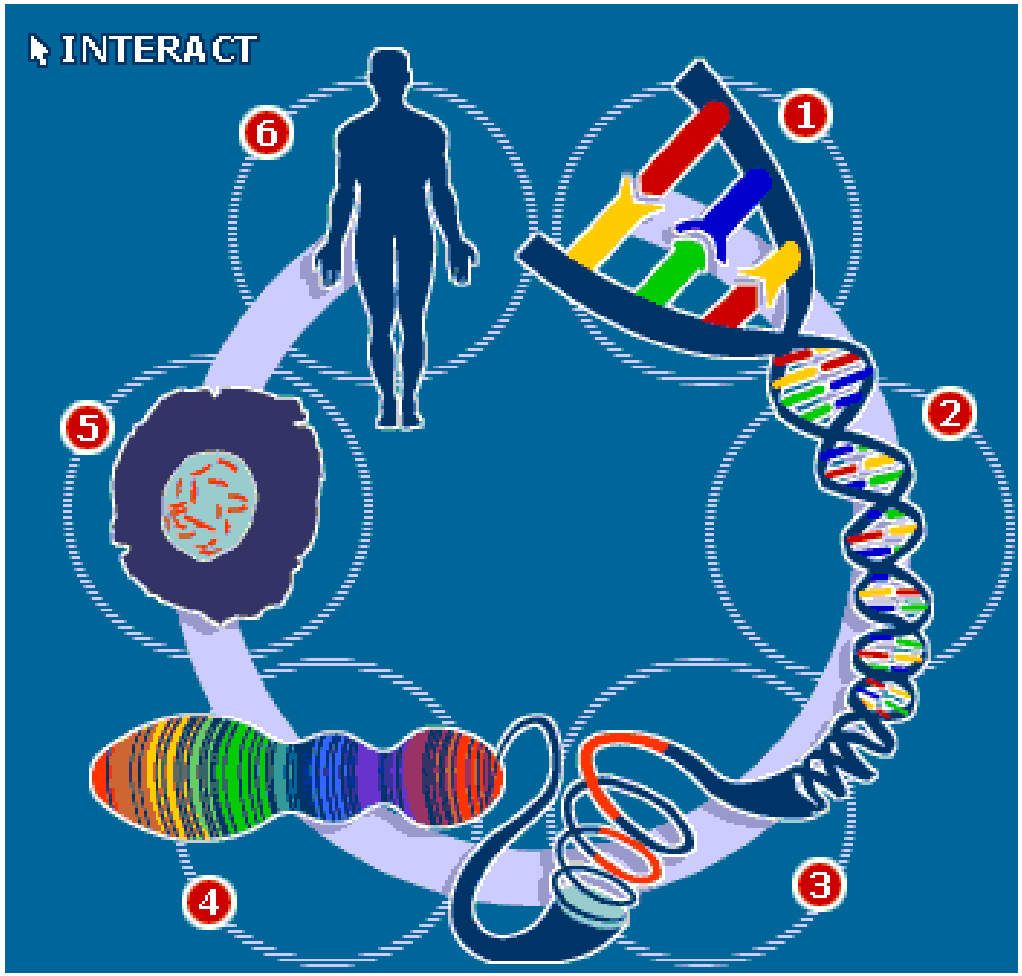
GENOMIKA

Az emberi genom felépítése



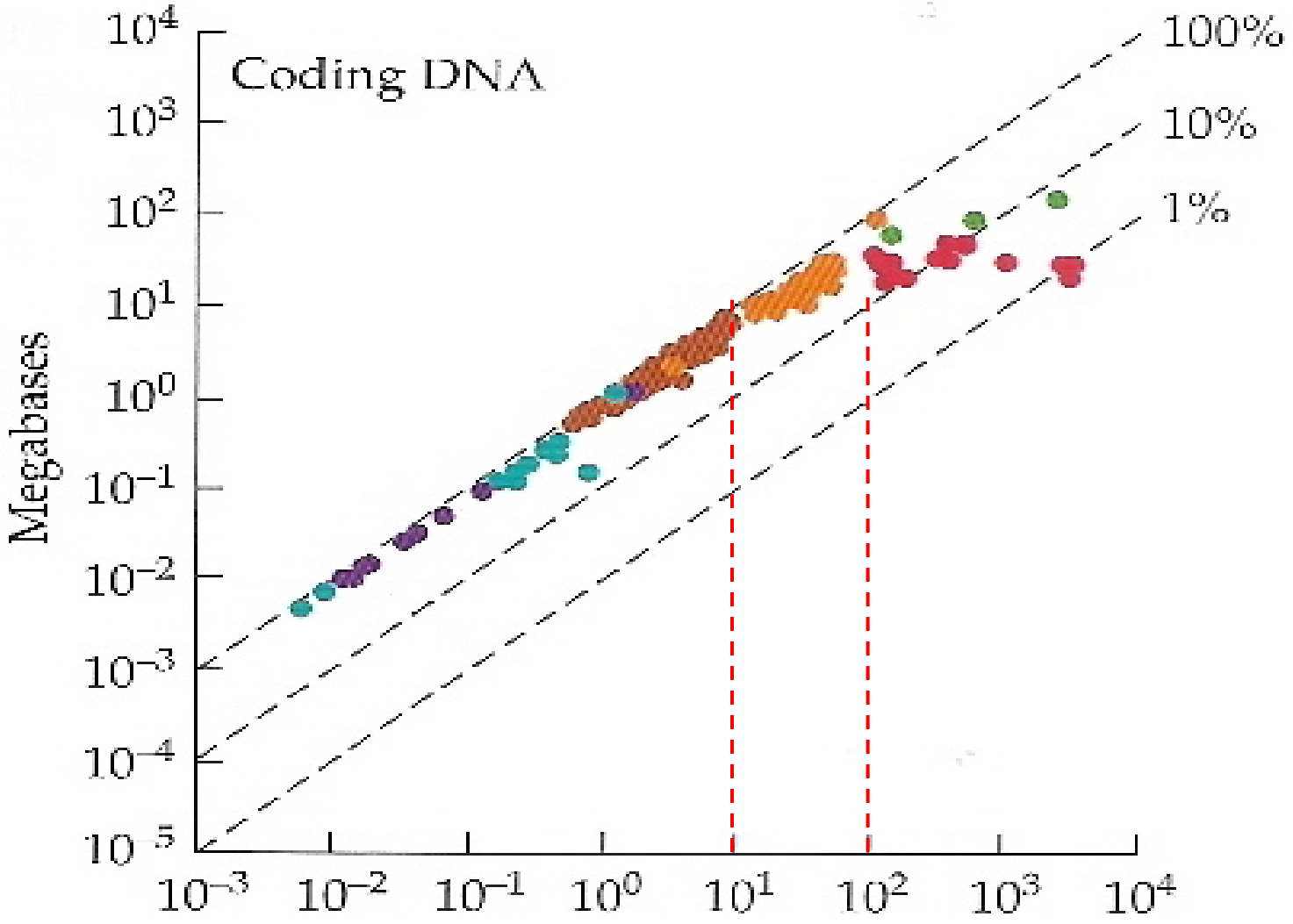
ELTE TTK Genetikai Tanszék

A humán genom projekt eredményei

- Előzetes publikálás 2001-ben (Science, Nature)
- Az eddig leírt legnagyobb méretű teljes genom (~ 2900 Mb)
- Az eukarióta genomra jellemző szerkezeti és szerveződési tulajdonságok (modellszervezetek)
- RNS gének növekvő száma (tRNA, snRNA, miRNA, stb.)
- Orphan gének: humán gének ~ 1/3-a
- Génszám és szöveti-, sejti sokféleség közti összefüggés?
- Meglepően alacsony a fehérje kódoló gének száma:

  ~ 22000, < 1 % a genomnak protein kódoló

Kódoló szekvenciák vs. genom méret



The Origins of Genome Architecture, Lynch, 2006

TABLE 3.2 Haploid genome size, number of protein-coding genes, and average number of nucleotides per gene for some well-characterized eukaryotic genomes

	GENOME SIZE (MB)	GENE NUMBER	KILOBASES/GENE		
			TOTAL	CODING	NON-CODING
Unicellular species					
Encephalitozoon cuniculi	2.90	1997	1.45	1.01	0.44
Saccharomyces cerevisiae	12.05	6213	1.94	1.44	0.50
Schizosaccharomyces pombe	13.80	4824	2.86	1.43	1.43
Cyanidioschyzon merolae	16.52	5331	3.10	1.55	1.55
Cryptococcus neoformans	19.05	6572	2.89	1.62	1.27
Plasmodium falciparum	22.85	5268	4.34	2.29	2.05
Entamoeba histolytica	23.75	9938	2.39	1.14	1.25
Leishmania major	33.60	8600	3.91	2.15	1.76
Thalassiosira pseudonana	34.50	11242	3.07	0.99	2.08
Trypanosoma spp.	39.20	10000	3.92	1.96	1.96
Oligocellular species					
