

**KOALESCENCE**

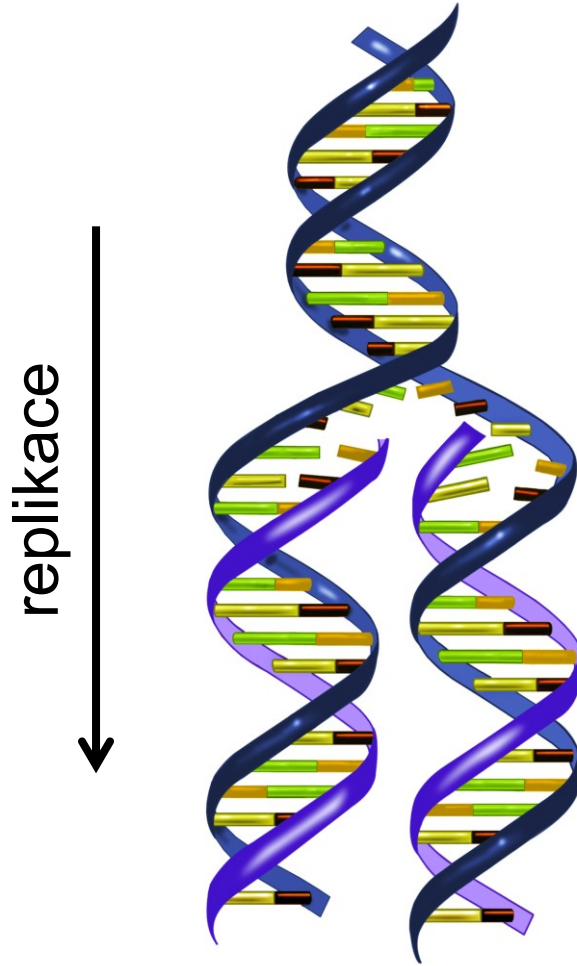
# **GENOVÉ GENEALOGIE A TEORIE KOALESCENCE**

Dosud: co se s populací stane v příští generaci?

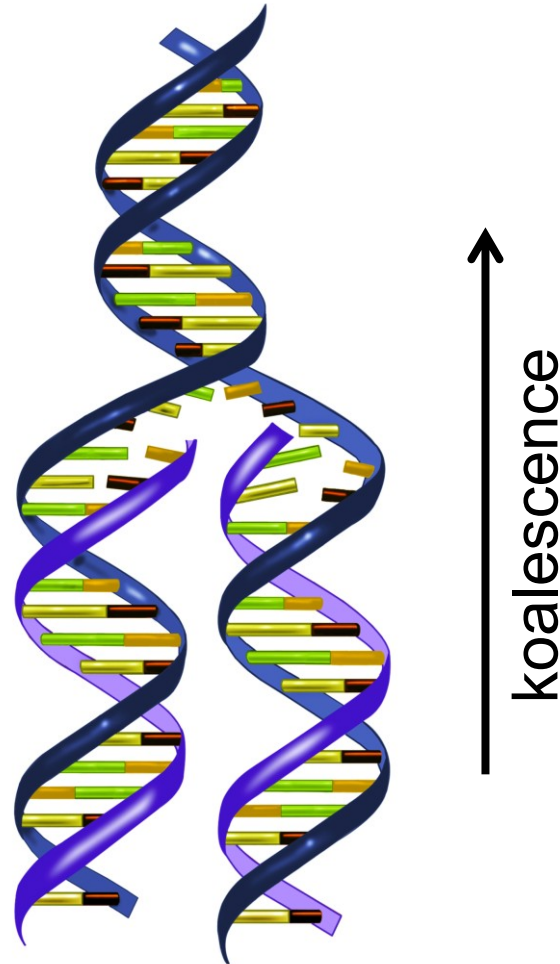
Teď: co předcházelo současnému stavu?

⇒ „pohled zpět“

# Základní premisa populační genetiky: DNA replikuje...

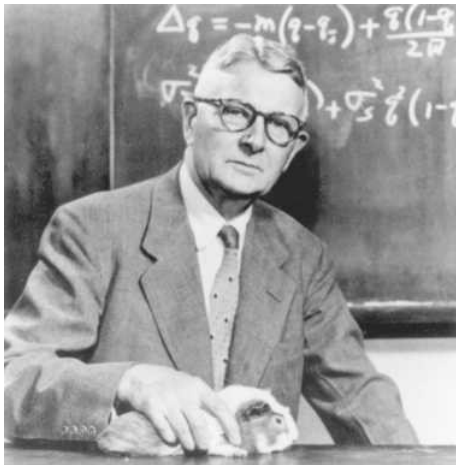


Základní premisa populační genetiky: DNA replikuje...

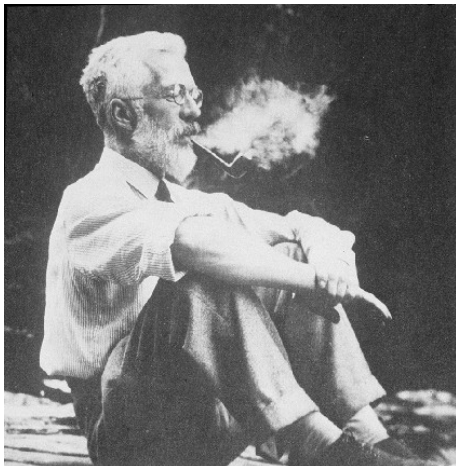


... opakem replikace je koalescence

# Wrightův-Fisherův model:



Sewall Wright

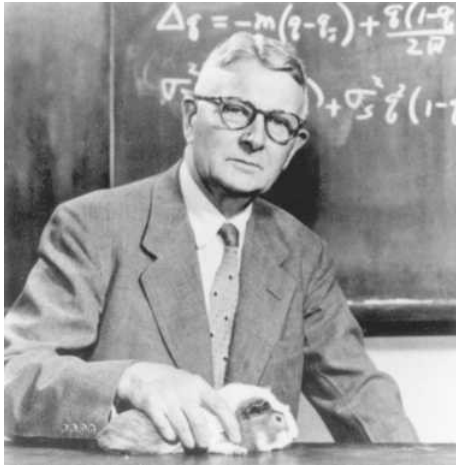


Ronald A. Fisher

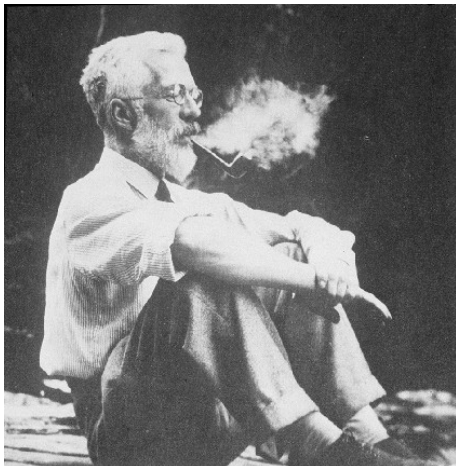
## WF populace – opakování:

- diploidní, hermafrodit
- velikost omezená, žádné fluktuační  $N$
- náhodné oplození
- kompletní izolace (žádný tok genů)
- diskrétní generace
- žádná věková struktura
- žádná selekce
- rozptýlení výběru gamet do další generace  
→ Poissonovo rozdělení

# Sortování linií ve WF modelu:



Sewall Wright



Ronald A. Fisher

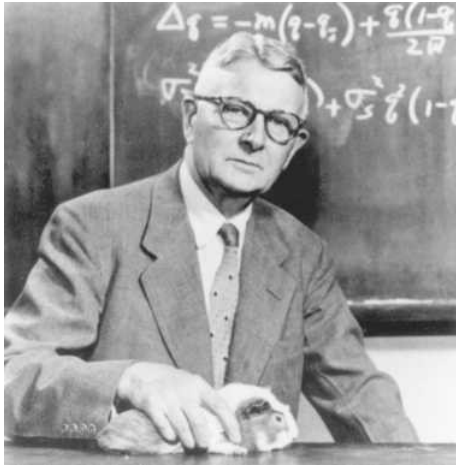
generace 1



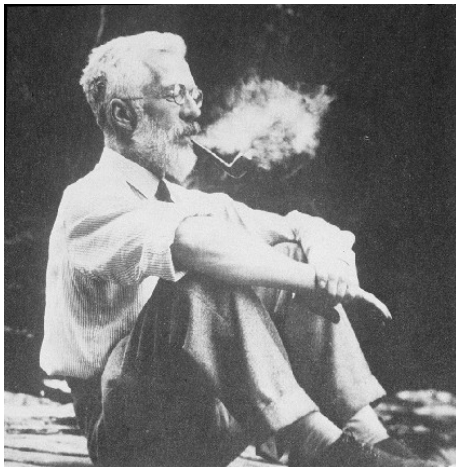
čas



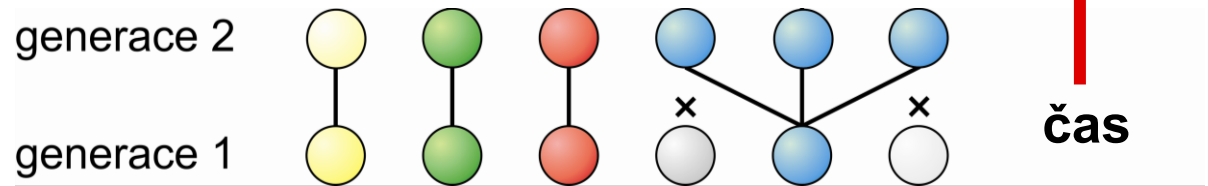
# Sortování linií ve WF modelu:



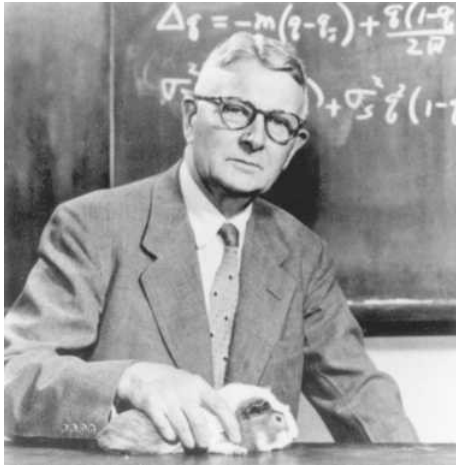
Sewall Wright



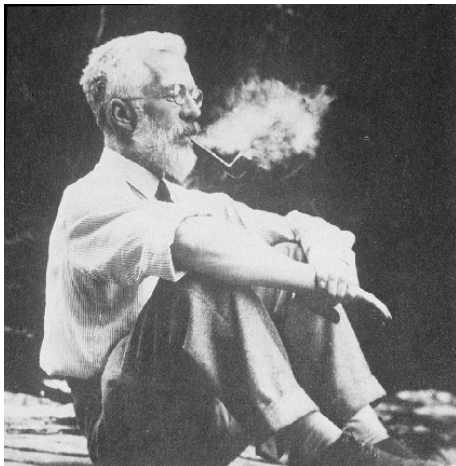
Ronald A. Fisher



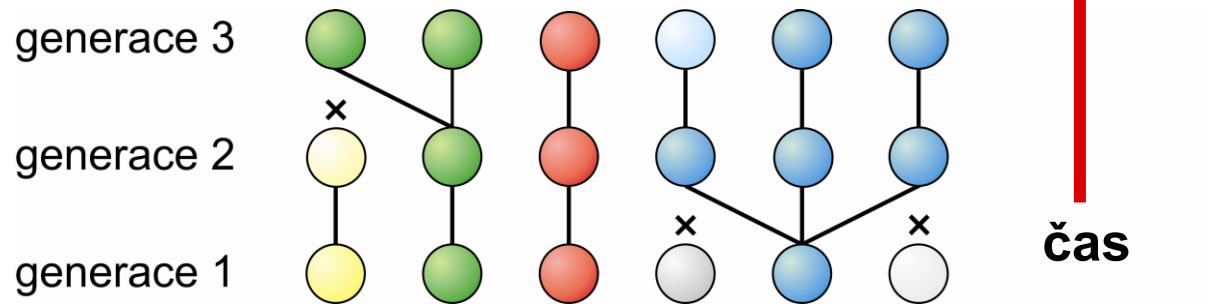
# Sortování linií ve WF modelu:



Sewall Wright

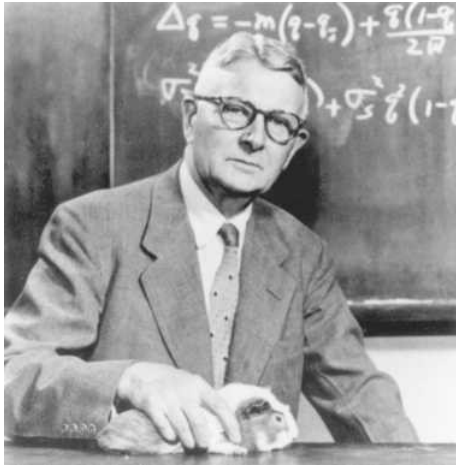


Ronald A. Fisher

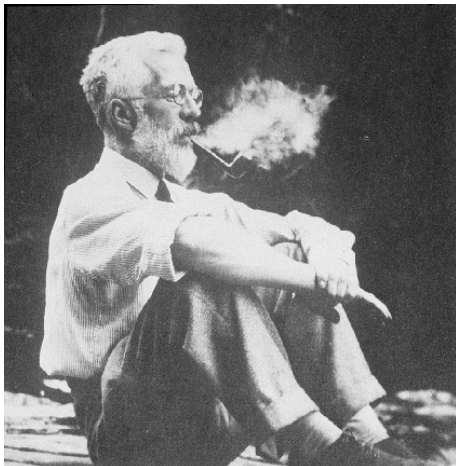




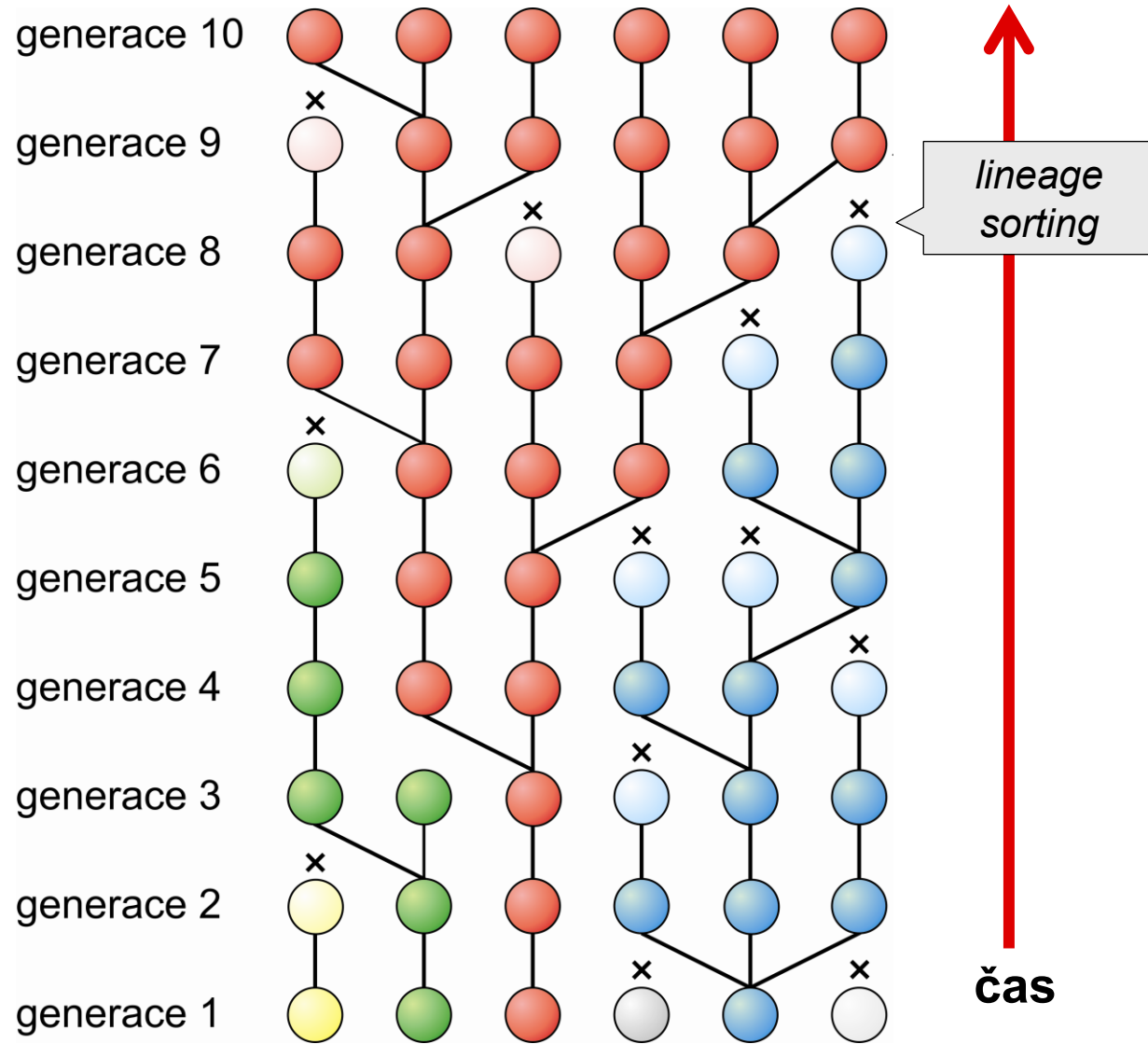
# Sortování linií ve WF modelu:



Sewall Wright



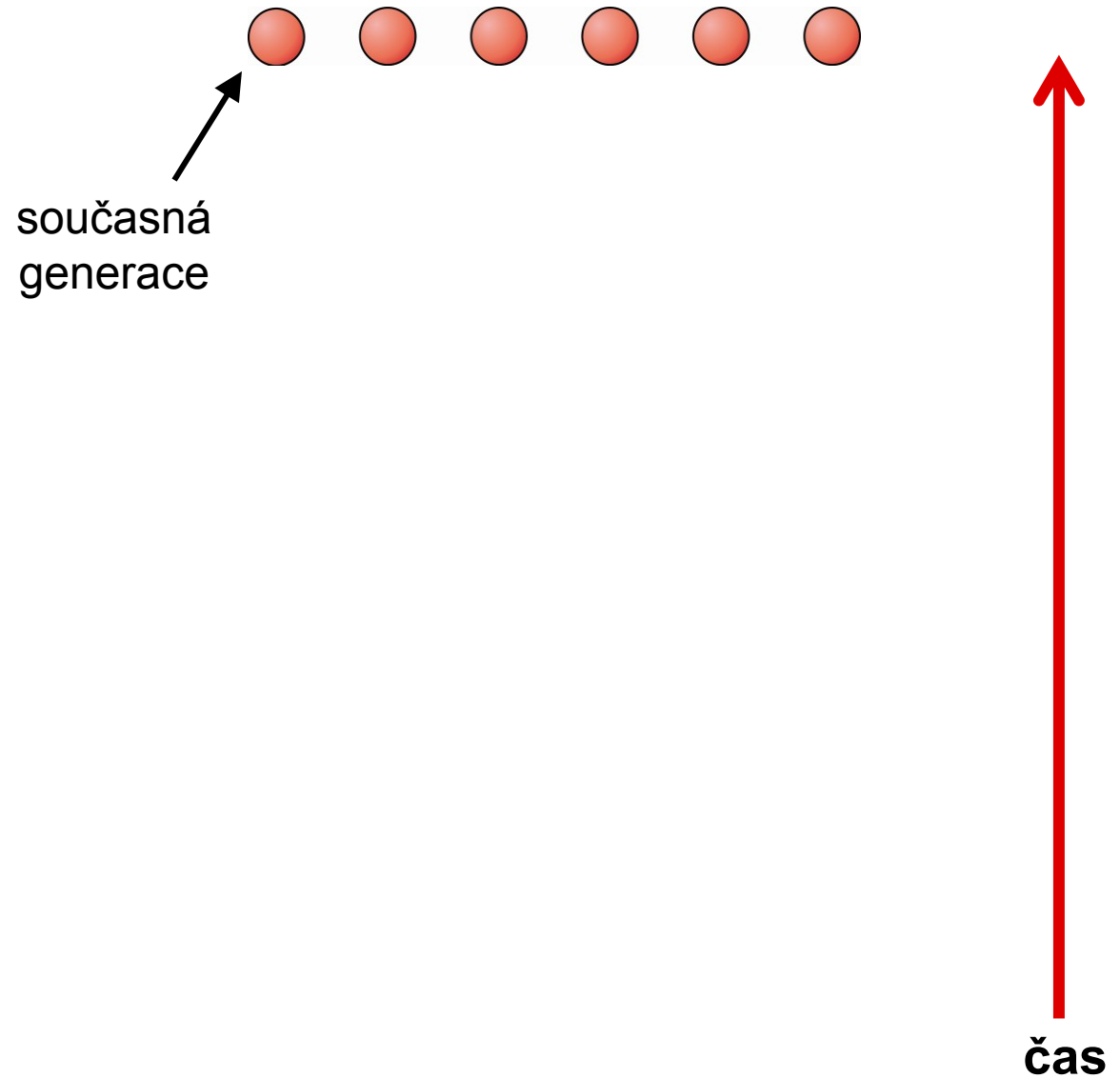
Ronald A. Fisher



# Koalescence:



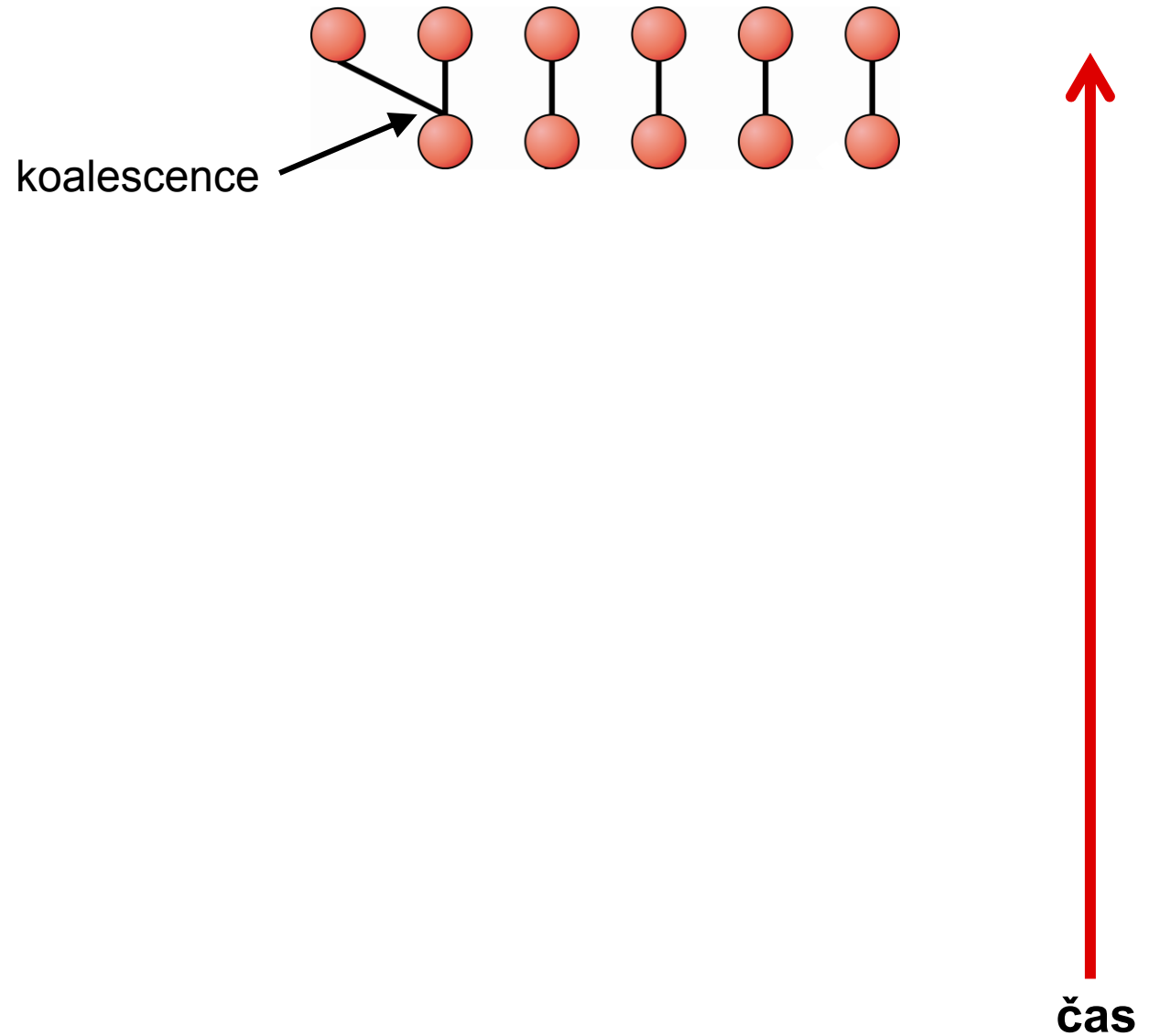
John F.C. Kingman



# Koalescence:



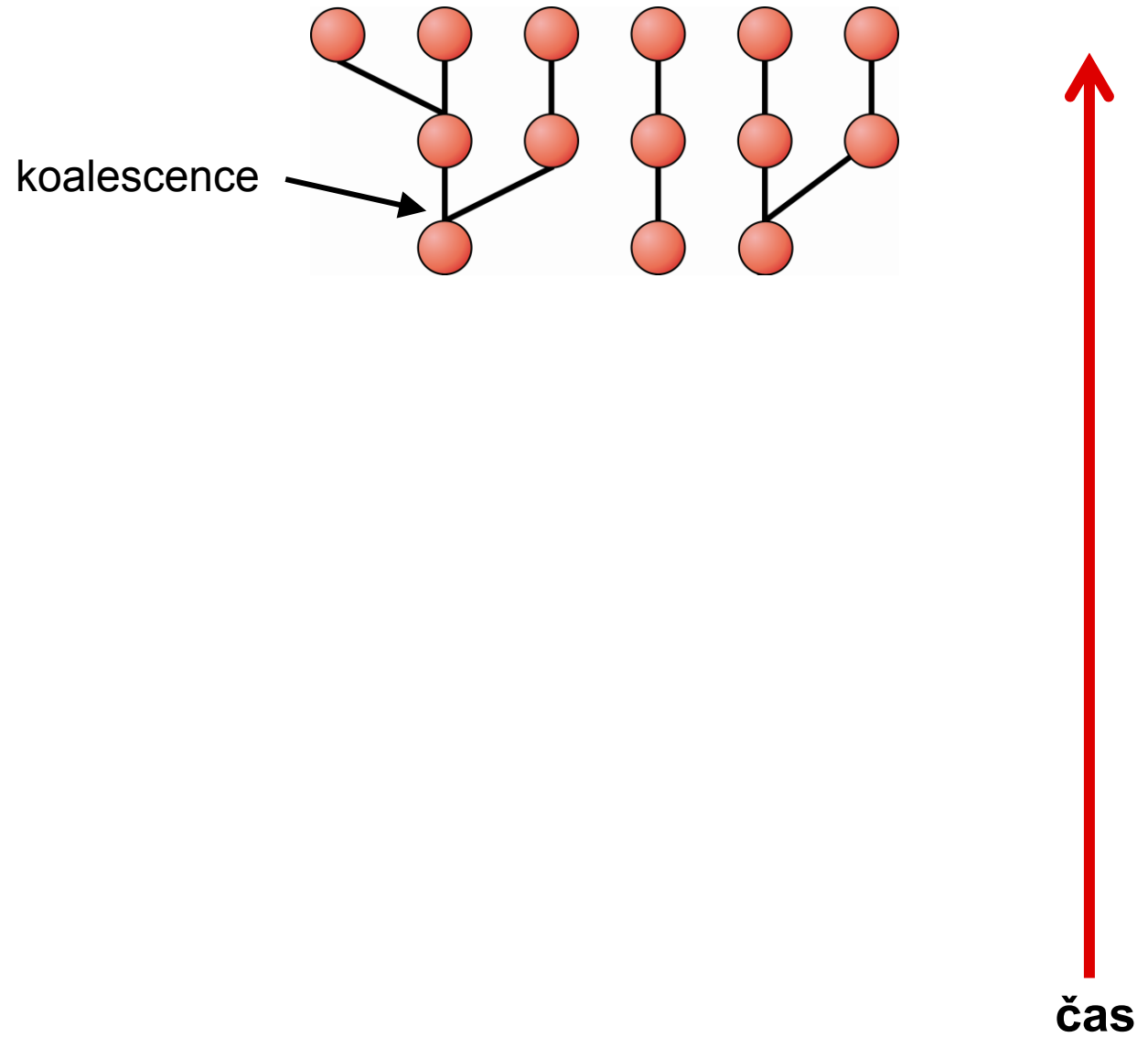
John F.C. Kingman



# Koalescence:



John F.C. Kingman

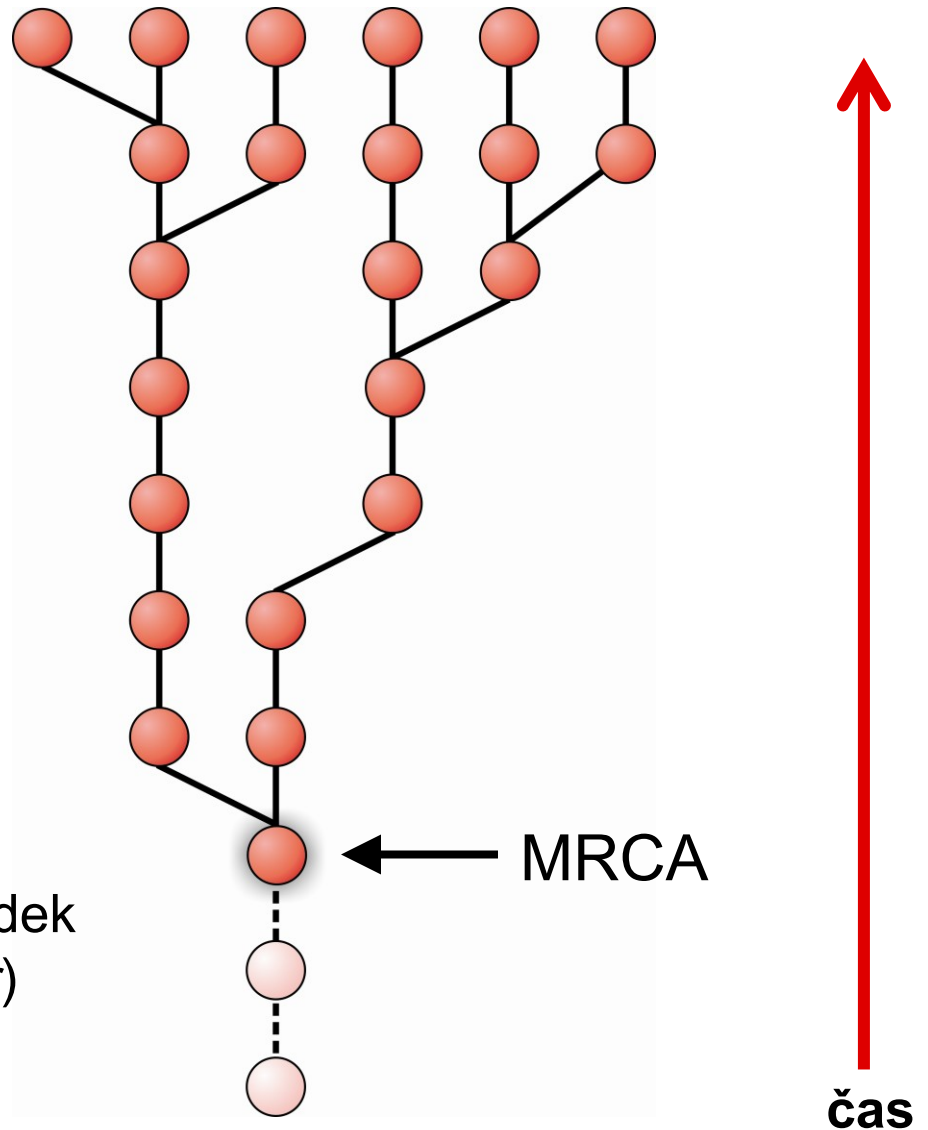


# Koalescence:



John F.C. Kingman

MRCA = nejblížší společný předek  
(*most recent common ancestor*)

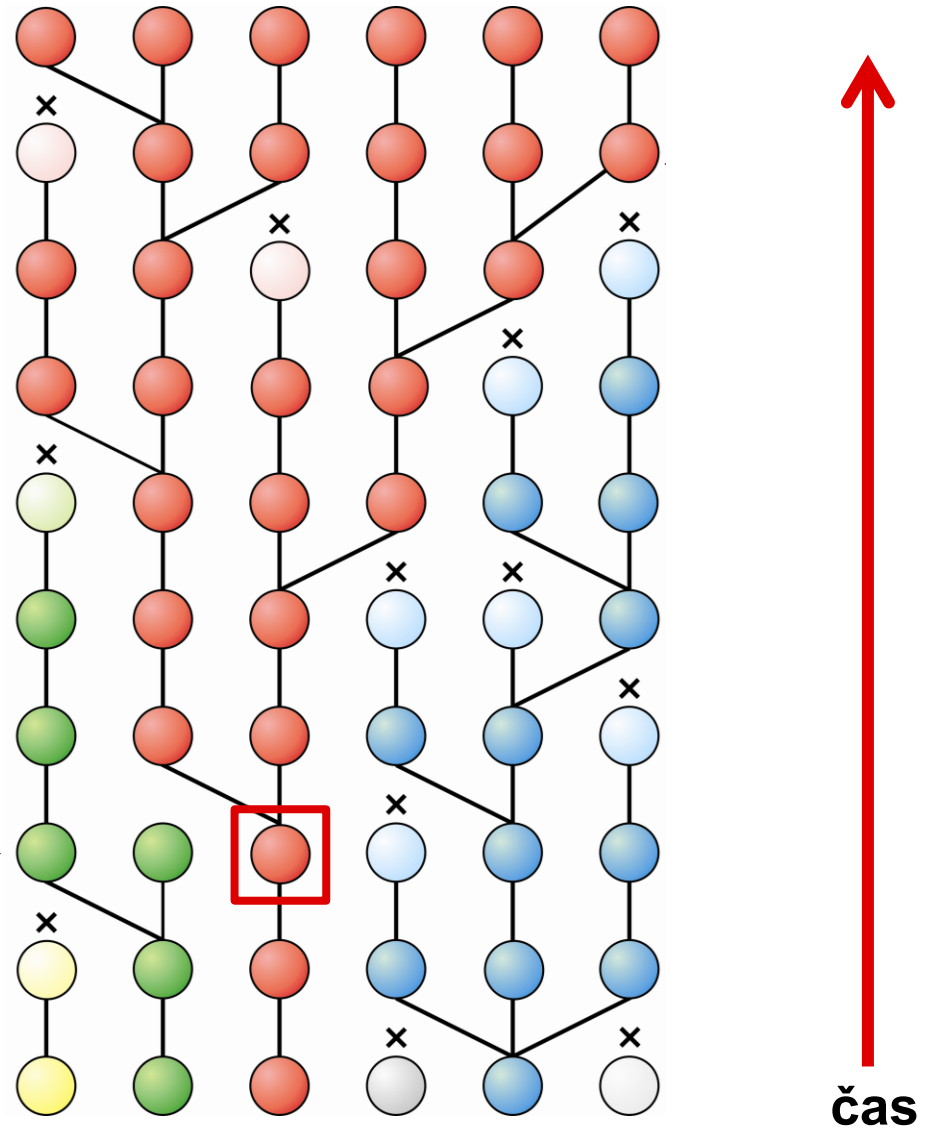


# Koalescence:



John F.C. Kingman

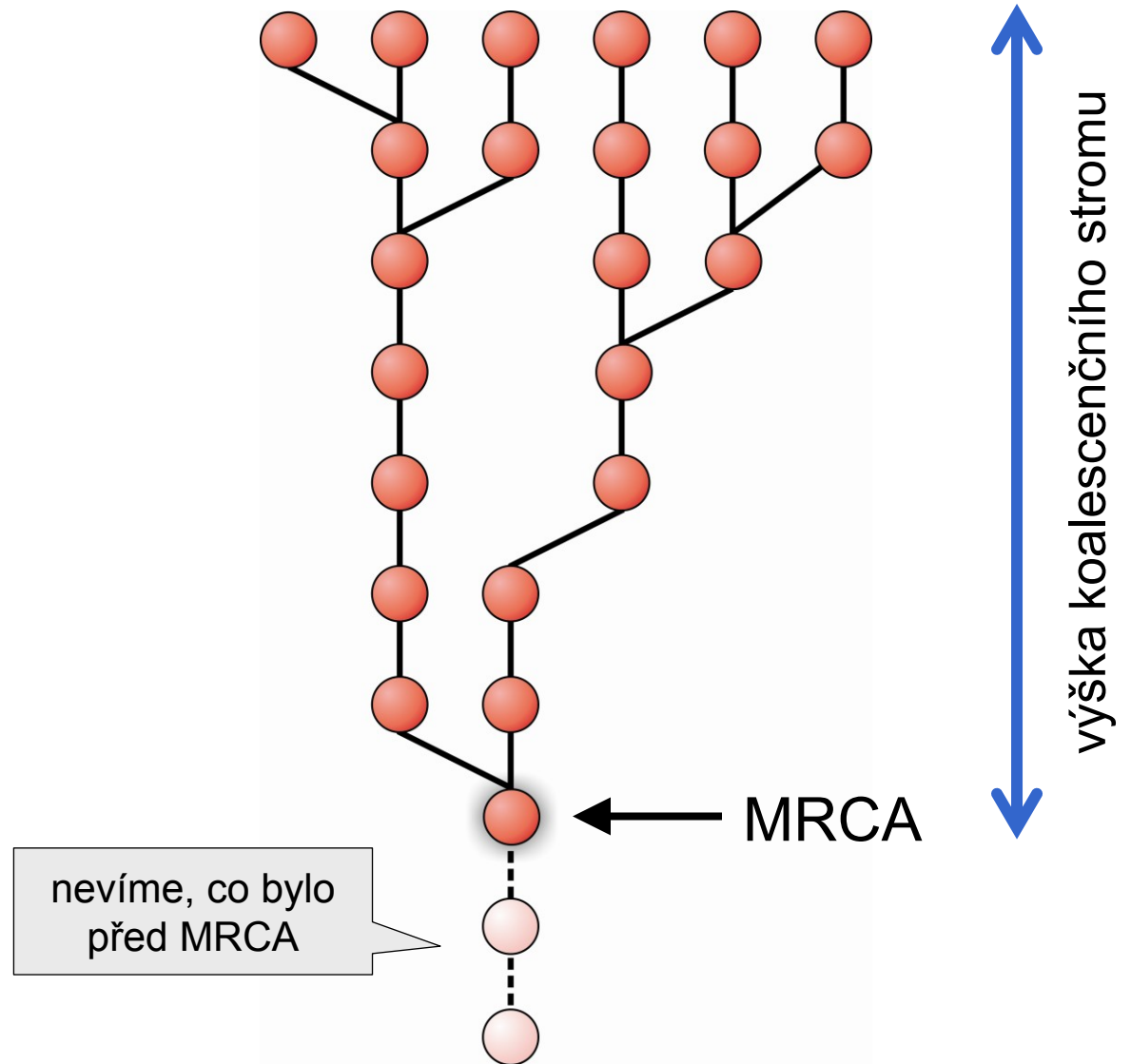
nevíme, kolik kopií bylo v generaci MRCA



