

**Eduard Kejnovský (Zdeněk Kubát) +
Roman Hobza**

EVOLUČNÍ GENOMIKA

IV. EVOLUCE GENŮ



OSNOVA

1. Definice genu, historie
2. Struktura genu
3. Vznik nových genů
4. Velikosti genů
5. Introny – staré nebo mladé

Definice genu

= základní jednotka genetické informace zapsaná v NK

Podle širší definice:

1. všechny sekvence DNA potřebné k syntéze proteinu nebo RNA, tedy i regulační a signální sekvence (nejširší)
2. transkribované sekvence (nezahrnuje regulační oblasti)
3. úseky přímo kódující peptid (nejužší) nebo pořadí bází ve funkčních molekulách RNA

Typy genů (širší definice):

Geny strukturní, geny pro RNA a geny-regulační sekvence

Historie konceptu genu

Klasické období genetiky

- **Mendel, Bateson**: buněčné elementy, faktory určující vlastnosti
- **Boveri, Sutton** (1902-3): chromozomová teorie dědičnosti (chromozomy se přenášejí při mitóze a meióze)
- **Bateson** (1905): hrachor, vazba genů (odporuje Mendlovým zákonům)
- **Johansen** (1909): zavedl pojem „gen“ (Hugo de Vries - pangen)
- **Morgan** (1910) (**Lock, 1906**): geny jsou uspořádány lineárně na chromozomech, vazbové skupiny
- **Sturtevant** (1913): První genetická mapa *D. melanogaster*
- **Dobzhansky** a další (20. léta): cytologické pokusy s X-rays, indukce zlomů, přestaveb, důkaz genů na chromozomech
- **Griffith** (1928): transformační experimenty u bakterií – přenos genů
- **Muller** (20.- 30. léta): geny jsou neviditelné body na chromozomech (dědičnost, rekombinace, mutace, funkce)

Neoklasické období genetiky

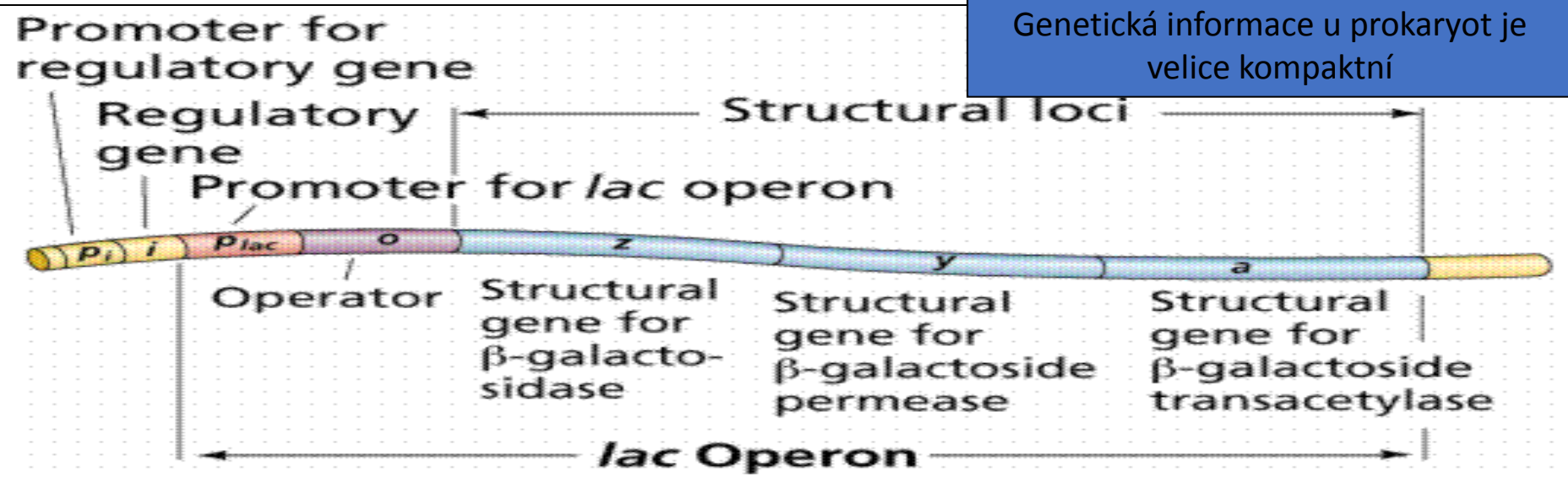
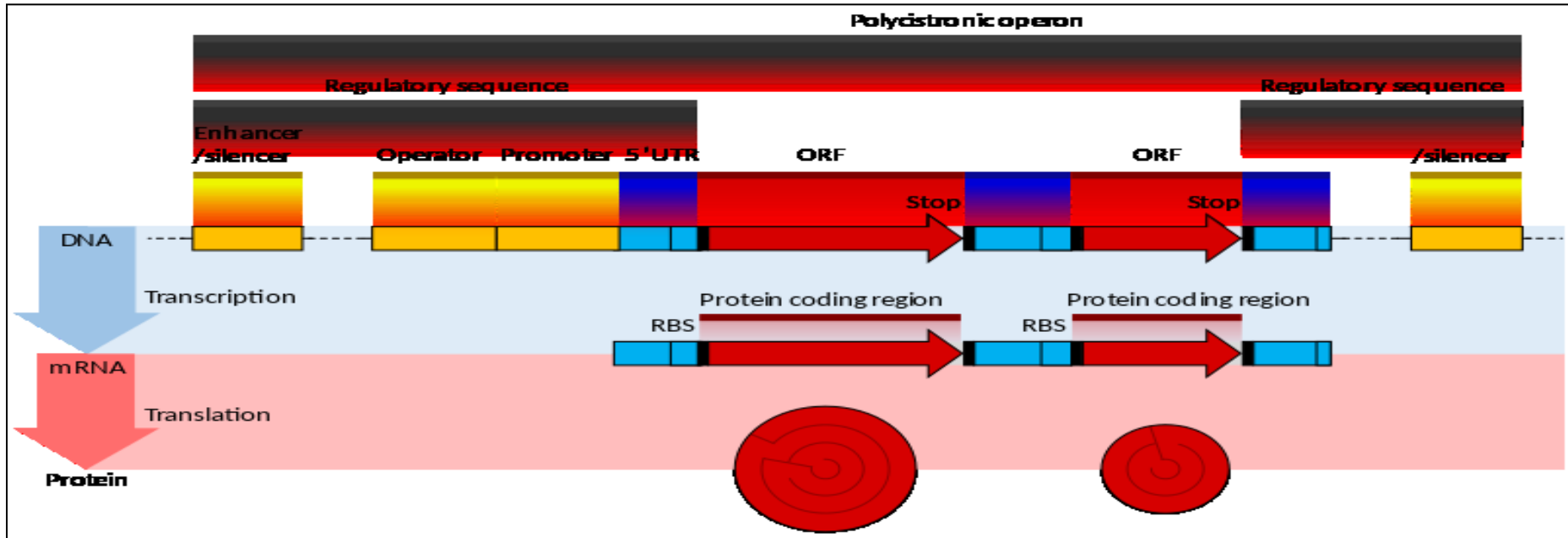
- **40. léta**: geny mohou být rozděleny rekombinací na segmenty – geny mají délku (DNA či proteiny?)
- **Avery, MacLeod a McCarthy** (1944): substancí zodpovědnou za transformaci je DNA
- **Hershey a Chaseová** (1952): genetickou informaci nese DNA (multiplikace bakteriofága zajištěna DNA)
- **Beadle a Tatum** (1941): „one gene – one enzyme“ (souvislost mezi geny a proteiny)
- **Watson a Crick** (1953): struktura DNA
- **Crick** (1958): centrální dogma MB a teorie proteosyntézy
- **Meselson a Stahl** (1958) semikonzervativní replikace
- **Jacob a Monod** (1961): mRNA, operonová teorie
- **Nirenberg, Khorana, Ochoa** (1966) genetický kód
- **60. léta**: gen – RNA – polypeptid

Moderní genetika

- **Cohen** (1973): rekombinantní molekuly DNA (genové inženýrství)
- **složené geny** (1977, Philip Sharp + Richard Roberts, NC 1993)
- **smRNA** (1998, Andrew Fire + Craig Mello, NC 2006)
- **genomově-centrický pohled** (např. Heng 2009, BioEssays)

STRUKTURA GENU

Geny prokaryot jsou uspořádány do operonů



Hledání genů u prokaryot

- **ORF (otevřené čtecí rámce)**

- start kodon je následován nejméně 60 AK, poté stop kodon
- homologie se známými ORF

- **Signální sekvence**

- Transkripce - konsensus promotorové a terminační sekvence
- Translace - vazebné místo na ribozóm: Shine-Dalgarnova sekvence

- **Rozdíly v obsahu bází** mezi kódující a nekódující sekvencemi DNA - obsah GC, tzv. codon bias

Komplikace u eukaryot

- **Složené geny (split genes)**

- introny a exony,
- obratlovci - délka genu 30kb/1-2kb je kódující
- např. gen pro dystrophin 2.4 Mb, desítky exonů, introny až 32kb

- **Velké genomy** - u rostlin až 110 000 Mb (*Fritillaria assyriaca*)

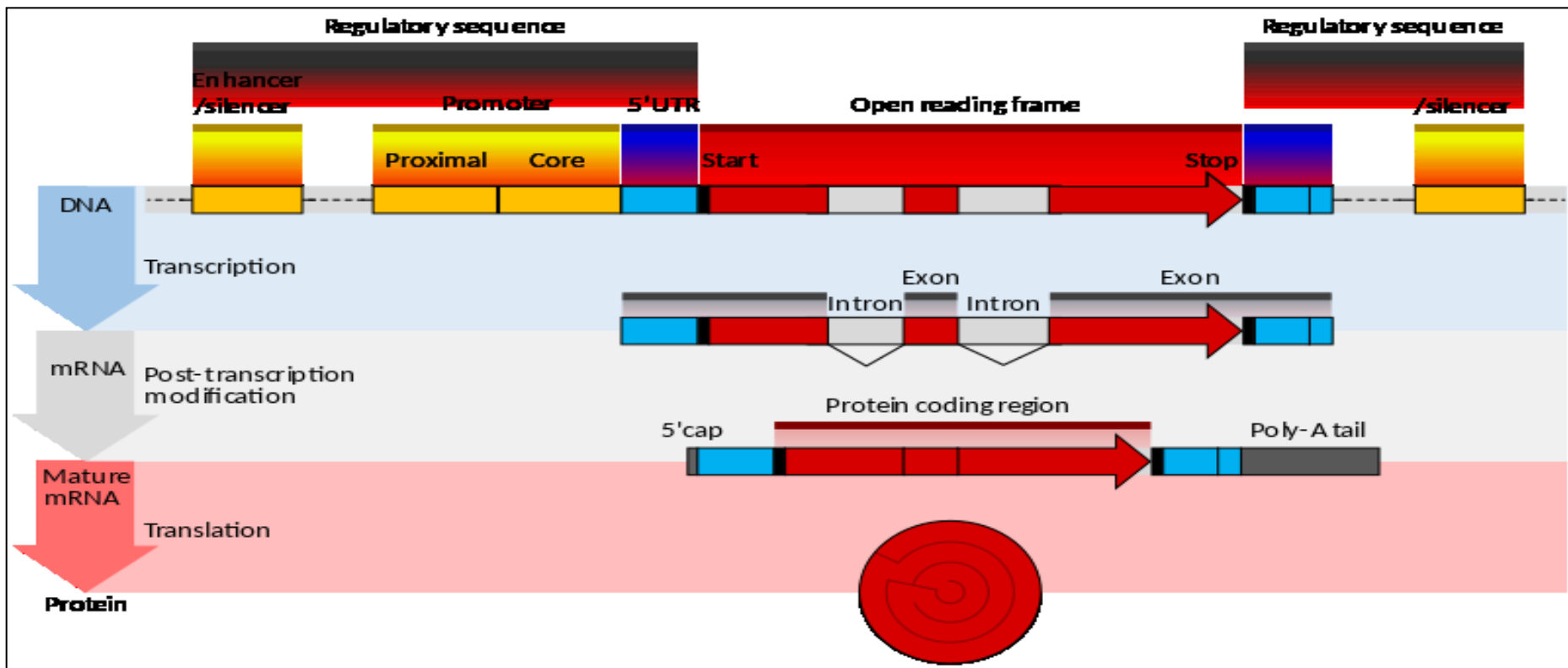
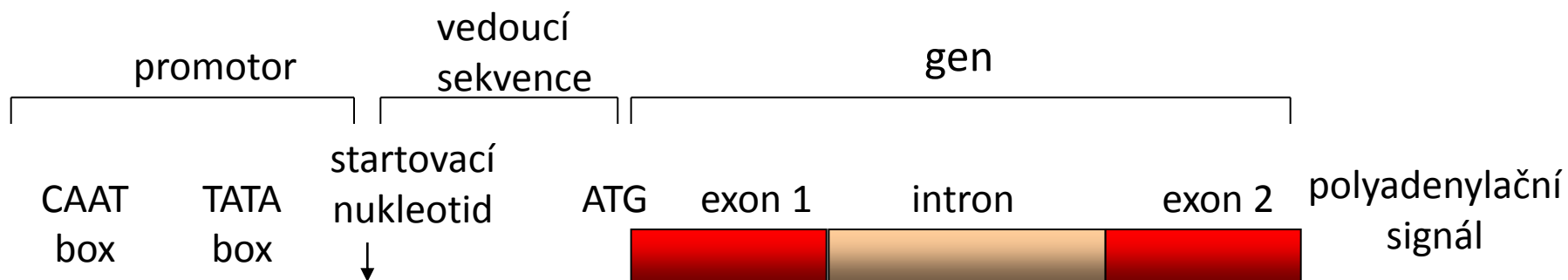
- **Většina DNA je nekódující**

- introny, regulační oblasti, “junk” DNA
- asi 1.5% kódující (člověk)

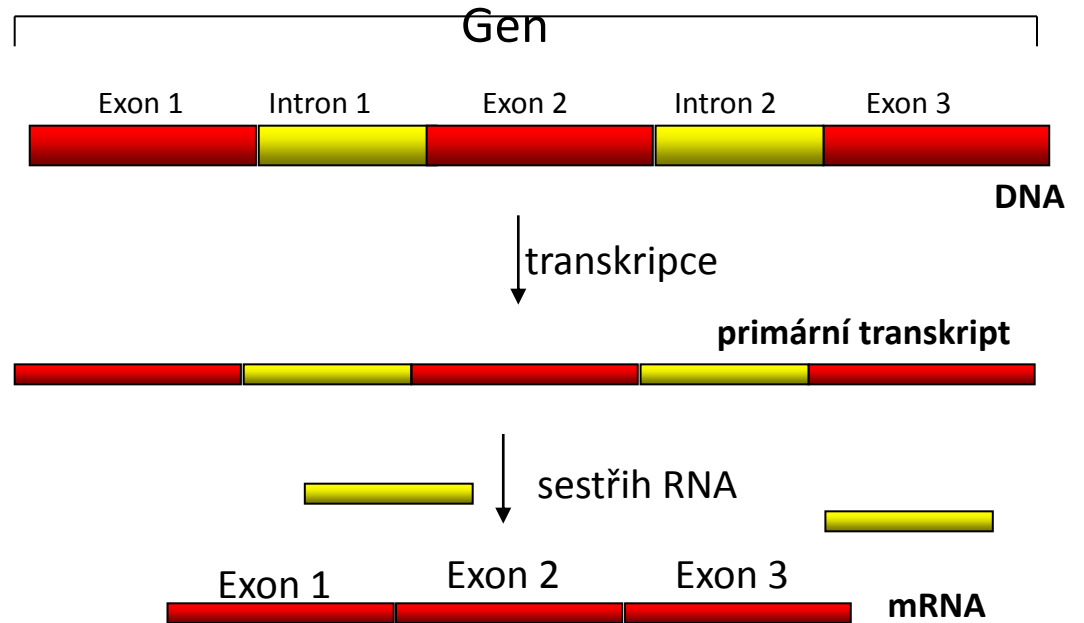
- **Složitá regulace** genové exprese (modifikace chromatinu, metylace DNA, RNAi, alternativní sestřih)

