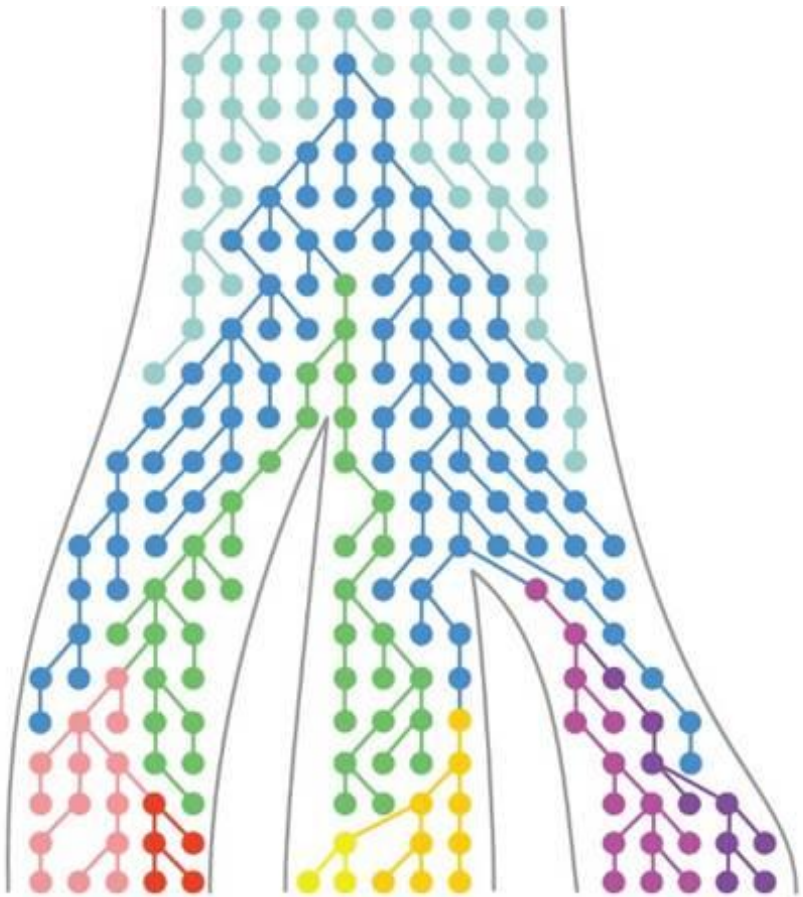
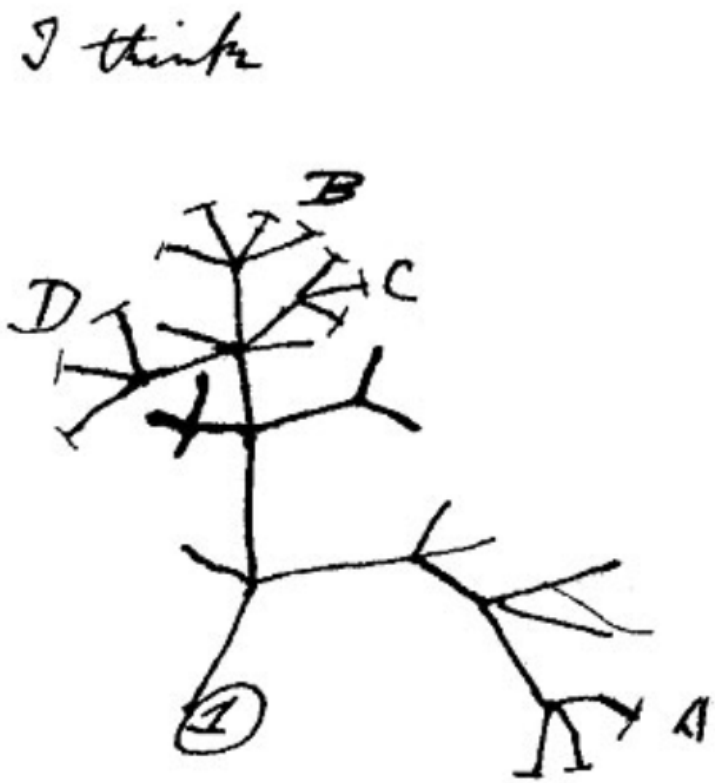
Genové a druhové stromy (genealogie a fylogeneze)

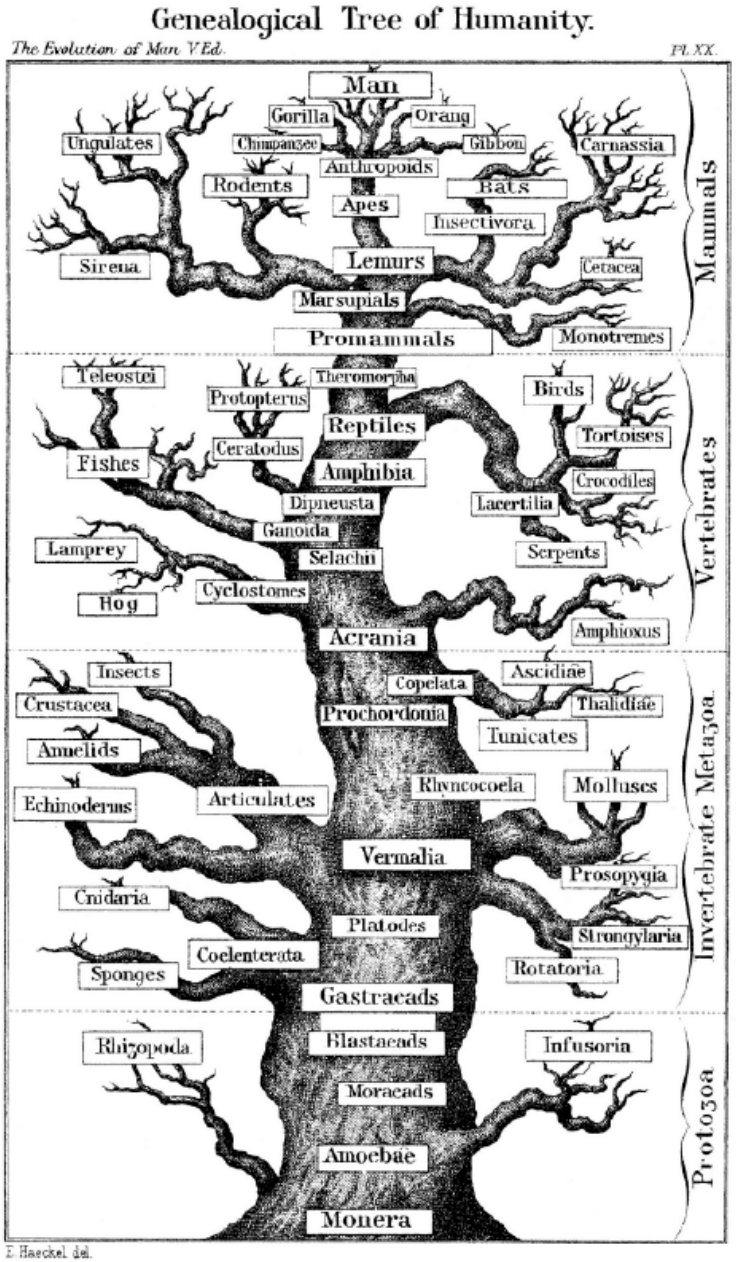


idea fylogenetického stromu (tj. evolučního), stromu života:

Darwin, 1859:
(mmch. jediný obrázek v jeho knize o původu druhů)



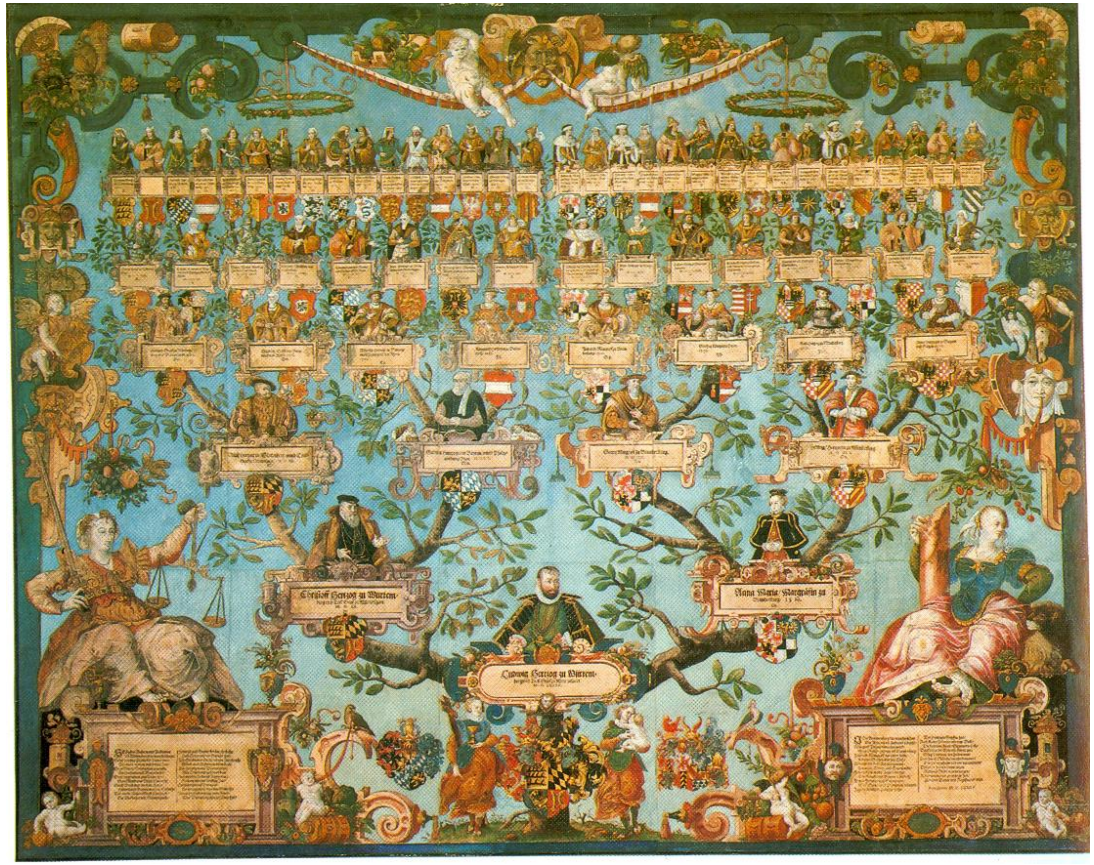
=> v této přednášce jsme na mezidruhové úrovni

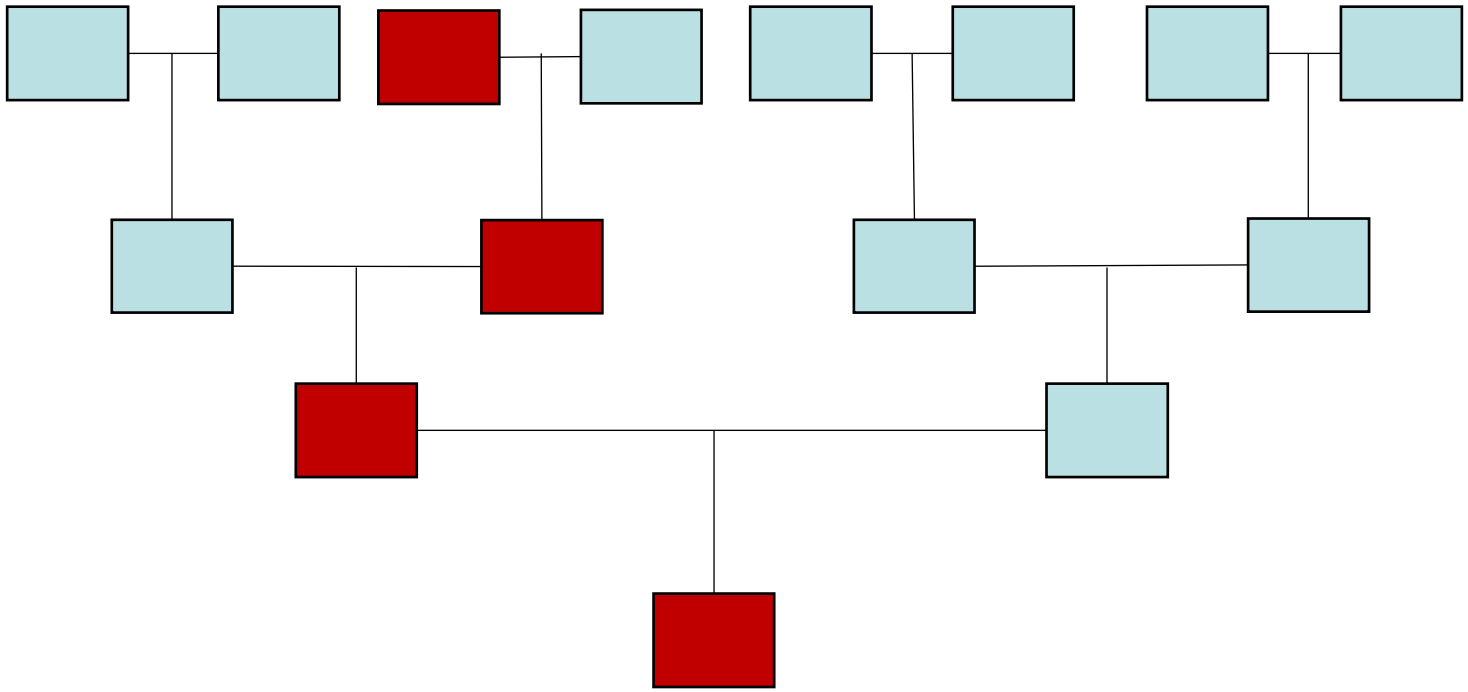


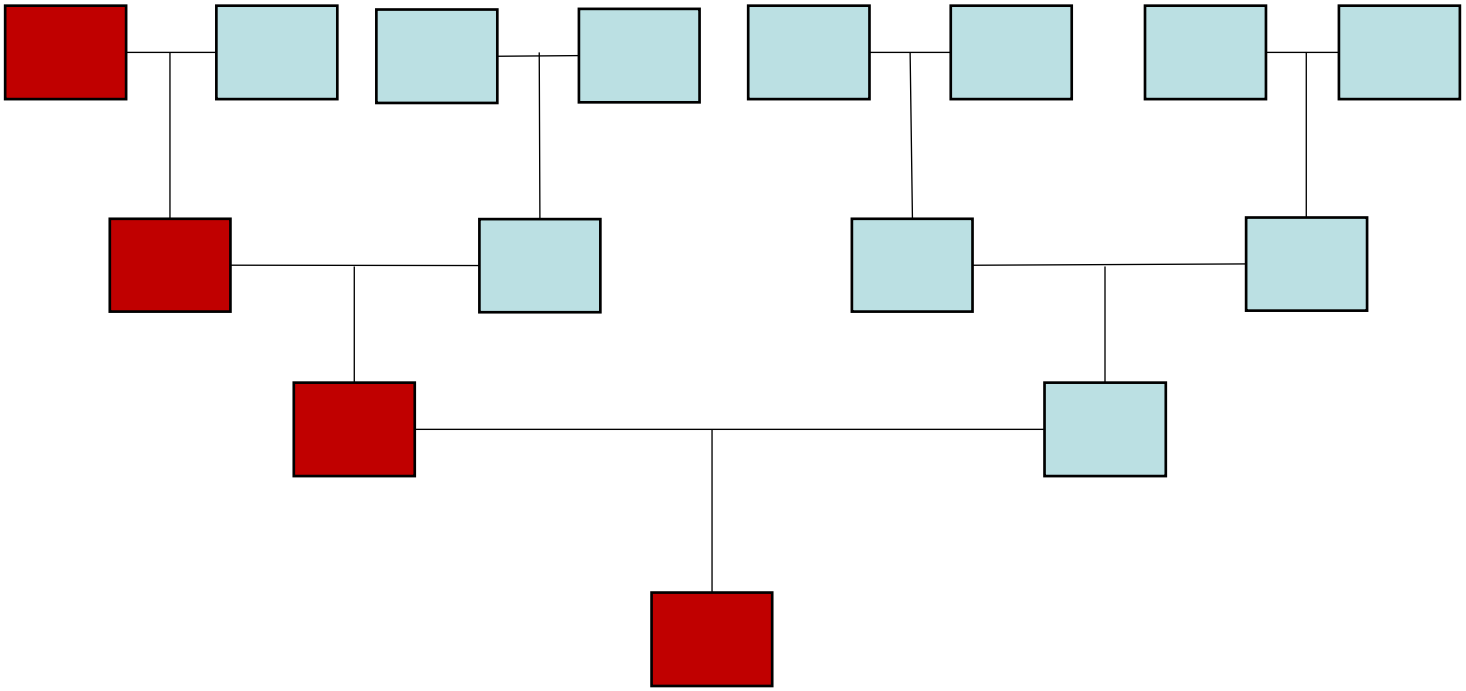
Heckel, 1879: Tree of Life

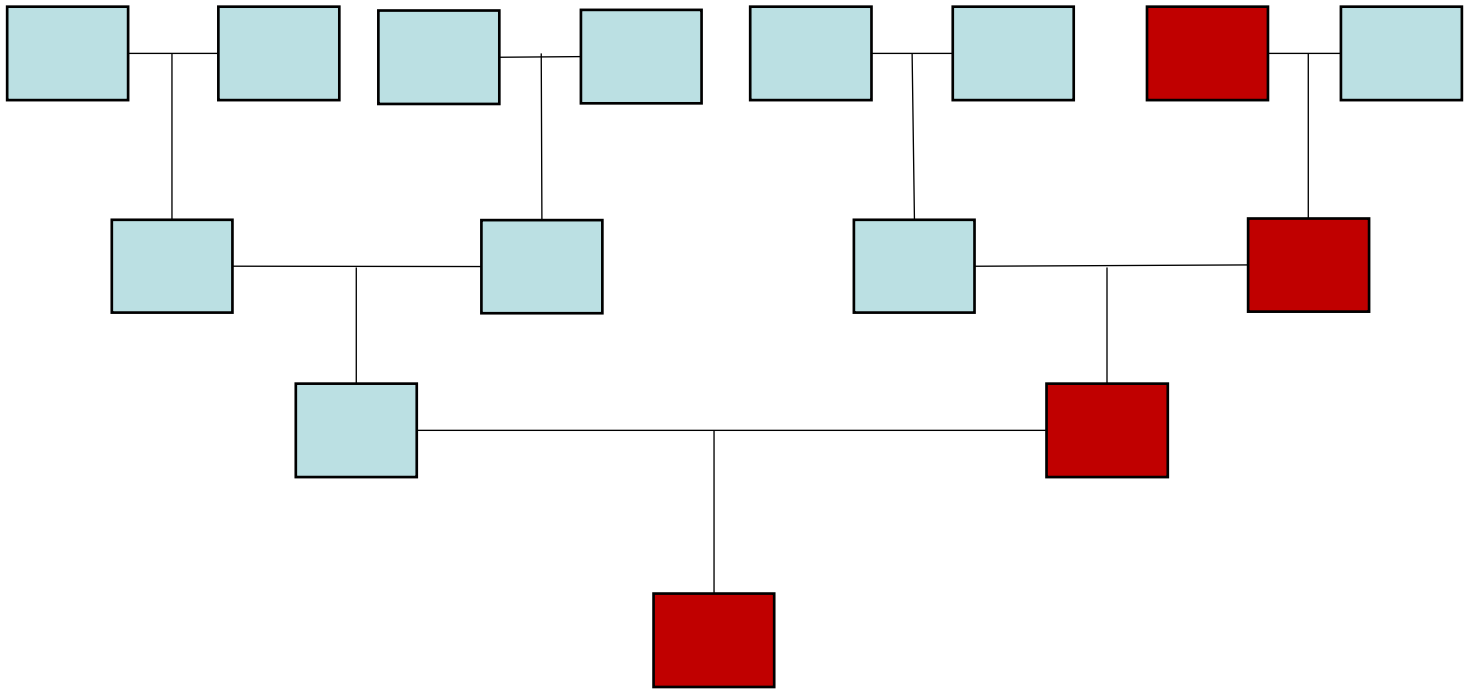
Genové stromy (genealogie)

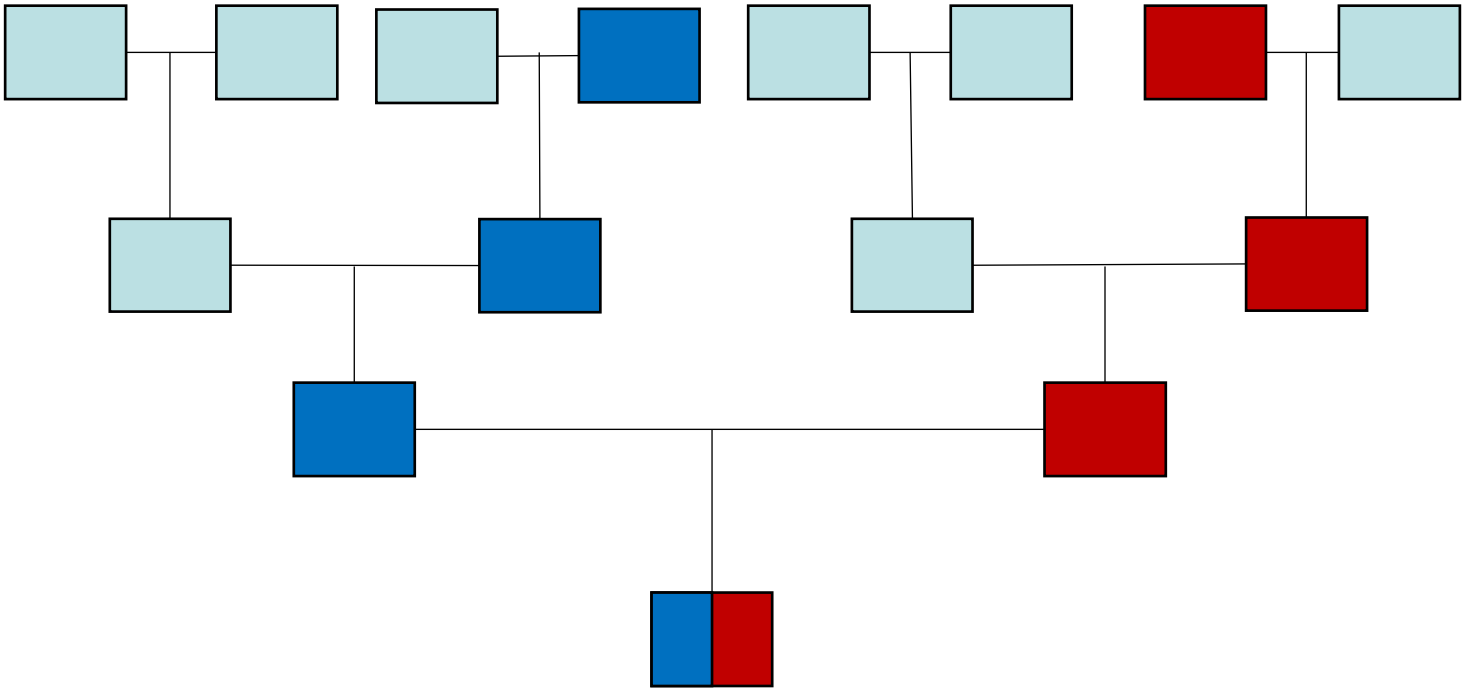
- Rodokmeny jednotlivých kopií určitého genu v populaci.
- Popisují vztahy mezi kopiemi určitého genu v populaci napříč generacemi.



