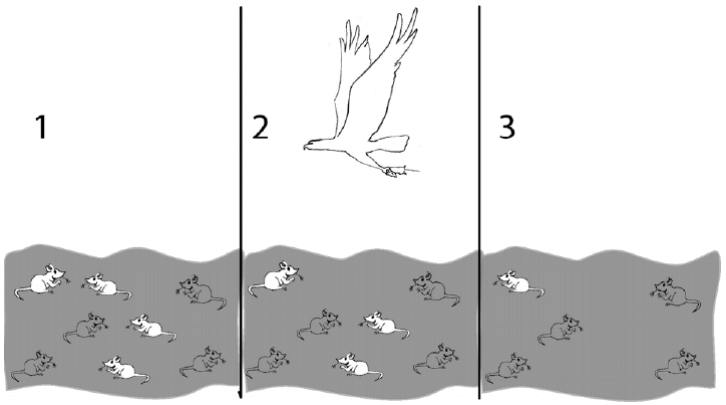
Sekvence jako data

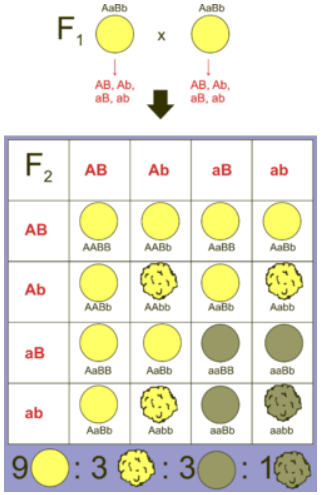
→ molekulární evoluce

Nejprve však malé **Opakování**

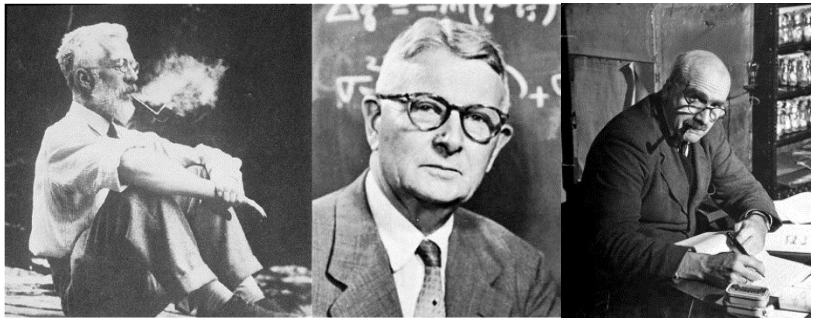
- Darwin 1859 On the Origin of Species by Means of Natural Selection
  - Natural selection
  - Variabilita, selekce (survival of the fittest)



- Mendel 1866 Versuche über Pflanzenhybriden
  - Dědičnost, znaky (Merkmal)



Moderní syntéza



- Fisher, Wright, Haldane, první polovina minulého století

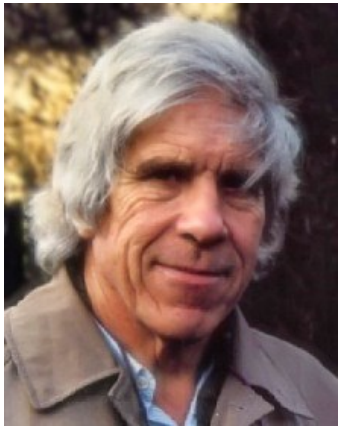
- Darwin + Mendel + matematické modely

- Evolution = changes of allele frequencies over time

  - *K dispozici jen fenotypy*
  - *1944 Avery, MacLeod a McCarty - DNA jako nositel dědičné informace*
  - *1953 struktura DNA*
  - *Šedesátá léta - první sekvence, genetický kód*
  - *Sedmdesátá léta - sekvenování (Sanger)*
  - *Osmdesátá léta - PCR*

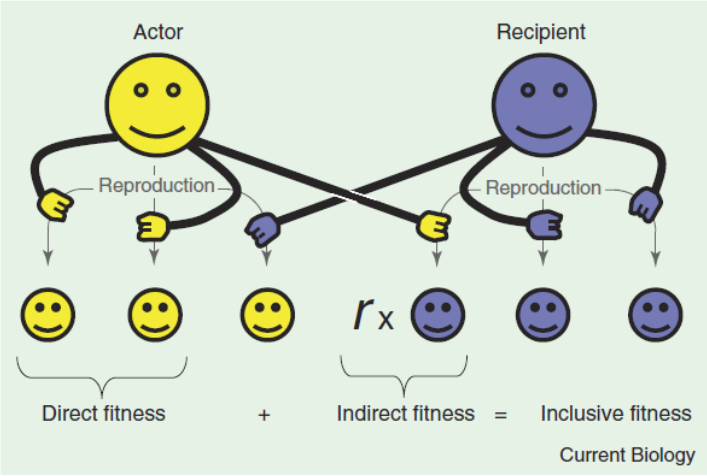
- Minulá přednáška

Rozšíření



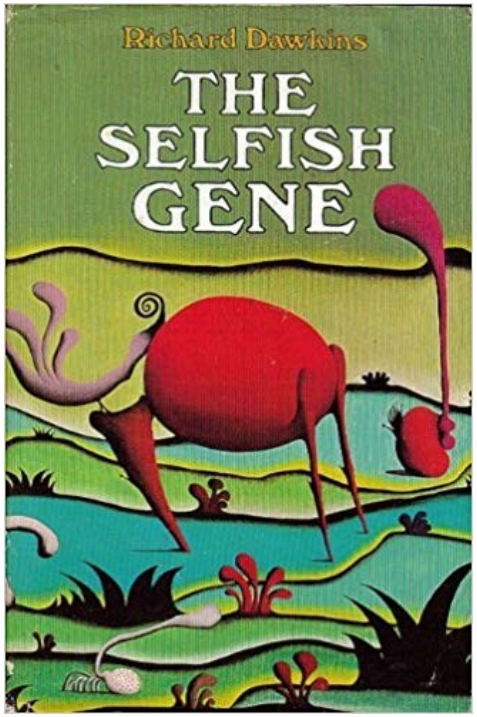
- **W. D. Hamilton, 1964, Inclusive fitness of an allele**
  - Inclusive fitness = Direct fitness + Indirect fitness
  - Direct fitness – fitness of the individual bearing the allele
  - Indirect fitness – fitness of other individuals carrying copies of the same allele
  - **Hamilton's rule**
  - a behaviour or trait will be favoured by selection (altruism spreads), when

    $rb - c > 0$
  - *c - the fitness cost to the actor*
  - *b - the fitness benefit to the recipient*
  - *r - their genetic relatedness*
  - *r - fraction of genes that are identical by descent*
  - Pozor je například i reciproční altruismus



Popularizace

- R. Dawkins 1976 Sobecký gen (The Selfish Gene)
- Gen ve smyslu alela
- Organismy jako „vehikuly ovládané sobeckými alelami“

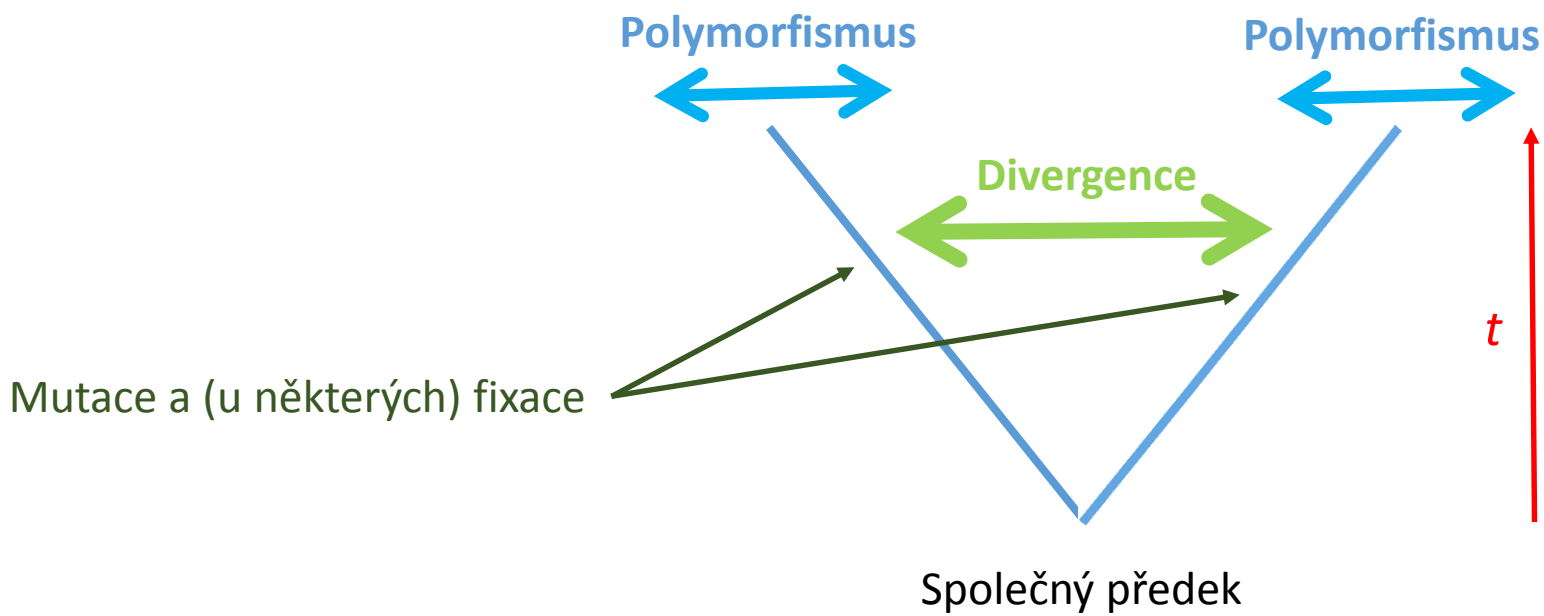


Examining a cricket in Oxford in the 1970s

Divergence a polymorfismus

- **Polymorfismus**
  - existence rozdílů uvnitř populace
    (více alel na lokusu, více variant na dané pozici DNA)
  - Neproběhla fixace

- **Divergence**
  - Fixované rozdíly nahromaděné v liniích od rozdělení od společného předka



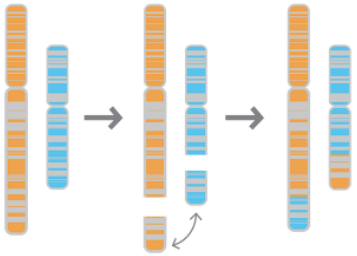
Rozdíly v sekvencích DNA
Mutace

```
cow       ATG---ACTAACATTCGAAAGTCCCACCCACTAATAAAAATTGTAAAC
sheep     ATG---ATCAACATCCGAAAAACCCACCCACTAATAAAAATTGTAAAC
goat      ATG---ACCAACATCCGAAAGACCCACCCATTAATAAAAATTGTAAAC
horse     ATG---ACAAACATCCGGAAATCTCACCCACTAATTAAAATCATCAAT
donkey    ATG---ACAAACATCCGAAAATCCCACCCGCTAATTAAAATCATCAAT
ostrich   ATGGCCCCCAACATTCGAAAATCGCACCCCCTGCTCAAAATTATCAAC
emu       ATGGCCCCTAACATCCGAAAATCCCACCCTCTACTCAAAATCATCAAC
turkey    ATGGCACCCAATATCCGAAAATCACACCCCCTATTAAAAACAATCAAC
```

```
1 ATTGCCGACCCTAGCCTAGGCTATTTGACCTGAGATTCCATTAATTGGGCGC
2 ---------T-----------------------------------------
3 ---------T------C----------------------------------
```

- Mutace jsou různé
  - Chromosomové přestavby (včetně inverzí)
  - Delece, inzerce (třeba transposony, některé pseudogeny)
  - A další
- Nás budou zajímat záměny nukleotidů (ATCG)

Fig. A Reciprocal translocation



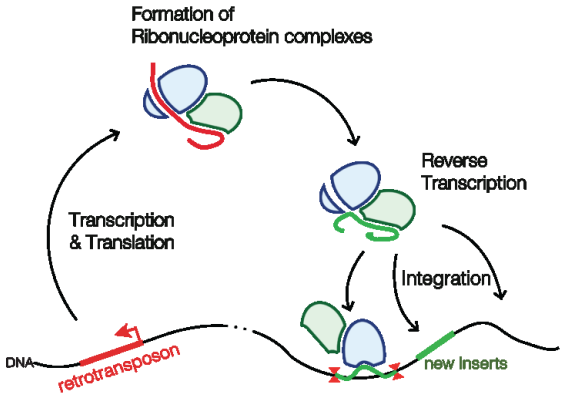
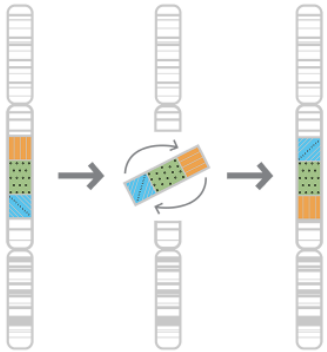


Fig. B Inversion



Sekvence DNA

```
cow       ATG---ACTAACATTCGAAAGTCCCACCCACTAATAAAAATTGTAAAC
sheep     ATG---ATCAACATCCGAAAAACCCACCCACTAATAAAAATTGTAAAC
goat      ATG---ACCAACATCCGAAAGACCCACCCATTAATAAAAATTGTAAAC
horse     ATG---ACAAACATCCGGAAATCTCACCCACTAATTAAAATCATCAAT
donkey    ATG---ACAAACATCCGAAAATCCCACCCGCTAATTAAAATCATCAAT
ostrich   ATGGCCCCCAACATTCGAAAATCGCACCCCCTGCTCAAAATTATCAAC
emu       ATGGCCCCTAACATCCGAAAATCCCACCCTCTACTCAAAATCATCAAC
turkey    ATGGCACCCAATATCCGAAAATCACACCCCCTATTAAAAACAATCAAC

1 ATTGCCGACCCTAGGCTAGGCTATTTGACCTGAGATTCCATTAATTGGGCGC
2 ---------T------------------------------------------
3 ---------T------C-----------------------------------
```

- Klasická populační genetika – alely
- Sekvence – nukleotidové polymorfismy (SNPs)
- Jednotlivé SNP – problém s LD
- Celé haplotypy – ztrácíme spoustu informace
- → jiné přístupy a modely

The neutral theory of molecular evolution (1968)

- Evoluce na molekulární úrovni
- Velká variabilita (proteinů)
- Selekce (balancing)?
- Lokusů příliš mnoho
- → extinkce
- Většina mutací asi neutrální (tedy hodně malé ***s***)

