Genomika

Hogyan változtak az elképzelések a genom tartalmáról, a szerveződési komplexitás és génszám közti összefüggésről?

ELTE TTK Genetikai Tanszék

2019.10.03.

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

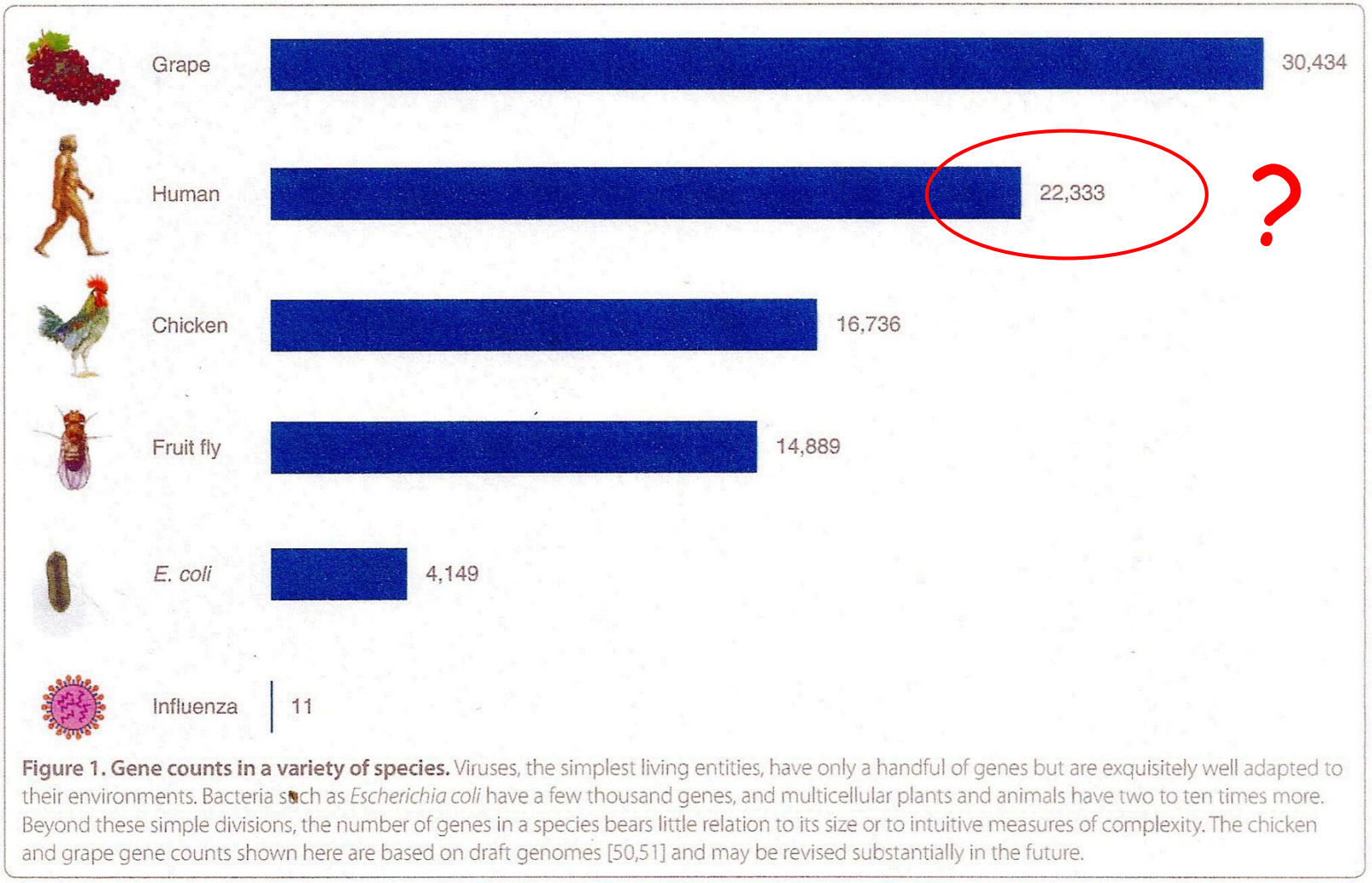


Figure 1. Gene counts in a variety of species. Viruses, the simplest living entities, have only a handful of genes but are exquisitely well adapted to their environments. Bacteria such as *Escherichia coli* have a few thousand genes, and multicellular plants and animals have two to ten times more. Beyond these simple divisions, the number of genes in a species bears little relation to its size or to intuitive measures of complexity. The chicken and grape gene counts shown here are based on draft genomes [50,51] and may be revised substantially in the future.

(Pertea & Salzberg, Genome Biology 2010, 11:206)

Első becslések a genom méretéről és a gének számáról

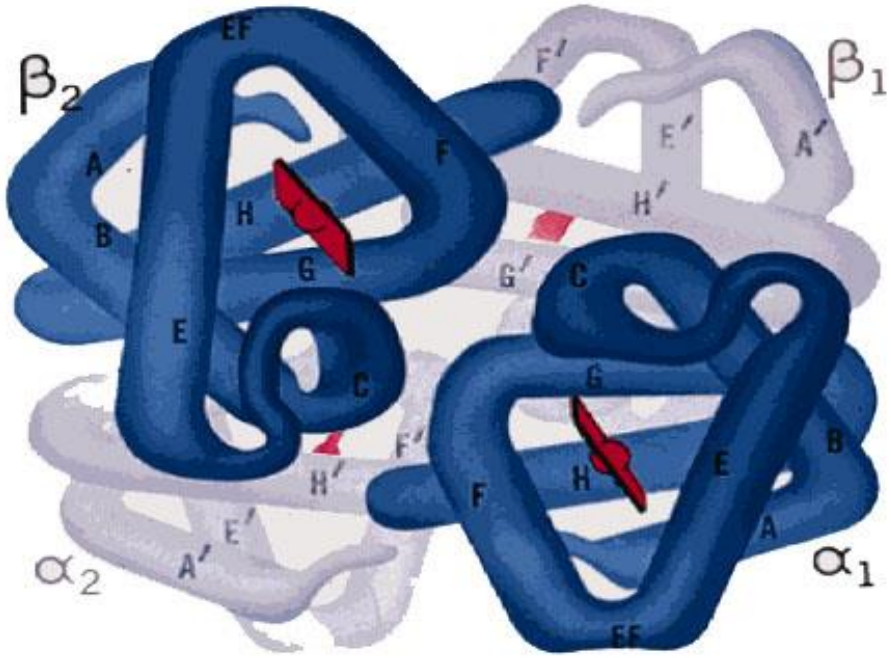
1964: F. Vogel (Heidelberg)

-Hemoglobin α és β lánc

-leegyszerűsített feltevések

-Haploid genom: 3×10^9 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

2001, Human Genome Project: csökkenő génszám, növekvő bizonytalanság

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 átírásáért felelős genetikai információt tartalmazza" (mRNS, rRNS, snRNS, tRNS, …)

ORF
- *open reading frame* (nyitott leolvasási keret)
- gének annotálása: bioinformatikai módszerekkel predikció (konzervált szekvencia motívumok alapján)

