

# GENETIKA 1 EA

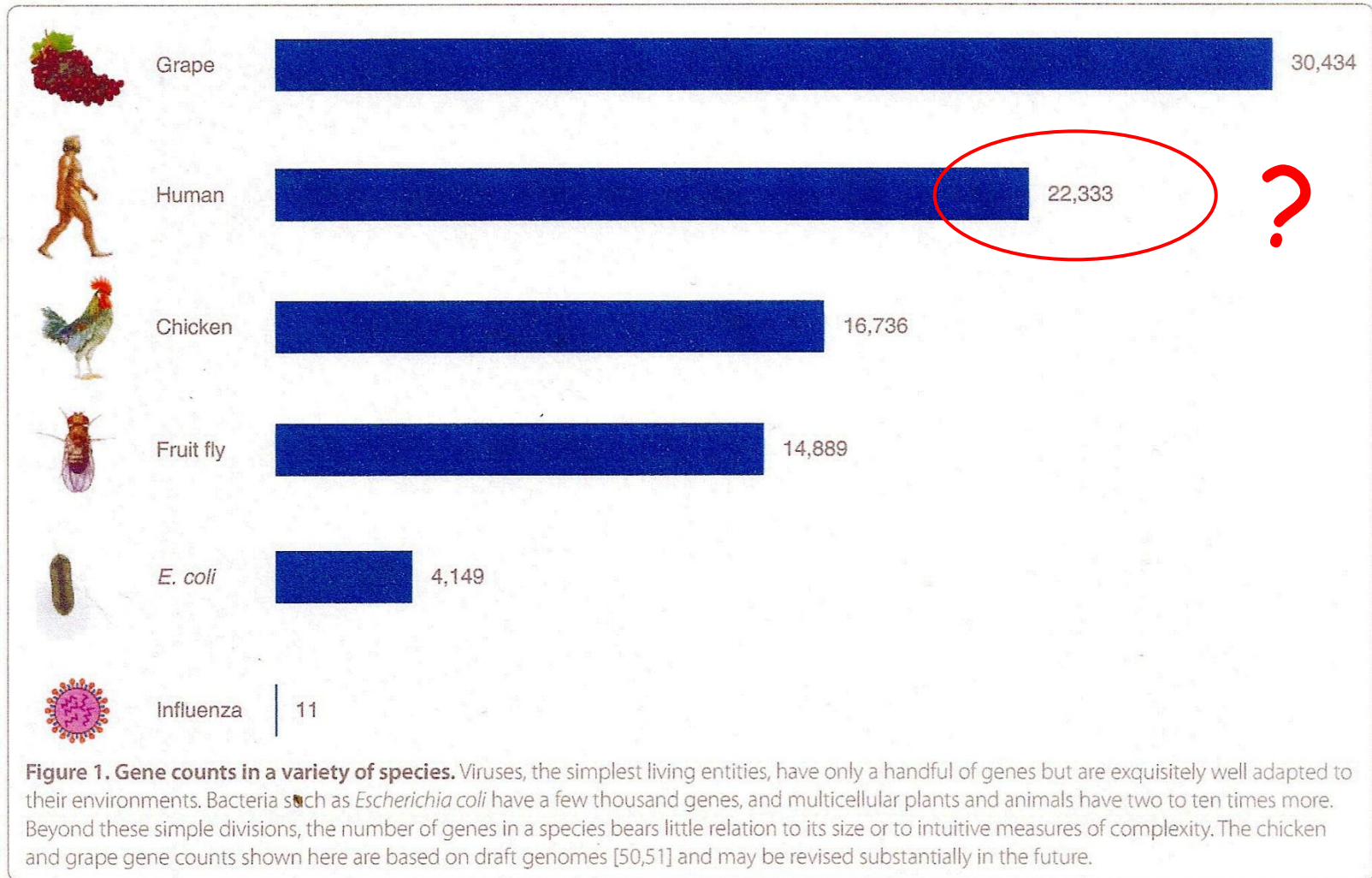
## Genomikai alapok: genom szerveződés és variabilitás

Genomtartalom, génszám, szerveződési  
komplexitás, genetikai variabilitás

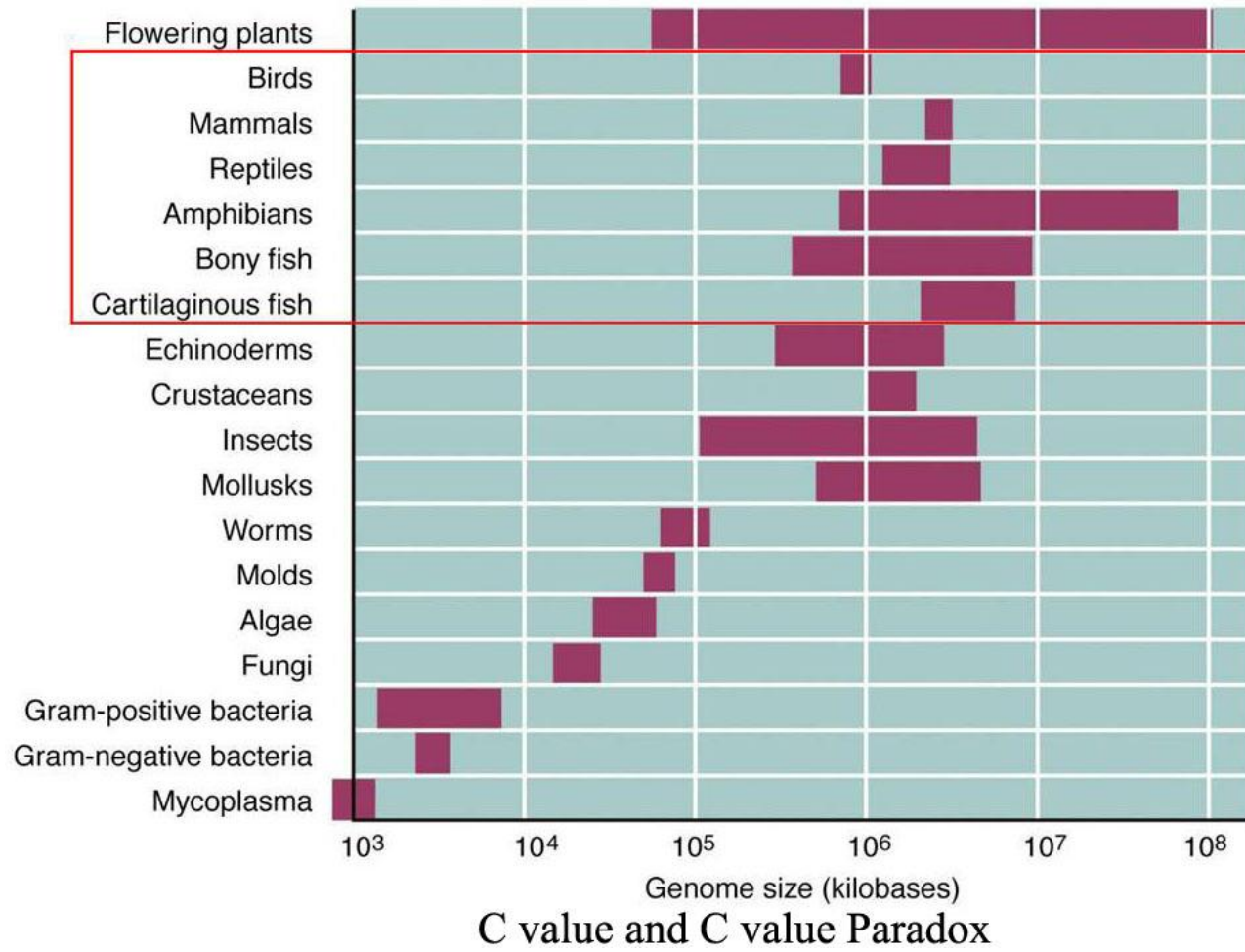
ELTE TTK Genetikai Tanszék

2024.10.22.

# Humán genom: „... valahol a csirke és a szőlő között?”



# C value paradox: C-érték paradoxon vagy rejtély?



## C value: egy haploid genom teljes DNS-tartalma

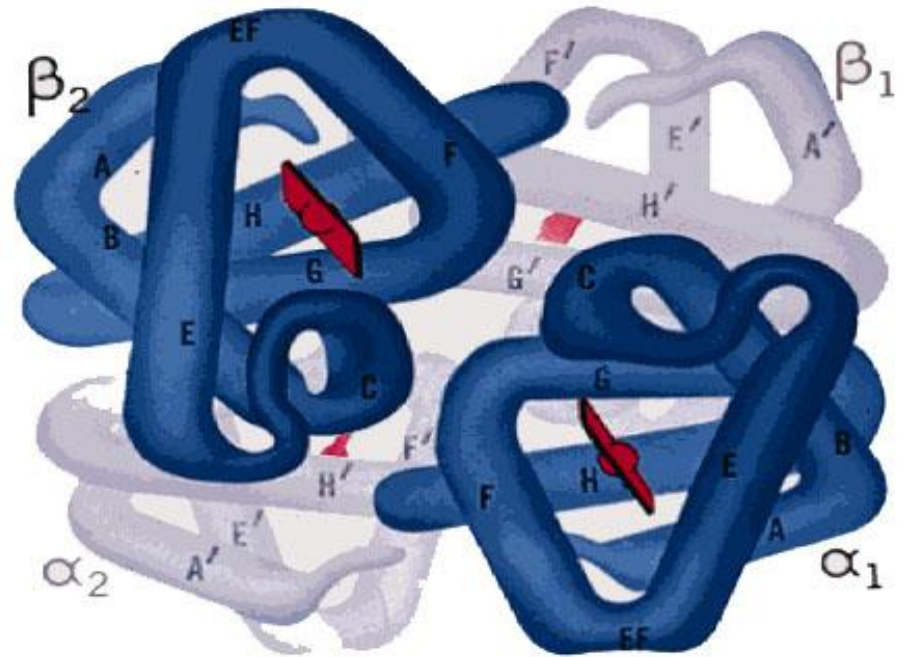
**C-value:** the „constant” (or „characteristic”) value of haploid DNA content per nucleus - measured in picogram ( $1 \text{ pg} \approx 1 \text{ Gb}$ ).