# Koalescence:



John F.C. Kingman

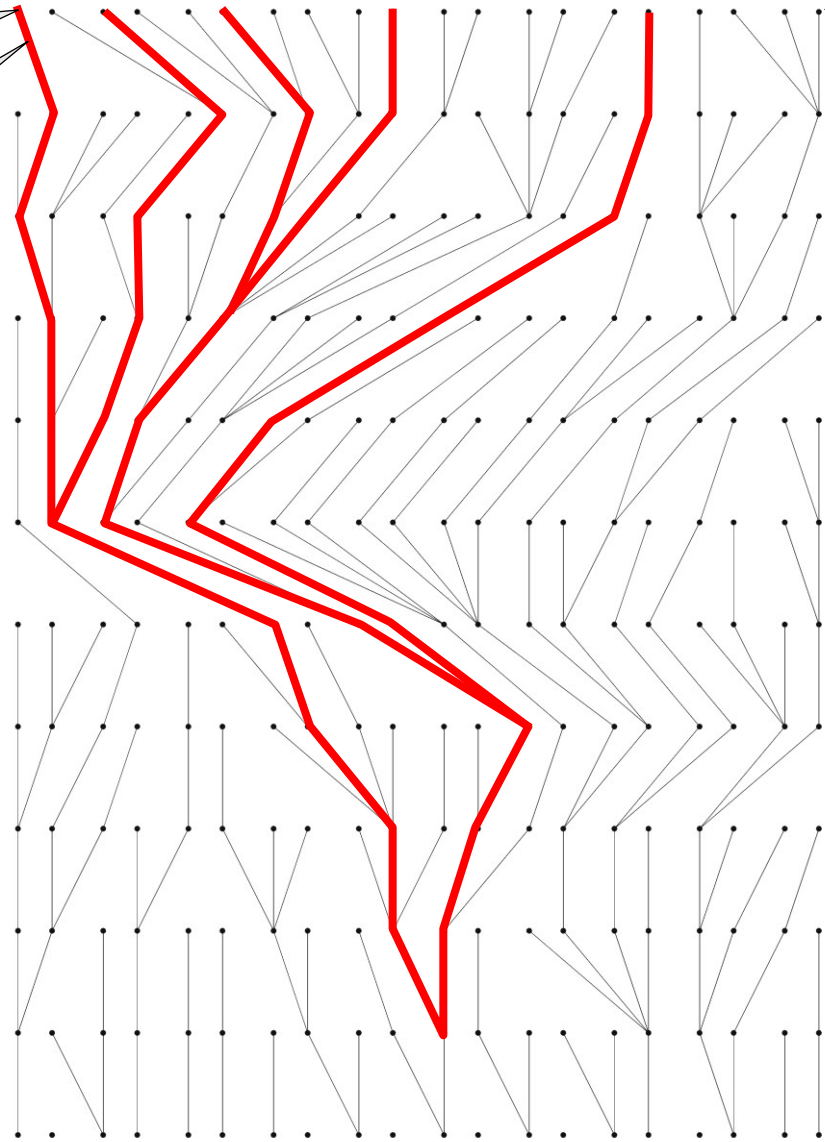




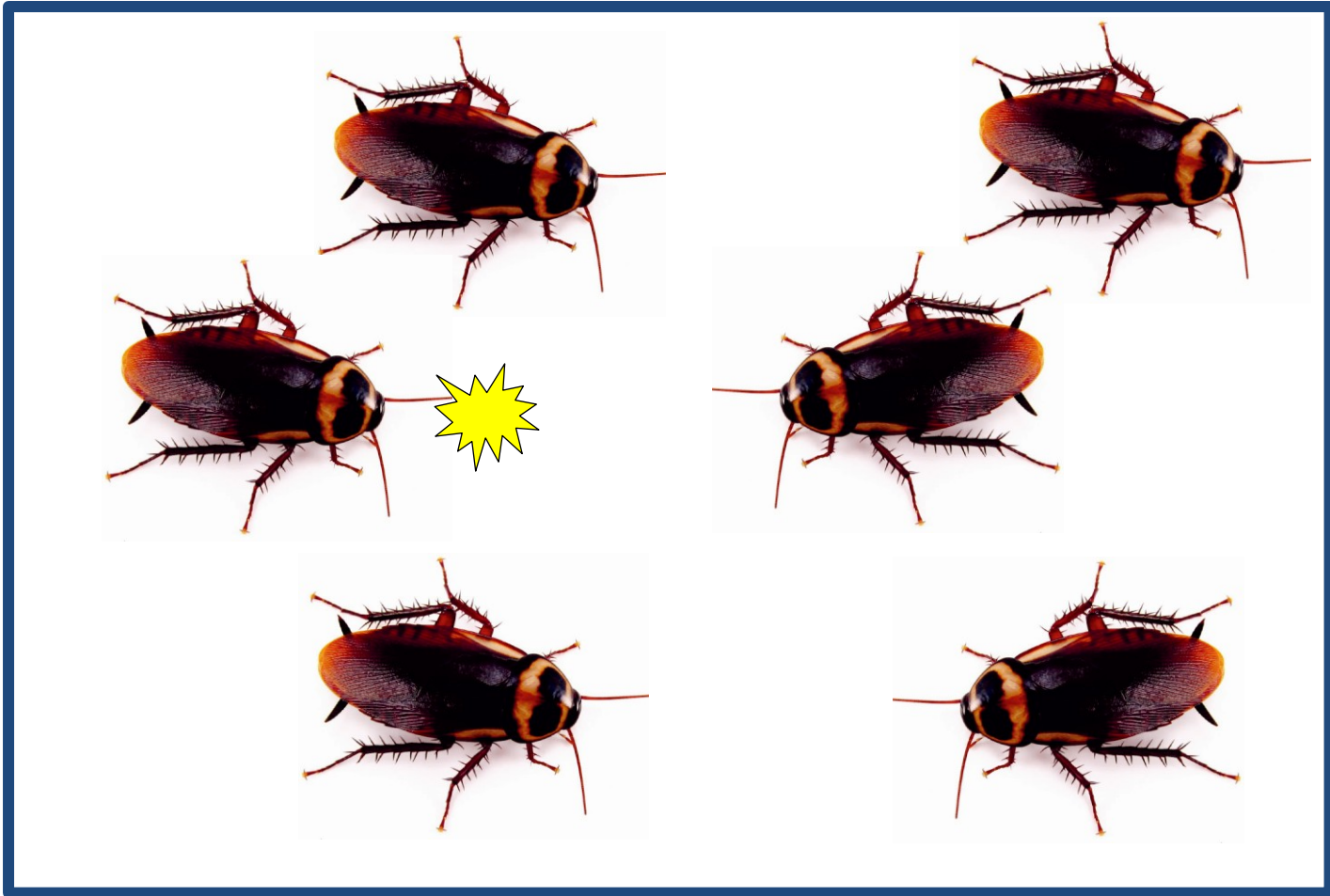
$n = 5$  kopií  
ve vzorku

většinou  
 $n \ll N$

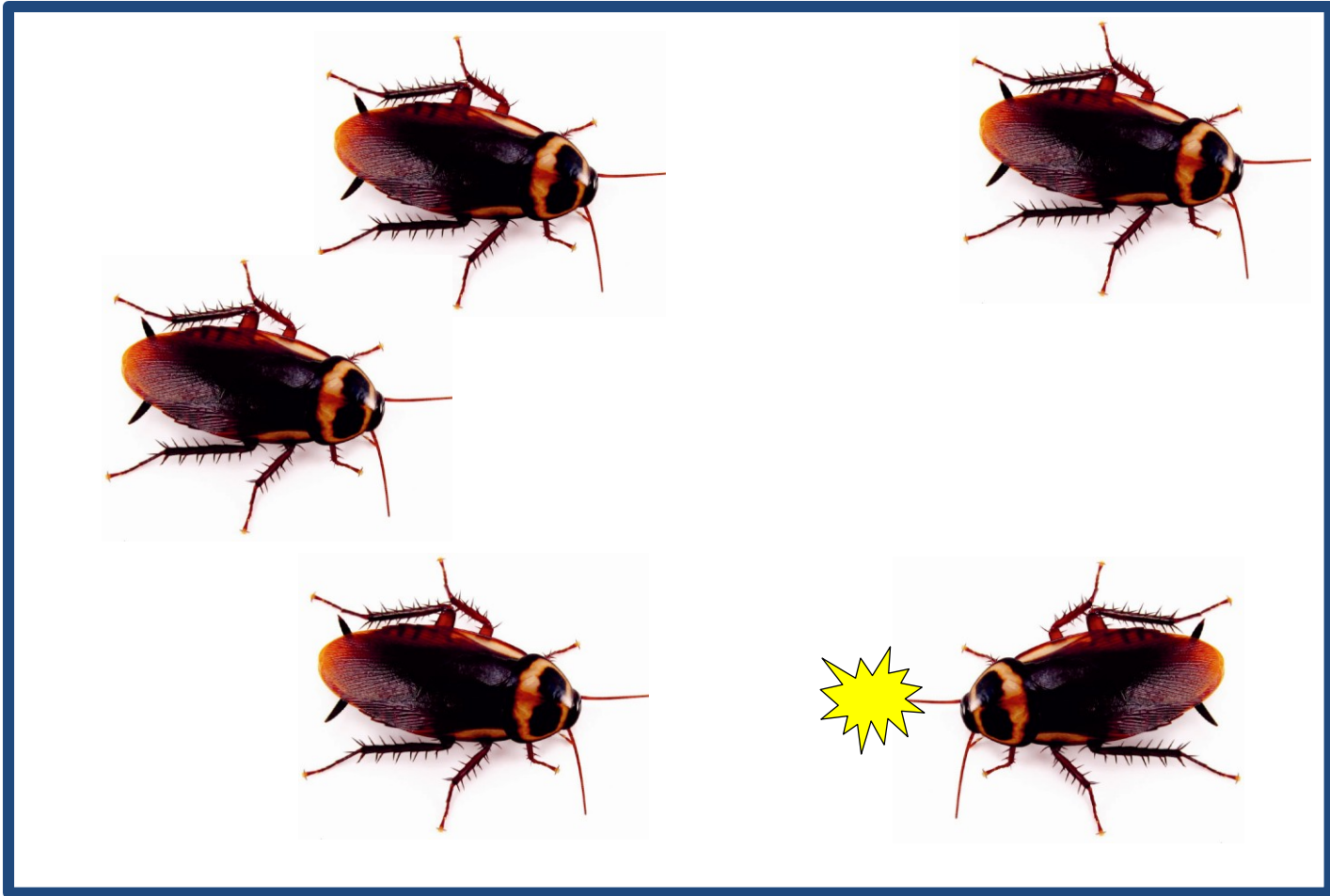
$N = 20$  kopií  
v populaci







Pravděpodobnost setkání 2 švábů je  $n(n - 1)/4N$ , kde  $n$  = počet švábů v krabici,  $N$  = velikost krabice



při koalescenci se počet švábů (kopií) sníží o 1 ...



s tím, jak klesá počet  
švábů ( $n$ ), roste čas k  
dalšímu kontaktu  
(koalescenci)

při koalescenci se počet švábů (kopií) sníží o 1 ...



... až nakonec zůstane jen 1 kopie

rozdělení času mezi koalescencemi je přibližně exponenciální:

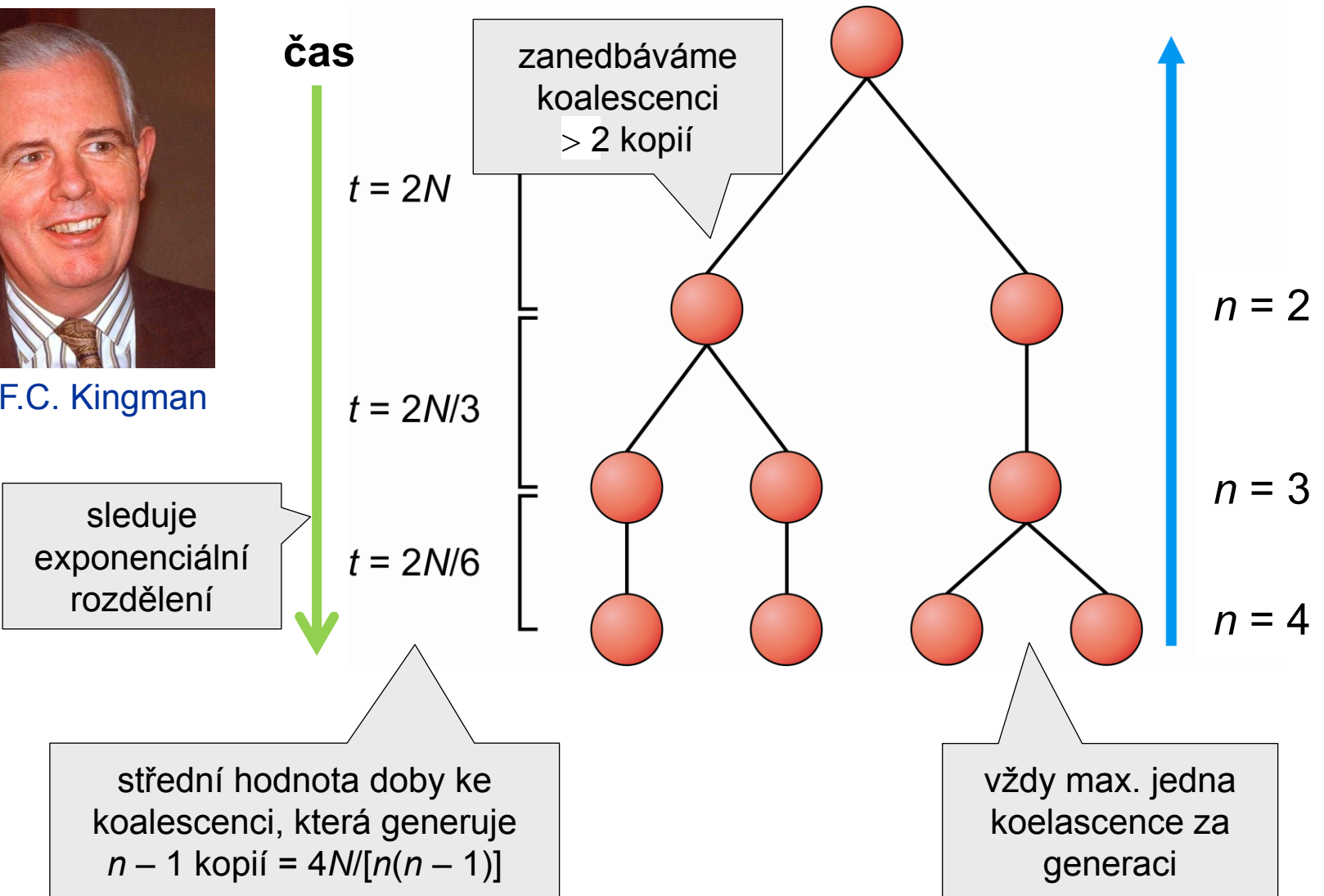


\*) viz počet švábů v krabici

# Kingmanova koalescence:



John F.C. Kingman



pravděpodobnost, že dvě kopie splynou v následující generaci =  
identita původem =  $1/(2N_{eF})$  ... v ideální populaci =  $1/(2N)$   
pro diploidní gen

obecně  $1/(xN)$ , kde  $x$  = ploidie

pravděpodobnost, že v té době nedojde ke koalescenci =  $1 - 1/(xN)$

⇒ pravděpodobnost koalescence před  $t$  generacemi =

$\Pr[\text{žádné koalescence pro prvních } t - 1 \text{ generací}] \times \Pr[\text{koalescence v gen. } t]$

pro diploidní gen:

$$Pr = \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right)$$

průměrná doba ke koalescenci 2 kopií =

$$= \sum_{t=1}^{\infty} t \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right) = 2N$$