- **Regulační sekvence mohou být daleko** od start kodonu

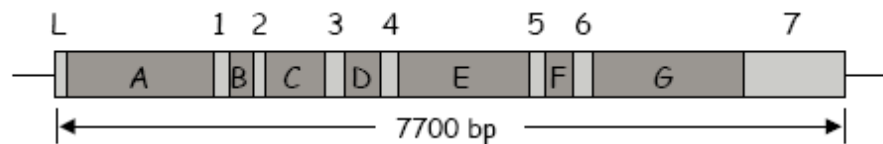
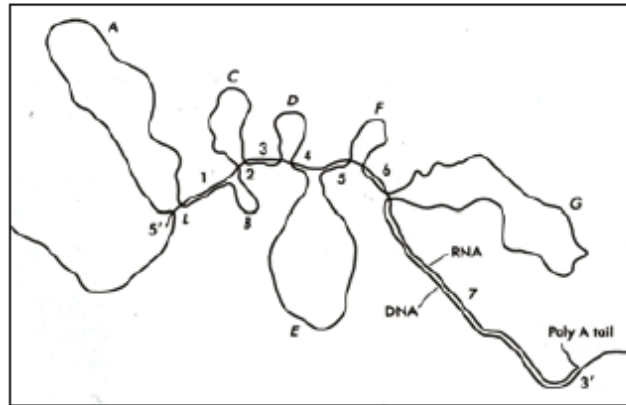
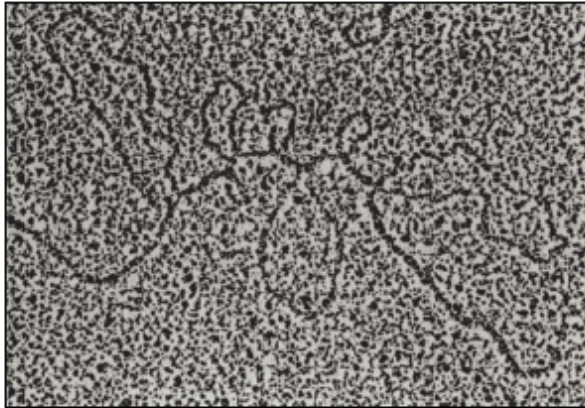
Obecné schéma eukaryotického genu



Složený gen



Genetická informace u eukaryot je fragmentována



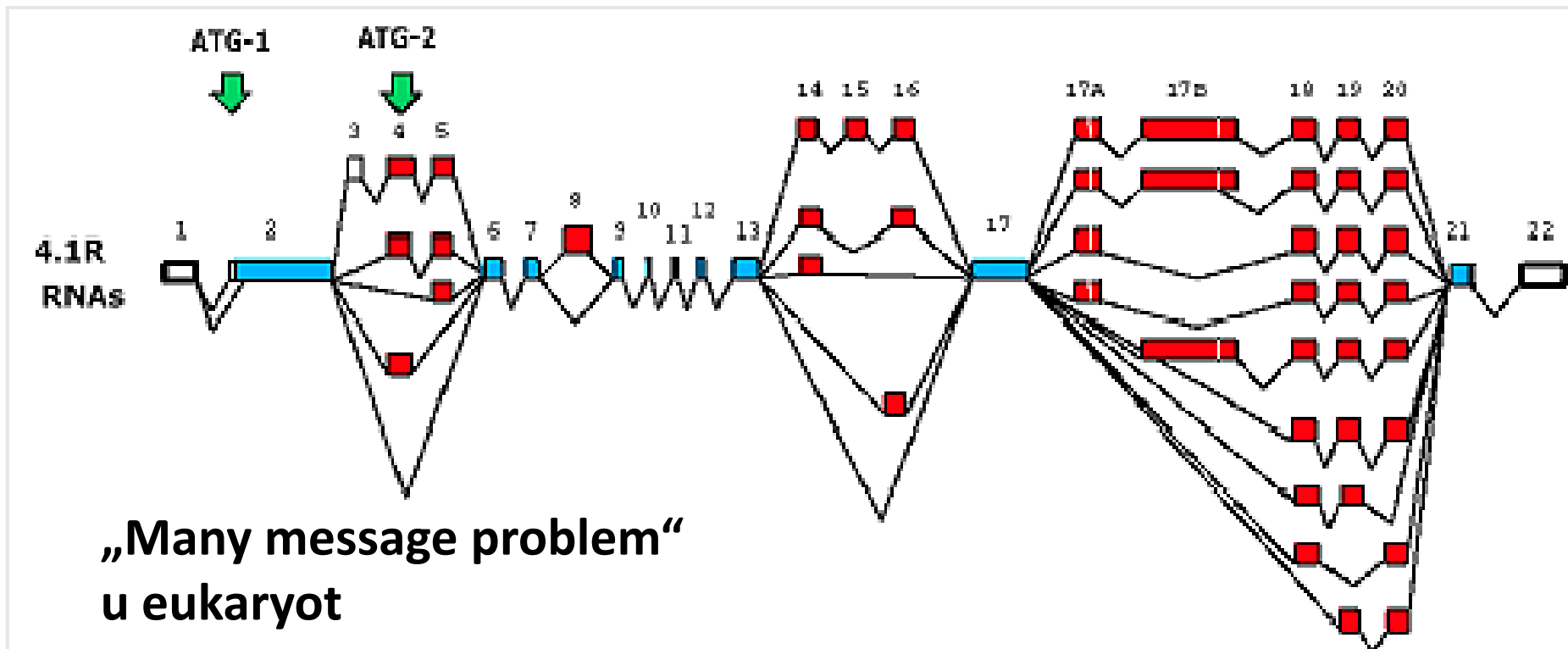
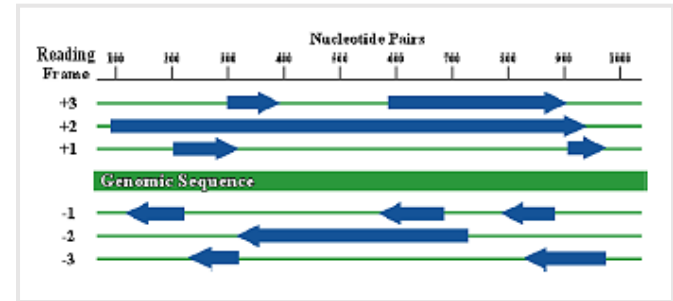
Ovalbumin gene from chicken

exon – expressed sequence
intron – intervening sequence

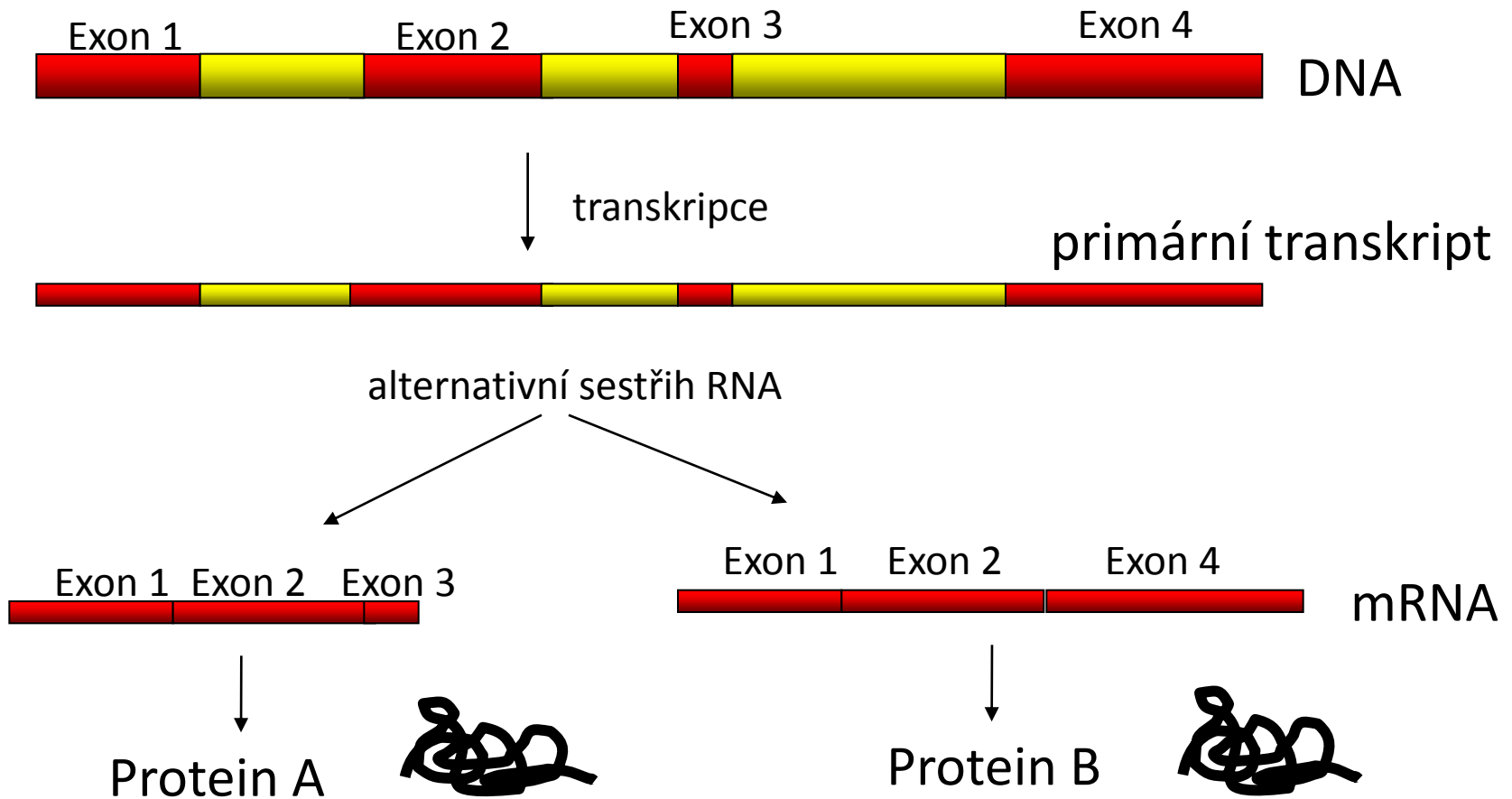
Jak se hledají geny?

- otevřené čtecí rámce (ORF)
- obsah a distribuce nukleotidů, „genové rysy“
- používání kodonů
- hranice exon-intron
- promotory, regulační sekvence
- homologie v databázích, EST

ORF:



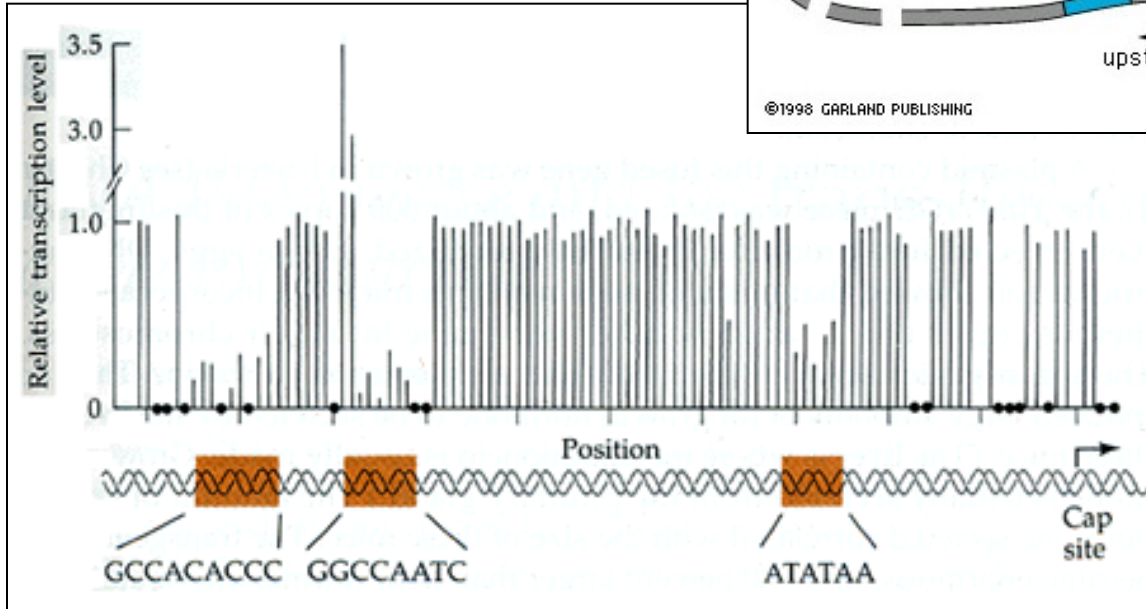
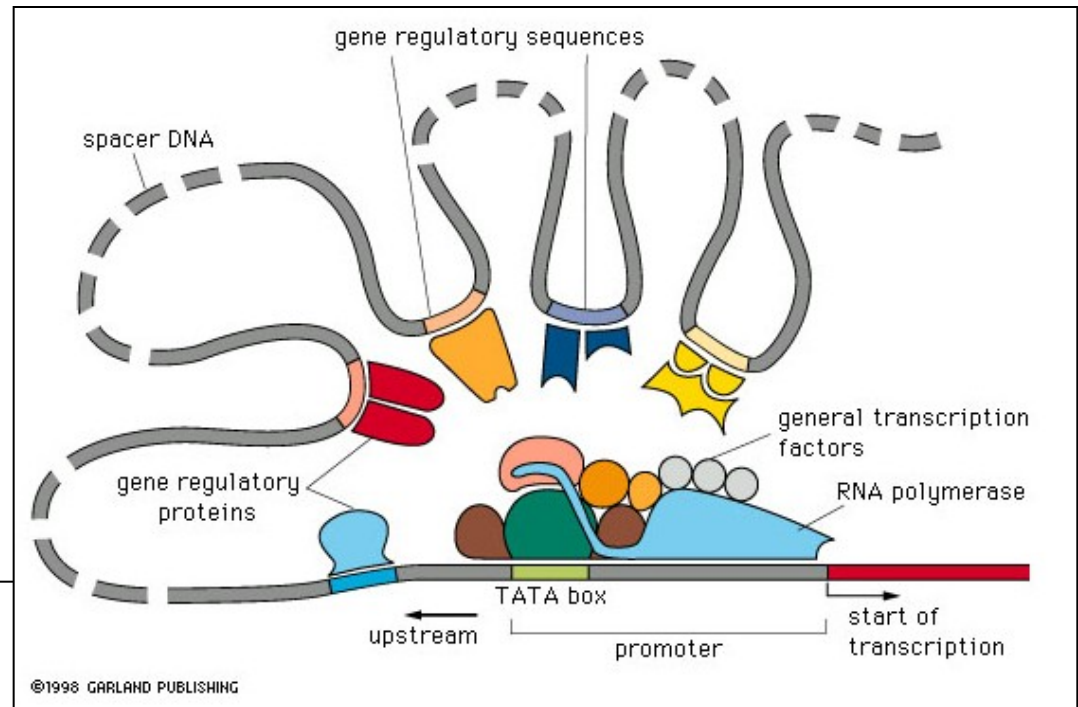
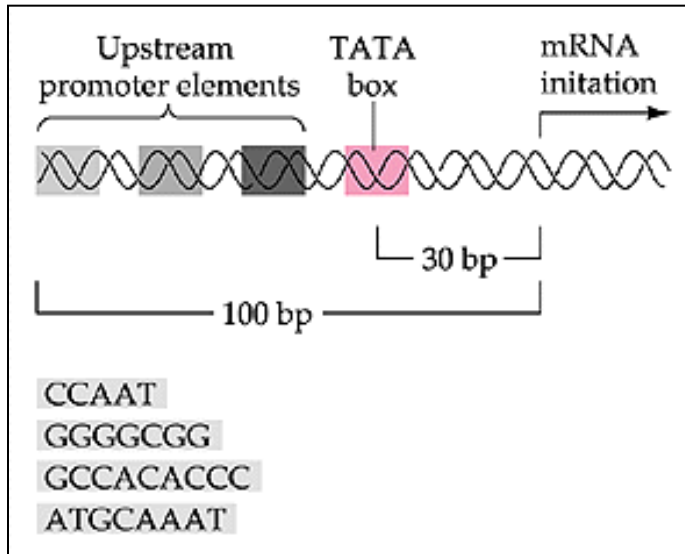
Alternativní sestřih