- $1C$ : fajra jellemző érték - konstans a faj minden organizmusára.
- Extrém módon variábilis lehet akár rokon fajok között is.
  - ( $<10^6$  bp a mycoplazmában és  $>10^{11}$  bp egyes növényekben)
- Egyszerűbb eukariótákban összefüggés van az organizmus komplexitása és a genom mérete között.
- Magasabb komplexitású eukarióták között nem állapítható meg összefüggés a szervezeti komplexitás és genom méret között.
  - (Mollusks, Insects, Vertebrata, Plants)

# Első becslések a humán genom méretéről és a gének számáról

1964: F. Vogel (Heidelberg)

- Hemoglobin  $\alpha$  és  $\beta$  lánc
- leegyszerűsített feltevések
- Haploid genom:  $3 \times 10^9$  bp
- Gének száma: 6,7 millió ???



1990: NIH/DOE report on Human Genome Project

- becslés: 100 000 gén, az átlagos gén méret (30 000 bp) alapján

# A génfogalom fejlődése: mit nevezünk ma egy génnek?

A Gén fogalmának jelentős átalakulása az elmúlt száz esztendőben:  
protein/RNS kódolás, intron/exon fogalom, szabályozó funkciók, stb.

## Gén

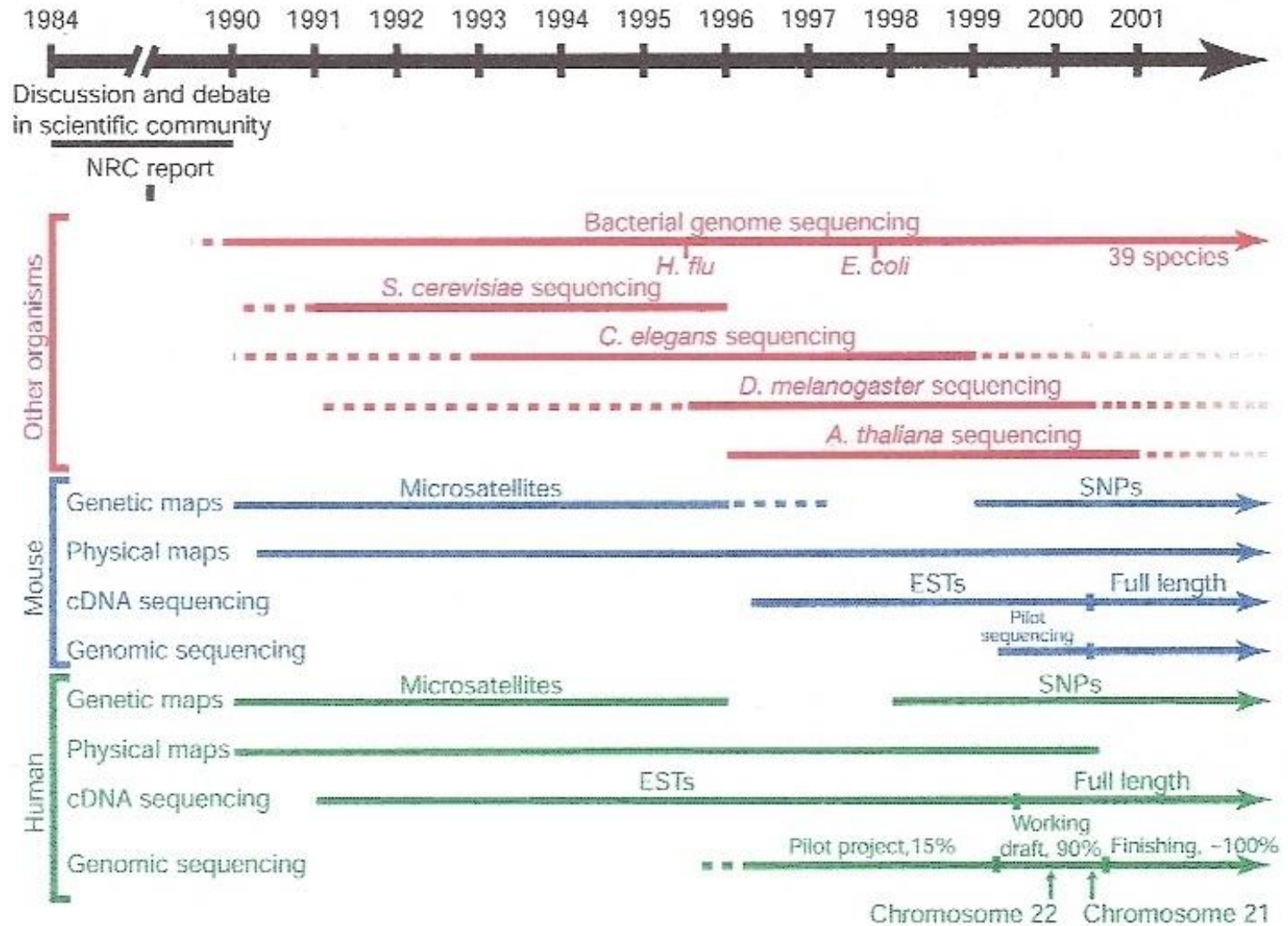
- 1950-es évektől
- a kódolási szabályok felismerése
- a DNS azon szakasza, amely egy transzkriptum (mRNS, rRNS, snRNS, tRNS, ...) átírásáért felelős genetikai információt tartalmazza

## ORF

- *open reading frame* (nyitott leolvasási keret)
- 1990-es évektől (genomika korszaka)
- gének annotálása: bioinformatikai módszerekkel prediktálnak transzkriptumok átírását végző DNS szakaszokat (konzervált szekvencia motívumok alapján)

Egy gén a genetikai állományunk jól körülhatárolt szakasza, mely mRNS-ként átíródik és egy v. több fehérjét kódol. (pl. alternatív splicing: izoformák)

# Korai genom projektek időskálán



# Humán Genom Projektek

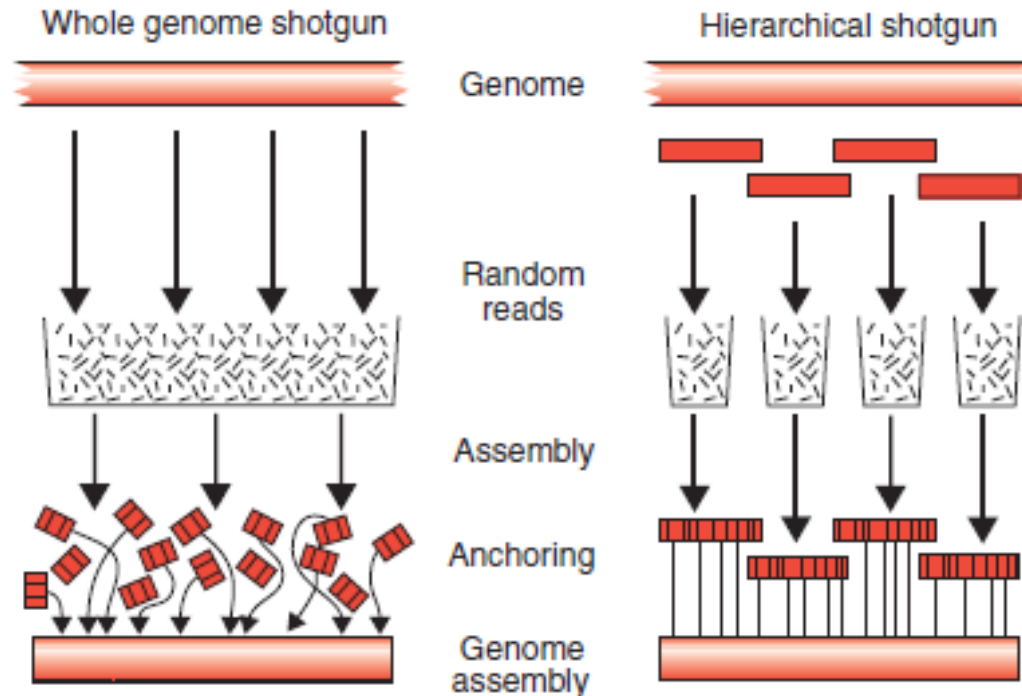
- **HUGO:** Human Genome Organization
  - US DOE és NIH, UK MRC és WTSI, CEPH , FMDA, Japán, Európai Közösség (élesztő genom), Németország, Kína
  - 1990-1995: genetikai és fizikai térképezés
    - betegség gének, fizikai pontok fixálása, modell szervezetek
  - large-scale sequencing: két fázisú
    - „hierarchikus „shotgun” szekvenálás”
- **Celera Genomics:**
  - Applied Biosystems., TIGR (C. Venter)
  - 1998-2001: „whole genome shotgun”
  - „DNS szekvencia gyárak”



Technology speeds science. ABI sequencers at Venter Insitute, 2007.



# „shotgun” genom szekvenálási stratégiák

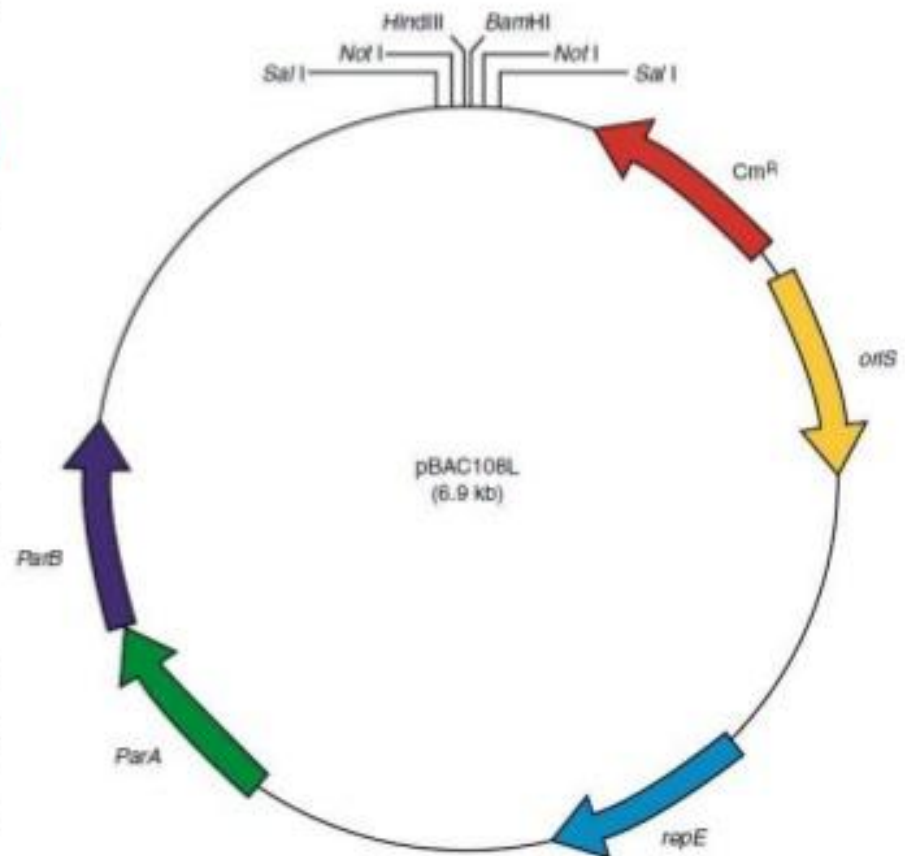


**Figure 9.11.** Assembling genomic data using the hierarchical and whole genome shotgun approaches. Adapted from Waterston, Lander and Sulston (2002), with permission

## Common gene components

Bacterial artificial chromosome is another cloning vector system in *E.coli* (**pBAC108L**), developed by Melsimon and his colleagues in 1992, have

- ❑ **HindIII and BamHI**: the cloning sites
- ❑ **CmR**: the **chloramphenicol resistance gene**, used as a selection tool.
- ❑ **oriS**: the origin of replication
- ❑ **repE**: for plasmid replication and regulation of copy number.
- ❑ **ParA and ParB**: the genes governing partition of plasmids to daughter cells during division and ensures stable maintenance of the BAC.



