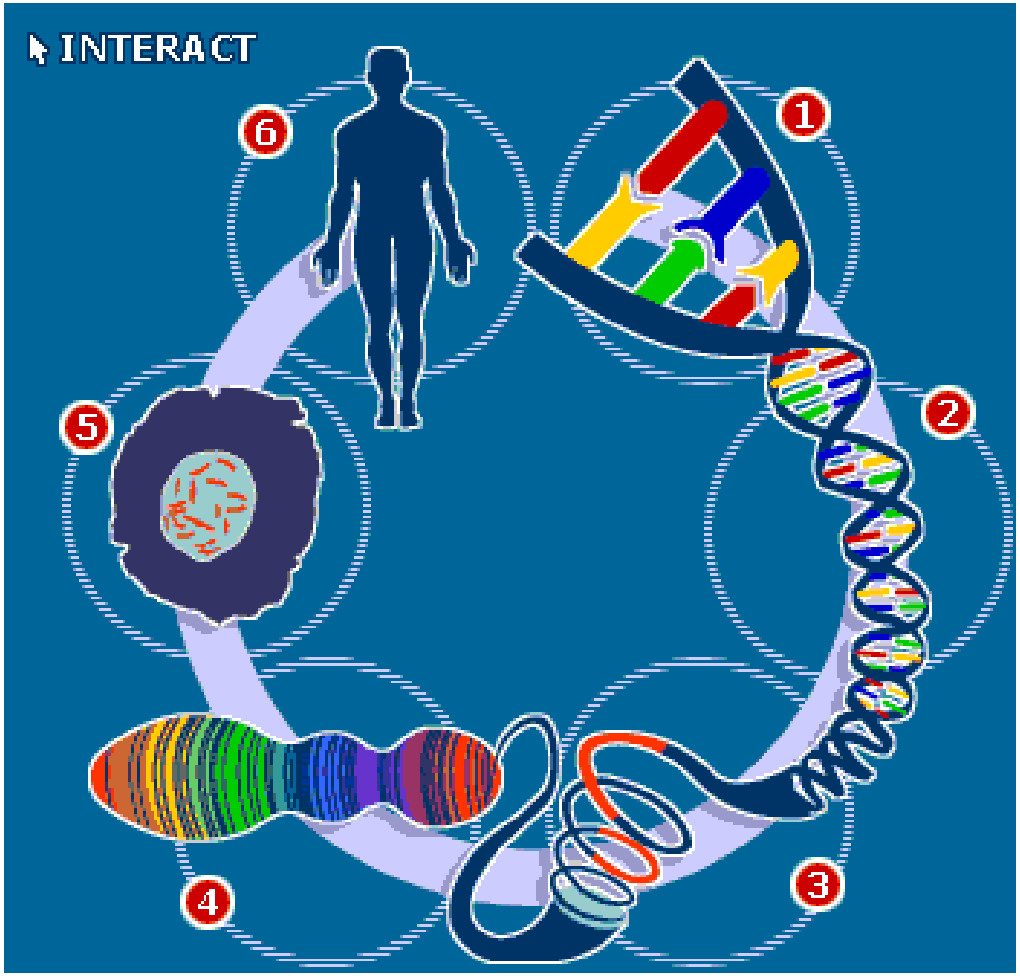
GENOMICS course

The structure of the human genome



Department of Genetics, Faculty of Science
Eötvös Loránd University

The Human Genome Project - results

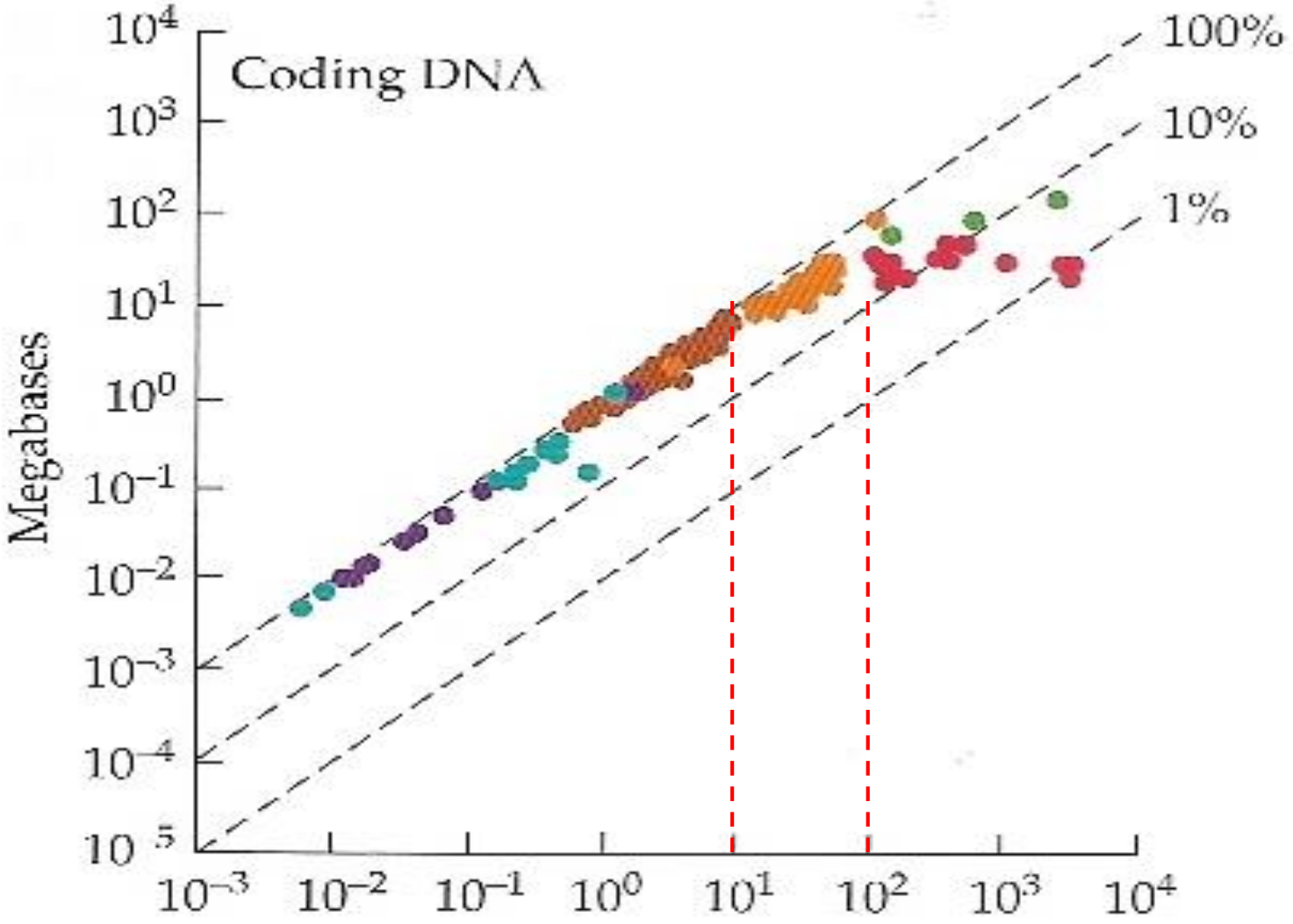
- Draft sequence published in 2001 (Science, Nature)
- Larger than any of other well-characterized (~ 2900 Mb)
- Structure and organization similare to eukaryotes