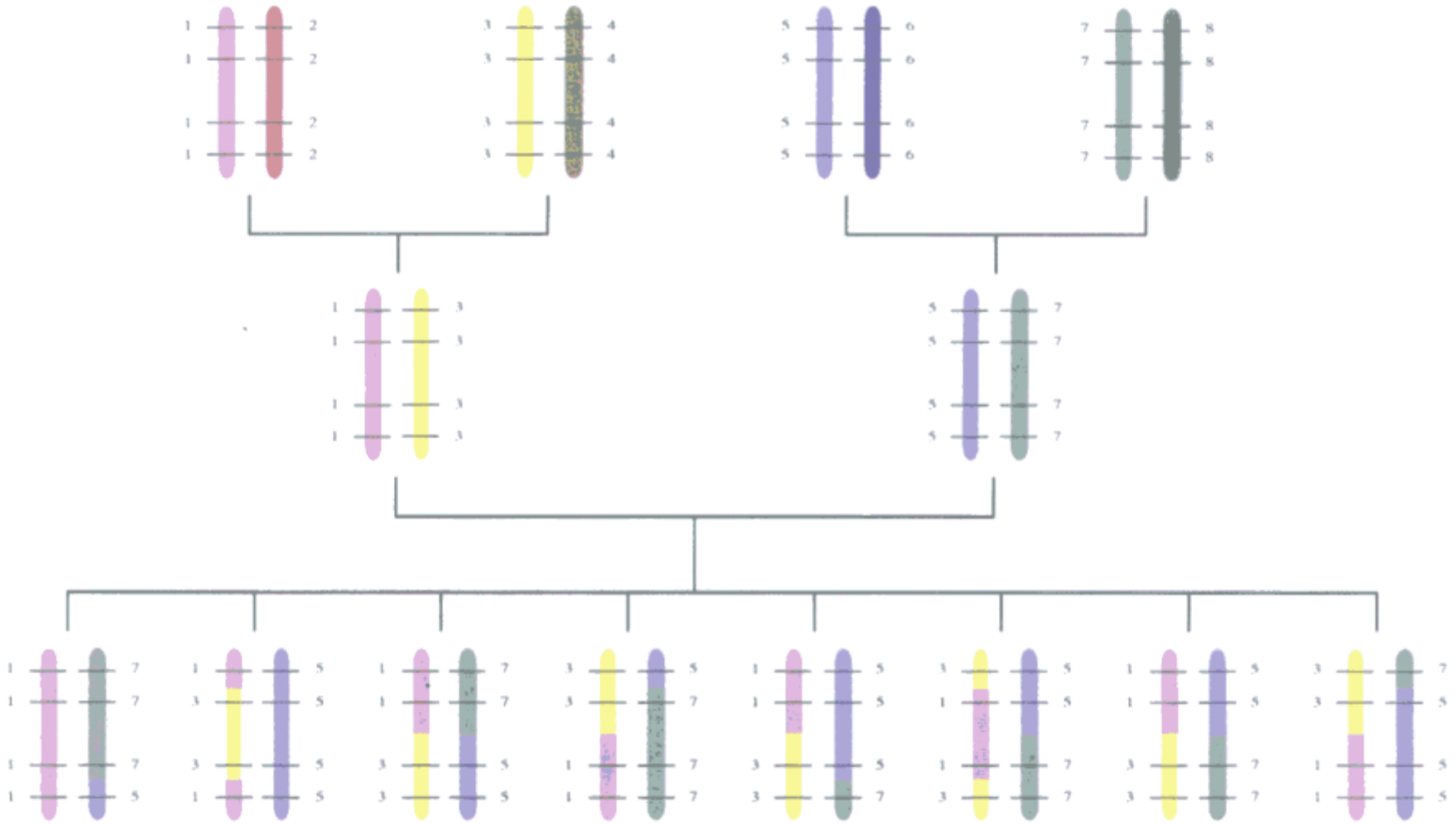
„Egy gén a genetikai állományunk jól körülhatárolt szakasza, mely mRNS-ként átíródik és egy v. több fehérjét kódol." (pl. alternatív splicing: izoformák)

És a nem protein kódoló RNA gének (pl. lncRNAs, miRNAs, snRNAs, piwiRNA)?

Human Genom Project (HUGO)

- Első kezdeményezés: 1980-as évek eleje
- orvosbiológiai megközelítés, infrastruktúrális beruházások
- Folyamatban lévő genom szekvenálási projektek
- λ-fág, SV40, humán mitokondriális genom
- Genetikai és fizikai térképezések
- Botstein et al., 1980; Coulson et al., 1986;
- DNS-szekvenálási technológia és bioinformatika
- shotgun sequencing, automated sequencing, ESTs, STSs,
- NRC Report 1988, US DOE, NIH,
- genetikai és fizikai térkép, parallel projektek modell organizmusokon, technológiai fejlesztések, bioetika

Meiotic Breaks – Genetic Linkage Maps



Universal Landmark

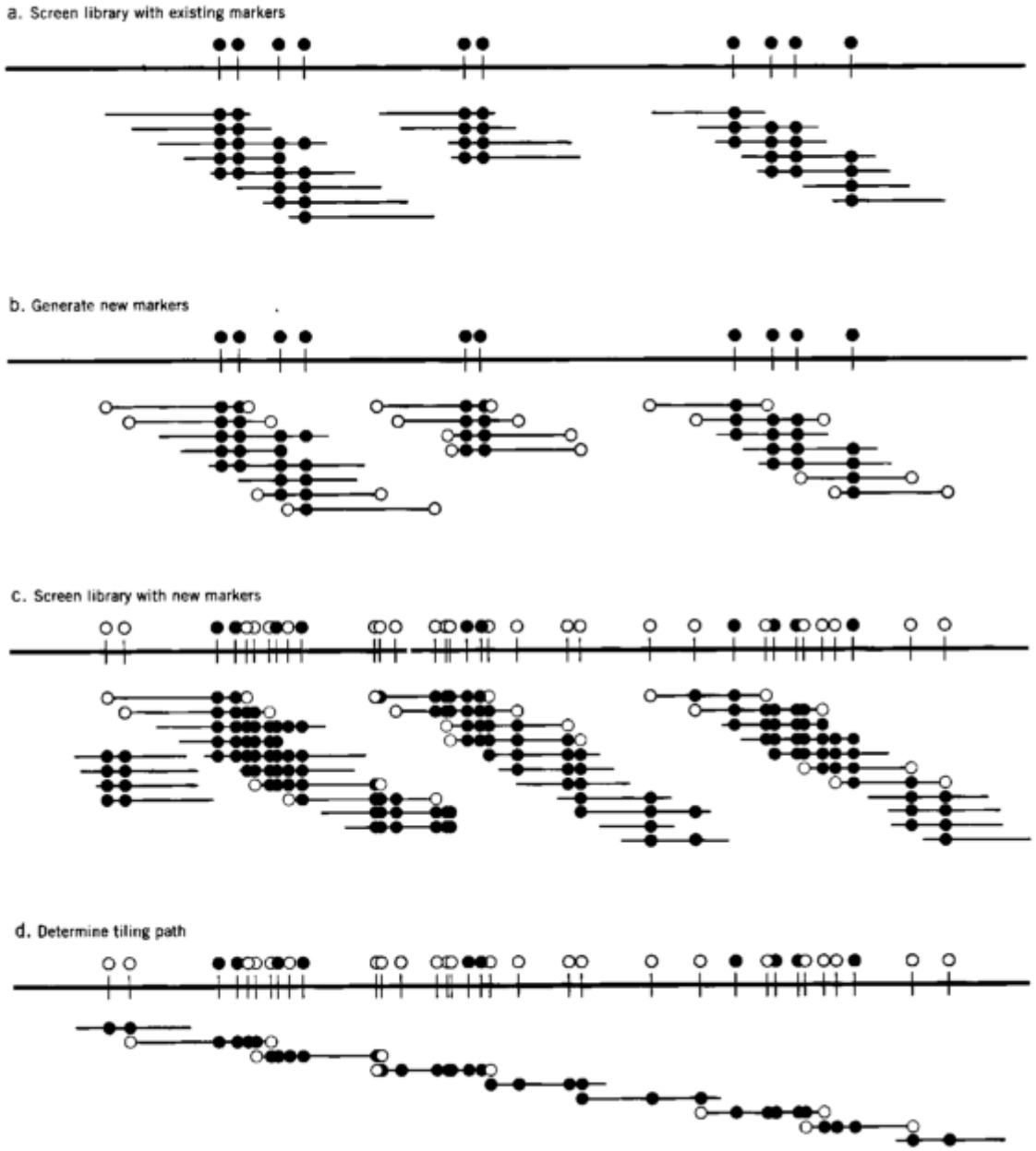
Sequence Tagged Site (STS) 1989

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.