**Fig: Map of the BAC vector, pBAC108L**

# Universal Landmark

## Sequence Tagged Site (STS) 1989

Replaces cloned DNA probe mapping landmarks with PCR assays.

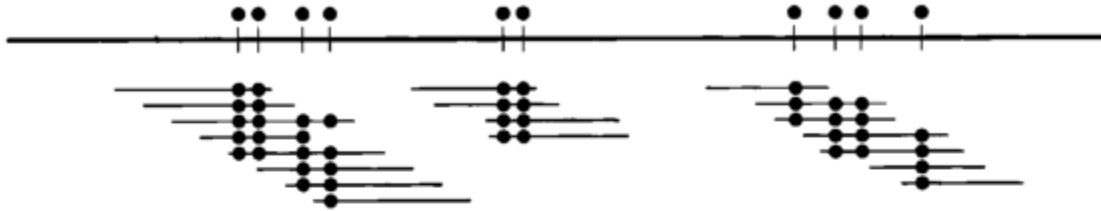
Each STS is uniquely described by a pair of oligonucleotides, a product size, and PCR reaction conditions. Can be stored and distributed electronically.

Enables merging of mapping data obtained from many labs using many different methods into a single consensus map of landmarks along a chromosome.

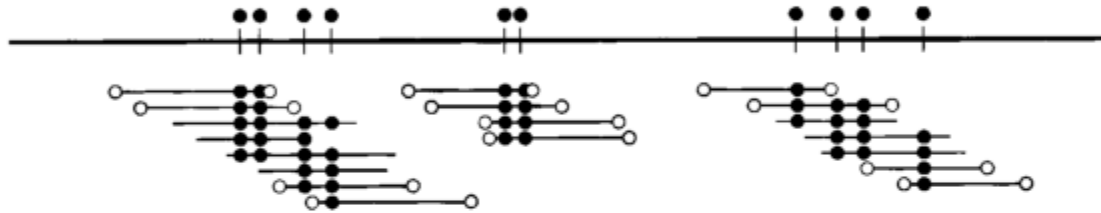
Eliminates the need for huge collections of cloned probe segments upon which prior maps depended.

# Clone ends – Clone-based Physical Map

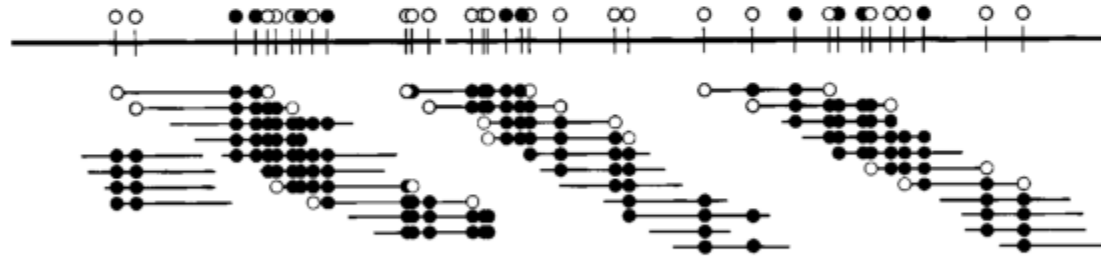
a. Screen library with existing markers



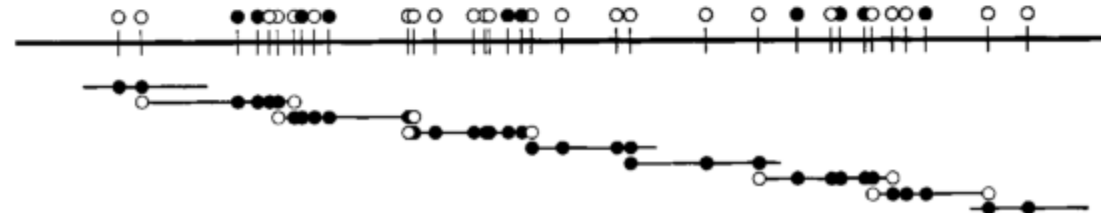
b. Generate new markers



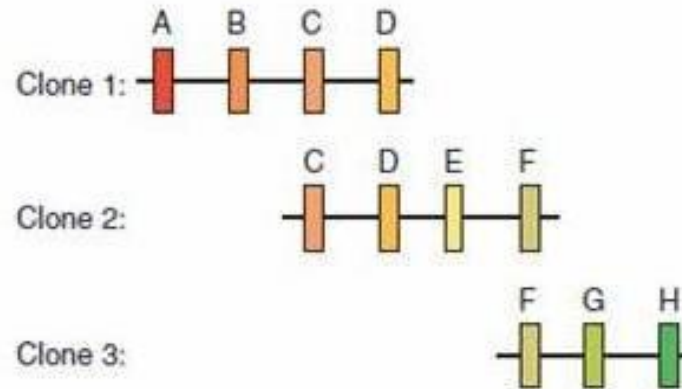
c. Screen library with new markers



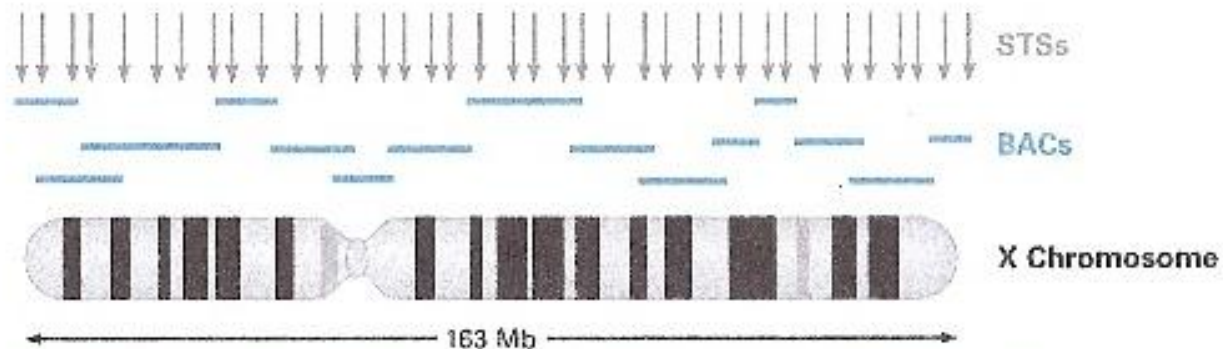
d. Determine tiling path



# Genomok fizikai térképezése

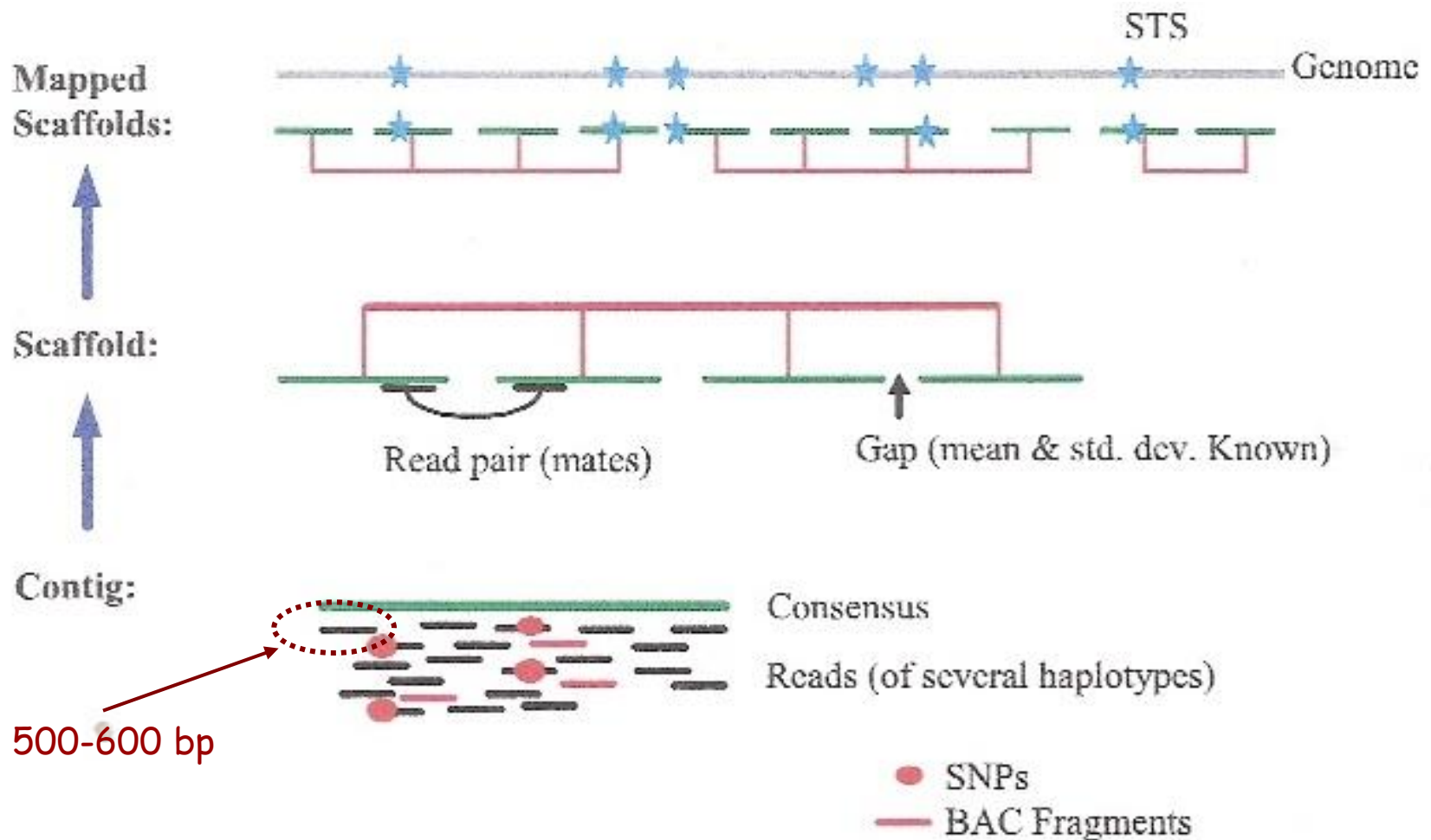


**Figure 9.5.** Aligning clones by STS mapping. Each clone contains several STSs. Clone 1 has four (A, B, C and D). Clone 2 also contains STSs C and D. Therefore clones 1 and 2 overlap with each other