  (see model organisms' genomes)
- Emerging number of RNA genes

  (siRNA, miRNA, piRNA, lncRNA etc.)
- Correspondence between gene no., cell- and tissue types and organismal complexity?
- Surprisingly low volume of protein coding genes:

  ~ 20.000, about 1 % of DNA in genome is protein coding

Coding sequences vs. genome size



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.

Gene number

vs.

Coding sequence length

Genome size

vs.

Non-coding sequence length

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

Gene duplication, functional gene diversity

- ~ 4000 pairs of duplicated human genes

  (without multigene families)
- 5% of human genome is recent segmental duplication
- Duplication rate: 0,01/ gene/ million year; silencing: 10M

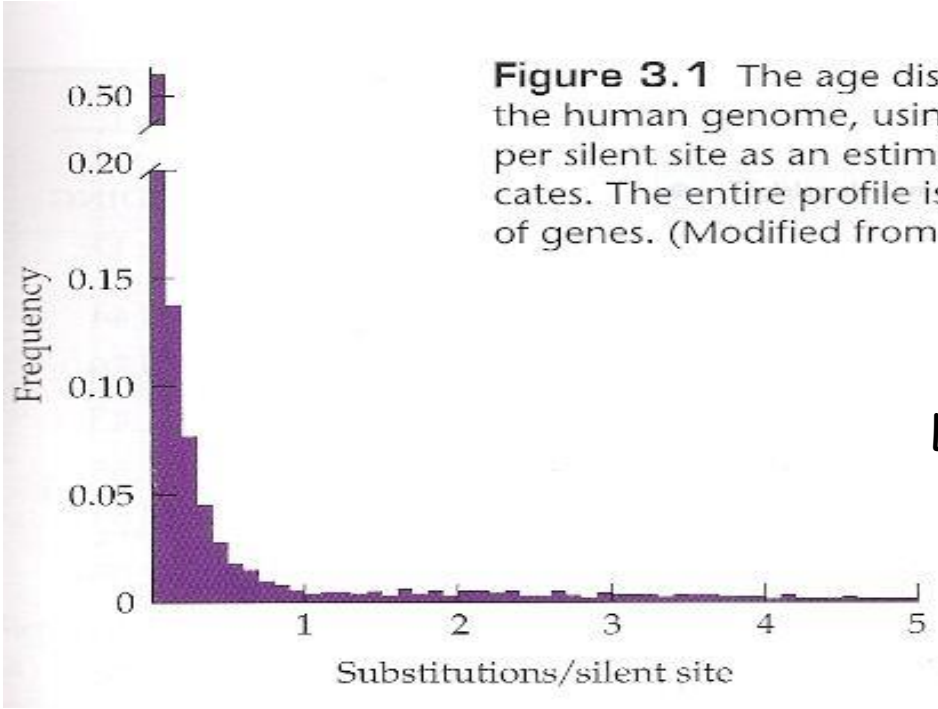


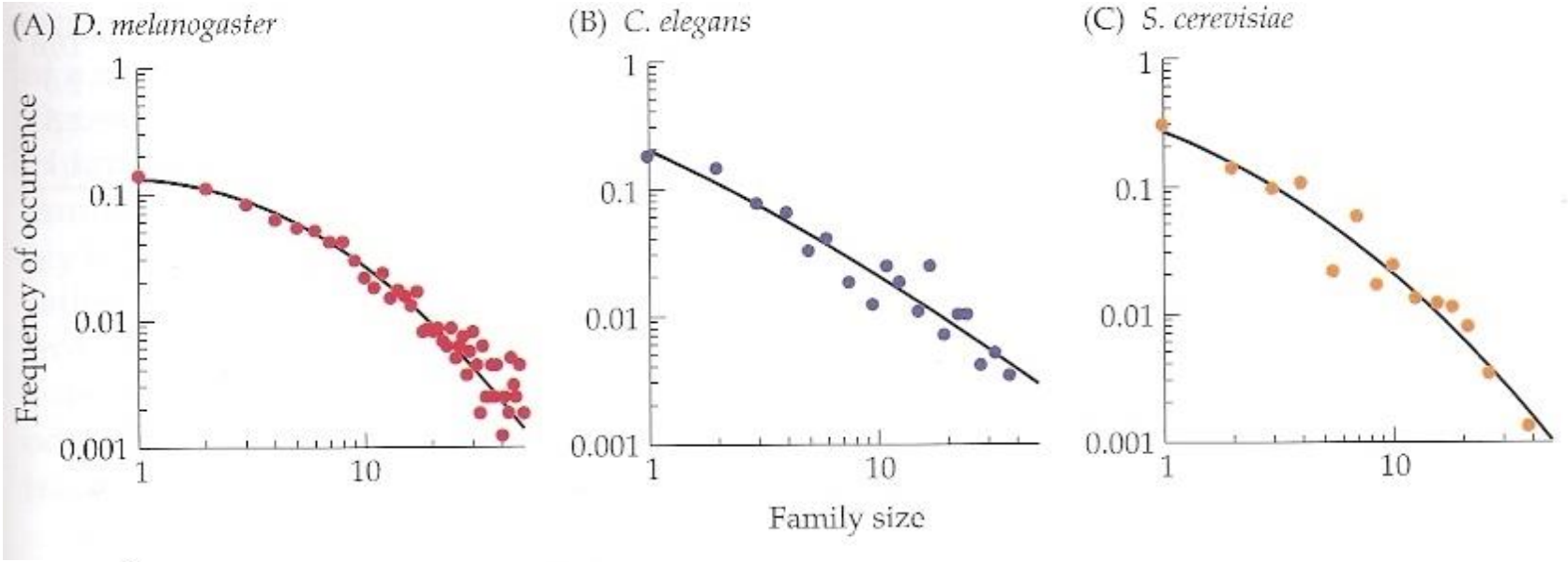
Figure 3.1 The age distribution of duplicate genes in the human genome, using the number of substitutions per silent site as an estimate of the age of a pair of duplicates. The entire profile is based on a survey of 3892 pairs of genes. (Modified from Lynch and Conery 2003a.)