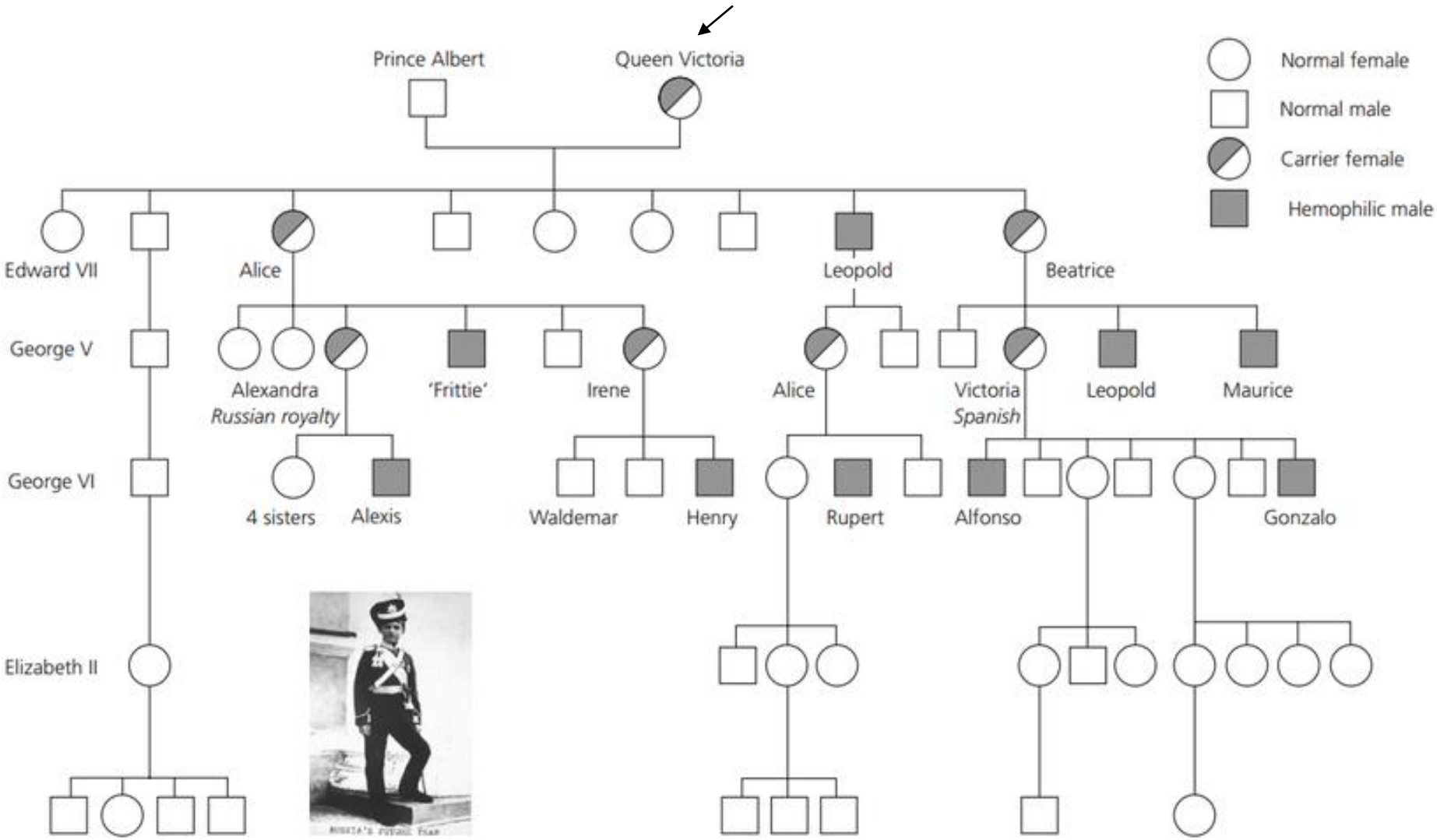






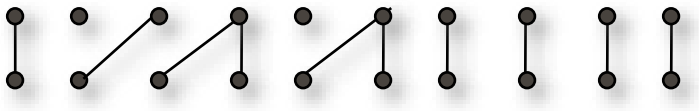


Poslední společný předek **„Most Recent Common Ancestor“ (MRCA)**

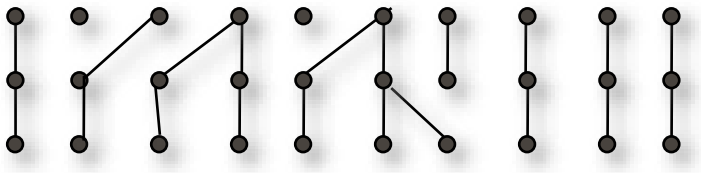


Dědičnost hemofilie v sasko-koburské dynastii

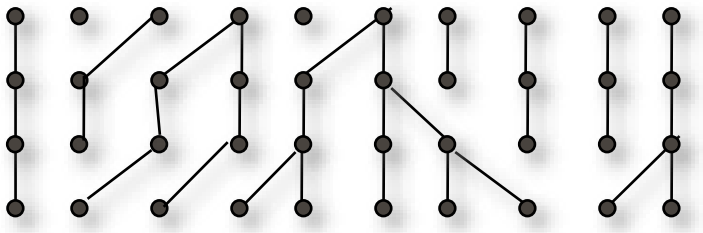
Genové genealogie



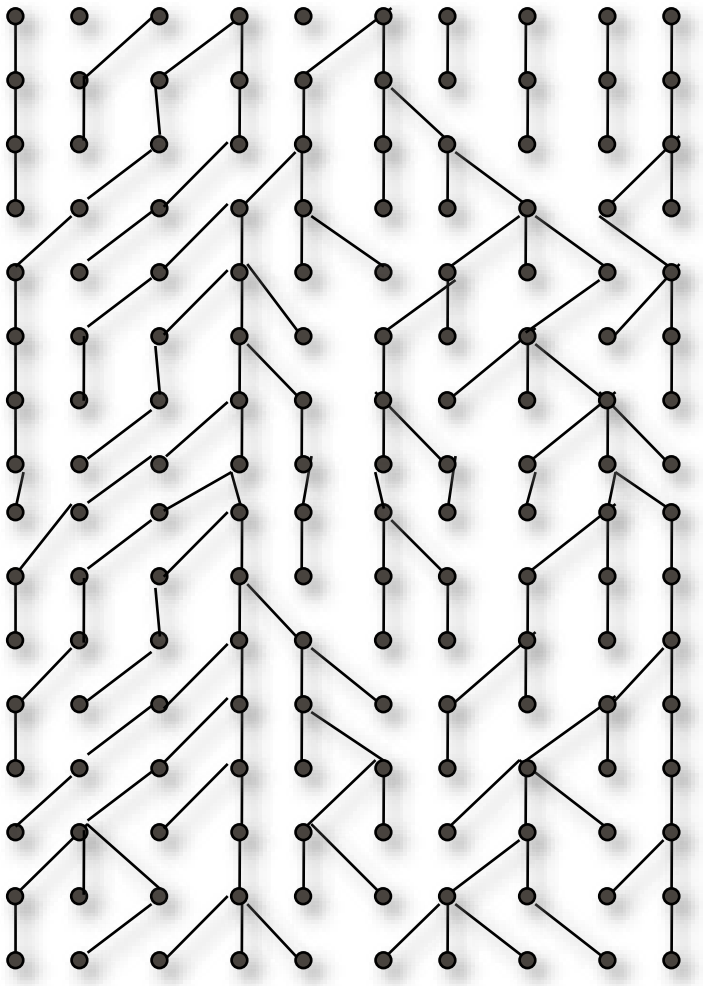
Genové genealogie



Genové genealogie

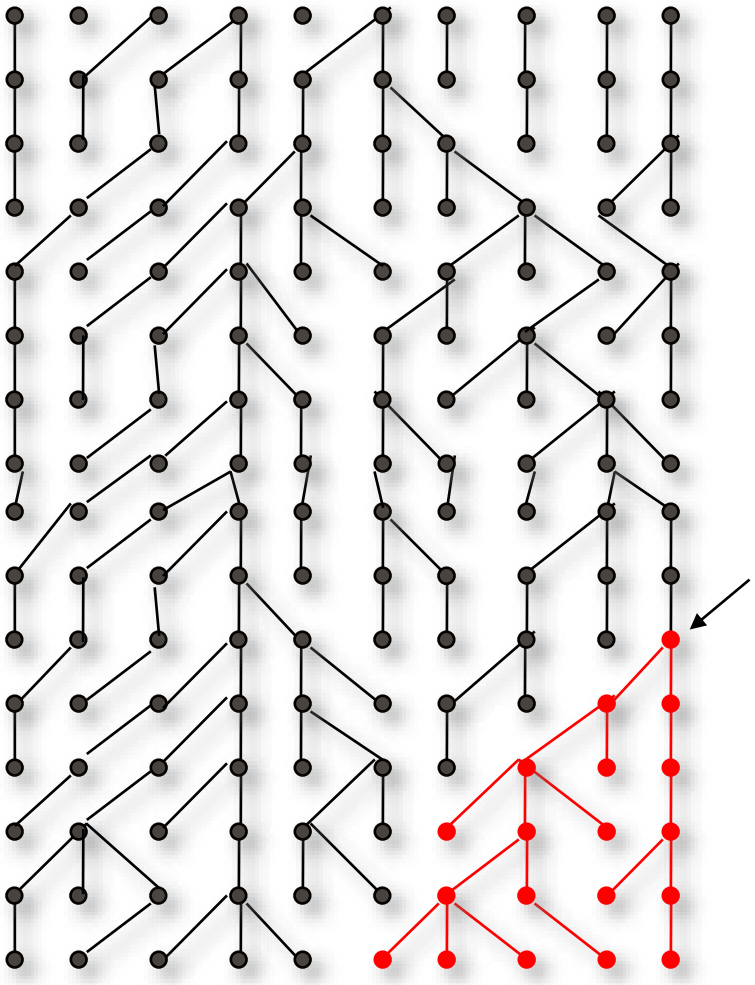


Genové genealogie



Koalescence
= splývání linií

Genové genealogie



Koalescence
= splývání linií

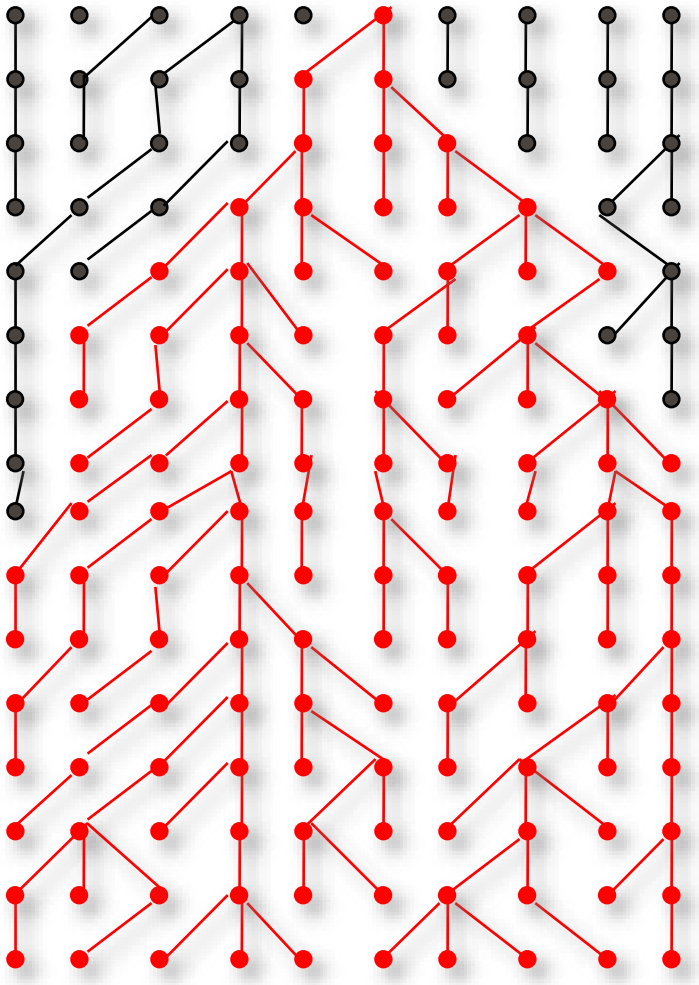
Most recent common ancestor (MRCA)
Poslední společný předek

Genové genealogie

Mitochondriální Eva

= Poslední společný předek pro současnou lidskou mt DNA

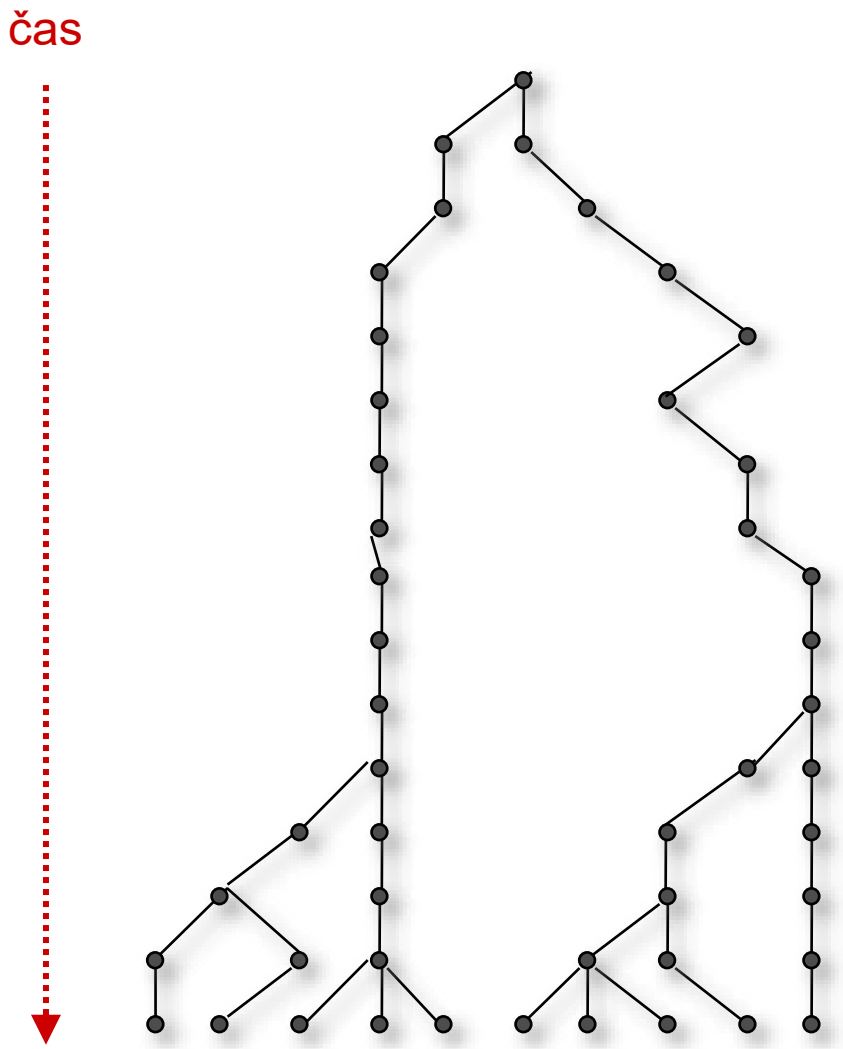
Poslední společný předek (MRCA)