**FIGURE 1.3** • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout the chromosome and on overlapping BACs.

# Teljes genom összeszerelés



**Fig. 3.** Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

# Humán Genom Projekt

# Science

16 February 2001

Vol. 291 No. 5507  
Pages 1145-1434 \$9

## THE HUMAN GENOME



 AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

15 February 2001

# nature

£5.45 €6.23 ¥54.00 US \$10.00

[www.nature.com](http://www.nature.com)

## the human genome

**Nuclear fission**  
Five-dimensional  
energy landscapes

**Seafloor spreading**  
The view from under  
the Arctic ice

**Career prospects**  
Sequence creates new  
opportunities

**naturejobs**  
genomics special

# Hol tartunk most?

**2001,** Human Genome Consortium: 30 000 - 40 000 protein kódoló gén

Celera Consortium: 26 500 „erős” + 12 000 „gyenge” bizonyíték

**2004,** Human Genome Consortium: 20 000 - 25 000 gén

- kevesebb mint az Arabidopsis → szervezeti komplexitás?

**2010,** Ensembl: 22 619 / NCBI: 22 333 protein kódoló gén

CCDS: 18 173 (<http://www.ncbi.nlm.nih.gov/CCDS/CcidsBrowse.cgi>)

fals pozitívak: retrotranszpozonok, pszeudogének, „orphan” DNS

**2022.09.12.:** CCDS GeneID: 19 030 genes > 1 CCDS ID: 7 869



# Annotált RNS gének növekvő száma

Comprehensive Human Expressed SequenceS; <http://ccb.jhu.edu/chess>

... human gene catalogs now contain more RNA genes than protein-coding genes (Salzberg, 2018)

**Table 1** Gene annotations in Gencode, Ensembl, RefSeq, and CHES

	Gencode <sup>a</sup>	Ensembl <sup>b</sup>	RefSeq <sup>c</sup>	CHES <sup>d</sup>
Protein-coding genes	19,901	20,376	20,345	21,306
lncRNA genes	15,779	14,720	17,712	18,484
Antisense RNA	5501		28	2694
Miscellaneous RNA	2213	2222	13,899	4347
Pseudogenes	14,723	1740	15,952	
Total transcripts	203,835	203,903	154,484	323,827

# Új gének és gén átrendeződések

- CGH analízisek: rokon fajok között kb. azonos génszám
- *de novo* gén keletkezés: génduplikáció és specializáció
- génszám eltérések egyének között: segmental duplications
- **large-scale copy number variations (CNVs)**
- emberi „pángenom”: változatok rasszok, csoportok között.

(Izd. Li R, et al., 2010, Nat Biotechnol, 28:57-63)

- kb. 40 Mb új szekvencia, + 1,3 %
- ***de novo* eredet: új humán gének?** (Knowles and McLysaght, 2009)

# Új gének keletkezése

**Table 1.** Novel human protein-coding genes and supporting evidence.

Gene name	Ensembl ID	Length (codons)	Longest chimp ORF <sup>a</sup>	Expression support and tissue <sup>b</sup>	Primate shared disablers <sup>c</sup>	Other major sequence differences	Presence of enabler in other human complete genome sequences <sup>d</sup>	HapMap SNPs
<i>CLLU1</i>	ENSG00000205056	121	42	EST/cDNA: Blood ( <u>AJ845165</u> , <u>AJ845166</u> ); UniGene: Blood, embryonic tissue, eye, lymph, lymph node, muscle, pharynx, tonsil (Hs.339918)	1-bp indel <sup>e</sup>	Macaque: 4- and 1-bp indels	Sequence available and enabler conserved in all	1 syn.; 1 nonsyn.
<i>C22orf45</i>	ENSG00000178803	159	87 (25 amino acids align with human sequence)	EST/cDNA: Kidney, other ( <u>AX747284</u> , <u>AK091970</u> , <u>DA635985</u> ); ArrayExpress: Sperm, lung (E-GEOD-6872, E-GEOD-3020)	Premature stop codon	Chimp: 1-bp indel; Macaque: lacks ATG start codon; 4-bp indel	Reverse strand is available and conserved in Venter	1 nonsyn.
<i>DNAH10OS</i>	ENSG00000204626	163	90 (75 amino acids align with human sequence)	EST/cDNA: Hippocampus ( <u>AK127211</u> ); UniGene: Blood, embryonic tissue, eye, lymph, lymph node, muscle, pharynx, tonsil (Hs.339918)	10-bp indel	Chimp: 2- and 1-bp indels; Macaque: lacks ATG start codon; 13-, 8-, 1-, and 1-bp indels	Reverse strand is available and conserved in Venter, Watson and HuAA	1 syn.; 1 nonsyn.

<sup>a</sup>Length in codons of longest in-frame (alignable) ORF starting from any ATG in the region.

<sup>b</sup>Type of data/database is listed followed by tissue information with database identifiers in parentheses. Underlined accession numbers are full-length, spliced cDNA.

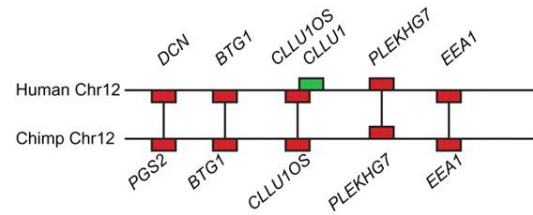
<sup>c</sup>Shared disablers are sequence differences shared by chimp, gorilla, orangutan, gibbon, and macaque that eliminate the capacity to produce a protein similar to the human protein.

<sup>d</sup>Independently sequenced whole genomes: Venter, Watson, HuAA, HuBB, HuCC, HuDD, and HuFF. All data are listed where available.

<sup>e</sup>Not shared with orangutan.

# Sequence changes in the origin of *CLLU1* from noncoding DNA. (A) Region of conserved synteny between human and chimp chromosomes 12.

A



B

Start

Human  
Chimpanzee  
Macaque

```
GTTTGGAGG - - - ATGTTCAAC AAATGCTCCTTTCA TTCTCTATTTACAGACC TGCCGCA
GTTTGGAGG - - - ATGTTCAAT AAATGCTGCTTTCA CTCTCTATTTACAGACC TGCCGCA
GTTTGGAGG - - - ATGCTCAAT AAATGCTCCTTTCA TTCTCTCATTACAAACT TGCCGCA
```

Human  
Chimpanzee  
Macaque

```
GACAATTC TGCTAGCAGCC TTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
GACAATTC TGCTAGCAGCC TTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
GACAATTC TGCTAGCAGCC TTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
```

Human  
Chimpanzee  
Macaque

```
GATCTGGAGACTAA - CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
GATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
TATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
```

Human  
Chimpanzee  
Macaque

```
CAGAATACGATTTAGCAAATTACTCTTAAAGATAT TATTTTACATTTCTATATTTCTCCTA
CAGAATACGATTTAGCAAATTACTCTTAAAGATACTATTTTACATTTCTATATTTCTCCTA
CAGAATA TGATTTAGCAAATTACTCTTAAAGATAT TATTTTGCAC TTCTATATTTCTCCTA
```

Human  
Chimpanzee  
Macaque

```
CCCTGAGTTGATGTGTGAGCAATATGTCAC TTTCATAAAGCCAGGTATACA - - - TTATG
CCCTGAGTTGATGTGTGAGCCGATGTCAC TTTCATAAAGCCAGGTATACA - - - TTATG
CCCTGAGTTGATGTGTGAGCAATATGTCAC TTTCACAAAGCCAGGTATATATACATTACG
```

Human  
Chimpanzee  
Macaque

```
GACAGGTAAGTAAAAAACATATTTATTTATTTCTACGTTTTGTCCAAAAATTTTAAATTTCT
GACAGGTAAGTAAAAAACATATTTATTTCTACGTTTTGTCCAAAGAAATTTTAAATTTCT
GACAGGTAAGTAAAAAACATATTTATTTATTTCTACGTTTTGTCCAAAGATTTTAAATTTCT
```

Human  
Chimpanzee  
Macaque

```
AACTGTTGCGCGTGTGTTGGTAA - - - TGTA AAACAAACTCAGTACA
AACTGTTGCGCGTGTGTTGGTAA - - - TGTA AAACAAACTCAGTACA
AACTGTTGCGCATGTGTTGGTAA - - - CGTA AAACAAACTCAGTACG
```