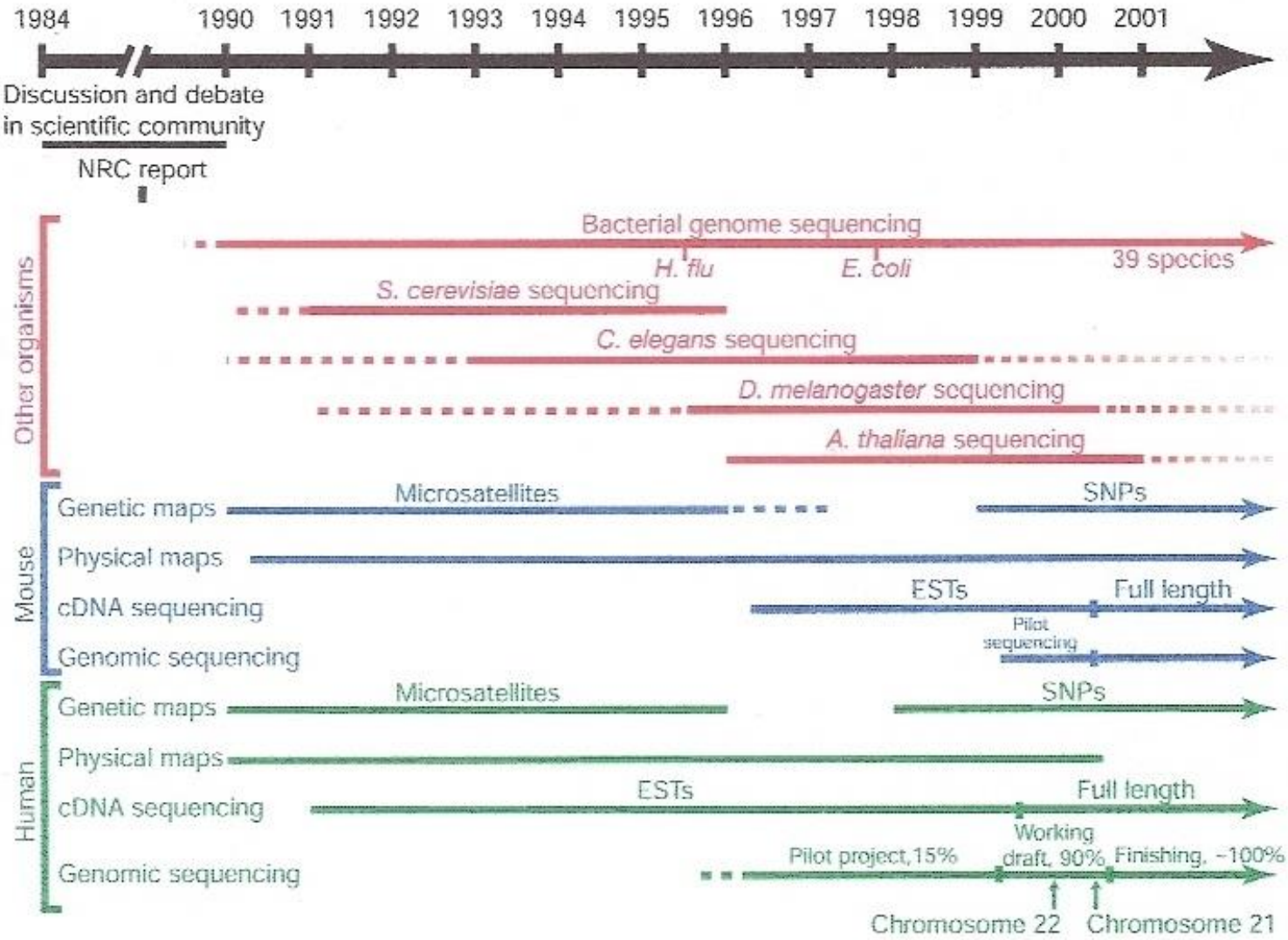
Enables merging of mapping data obtained from many labs using many different methods into a single consensus map of landmarks along a chromosome.

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



Clone ends – Clone-based Physical Map

Genom projektek időskálán – a „hőskor”



Humán Genom Projekt
- résztvevők és módszerek

- **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 „shotgun" szekvenálás
- 2001: draft genom szekvencia, 2003: teljes genom szekvencia

- **Celera Genomics**:
- Applied Biosystems., TIGR (C. Venter)
- 1998-2001: „whole genome shotgun"
- ABI PRISM 3700 DNA Analyzer