Motoo Kimura

Mutační rychlost

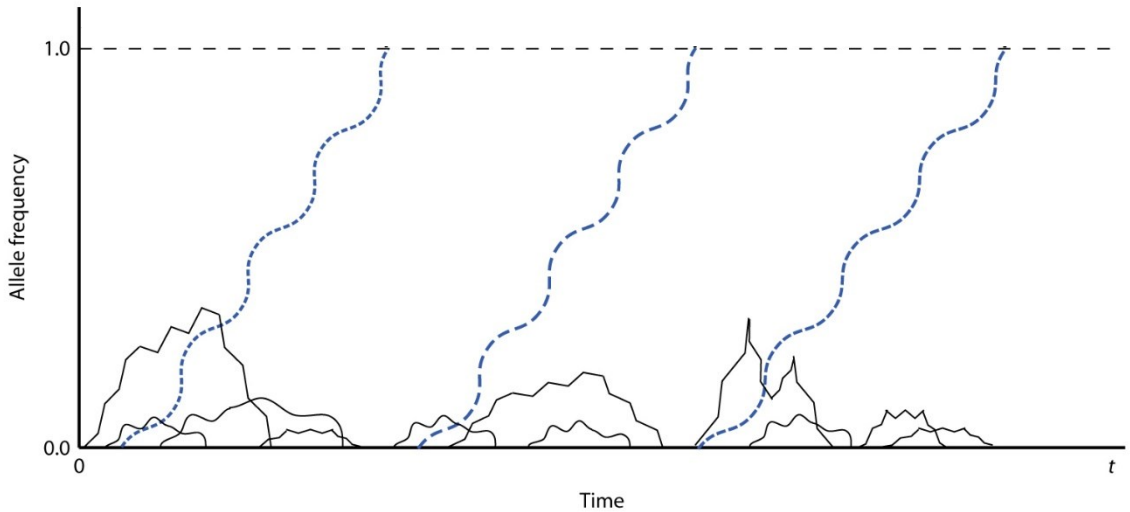
- μ neutral mutation rate

- pravděpodobnost vzniku mutace (na určitém místě)
  - Člověk 10^{-8}
  - Genom přes 3 giga (10^9)

- V populaci $2N_e\mu$ (2 – diploidní organismy)

- Větší populace – vznikne víc mutací

Drift a fixace alely



- Drift závisí na velikosti populace

- Pravděpodobnost fixace $1/2N_e$ (nové neutrální mutace)

Substituční rychlost

- **Substituce** – úplné nahrazení původní varianty jinou (vzniklou mutací), fixace
- Akumulace nových fixovaných variant
- Mutace $2N_e\mu$ a drift $1/2N_e$
- $2N_e\,\mu/2N_e = \mu$ *(mutační rychlost)*
- Substituční rychlost nezávisí na velikosti populace
- Stejná akumulace ve velkých i malých populacích
- Nulový model, neutrální mutace

Odhad genetické divergence

sekvence druhu 1 AGTGAGTCGTCAGTACTGCTG
sekvence druhu 2 ACTTAGCCGTGAGTACAGCTA

$$D = \frac{6}{21} = 0{,}2857\ (28{,}57\%)$$

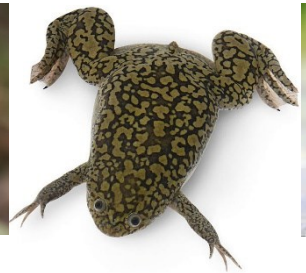
Proporce nukleotidových míst, které se liší mezi dvěma sekvencemi DNA.

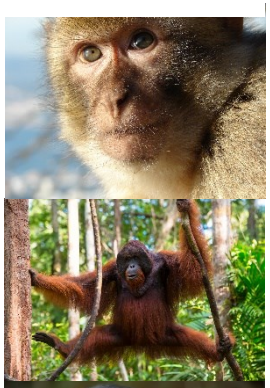
Někdy také ***p* distance** (proportion).

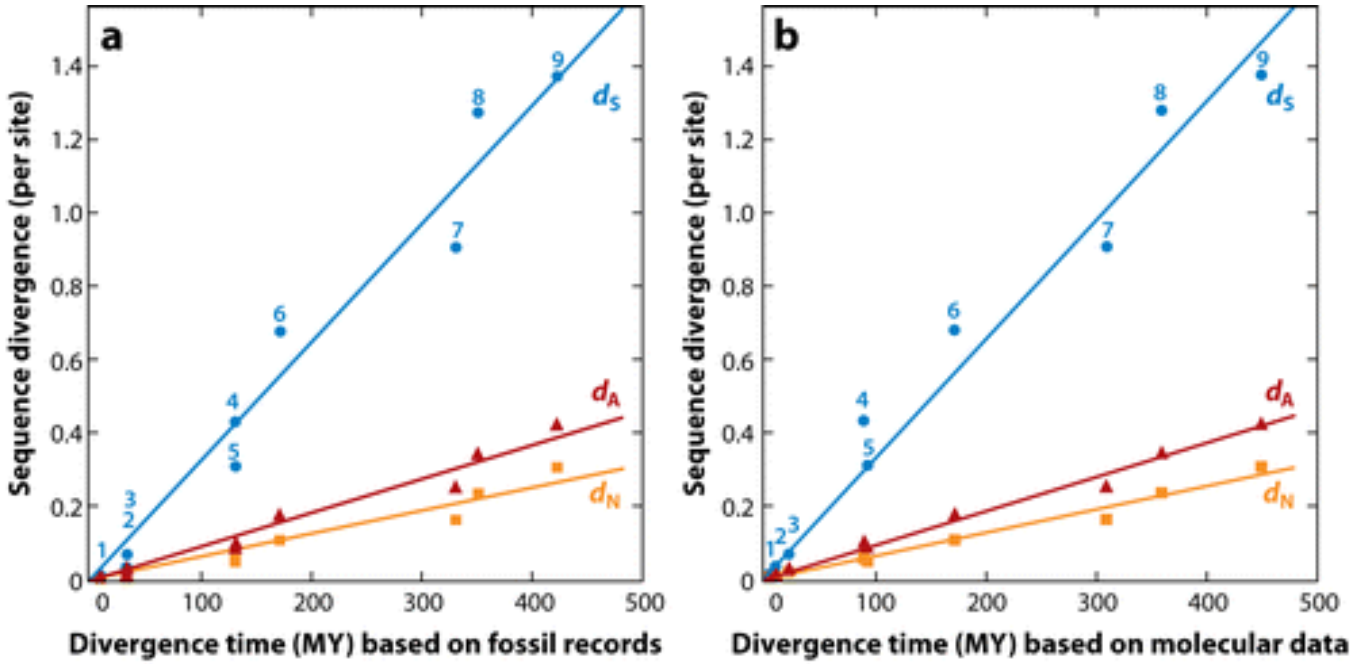
AGTGAGTCGTCAGTACTGCTG
sekvence druhu 1 ACTGAGTCGACAGTACTGCTG
AGTGACTCGTCAGTACTGCTG

ACTTAGCCGTGAGTACAGCTA
sekvence druhu 2 ACTTAGCGGTGAGTACAGCTA
ACTTAGCCGTGAGTACAGCAA

Dxy, průměrná divergence
(avergage pairwise divergence)







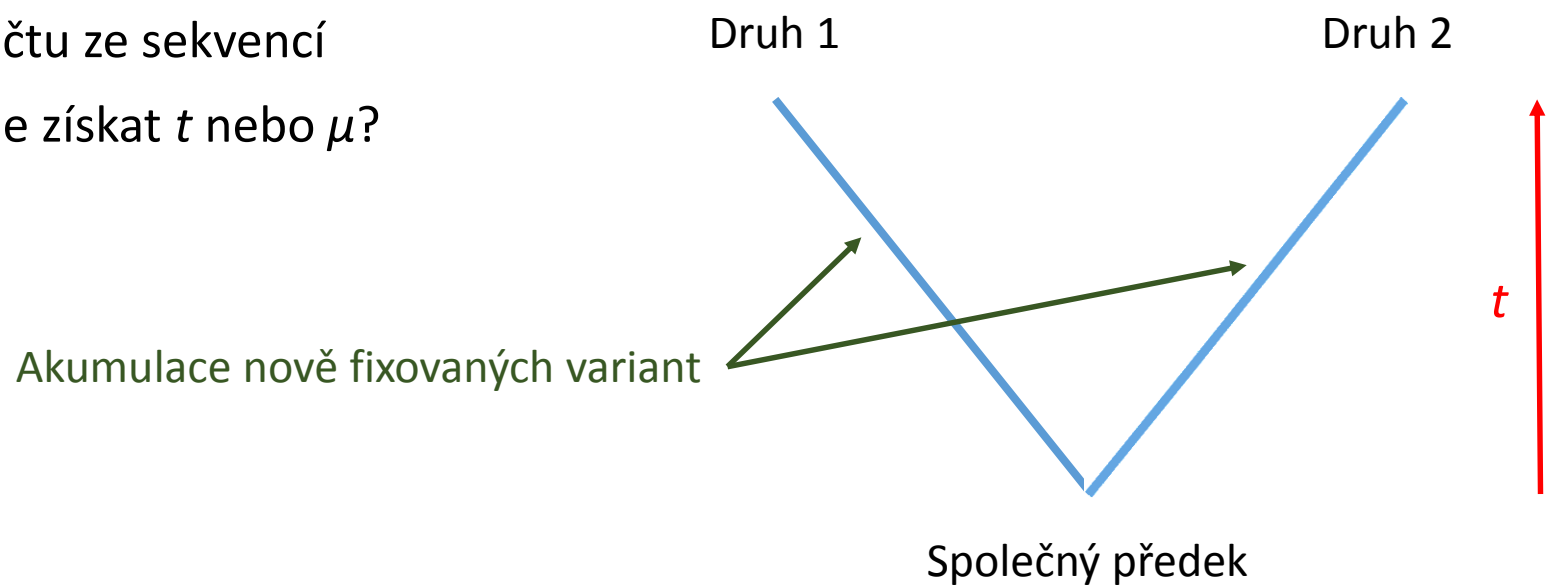
Nei M, et al. 2010.
Annu. Rev. Genomics Hum. Genet. 11:265–89

Molekulární hodiny

- Očekávaná divergence mezi druhy $E(d) = 2\,\mu t$
- *Divergence = průměrný rozdíl (v procentech)*

- → $t = d/2\mu$

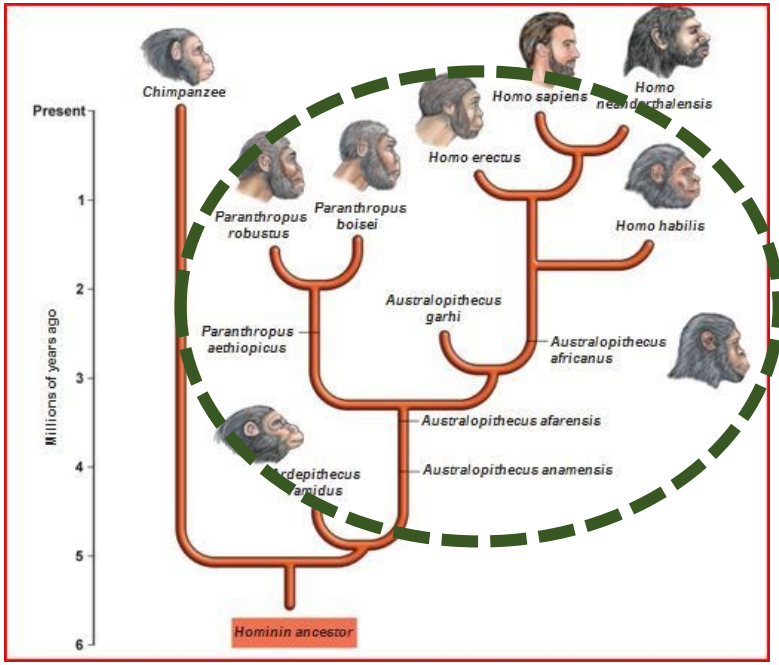
- Z d a μ tedy vypočtu t
- … a naopak z d a t tedy vypočtu μ
- d spočtu ze sekvencí
- Jak ale získat t nebo μ?



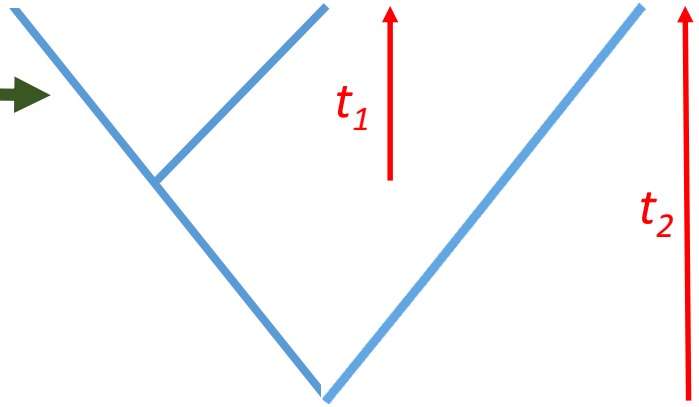
Kalibrace molekulárních hodin

- μ ← srovnání rodičů a potomků,
- mutací málo → mnoho potomků nebo hodně (bezchybných) sekvencí

- Kalibrace pomocí fosilií – většinou spíš jen limity



Dr. Emile Zuckerkandl and Dr. Linus Pauling



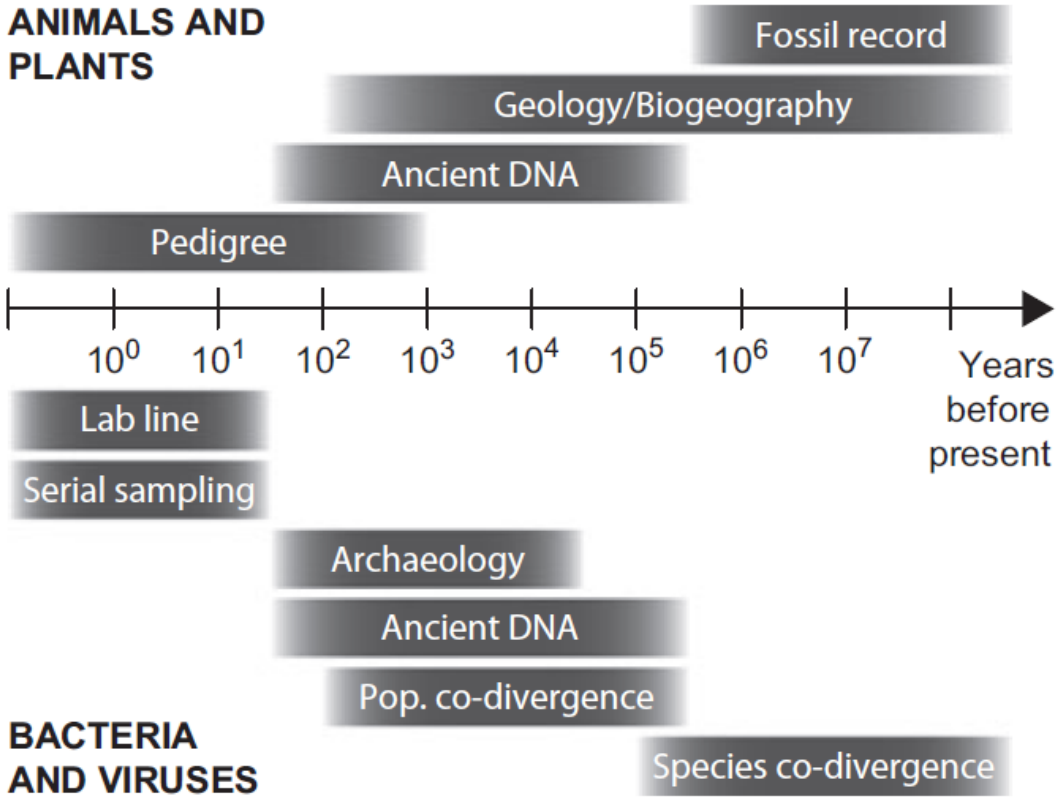
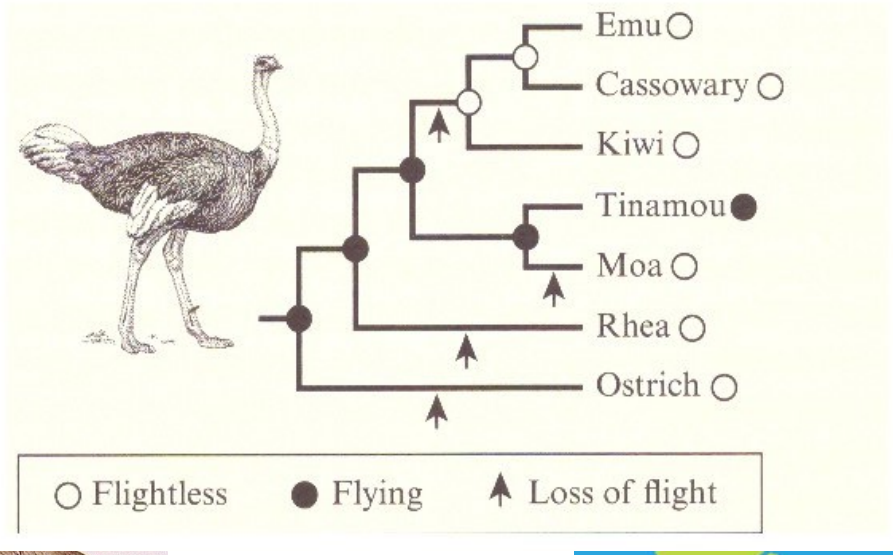


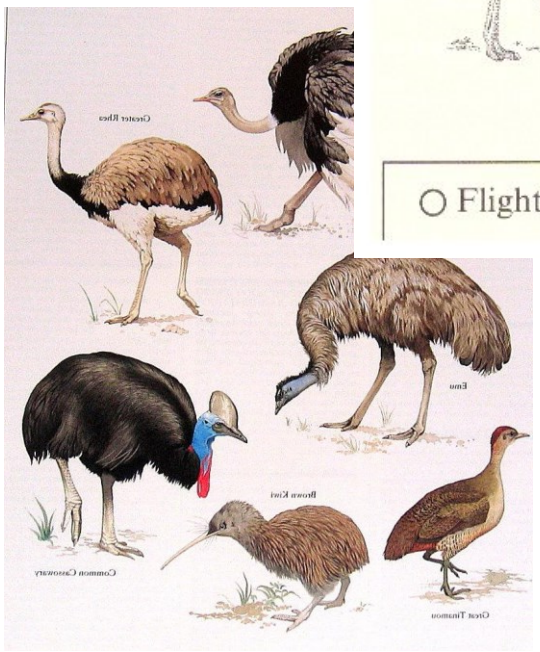
Fig. 2 Typical age ranges of different forms of calibrating information.