exitrony

= introny bez stop kodonů

Struktura promotoru



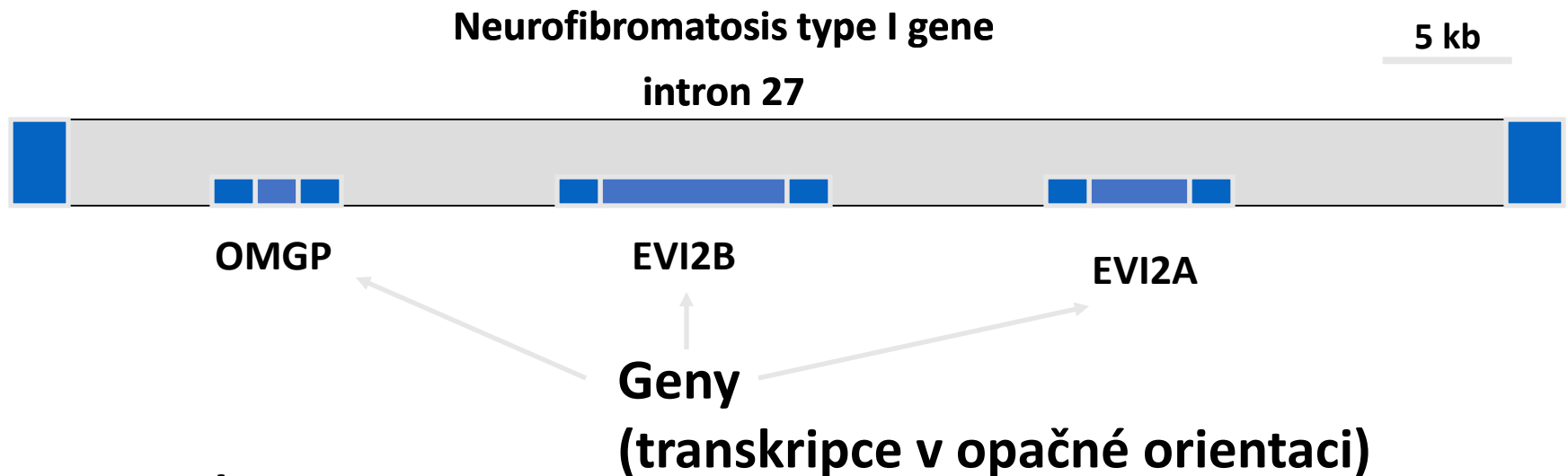
mutace v kritických
místech blokují
transkripci

Geny v genech a jiné podivnosti

- překrývající se geny:

met val Gen A
GTTTATGGTA
val tyr gly Gen B

- geny uvnitř jiných genů:



- pseudogeny:

Pseudogeny

Definice:

- sekvence podobná genu, nekóduje funkční produkt
- nefunkční relikv původně funkčního genu

Problém definice:

- pseudogen může plnit důležitou funkci (nekódující RNA, regulační sekvence, stabilita RNA svého homologa),
- komplikují mol-biol. studie

An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene

Vznik:

- (a) duplikace a degenerace jedné kopie – „non-processed“
- (b) retrotransposice – „processed“ pseudogen
- člověk má 19 000 pseudogenů, pravidlo 50:50

Nekódující RNA

microRNAs – 22bází

siRNAs – small-interfering, 20-25 bází

piRNAs – PIWI-interacting RNA

snoRNAs – small nucleolar RNA

snRNAs – small nuclear RNA

exRNAs – extracellular RNA

scaRNAs – small Cajal body-specific RNA

long ncRNAs – delší než 200 bp, role v regulaci transkripce, některé translatovanaé, Xist and HOTAIR.

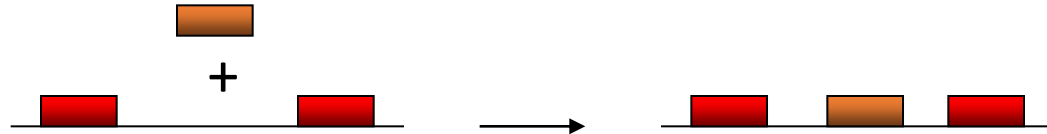
VZNIK NOVÝCH GENŮ

Každý gen vzniká z genu (nebo *de-novo*)

- geny jsou si podobné, duplikace a postupná **divergence** genů, genealogické stromy
- genové **rodiny** a nadrodiny
- počet genů u eukaryot: 10 000 – 40 000
- počet základních **modulů** malý: stovky-max tisíce vzájemně nepříbuzných exonů, nejmenší jsou genové moduly
- Ale některé geny vznikají *de-novo* z „junk DNA“ !!!

Vznik nových genů

(a) Přeskupování
exonů:



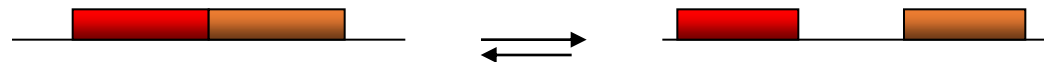
(b) Duplikace
genů:



(c) Retrotranspozice:

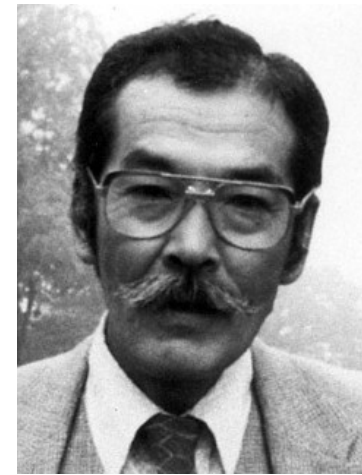


(d) Fúze a štěpení
genů:



Evoluce genovou duplikací

- duplikace je základem diverzifikace
- zrod nových genů u rostlin, kvasinky a drosophily je 10x pomalejší než u *C. elegans*
- poločas rozpadu genů delší u rostlin, duplikáty přetrvávají, mechanismy retence duplikátů?



(Ohno, 1970)

- disperzní x tandemové kopie – rychlost asymetrické evoluce, často u rostlin zůstávají v tandemu
- v nerekombinujících oblastech – rychlejší evoluce

- **Duplikace části genu:**

duplikace domén/vnitřní části genu → zvýšení funkce nebo nová funkce prostřednictvím nových kombinací

- **Duplikace celého genu (genová rodina)**

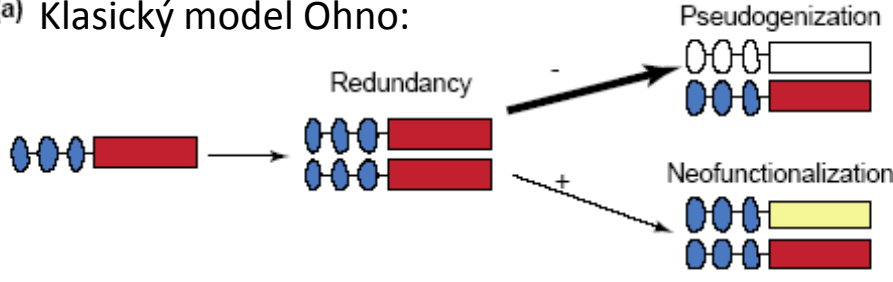
stejná kopie: zvýšení dávky genu,
rozdílnění kopií: nové funkce

- **Duplikace klastru genů**



Genová duplikace: pseudogenizace, neofunkcionalizace, subfunkcionalizace

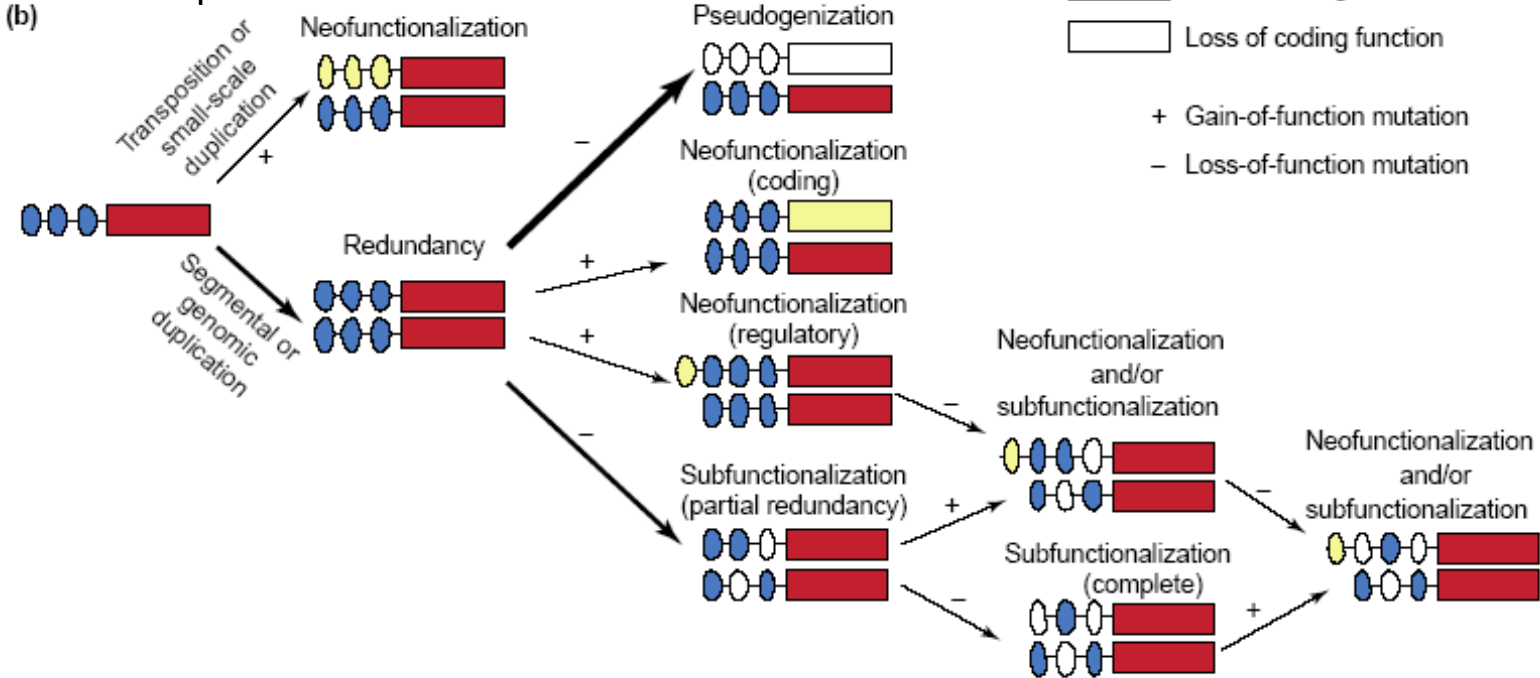
(a) Klasický model Ohno:



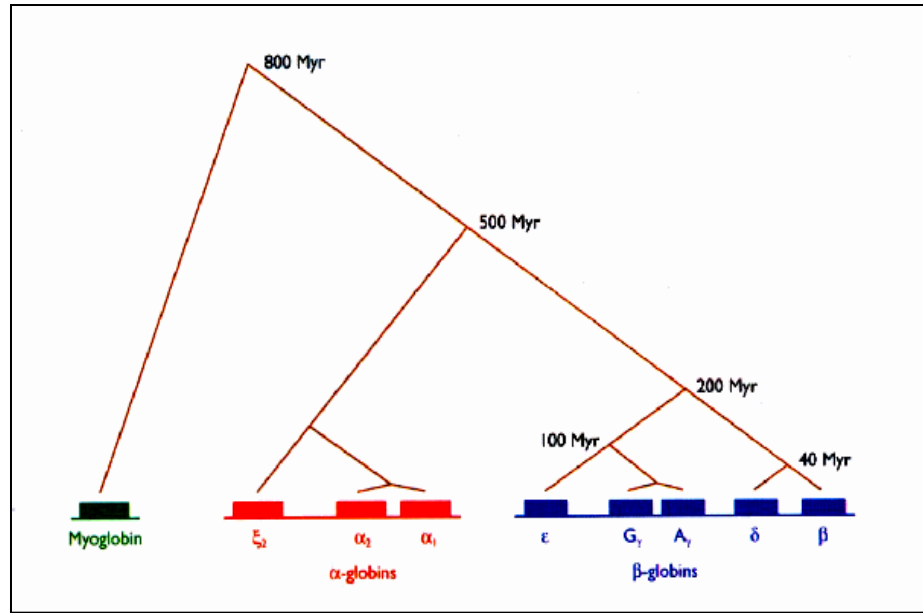
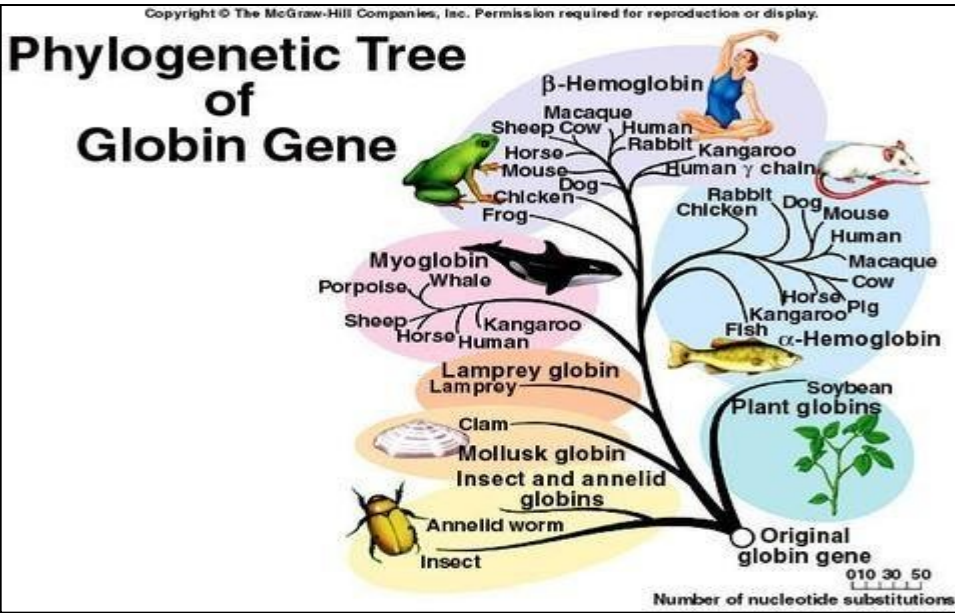
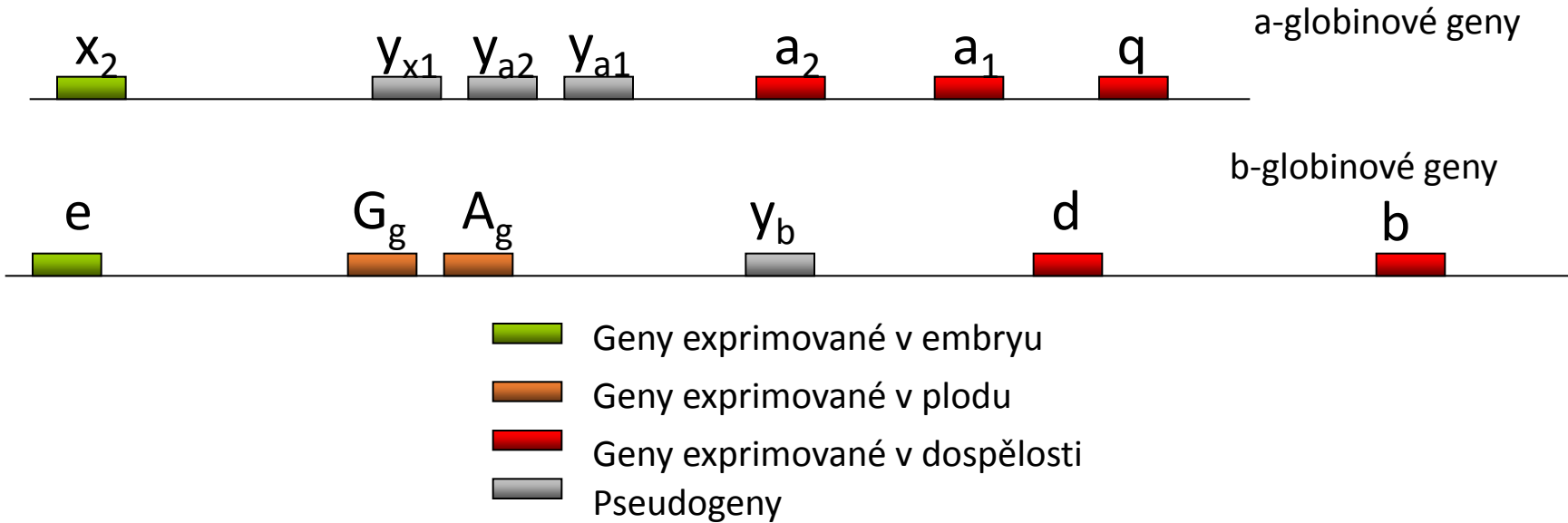
- Regulatory subfunction
- Gain of regulatory subfunction
- Loss of regulatory subfunction
- Coding function
- Gain of coding function
- Loss of coding function