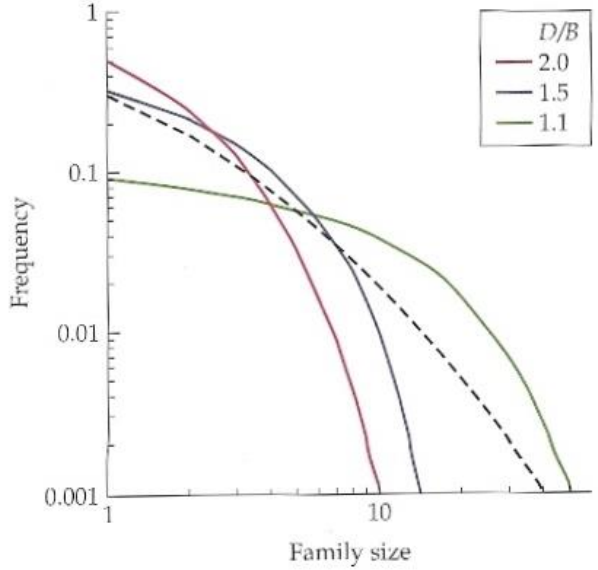
L-shaped age distribution

Gene duplications, functional gene diversity

- Duplication rate: 0,01/ gene/ million year
- Gene duplicates switch off (death) in average: 10 million years
- A significant fraction of genes is nonessential (redundant)
- Stochastic gene expansion and contraction: adaptation?
- Balanced gene expansion and contraction of gene families in mouse and human lineages: *stochastic equlilibrium birth/death process of gene families* (ie. olfactory receptor genes).
- No need for adaptation (but can be: immunity, reproduction)