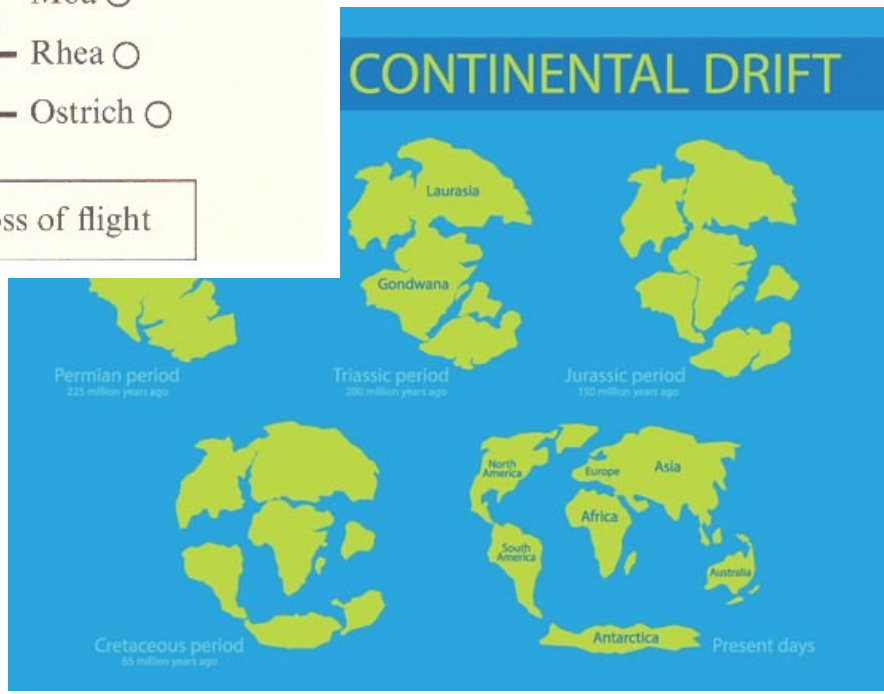
Kalibrace pomocí geologických událostí biogeografie, vikariance

- Vznik panamské šíje, rozpad Gondwany…
- Problematické, často podcenění disperze
- Příklad: běžci a rozpad Gondwany



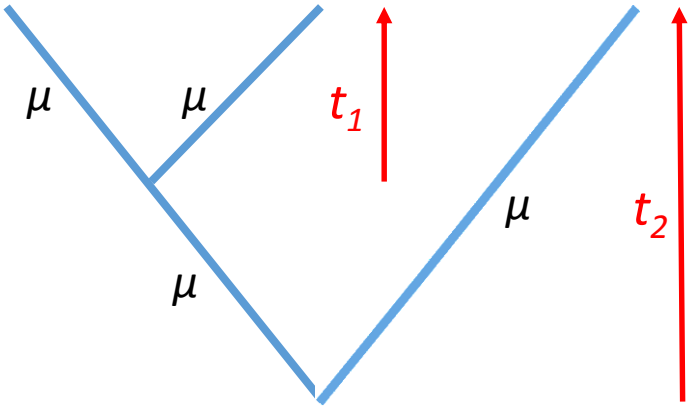
Alfred Wegener 1915



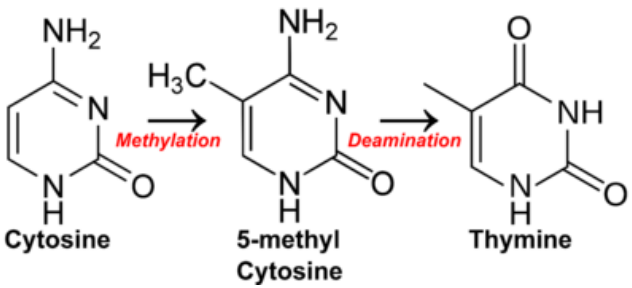


Předpoklad konstatní mutační a substituční rychlosti

- Zásadní pro molekulární hodiny
- Tedy předpoklad, že vše mutuje stejně rychle a nic není pod selekcí

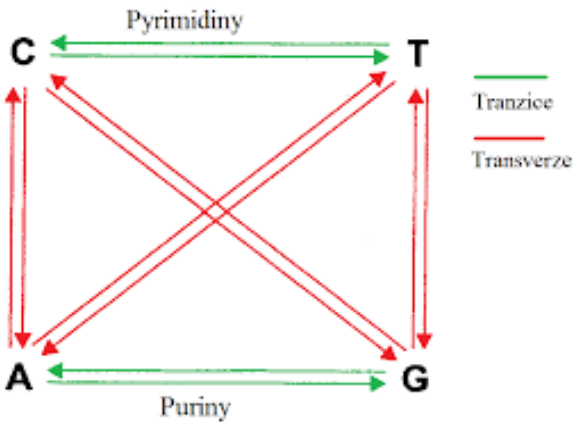


Mutační rychlost

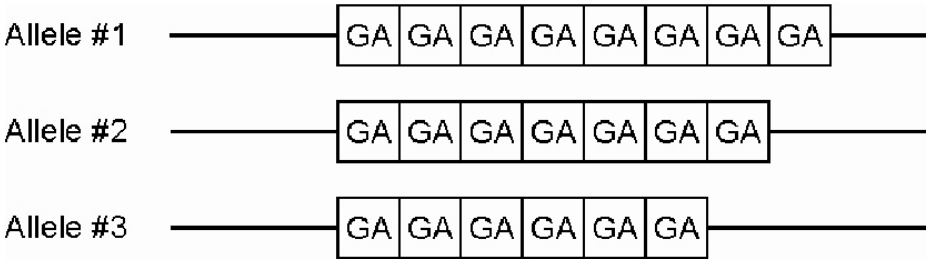


Je odlišná pro různé typy mutací

- rozsáhlejší mutace obecně méně časté
- mutační „hotspots“: CpG u živočichů, mikrosatelity.
- tranzice vznikají častěji než transverze

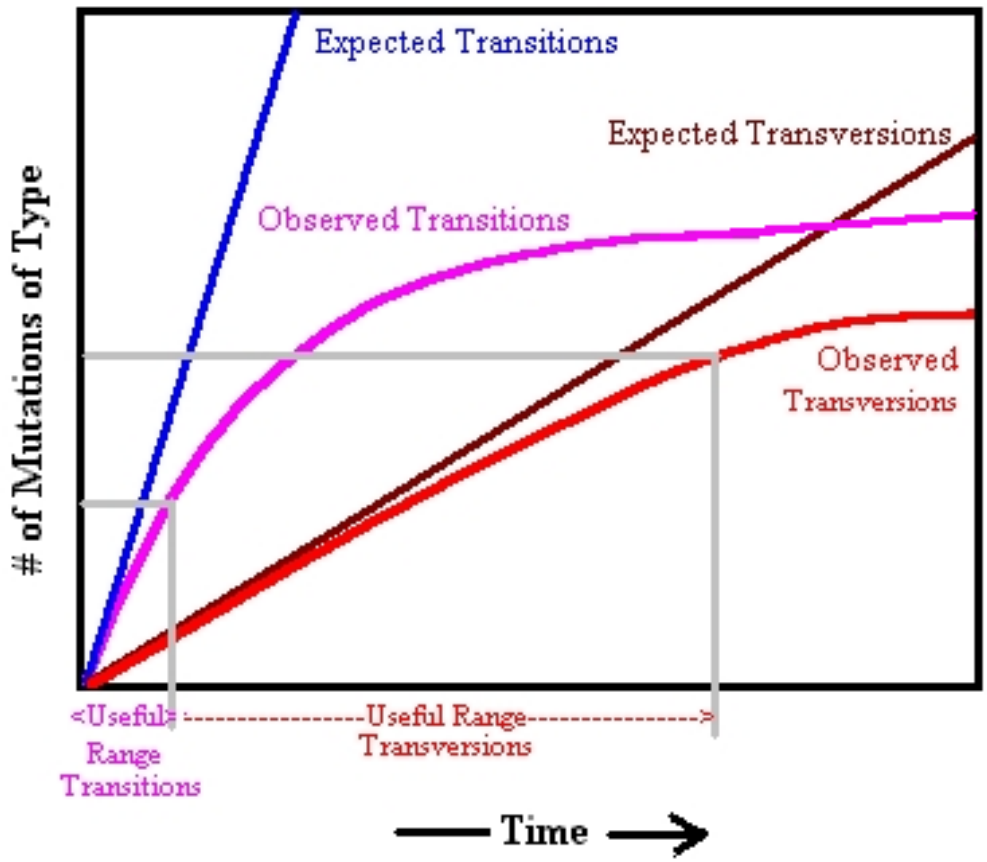


mikrosatelit



Saturace mutací

- Po určité době divergence přestane stoupat. Je to kvůli mnohočetným substitucím ve stejné pozici. Došlo k saturaci.



Nukleotidové substituční modely

Umožňují korekci na mnohonásobné substituce a odhad skutečné divergence.

Jukes & Cantor (1 parameter) model
-Nejjednodušší jednoparametrový model
-Předpokládá, že všechny možné substituce stejnou pravděpodobnost (α)
a všechny nukleotidy mají stejnou frekvenci.

Kimura (2 parameter) model
-Pravděpodobnost tranzic větší než pravděpodobnost transverzí

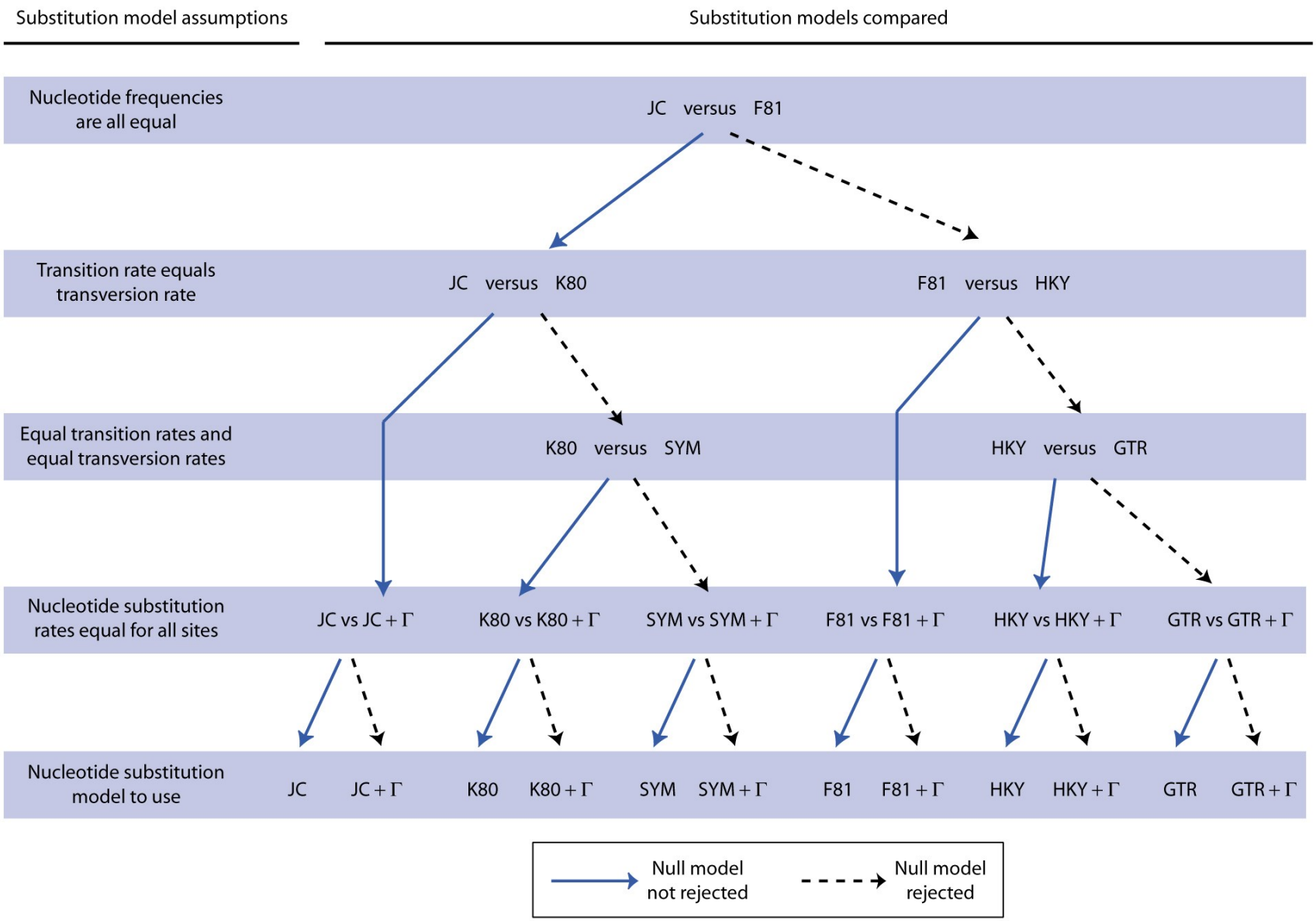
Felsenstein 81 (4 parameter) model
-Umožňuje nastavit různé frekvence nukleotidů

HKY-Hasegawa,Kishino,Yano (5 parameter) model
-Zohledňuje odlišnou pravděpodobnost transic a transverzí i různé frekvence nukleotidů

HKY + Γ
-Gama parametr umožňuje modelovat odlišnou rychlost substitucí v různých nukleotidových pozicích

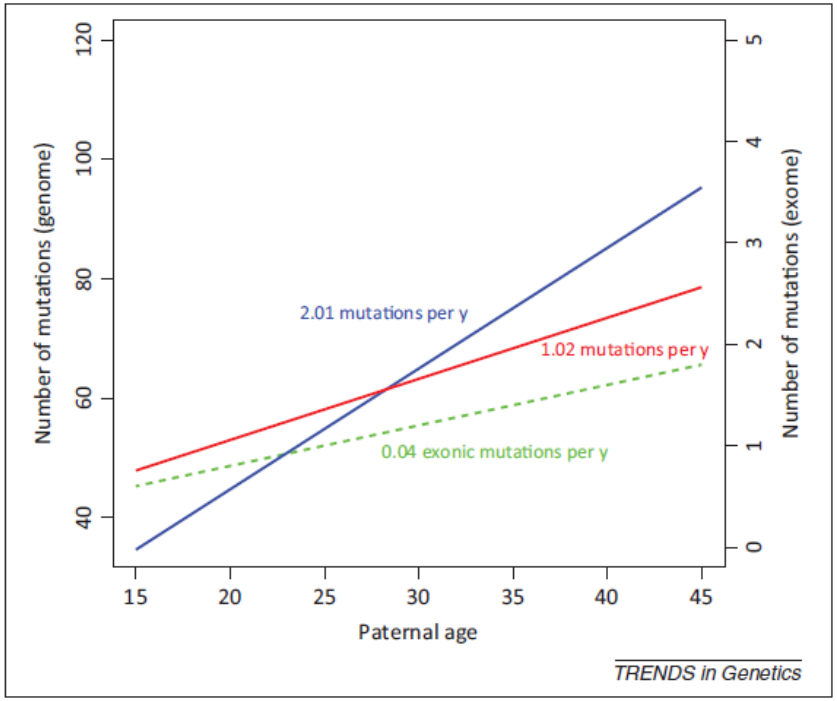
Program ModelTest: vybere model, který nejlépe sedí na data

Nukleotidové substituční modely



Většinu mutací dědíme od otců

- Většina nukleotidových substitucí vzniká v samčí germinální linii (více replikací DNA) (**male-driven evolution**).
- Počet nově vzniklých nukleotidových substitucí vzrůstá s věkem otce (**paternal age effect**).
- Rozdíl v mutační rychlosti mezi autosomy, mtDNA a pohlavními chromosomy.