Ustilago maydis	19.68	6572	2.99	1.84	1.15
Aspergillus nidulans	30.07	9541	3.15	1.57	1.58
Dictyostelium discoideum	34.00	9000	3.78	2.45	1.33
Neurospora crassa	38.64	10082	3.83	1.44	2.39
Land plants					
Arabidopsis thaliana	125.00	25498	4.90	1.80	3.10
Oryza sativa	466.00	60256	7.73	1.18	6.55
Lotus japonicus	472.00	26000	18.15	1.35	16.80
Animals					
Caenorhabditis elegans	100.26	21200	4.73	1.25	3.48
Drosophila melanogaster	137.00	16000	8.56	1.66	6.90
Ciona intestinalis	156.00	16000	9.75	0.95	8.80
Anopheles gambiae	278.00	13683	20.32	1.64	18.68
Fugu rubripes	365.00	38000	9.61	0.93	8.68
Bombyx mori	428.70	18510	23.16	1.66	21.50
Gallus gallus	1050.00	21500	48.84	1.44	47.40
Mus musculus	2500.00	24000	83.33	1.30	82.03
Homo sapiens	2900.00	24000	96.67	1.33	95.36

Source: Lynch 2006a.

Gének száma

vs.

Kódoló szekvenciák hossza

Genom méret

vs.

Nem-kódoló szekvenciák hossza

TABLE 3.1 Approximate fractional composition of the human genome

TYPE OF DNA	FRACTION
Coding exons	0.008
Internal introns	0.308
5′ Untranslated regions	
Exons	0.045
Introns	0.002
3′ Untranslated regions	
Exons	0.006
Introns	0.001
Intergenic DNA	0.683
Conserved noncoding DNA	0.016
Pseudogenes	0.007
Mobile genetic elements	0.446

Note: Derived from various references given in the text. Intergenic DNA is all DNA except coding exons and internal introns. The fractions do not sum to one because mobile elements, pseudogenes, and transcription factor binding sites reside in introns, UTRs, and/or intergenic DNA.