- + Gain-of-function mutation
- Loss-of-function mutation

Moderní pohled:

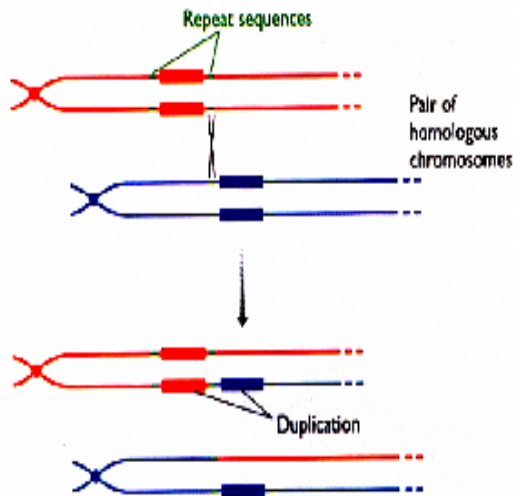


Globinová genová rodina – vznik duplikací

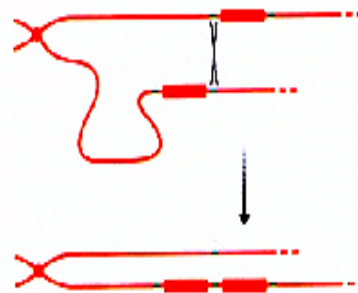


Mechanizmy duplikace genů

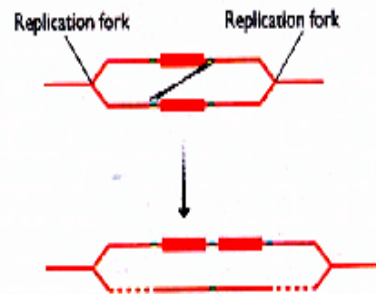
(A) Unequal crossing over



(B) Unequal sister chromatid exchange



(C) During DNA replication



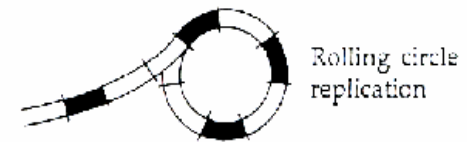
Chromosomal rDNA repeats



Replication and circulatization



Extrachromosomal circular DNA



Rolling circle replication

Amplification



Linearization and integration into genome



1. nerovnoměrný crossing-over (různé chromosomy)
2. nerovnoměrná výměna mezi sesterskými chromatidami
3. duplikace při replikaci
4. mechanismus otáčivé kružnice

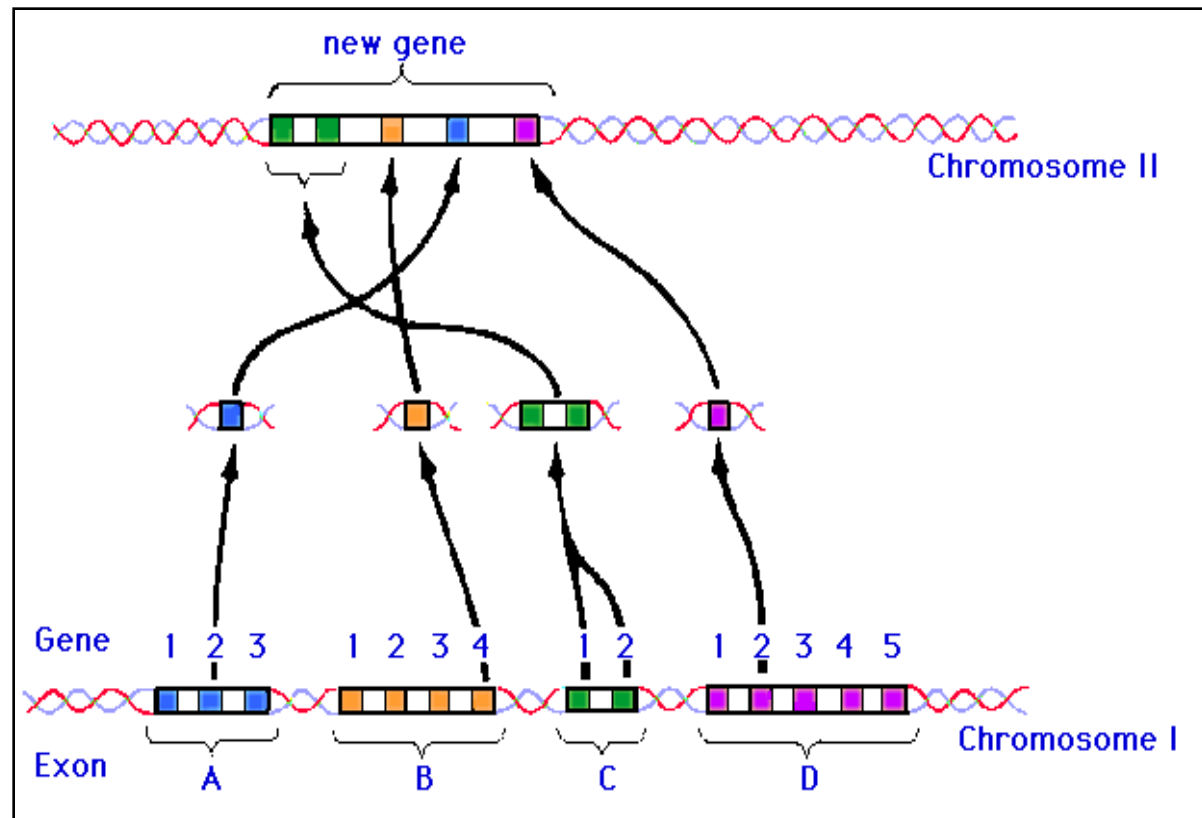
Původ nových genů: Přeskupování exonů (exon shuffling)

- exony různých genů jsou spojeny dohromady za vzniku nového genu
- exon může být duplikován za vzniku nové exon-intronové struktury
- kombinace domén různých proteinů – mozaikový protein

Mechanismy:

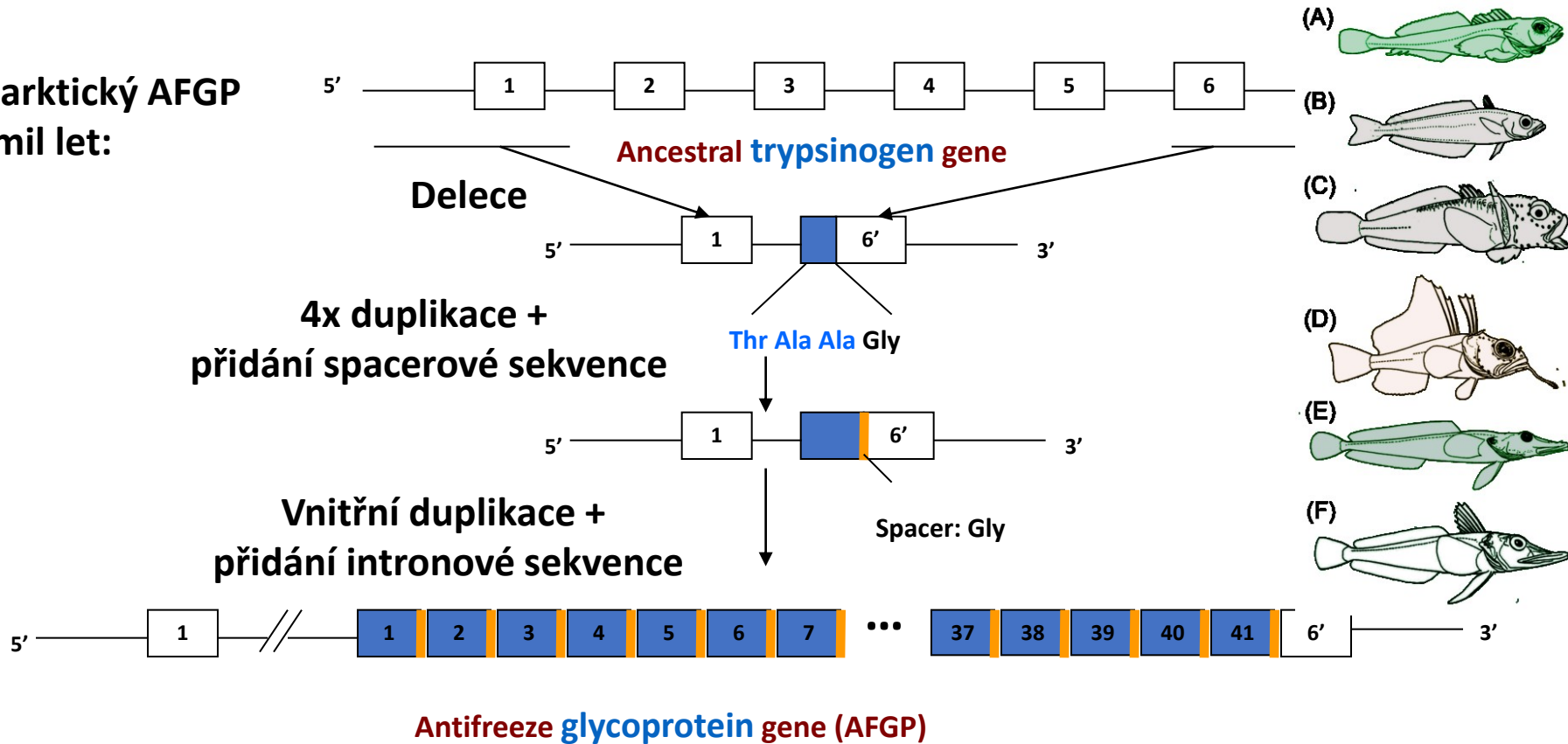
Ektopická rekombinace

Nelegitimní rekombinace



Vznik nového genu na příkladu AFGP

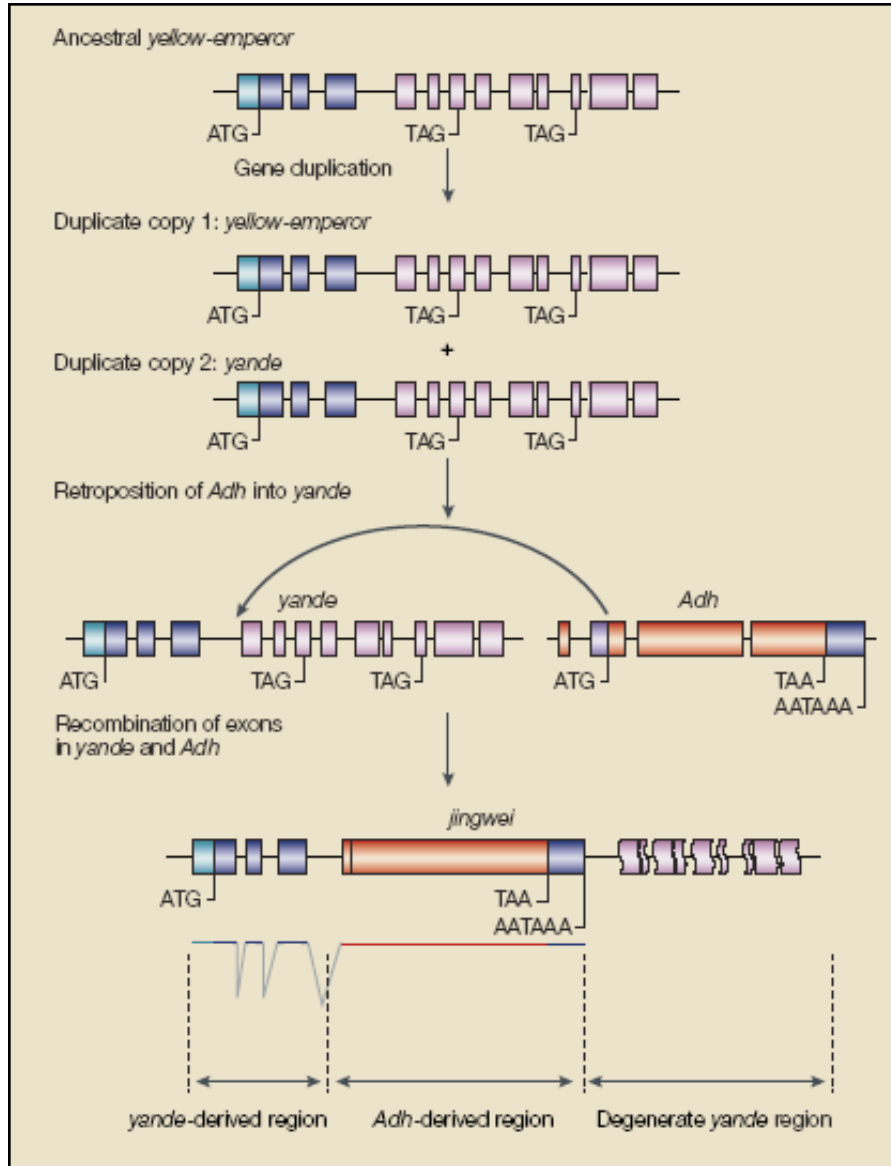
Antarktický AFGP
10 mil let:



- brání zmrznutí tělních tekutin, růstu krystalků ledu
- vznikl před 10 mil let, první zamrznutí polárních oblastí
- vznik z trypsinogenu, zachován 5' a 3' konce (sekrece)
- amplifikace (Thr-Ala-Ala)_n, kde n=4-55
- konvergentní evoluce – antarktický a arktický

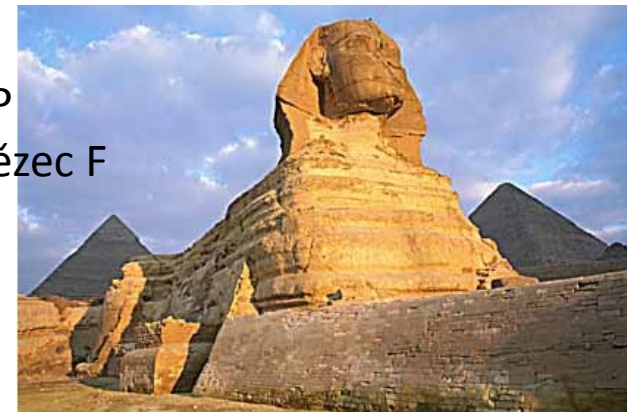
Antarktické ryby
Notothenioidei
Řád Ostnoploutví

Původ genu *Jingwei* + *Sfinx* retrotranspozicí do intronu



- vznik před 2 mil let, drosophila
 - základem yellow emperor
 - duplikace a retro-včlenění *Adh*
 - *Adh* terminační signál
 - degenerace exonů na 3'konci
 - nová kombinace exonů
- pohádka o princezně Jingwei: reinkarnace utonulé princezny v krásného ptáka podobně jako odhalení fungujícího genu v původně objeveném pseudogenu

Gen *Sfinx*:
rRNA gen+ATP
syntázový řetězec F



Původ genu SETMAR – „recyklace“ transposonu

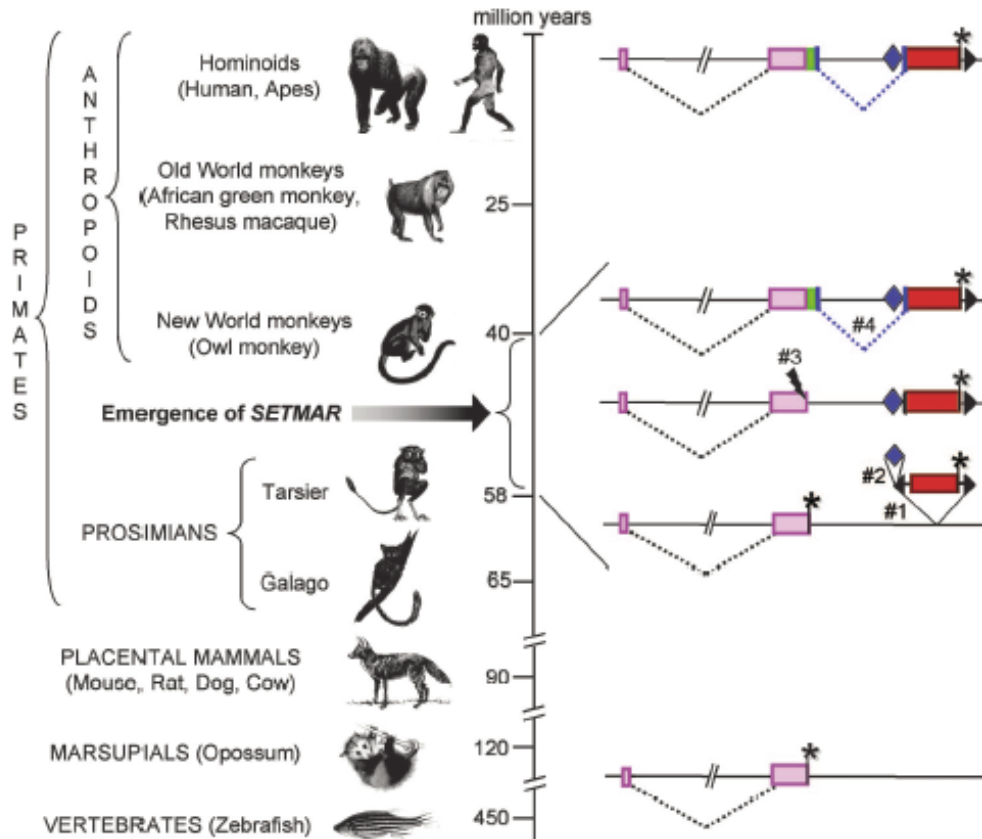
Birth of a chimeric primate gene by capture of the transposase gene from a mobile element

Richard Cordaux*, Swalpa Udit†, Mark A. Batzer*, and Cédric Feschotte†*

Histon metyltransferáza

+

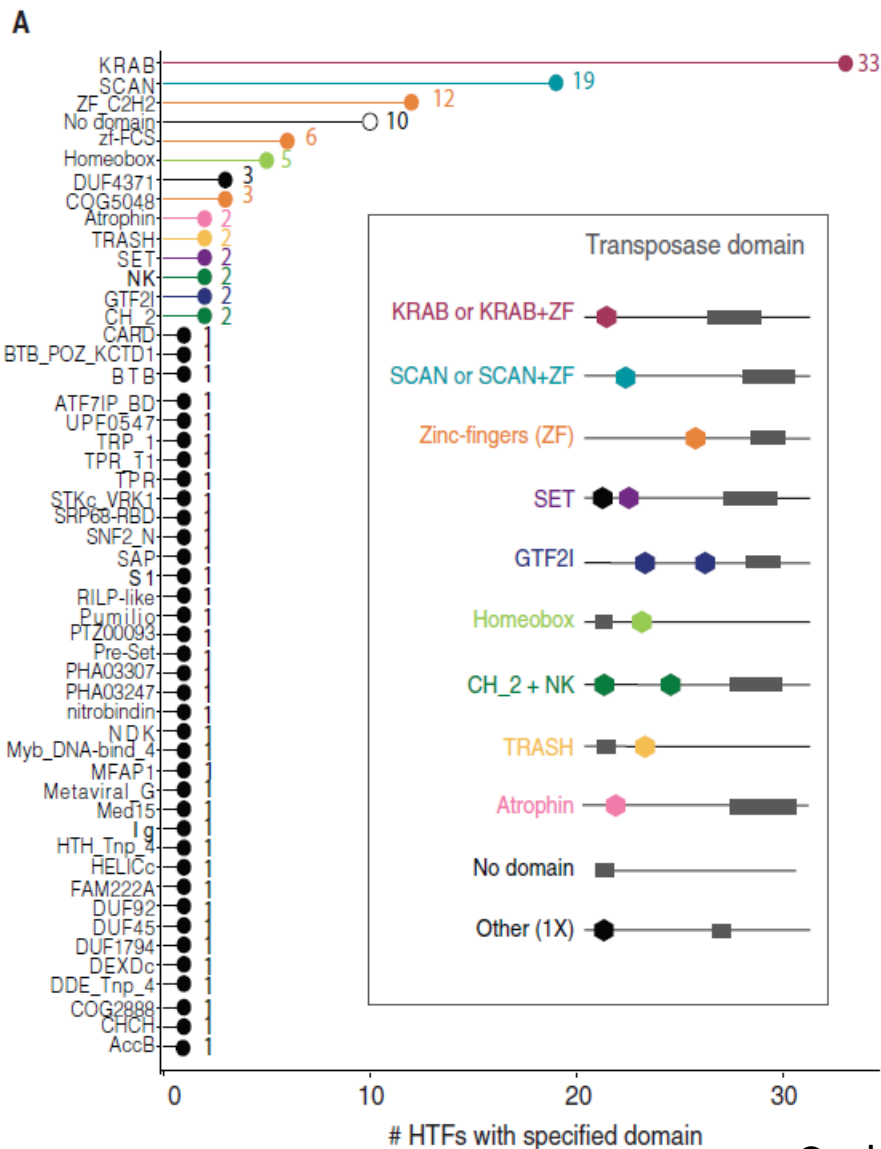
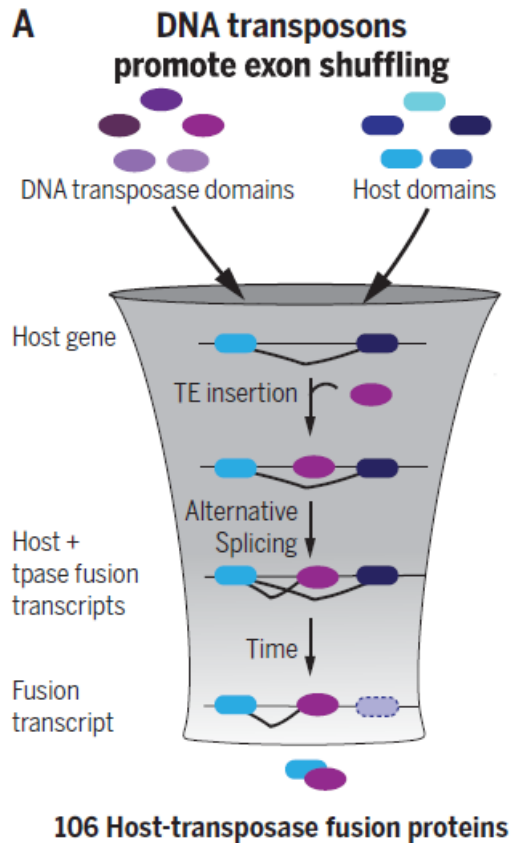
transpozáza



- zrušení stop
- vznik nového stop
- exonizace
- degenerace TIRu
- vznik intronu
- DNA vazebná doména Tn zachována
- TIR místa v genomu
- 50 mil let

Fig. 1. Milestones leading to the birth of *SETMAR*. The structure of the *SETMAR* locus (Right) and a simplified chronology of the divergence time of the species examined relative to hominoid primates (Left) are shown. Pink boxes represent the two *SET* exons, which are separated by a single intron (interrupted black line) and form a “*SET*-only” gene whose structure is conserved in all nonanthropoid species examined and terminated with a stop codon (*) located at a homologous position (except in cow; see Fig. 2a). The *Hsmar1* transposon (event 1) was inserted in the primate lineage, after the split between tarsier and anthropoids, but before the divergence of extant anthropoid lineages. The transposon is shown here with its TIRs (black triangles) and transposase coding sequence (red box). The secondary *Alu5x* insertion within the TIR of *Hsmar1* (event 2) is represented as a blue diamond. The position of the deletion removing the stop codon of the “*SET*-only” gene (event 3) is indicated as a lightning bolt. The *de novo* conversion from noncoding to exonic sequence is shown in green, the creation of the second intron is represented as a dashed blue line (event 4), and the splice sites are shown as thick blue lines.

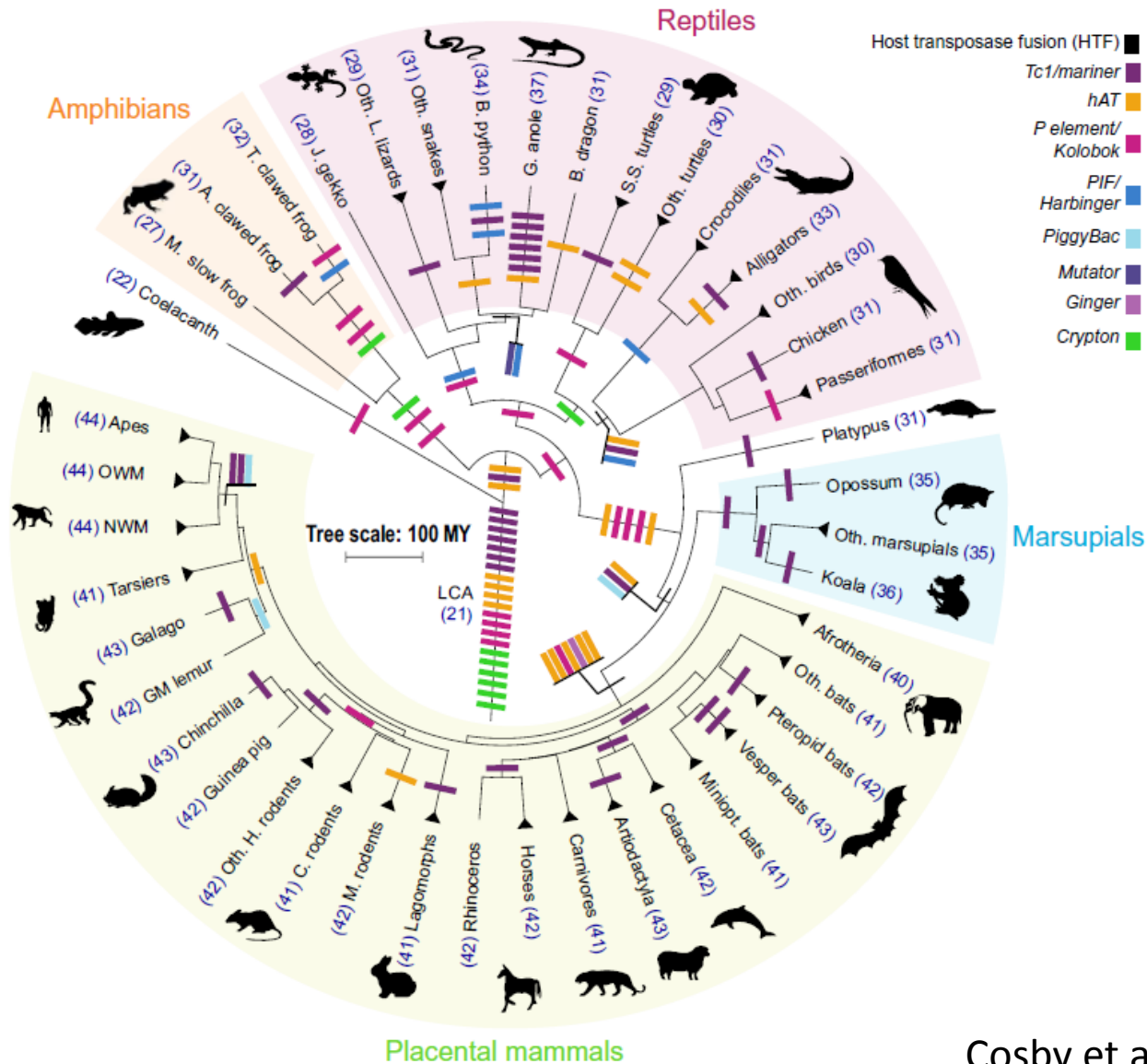
Mnohonásobný vznik transkripčních faktorů začleněním transpozázy u obratlovců