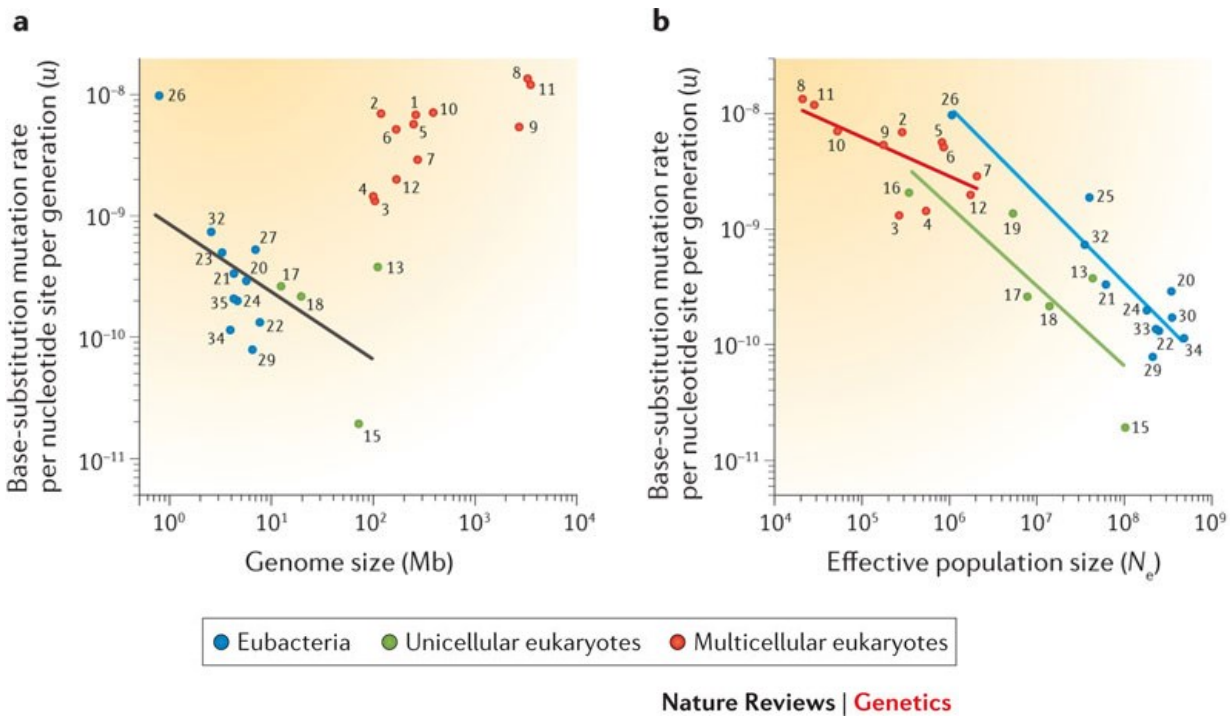


Campbell and Eicher, 2013

Mutační (substituční) rychlost

Liší se u různých organismů

- velikost genomu
- životní strategie (velikost těla, intenzita metabolismu, generační doba, dlouhověkost, efektivní velikost populace…)



Lynch et al. 2016

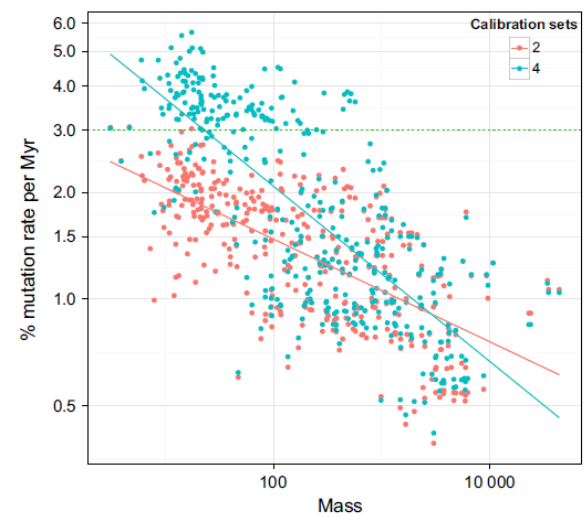
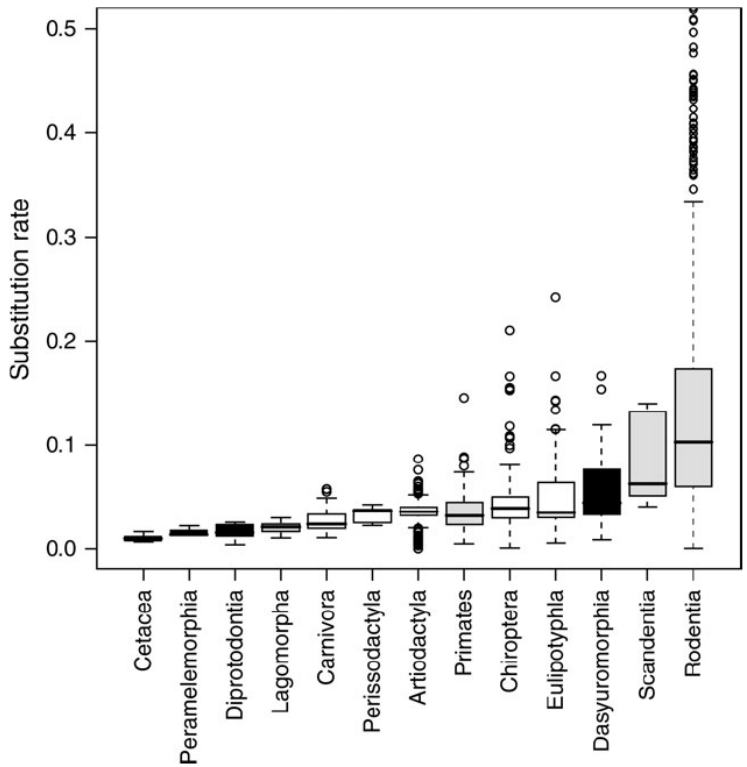


Fig. 1 Correlation between body mass and third codon position substitution rates. The dotted green line indicates an approximation of the rate corresponding to a 2% divergence per million years.

Rozdíly v substituční rychlosti

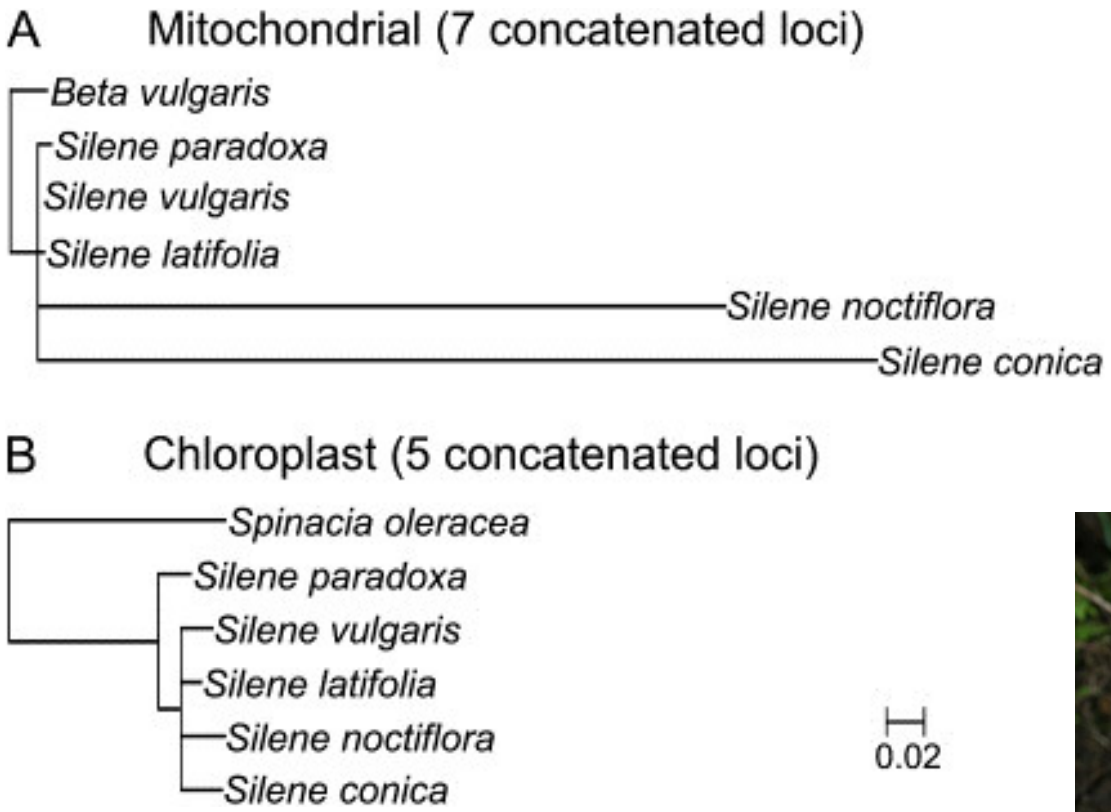
Mezi taxony (řádové rozdíly na jednotku času)

Příklad – savci a cytochrom b



Gerbillus nanus
0.741 substitution/site/Myr

Balaenoptera borealis
0.007 substitution/site/Myr



Branch lengths correspond to the number of synonymous substitutions per site for mitochondrial DNA (A) and chloroplast DNA (B)

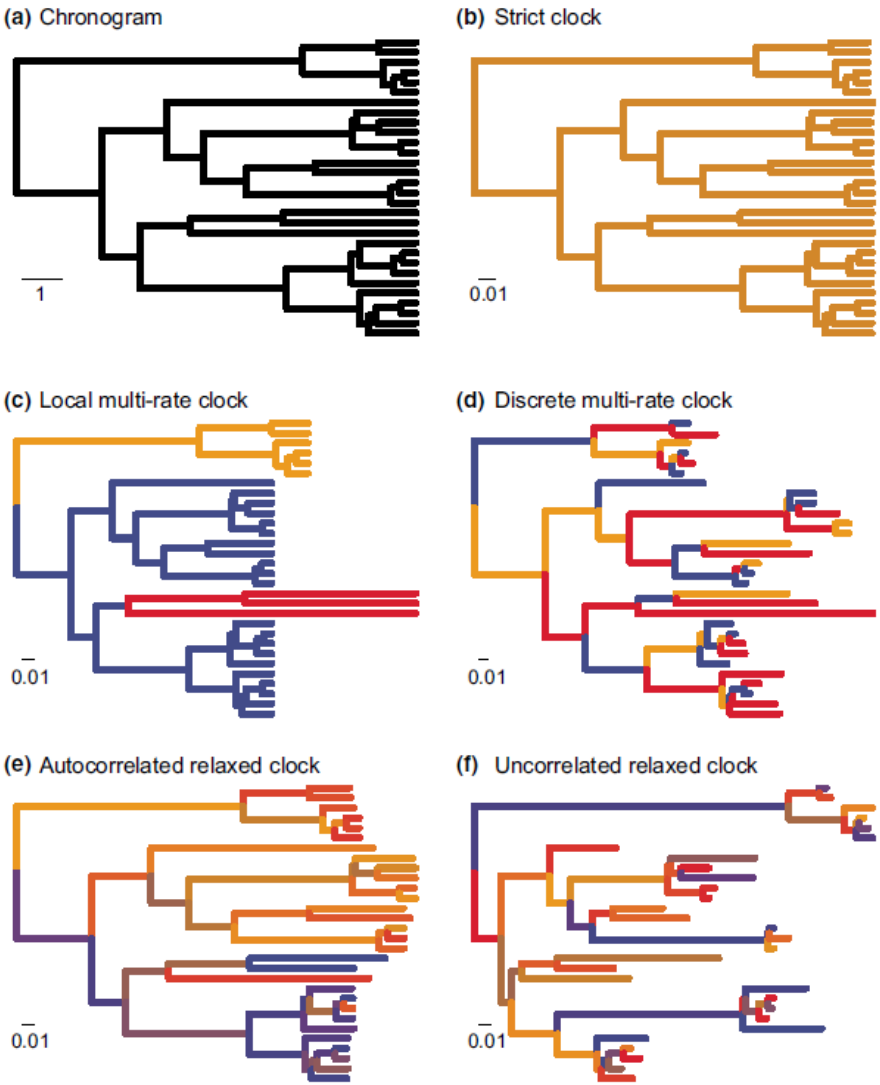
Molekulární hodiny

Striktní molekulární hodiny

- Substituční rychlost homogenní napříč fylogenezí.

Relaxované molekulární hodiny

- Berou v potaz možnou variabilitu v substituční rychlosti mezi liniemi.