The Origins of Genome Architecture, Lynch, 2006

Génduplikációk, funkcionális géndiverzitás

- ~ 4000 pár humán duplikált gén (multigén családok nélkül)
- Genom 5 %-a recens szegmentális duplikáció
- Duplikációs ráta: 0,01/ gén/ millió év; Silencing: 10M év

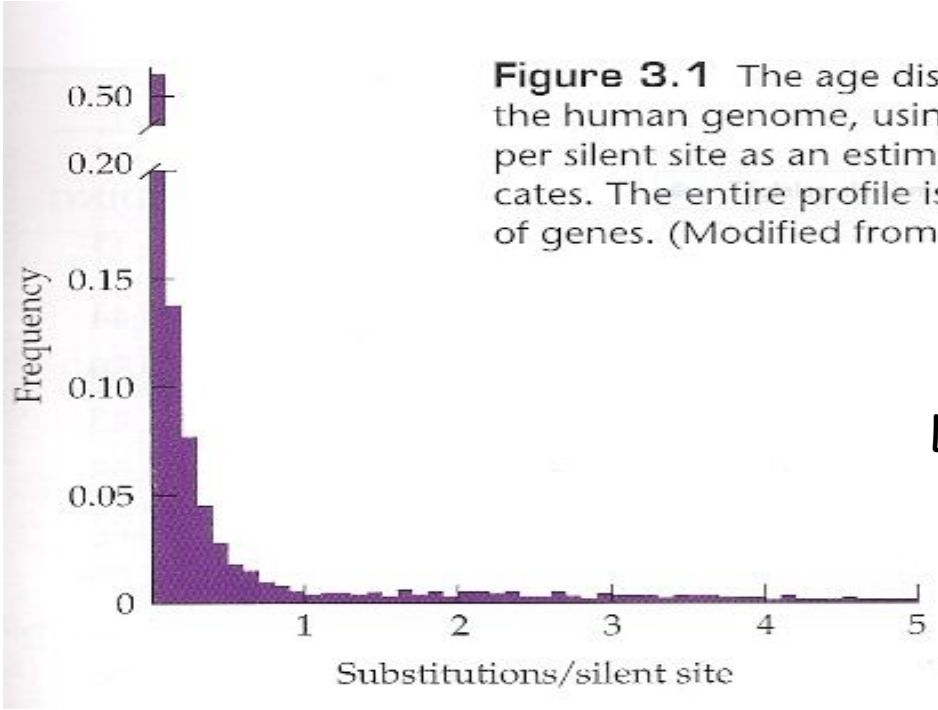


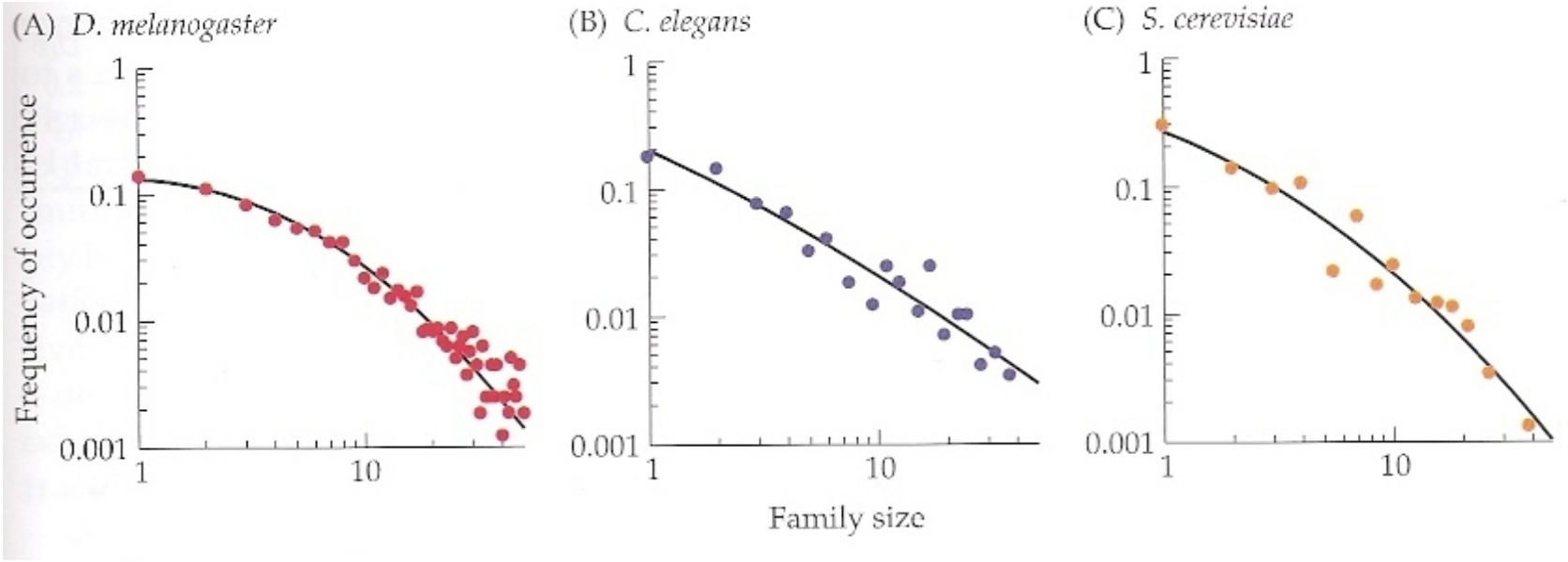
Figure 3.1 The age distribution of duplicate genes in the human genome, using the number of substitutions per silent site as an estimate of the age of a pair of duplicates. The entire profile is based on a survey of 3892 pairs of genes. (Modified from Lynch and Conery 2003a.)