Adam chromosomu Y

= Poslední společný předek pro současný lidský chromosom Y

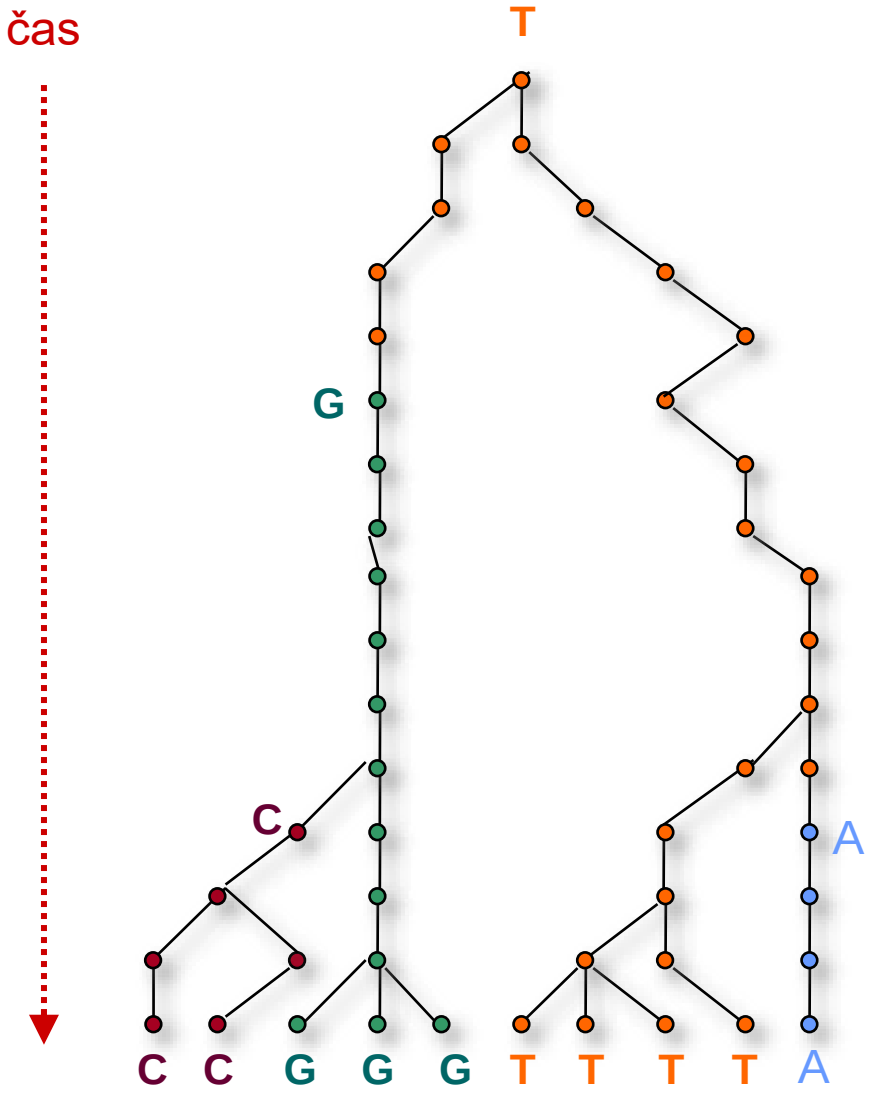
Genové genealogie



Vycházíme ze současnosti a sledujeme, co se dělo s populací v minulosti

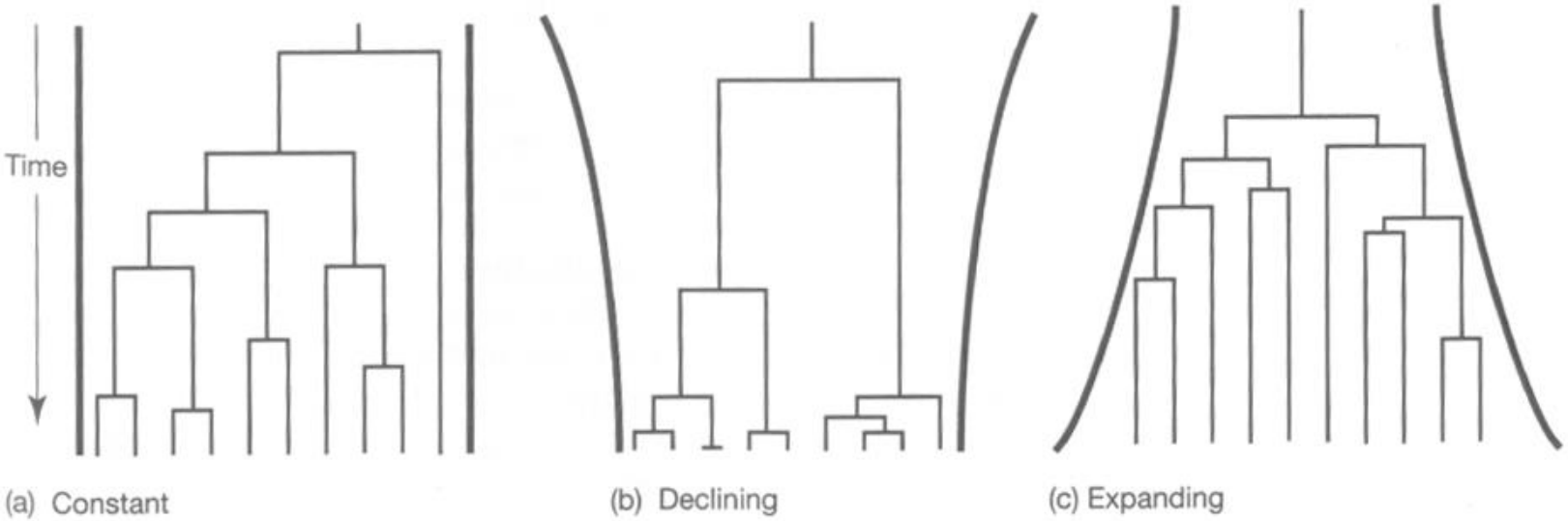
Genové genealogie

Můžeme rekonstruovat díky mutacím

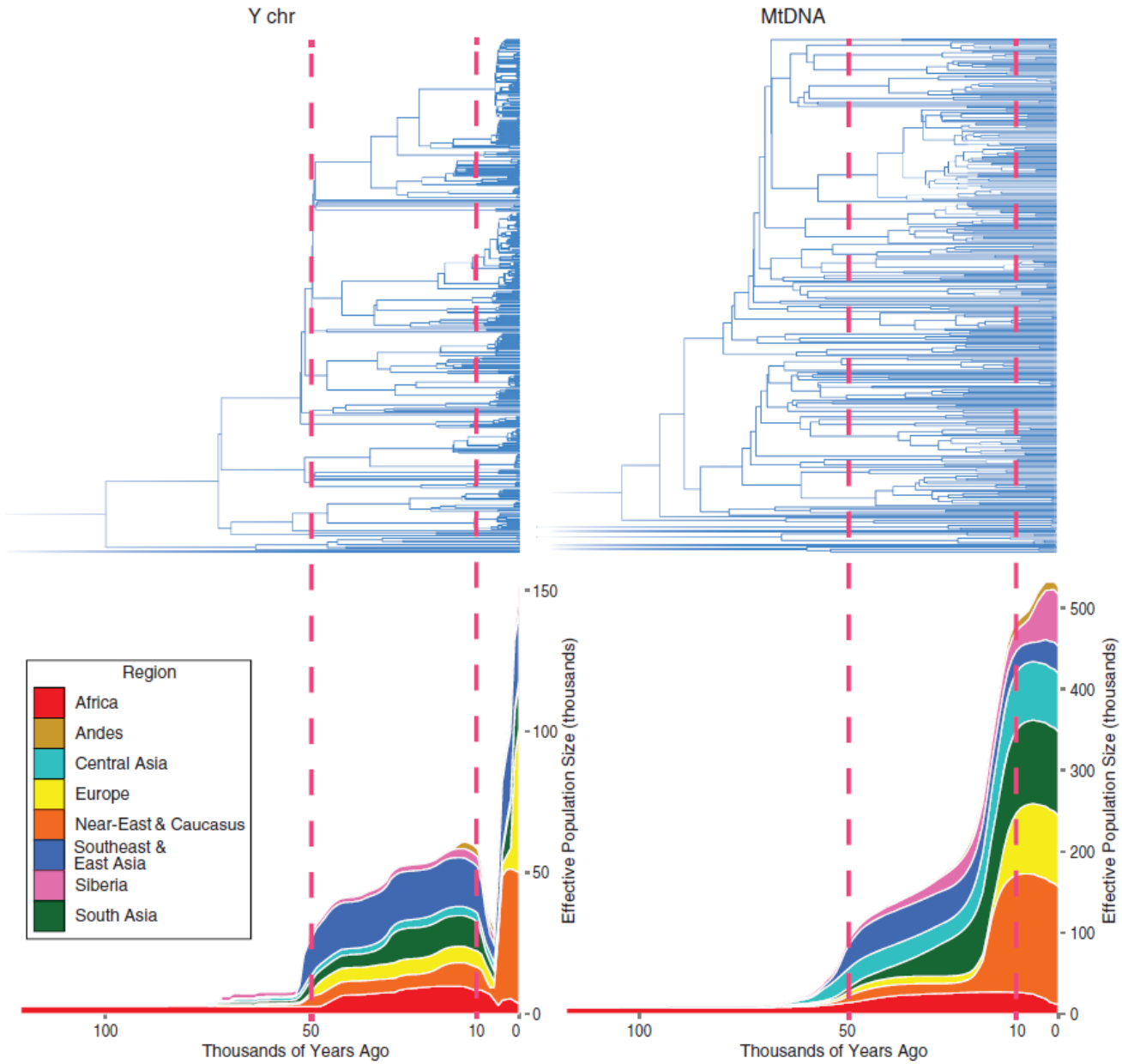


Průběh genealogií je ovlivněn demografickými faktory

změny ve velikosti populace



Z průběhu genealogií lze zjistit změny ve velikosti populací v minulosti



Karmin et al. 2015

Bayesian Skyline Plot

- Změny velikostí populací

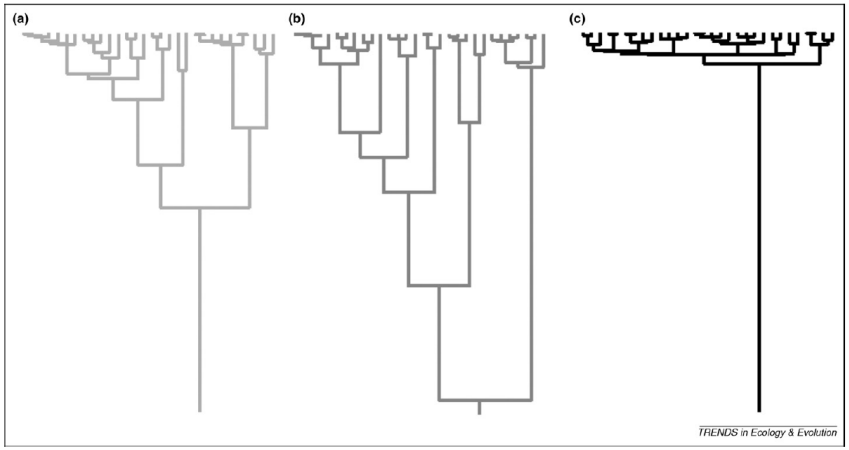
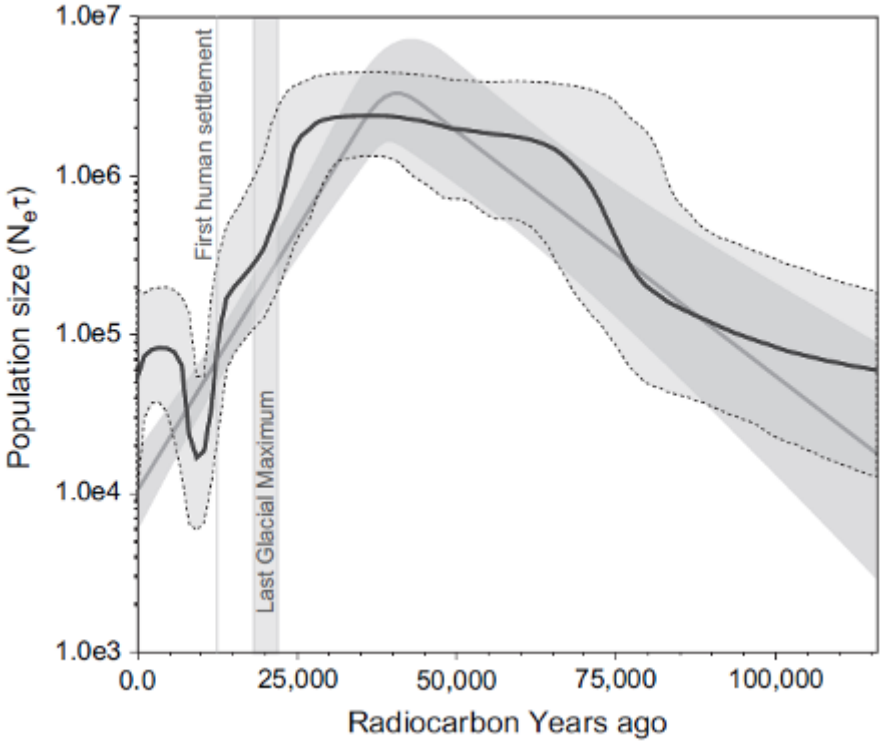
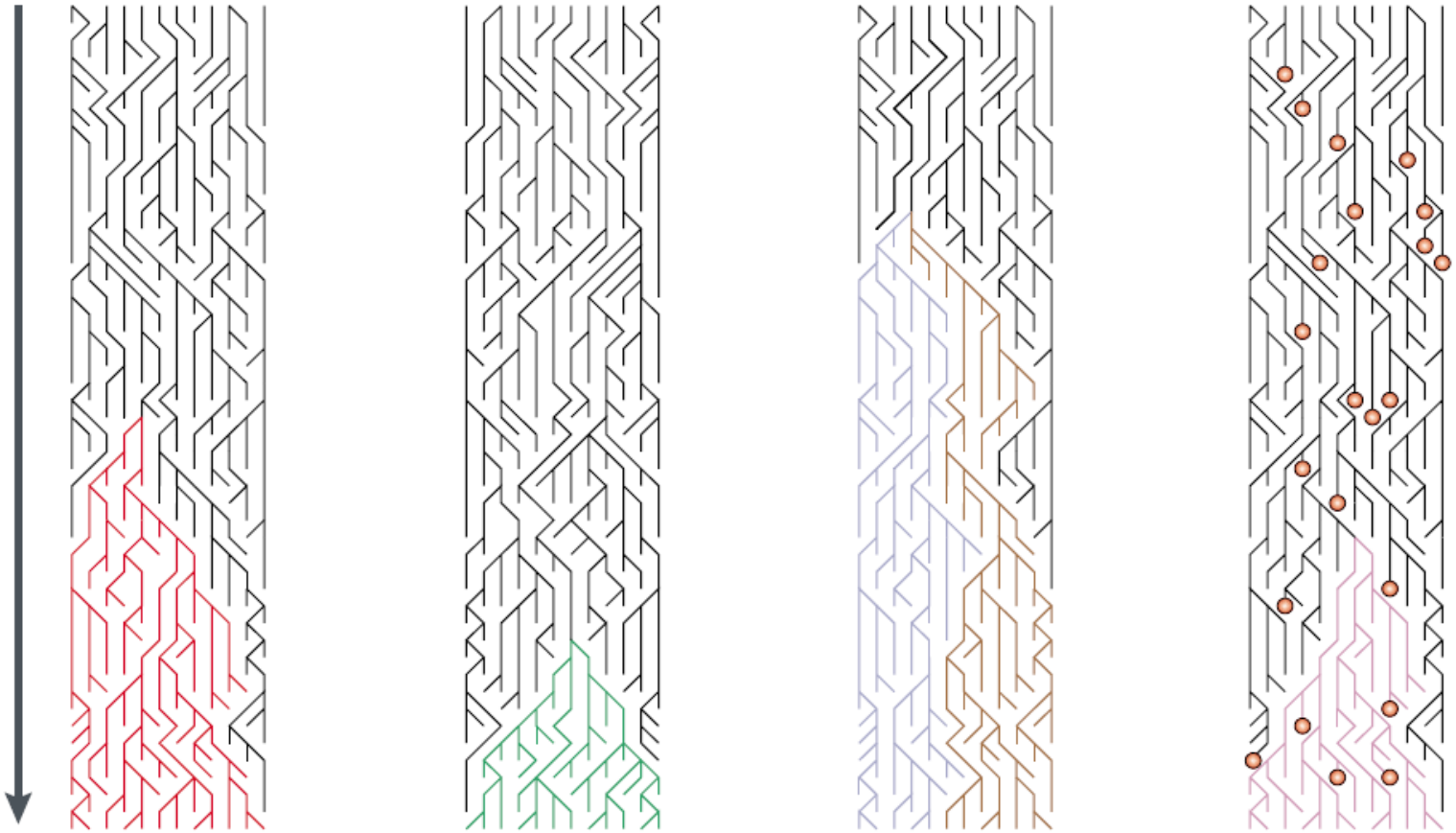


Figure 2. Growth signature in genealogies. Genealogies sampled from (a) constant-size, (b) exponentially shrinking and (c) exponentially growing populations.



Bison priscus

Genealogie a selekce



Neutrální

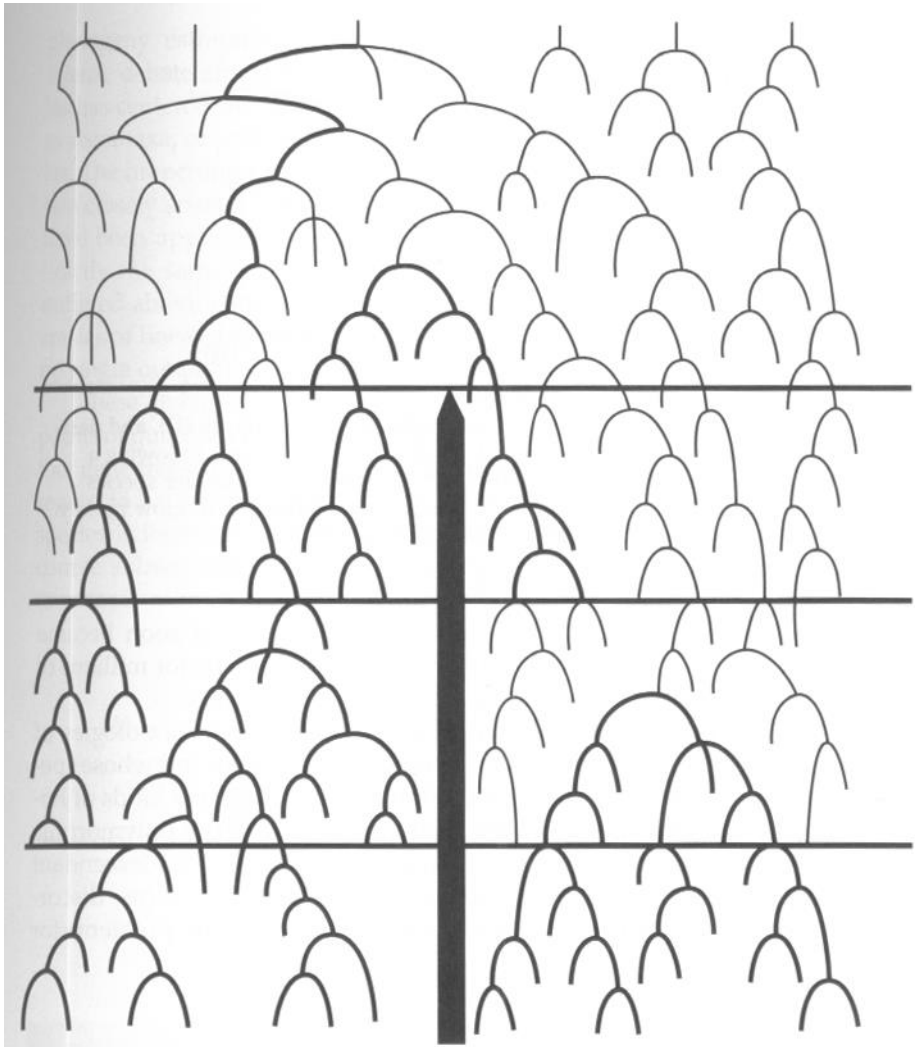
Pozitivní selekce

Balancing selekce

Negativní selekce

Sortování linií

ancestrální polymorfismus



druh 1 druh 2

polyfilie

parafilie

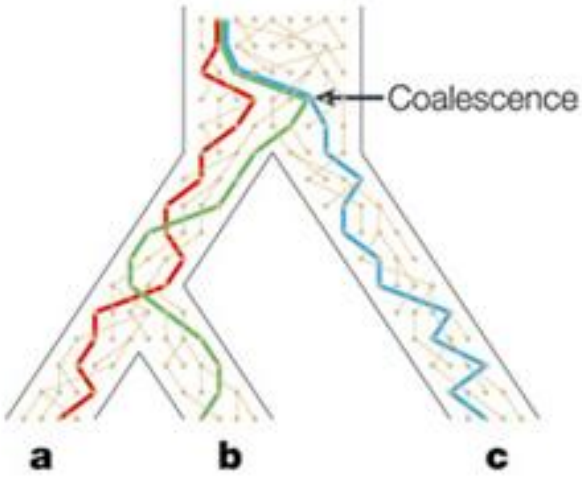
monofilie

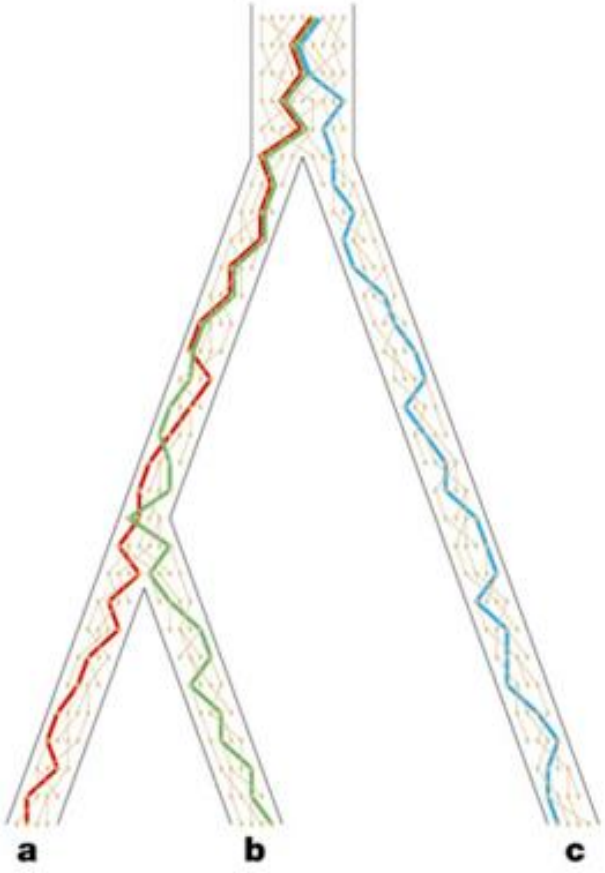
Průměrná doba fixace neutrální mutace

$t = 4N_e$ generací

pro chrX: $t = 3N_e$
pro chrY, mt DNA: $t = 1N_e$

- U druhů s velkou N_e přetrvává neutrální polymorfismus delší dobu.
- Pomalejší sortování genealogií





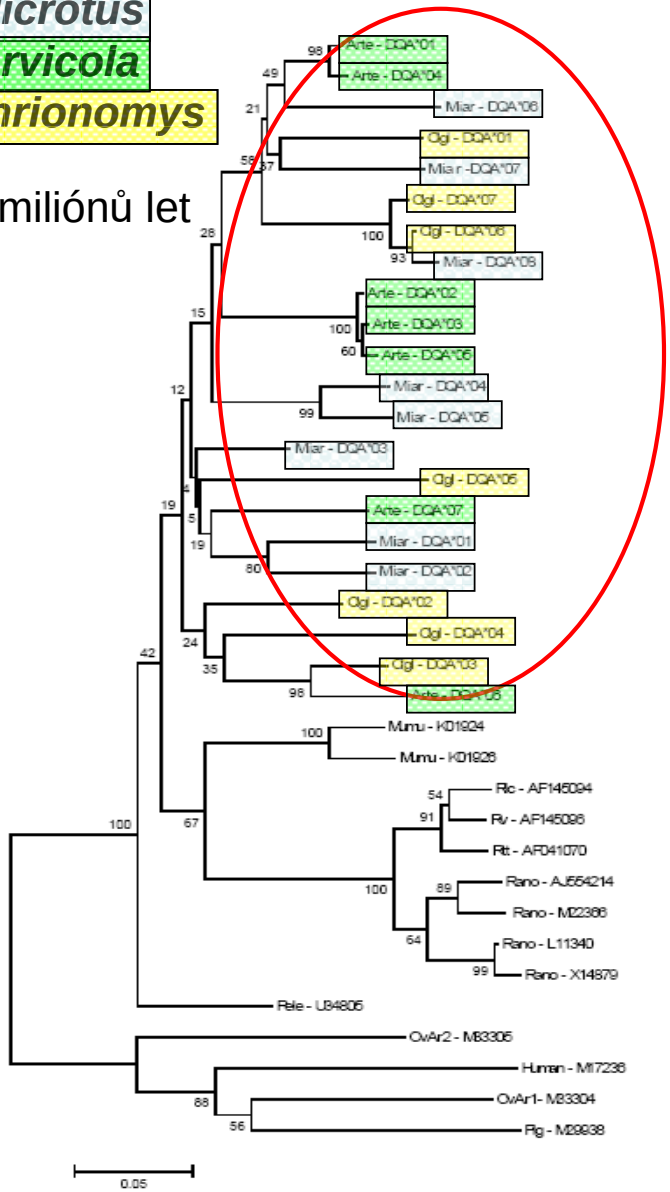
Trans-species polymorfismus:

Microtus

Arvicola

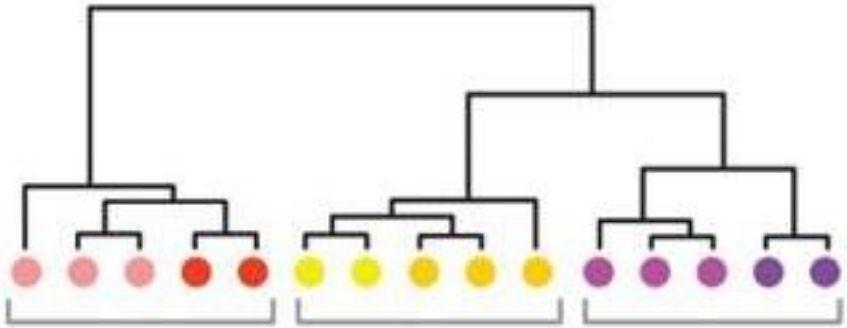
Clethrionomys

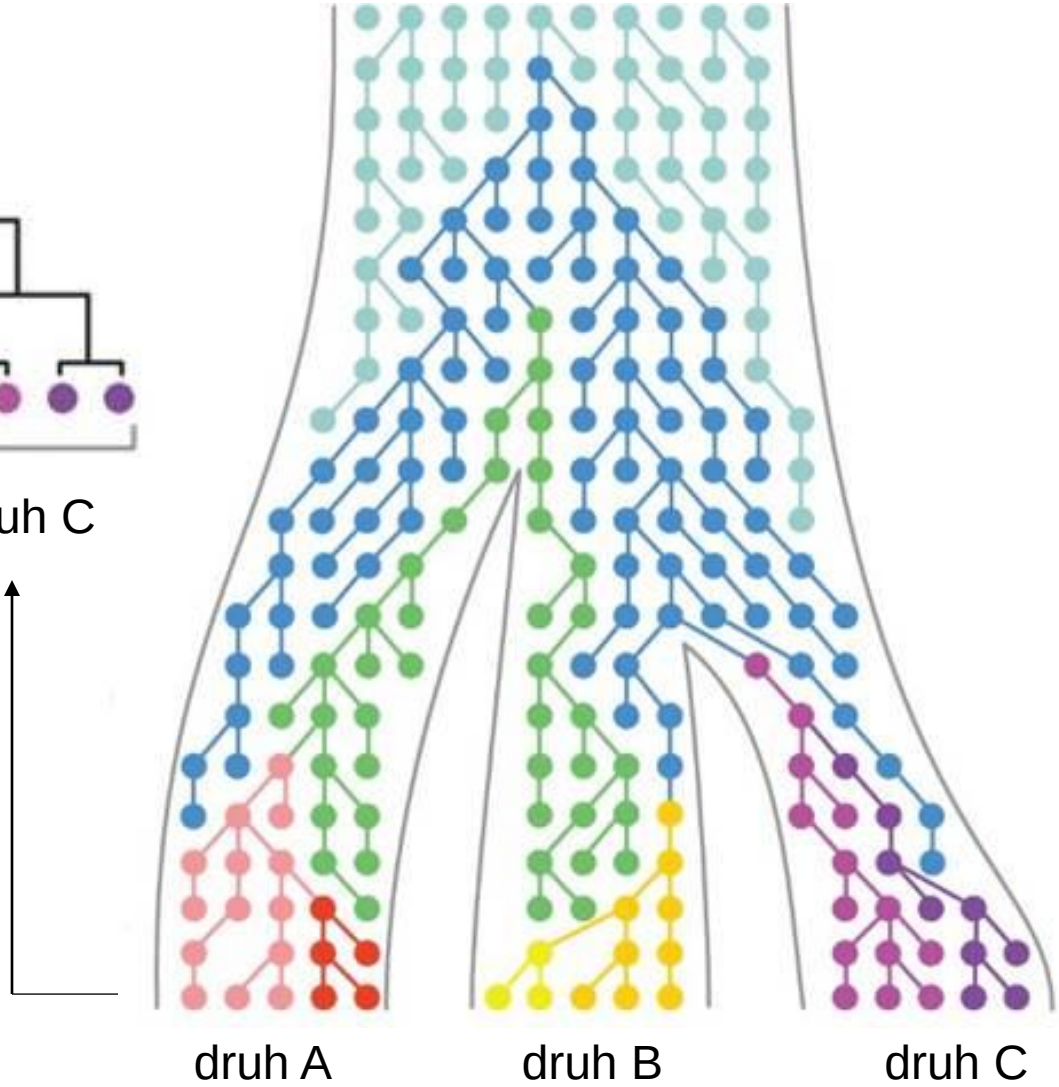
7-8 miliónů let



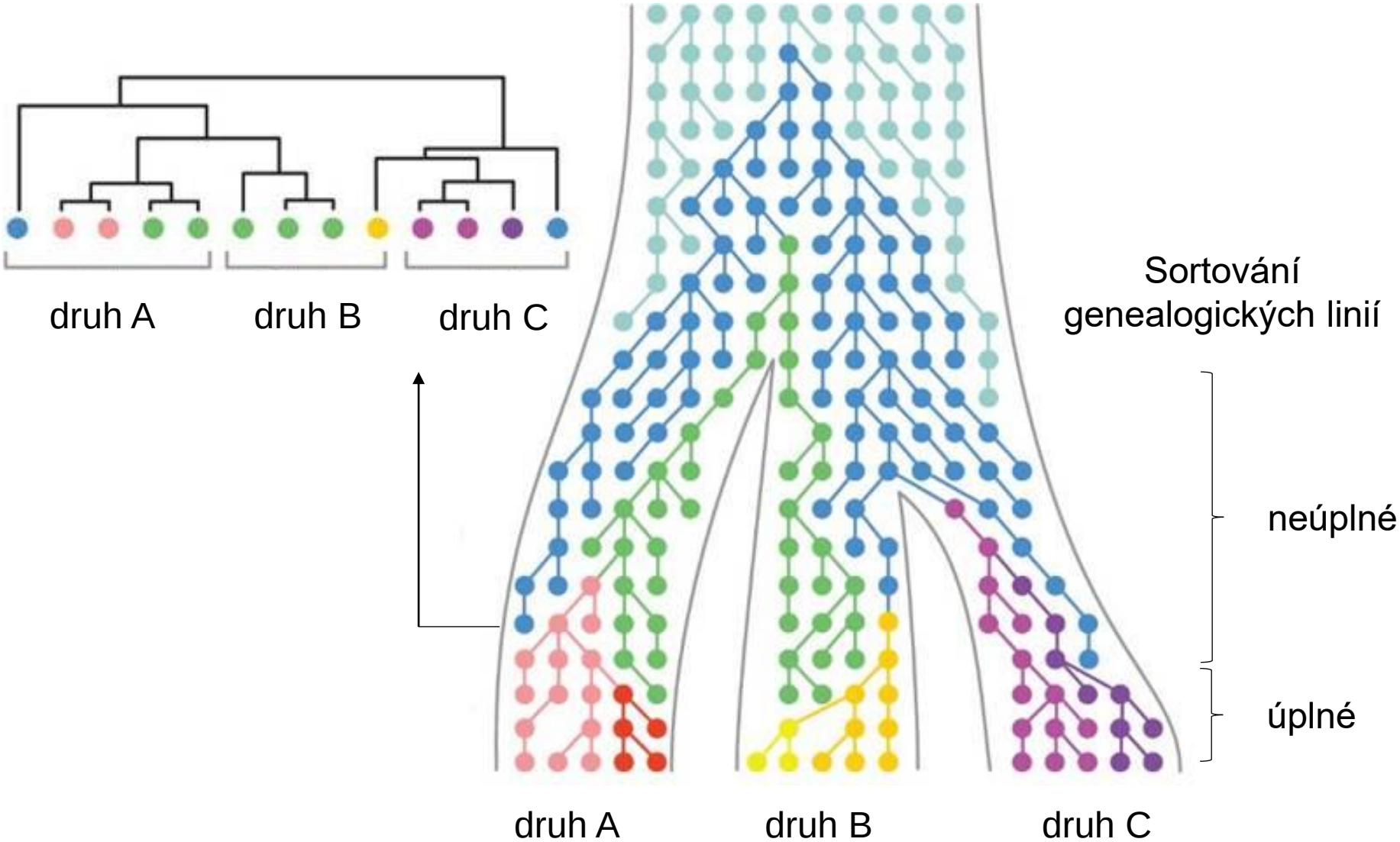
- MHC: stejné nebo podobné alely byly nalezeny u různých druhů
- když působí selekce, tak se výhodné sekvence udrží i přes období divergence druhů