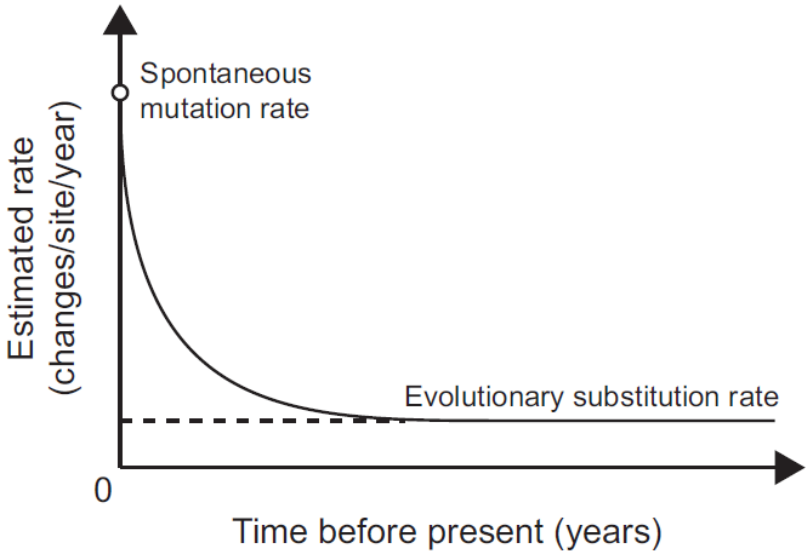


Ho and Duchene, 2014

Nearly neutral theory…

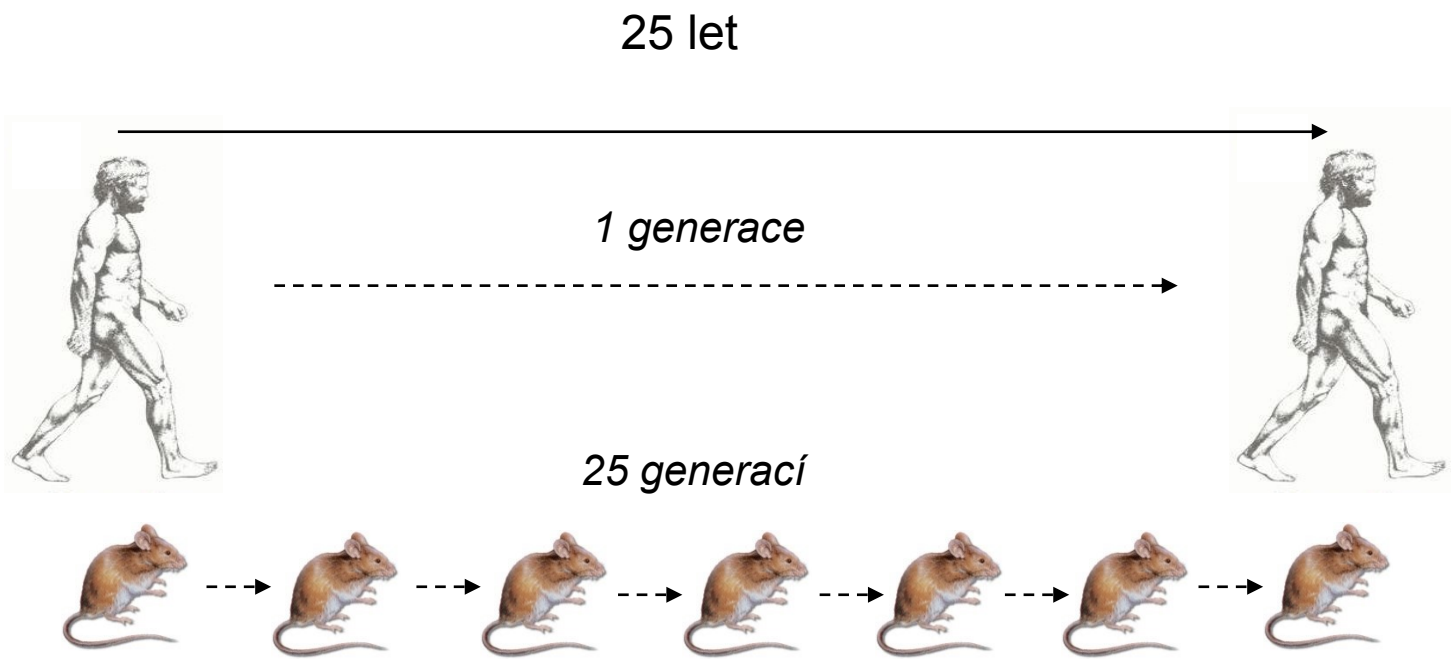
Tomoko Ohta

- Řada mutací nemusí být neutrální
- Slightly or weakly deleterious mutations
- Vliv driftu i selekce $2s = 1/2\ N_e$ (*s* selekční koeficient, viz minulá přednáška)
- Druhy s **kratší** generační dobou
  - → větší N_e
  - → **selekce** proti škodlivým mutacím
  - → **zpomalení** mol. Evoluce
- Druhy s **delší** generační dobou
  - → menší N_e
  - → převládne **drift**
- Toto je ale na **počet generací**

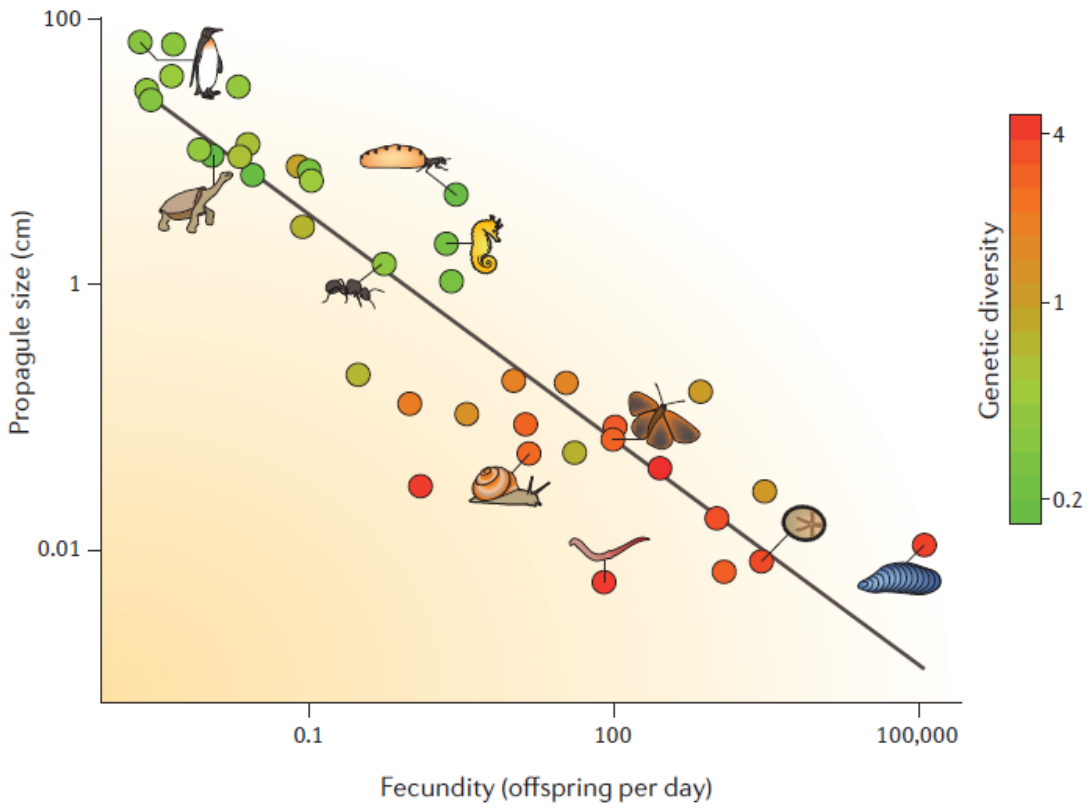


Generační doba

- Druhy s kratší generační dobou (obvykle menší velikost těla) mají vyšší substituční rychlost (na rok). Probíhá u nich více buněčných dělení v germinální linii za jednotku času.



Genetická diverzita organismů závisí na životních strategiích organismů



Drosophila 3 % genomu variabilní
Člověk 0,1 %

Diverzitu populace vyjadřije population (scaled) mutation rate parameter *Θ*

- Pravděpodobnost rozdílu náhodně vybrané dvojice sekvencí (v populaci)
- Tedy (teoretická) míra diverzity populace

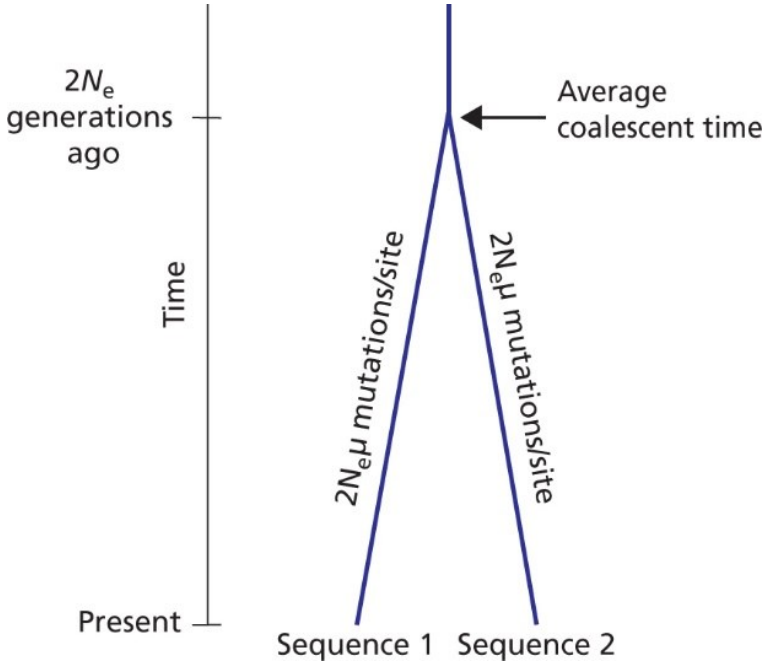
- Od společného předka $2N_e$ generací *(v průměru, odvození jindy)*
- Akumulace v obou 2 liniích
- Rychlost μ

- $\rightarrow 2 \times 2N_e\,\mu$

- $\boldsymbol{\Theta} = 4N_e\,\mu$

- Obvykle neznáme N_e ani μ

- Odhady pomocí sekvencí



Sekvence – ukazatele diversity

- **Haplotype** (Gene) **diversity**

  – pravděpodobnost, že dvě náhodně vybrané sekvence jsou různé

- **Nucleotide diversity** π

  – pravděpodobnost, že náhodně vybrané místo u dvou náhodně vybraných sekvencí je různé

- Haplotypová diverzita > nukleotidová diverzita

```
1 ATTGCCGACCCTAGGCTAGGCTATTTGACCTGAGATTCCATTAATTGGGCGC
2 ---------T------------------------------------------
3 ---------T------C-----------------------------------
```

Odhad nukleotidové diverzity (π)

Sequence 1	A A T G T C A A C G
Sequence 2	A A T G T C A A C G
Sequence 3	A T T G T C A A C G
Sequence 4	A T T G T G A T C G
	* * *
Site number	1 2 3 4 5 6 7 8 9 10

Nucleotide diversity (π):

1 A A T G T C A A C G
2 A A T G T C A A C G $d_{12} = 0$

1 A A T G T C A A C G
3 A T T G T C A A C G $d_{13} = 1$

2 A A T G T C A A C G
3 A T T G T C A A C G $d_{23} = 1$

1 A A T G T C A A C G
4 A T T G T G A T C G $d_{14} = 3$

2 A A T G T C A A C G
4 A T T G T G A T C G $d_{24} = 3$

3 A T T G T C A A C G
4 A T T G T G A T C G $d_{34} = 2$

$$\Sigma d_{ij} = 0 + 1 + 3 + 1 + 3 + 2 = 10$$

Number of pairs of sequences compared = $[n(n-1)]/2 = [4(3)]/2 = 6$
$\hat{\pi}$ = 10 differences/6 pairs = 1.67 average pairwise differences
$\hat{\pi}$ = 1.67 avg. differences/10 sites = 0.167 pairwise differences per site

Počet segregujících míst *s*

Sequence 1	A	A	T	G	T	C	A	A	C	G
Sequence 2	A	A	T	G	T	C	A	A	C	G
Sequence 3	A	T	T	G	T	C	A	A	C	G
Sequence 4	A	T	T	G	T	G	A	T	C	G
		*				*		*		
Site number	1	2	3	4	5	6	7	8	9	10

Segregating sites (S and p_S):

Sites 2, 6, and 8 have variable base pairs among the four sequences (columns marked with *). These are segretating sites. Therefore, for these sequences $S = 3$ segregating sites and $p_S = 3/10 = 0.3$ segregating sites per nucleotide site examined.

- Nukleotidová diverzita nezávisí (moc) na velikosti vzorku
- Počet segregujících míst závisí (hodně) na velikosti vzorku

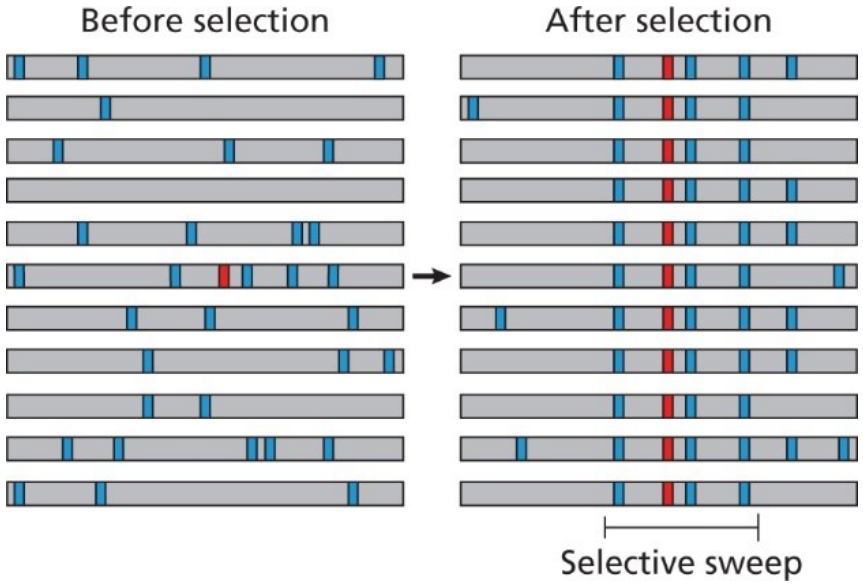
```
1 ATTGCCGACCCTAGGCTAGGCTATTTGACCTGAGATTCCATTAATTGGGCGC
2 ---------T------------------------------------------
3 ---------T------C-----------------------------------
```

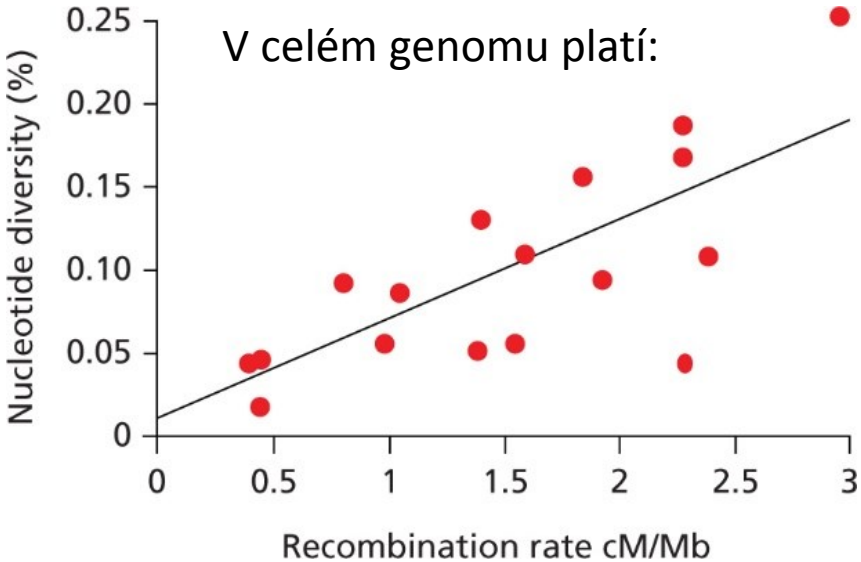
Odhad parametru *Θ*

Celkové diverzity (polymorfismu) populace

- Neutralita → odpovídá si:
  - ***Θ*** population mutation rate parameter („teoretická diverzita“)
  - ***s*** počet segregujících míst (proporce)
  - ***π*** nukleotidová diverzita
- $\boldsymbol{\pi} = \boldsymbol{\Theta} = 4N_e\,\mu$
- $\boldsymbol{s} = a\boldsymbol{\Theta}$ škálování přes velikost vzorku, a = 1 + ½ + … + 1/(n-1)
- ***π*** *a* ***s*** *v*ypočítáme z alignmentu sekvencí
- Někdy se ale odhady pomocí ***π*** *a* ***s*** liší → neplatí neutralita
- Využití pro detekci selekce