C

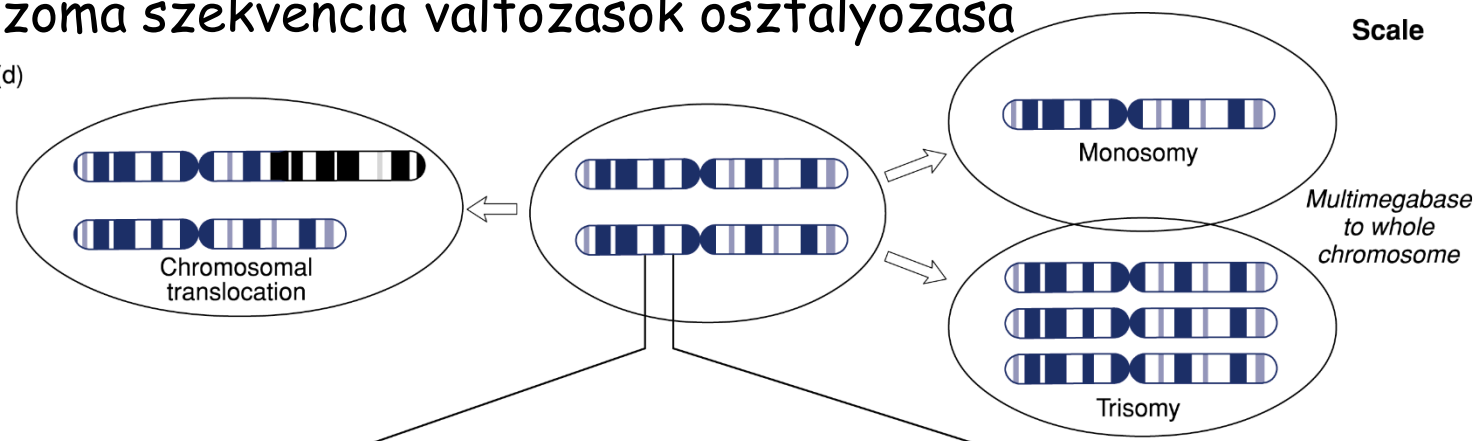


Knowles D G , McLysaght A Genome Res. 2009;19:1752-1759



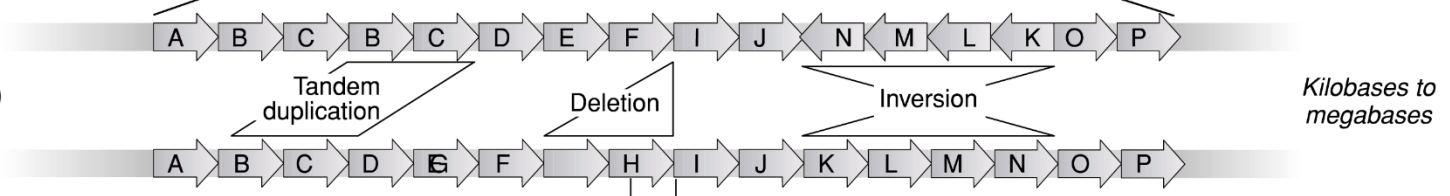
# Kromoszóma szekvencia változások osztályozása

(d)



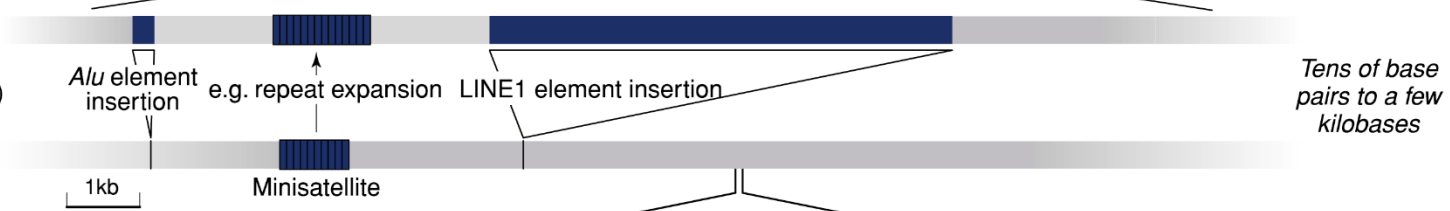
Scale

(c)



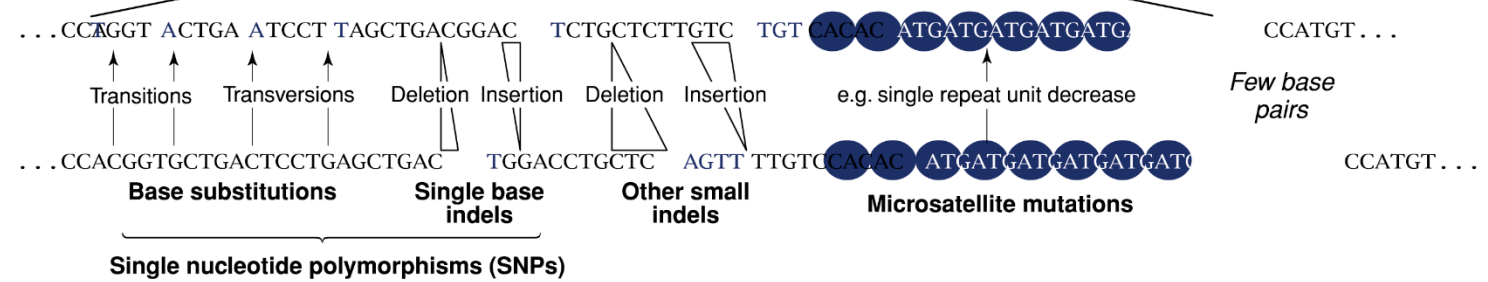
Kilobases to megabases

(b)



Tens of base pairs to a few kilobases

(a)



Few base pairs

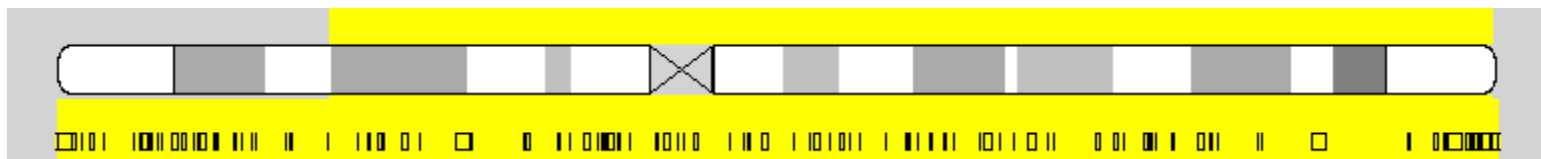
# Copy Number Variation (CNV)

## Kópia-szám variabilitás

A diploid szervezetek alapesetben minden génből két másolatot hordoznak (homológ párok). Az emberi genom vizsgálata során felismerték, hogy hosszú (ált. több Kb vagy Mb) DNS szakaszok előfordulhatnak kettőnél több példányban is. Ezeket copy-number variation (CNV)-nak nevezték el. Az egyes egyének között a CNV mintázat különböző lehet.

Kb. 300 emberen végzett vizsgálatban 1447 CNV-t mutató genomikus szakaszt azonosítottak, ez kb. a genom 12%-át fedi le.

Sikerült néhány CNV-t betegségekkel kapcsolatba hozni. Pl. a prosztataraák betegség az UGT2B17 gén kópia szám változataival hozták kapcsolatba. Vagy a HIV fertőzéssel szembeni ellenálló képesség a CCL3L1 gén több mint két példányával kapcsolatos.



Az ember 20. kromoszómáján kimutatott CNV-k helyzete és kiterjedése

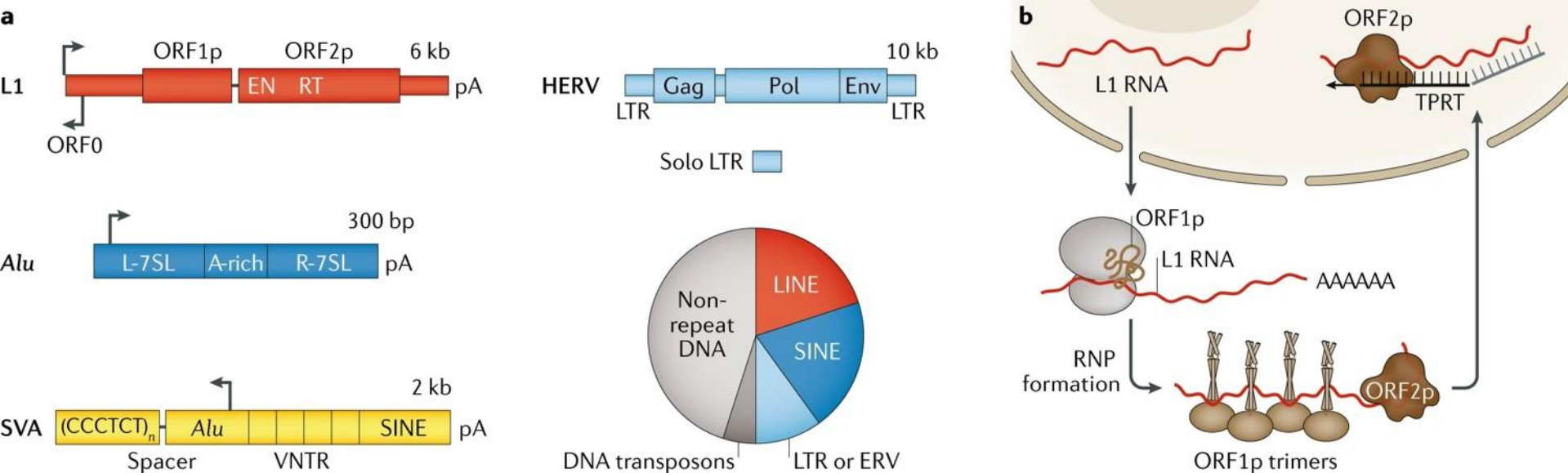
# Mobilis Genetikai Elemek (Transzpozonok)

**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 of complete copies
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
DNA transposon copies	300,000	3%	complete copies, yes	2–3 kb

Incomplete elements, incapable of autonomous transposition, are common (see [Section 3.5](#)). [Data from International Human Genome Sequencing Consortium (2001) *Nature* 409, 860.]