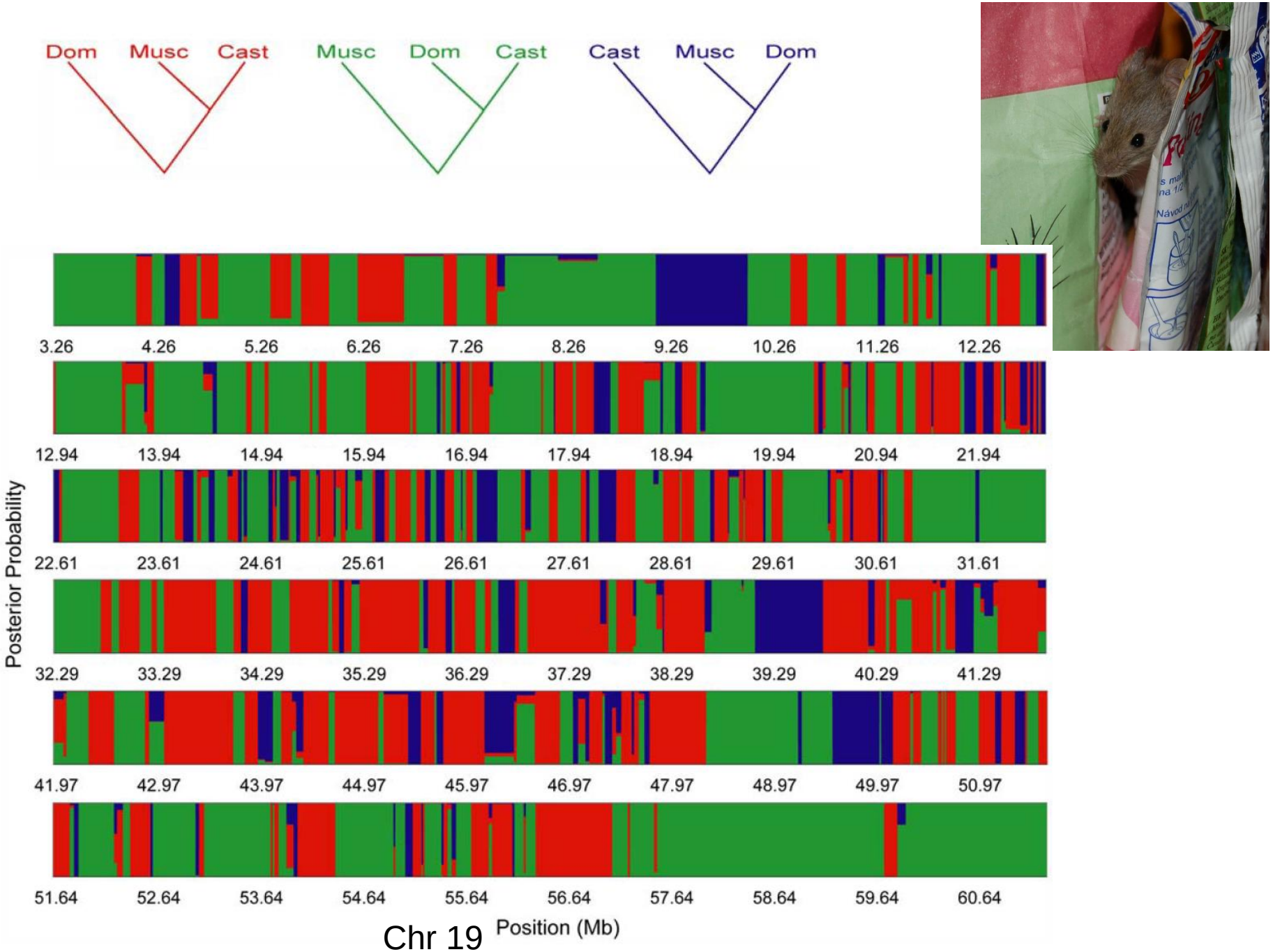
Nesrovnalost mezi genovým a druhovým stromem způsobená neúplným sortováním genealogických linií





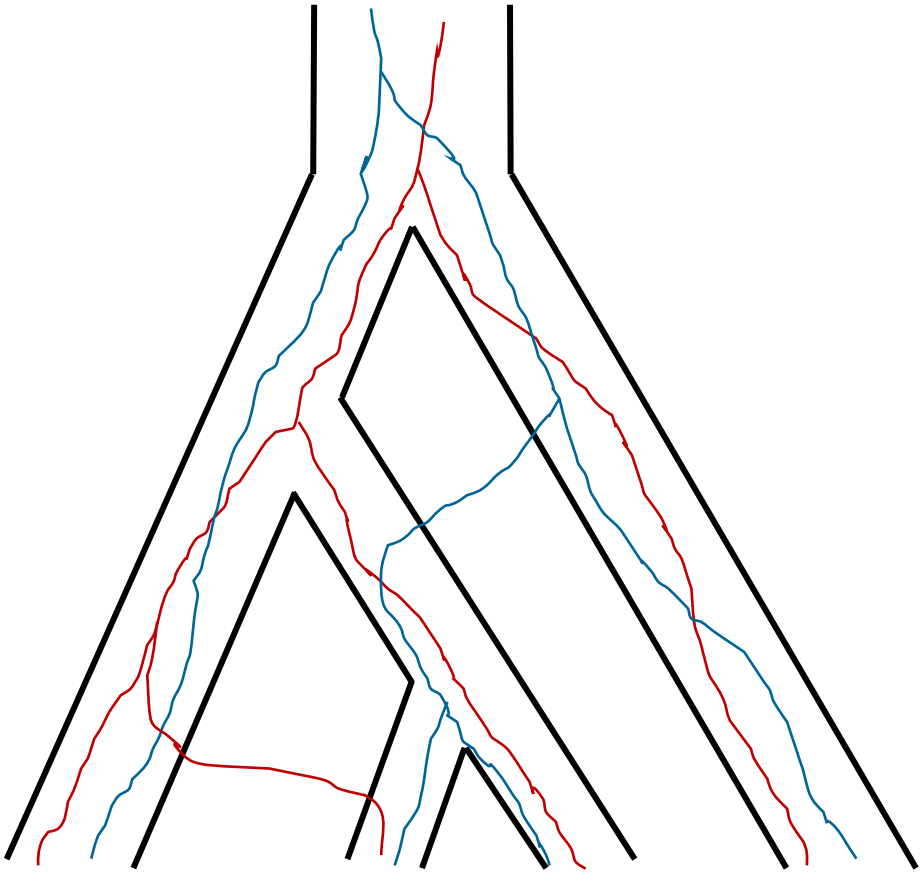
Nesrovnalost mezi genovým a druhovým stromem způsobená neúplným sortováním genealogických linií



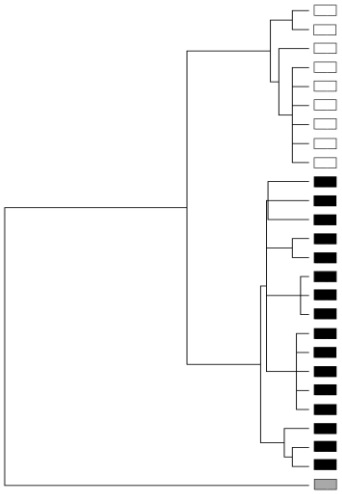

Dom
Musc
Cast
Musc
Dom
Cast
Cast
Musc
Dom
3.26
4.26
5.26
6.26
7.26
8.26
9.26
10.26
11.26
12.26
12.94
13.94
14.94
15.94
16.94
17.94
18.94
19.94
20.94
21.94
22.61
23.61
24.61
25.61
26.61
27.61
28.61
29.61
30.61
31.61
32.29
33.29
34.29
35.29
36.29
37.29
38.29
39.29
40.29
41.29
41.97
42.97
43.97
44.97
45.97
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
Chr 19
Position (Mb)

Nesrovnalost mezi genovým a druhovým stromem způsobená mezidruhovou introgresí

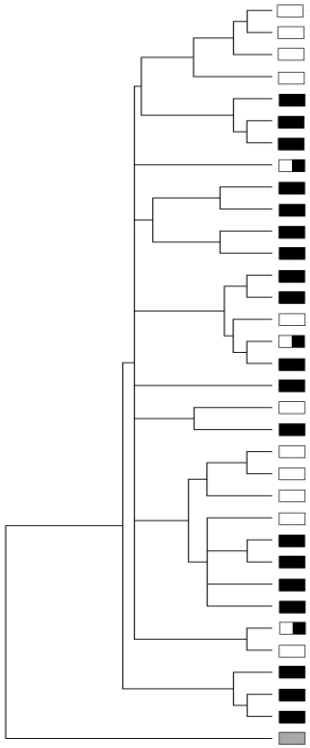
- Geny mají nezávislé osudy nejen v rámci populace, ale do určité míry i na napříč fylogenezí.
- Různé geny mohou ukazovat různé fylogeneze.
- Na druhové stromy je třeba nahlížet jako na výsledek demokratické většiny genových stromů.



Jak zjistit, zda je nesrovnalost mezi genovým a druhovým stromem způsobená introgresí či ancestrálním polymorfismem?



Z-linked locus

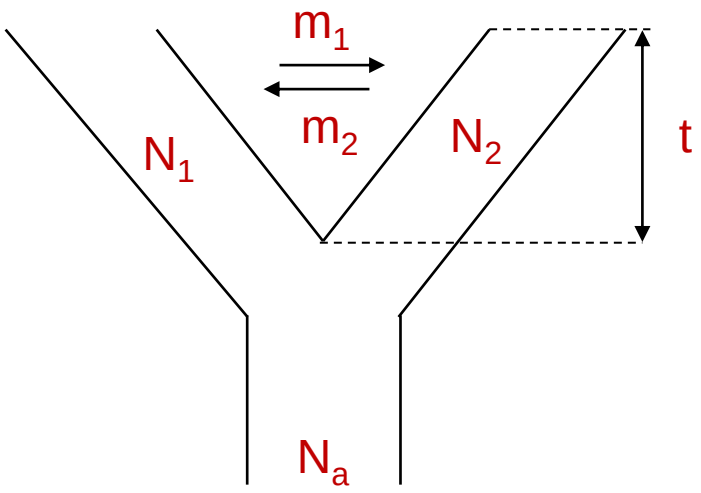


autosomal locus

Slavík obecný

Slavík tmavý

Koalescenční model izolace s migrací (IM)



Program IMa2 (Jody Hey)

Na základě genetických dat (sekvence) program odhadne parametry IM modelu.

Heuristic approach to estimate posterior probabilities of parameters
Markov-chain Monte Carlo (MCMC)

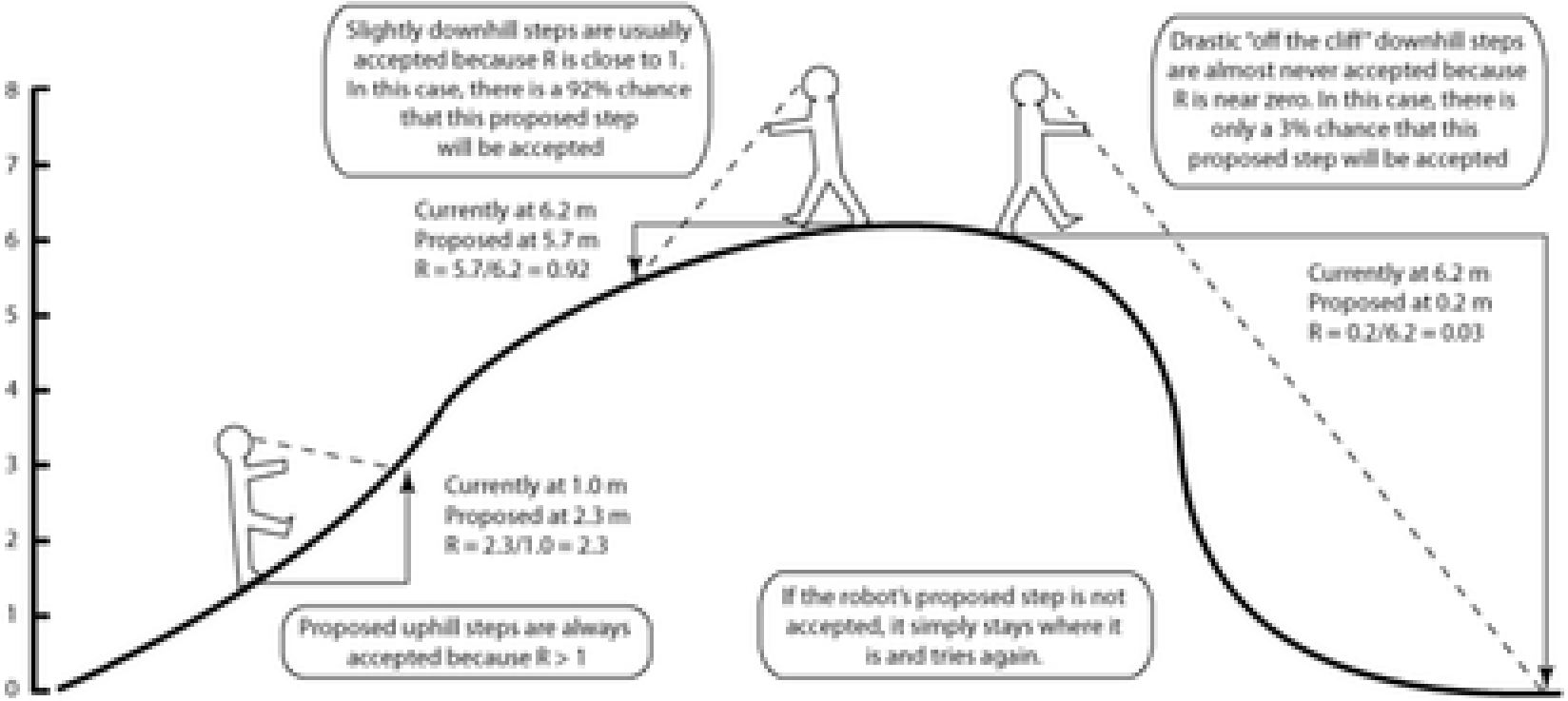
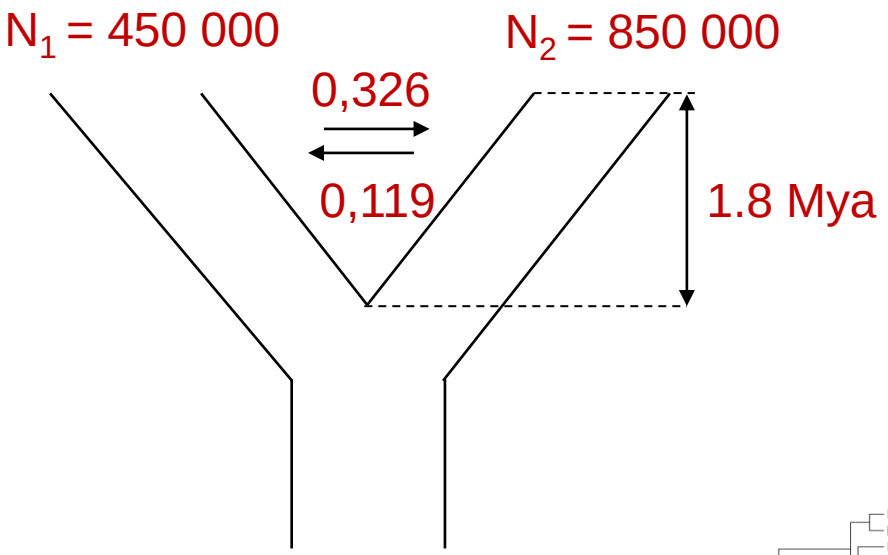


Illustration of MCMC method process (Lewis, 2011)

Model izolace s migrací

Slavík obecný

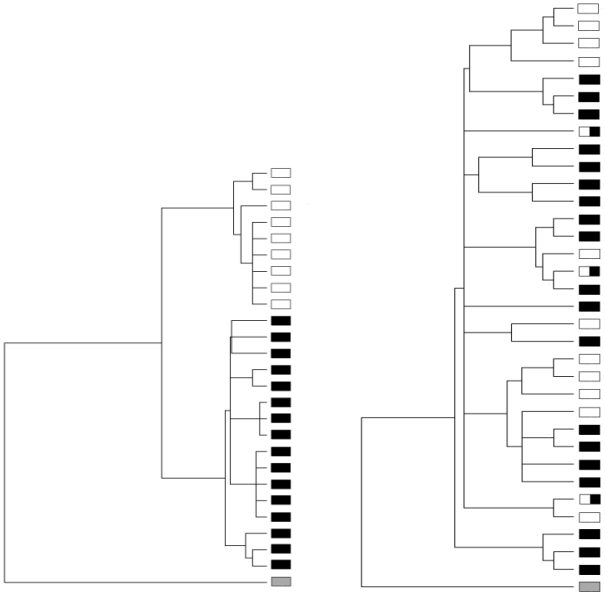
Slavík tmavý



- 12 intronic loci

Large Z-effect

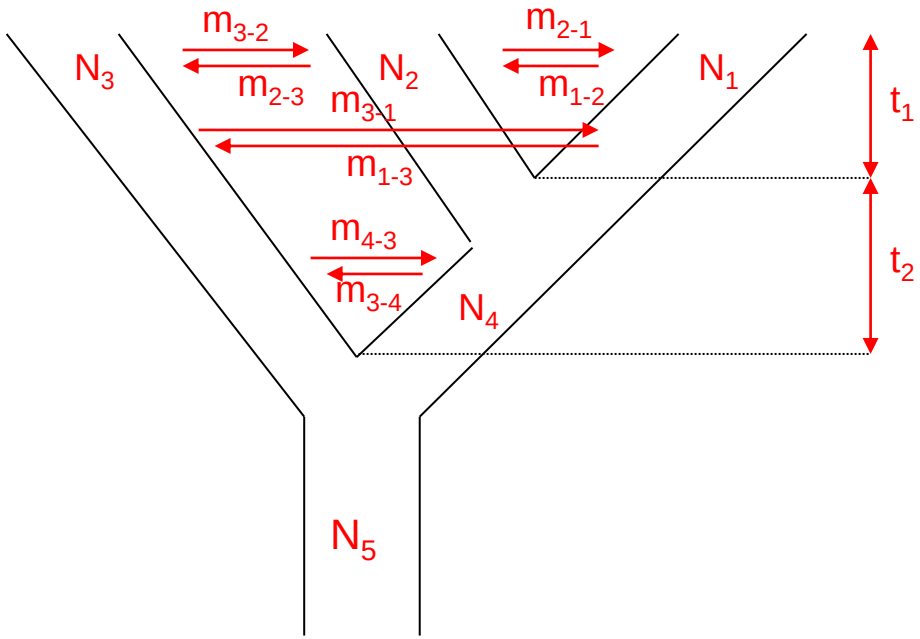
autozomy: m_1 = 0,081, m_2 = 0,763
chromozom Z: m_1 = 0,034, m_2 = 0,002



chromozom Z autozom

Storchová *et al.* 2010. Evolution

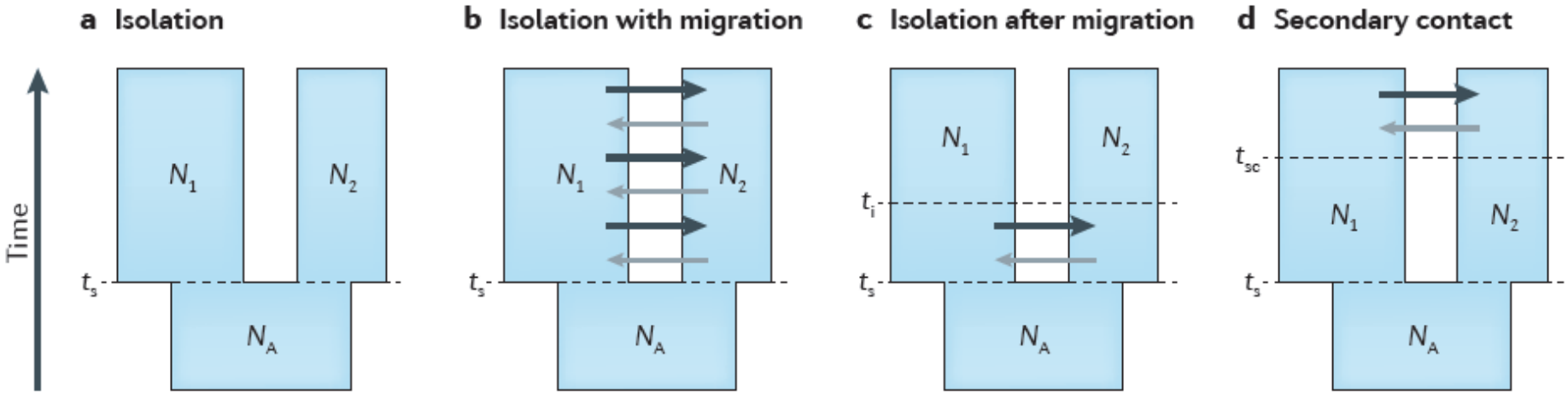
Isolation with migration (IM) model for multiple populations



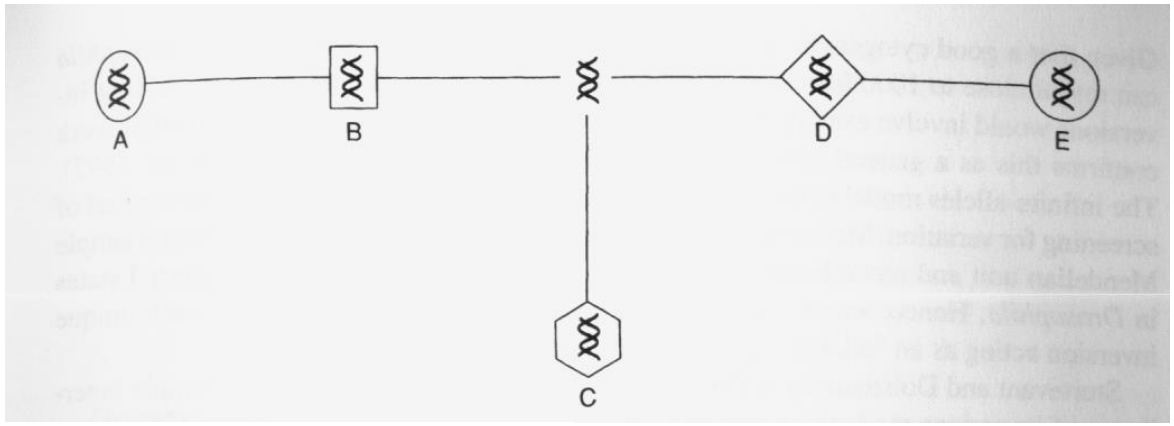
Programs:
IMa – two populations
IMa2 – multiple populations
IM – locus specific migration rates

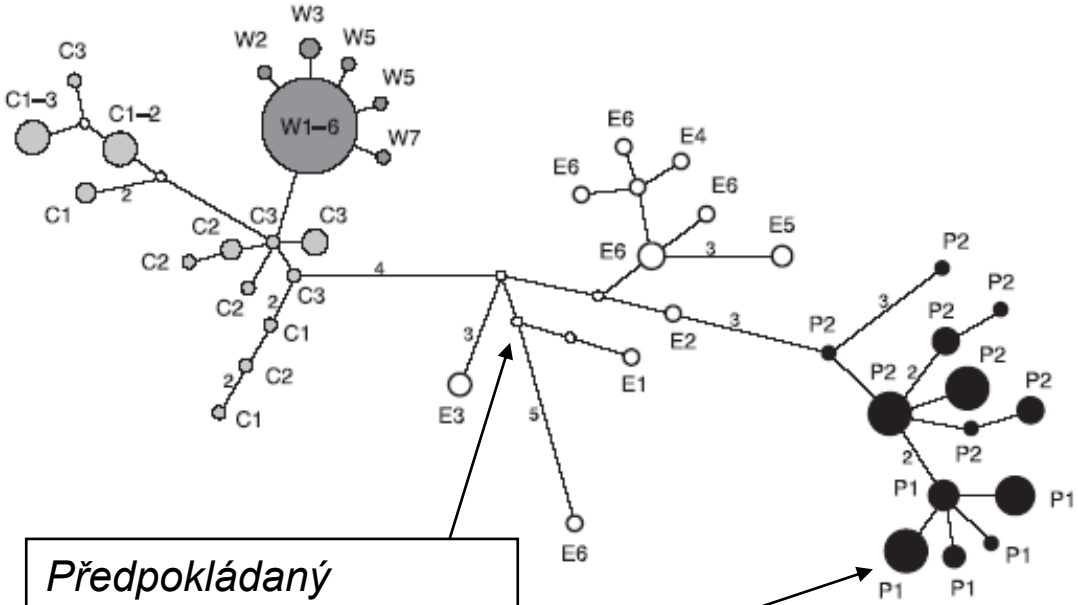
Hey J (2010).
Mol.Biol.Evol.