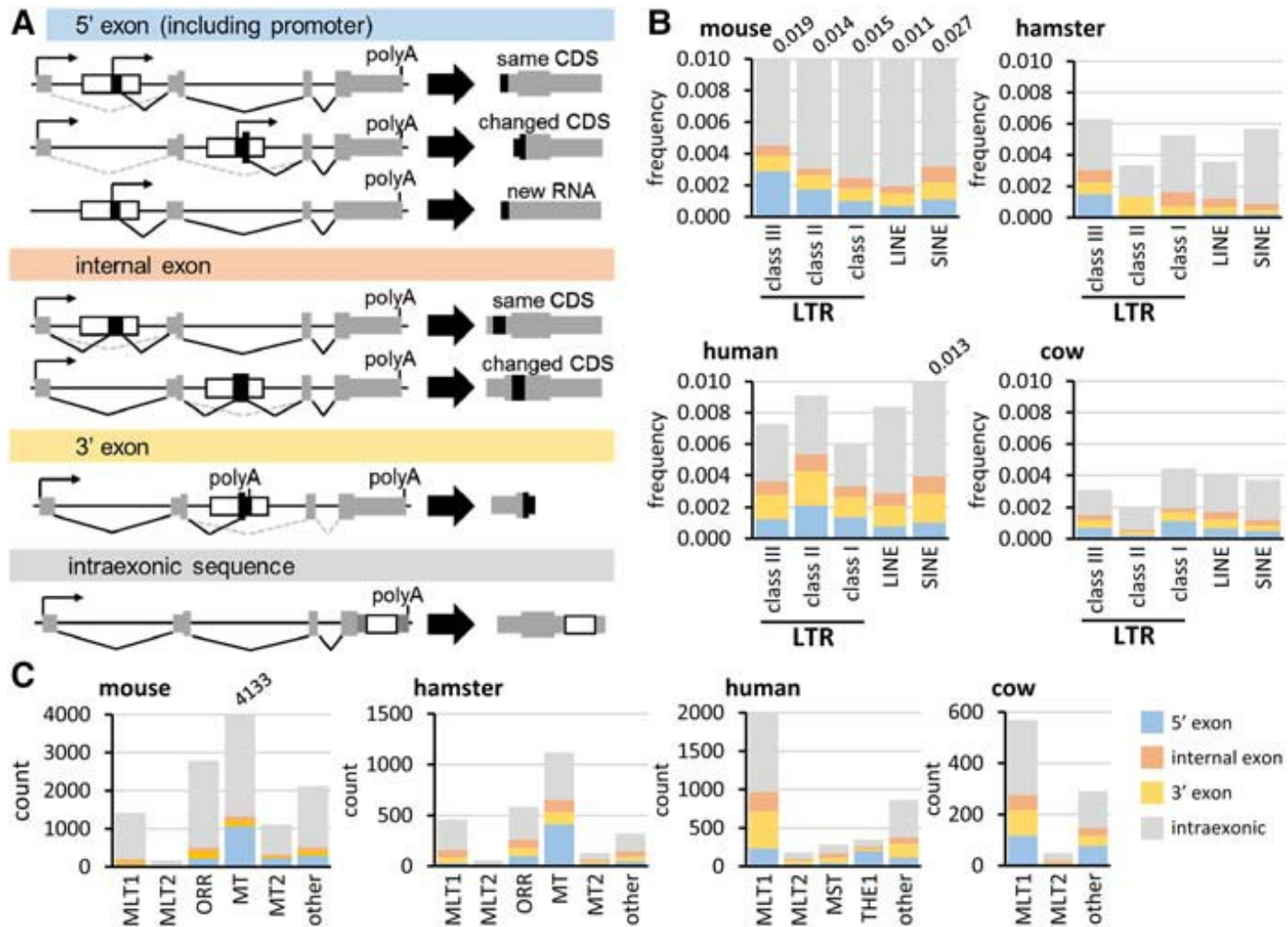
Mnohonásobný vznik transkripčních faktorů začleněním transpozázy u obratlovců

Fig. 1. Gene birth by transposase capture in tetrapods.

Tetrapod phylogenetic tree with boxes representing HTF fusion genes. Colors indicate the transposase superfamily assimilated. Numbers in parentheses indicate the number of HTF genes identified in the specified lineage. OWM, Old World monkeys; NWM, New World monkeys; GM, gray mouse; Oth., other; H., hystricoid; C., castorid; M., muroid; Miniopt., miniopterid; Vesper, vespertilionid; S.S., soft-shelled; B., bearded dragon; G., green; B., Burmese python; L., lacertid; J., Japanese; T., tropical; A., African; M., mountain; LCA, last common ancestor; MY, million years.



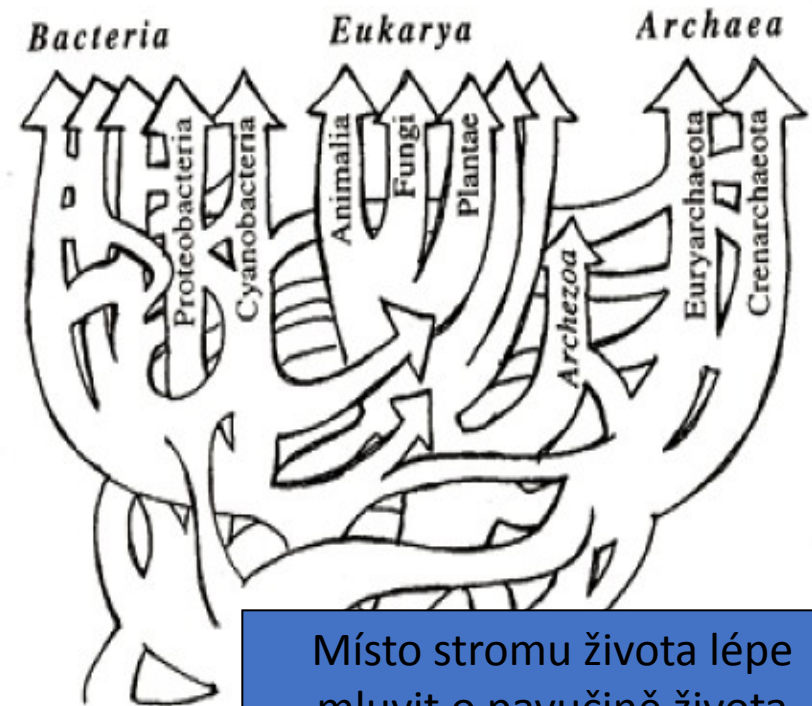
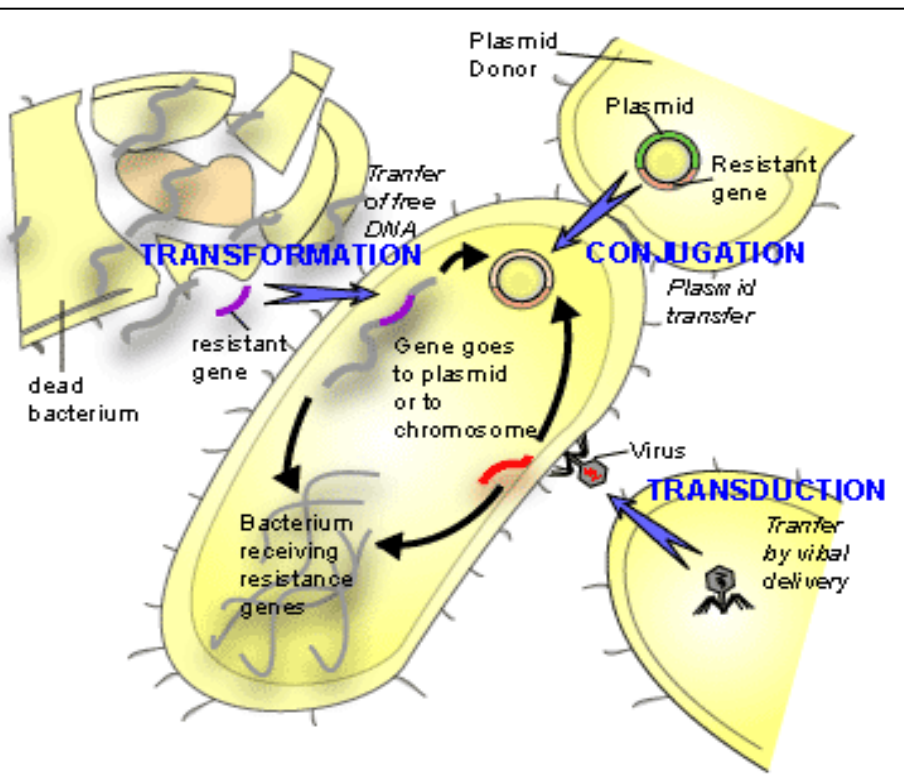
Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes (Franke et al., 2017)



Původ nových genů: Horizontální přenos

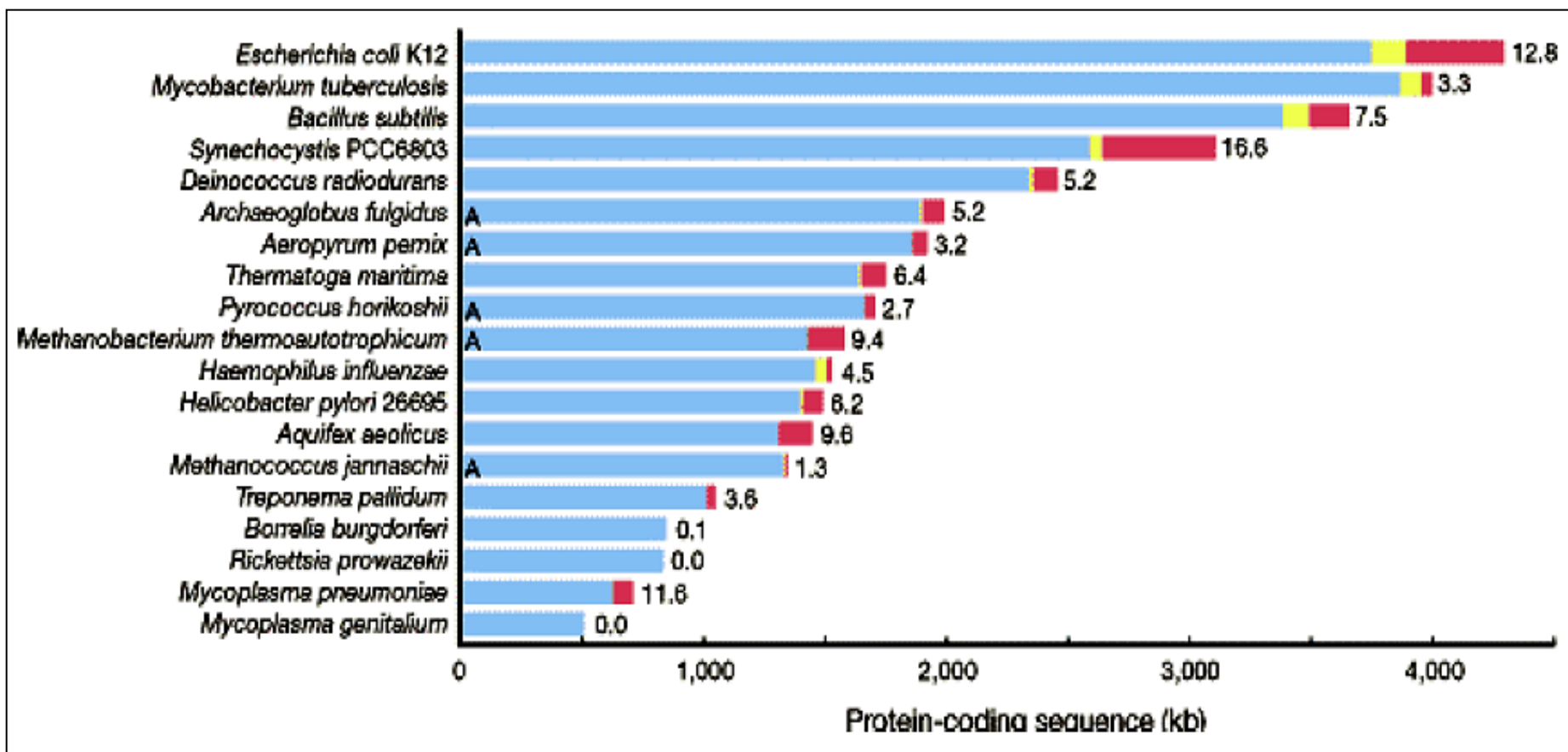
- vertikální (sexualita) a horizontální přenos (mezi druhy)
- bakterie - konjugace, transdukce a transformace
- vířníci pijavenky (Bdelloidea) – z bakterií, hub, řas, prvoků
- vnitrobuněčný parazitismus (Wolbachia)
- DNA transposony
- endosymbióza – promiskuitní DNA
- GMO organizmy si budou vyměňovat geny s ne-GMO

Přírodní genetické inženýrství je časté, dokonce i mezi evolučně vzdálenými taxony



Místo stromu života lépe mluvit o pavučině života

Horizontální genový přenos u bakterií



Metody studia

přímé:

Subtraktivní hybridizace

Microarrays

nepřímé:

Zastoupení kodonů (codon bias)

GC obsah

Konzervativní pořadí genů

Vysoká homologie se vzdáleným druhem

Původ nových genů: Štěpení a fúze genů (na základě studia ortologů)



Větší genom – více fúzí

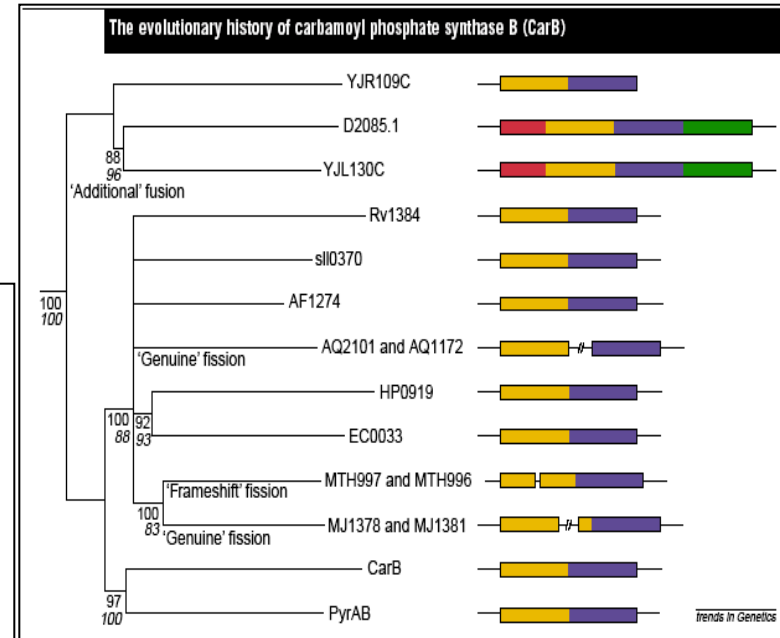
Number of gene organizations resulting from fission and fusion

| Species ^c | Genome size ^a | Fusion | Fission | | |
|---|--------------------------|--------|---------|----------------------|-------------------------|
| | | | Total | Genuine ^b | Frameshift ^b |
| <i>Mycoplasma genitalium</i> | 468 | 2 | 2 | 1 | 1 |
| <i>Mycoplasma pneumoniae</i> | 677 | 2 | 1 | 0 | 1 |
| <i>Rickettsia prowazekii</i> | 834 | 6 | 2 | 0 | 2 |
| <i>Borrelia burgdorferi</i> | 850 | 3 | 1 | 1 | 0 |
| <i>Chlamydia trachomatis</i> | 876 | 8 | 0 | 0 | 0 |
| <i>Treponema pallidum</i> | 1031 | 6 | 0 | 0 | 0 |
| <i>Aquifex aeolicus</i> | 1522 | 12 | 13 | 8 | 5 |
| <i>Helicobacter pylori</i> 26695 | 1590 | 9 | 0 | 0 | 0 |
| <i>Haemophilus influenzae</i> | 1717 | 18 | 13 | 3 | 10 |
| <i>Methanococcus jannaschii</i> | 1735 | 12 | 7 | 5 | 2 |
| <i>Methanobacterium thermoautotrophicum</i> | 1871 | 16 | 18 | 5 | 13 |
| <i>Pyrococcus horikoshii</i> | 2061 | 4 | 3 | 3 | 0 |
| <i>Archaeoglobus fulgidus</i> | 2407 | 19 | 9 | 8 | 1 |
| <i>Synechocystis</i> PCC6803 | 3168 | 24 | 4 | 4 | 0 |
| <i>Mycobacterium tuberculosis</i> | 3924 | 36 | 4 | 1 | 3 |
| <i>Bacillus subtilis</i> | 4100 | 19 | 1 | 1 | 0 |
| <i>Escherichia coli</i> | 4290 | 33 | 10 | 2 | 8 |

^aGenome size in number of predicted genes.