# A humán genom transzpozábilis elemeinek szerkezete

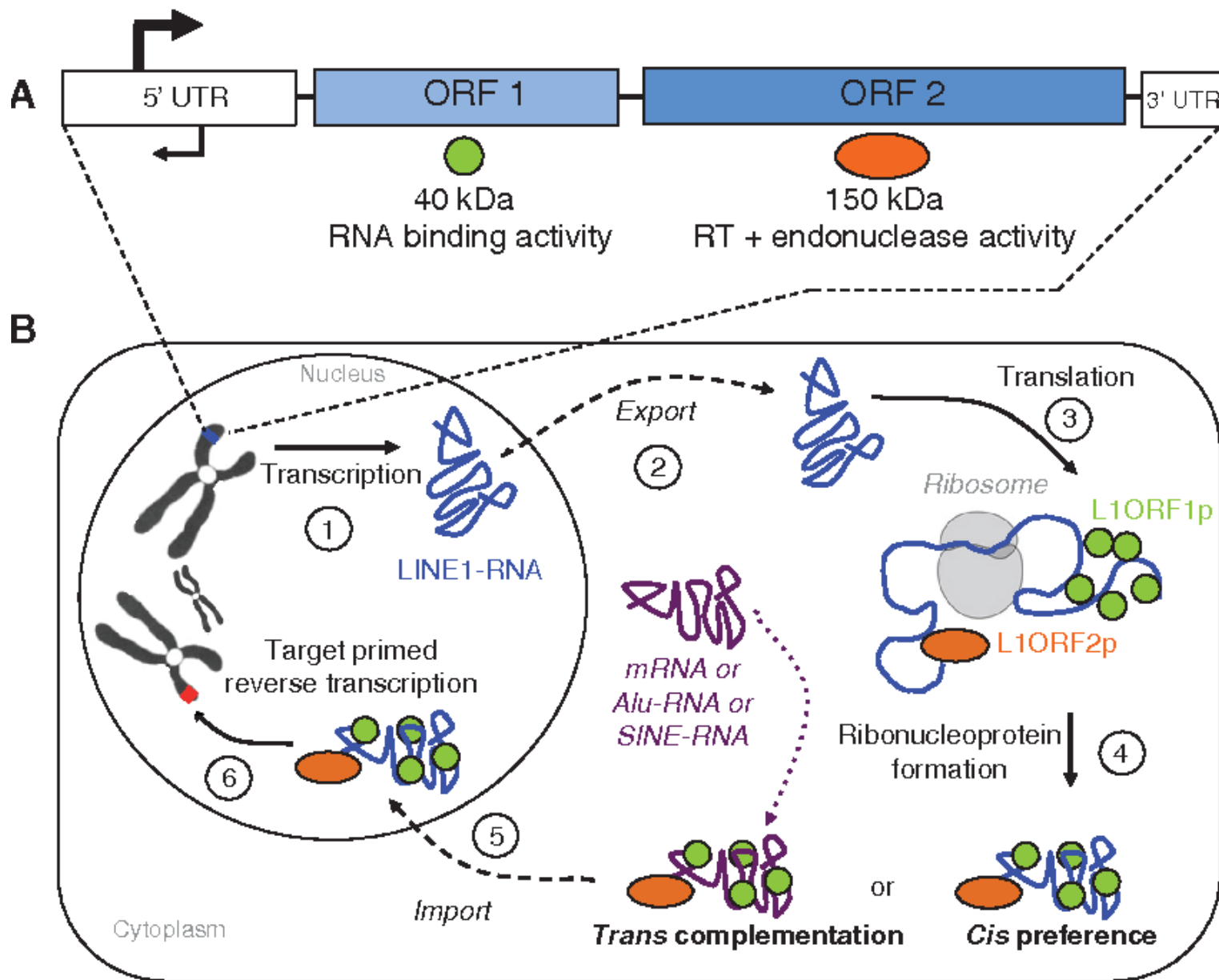


A schematic of common human transposable elements with their full-length size denoted. Long interspersed element 1 (LINE-1 or L1) encodes two open reading frames (ORFs). ORF2p protein has endonuclease (EN) and reverse transcriptase (RT) domains.

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.

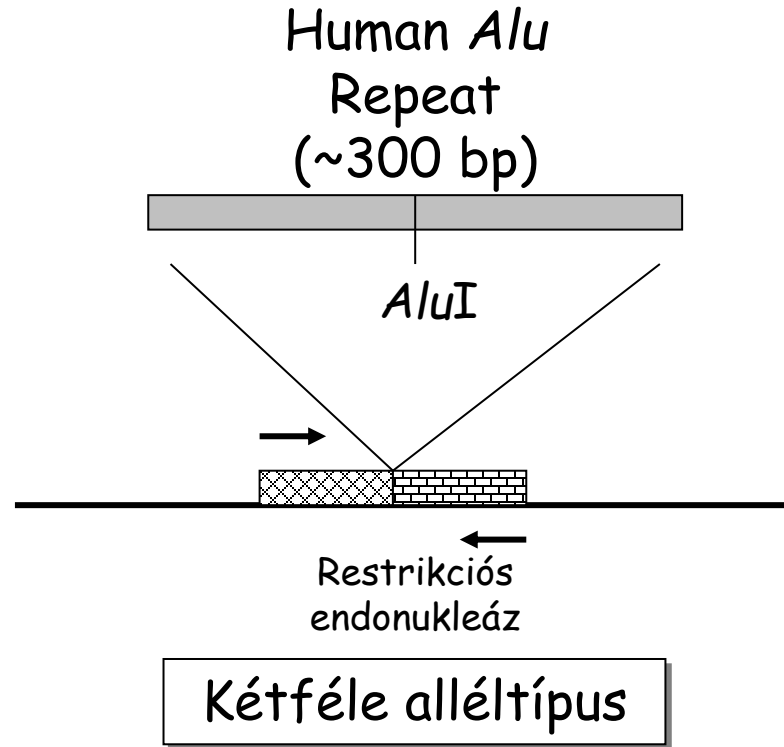
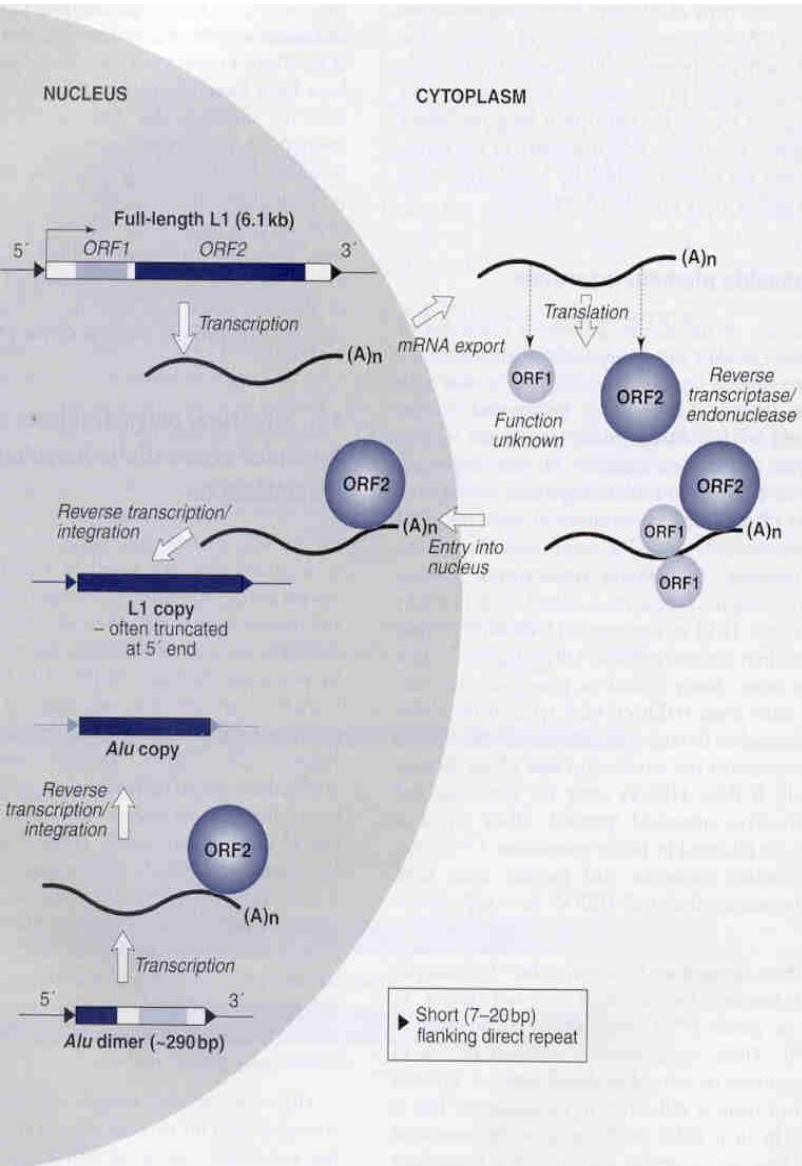
The pie chart shows the proportion of the human genome made up of these repetitive sequences.



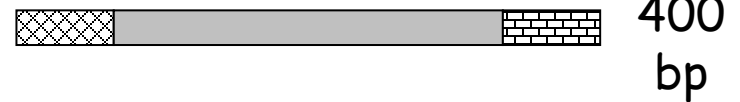


# Mobilis elemek: biallélikus polimorfizmus a DNS-ben

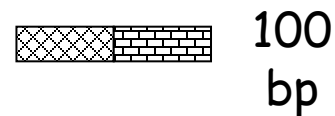
LINE és SINE elemek: long ill. short beépült szekvenciák



"long" (+)  
allél



"short" (-)  
allél



**TABLE 3.1** Approximate fractional composition of the human genome

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

*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.

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

Source: Lynch 2006a.

Gének száma

vs.