Singleton and multigene lineages distribution





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

x: no. of families of size

B: probability of duplication

D: probability of deletion

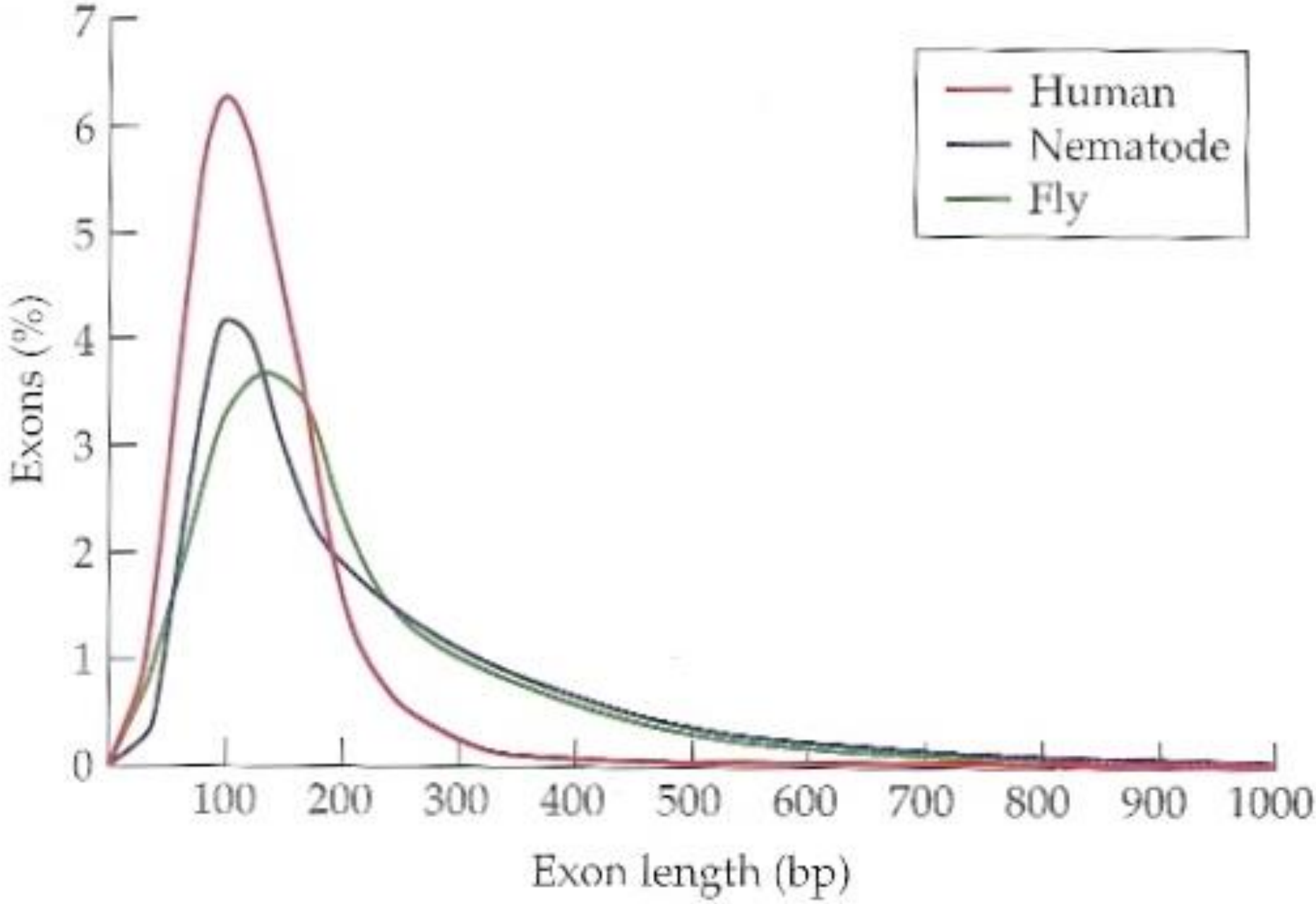
D > B; D/B: 1.5-2.0

The Origins of Genome Architecture, Lynch, 2006

Introns and Exons

- Most eukaryotes produce proteins with similar average length, but variation exist in noncoding intragenic portions of genes.
- Average human gene: 7.7 introns, 0.15 kb exon, 4.66 kb intron
- Invertebrata: less intronic sequences, average exon size usually larger than in humans

  (*Saccharomyces*: intron-free, *C. elegans*: 5.2 intron / with 120 bp
- Human genome: reduction in variance of exon size (< 300 bp)
- Splicing mechanism: „exon scanning"



The Origins of Genome Architecture, Lynch, 2006

Introns and Exons

- Diversifying gene functions without increasing gene no.
  i.e. alternative intron-exon junctions: majority of genes.
- Approx. 20 % of alternative splicing is tissue-specific.
- Functional proteins >> No. of genes.
- Alternative splicing and organismal complexity?

- *C. elegans, Drosophila:* 20% of genes, 1.3 transcript /gene.
- *Humans:* more than 50% of genes, 2.6 transcript variants /gene.
- Functional domens approx. 2 times more than in invertebrata.

- ~ 1,5 – 2 times more genes, and 50.000 additional proteins.
- Human and mouse lineage: 70% of minor splice variants are *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

Regulatory elements

- Organismal complexity: non-coding DNA /gene,
- variability: unicellular– multicellular– vertebrata– human *(table)*,
- Complex identification (ORF?), orthologous sequences?
- *transcription factor binding sites, exon-intron boundaries, transcription termination etc.*
- Conservative estimate: 2.0 kb/gene in average?
- Mouse/human: 66.000 conserved intergenic blocks (150 bp)
- *90-100% sequence identity, stringent selective constraint.*
- Gene expression: enhancer and repressor binding sites.
- Functional RNA transcripts?

Mobile Genetic Elements

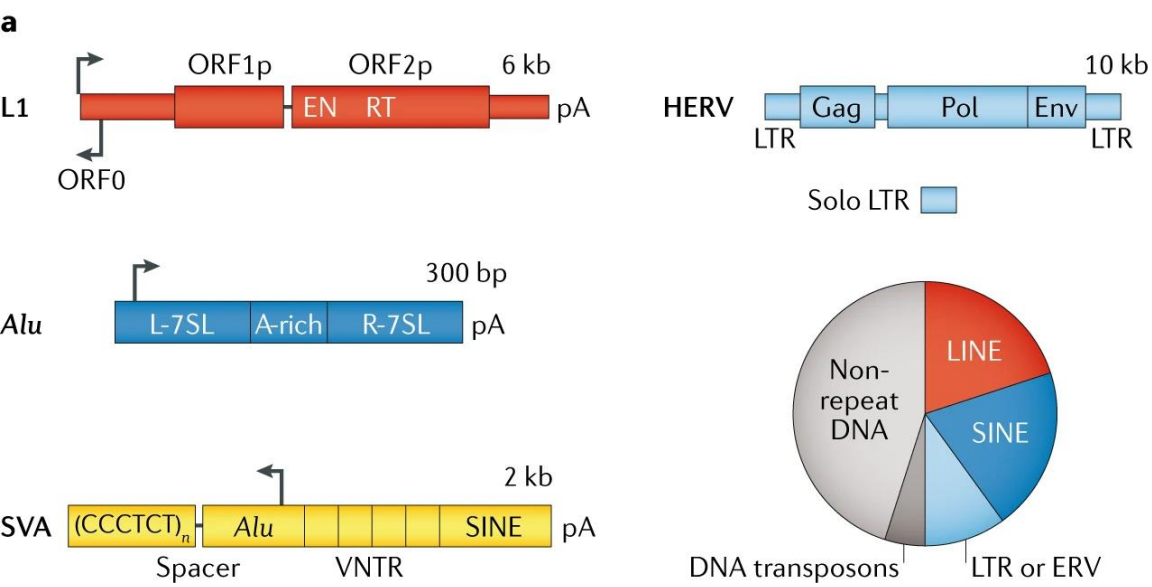
- Extra volume in human genome: 100/gene (~ half of genome)
- Human genome: ~75% is the product of past mobile element activities
- Mutagenic side effects: inzertion, non-homologous recombination, → negative consequence for the host
- Retrotransposons: „copy-and-paste", LINEs, SINEs, LTRs
- Transposons: „cut-and-paste"

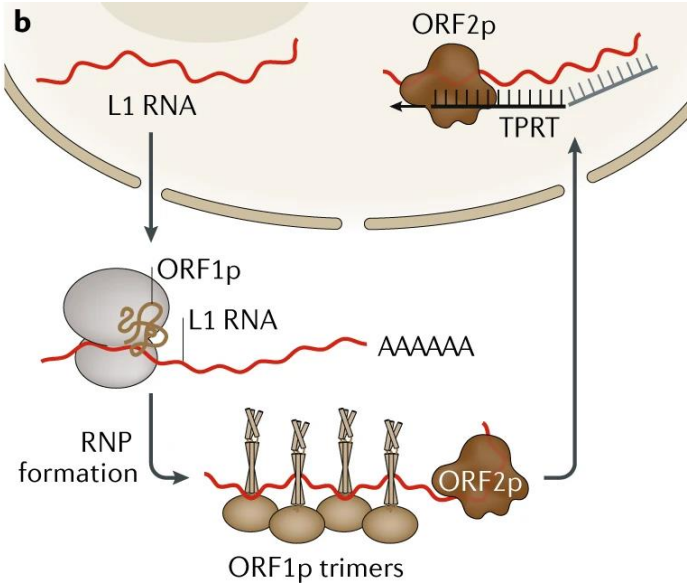
TABLE 2.2: CLASSES OF DISPERSED REPEATS IN THE HUMAN GENOME.