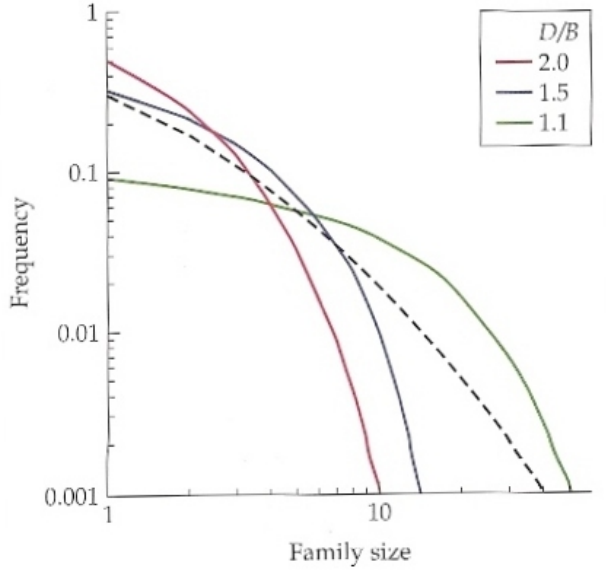
L-alakú koreloszlás

Génduplikációk, funkcionális géndiverzitás

- Duplikációs ráta: 0,01/ gén/ millió év
- Génduplikátum kikapcsolása átlagosan kb. 10 millió év
- Gének jelentős hányada nem nélkülözhetetlen (redundáns)
- Génexpanzió és kontrakció sztochasztikusan: adaptáció?
- Kiegyensúlyozott génexpanzió és kontrakció a humán és egér vonalak között – olfactory receptor gene inaktiváció
- Nem szükséges adaptív relevancia (de lehet: immun, repr.)