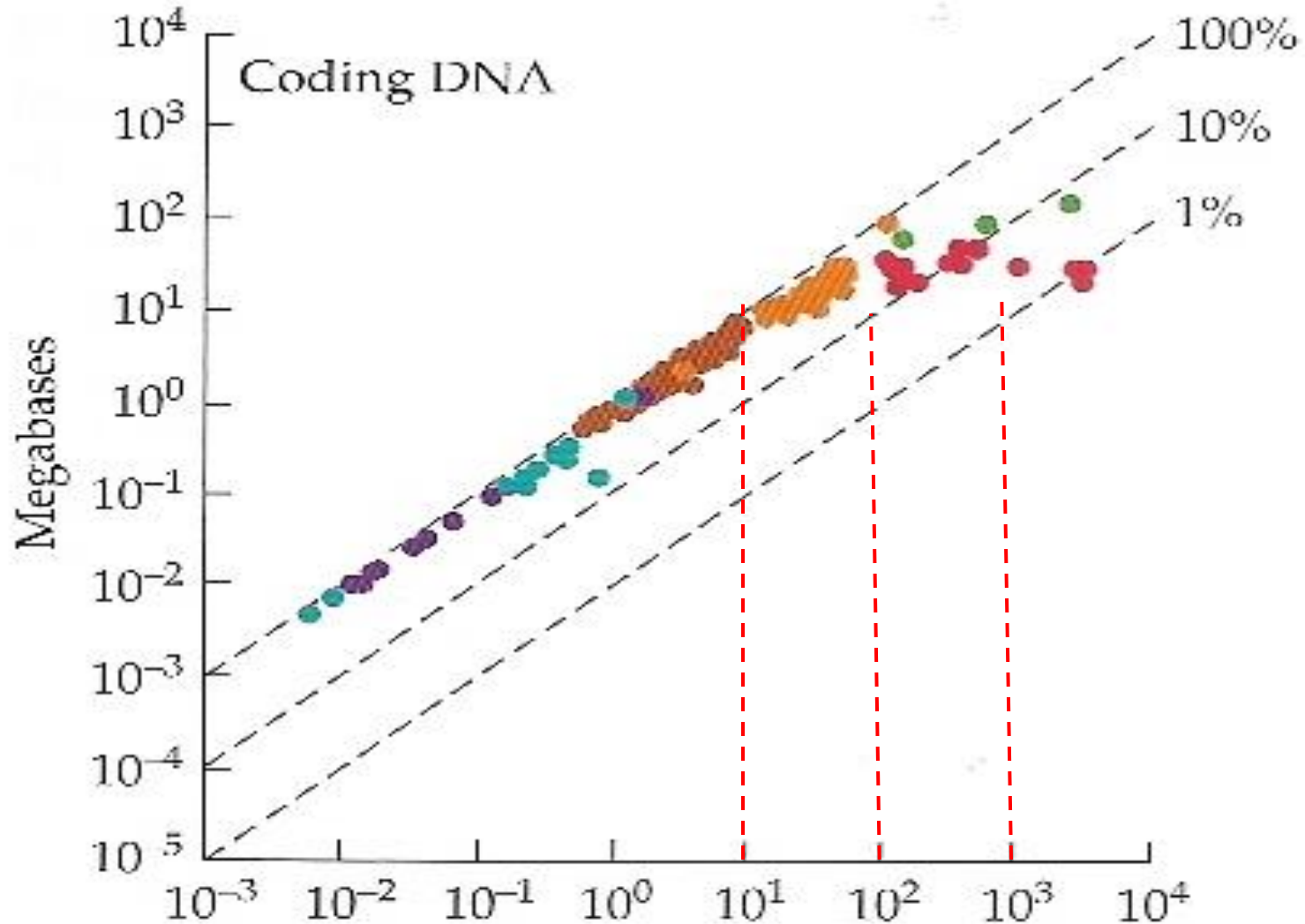
Kódoló szekvenciák  
hossza

Genom méret

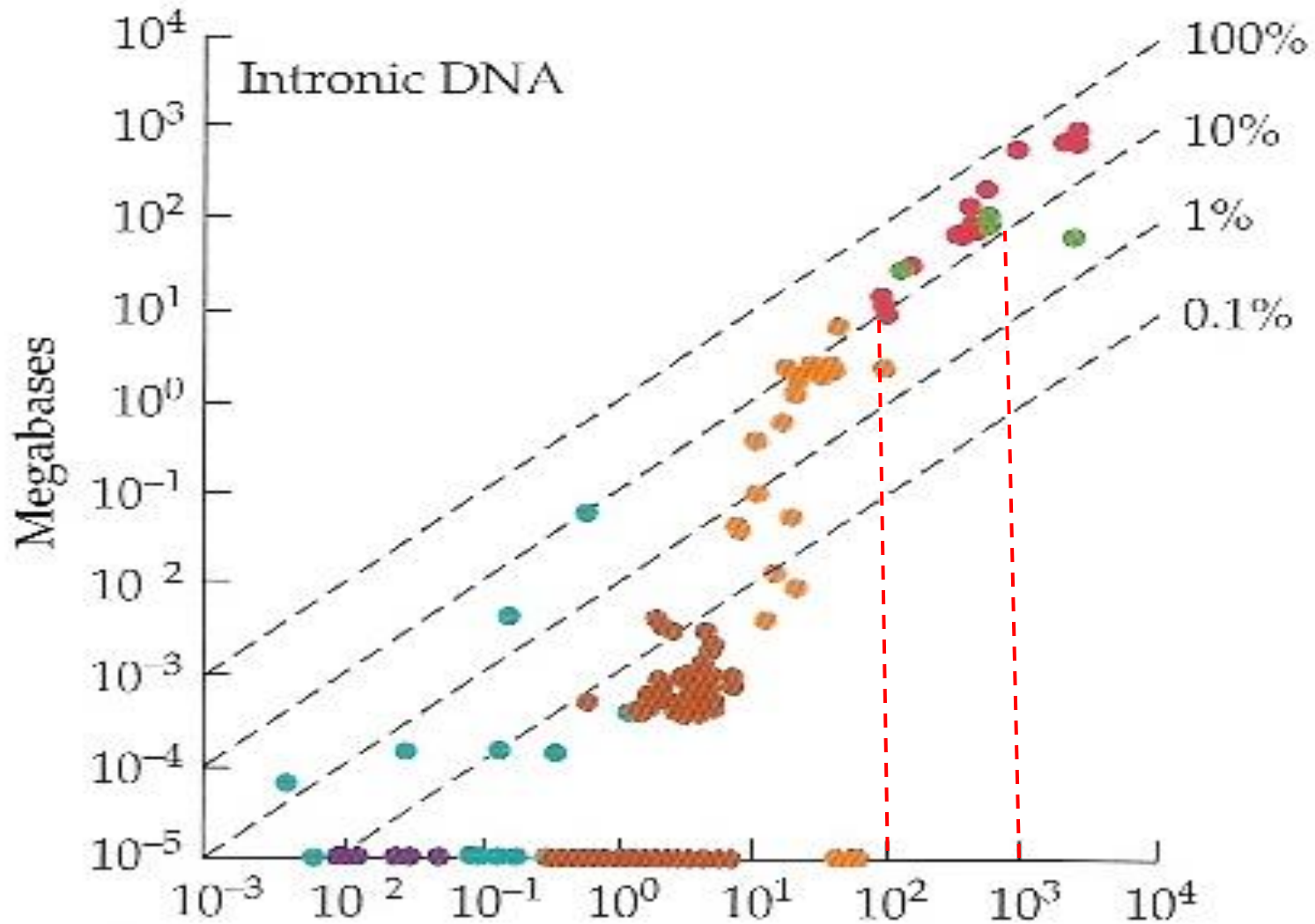
vs.

Nem-kódoló  
szekvenciák hossza

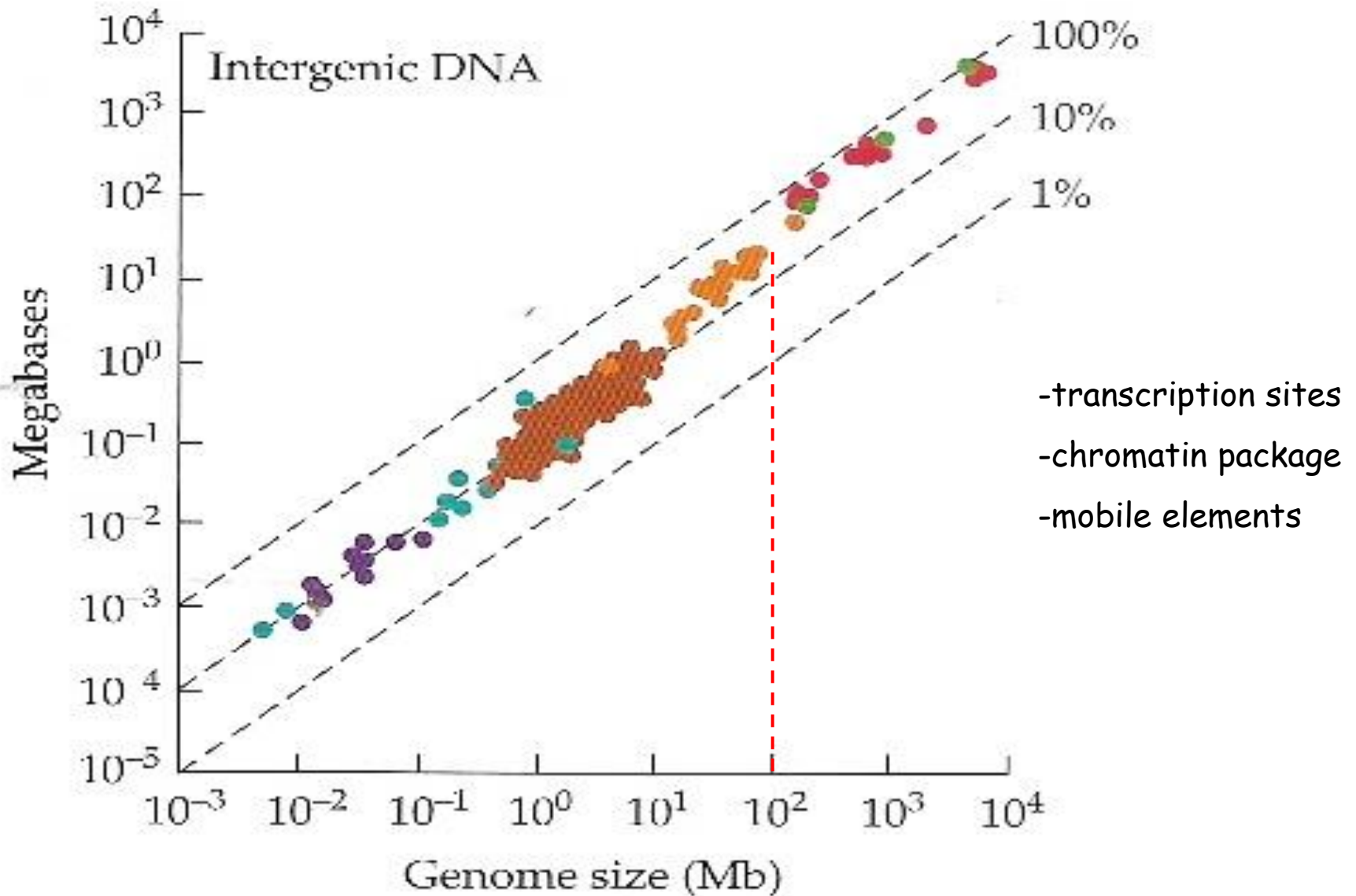
# Genom méret vs. kódoló szekvenciák



# Genom méret vs. intronok



# Genom méret vs. intergénikus DNS



# Genom méret és szerkezeti komplexitás

- **C-value paradoxon:** haploid genom méret/sejt - **nincs összefüggés a szerkezeti komplexitás és a genom méret között!**
- **Prokarióta:** 350-8000 gén, 0.5 - 9 Mb genom
- **Multicelluláris Eukarióta:** > 13.000 gén, > 100 Mb genom
- Nem-kódoló DNS expanzió (intronok, mobilis elemek, pszeudogének)
- Organizmus mérete vs. sejttípusok száma - pozitív korreláció
- Génszám / genom méret vs. multicellularitás / szerkezeti komplexitás

*Van korreláció? Nem a genom mérettől v. génszámtól függ, hanem ahogy a gének működnek (transzkripciós szabályozás, alternatív splicing, stb.)*