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).

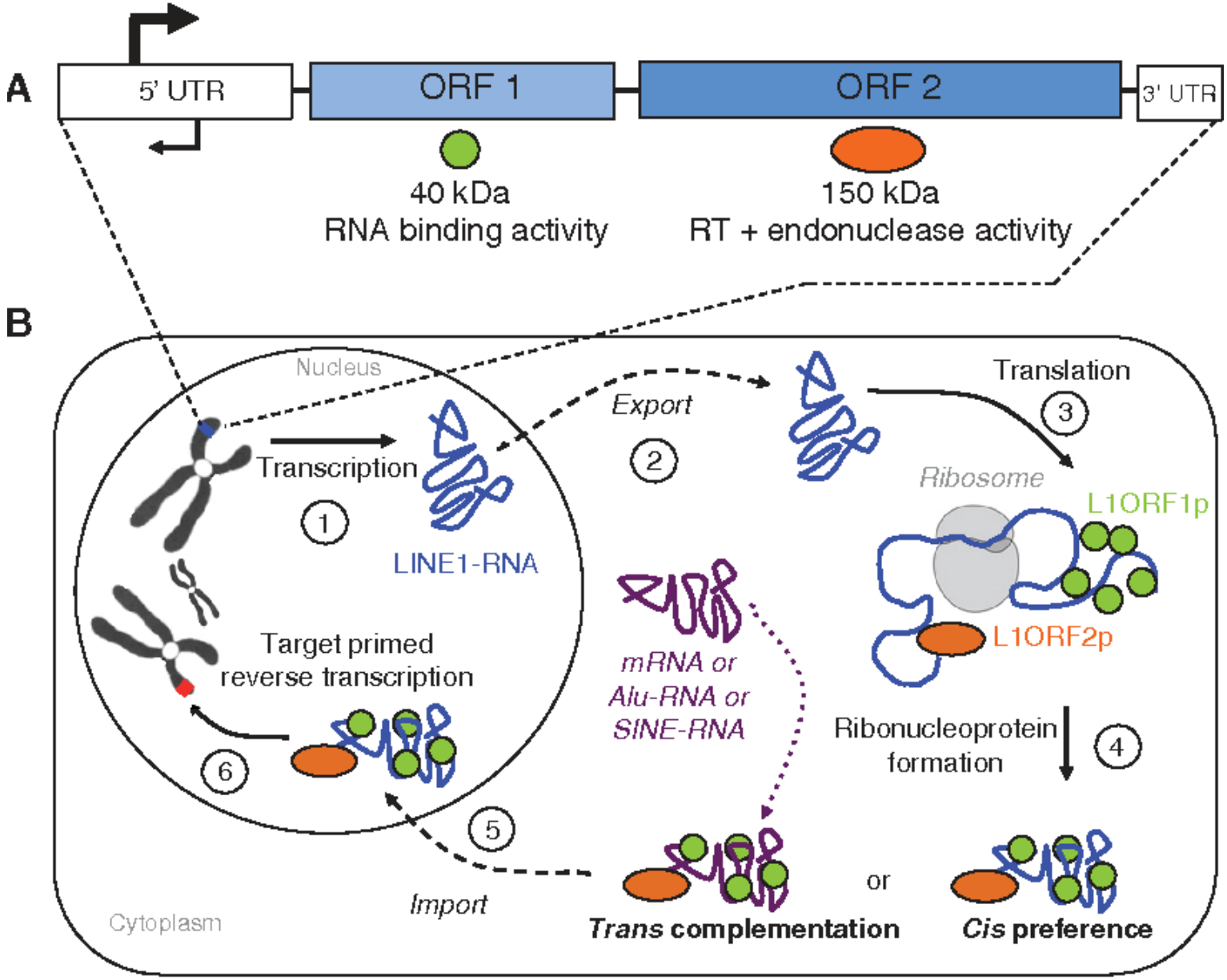
Structure of Transposable Elements





A schematic of common human transposable elements with their full- length size denoted. Long interspersed element 1 (LINE-1 or L1) encodes two open reading frames (ORFs). ORF2p protein has endonuclease (EN) and reverse transcriptase (RT) domains.
Alu elements are bipartite, with the two arms derived from 7SL RNA separated by an A- rich region.
SVA is a composite element containing variable number tandem repeats (VNTRs). Human endogenous retroviruses (HERVs) are flanked by long terminal repeats (LTRs) and encode three essential viral proteins, including envelope (Env). ERVs also exist in the genome as solo LTRs.
The pie chart shows the proportion of the human genome made up of these repetitive sequences.