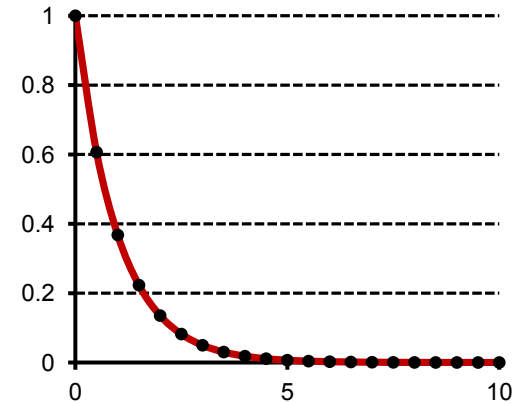
čas od jedné koalescence ke druhé se řídí geometrickým rozdělením, které se dá aproximovat exponenciálním

⇒ střední hodnota času do příští koalescence:

$$E[T(2)] = \frac{2N}{\binom{n}{2}} = \frac{2N}{\frac{n(n-1)}{2}} = \frac{4N}{n(n-1)}$$

$$\binom{n}{2} = \frac{n!}{(n-2)!2!} = \frac{n(n-1)}{2}$$

s každou koalescencí se počet kopií sníží o 1, tj.



$$E[T(n)] = \frac{4N}{n(n-1)} + \frac{4N}{(n-1)(n-2)} + \frac{4N}{(n-2)(n-3)} + \dots + \frac{4N}{2}$$

⇒ proces koalescence se postupně zpomaluje

doba koalescence MRCA všech  $n$  kopií =  $4N \left(1 - \frac{1}{n}\right)$

pro  $n = 2 \rightarrow t = 2N$   
 pro velká  $n \rightarrow t \approx 4N$   
 (obecně  $t \approx 2 \times nN$ )

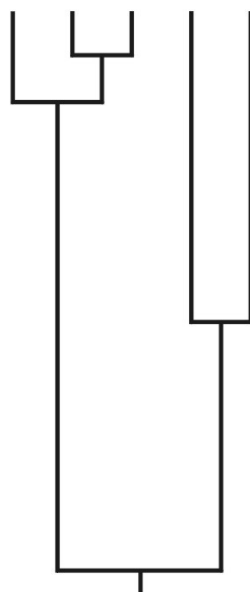
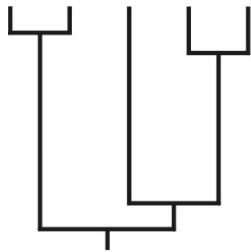
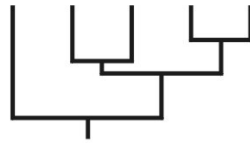
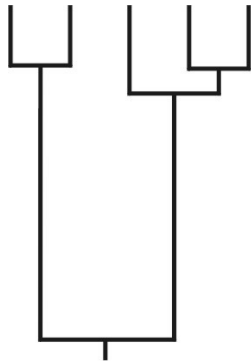


splynutí posledních  $k$  kopií zabere

$$\left(1 - \frac{1}{k}\right) \left(1 - \frac{1}{n}\right) \text{ generací}$$

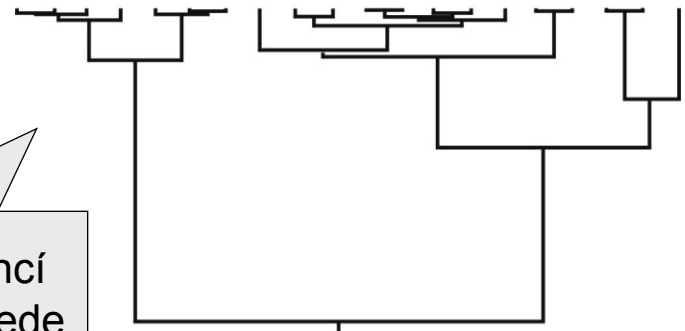
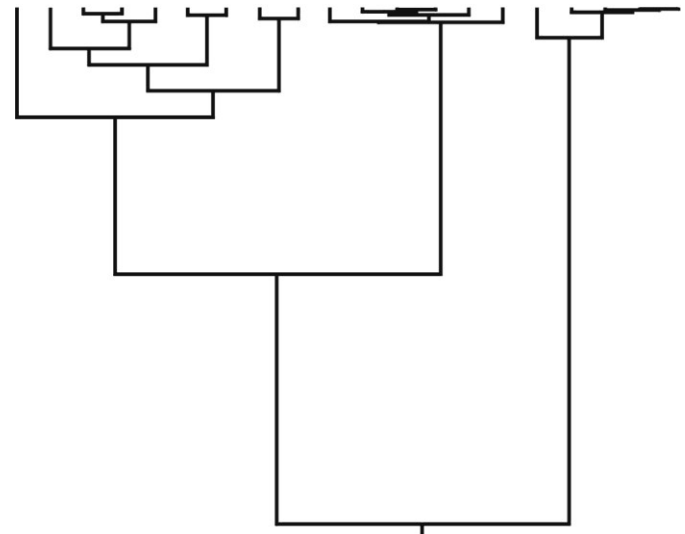
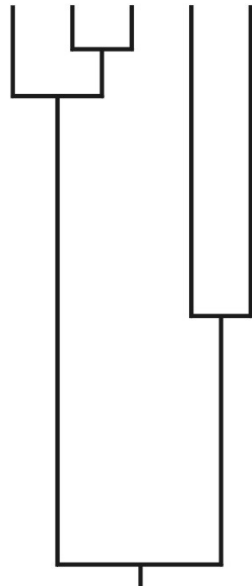
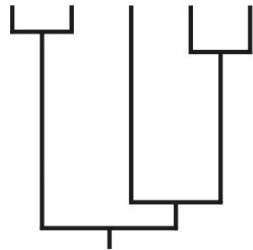
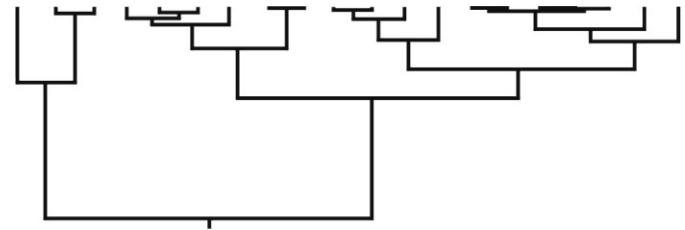
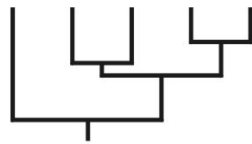
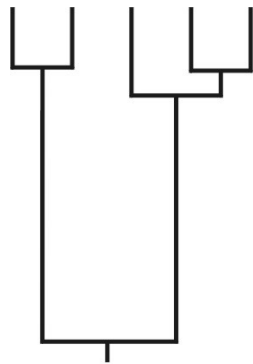
$\Rightarrow$  prvních 90 % kopií splyne během 9 % celkového času, zbývajících 91 % času se čeká na splynutí posledních 10 % kopií!

jestliže máme 100 kopií ve vzorku, pravděpodobnost, že přidáním 101. kopie dospějeme k hlubšímu kořenu, je pouze 0,02%  $\Rightarrow$  přidání další genové kopie pravděpodobně nepovede ke starší koalescenci



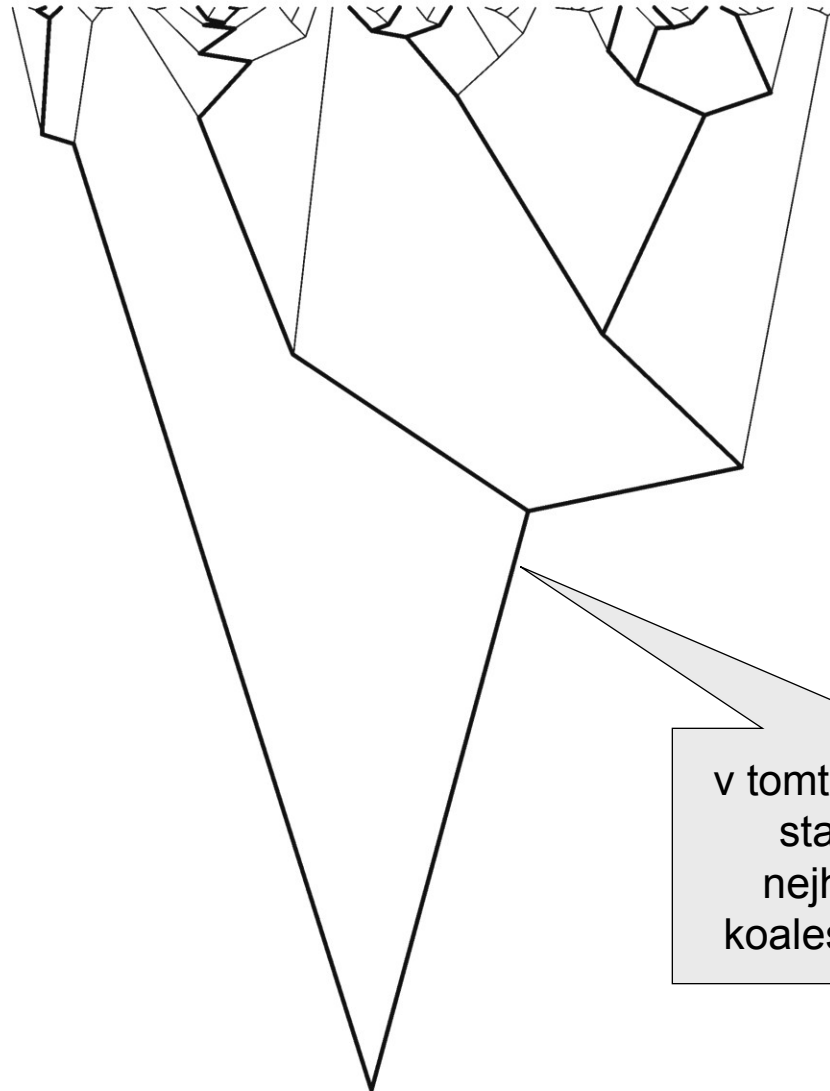
doba MRCA je vysoce variabilní...

s klesajícím počtem volných kopií se proces zpomaluje ...



přidání dalších sekvencí  
pravděpodobně nepovede  
k hlubší koalescenci ...

50 genových kopií, 10 náhodně vybraných:



v tomto případě 10 kopií  
stačilo k nalezení  
nejhlubšího kořene  
koalescenčního stromu

Pokud nás zajímají „staré“ koalescence, nepotřebujeme velké vzorky

např. pouhé 2 kopie poskytují v průměru 50 % koalescenčního času pro celou populaci!

Naopak pokud nás zajímá čas do první koalescence z  $n$  na  $n - 1$ , odhad  $N/[n/(n - 1)]$  je citlivý vůči  $n$

např. rozptyl průměrné doby první a poslední koalescence pro 10 genů je  $0,0444N$  až  $3,60N$ ; zvýšením  $n$  na 100 genů, rozmezí bude  $0,0004N - 3,96N$

zvýšením  $n$  10× se  
rozdíl zvýší 100× ...

... pro poslední  
koalescenci prakticky  
žádný rozdíl

**Z toho plyne, že pro odhady starých evolučních genových událostí stačí malé vzorky, pro odhady recentních událostí jsou velké vzorky nezbytné**

## Koalescence pro různé typy ploidie:

střední hodnota času koalescence (viz výše) pro velká  $n = 2xN_e$

(pro 2 kopie =  $2N_e$ )

autozomální lokus:  $4N_e$

chr. X (při poměru pohlaví 1:1  $\rightarrow x = 1,5$ ):  $3N_e$

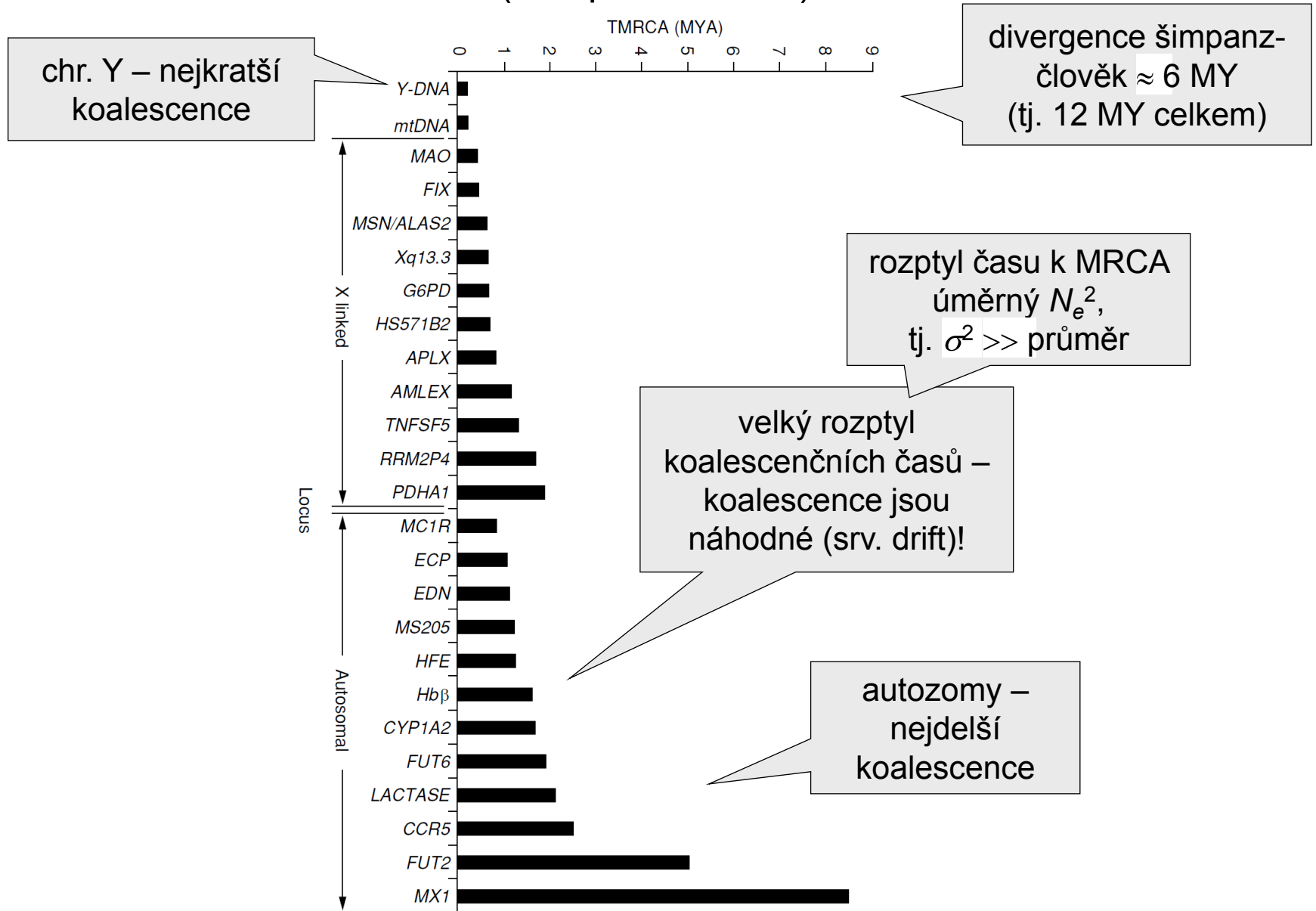
mtDNA, chr. Y (při poměru pohlaví 1:1):  $N_e$