Approximate Bayesian Computation (ABC)



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

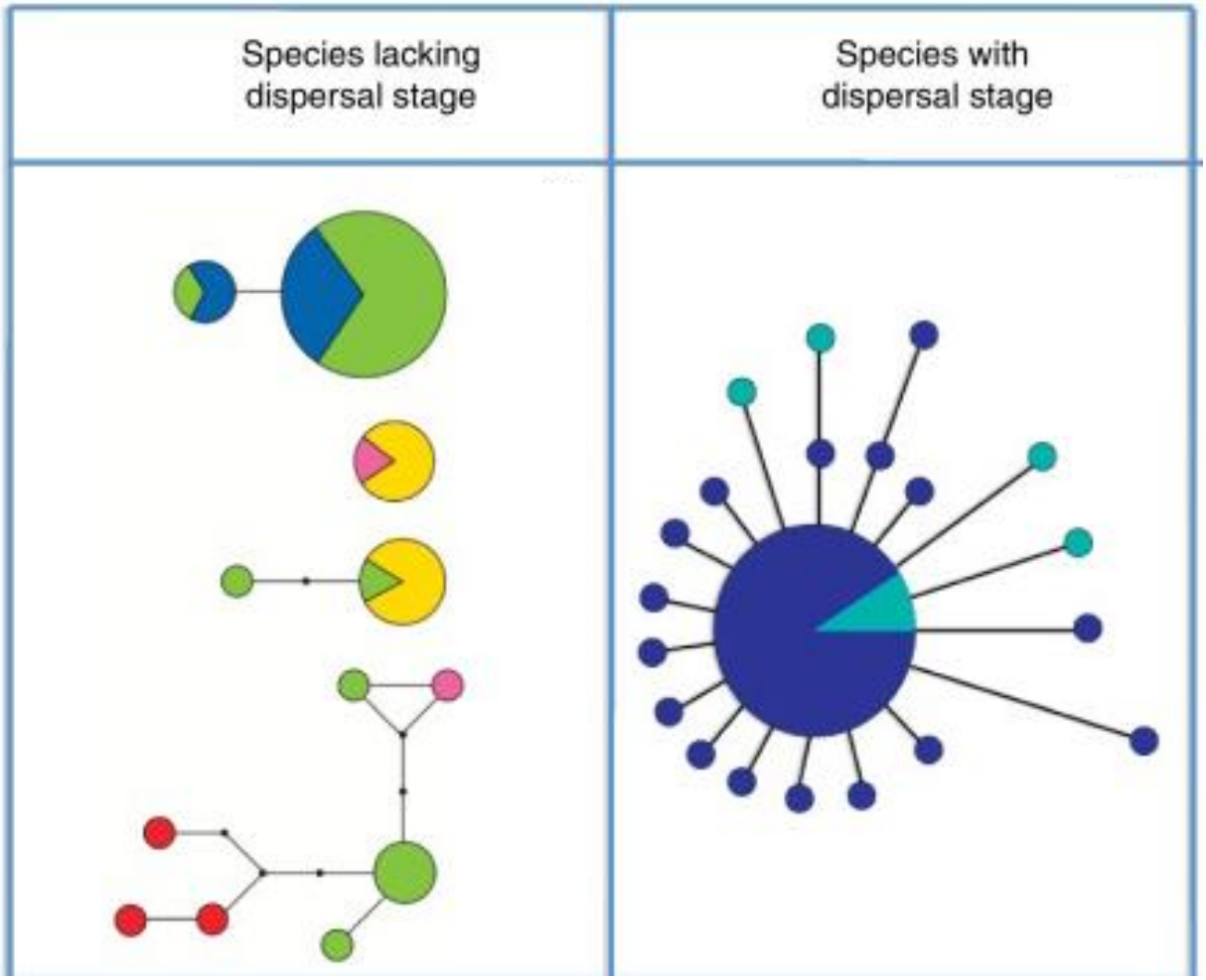




Toto jsou sice stromy,
ale bežně se jim říká sítě.

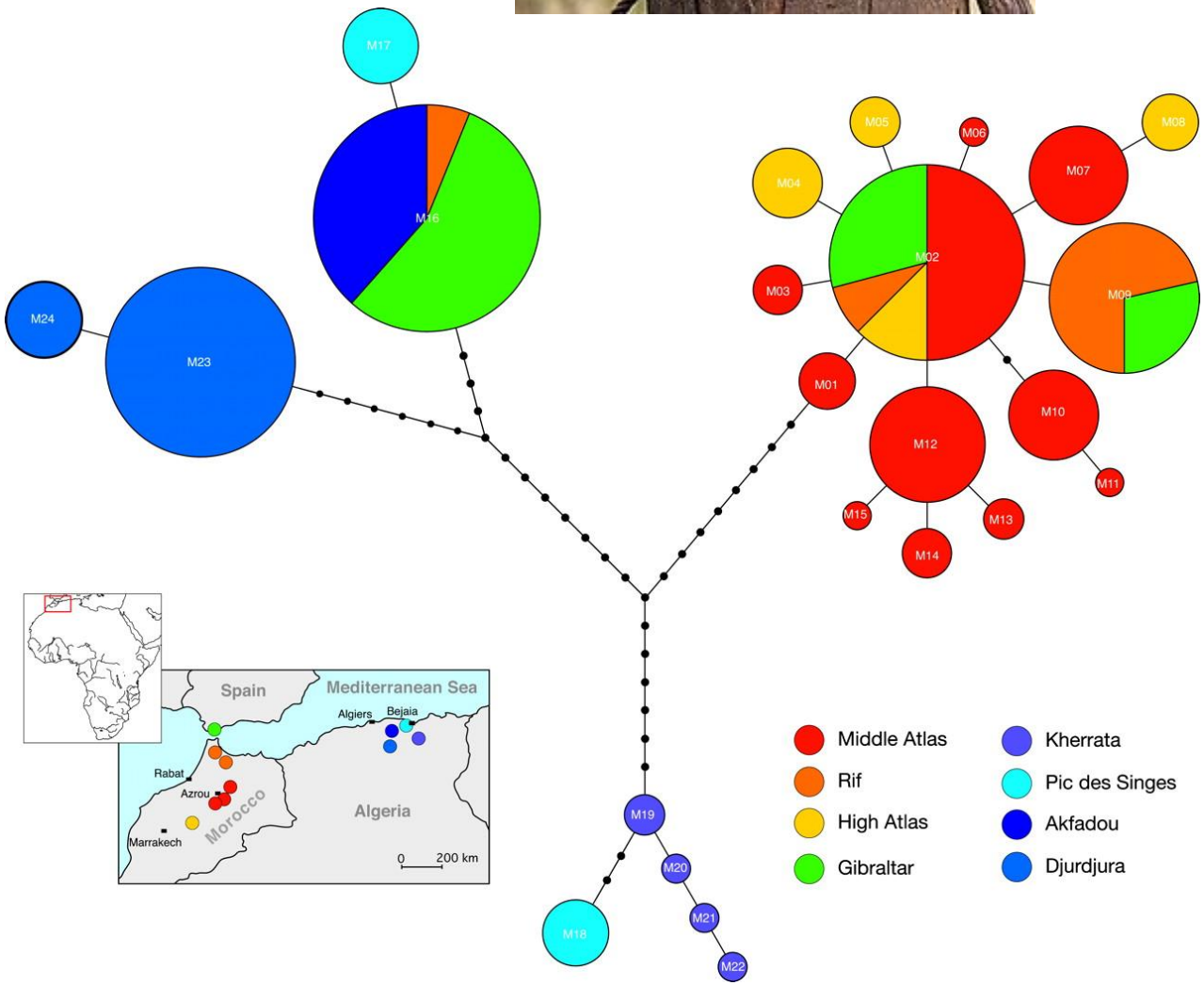
Předpokládaný
nikoliv však pozorovaný
haplotyp

Konkrétní zjištěný haplotyp
Velikost kroužku = počet jedinců s tímto haplotypem


Species lacking dispersal stage
Species with dispersal stage

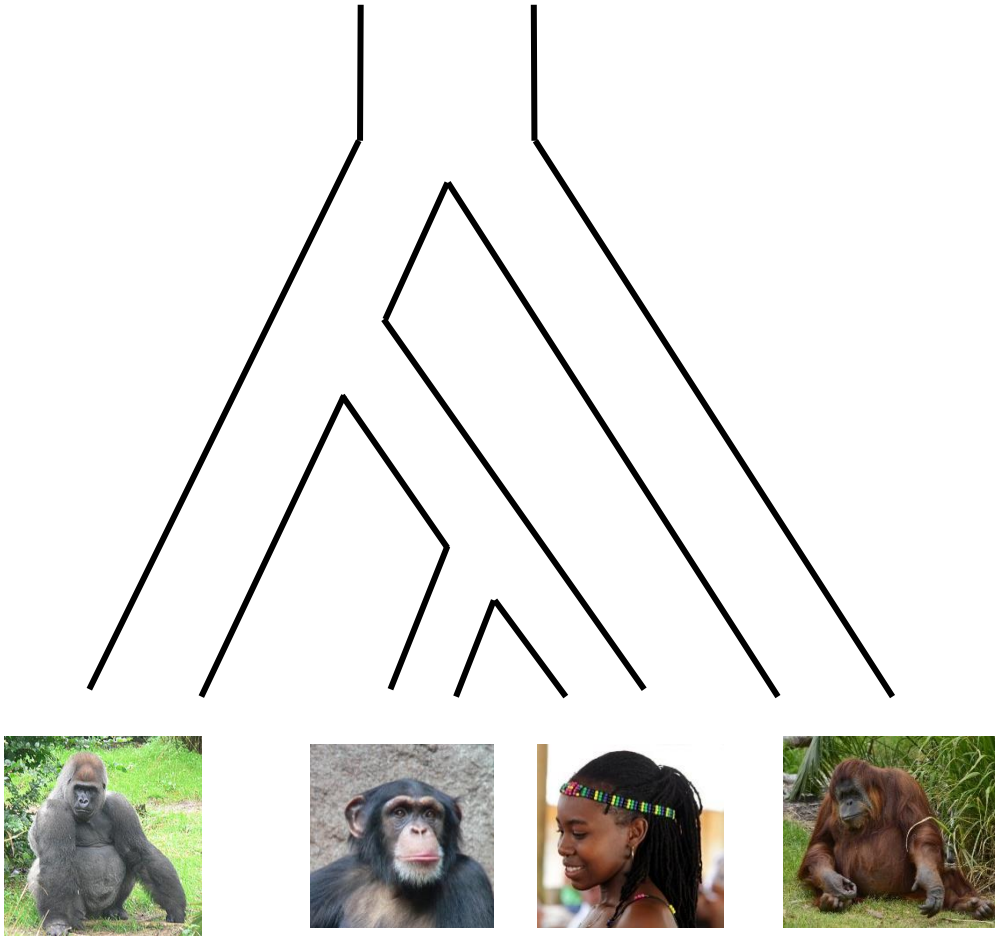
Phylogeography

Barbary macaques (*Macaca sylvanus*) and the origin of the Gibraltar colony



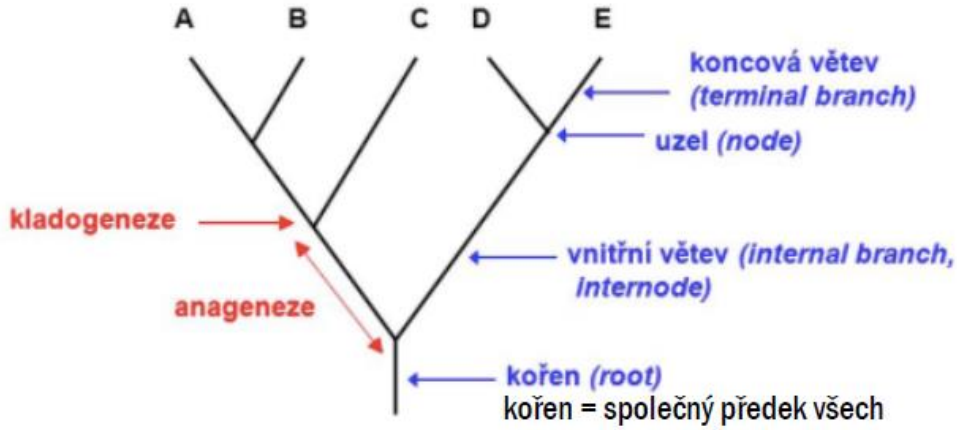
Fylogeneze

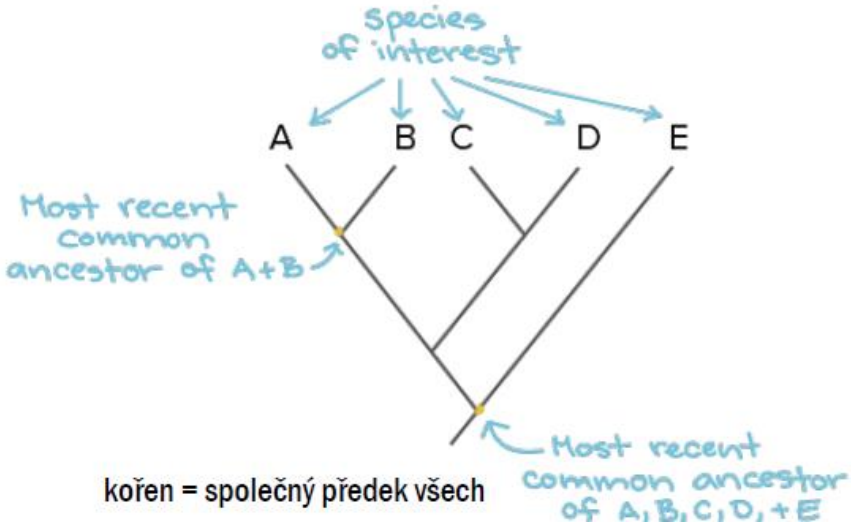
Vztahy mezi druhy a vyššími taxonomickými jednotkami.

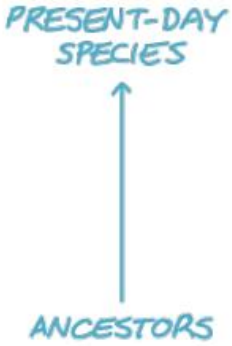


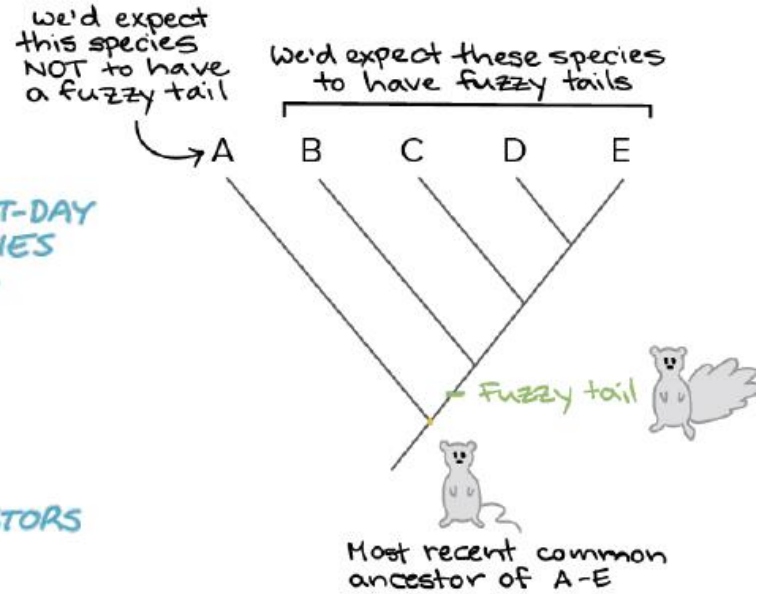
fylogenetický strom = záznam evoluce zkoumaných taxonů:

fylogeneze (to štěpení) vs. fylogenetika (přístup, který to zkoumá)





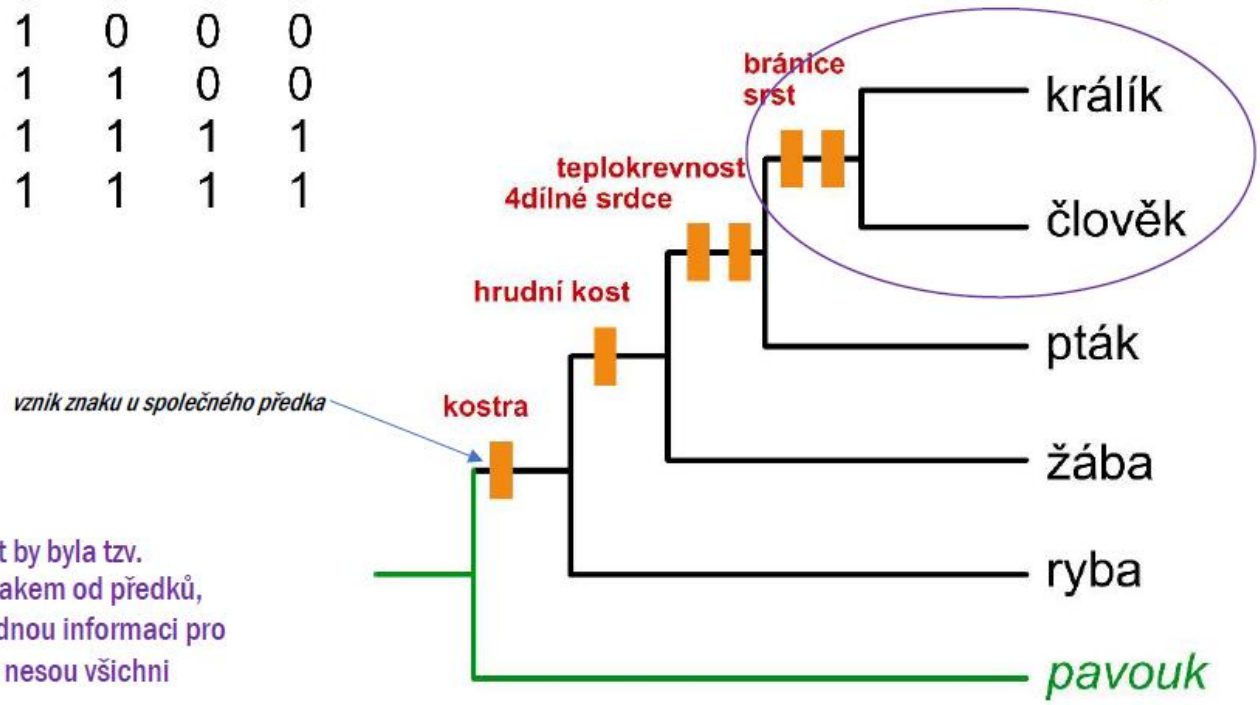




znaky pro rekonstrukci fylogeneze – např. morfologie, anatomie, fyziologie, atd. – může to být cokoliv

	kostra	teplokrevnost	hrudní kost	čtyřdílné srdce	bránice	srst
ryba	1	0	0	0	0	0
žába	1	0	1	0	0	0
pták	1	1	1	1	0	0
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1

apomorfie
(nově odvozený znak, nese informaci o příbuznosti), funkčně jde o **homologii**

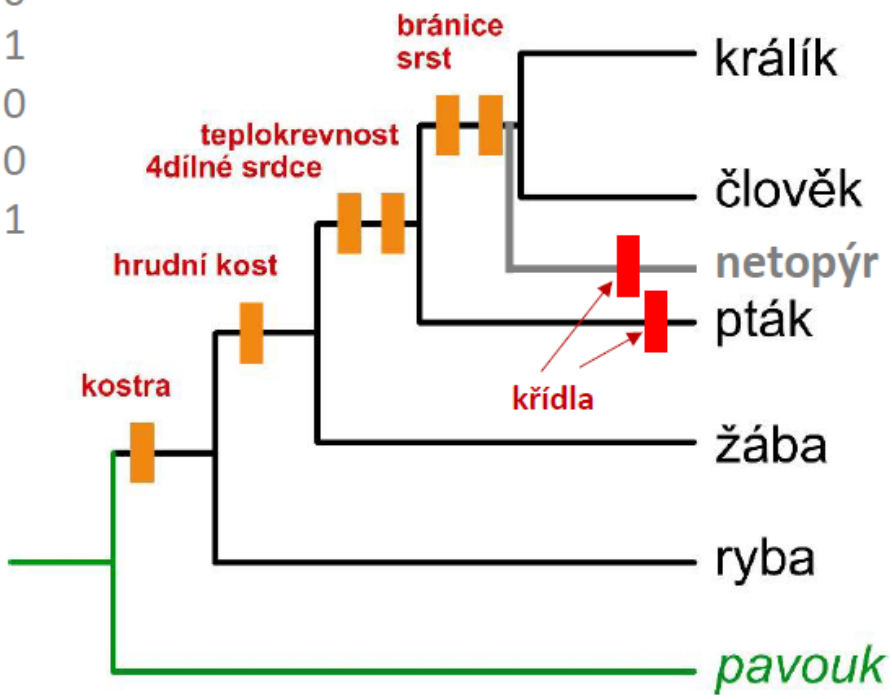


naopak: např. mnohobuněčnost by byla tzv. **pleziomorfie**, tedy zděděným znakem od předků, který v tomto kontextu nemá žádnou informaci pro rekonstrukci evoluce, jelikož ho nesou všichni

znaky – např. morfologie, anatomie, fyziologie, atd. – může to být cokoliv

	kostra	teplokrevnost	hrudní kost	čtyřdílné srdce	bránice	srst	"křídla"
ryba	1	0	0	0	0	0	0
žába	1	0	1	0	0	0	0
pták	1	1	1	1	0	0	1
králík	1	1	1	1	1	1	0
člověk	1	1	1	1	1	1	0
netopýr	1	1	1	1	1	1	1

přítomnost znaků "bránice" a "srst" nám pomůže odhalit to, že křídla nevznikla jednou, ale nezávisle dvakrát (a tedy že nejsou homologní, ale je to tzv. homoplázie)



!!polarizace znaků – co je ancestrální víme až poté co známe fylogenezi, založenou i na jiných znacích...

znaky molekulární = sekvence DNA – alignment = seřazení do matice

tvorba matice pro následné analýzy = srovnání sekvencí pod sebe...

```
AATGCCCTAAA                AATGCC-CTAAA
AATGCGGCTAAA   ------>     AATGCGGCTAAA
AACGCGCTAAA                AACGCG-CTAAA
ATGCTAA                    -ATG---CTAA-
```

při alignmentu: program hledá nejlepší rozmístění, dostává body:
za každou shodnou bázi plusové body,
za každou mezeru mínus body,
za otevření mezery speciální mínus body...