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

„shotgun" genom szekvenálási stratégiák

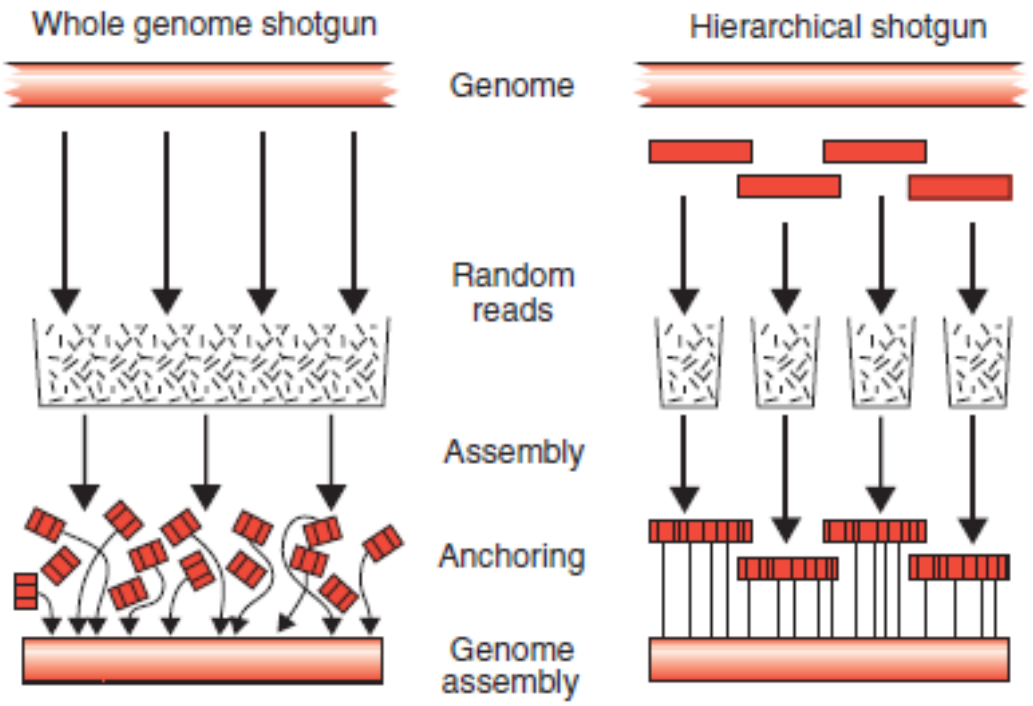
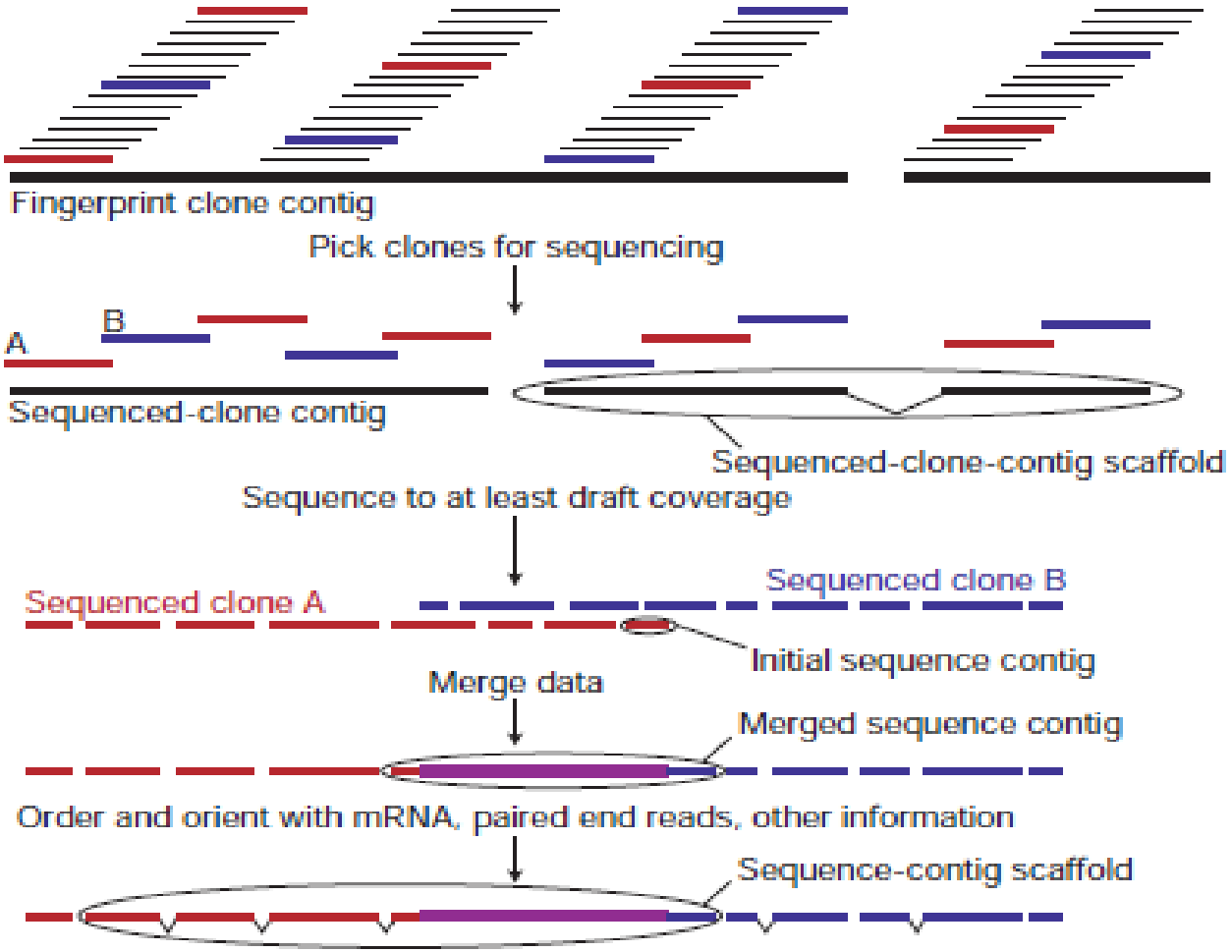


Figure 9.11. Assembling genomic data using the hierarchical and whole genome shotgun approaches. Adapted from Waterston, Lander and Sulston (2002), with permission

RJ Reece: Analysis of Genes and Genomes, 2004

Genom szekvenciaváz összeállítása



International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome, Nature 409, 860 (2001)

Teljes genom összeszerelés

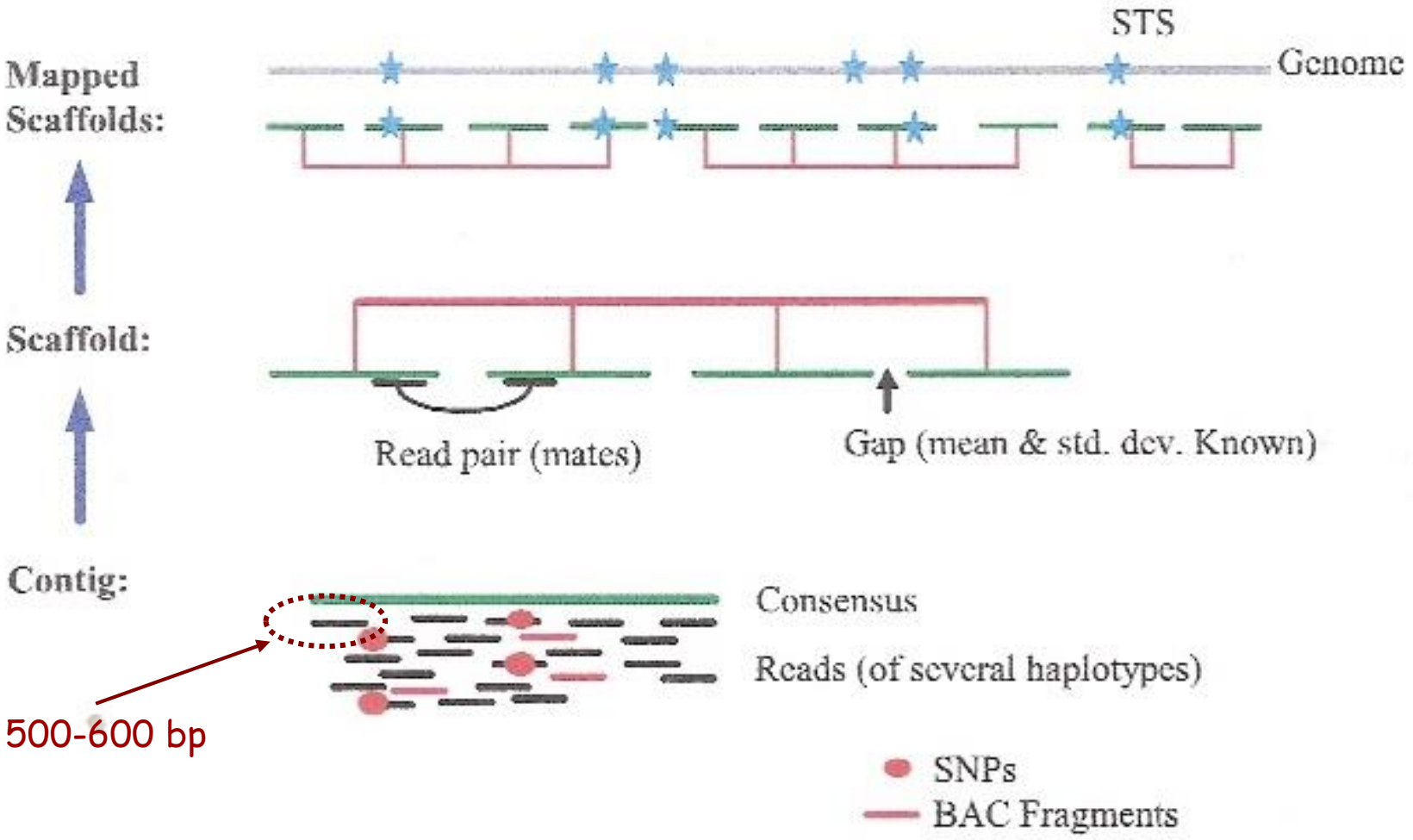


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair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