Payer and Burns, 2019, Nat Rev Genet



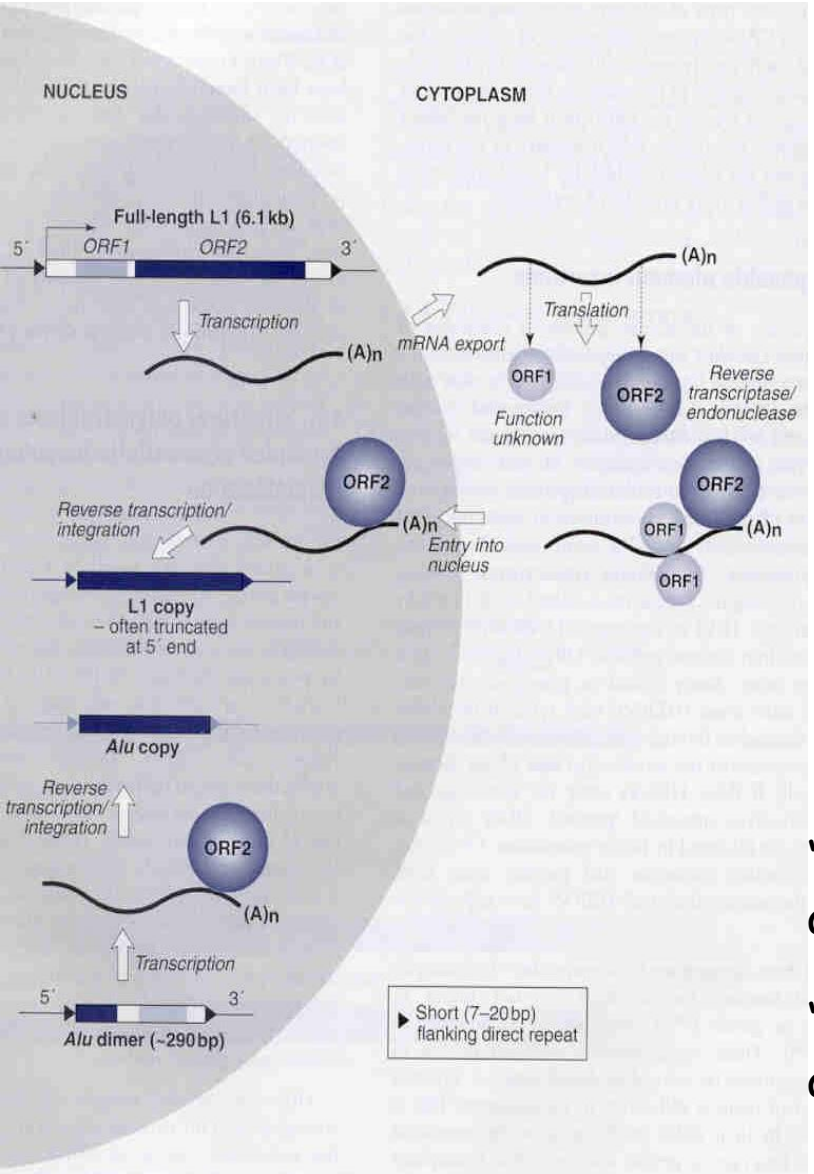
Regulation of LINE-1 in mammals
Maxime Bodak / Jian Yu / Constance Ciaudo
Published Online: 2014-09-25 | DOI: https://doi.org/10.1515/bmc-2014-0018

Long Interspersed Nuclear Elements - LINEs

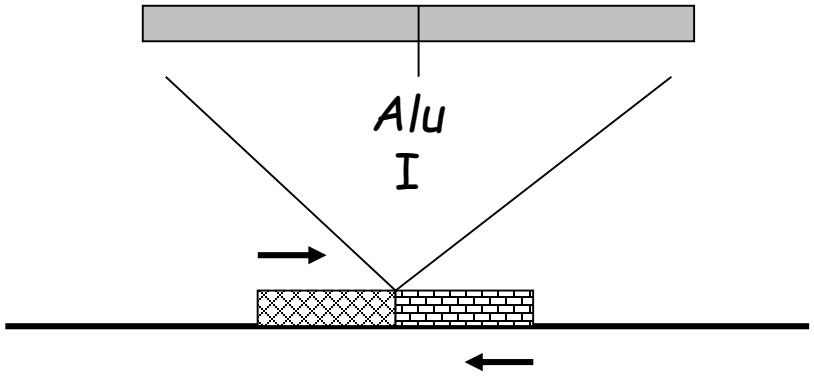
- **LINEs v. *Kpn*:** approx. 20% and 870.000 copies
- around 100 LINE sequences active as retrotransposons
- ~ 6.0 kb internal 5' promoter, 2 ORFs (RNA-binding protein, endonuclease + revers transcriptase), poly(A)-tail,
- Target-primed revers transcription: TT | *AAAA* – target
- Sloppy process of copying

  (transcription „read-through", truncated insertion „dead-on-arrival", local rearrangements, other defective LINEs can be mobilized)
- LINEs are incapable of cleaving themselves from host DNA
- Ancient relics an relative new sequences

Mobile elements: biallelic length-polimorphism



Human *Alu* Repeat
(~300 bp)