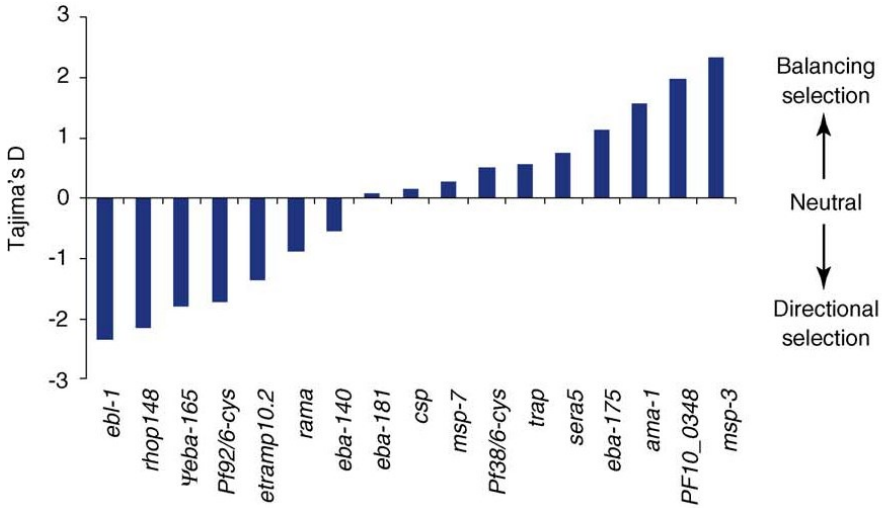
Selective sweep



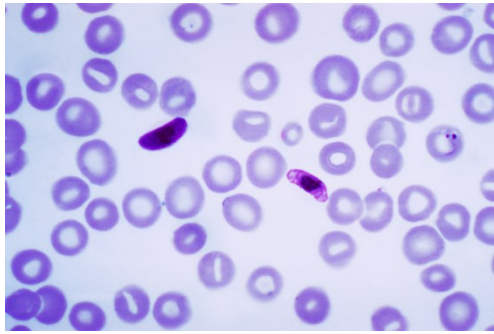


- Při nízké rekombinaci zůstane okolo déle snížená variabilita
- Ve „vymetené“ oblasti se objevují v nízké frekvenci nové mutace
- Přidají segregační místa (zvýší ***s***), málo ovlivní ***π*** nukleotidovou diverzitu

Tajima's D test



- $D = \frac{\pi - \frac{s}{a}}{\sqrt{V(\pi - \frac{s}{a})}}$
  - ***s*** počet segregujících míst (proporce)
  - ***π*** nukleotidová diverzita
  - ***a*** ošetření velikosti vzorku
  - ***V*** variance (rozptyl)



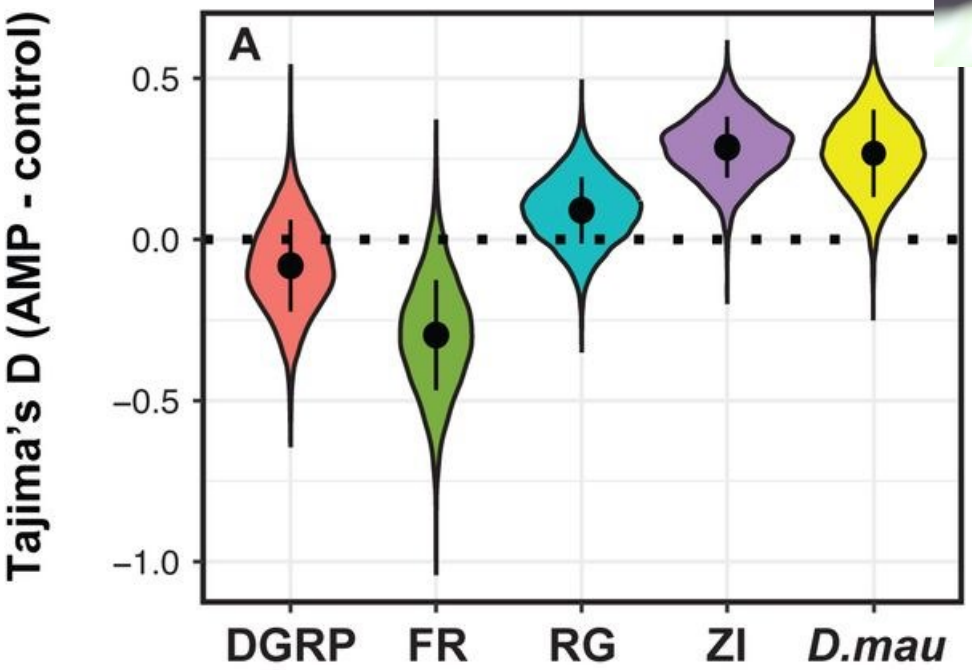
Plasmodium falciparum

- Ukazuje odchylku od nulového modelu
- D < 0 sweep (a obecně pozitivní, negativní i purifikující selekce)
- D > 0 balancing selection

- Pozor! Ovlivní i demografické události (expanze, bottleneck)