Singleton és multi- géncsaládok eloszlása





N(x): $x^{-b;}$ b: ~ 1.5-2.0

x: géncsalád tagjainak száma

B: génduplikáció valószínűsége

D: géndeléció valószínűsége

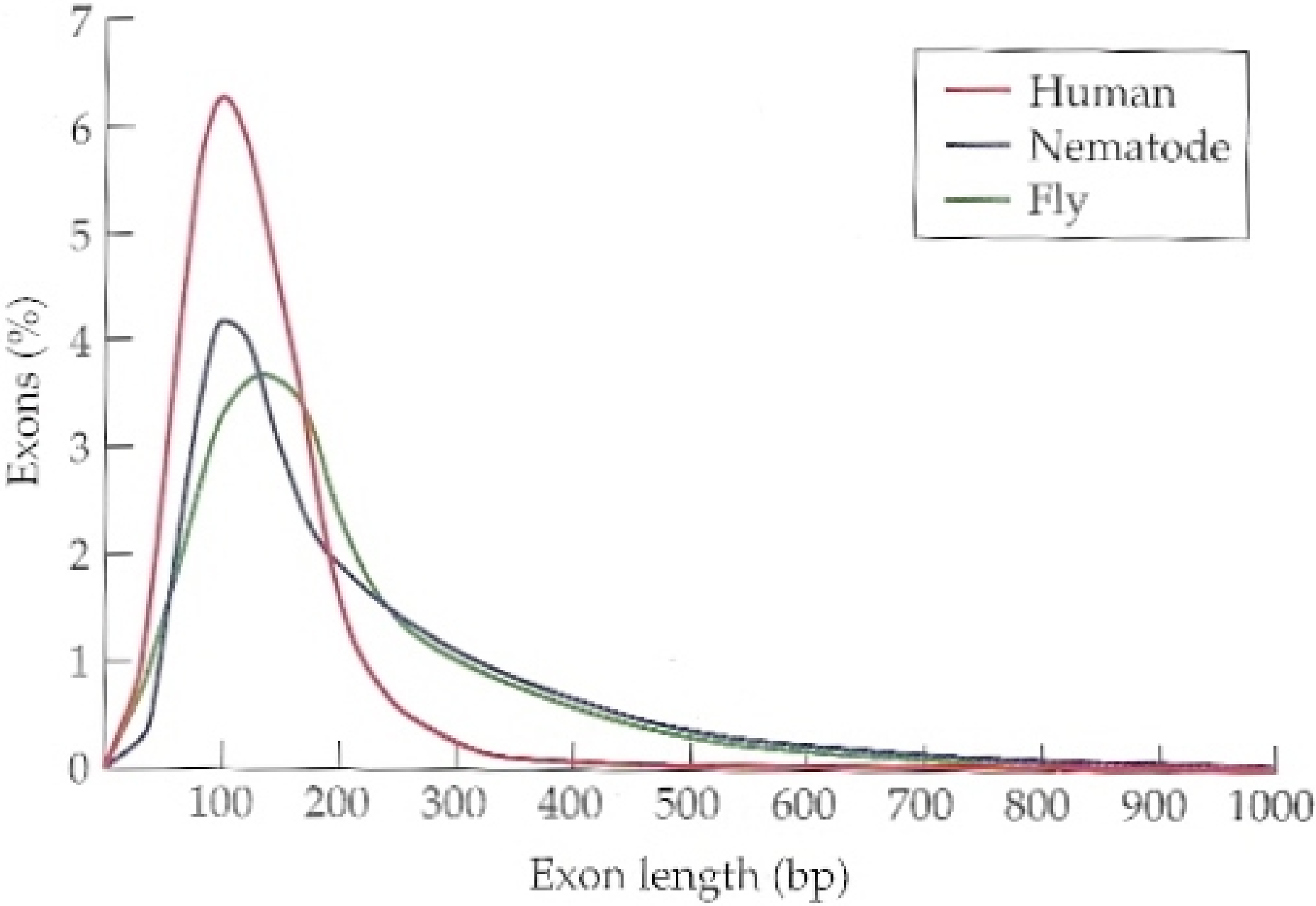
D > B; D/B: 1.5-2.0

The Origins of Genome Architecture, Lynch, 2006

Exonok és Intronok

- Eukarióta protein szekvenciák átlagos hossza azonos, viszont a nem kódoló intragenikus szakaszok varianciája nagy
- Átlagos humán gén: 7.7 intron, 0.15 kb exon, 4.66 kb intron
- Invertebrata: lényegesen kevesebb intronikus szakasz, átlagos exon méret nagyobb mint az ember esetében

  (*Saccharomyces:* intron mentes, *C. elegans:* 5.2 intron /120 bp
- Humán genom: exon méret redukált varianciája (< 300 bp)
- Splicing mechanizmus: „exon scanning"



The Origins of Genome Architecture, Lynch, 2006

Exonok és Intronok

- Génfunkció diverzifikálása génszám növelése nélkül.
- Alternatív splicing az emberi genomban: gének többsége.
- Alternatív splice variánsok 20 %-a szövet-specifikus.
- Funkcionális proteinek >> gének száma.
- Alternatív splicing és szervezeti komplexitás?

- *C. elegans, Drosophila:* gének 20 %, 1.3 transzkript /gén.
- *Humán:* gének több mint 50 %, 2.6 transzkript variáns /gén.
- Funkcionális domének száma kb. 2 x invertebrata (pl. Drosi)

- ~ 1,5 – 2 x több gén, de 50.000 addícionális fehérje.
- Humán és egér vonal: minor splice variánsok 70 %-a *de novo*.

TABLE 3.3 Average amount of DNA per gene (in kilobases) associated with coding exons, internal introns, and intergenic spacers (outside points of translation initiation and termination)

			INTERGENIC	
	EXON	INTRON	REGULATORY	OTHER
Saccharomyces	1.44	0.02	0.11	0.37
Aspergillus	1.57	0.27	0.03	1.55
Plasmodium	2.29	0.25	0.04	1.76
Caenorhabditis	1.25	0.64	0.43	2.41
Drosophila	1.66	2.93	1.37	2.60
Homo/Mus	1.32	32.27	1.95	61.14

Note: Exonic and intronic DNA includes only that associated with the coding region, i.e., excludes UTR regions, which are included in the intergenic categories. Estimates for the intergenic regulatory DNA category are based on islands of observed intergenic sequence conservation among closely related species: *Saccharomyces* (Kellis et al. 2003); *Aspergillus* (Galagan et al. 2005); *Plasmodium* (van Noort and Huynen 2006); *Caenorhabditis* (Webb et al. 2002); *Drosophila* (Bergman and Kreitman 2001; Andolfatto 2005); *Homo/Mus* (Shabalina et al. 2001). Intergenic other refers to all DNA between the stop codon of an upstream gene and the start codon of the following gene that is not discernable as intergenic regulatory. Qualitatively similar results have been obtained with other methods (e.g., Siepel et al. 2005).

The Origins of Genome Architecture, Lynch, 2006

Szabályozó elemek a genomban

- Szervezeti komplexitás: non-coding DNA /gén,
- variabilitás: egysejtű – többsejtű – gerinces – humán *(tábl.)*,
- Komplex identifikáció (ORF?), ortológ szekvenciák?
- *transzkripciós faktor kötőhelyek, exon-intron határ motívumok, transzkripció termináció,*
- Konzervatív becslés: átlagosan 2 kb / gén, *mi a maradék?*
- Egér /ember: 66.000 konzervált intergenikus blokk (150 bp)
- *90-100 % szekvencia azonosság, erős szelekciós nyomás*
- Génexpresszió: enhancer és repressor kötőhelyek
- Funkcionális RNS átíródás?