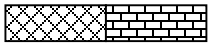


Two types of alleles

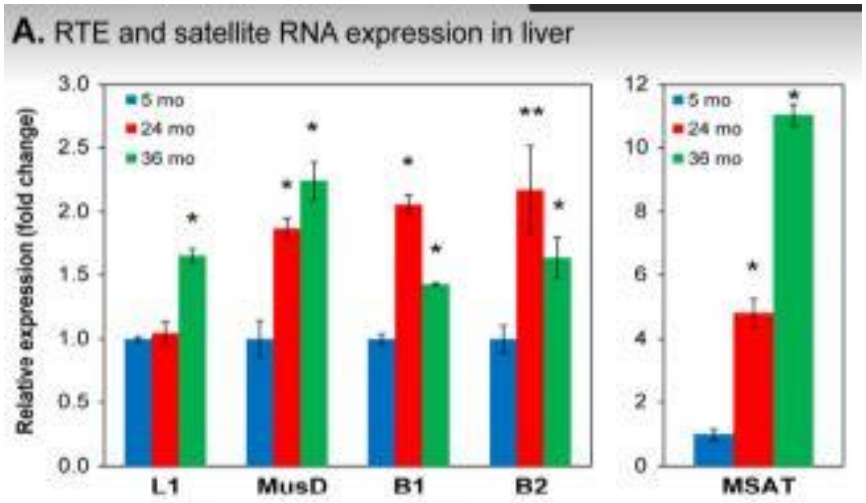
"long" (+) allel

400 bp

"short" (-) allel



100 bp



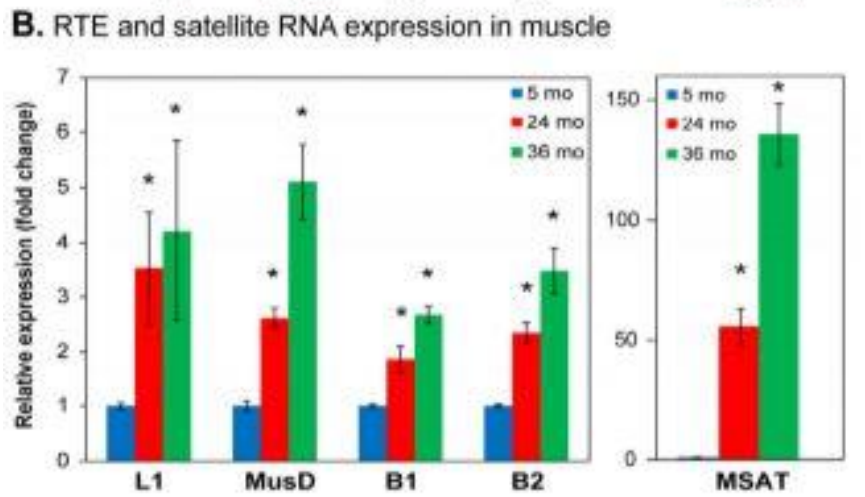


Figure 4. qPCR analysis of RNA expression of representative RTEs and SEs. Total RNA was extracted from (**A**) liver and (**B**) skeletal muscle, quantified by qPCR using indicated primers (Table S1) and normalized to GAPDH. Data were additionally normalized to the 5 month value for each element (shown as 1.0). L1, LINE L1; MusD, LTR RTE MusD/ETn; B1, SINE B1; B2, SINE B2; MSAT, major (also known as γ) SE. (*) $p<0.01$; (**) $p\leq0.05$.

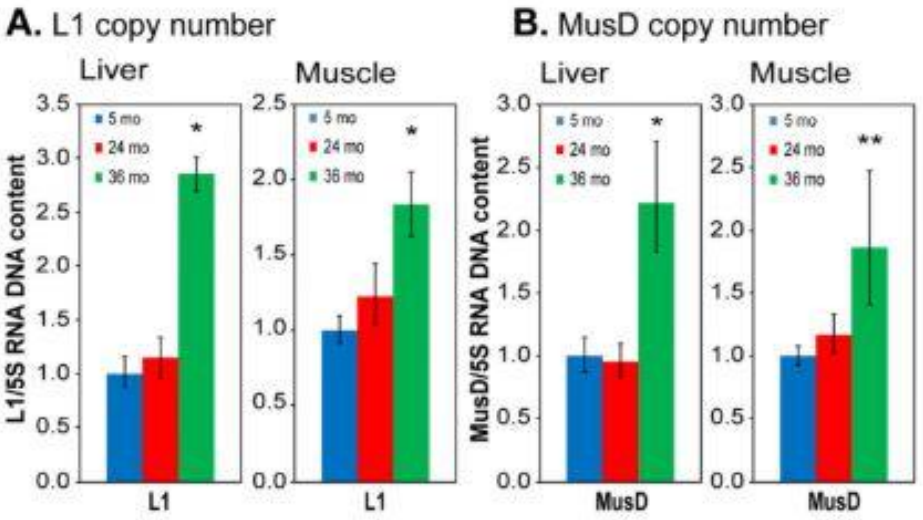
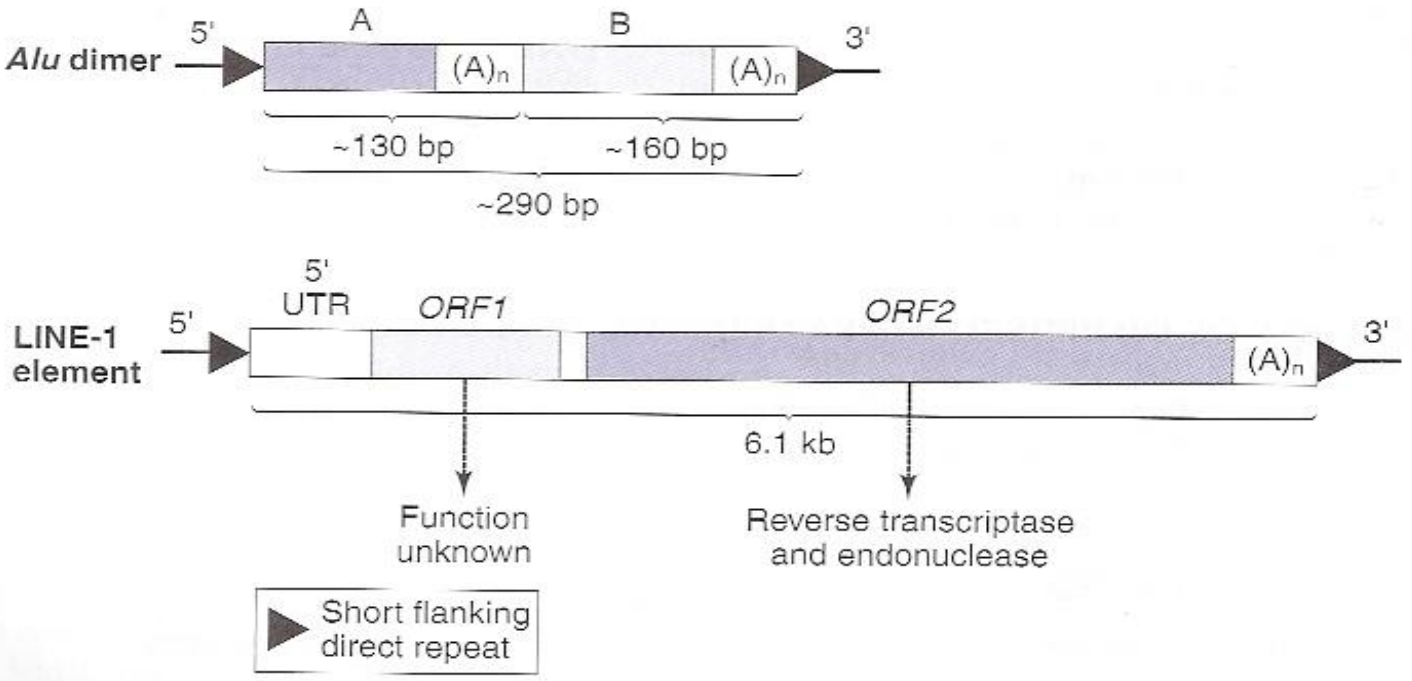