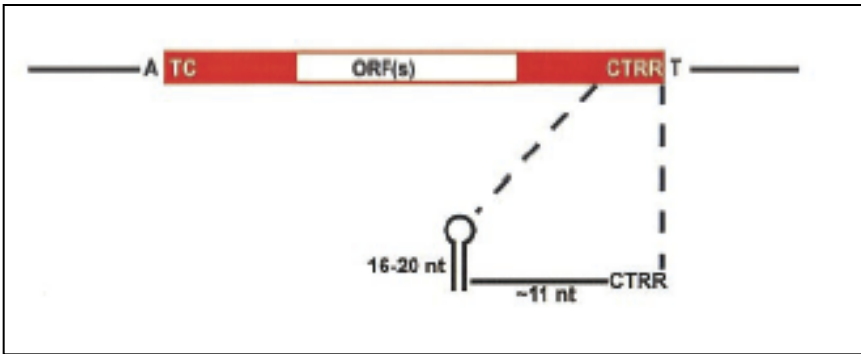
^bFor a definition of the subdivision in 'genuine' and 'frameshift', see text.

^cThermophilic species are shown in bold.



- častější fúze než štěpení
- štěpení u termofilů

Napomáhají Helitrony vzniku nových genů?



Rolling-circle transposons in eukaryotes

Vladimir V. Kapitonov* and Jerzy Jurka

Genetic Information Research Institute, 2081 Landings Drive, Mountain View, CA 94043

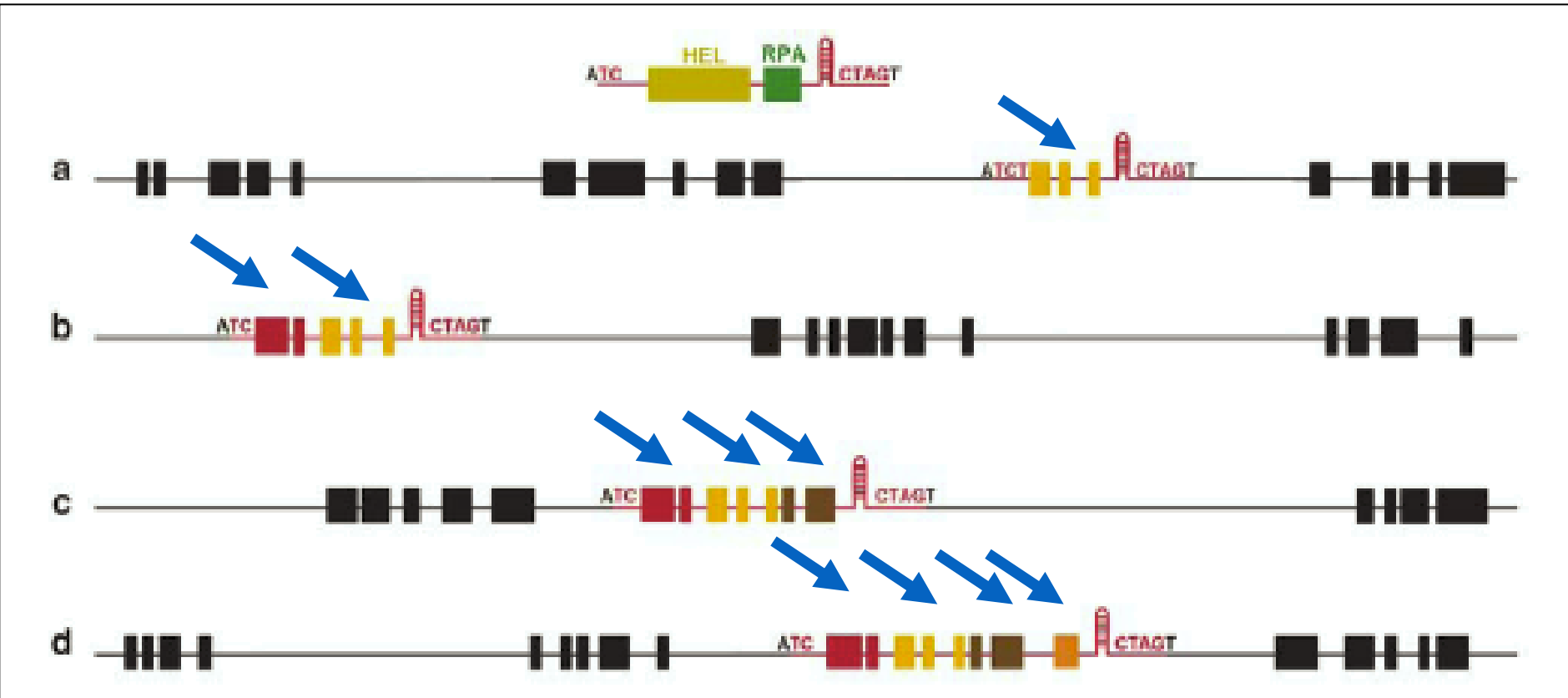
Communicated by Margaret G. Kidwell, University of Arizona, Tucson, AZ, May 29, 2001 (received for review April 10, 2001)

All eukaryotic DNA transposons reported so far belong to a single category of elements transposed by the so-called "cut-and-paste" mechanism. Here, we report a previously unknown category of eukaryotic DNA transposons, *Helitron*, which transpose by rolling-circle replication. Autonomous *Helitrons* encode a 5'-to-3' DNA helicase and nuclease/ligase similar to those encoded by known rolling-circle replicons. *Helitron*-like transposons have conservative 5'-TC and CTRR-3' termini and do not have terminal inverted repeats. They contain 16- to 20-bp hairpins separated by 10-12 nucleotides from the 3'-end and transpose precisely between the 5'-A and T-3', with no modifications of the AT target sites. Together with their multiple diverged nonautonomous descendants, *Helitrons* constitute ~2% of both the *Arabidopsis thaliana* and *Caenorhabditis elegans* genomes and also colonize the *Oriza sativa* genome. Sequence conservation suggests that *Helitrons* continue to be transposed.

and best illustrated by a recent study of *Sleeping Beauty*, a Tc1-like transposon from fish (13), reconstructed from its inactive copies and demonstrated to be transpositionally active in a test tube. Another much more ancient example is a PiggyBac-like DNA transposon, *Looper*, discovered in the human genome [V.V.K. and J.J. Rebase Update (1998) www.girinst.org/Rebase.Update.html], whose consensus sequence is based on a multiple alignment of the inactive copies, which are ~100 million years old. All genomic copies of *Looper* are mutated to the extent that no traces of its transposase could be detected at the sequence level. However, the transposase re-emerged from the virtual background noise after reconstructing the consensus sequence.

Materials and Methods

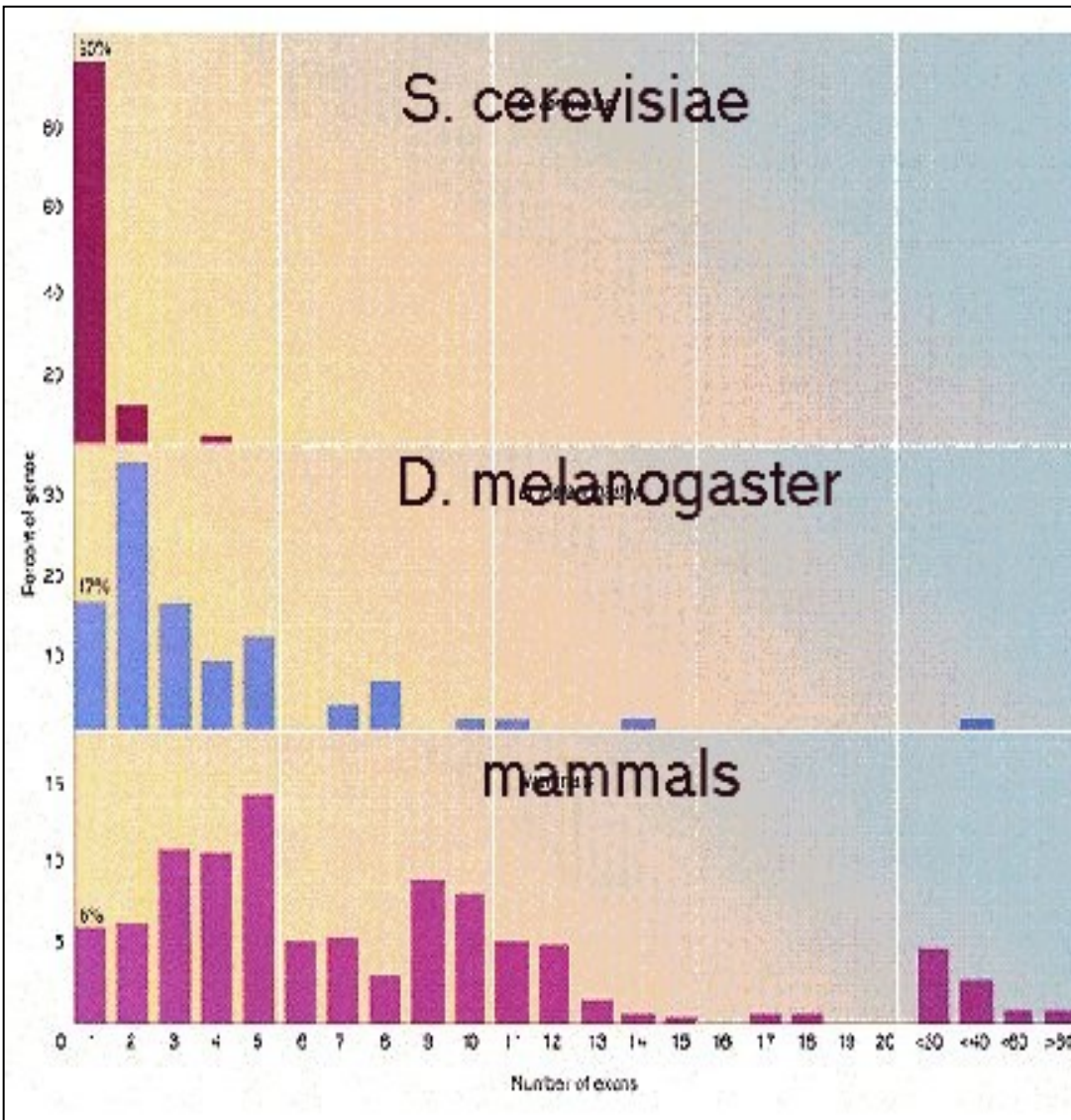
Computational Analysis. TEs reported in the manuscript were identified by running DNA sequences of prospective TEs against



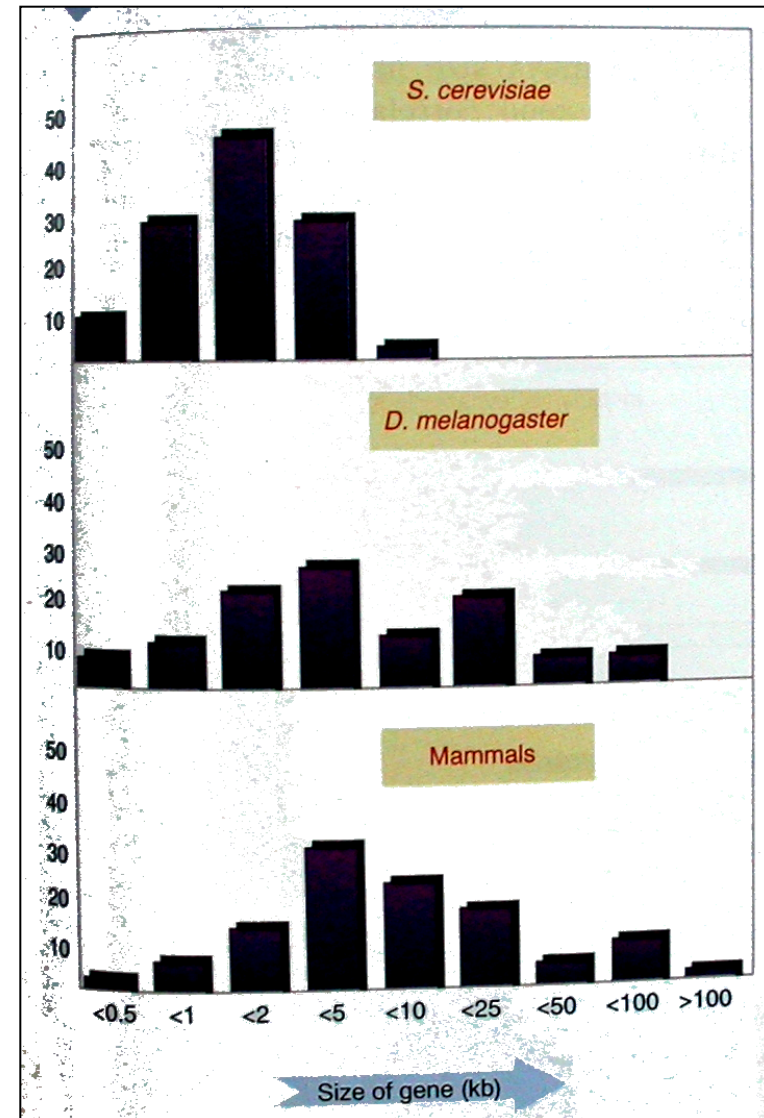
VELIKOSTI GENŮ

Počty exonů jsou nejvyšší u savců

Počet exonů



Délka genu (kb)