Drosophila melanogaster a *mauritania* a geny pro antimikrobiální peptidy

D > 0 balancing selection



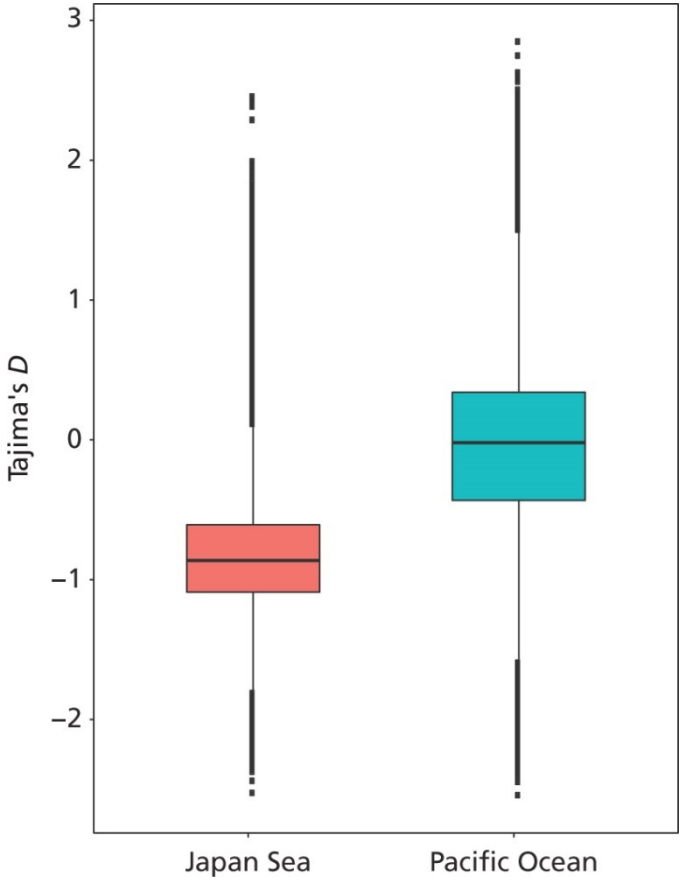
Gaterosteus aculeatus

Gaterosteus nipponicus

Současná sympatrie

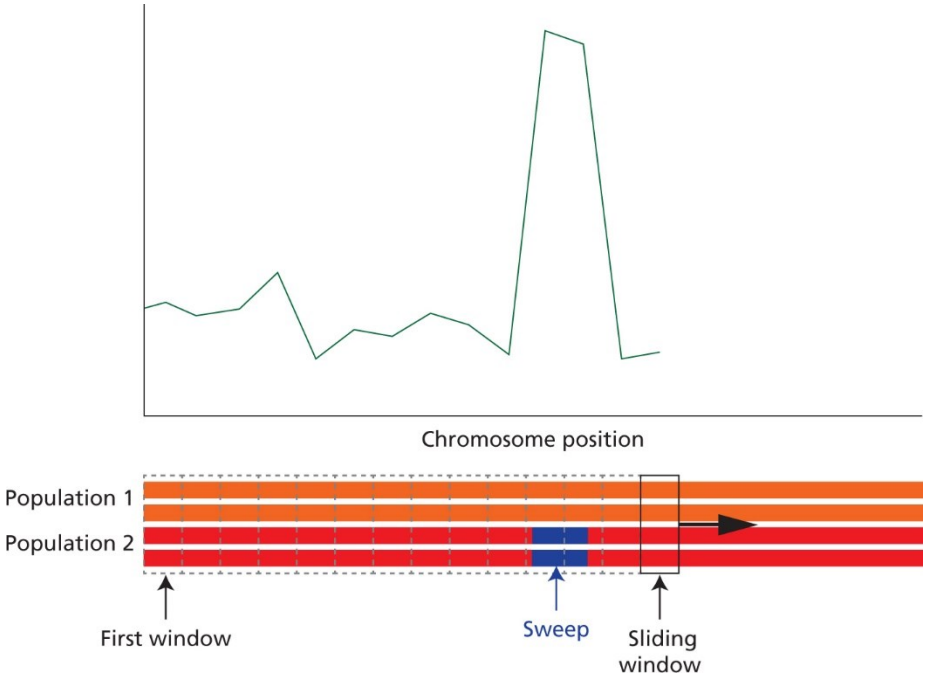
Hybridní sterilita

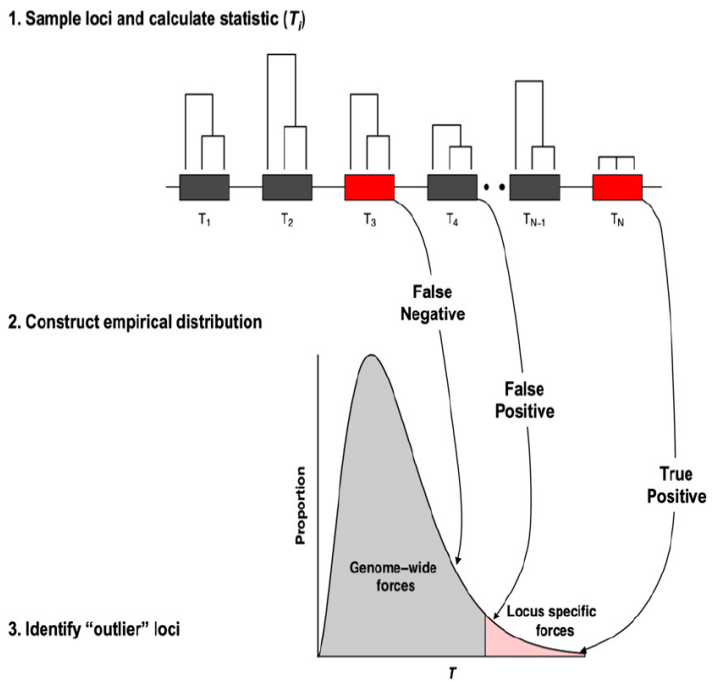
Bottleneck u G. nipponicus

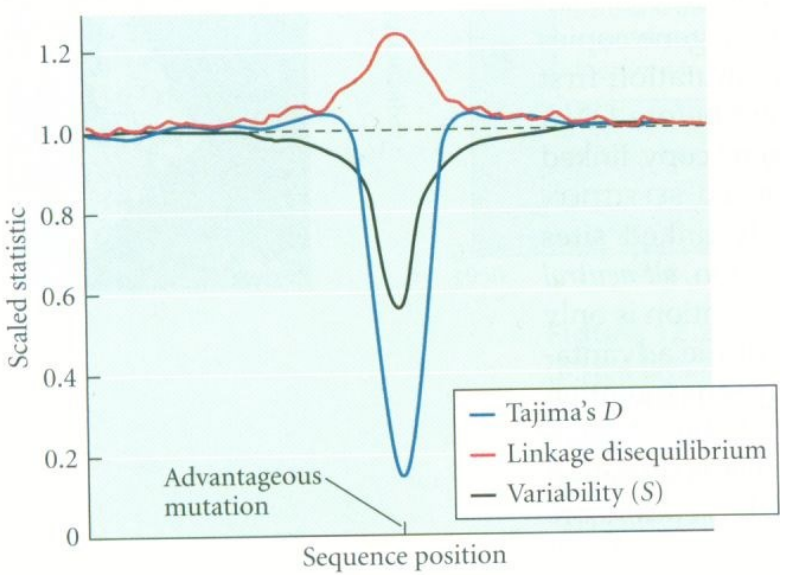


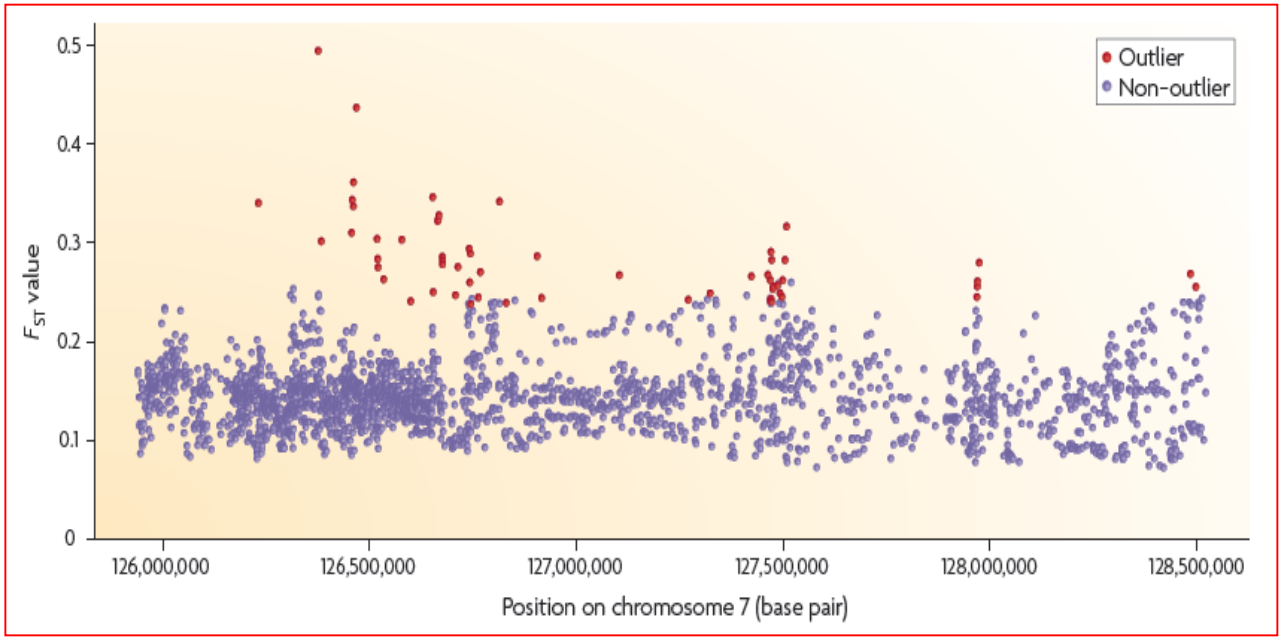
Genome scans

- Sliding window
- Různé statistiky
- Empirická distribuce → outliers



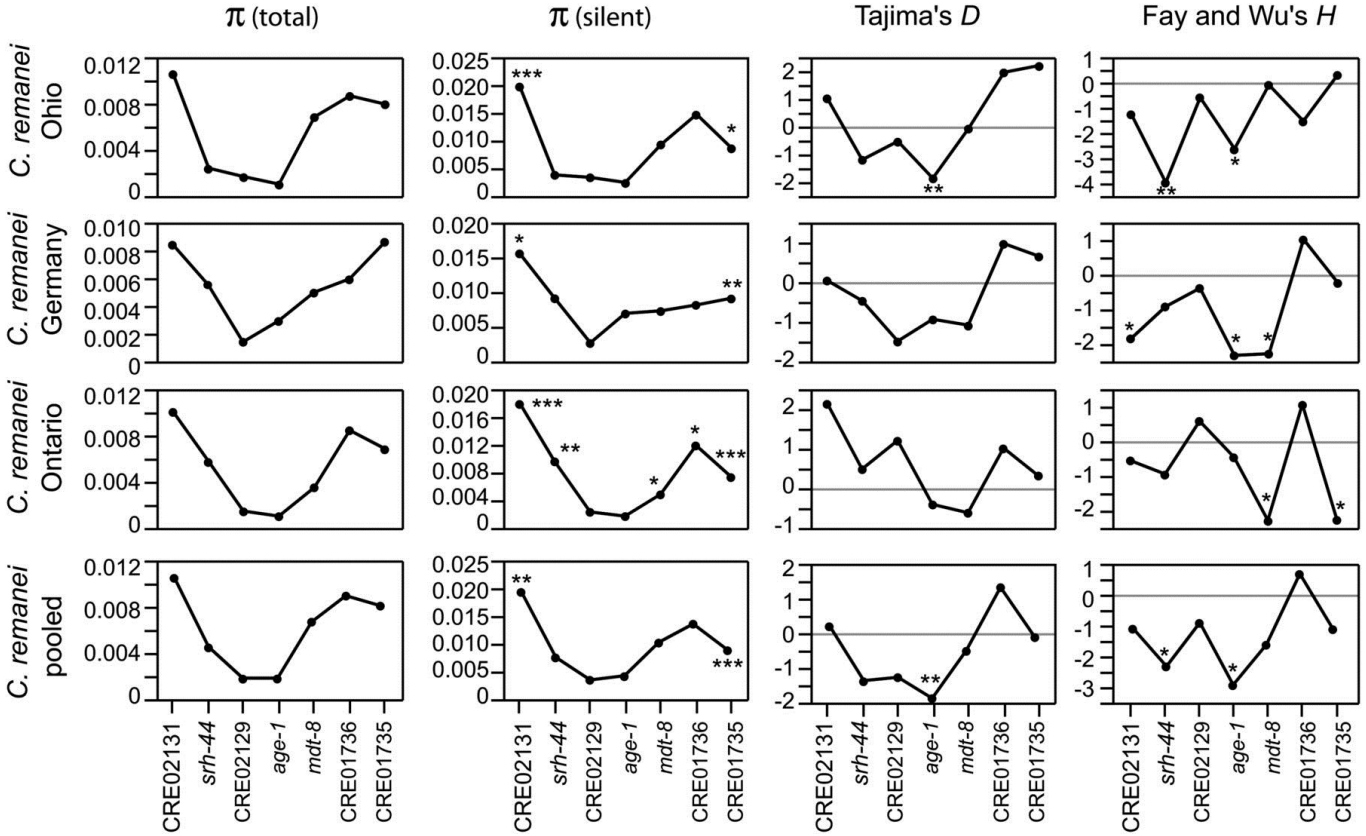




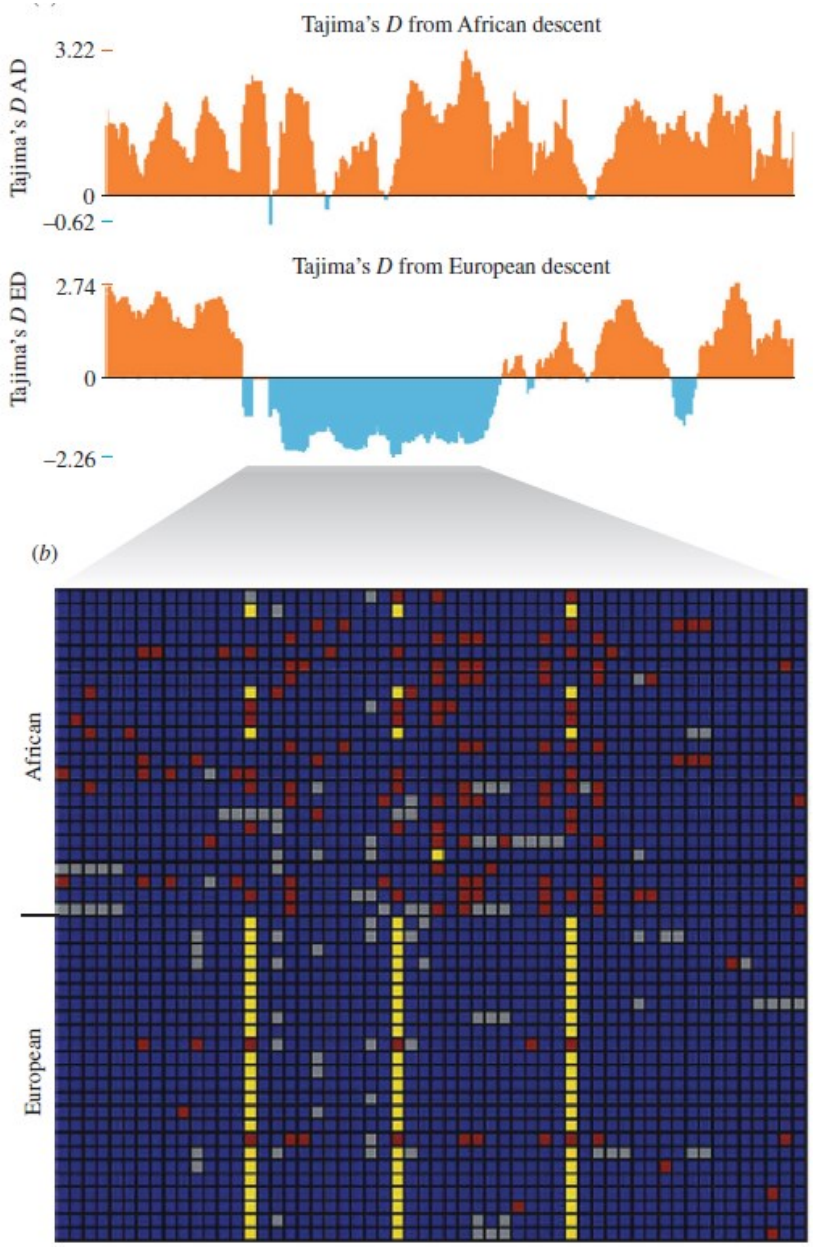

Outlier
Non-outlier
0.5
0.4
0.3
0.2
0.1
0
F_{ST} value
126,000,000
126,500,000
127,000,000
127,500,000
128,000,000
128,500,000
Position on chromosome 7 (base pair)

age-1 u *C. remanei*

lokus signální dráhy spojené s dlouhověkostí



Richard Jovelin et al. G3 2014;4:1123-1133



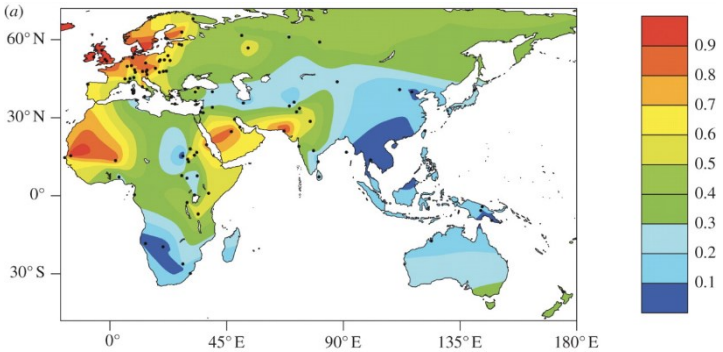
CLSPN

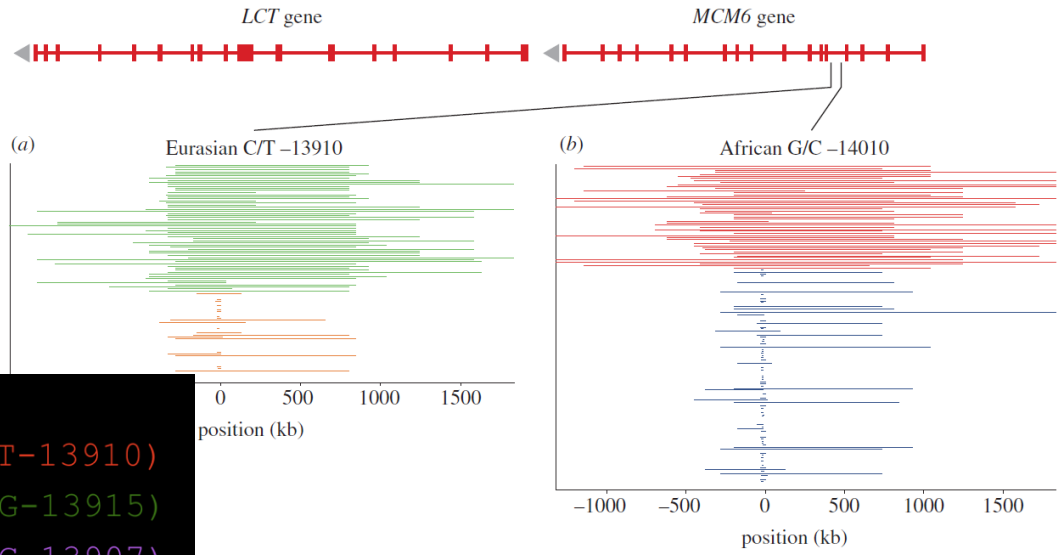
Claspin

Required for checkpoint mediated cell cycle arrest in response to inhibition of DNA replication or to DNA damage induced by both ionizing and UV irradiation.

Haplotypové bloky

- Například Lactase Persistence
- Testy na LD
- Nedávné události (desítky tisíc let)

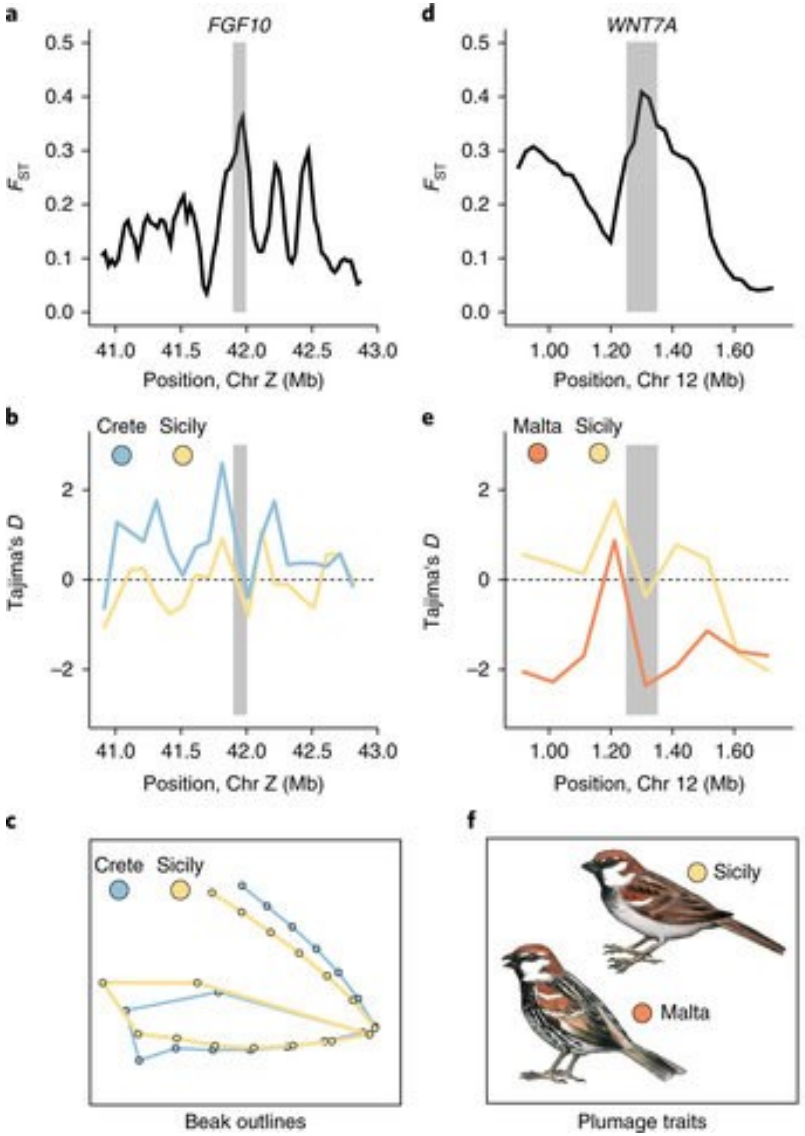




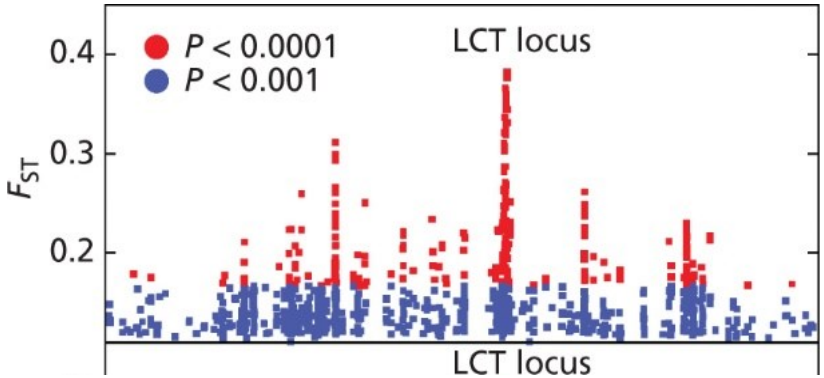

```
Europe:                T        (T-13910)
Kenya:           G              (G-13915)
Sudan:                     G    (G-13907)
```

Lokální adaptace a F_{ST}

- Jiné frekvence alel na genech pro lokální adaptace
- → populace zde více „odlišeny“
- → vyšší $\boldsymbol{F_{ST}}$