Figure 6. qPCR analysis of DNA to assess RTE genome copy number. (**A**) L1; (**B**) MusD. Total DNA was extracted from tissues of the same animals and tissues as used in Figure 4. Relative copy numbers were quantified using a multiplex TaqMan qPCR assay with the indicated primers (Table S1) and normalized to 5S ribosomal DNA. Data were additionally normalized to the 5 month value for each element (shown as 1.0). 5S DNA copy number was independently verified not to vary with age or between animals or tissues using qPCR against known single copy sequences. Means and standard deviations are shown. (*) $p<0.01$; (**) $p\leq0.05$.

Aging (Albany NY). 2013 Dec;5(12):867-83.
Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues.
De Cecco M[1], Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA

Short Interspersed Nuclear Elements - SINEs

- **SINEs / *Alu*:** 1.500.000 copies, 70 % *AluI*, 300 bp,
- Primate-specific, *Alu I* : AGCT, polimorphisms,
- Noncoding sequence, no self-mobilization
- *Alu* – LINE-1 retrotransposition, 0.05 /genome / generation



Human Evolutionary Genetics, Jobling et al, 2004

LTRs and Transposons

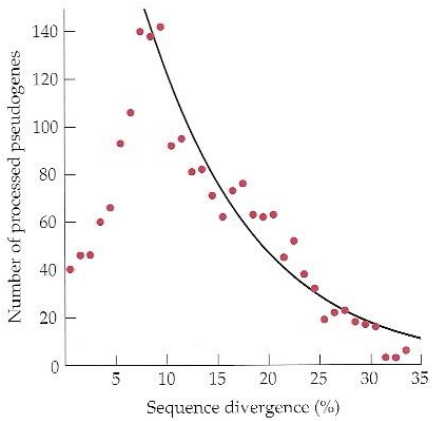
- **LTRs:** Long Terminal Repeats, Retroviral origin
- HERVs: Human Endogenous Retroviruses
- Revers transcription and integration: its own primer site
- Identical sequences: dsDNA → nucleus, mutation, divergence

(substitution rate: 1.25×10^{-9}, neutral sites, chimpanzee-human)

- *env* gene: movement among cells, not only vertical transfer
- **Transposons:** „cut-and-paste", multi families
- TIRs (terminal inverted repeats, 10-500 bp), small duplication
- transposase enzyme – TIR binding – excizition / insertion
- DNA repair, homologous chromosomes, multiplication

Pseudogenes

- Failed gene duplication events, in noncoding DNA.
- Processed and nonprocessed pseudogenes.
- cDNA reintegration, missing sequences (introns, regulatory elements), dead-on-arrival, poly(A), retrotransposons.
- DNA tandem duplication, usually dead-on-arrival.
- approx. 15.000 /genome, 0.5 /gene (differences between gene types: ribosome protein coding genes: 26 /gene)



Age distribution of ribosome protein pseudogenes

- Human, chimp and mouse genomes too
- Substitution rate: 1.25×10^{-9} / silent sites
- $(1+0.25)\times(1.25\times10^{-9}) = 1.56 \times 10^{-9}$ /year
- 9 % ~ 50 MYA: „genomic uphealing"

The Origins of Genome Architecture, Lynch, 2006

Content of known and proposed functional noncoding DNA sequences in the human genome

DNA elements	Size, kb	Totally in the genome*		Functional elements and/or functions
		nucleotides, Mb	share, %	
Mobile genetic elements	<1-25	1395	45	tissue-specific regulation of protein-encoding gene transcription; epigenome maintenance and establishment of borders between functional domains of chromosomes
Introns	<0.1-1000	744	24	5-fold increase in the information capacity of the genome through alternative splicing, including intergenic splicing; IME; recombination of allele genes. Introns can contain transcription promoters, terminators, enhancers, and silencers
Conserved sequences evolving slowly		130	4.2	exons (30%), introns (30%), and intergenic sequences (40%), including DNase hypersensitivity sites, transcription factor binding sites, promoters, UTRs, enhancers, insulators, and lncRNAs
rapidly		254	8.2	
Centromeric satDNA	250-5000	155	5	site of kinetochore assembly; involvement of satDNA transcripts in chromatin heterochromatization and regulation of development
Enhancers	<1-50	93	3	assembly of protein complexes, which activate or inhibit transcription, including tissue-specific transcription
CpG islands and ICR	0.2-2	31	1	regulation of gene transcription through methylation/demethylation of CpG and adjacent sequences in the process of imprinting as well
5′-UTR	0.02-3 (0.21**)	4	<0.1	regulation of translation
3′-UTR	1.3**		<0.1	regulation of gene expression at posttranscriptional and translational levels
Telomeric tDNA	10-15	0.23-0.35	<0.1	maintenance of chromosome integrity and regulation of cell division number
Pseudogenes	0.83**	11.9	9	regulation of protein-encoding gene transcription (their RNAs can act as traps for miRNAs or sources for siRNAs)
Insulators	1**	<0.1	<0.1	prevention of nonspecific effects of enhancers on promoters; separation of functional domains of chromosomes; regulation of V(D)J recombination in immunoglobulin loci
S/MAR	5	<0.1	<0.1	organization of functional domains of chromosomes in interphase nuclei
Promoters		<0.1	<0.1	regulation of transcription
Noncoding RNA genes		<0.1-0.23	>90 ?	regulation of gene expression at all levels

Patrushev and Kovalenko, 2014