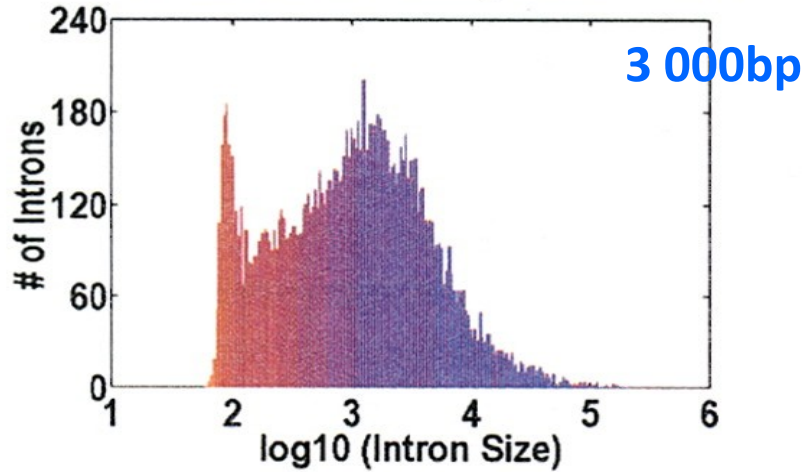


Velikosti intronů



Homo sapiens

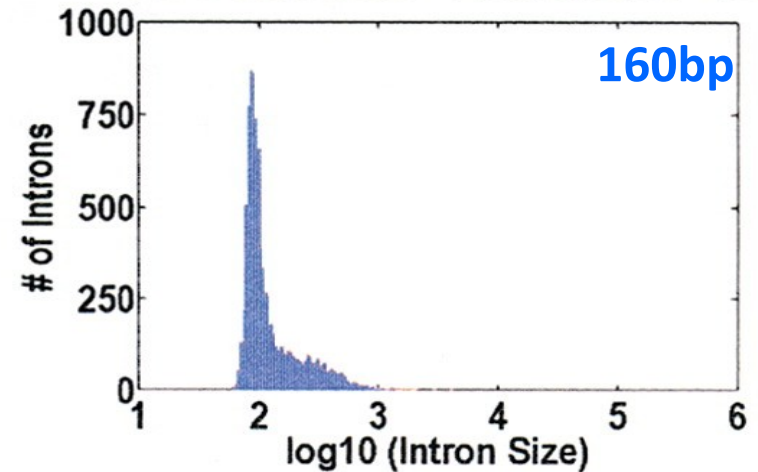
Intron Size: mean = 3116, median = 1044



(a)

Arabidopsis thaliana

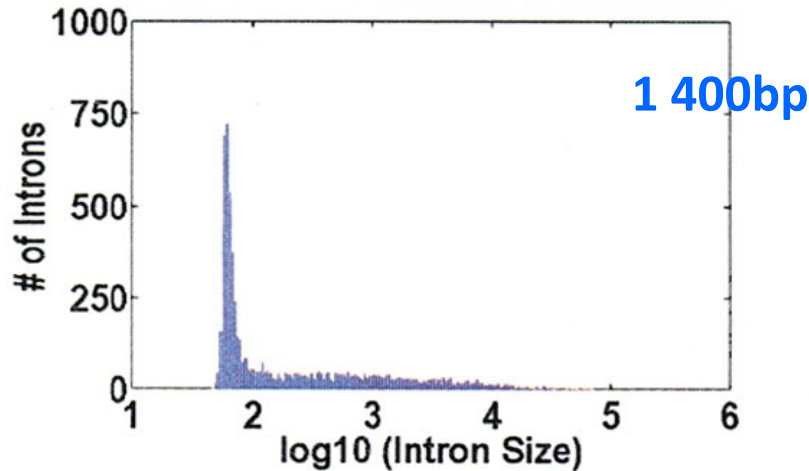
Intron Size: mean = 159.3, median = 98



(b)

Drosophila melanogaster

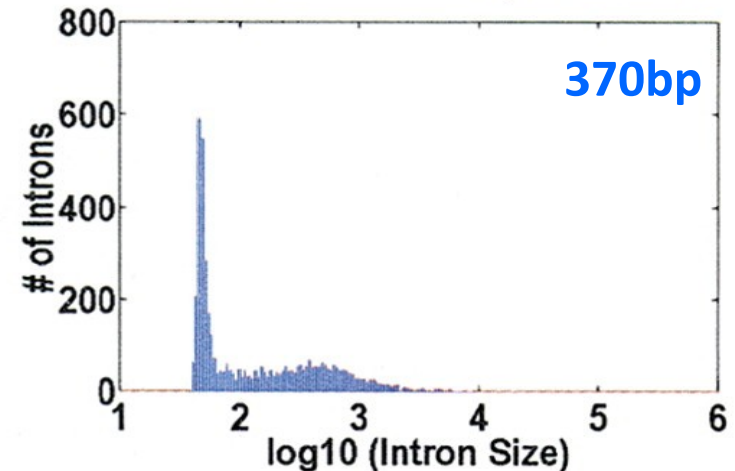
Intron Size: mean = 1411, median = 86



(c)

Caenorhabditis elegans

Intron Size: mean = 372, median = 85



(d)

Dystrofinový gen – obří gen

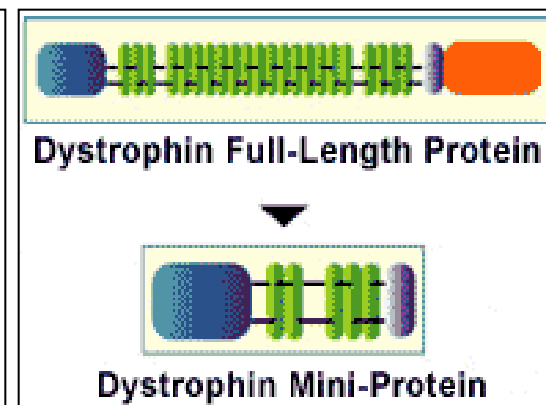
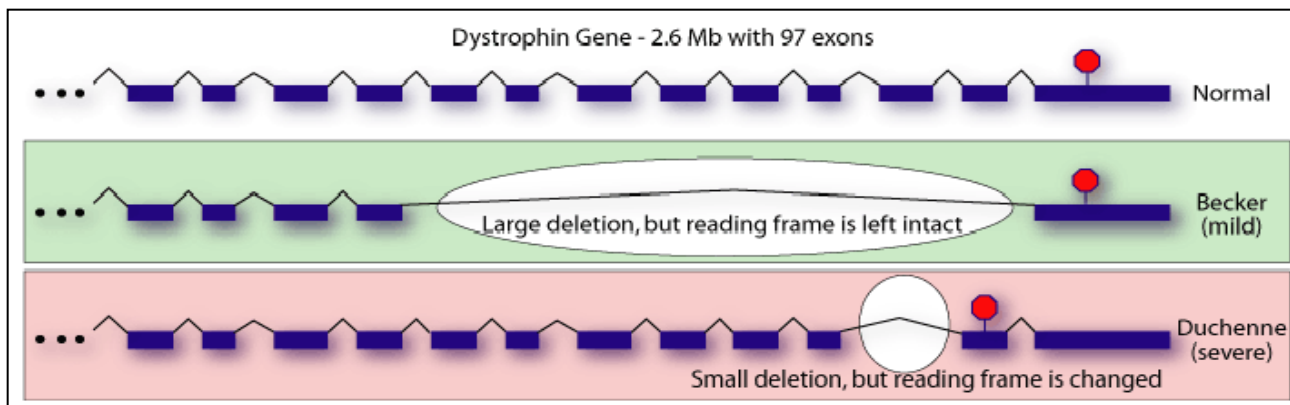
79 exonů, nejdelší známý gen

8 promotorů, exprese ve svalech a mozku

2.5 Mb dlouhý (0.1% genomu), 14kb mRNA

delece: Duchenne MD nebo Becker MD

Poloha Xp21, 1:3500 u mužů



INTRONY – STARÉ NEBO MLADÉ

Hypotézy původu intronů

„Intron first“:

- původní organizmy obsahovaly introny
- prokaryota je ztratila

„Intron late“:

- původní organizmy introny neobsahovaly
- eukaryota je získala

Význam intronů:

1. Introny **užitečné nejsou**, ale organizmy se jich nedokáží zbavit
2. Introny **mají funkční** význam pro organizmy, jsou užitečné

Introny byly v genech již na počátku ("intron first")

- studium vnitřní periodicity genů – stejné motivy v exonech i v sousedních intronech
- malá pravděpodobnost dlouhých úseků bez stop-kodonů,
- evoluční výhoda enzymatického aparátu, který vystřihne oblasti se stop-kodony a sestaví dlouhou mRNA

Introny byly do genů vloženy až dodatečně ("intron late")

- Existuje řada různých intronů lišících se mechanismem vystřihování z RNA – vznikaly nezávisle
- Distribuce intronů v rámci fylogenetických stromů svědčí o dodatečném vložení spíše než o opakovaném nezávislém vymizení

„Introns first“ versus „introns late“

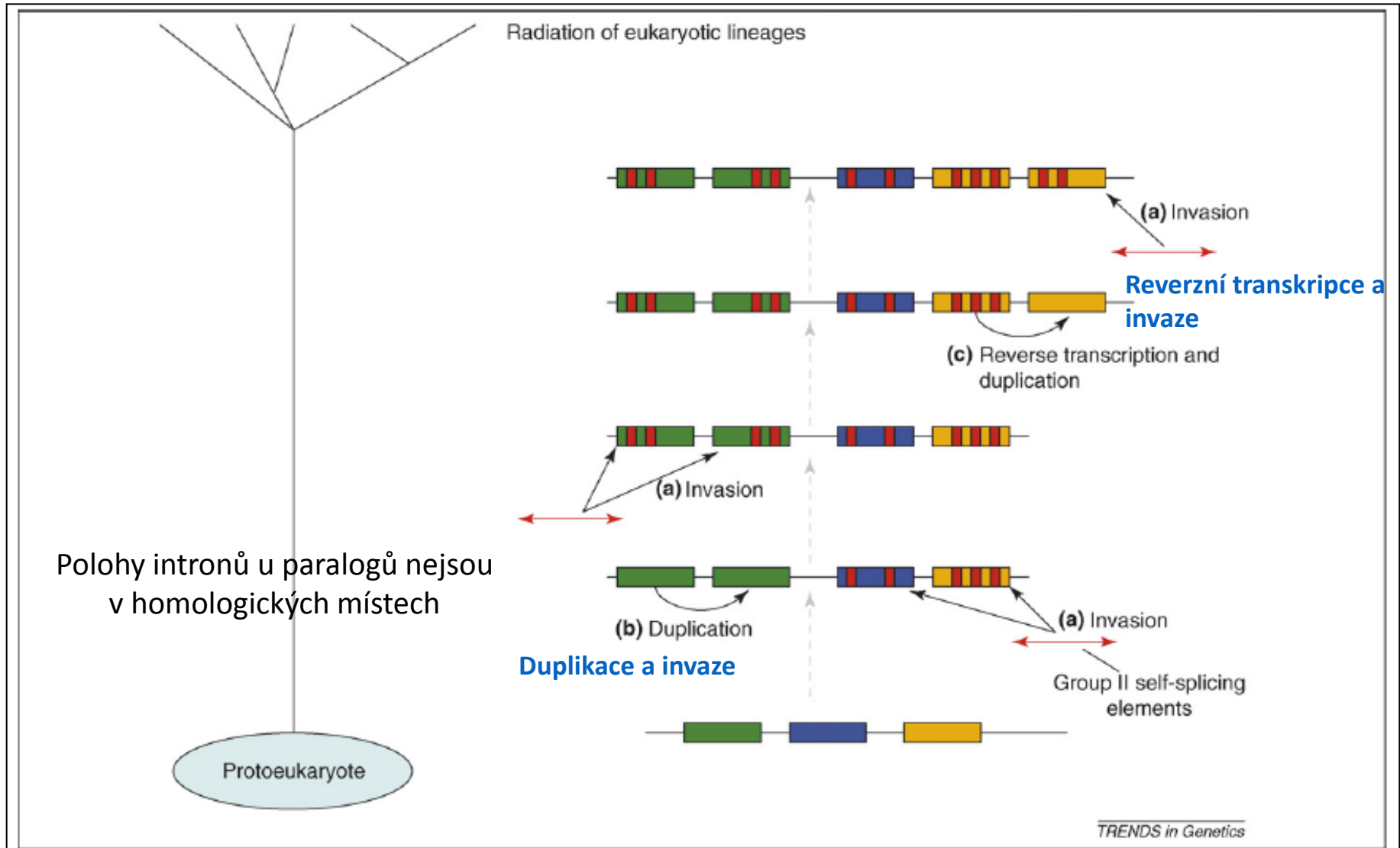
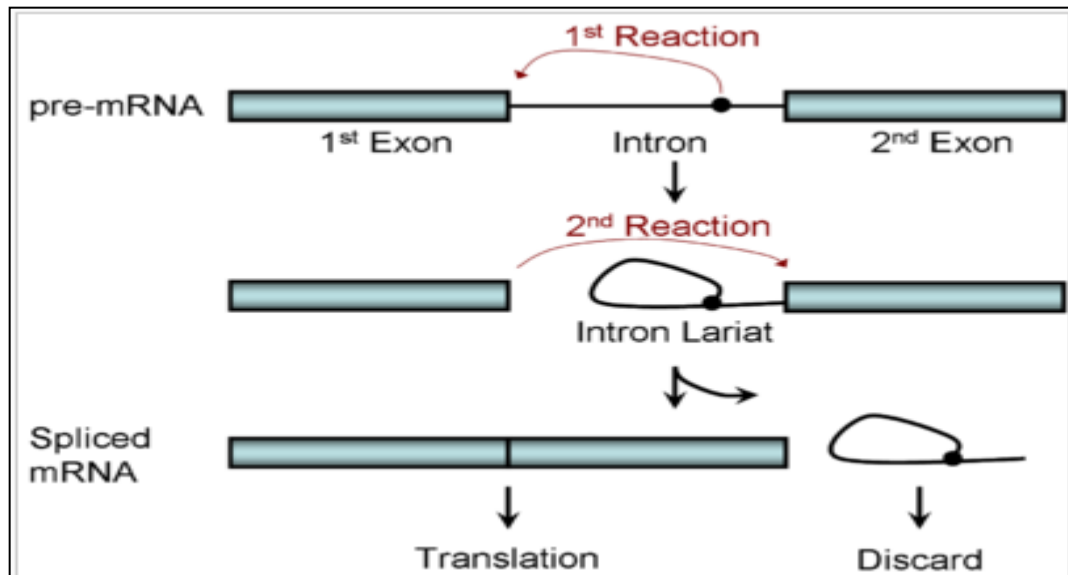


Figure 2. The processes that probably account for the lack of conservation of intron positions between ancient eukaryotic paralogs. (a) Ongoing invasion of group II self-splicing elements into eukaryotic genes, giving rise to spliceosomal introns. (b) Duplication of an intronless gene followed by differential insertion of introns into the paralogs. (c) Reverse-transcription-mediated duplication of an intron-containing gene, yielding an intronless paralog that, subsequently, accumulates introns in different positions. A schematic tree of eukaryotic evolution is shown, emphasizing that all of these processes are attributed to the time between the emergence of the eukaryotes and the radiation of the known eukaryotic lineages.

Introny jsou genomovými parazity

- Šíří se pouze v rámci genomu, vertikální přenos, aby nezabíjeli buňku, před translací se vystříhnou
- schopnost **samosestřihu**
- **Splicesom** – komplex kódovaný buňkou, původně parazitickými introny, kódují enzymy pro šíření v rámci genomu



Introny jsou **užitečné** pro organizmy

1. Zvyšují evoluční potenciál organismu

- souvisí se vznikem eukaryot, v pozadí adaptivní radiace eukaryot
- nenáhodná distribuce, odděluje funkční domény proteinů
- snižuje pravděpodobnost rekombinace v exonech (doménách)
- stavební charakter genů urychluje evoluci nových proteinů

2. Souvisí s existencí histonů

- oblasti v kontaktu s histony nepřístupné
- introny zpřístupňují regulační oblasti

3. Snižují riziko nelegitimní rekombinace

- paralogy a riziko nelegitimní rekombinace
- včlenění intronů do různých míst diferencuje geny, snižuje riziko NR

Geny na chromosomu Y degenerují,

mají delší introny