=> cílem je získat co nejvíce bodů!

metody výpočtu fylogenetických stromů:

1) distanční metody
2) maximální úspornost = parsimonie
3) maximální věrohodnost = likelihood
4) Bayesovské metody výpočtu

Distanční metody: UPGMA (Sokal & Michener 1958), Neighbor-Joining (Saitou and Nei, 1987)

čistě hierarchické clusterování

již částečné zanesení evoluční info o potenciálně odlišné rychlosti změn v různých větvích

z DNA či morfologie:

```
AATGCC-CTAAA
AATGCGGCTAAA
AACGCG-CTAAA
-ATG---CTAA-
```

	kostra	teplokrevnost	hrudní kost	čtyřdílné srdce	bránice	srst
ryba	1	0	0	0	0	0
žába	1	0	1	0	0	0
pták	1	1	1	1	0	0
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1

příklad

distanční metody výpočtu stromů:

- výpočet vzdáleností každé sekvence od každé – vznikne matice vzdáleností:

morfo_priklad.txt - Poznámkový blok

Soubor Úpravy Formát Zobrazení Nápověda

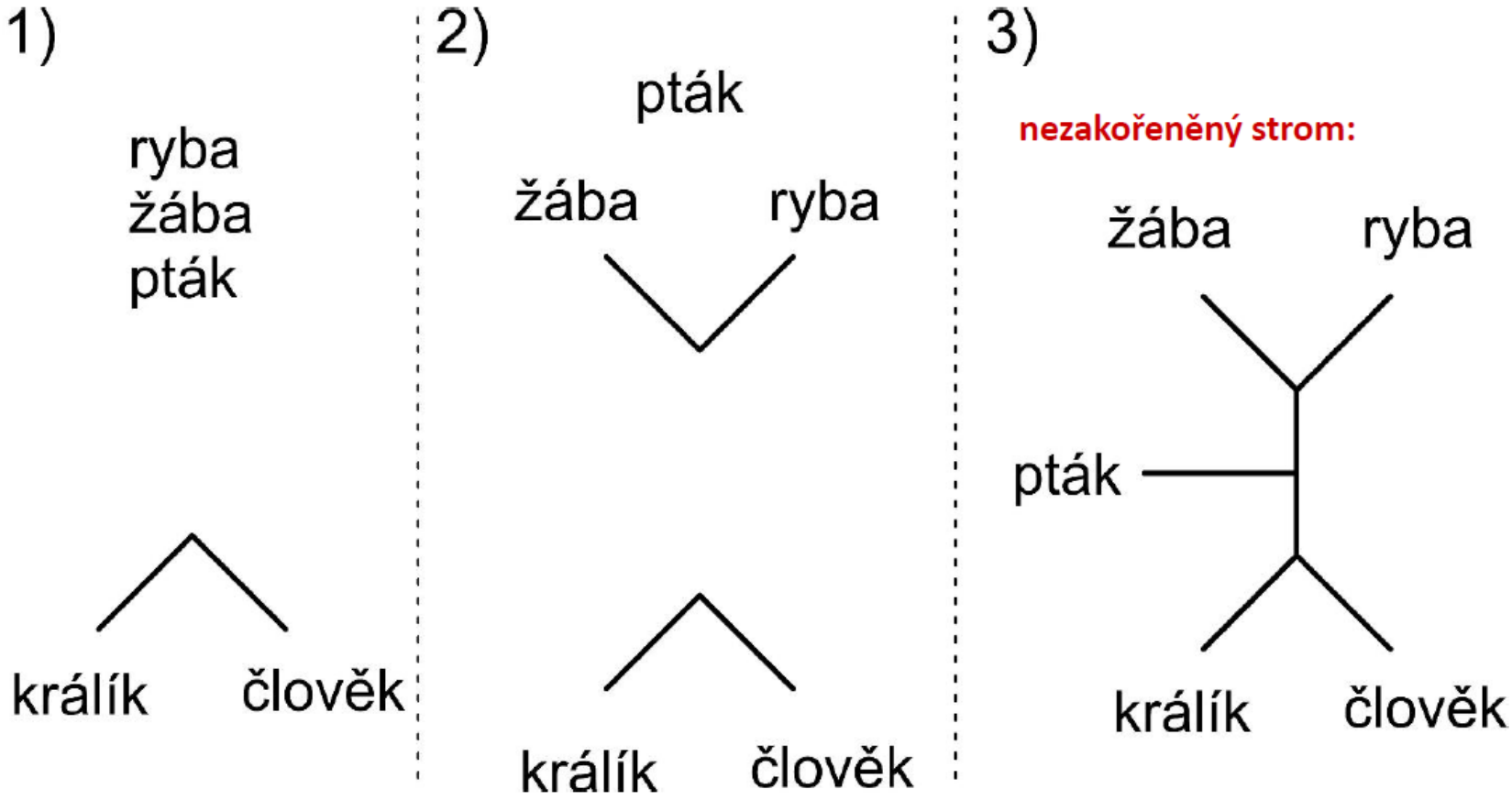
```
#NEXUS

 begin data;
 dimensions ntax=5 nchar=6;
 format datatype=standard;
  matrix
ryba     100000
zaba     101000
ptak     111100
kralik   111111
clovek   111111;
end;
```

```
Pairwise distances between taxa

   Below diagonal: Total character differences
   Above diagonal: Mean character differences

                 1        2        3        4        5
  1 ryba         - 0.16667 0.50000 0.83333 0.83333
  2 zaba         1        - 0.33333 0.66667 0.66667
  3 ptak         3        2        - 0.33333 0.33333
  4 kralik       5        4        2        - 0.00000
  5 clovek       5        4        2        0        -

Mean character difference distance matrix written to
"I:\_ZUZA\_prezentace\CZU_geneticke metody\pokus_muj\
```

nejmenší vzdálenost = spojení

=> matice se pak znovu přepočítá, králík+člověk se teď berou jako jeden taxon s průměrnou hodnotou

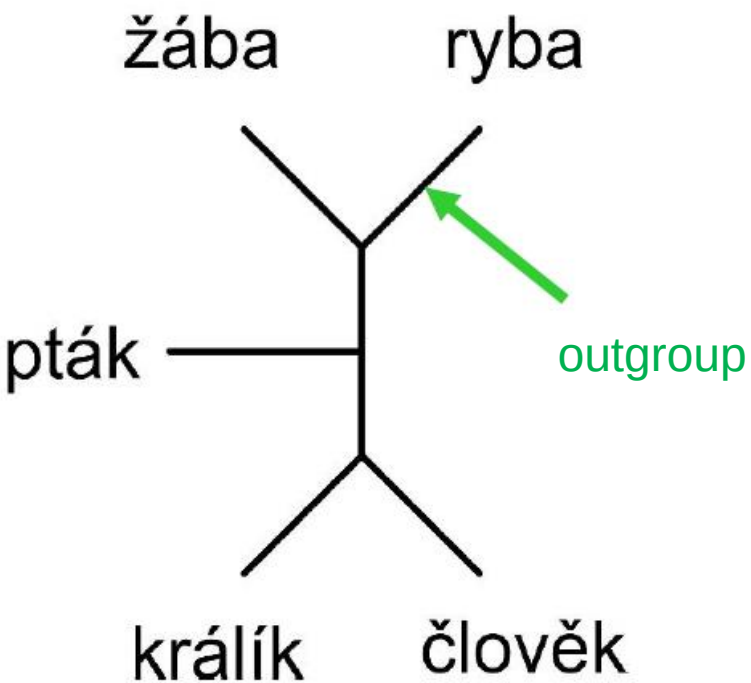
distanční metody výpočtu stromů:



-nyní se vezmou druhy s nejmenší vzdáleností a spojí se, mají společnou vzdálenost a pokračuje se, dokud nejsou všechny druhy ve stromu...

zakořeněné a nezakořeněné fylogenetické stromy:

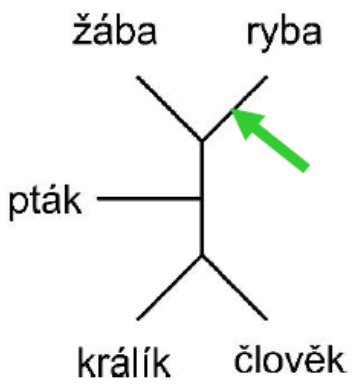
nezakořeněný strom:



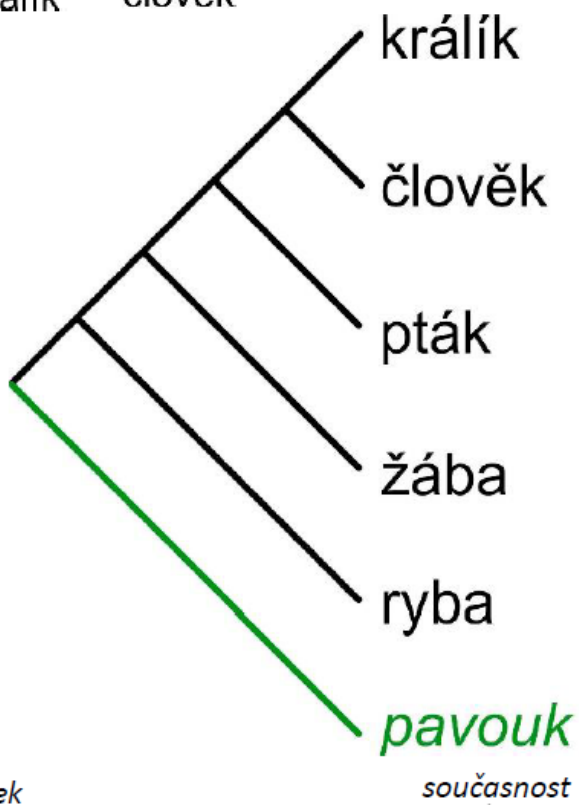
	kostra	teplokrevnost	hrudní kost	čtyřdílné srdce	bránice	srst
ryba	1	0	0	0	0	0
žába	1	0	1	0	0	0
pták	1	1	1	1	0	0
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1
pavouk	*0*	*0*	*0*	*0*	*0*	*0*

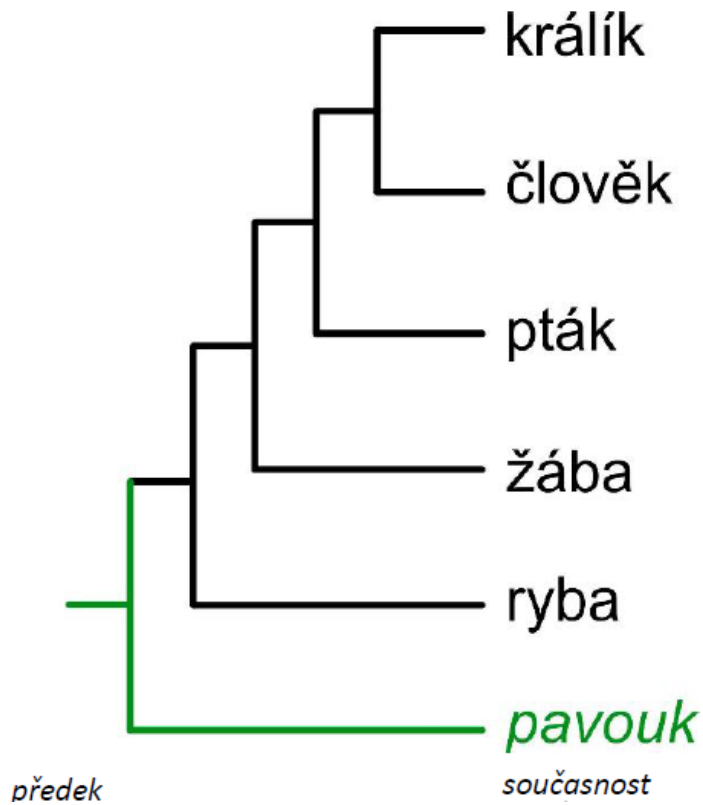
zakořeněné a nezakořeněné fylogenetické stromy:

nezakořeněný strom:



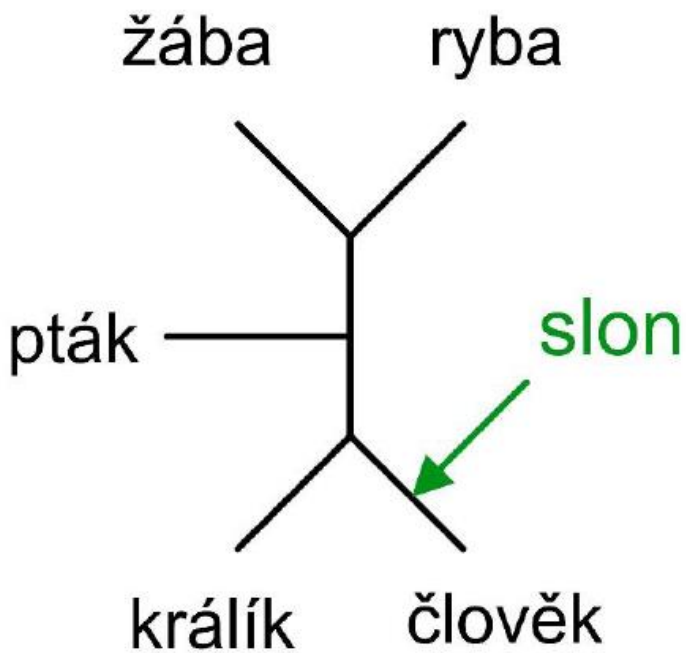
zakořeněné stromy:

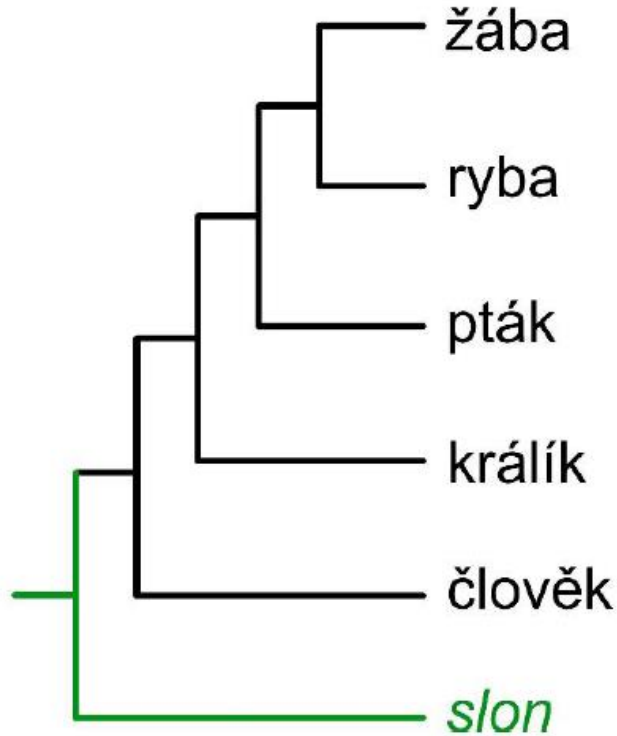




co se stane, když zvolíme špatný outgroup?

nezakořeněný strom:





Metoda maximální parsimonie = maximální úspornosti

- máme 3 zvířata: vlaštovka, netopýr a kočka
- bez jakékoliv matice: existují 3 možné způsoby, jak si mohou být příbuzní:

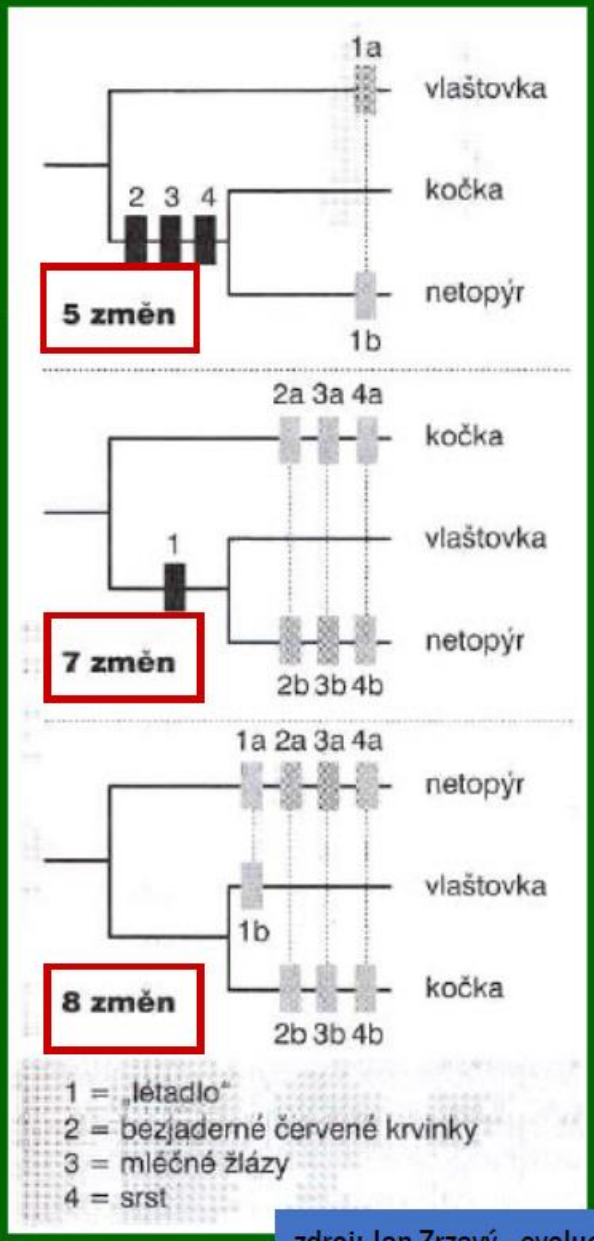
- maximální parsimonie = nejmenší počet změn znaků

· maximální parsimonie = nejmenší počet změn znaků:

	vlaštovka	netopýr	kočka
„létadlo"	ano	ano	ne
bezjaderné červené krvinky	ne	ano	ano
mléčné žlázy	ne	ano	ano
srst	ne	ano	ano

počet dichotomických stromů pro *n* druhů

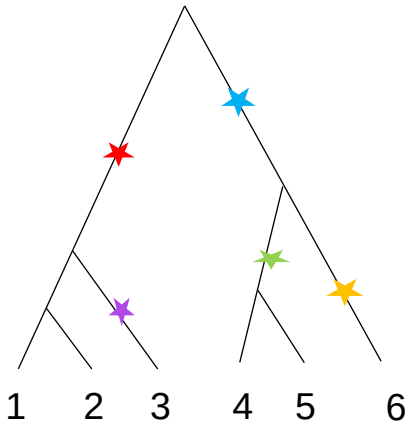
$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

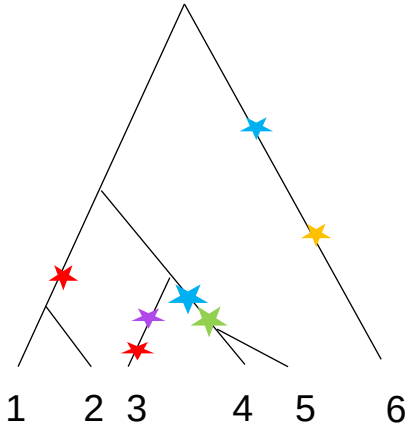


zdroj: Jan Zrzavý - evoluce

Metody rekonstrukce fylogeneze

Metoda maximální úspornosti (parsimonie)

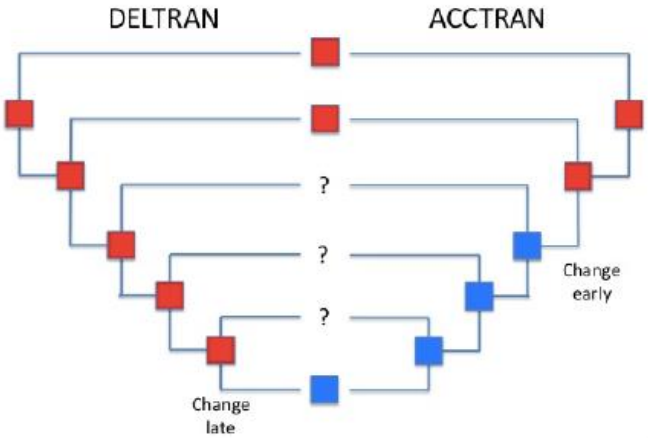




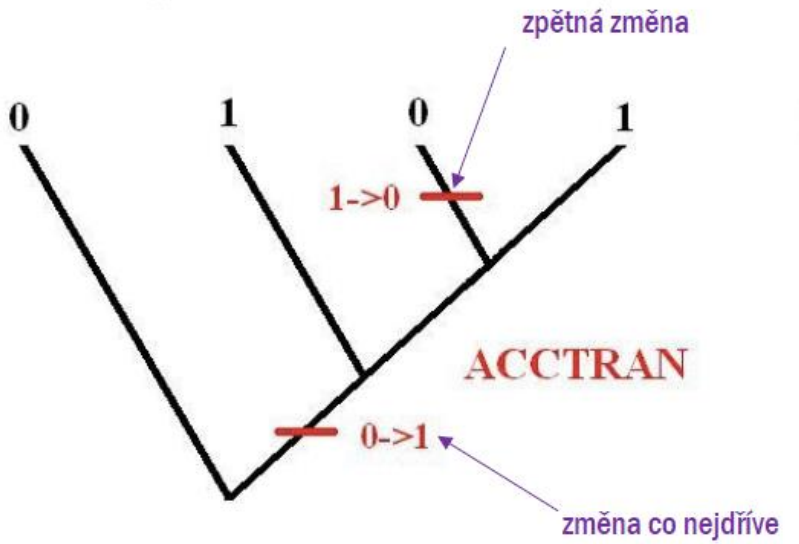
Sekvence	
1	TAGCA
2	TAGCA
3	CAGCA
4	TTCTA
5	TTCTA
6	TTGTC

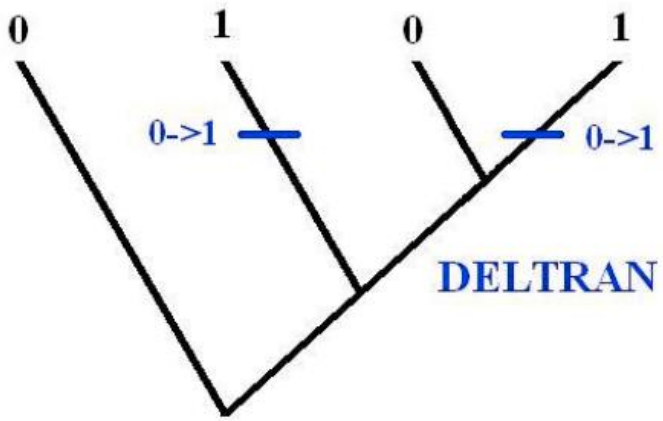
ACCTRAN/DELTRAN optimalizace – pro situace, kdy je počet záměn stejný, ale jsou různé varianty

Parsimony problem



ACCTRAN = preference změny co nejdříve
DELTRAN = změna co nejpozději





heuristický přístup – když nelze prohledat všechno:

problém: všechny stromy jdou udělat do max. 11 druhů...