RNS transzkripció

TABLE 3.4 A glossary of terminology for RNAs

mRNAs	Messenger RNAs; mature gene transcripts, after introns have been processed out of the mRNA precursor
miRNAs	Micro RNAs; generally 20–30 bp in length, and processed from transcribed "hair-pin" precursor RNAs; used in the regulation of gene expression by complementary binding to nearly identical motifs in the 3′ UTRs of transcripts
ncRNAs	Noncoding RNAs; loosely defined as any transcript that does not encode protein
rRNAs	Ribosomal RNAs; the RNA subunits of the ribosome
sRNAs	Small RNAs; a generic term that encompasses miRNAs and siRNAs
siRNAs	Small interfering RNAs; generally 20–30 bp in length, and processed from longer double-stranded RNAs by the RNA interference pathway; deployed in posttranscriptional gene silencing
snRNAs	Small nuclear RNAs; a heterogeneous group of small RNAs whose functions are confined to the nucleus, including those involved in splicing introns out of precursor mRNAs and in telomere maintenance
snoRNAs	Small nucleolar RNAs; involved in the chemical modifications made in the construction of ribosomes; often encoded within the introns of ribosomal protein genes
tRNAs	Transfer RNAs; serve as vehicles for delivering amino acids during the translation of an mRNA

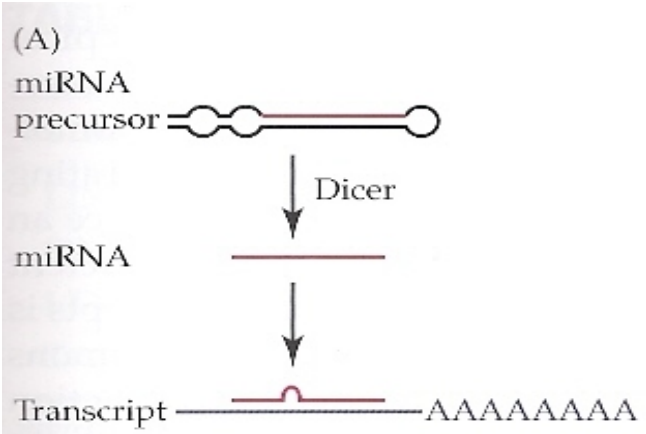
The Origins of Genome Architecture, Lynch, 2006

Génexpresszió szabályozása

Microarray technológia: RNS transzkripció nem kódoló (intergenikus) DNS szakaszokról is.

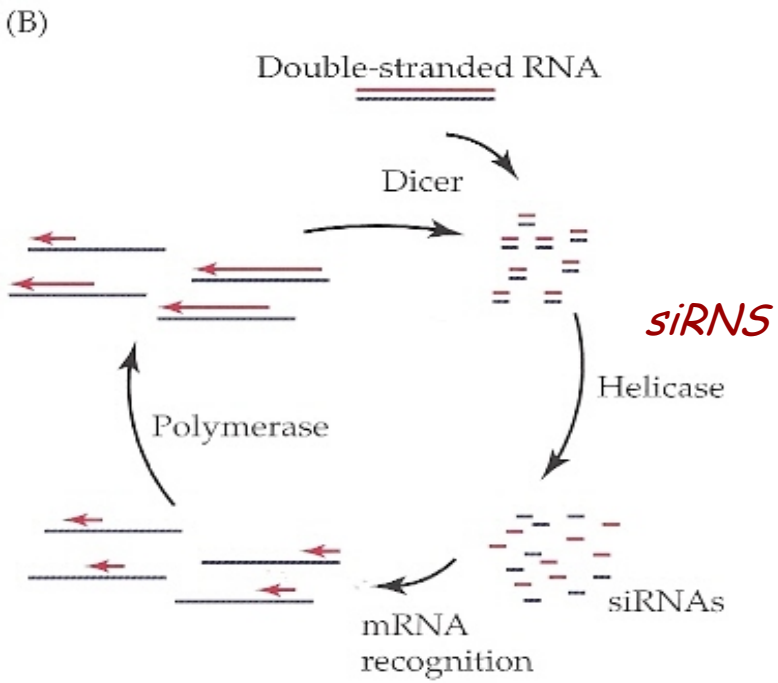
Nem kódoló RNS transzkriptumok ***szövet-specifitása*** !

miRNA, siRNA – transzlációs géncsendesítés és interferencia



miRNS:

- *stage or tissue specific*
- *complementarity to 3'UTRs*
- *translational silencing*



Mobilis Genetikai Elemek

- Extra mennyiség a humán genomban: 100/gén (~ 45 %)
- Humán genom ~ 75 %-a lehet mobilis elem maradványa
- Mutagén hatások: pl. inzerció, nem homológ rekombináció,

→ negatív konzekvencia a gazdára nézve

- Retrotranszpozon: „copy-and-paste", LINEs, SINEs, LTRs
- Transzpozonok: „cut-and-paste"

TABLE 2.2: CLASSES OF DISPERSED REPEATS IN THE HUMAN GENOME.

Class	Copy no. per haploid genome	Fraction of genome	Autonomous transposition or retrotransposition?	Length
LINEs	850 000	21%	Yes	Up to 6–8 kb
SINEs	1 500 000	13%	No	Up to 100–300 bp
Retrovirus-like elements	450 000	8%	Complete copies, yes	6–11 kb (1.5–3 kb)
DNA transposon copies	300 000	3%	Complete copies, yes	2–3 kb (80–3000 bp)

Values given in parentheses are lengths of incomplete elements, incapable of autonomous transposition (see Section 3.4). Adapted from Lander *et al.* (2001).