JC Venter, et al.: The Sequence of the Human Genome, Science 291, 1304 (2001)

STS genom térképezés

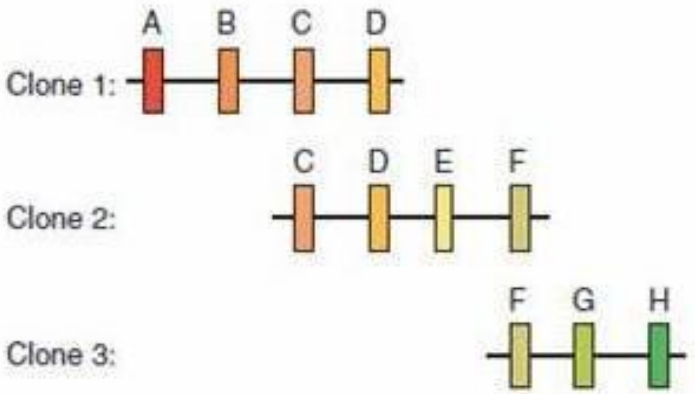


Figure 9.5. Aligning clones by STS mapping. Each clone contains several STSs. Clone 1 has four (A, B, C and D). Clone 2 also contains STSs C and D. Therefore clones 1 and 2 overlap with each other

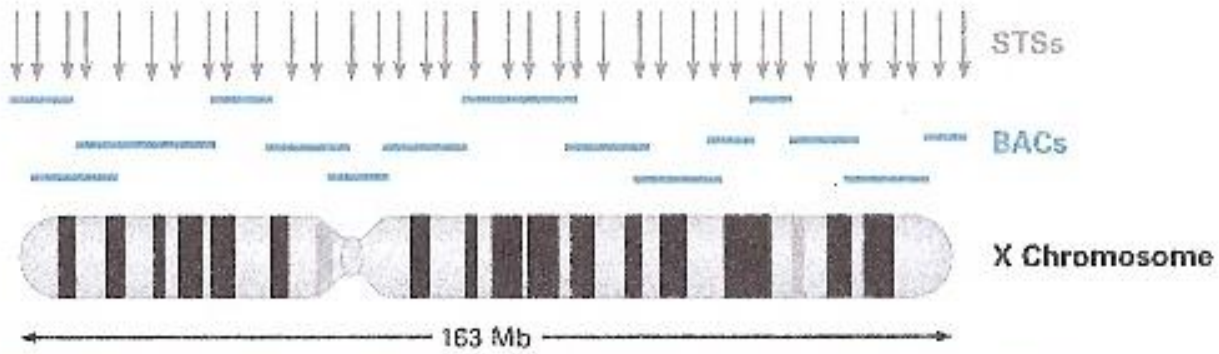
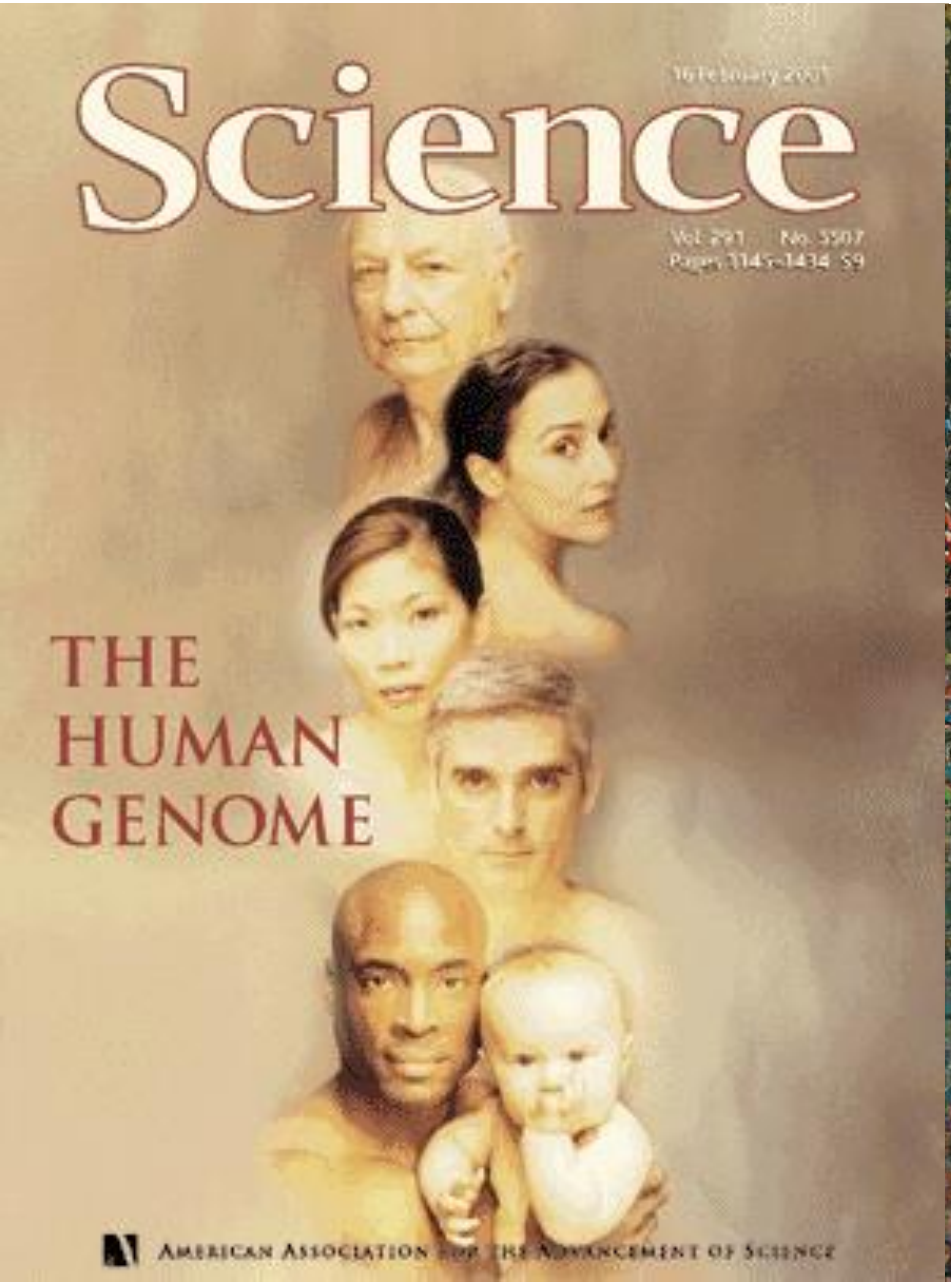
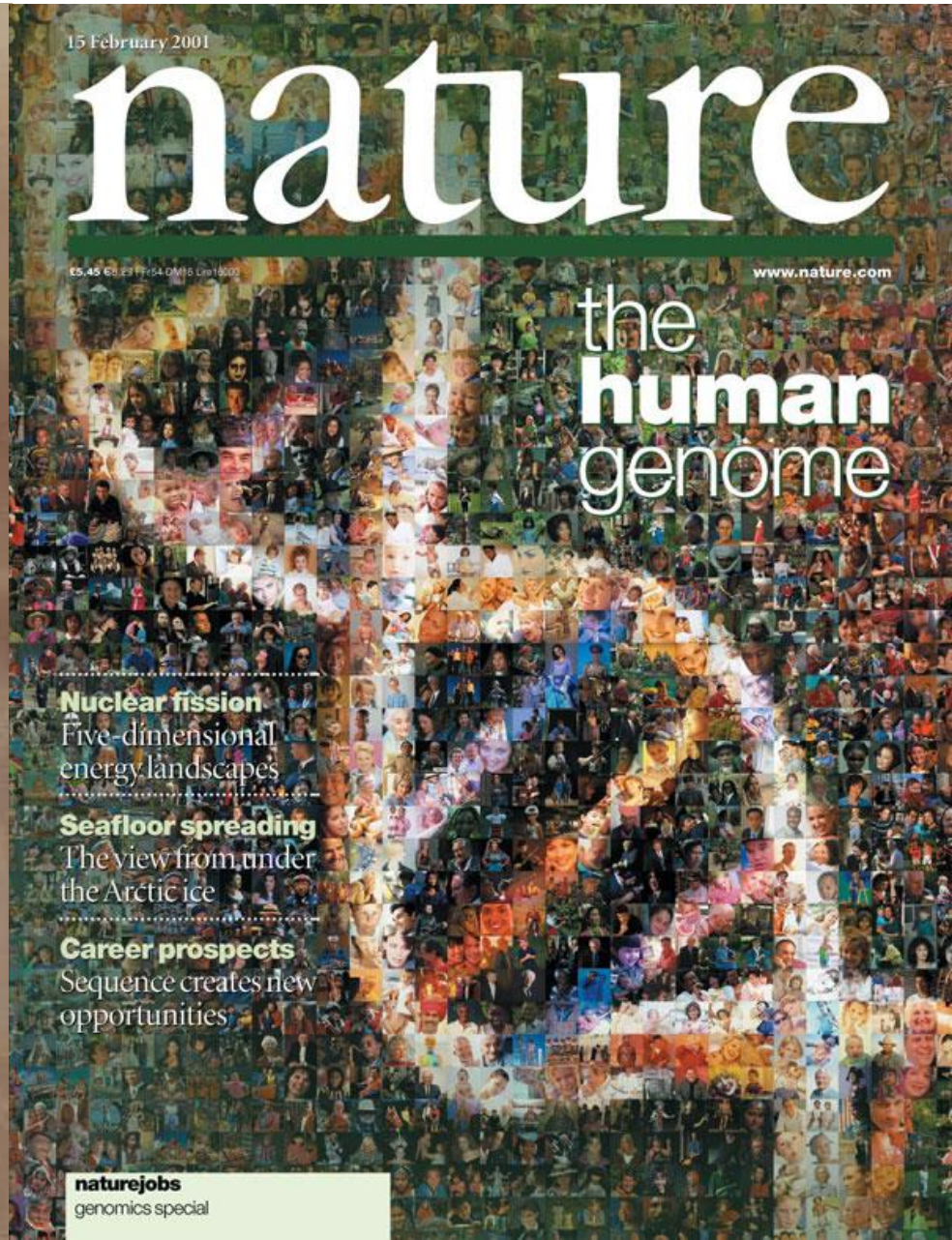