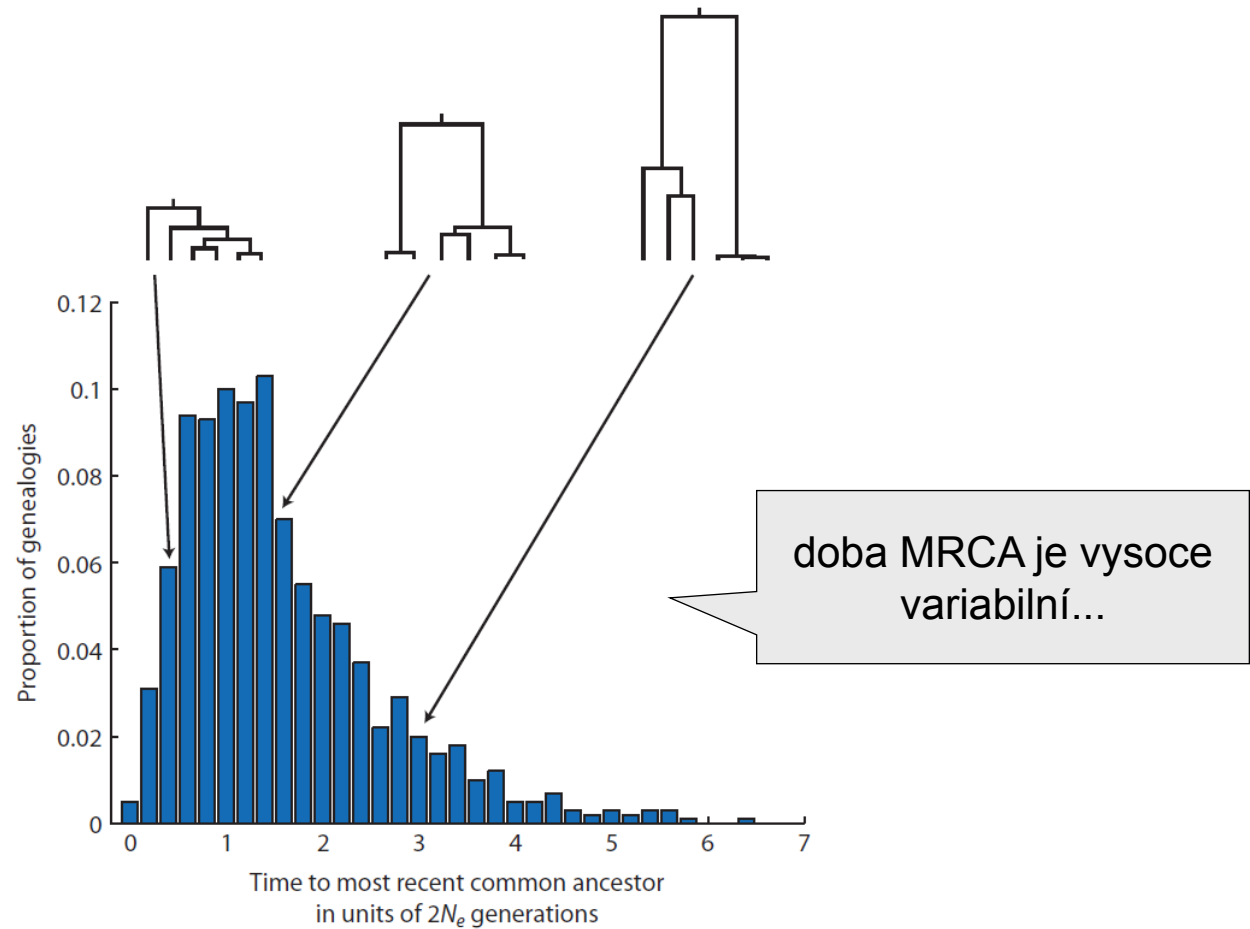
ALE: u savců většinou rozptyl reprodukční úspěšnosti samců  $>$  samic

$\Rightarrow N_e$  samců  $<$   $N_e$  samic

Z toho plyne, že ke koalescenci Y dochází dříve než u mtDNA  
a mnohem dříve než u ostatních lokusů (srv. mitochondriální  
Eva  $\times$  Y-Adam)

Př.: časy koalescence pro 12 autozomálních genů, 11 genů na X, mtDNA a Y u člověka (Templeton 2005):

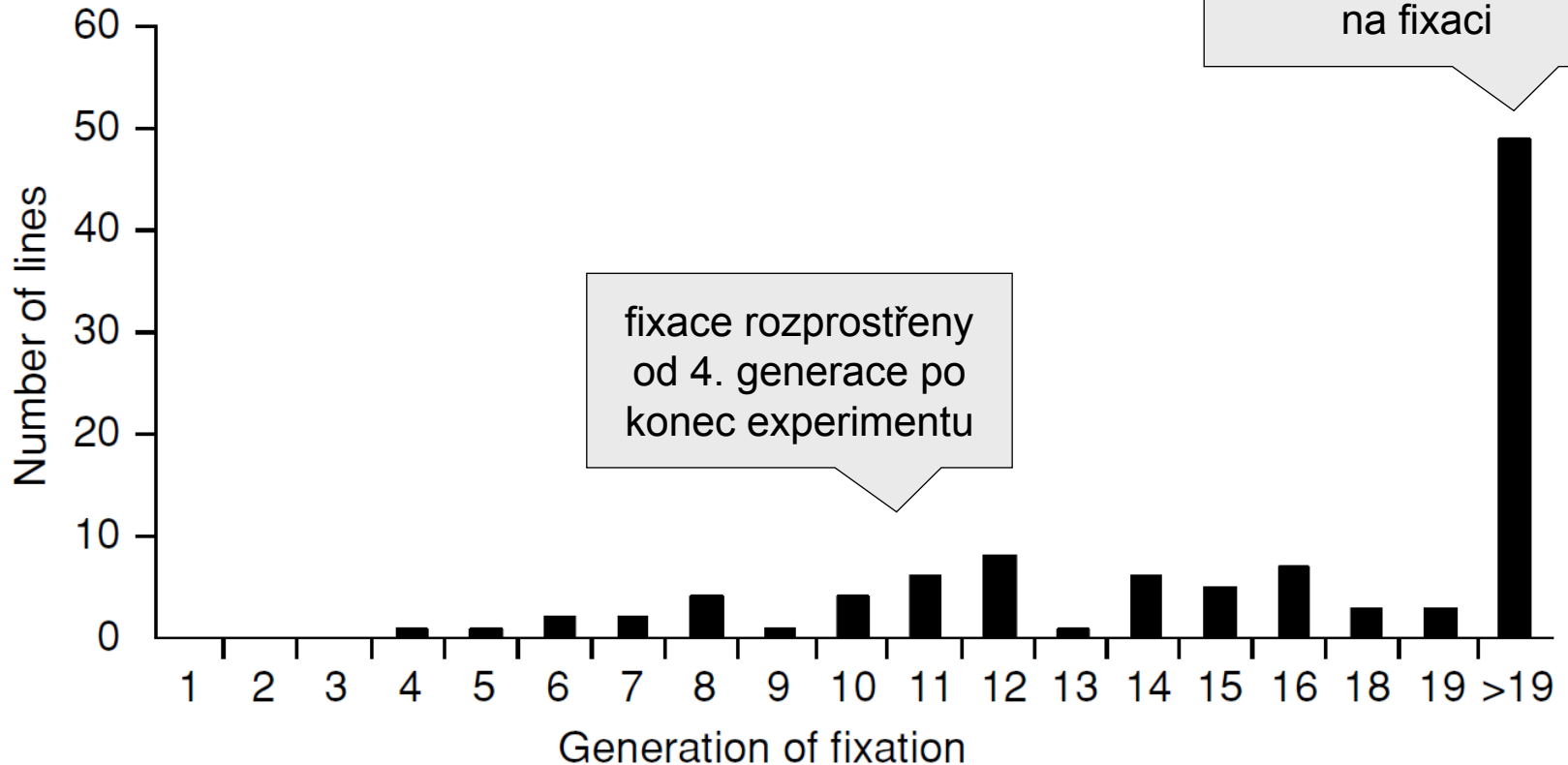




**Figure 3.27** The distribution of times to a MRCA (or genealogy heights) for 1000 replicate genealogies starting with six lineages ( $k = 6$ ). The distribution of total coalescence times has a large variance because the range of times is large and also asymmetric with a long tail of a few genealogies that take a very long time to reach the MRCA. The genealogies shown above the distribution are those for the tenth, fiftieth, and ninetieth percentile times to MRCA. In this example  $N_e = 1000$ .



Ještě jednou Buriho experiment (Buri 1956):



Přes identické podmínky velký rozptyl fixací; v přírodě jen 1 „experiment“  
→ musíme očekávat velkou „chybu“ v odhadech koalescencí, která  
nezávisí na velikosti vzorku ani na dalších typech chyb!

## Čas a koalescence:

V základní podobě je čas koalescencí měřen v jednotkách  $N$  generací

pokud existují rozdíly mezi příslušníky populace z hlediska jejich reprodukční úspěšnosti (rozptyl  $0 < \sigma^2 < \infty$ ), ale velikost populace je pořád stejná, koalescenční proces je měřen v jednotkách

$N/\sigma^2$  generací

**Vyšší rozptyl reprodukční úspěšnosti vede k rychlejším koalescencím (nižší  $N_e$ )**

Lineární změnou časové škály můžeme vzít v úvahu např. nediskrétní generace, oddělená pohlaví, odlišné systémy páření atd.

⇒ **můžeme využít i na reálné organismy**

## Při aplikaci na reálné organismy ale potřebujeme nezávislé odhady $N_e$ , $\sigma^2$ (popř. dalších parametrů)

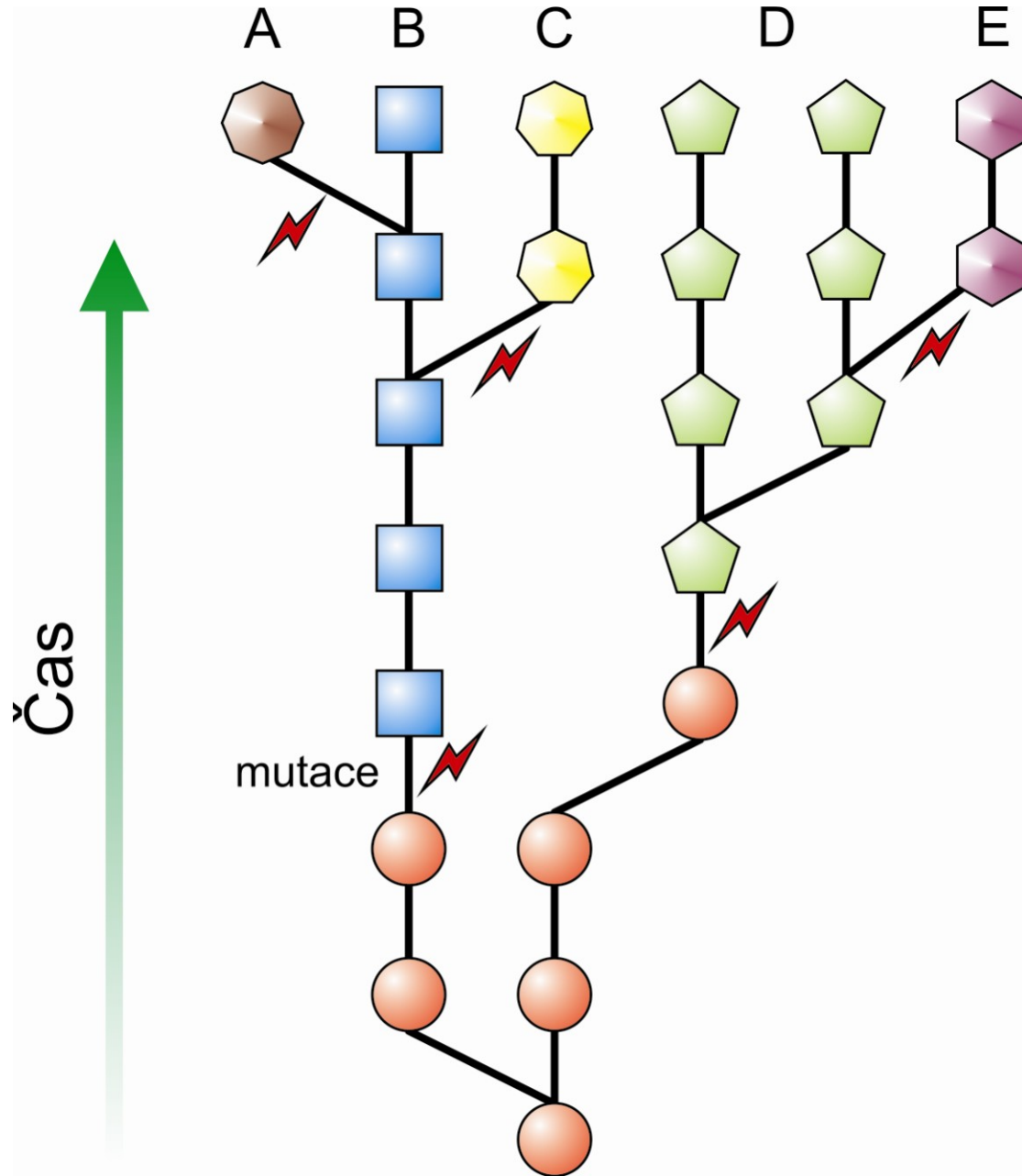
Tyto odhady nejsou vždy k dispozici, navíc jsou zdrojem další chyby (např. divergence sekvencí, datování fosilního záznamu, odhad generační doby atd.).