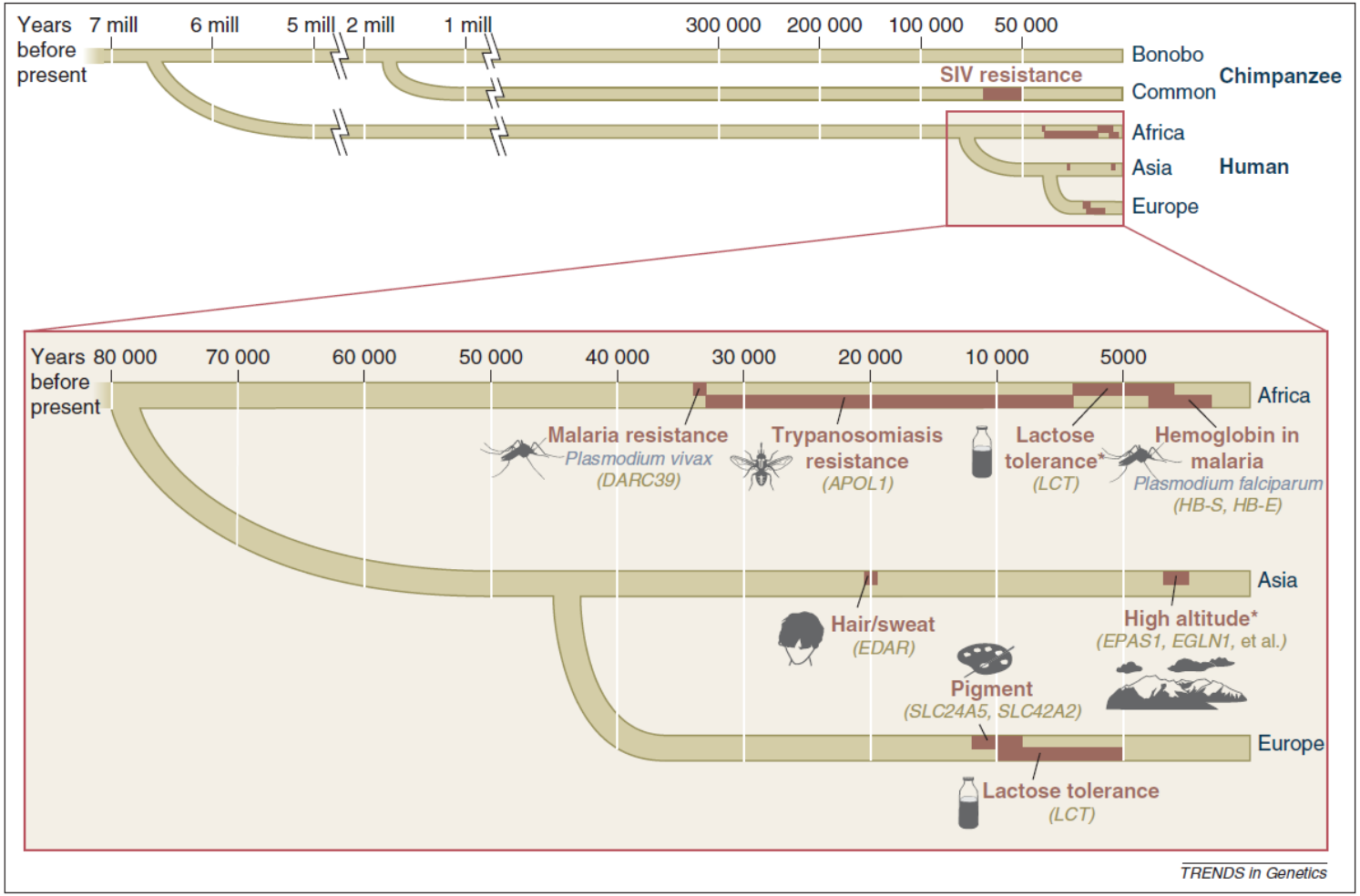
Runemark et al. 2018



Maasai

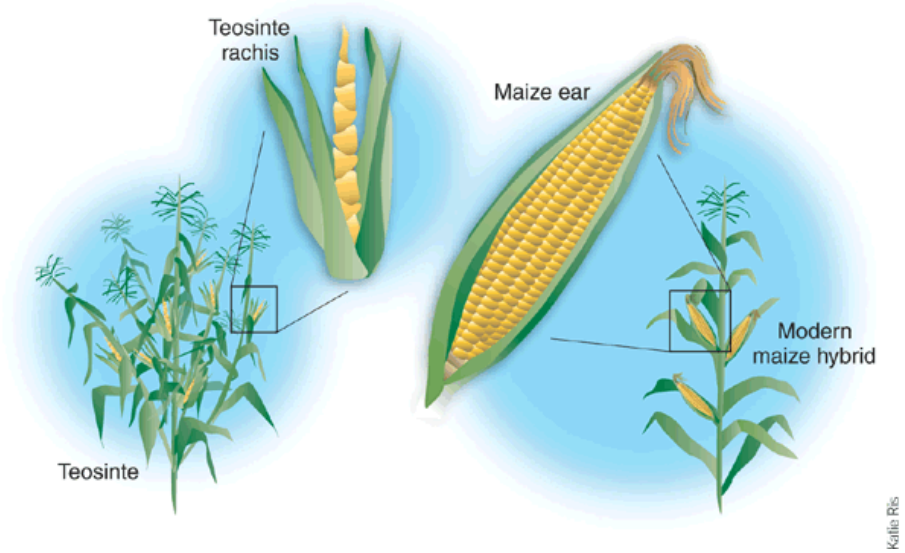
Luhya

Příklady (nedávné) pozitivní selekce *(Vitti et al. 2012)*



Další testy neutrality založené na frekvencích

- Je jich více
- **HKA test** (Hudson-Kreitman- Aguadé)
- Pozitivní korelace variability uvnitř populace (druhu) s variabilitou mezi populacemi (druhy)
- Tedy ***Dxy*** a ***s*** jako odhad ***Θ***

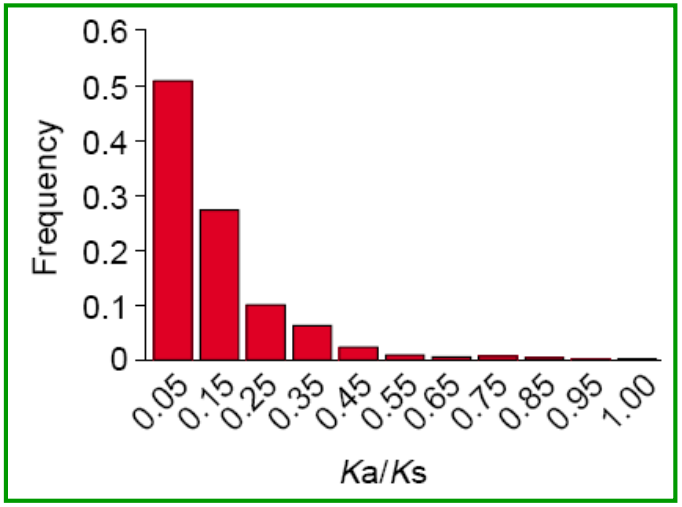


***I**nbred, **L**andrace, **T**eosinte*

Gene	P Values in HKA			Candidate Status by HKA
	I	L	T	
AY108876	0.010*	0.014*	0.225	Domestication
AY106123	0.155	0.022*	0.014*	–
AY106190	0.047*	0.047*	0.046*	–
AY108504	0.030*	0.013*	0.036*	–
AY108957	0.057	0.023*	0.028*	–
AY108255	0.323	0.158	0.848	–

Yamasaki et al. 2005

Selekce před delší dobou Ka/Ks (d*N*/d*S*)



Srovnání ortologních genů myši a potkana

- Poměr **nesynonymních** (*Ka*) a **synonymních** (*Ks*) mutací

- **Alignment** sekvencí stejného genu u více druhů
  → **rozdíly** (mutace)
  → jen některé mutace změní aminokyselinu

- **Korekce**
  jen 25 % možných záměn nezmění aminokyselinu
  některé typy mutací častější
  zpětné a konvergentní mutace (hlavně u více odlišných sekvencí)

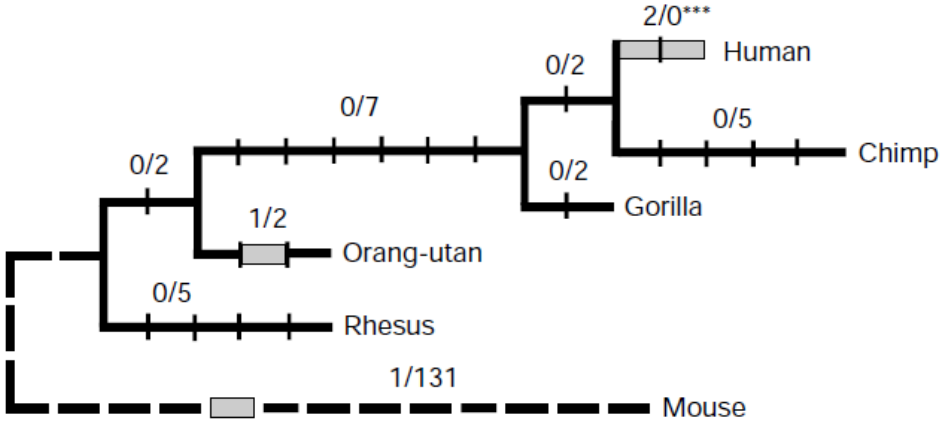
- Většinou *Ka/Ks* << 1, selekce eliminuje změny, „purifying selection“
  protein se nemění

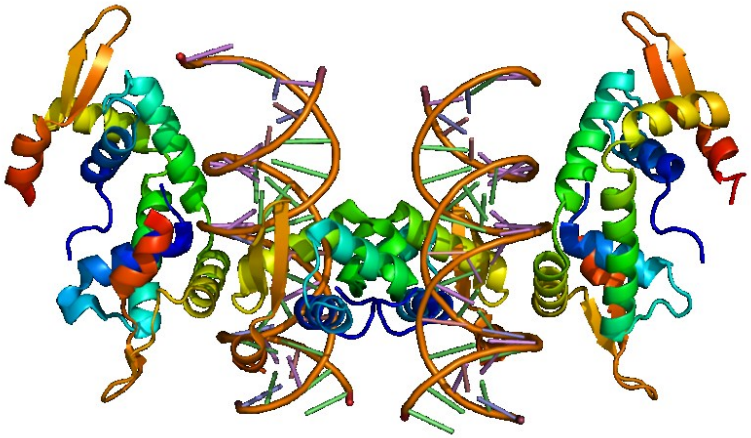
- *Ka/Ks* >> 1 „positive selection“, MHC, reprodukční proteiny

- Různé části proteinu mohou být pod různou selekcí (ABS u MHC)

FOXP2 Enard et al. 2002

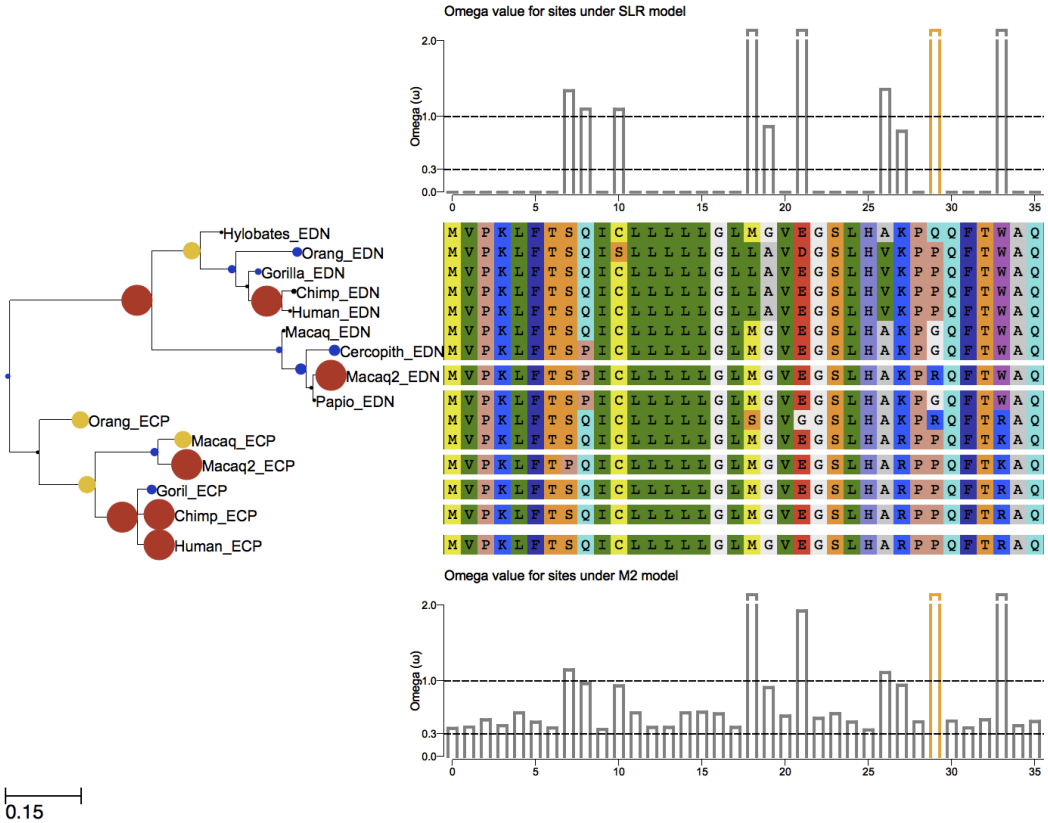
- Schopnost artikulovat
- Nesynonymní versus synonymní mutace





Bayesianské modely

- $\omega = d_N/d_S$
- pro jednotlivé kodony
- CodeML
- False positives



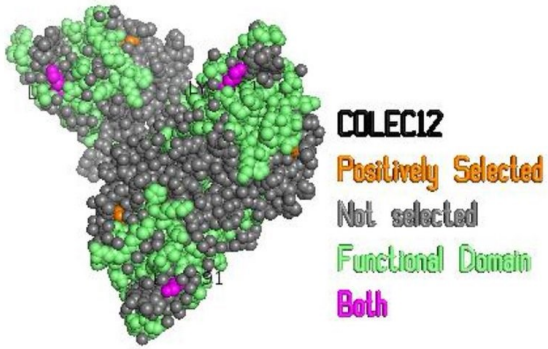
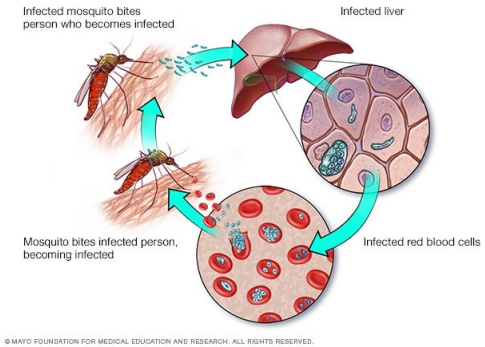


Figure 3. LYS.py visual output example of protein coloured according to its evolutionary positively selected amino acid residues and domain positions.

Testy založené na Ka/Ks (d_N/d_S)

- MK test (McDonald-Kreitman)

- Opět srovnání uvnitř a mezi (populacemi, druhy…)



Plasmodium falciparum

Kalciová pumpa end. Retikula, co lze vyřadit Artemisininem

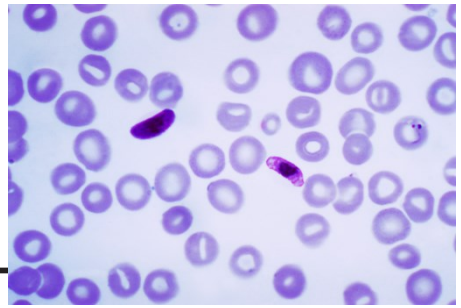


Table 4. d_N/d_S ratio and McDonald-Kreitman test of neutrality for the *pfserca* gene.

	d_N/d_S ratio		McDonald-Kreitman test					
			Interspecific fixed nucleotide differences		Intraspecific nucleotide polymorphisms			
	d_N/d_S	*P*-value[1]	*Syn*	*Nsyn*	*Syn*	*Nsyn*	*NI*	*P*-value[2]
French Guiana	0.53	0.511	43	15	3	6	5.7	0.022
Pará	0.59	0.707	43	15	1	3	8.6	0.070
Amazonas	∞	0.169	43	15	0	2	n.a.	0.077
Senegal	0.75	0.824	43	15	1	5	14.3	0.010
Eq. Guinea	0.22	0.318	43	15	2	3	4.3	0.136
Thailand	0.15	0.415	43	15	1	1	2.9	0.475
Cambodia	n.a.	n.a.	43	15	0	0	n.c.	n.c.

Shrnutí

- Od alel k sekvencím
- Neutrální teorie molekulární evoluce
- Molekulární hodiny
- Odhady celkové diverzity populací
- Detekce selekce na molekulární úrovni