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

exponenciální nárůst počtu potenciálních stromů

3 taxony: (6-3)!/2(1)! = 6/2 = **3 stromy**

6 taxonů: (12-3)!/2^4(4)! = **945 stromů**

9 taxonů: (18-3)!/2^7(7)! = **2‘027‘025 stromů**

12 taxonů: (24-3)!/2^{10}(10)! = **13‘749‘310‘575 stromů**

=> maximální počet, kdy je reálné prohledávat všechny kombinace je 11 taxonů

species	number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

heuristický přístup – když nelze prohledat všechno:

není tedy možné prohlédnout všechny stromy, spočítat pro ně počty evolučních změn a vybrat ten nejlepší

⇒heuristické hledání stromů

tj. - vytvoří se náhodný strom,
- spočítají se evoluční změny,
- náhodně se v něm přehodí dvě větve,
- spočítají se evoluční změny
- dál postupuje jen ten, který měl méně změn!!! atd. atd.
- na konci řady je strom s nejméně změnami

výběr náhodného stromu se opakuje několikrát, porovnání výsledných stromů
- může být i několik nejlepších stromů

- heuristický postup se uplatňuje u všech dalších metod, tj. nejen parsimonie, ale i max. likelihood a Bayesovském přístupu

metody výpočtu fylogenetických stromů:

1) distanční metody - na základě distancí, výsledkem 1 strom
2) maximální úspornost = parsimonie – více možných stromů, bereme konsensus
3) maximální věrohodnost = likelihood – výsledkem 1 strom
4) Bayesovské metody výpočtu – výsledkem 1 strom vč. statistické podpory

tyto dvě metody jsou statisticky nejodvozenější, používají obecný statistický aparát nevyvinutý přímo pro fylogenetiku

- existují myšlenkové školy, které uznávají jen první dvě metody, které mají dle nich jasné evoluční zadání: buď používám znaky a chci dosáhnout nejmenšího počtu změn, nebo dělám vzdálenosti (např. Willi Hennig Society, časopis Cladistics; jde rozhodně o menšinový názor, vadí jim i toto:

- metoda 3 a 4 jsou momentálně nejodvozenější a nejčastěji využívané

Maximum Likelihood method:

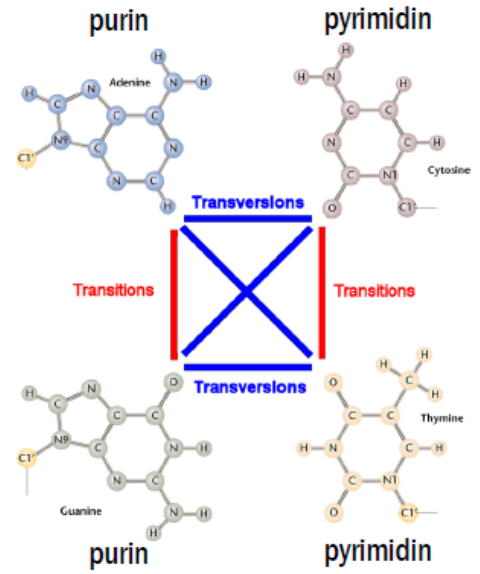
= the probability = LIKELIHOOD of observing our data (=matrix) under the given model, P(D | M)
model = tree + parameters (evolutionary events taken into account)

=> parameters taken into account in the model (e.g.

Bayesian approach:

= it can be seen as a modified version of likelihood
= **simplified:** the probability/likelihood of observing the tree given the data (matrix) + parameters
=> it's the opposite to the Maximum Likelihood, it uses so-called inverze probability

=> both approaches search for the best tree



the **transition** is more likely to be preserved in the DNA sequence than the **transversion** because the molecule doesn't change stability and the mutation doesn't lead tha often to the change of amino acid in the resulting protein

evoluční model pro ML (ale i Bayesovskou analýzu):

v modelu (paralela „cinknutosti“ mince) jsou matematicky zainkorporovány:

1) pravděpodobnosti jednotlivých záměn (A->T, A->C, A->G, T->C, T->G, C->G), maximálně může být až šest různých pravděpodobností;
(= šest různých substitučních typů)

2) poměry jednotlivých bází

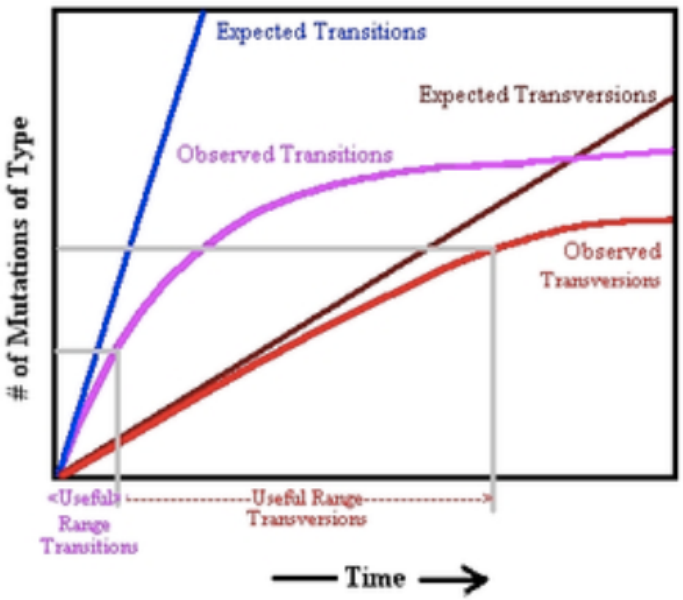
3) poměr nevariabilních míst vůči variabilním

4) pravděpodobnost záměny jednotlivých pozic (např. morfologie – stejná pro všechny znaky; DNA – různá, např. dle pozic v kodónu atd.)

5) korekce distance (vychází i z bodu 3).

<u>to vše je vlastně popis evoluce DNA</u>

p-distance vs. g-distance – pokus o optimální korekci; saturace



AATGCCGTATCGCGTTAATTTGAGCGCTTTCGAT

↓

AATGCCATATCGCGTTAAGTTGAGCCCTATCGAT

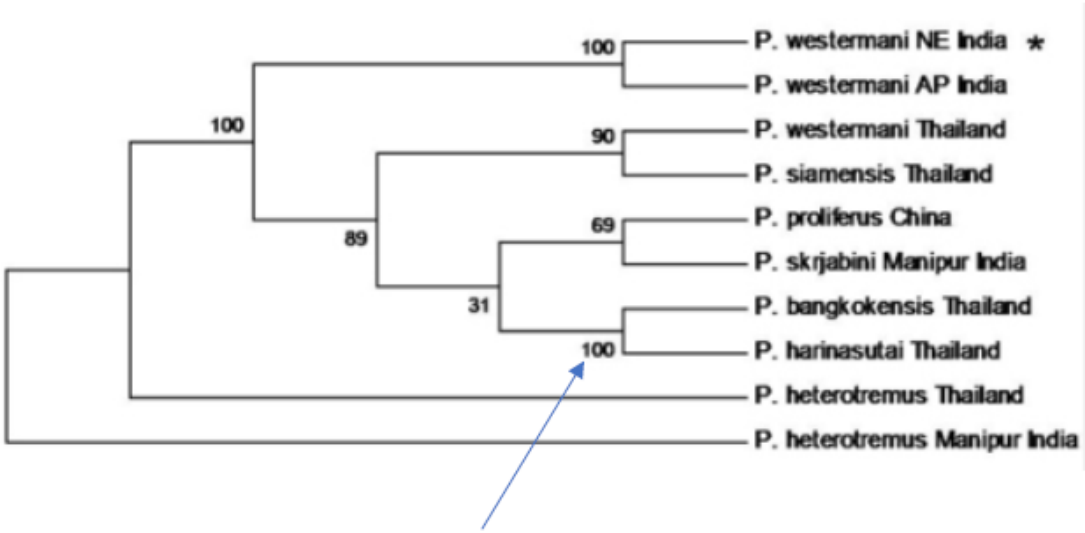
↓

AAAGCCATATCGCATTAAGTTGAGCGCTATCGAT

tato mutace nebude detekována, protože je zpětná, tedy změnila se na původní stav

=> saturované sekvence již nesou hodně šumu, saturované pozice pak málo fylogenetické informace. Nejčastěji jsou satuorvány 3. pozice kodónů. Model se snaží korigovat tento jev tím, že předpoví míru saturace

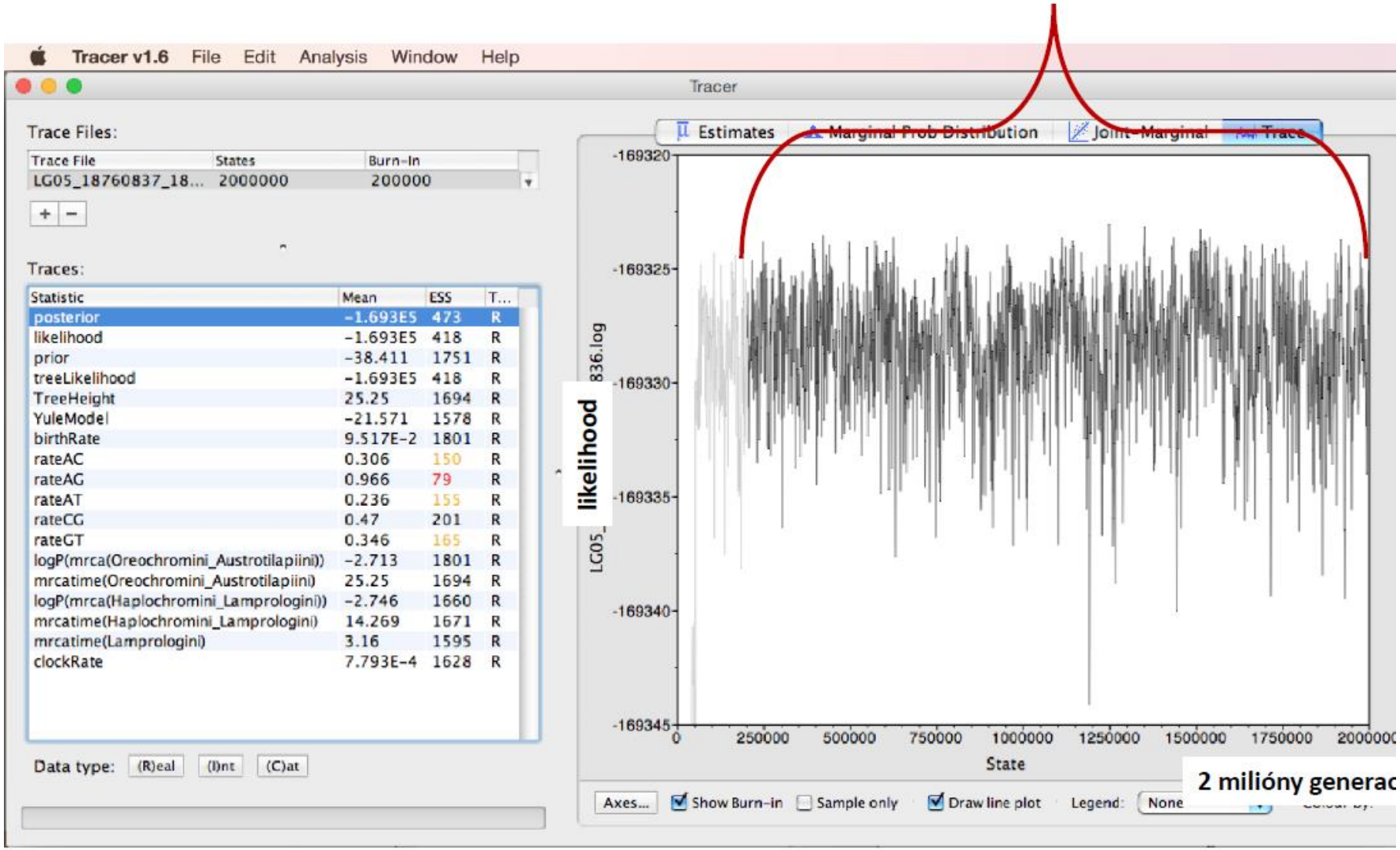
statistické podpory – míra důvěryhodnosti topologie; bootstrapování a BPP



statistická podpora existence uzlu

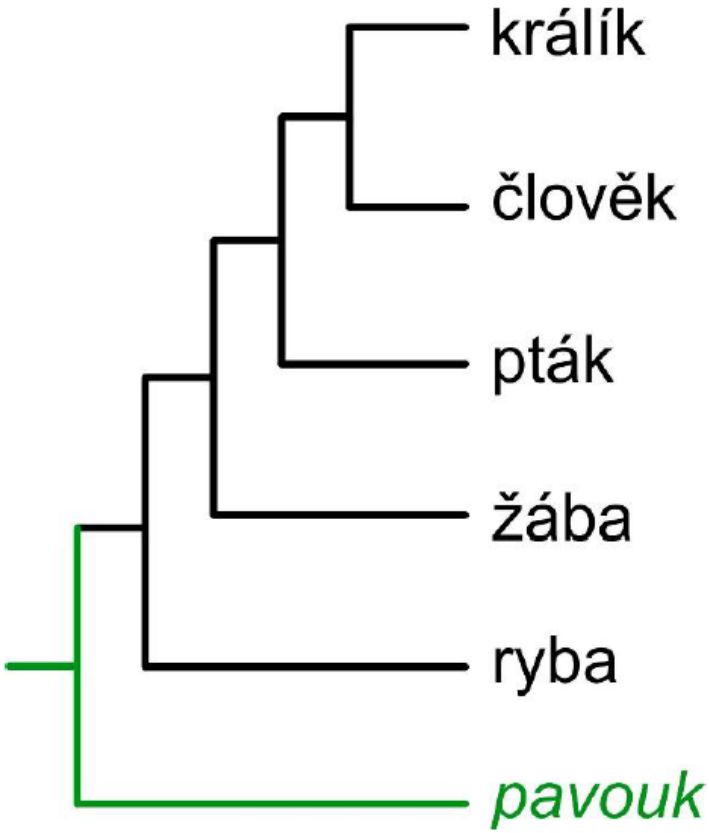
statistické podpory – BPP = Bayesian Posterior Probability

Bayesovská analýza – zpracuje všechny stromy v plateau fázi = kolik % stromů obsahuje daný uzel



statistická podpora – bootstrapování

využívá se pro metody Maximální parsimonie a Max. Likelihood



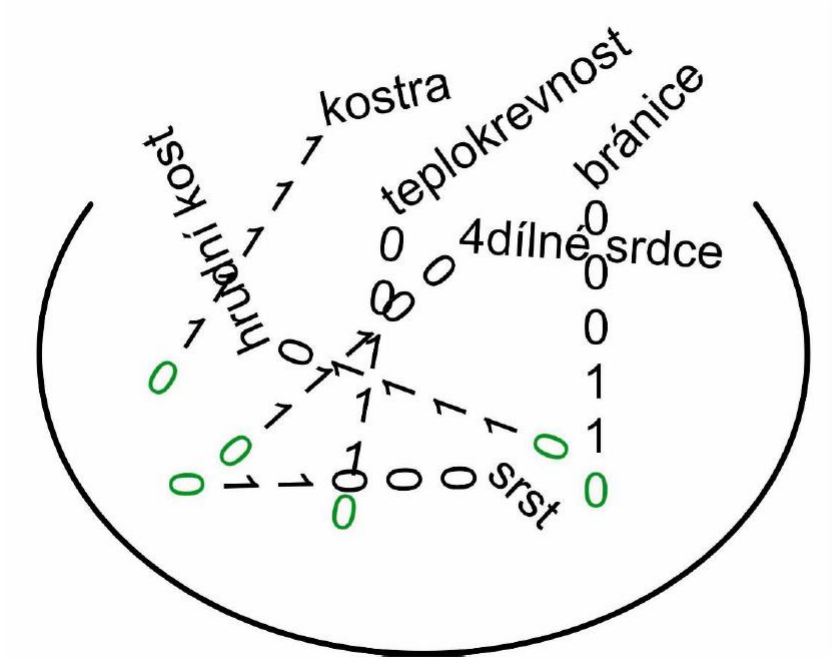
jaká je statistická podpora jednotlivých uzlů?

statistická podpora – bootstrapování

představte si situaci, kdy je některý uzel podpořen téměř všemi znaky, a jiný jen některými, které zrovna převáží...

	kostra	teplokrevnost	hrudní kost	čtyřdílné srdce	bránice	srst
ryba	1	0	0	0	0	0
žába	1	0	1	0	0	0
pták	1	1	1	1	0	0
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1
pavouk	*0*	*0*	*0*	*0*	*0*	*0*

	kostra	teplokrevnost	hrudní kost	čtyřdílné srdce	bránice	srst
ryba	1	0	0	0	0	0
žába	1	0	1	0	0	0
pták	1	1	1	1	0	0
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1
pavouk	*0*	*0*	*0*	*0*	*0*	*0*



následuje “losování”
a výroba tzv. **pseudomatic,**
které mají stejný počet
druhů i znaků;
znaky se mohou opakovat

Bootstrap - pseudomatice:

	srst	srst	bránice	teplokrevnost	teplokrevnost	teplokrevnost
ryba	0	0	0	0	0	0
žába	0	0	0	0	0	0
pták	0	0	0	1	1	1
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1
pavouk	0	0	0	0	0	0

strom 1

	bránice	4dílné srdce	teplokrevnost	teplokrevnost	srst	hrudní kost
ryba	0	0	0	0	0	0
žába	0	0	0	0	0	1
pták	0	1	1	1	0	1
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1
pavouk	0	0	0	0	0	0

strom 2

	kostra	hrudní kost	hrudní kost	teplokrevnost	bránice	bránice
ryba	1	0	0	0	0	0
žába	1	1	1	0	0	0
pták	1	1	1	1	0	0
králík	1	1	1	1	1	1
člověk	1	1	1	1	1	1
pavouk	0	0	0	0	0	0

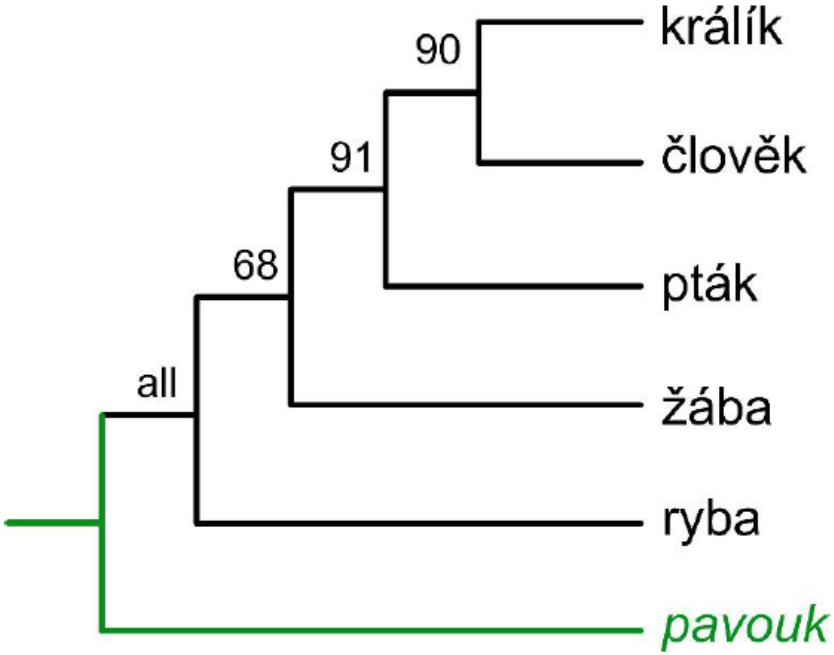
strom 3

500 - 1000x

...strom 1000

Bootstrap – výsledný strom:

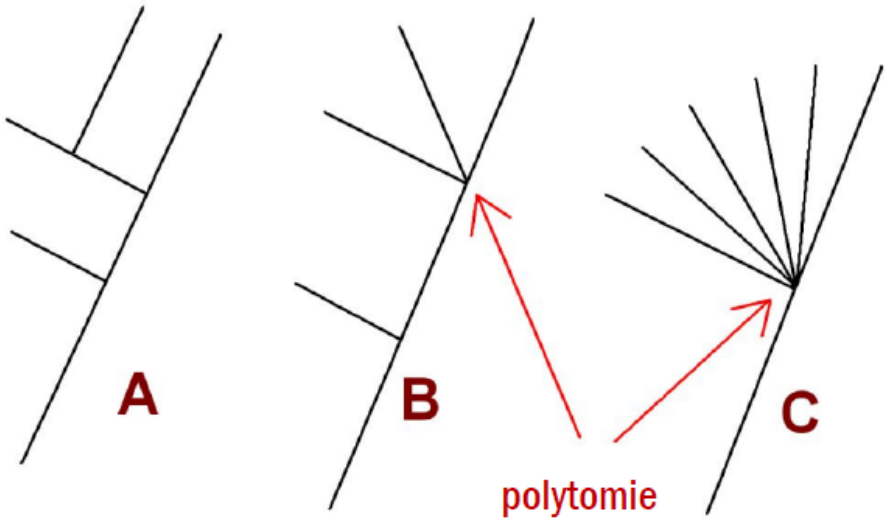
v kolika % z 1000 stromů z pseudomatic se uzel vyskytl ?