FIGURE 1.3 • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout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

Humán Genom Projekt





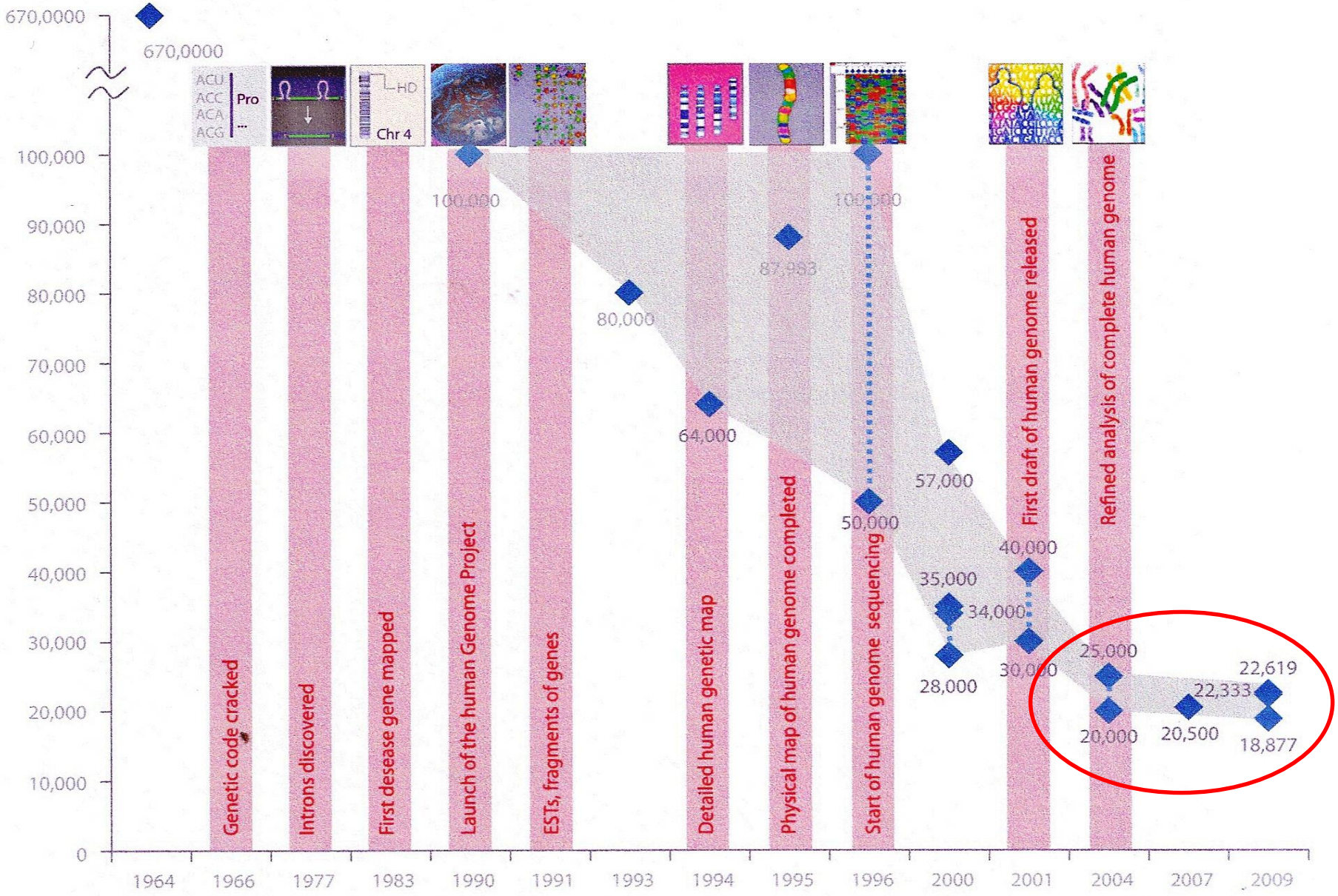


Figure 2. The trend of human gene number counts together with human genome-related milestones. Individual estimates of the human gene count are shown as blue diamonds. The range of estimates at different times is shown by the two vertical blue dotted lines. Note how this range has narrowed in recent years.