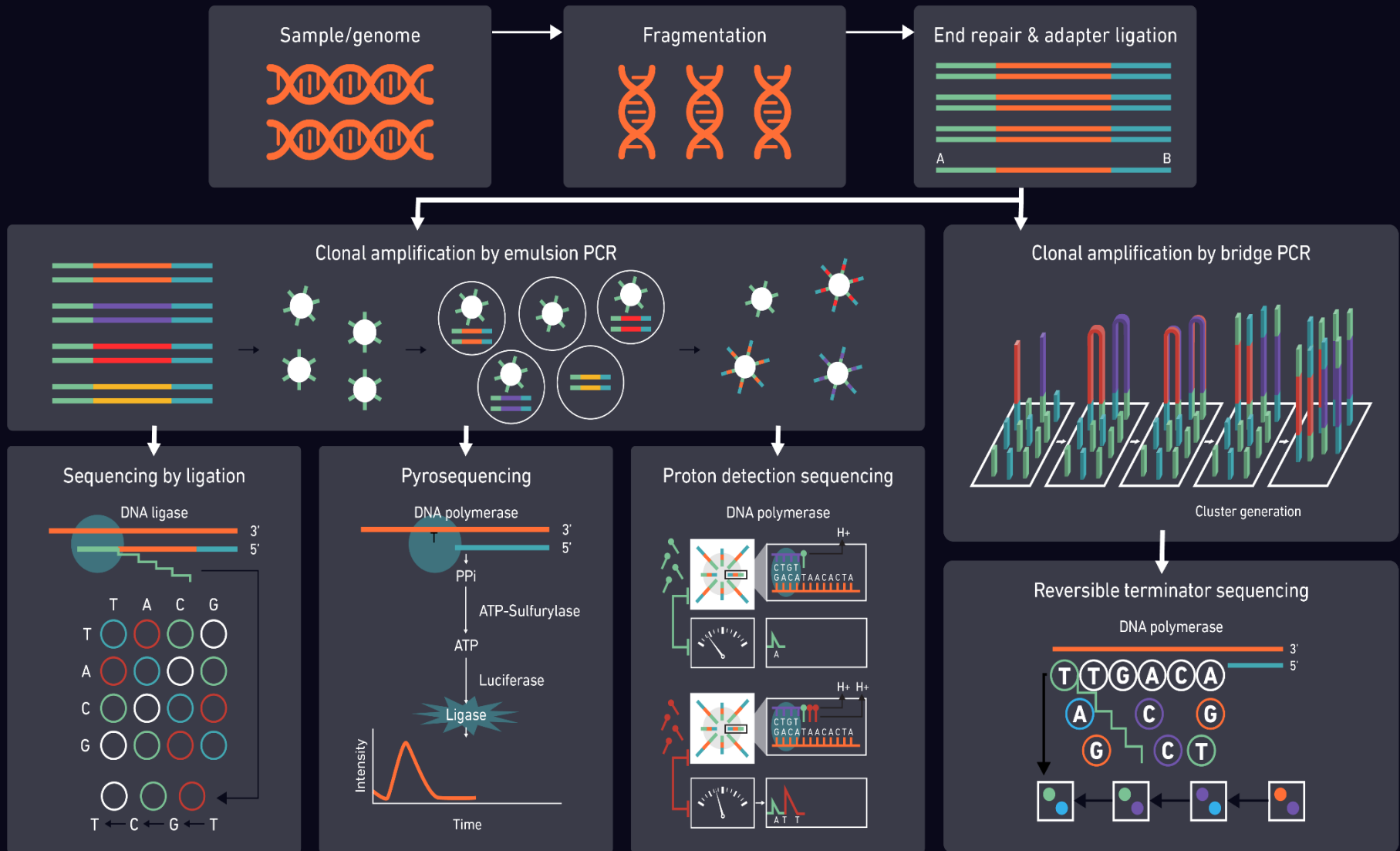


## Next Generation Sequencing –

# Massively Parallel Sequencing of clonally amplified (or single) DNA molecules

- Process millions of sequence reads in parallel
- Library preparation
- Specific adaptor oligos
- Little volume DNA template
- Produce shorter read lengths (35-400 bp)
- 100 Mb to several Gb nucleotid sequence determination

# Next Generation Sequencing – Massively Parallel Sequencing of clonally amplified (or single) DNA molecules



# A global reference for human genetic variation

The 1000 Genomes Project Consortium\*

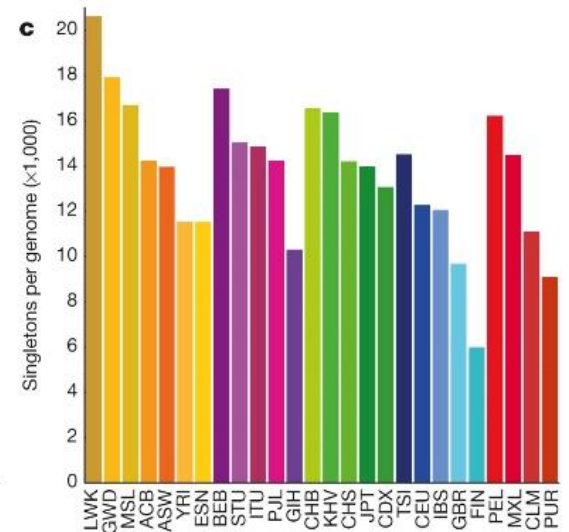
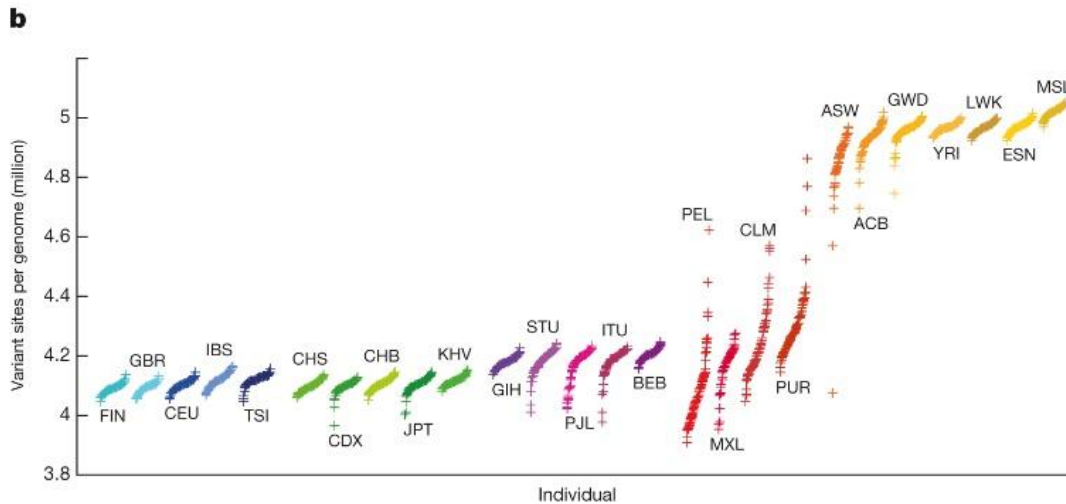
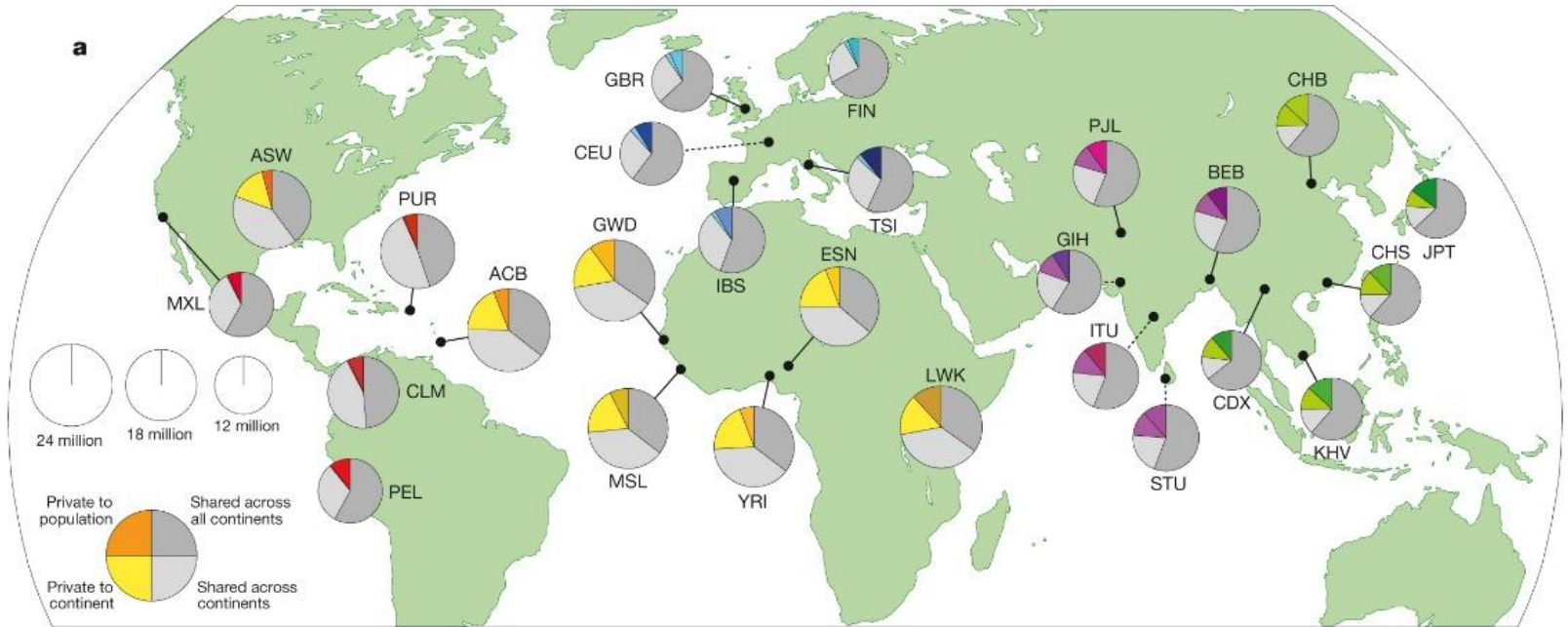
The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

## An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene-intersecting structural variants exhibiting population stratification and describe naturally occurring homozygous gene knockouts that suggest the dispensability of a variety of human genes. We demonstrate that structural variants are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of structural variant complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex structural variants with multiple breakpoints likely to have formed through individual mutational events. Our catalogue will enhance future studies into structural variant demography, functional impact and disease association.

# Population sampling



## Conclusions

- a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites.
- >> 99.9% of variants consist of SNPs and short indels.
- structural variants (CNVs) affect more bases:
  - typically 2,100 to 2,500 structural variants  
(1,000 large deletions, 160 copy-number variants, 915 Alu insertions, 128 L1 insertions, 4 NUMTs and 10 inversions)
- affecting 20 million bases of sequence.