Long Interspersed Nuclear Elements - LINEs

- **LINEs v. *Kpn*:** emberi genom 20 %-a, 870.000 kópia
- kb. 100 LINEs működőképes retrotranszpozonként

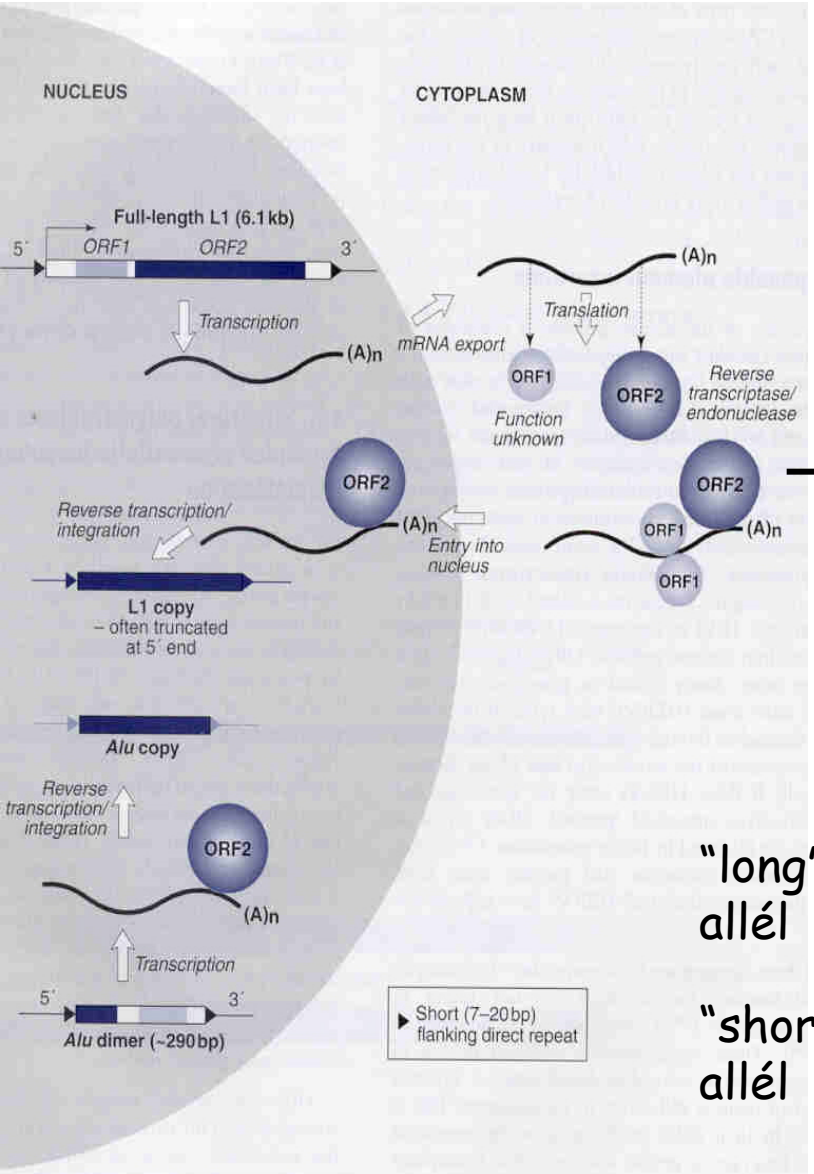
~ 6 kb, belső 5' promóter, 2 ORF (RNS-kötő fehérje, endonukleáz + reverz transzkriptáz), poly(A)-farok,

- Target-primed reverz transzkripció: TT | *AAAA* – target
- Hanyag másolás

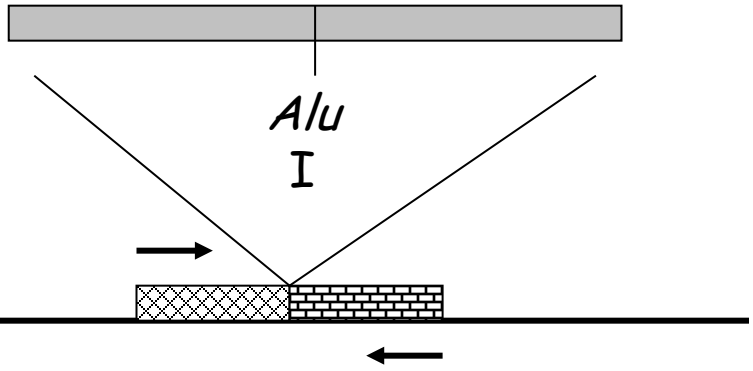
(transzkripciós „read-through", „dead-on-arrival", nagyobb szekvenciarészek átrendeződése, egyéb nem autonóm szekvenciákhoz való kötődés)

- Önmaguktól nem tudnak a genomból kivágódni (deléció)
- ősi és relatíve új szekvenciák (pl. LINE-1: 5 MYs)

Mobilis elemek: biallélikus hossz-polimorfizmus



Human *Alu* Repeat
(~300 bp)