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/CcdsBrowse.cgi)

fals pozitívak: retrotranszpozonok, pszeudogének, „orphan" DNS

2019.09.08.: CCDS GeneID: 19 029 genes > 1 CCDS ID: 7 869

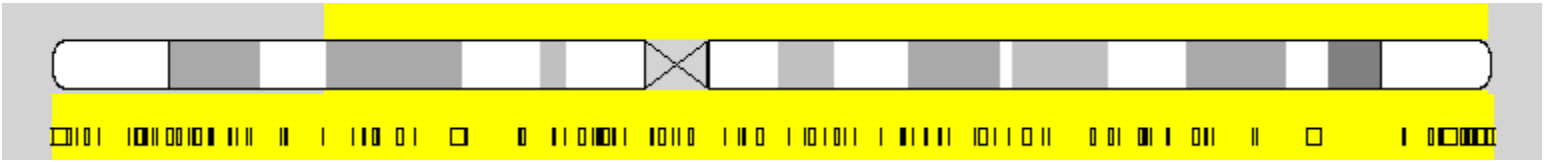
Copy Number Variation (CNV)

Kópia-szám variábilitás

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 prosztata rák 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

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 polymorphisms (CNVs)
- emberi „pángenom": változatok rasszok, csoportok között.

(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)

Emerging novel gene sequences

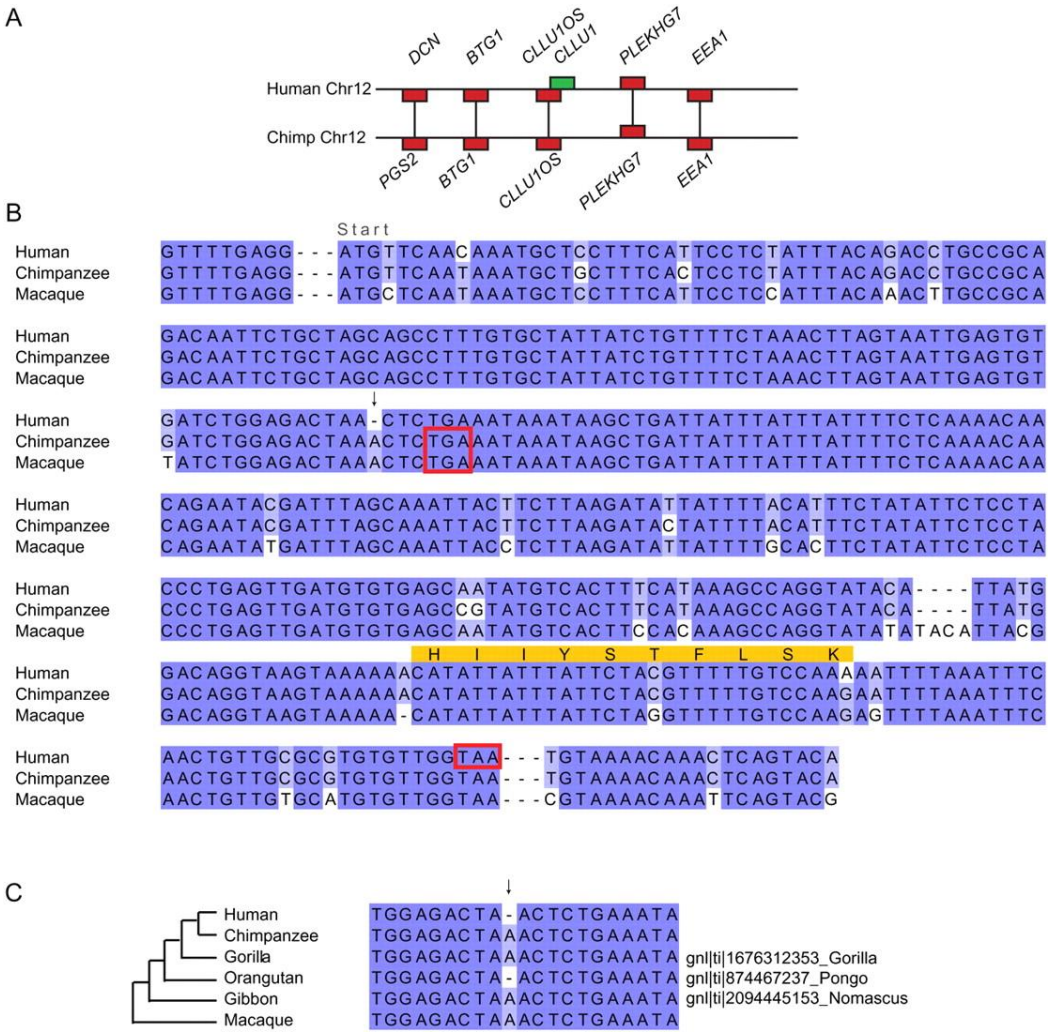
Table 1. Novel human protein-coding genes and supporting evidence.

Gene name	Ensembl ID	Length (codons)	Longest chimp ORF[a]	Expression support and tissue[b]	Primate shared disablers[c]	Other major sequence differences	Presence of enabler in other human complete genome sequences[d]	HapMap SNPs
CLLU1	ENSG00000205056	121	42	EST/cDNA: Blood (AJ845165, AJ845166); UniGene: Blood, embryonic tissue, eye, lymph, lymph node, muscle, pharynx, tonsil (Hs.339918)	1-bp indel[e]	Macaque: 4- and 1-bp indels	Sequence available and enabler conserved in all	1 syn.; 1 nonsyn.
C22orf45	ENSG00000178803	159	87 (25 amino acids align with human sequence)	EST/cDNA: Kidney, other (AX747284, AK091970, DA635985); 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.
DNAH10OS	ENSG00000204626	163	90 (75 amino acids align with human sequence)	EST/cDNA: Hippocampus (AK127211); 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.

[a]Length in codons of longest in-frame (alignable) ORF starting from any ATG in the region.
[b]Type of data/database is listed followed by tissue information with database identifiers in parentheses. Underlined accession numbers are full-length, spliced cDNA.
[c]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.
[d]Independently sequenced whole genomes: Venter, Watson, HuAA, HuBB, HuCC, HuDD, and HuFF. All data are listed where available.
[e]Not shared with orangutan.

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

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



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

An expanding number of RNA genes

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

Table 1 Gene annotations in Gencode, Ensembl, RefSeq, and CHESS

	Gencode[a]	Ensembl[b]	RefSeq[c]	CHESS[d]
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

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

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.