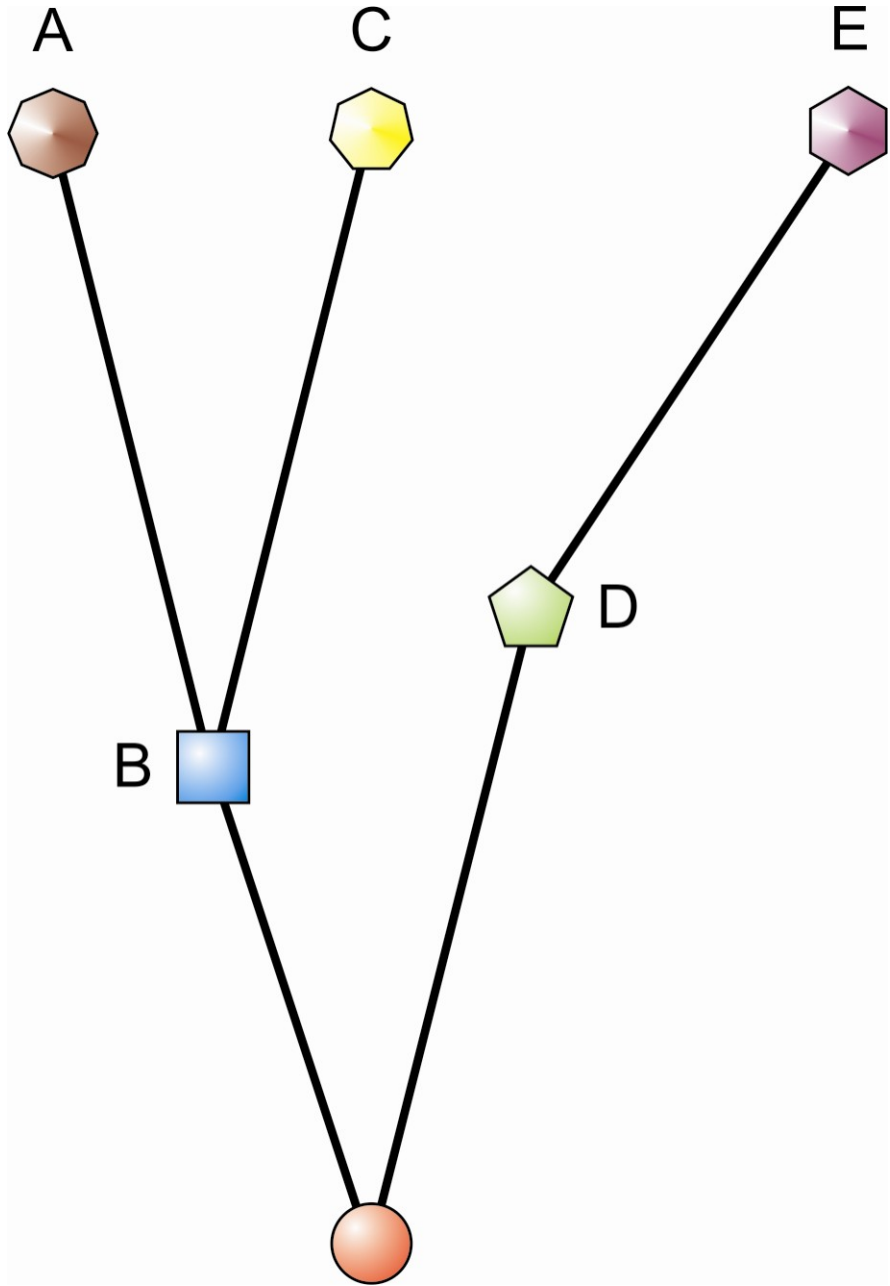
Navíc odhady koalescenčních časů často ignorují evoluční nahodilost fixačního/koalescenčního procesu (viz výše) – použité metody vycházející z NT jsou založeny na mezidruhové divergenci, tzn. považují fixaci alel za náhlý proces, který tím pádem nepřispívá do celkového rozptylu odhadu.

Např. odhad koalescence mtDNA člověka  $\approx 290\,000$  let  $\rightarrow$  vzhledem k rozptylu spojenému s molekulárními hodinami a koalescenčním procesem\*) 95% CI = 152 000 – 473 000 let (tj.  $> 300\,000$  let!)\*\*)

- \*) ignorujeme výběrovou chybu (*sampling error*), chybu měření (*measurement error*) a ne zcela přesně známou mutační rychlost  $\mu \Rightarrow$  ve skutečnosti by byl konfidenční interval mnohem větší
- \*\*\*) rozptyl v časech fixací/koalescencí způsobený driftem nelze odstranit jednoduše zvýšením velikosti vzorku ( $n$ ), protože populace představují pouze jednu realizaci evolučního procesu (můžeme ale použít víc lokusů)

**Při datování evolučních událostí pomocí vnitrodruhové molekulární variability bychom si měli být těchto skutečností vědomi!**





## Koalescence a neutrální mutace:

Jestliže  $\mu$  je velmi malé a  $N_e$  velmi velké, můžeme výskyt obou jevů během jedné generace zanedbat.

Pokud jdeme zpět v čase, dokud 2 kopie buď nesplynou, nebo jedna z nich nezmutuje, pak pravděpodobnost, že k mutaci došlo dřív než ke koalescenci, je

podmíněná  
pravděpodobnost

$$\Pr[\text{mutace před koal.} | \text{mutace, nebo koal.}] \approx \frac{\theta}{\theta + 1}$$

kde  $\theta = 4N_e\mu$

## Koalescence a neutrální mutace:

$$\frac{\theta}{\theta + 1}$$

≈ očekávaná heterozygotnost při náhodném oplození –  
jestliže dojde k mutaci před koalescencí, pak obě kopie  
musí reprezentovat odlišné alely (*infinite-alleles model*)

⇒ Při pohybu v čase dopředu i zpět je dopad rovnováhy  
driftu a mutace na genetickou variabilitu stejný.



# Pomocí koalescence můžeme definovat efektivní velikost populace $N_e$

~ velikost idealizované populace, která má stejný koalescenční čas jako skutečná populace

Takto definovaná  $N_e$  součástí softwarů založených na koalescenci

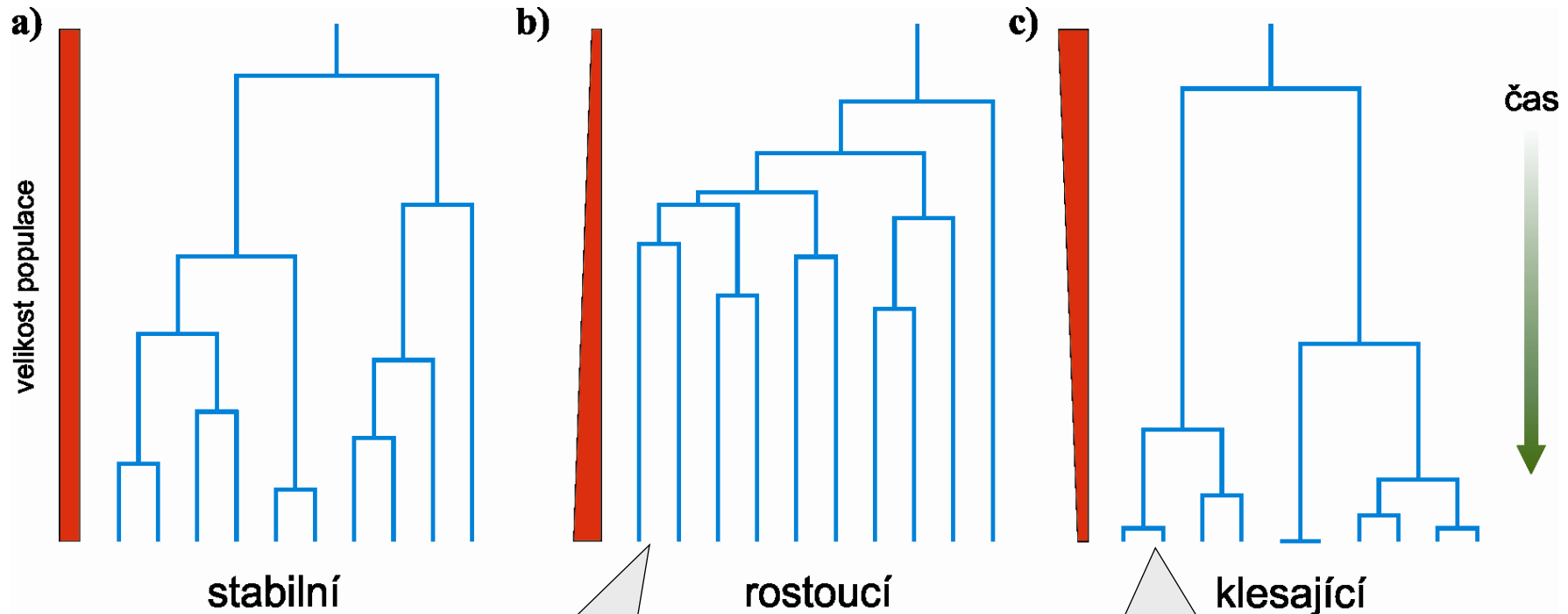
Např. generační čas vyjádřen v jednotkách  $4N_e\mu$ , kde  $N_e$  = koalescenční efektivní velikost populace

Pozor: také v tomto případě rozdíly mezi geny:

oblasti s nízkou rekombinací → nízká  $N_e$

oblasti s vysokou rekombinací → vysoká  $N_e$

# Vliv změn velikosti populace na tvar koalescenčního stromu

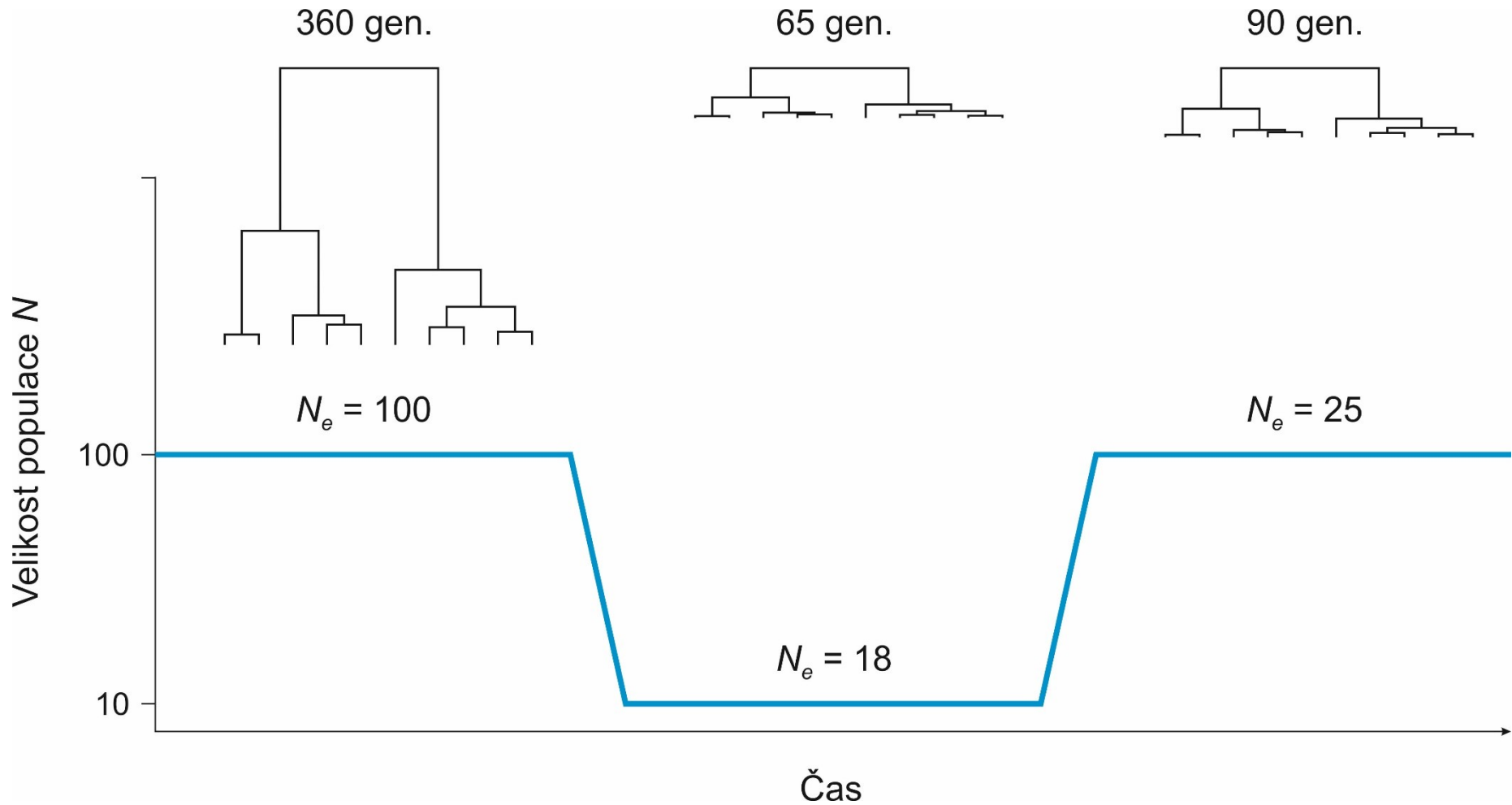


rostoucí populace:  
koalescence se  
postupně zpomalují

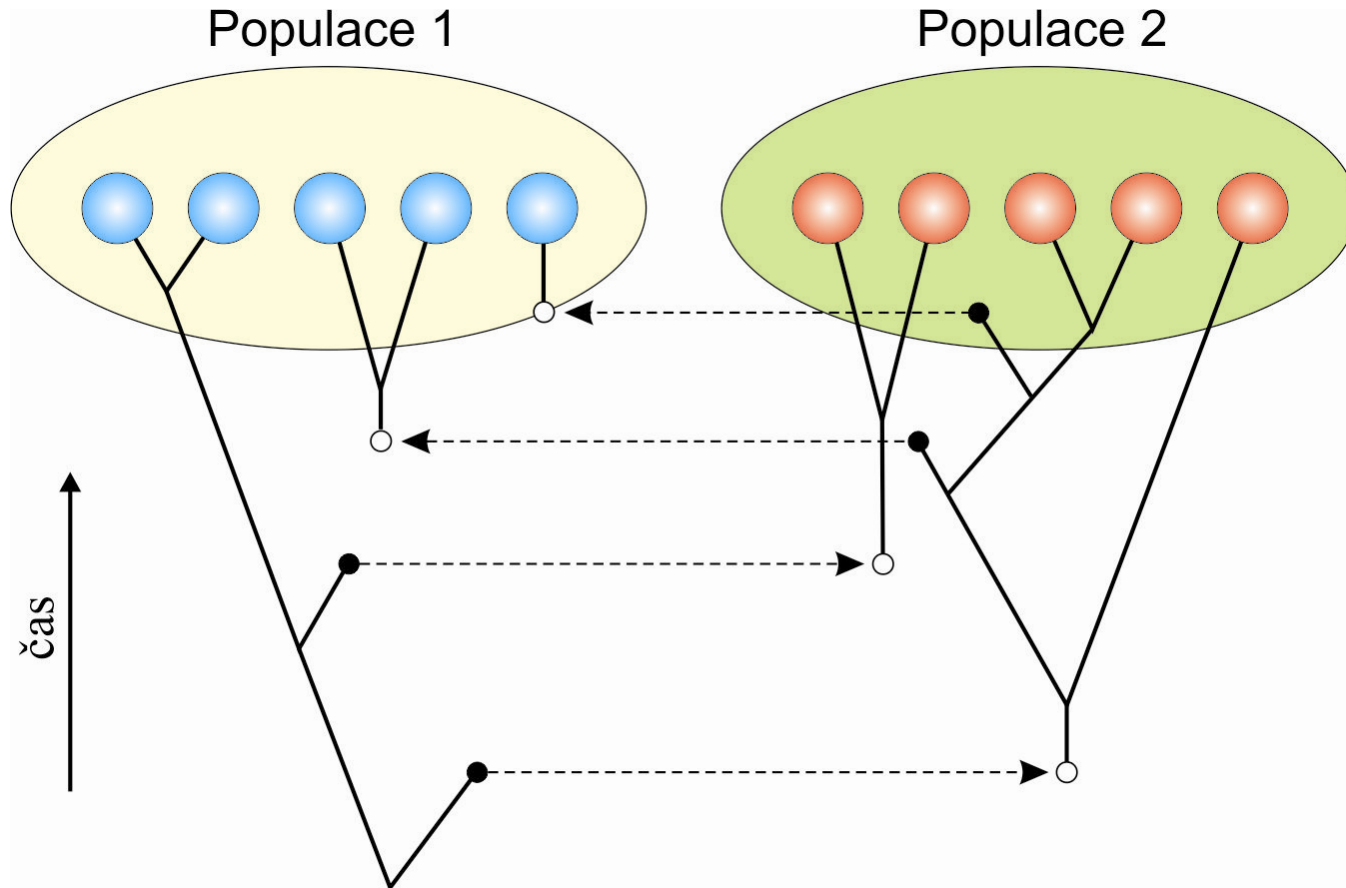
zmenšující se populace:  
koalescence se postupně  
zrychlují

# Vliv bottlenecku na koalescenci:

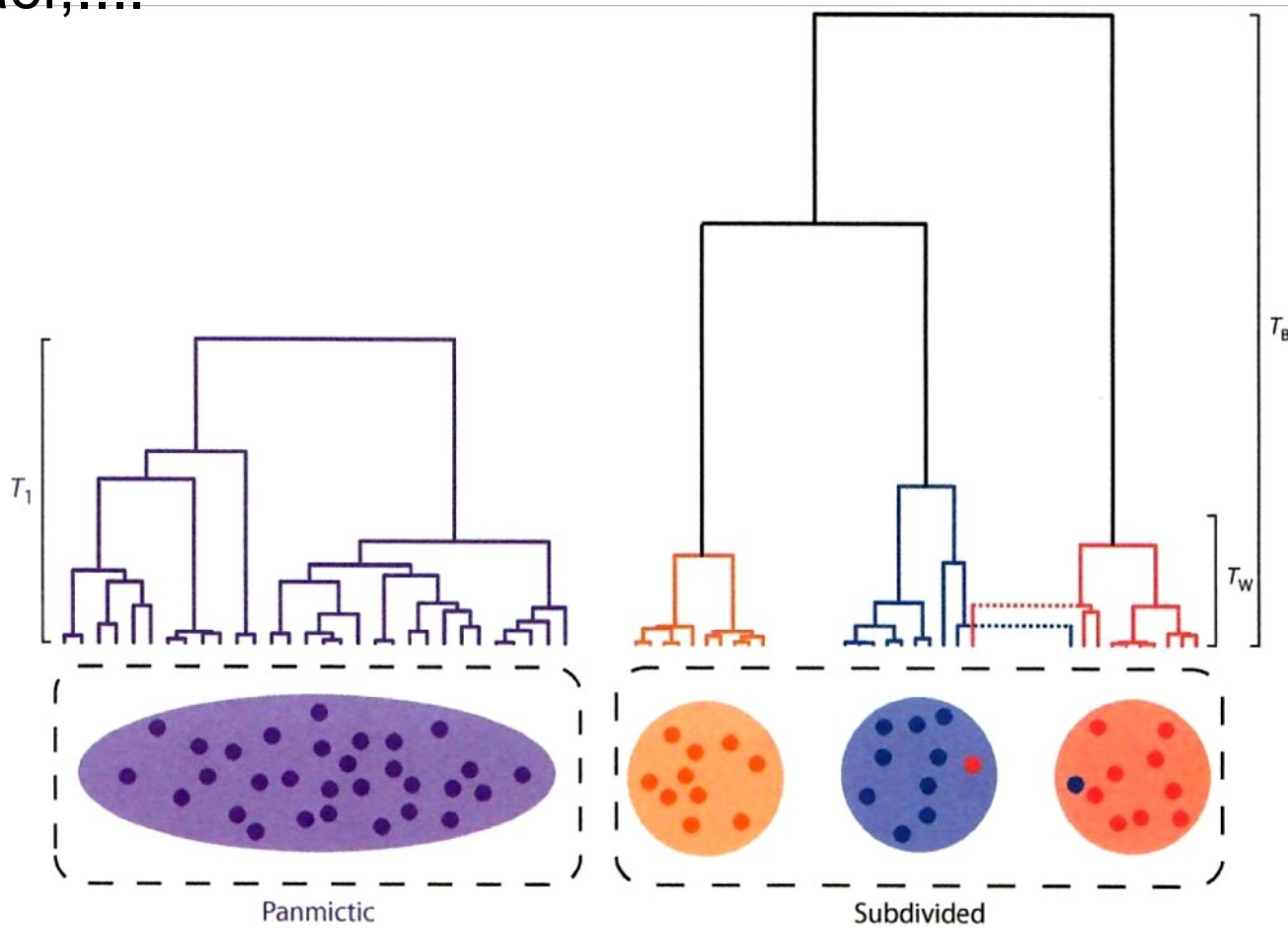
$n = 10$



# Koalescence a tok genů:

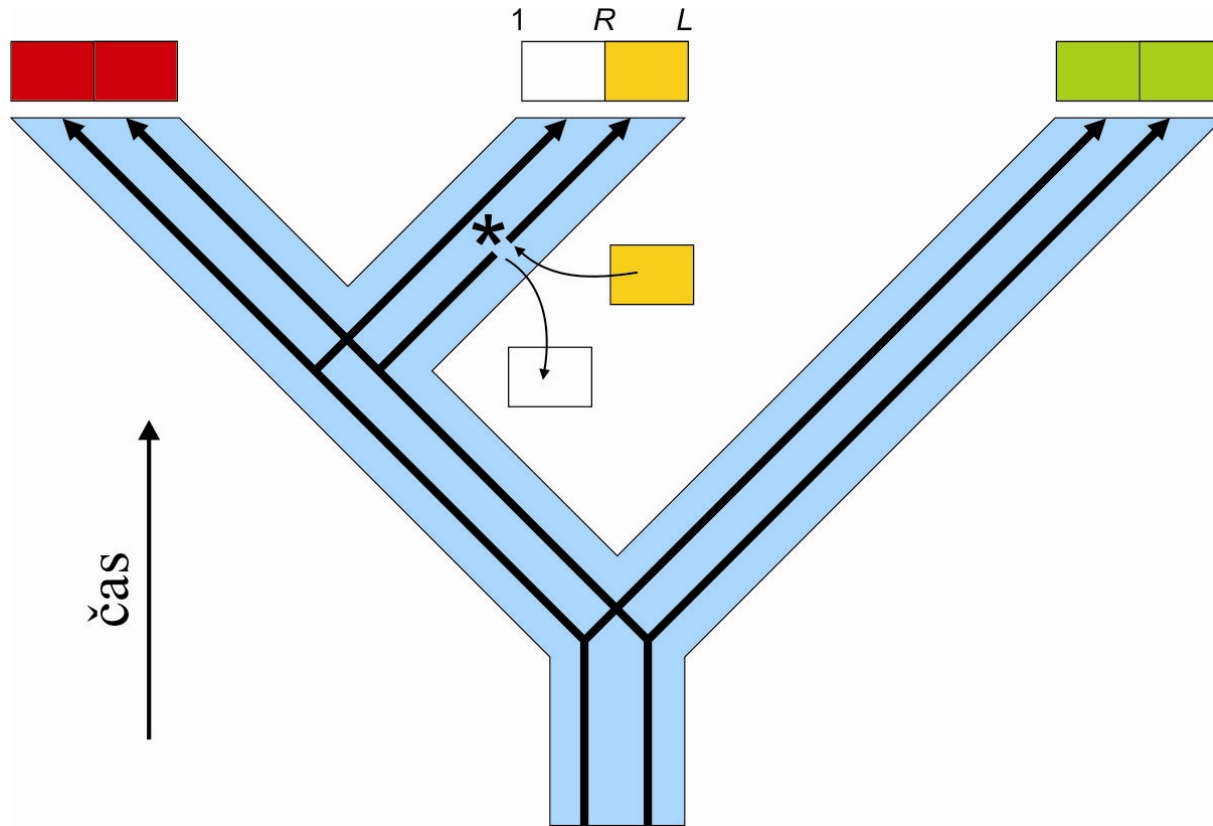


Slabá migrace vede k většině koalescencí uvnitř lokálních populací,.....

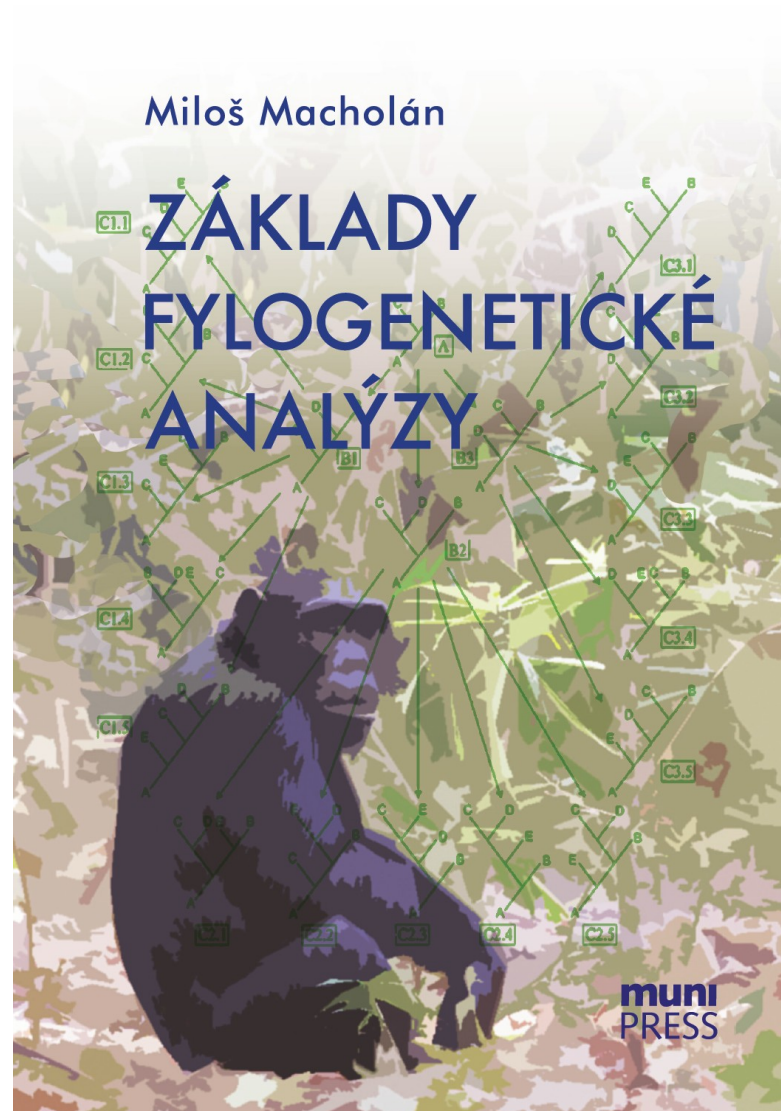


.... k prodloužení času k MRCA a zvýšení jeho rozptylu.

# Koalescence a rekombinace:



více podrobností viz



vliv selekce na koalescenci viz Selekcce