Kétféle alléltípus

"long" (+) allél

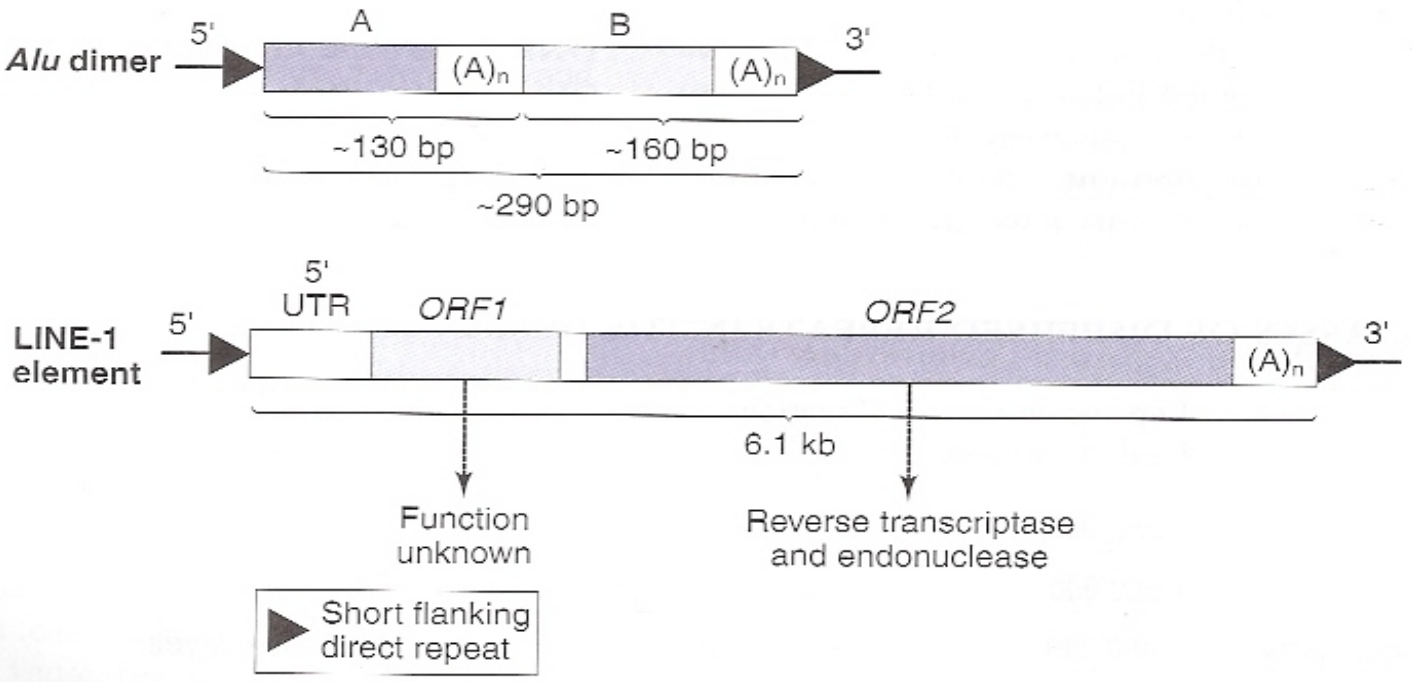
400 bp

"short" (-) allél

100 bp

Short Interspersed Nuclear Elements - SINEs

- **SINEs v. *Alu***: 1.500.000 kópia, 70 % *AluI*, 300 bp,
- Főemlős specifikus, *Alu I* : AGCT, polimorfizmusok,
- Nem kódoló szekvencia, önállóan nem mobilizálódik
- *Alu* – LINE-1 retrotranszpozíció, 0.05 /genom / generáció



Human Evolutionary Genetics, Jobling et al, 2004

LTRs és Transposons

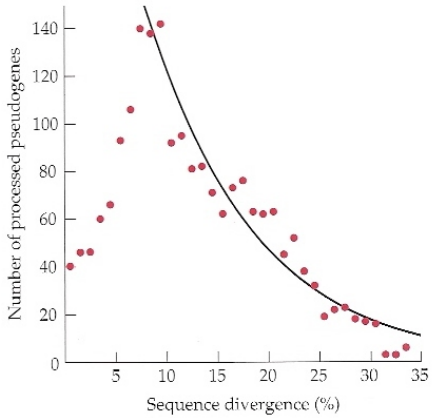
- **LTRs:** Long Terminal Repeats, Retrovírus eredet
- HERVs: Human Endogenous Retroviruses
- Reverz transzkripció és integráció: saját primer kötőhely
- Szöveghű másolás: ds-DNA → nucleus, mutáció, divergencia

(szubsztitúciós ráta: 1.25×10^{-9}, neutral sites, óvilági majmok)

- *env* gén: sejtek közötti mozgás, horizontális transzfer
- **Transzpozonok:** „cut-and-paste", több család
- TIRs (terminal inverted repeats), kisméretű duplikációk,
- transposase enzime – TIR kötőhely – excizió / inzerció
- DNA repair, homológ kiegészítés, sokszorozódás

Pszeudogének

- Hibás génduplikációk eredményei, nem kódoló DNS-ben.
- Processzált és nem processzált pszeudogének.
- cDNS reintegráció, hiányos szakaszok (intronok, szabályozó elemek), dead-on-arrival, poly(A), retrotranszpozonok.
- DNS tandem duplikáció, gyakran dead-on-arrival.
- kb. 15.000 /genom, 0.5 /gén, géntípusok közti különbségek
- riboszóma protein kódoló gének: 26 /gén, az összes 13 %-a,



Riboszóma protein pszeudogének koreloszlása

- Humán, csimpánz és egér genomban is
- Szubsztitúciós ráta: 1.25×10^{-9} silent sites
- $(1+0.25)\times(1.25\times10^{-9}) = 1.56 \times 10^{-9}$ /év
- 9 % ~ 50 MYA: „genomikai felfordulás"

The Origins of Genome Architecture, Lynch, 2006

Classes of interspersed repeat in the human genome