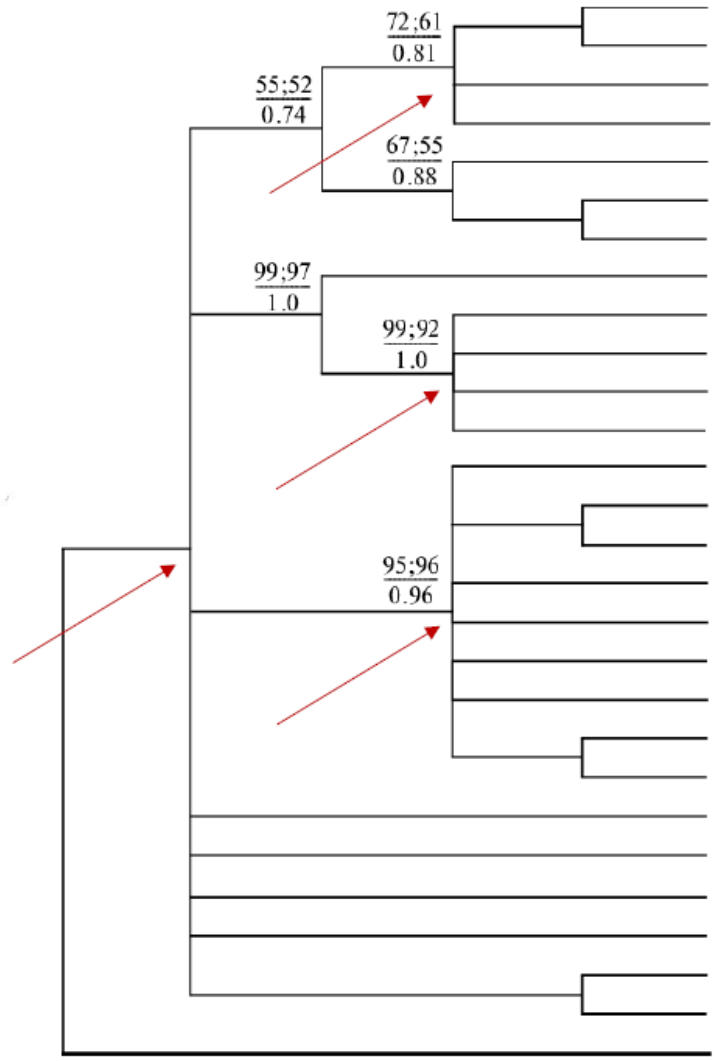
morfologie i DNA – některé znaky se opakují a jiné se do pseudomatice nedostanou, během 1000 opakování by mělo padnout dostatečné množství kombinací

pokud je topologie (=uzel) založena jen na málo znacích, bude podpora bootstrapu malá (pravděpodobnost, že se vylosují tyto znaky do pseudomatice je menší než když je topologie založena na mnoha znacích napříč datasetem).

rozřešený strom – založen na dichotomickém větvení:



reálný strom:

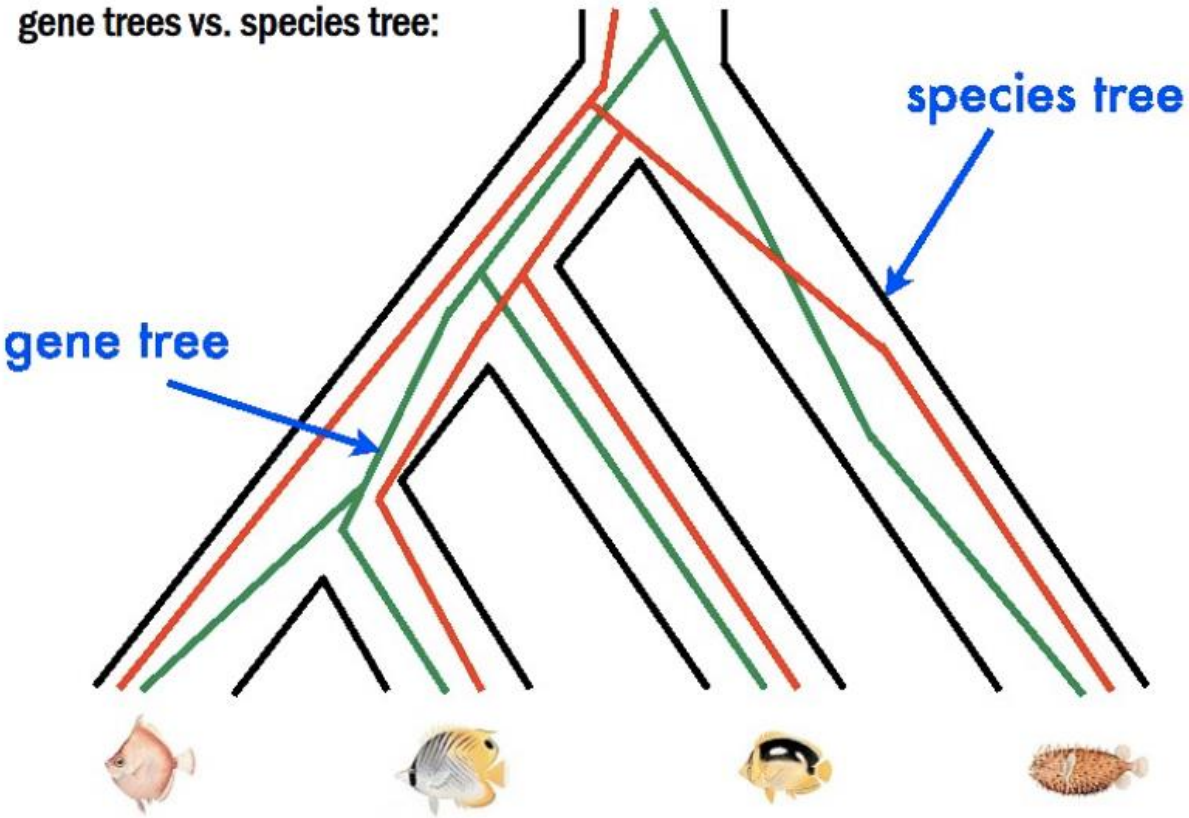


=> v uzlu, kde je polytomie nemáme dostatek informace abychom byli schopni vyřešit vztahy mezi potomky uzlu; jen velmi vyjímečně může jít o biologický jev (vyvinulo se více druhů najednou z 1 předka = např. adaptivní radiací)

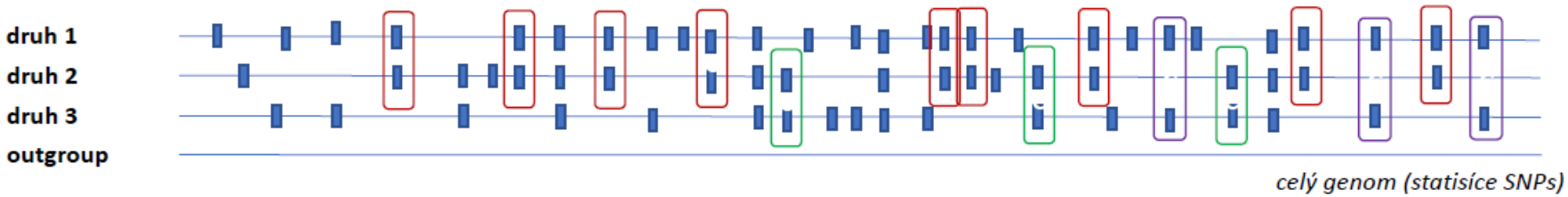
molekulární fylogeneze založená na sekvenci DNA různých genů:

- máme-li více genů, pak můžeme buď všechny sekvence jednoduše spojit za sebe (=konkatenovat) nebo počítat strom pro každý gen zvlášť

- počítat zvlášť je správnější – poté se udělá konsensus z jednotlivých genů

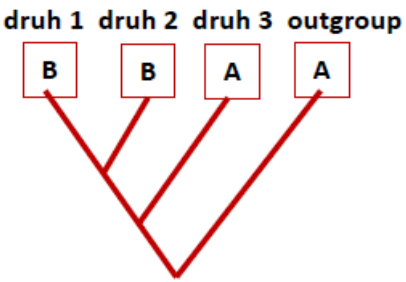


fylogenomika = fylogenetika na celých genomech!

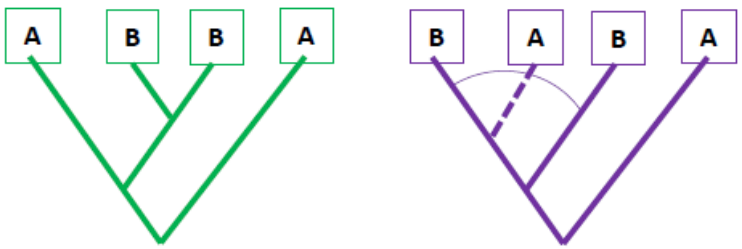


analýza podél genomu – jednotlivé záměny, či sliding window, postupné stromy, většina podporuje jednu topologii:

dominantní topologie:



alternativní topologie:

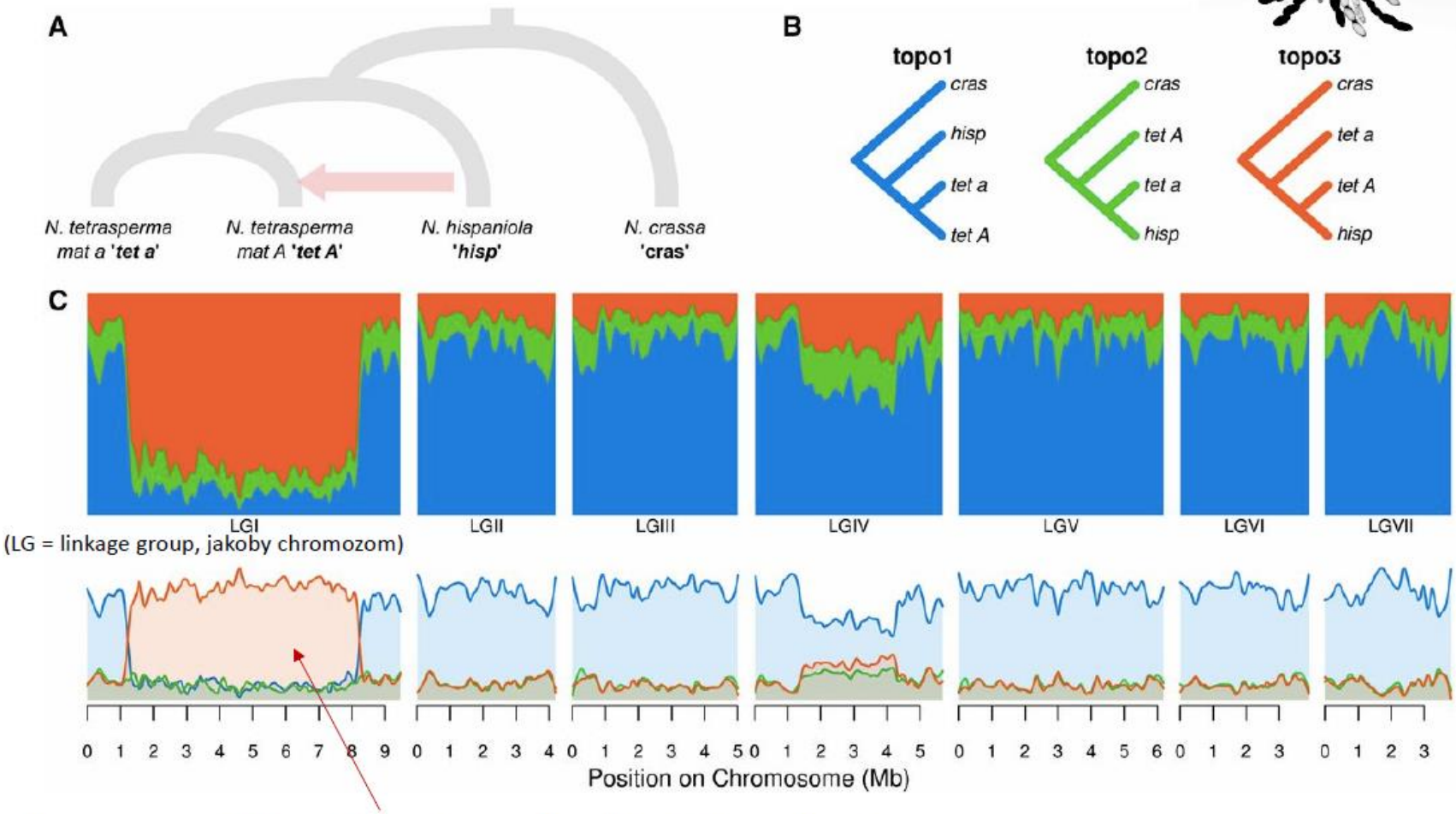


ABBA-BABA test:

=> testuje pozorované alternativní topologie – tedy ABBA nebo BABA => jsou-li v rovnováze, jde jen o stochasticitu, ale nejsou-li v rovnováze => odhalení případných hybridizací

fylogenomika = fylogenetika na celých genomech!

topologie se může měnit podél chromozomů

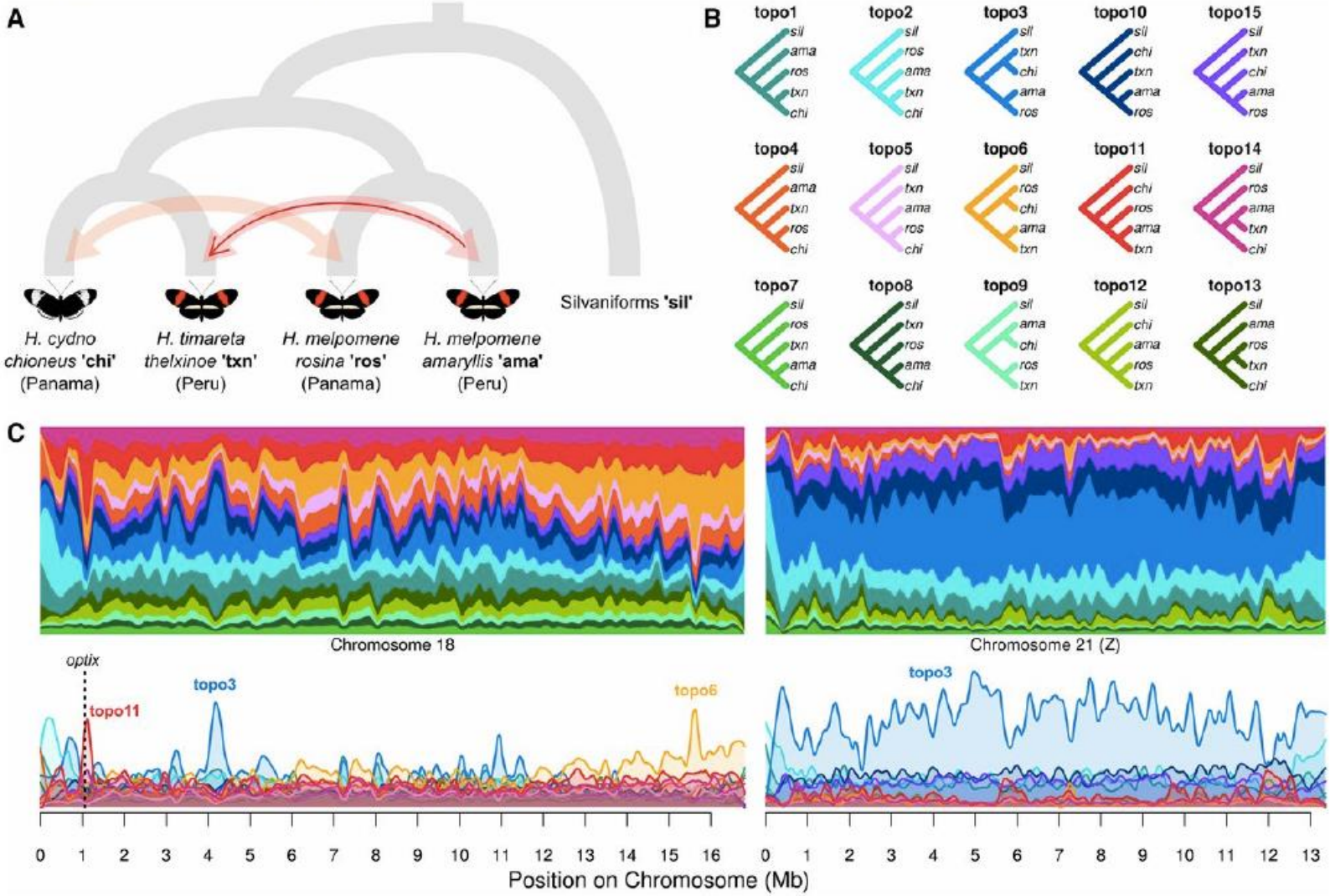


Neurospora – detekce introgrese a nerekombinujícího úseku v genomu

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, *206*(1), 429-438.

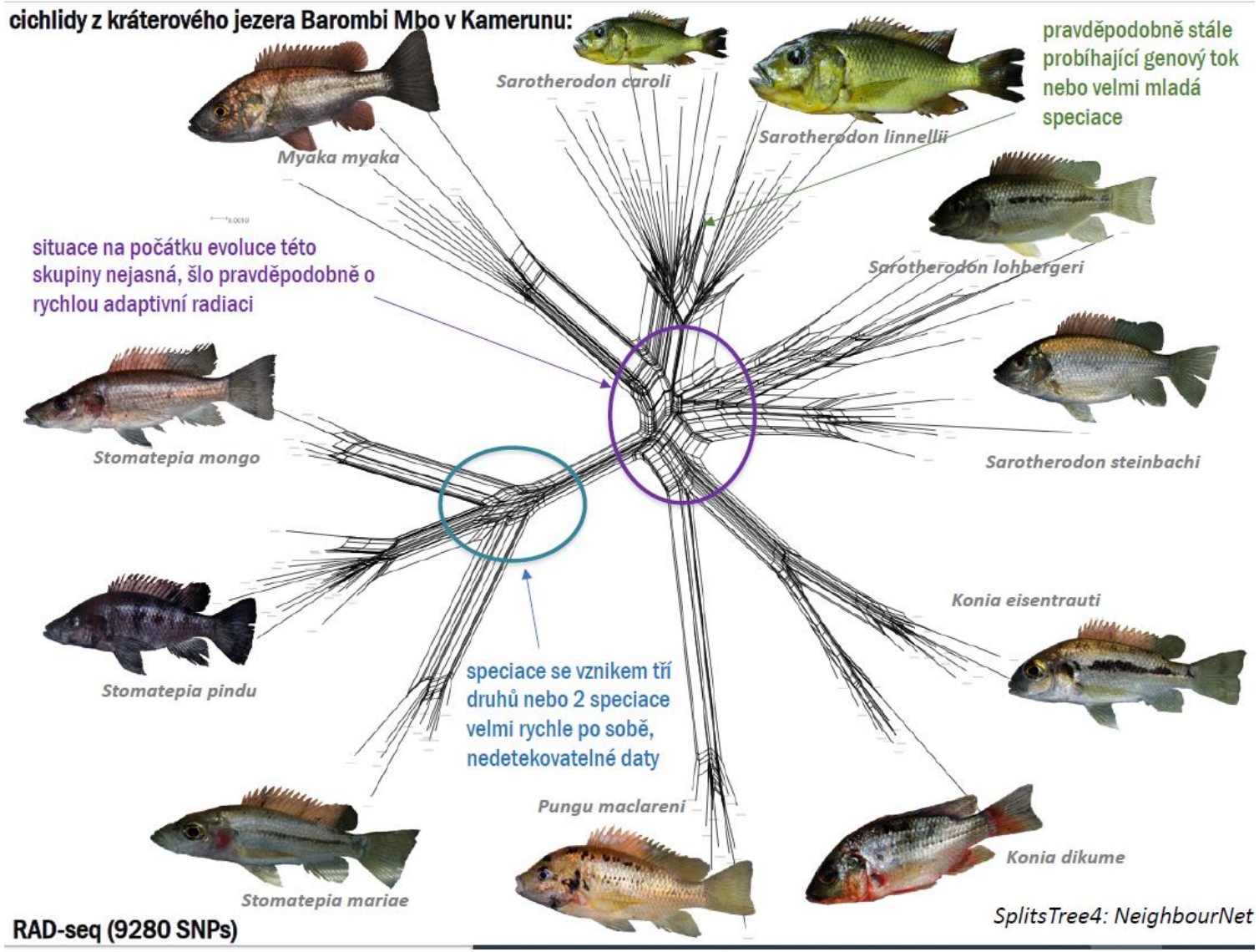
fylogenomika = fylogenetika na celých genomech!

topologie se může měnit podél chromozomů

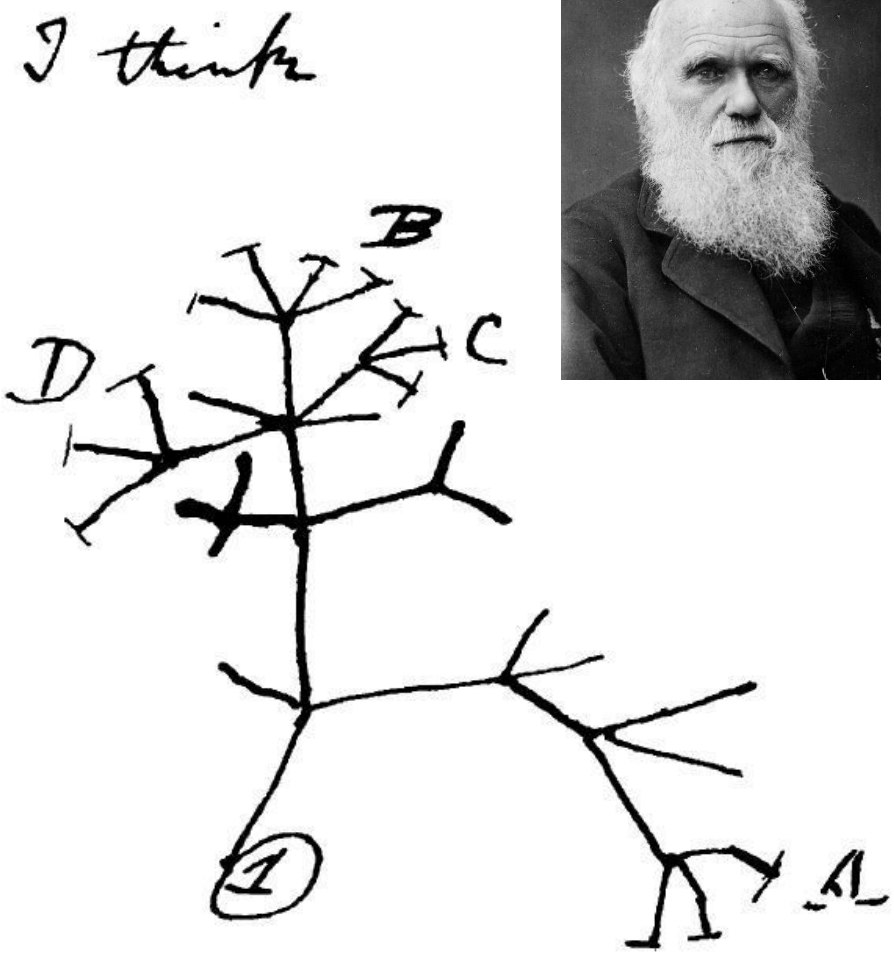


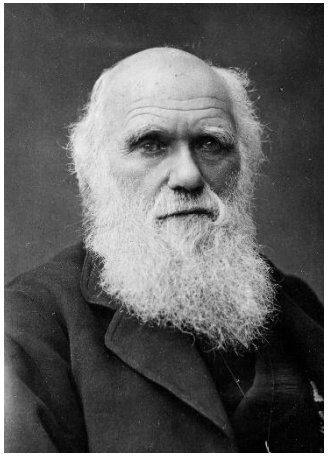
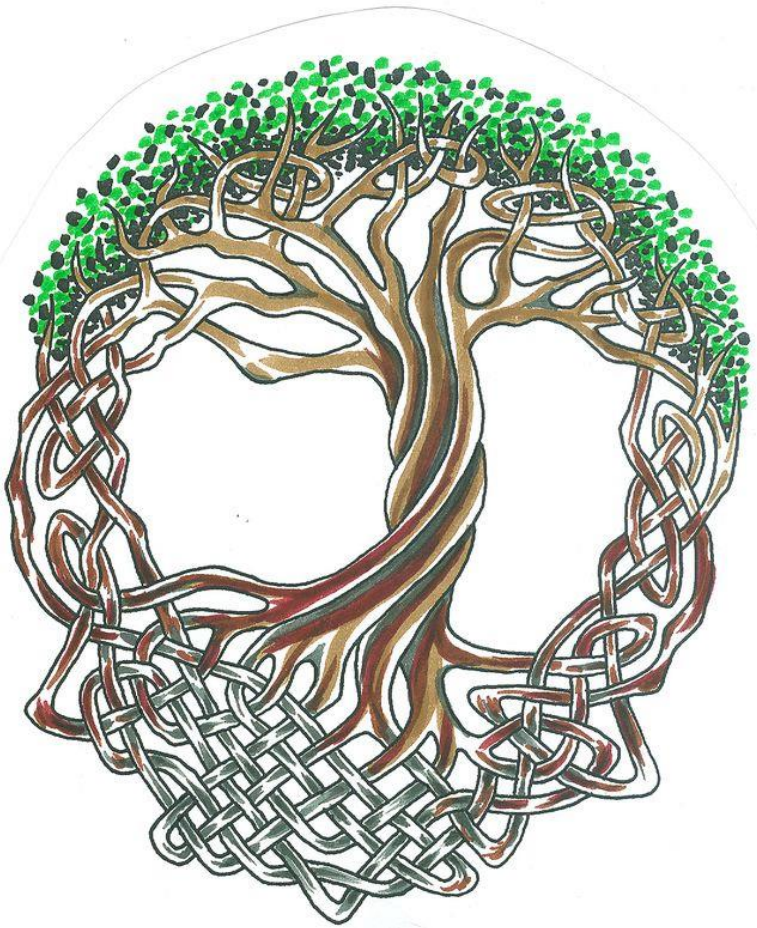
Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, *206*(1), 429-438.

Fylogenetické sítě



Strom života





Síť života