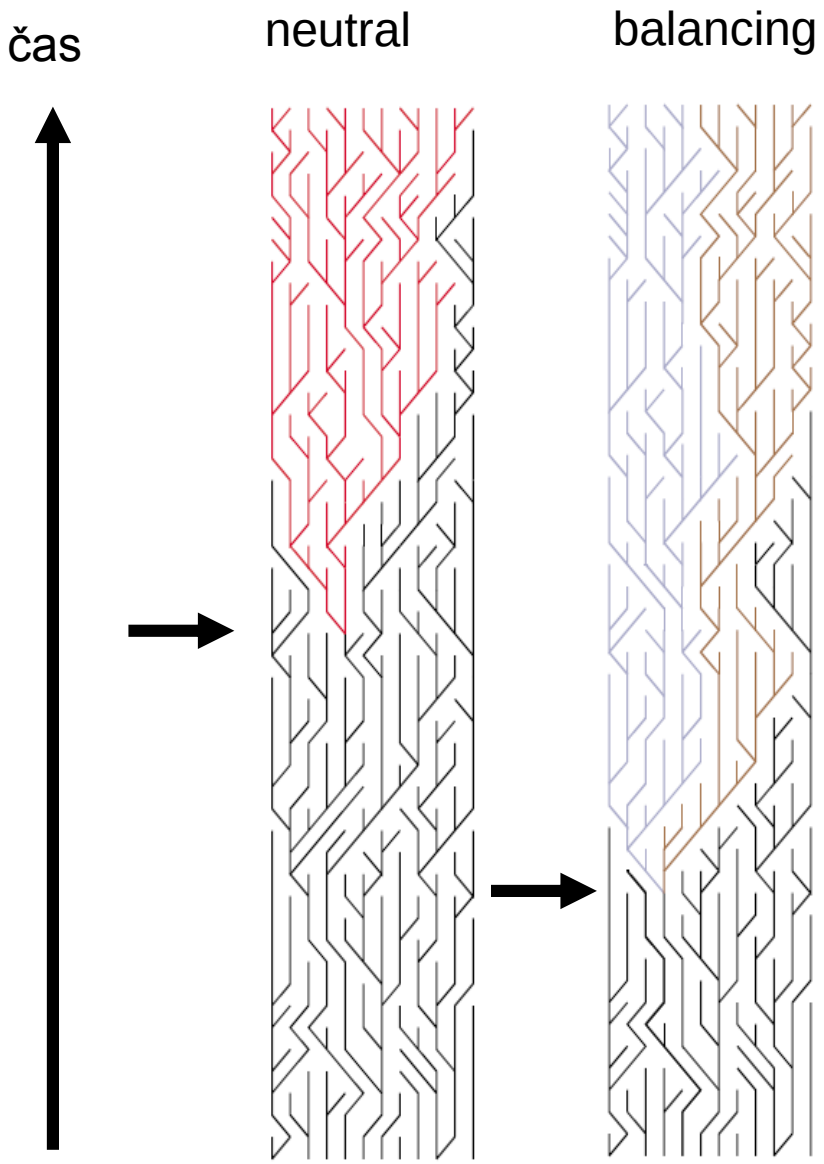
  *The fitness of a genotype depends on the genotype frequencies in the population.*

- Inverse frequency - dependent selection

  the rarer a phenotype → the greater its fitness

# Balancing selection

- Selekce na udržení více variant
- Overdominance, frekvenční závislost
- Společný předek variant hlouběji v minulosti
- Extrémní příklad – mezidruhové polymorfismy, např. MHC



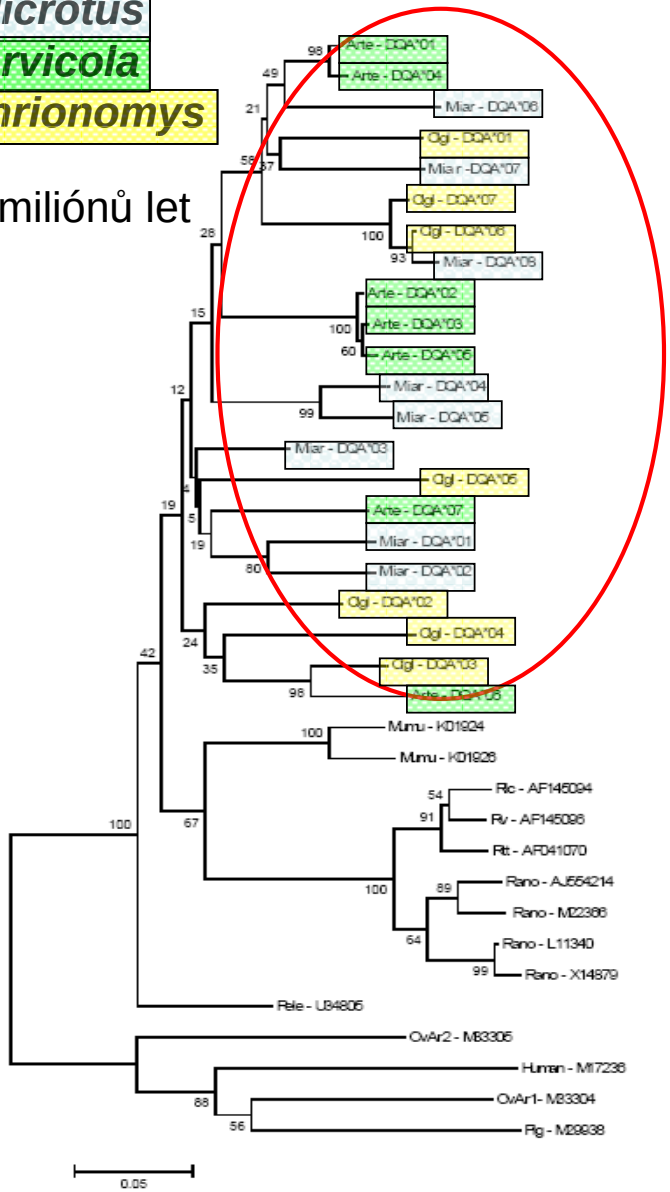
# Trans-species polymorfismus:

***Microtus***

***Arvicola***

***Clethrionomys***

7-8 miliónů let

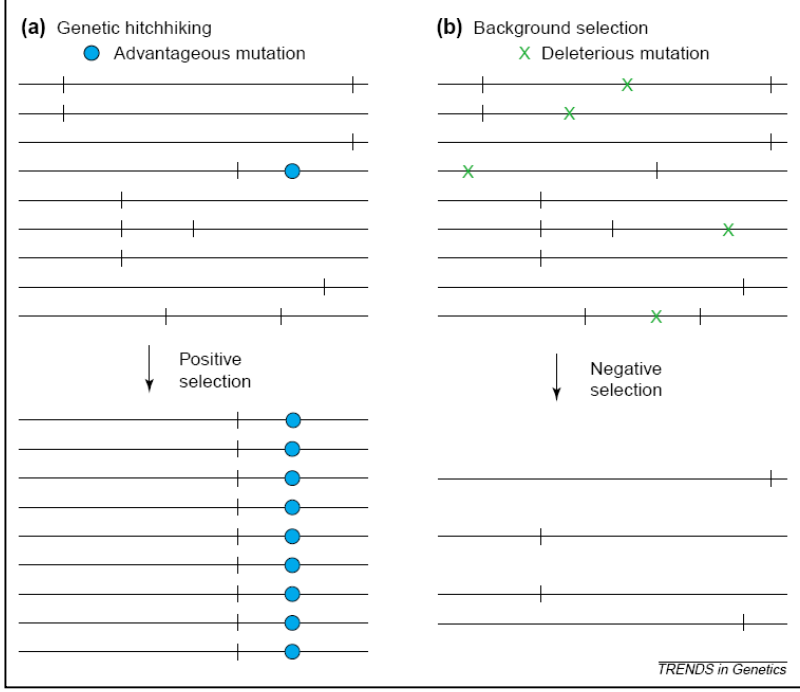


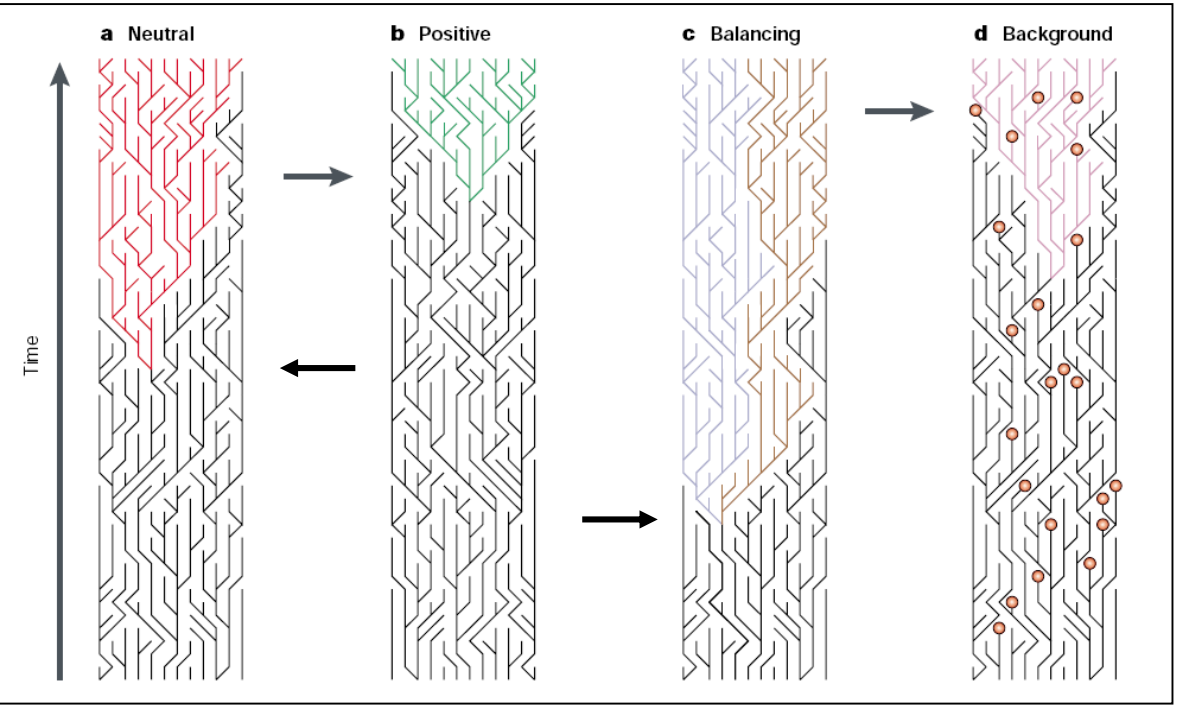
- MHC: stejné nebo podobné alely byly nalezeny u různých druhů
- když působí selekce, tak se výhodné sekvence udrží i přes období divergence druhů

# Selekce a variabilita

- **Positivní** i **negativní** snižují variabilitu

- **Balancing** zvyšuje variabilitu
  - udržení více linií
  - delší čas ke společnému předku
    → více času na mutace



# Zásadní problém

Vliv selekce často těžko odlišitelný od jiných vlivů.

??????????????????????????????????????

Vidím výsledek:

Selekce

Demografické změny

Populační struktury

Migrace …

??????????????????????????????????????

Pozor na analýzy založené na malém počtu genů.