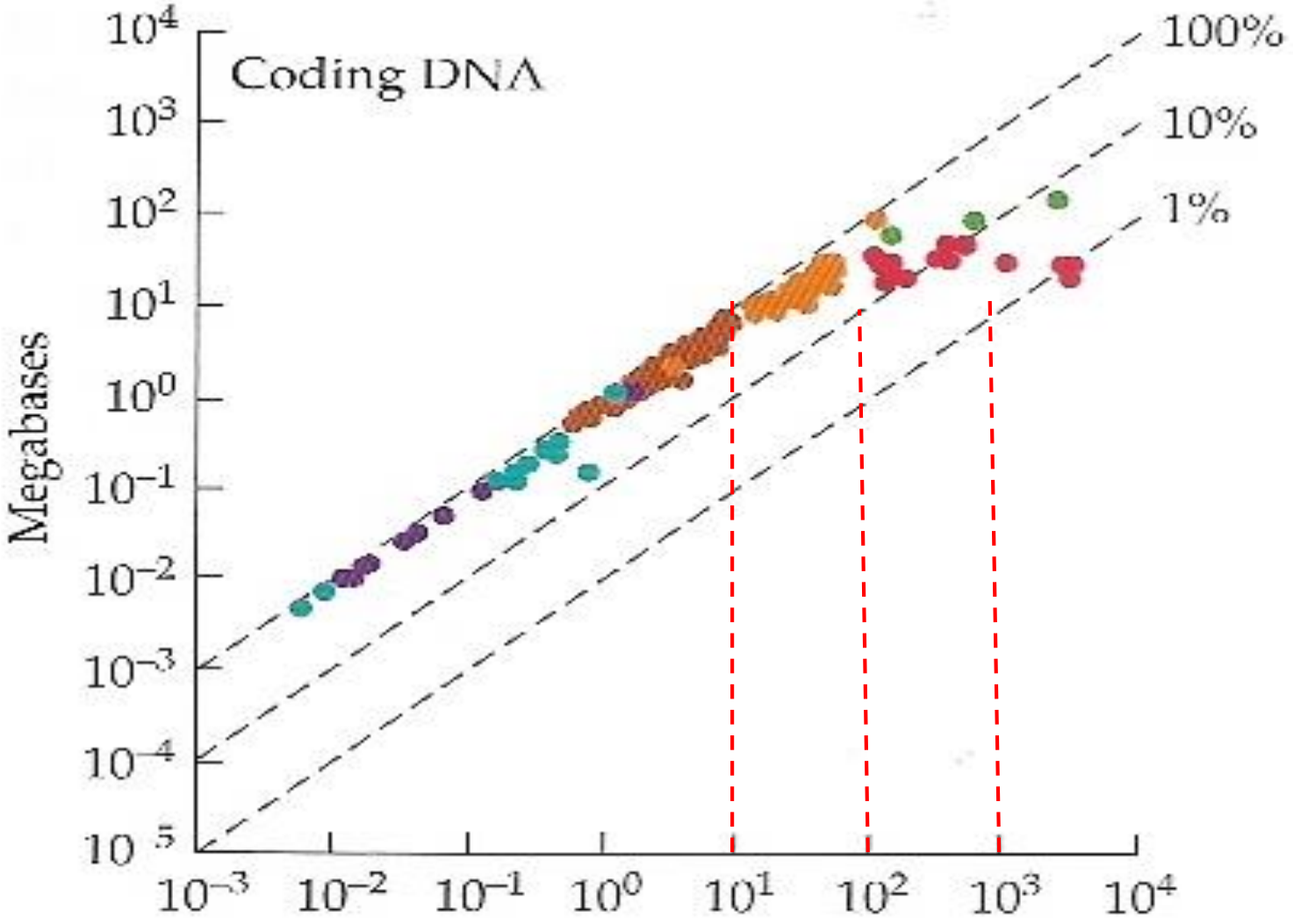
Kódoló szekvenciák hossza

Genom méret

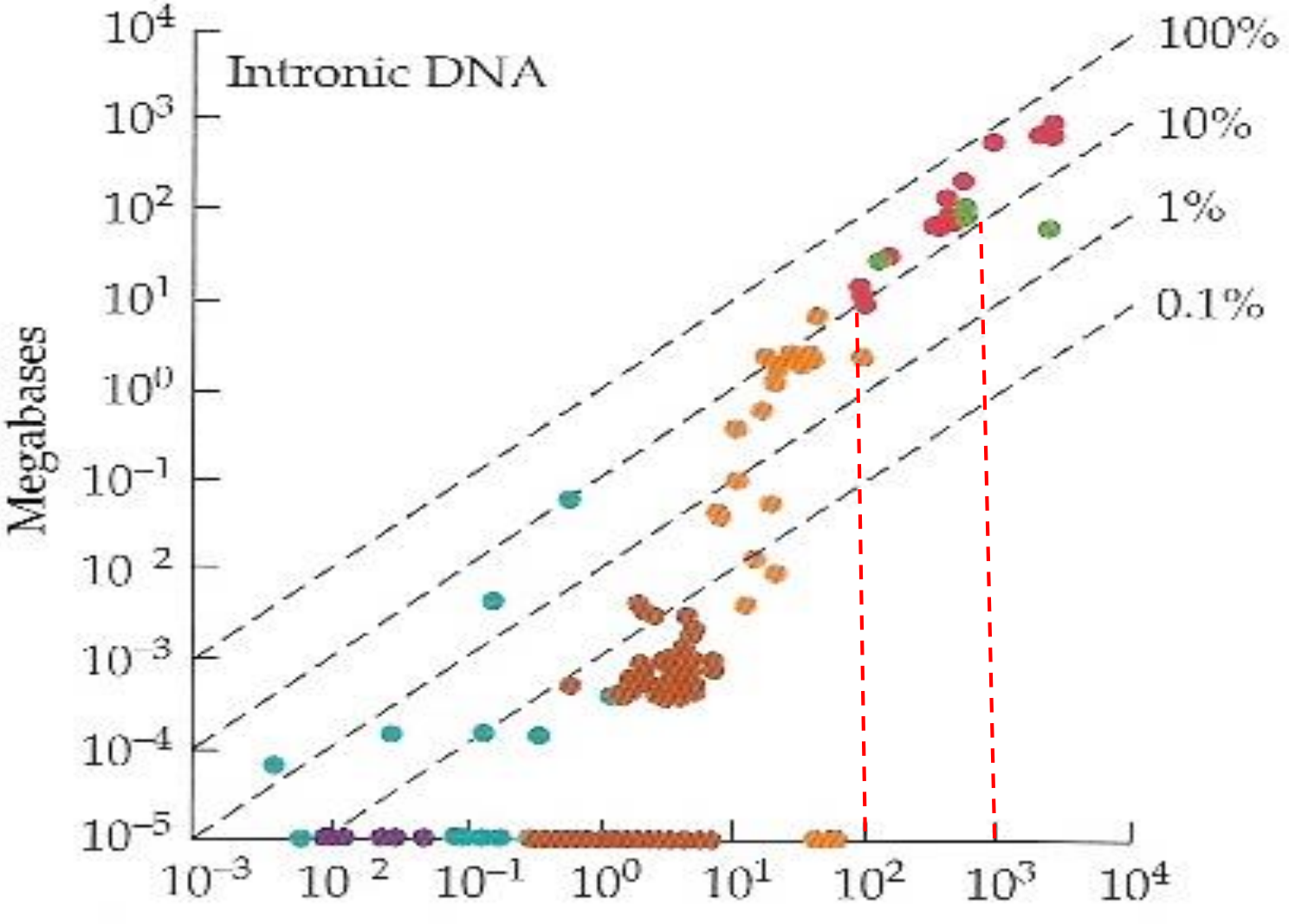
vs.

Nem-kódoló szekvenciák hossza

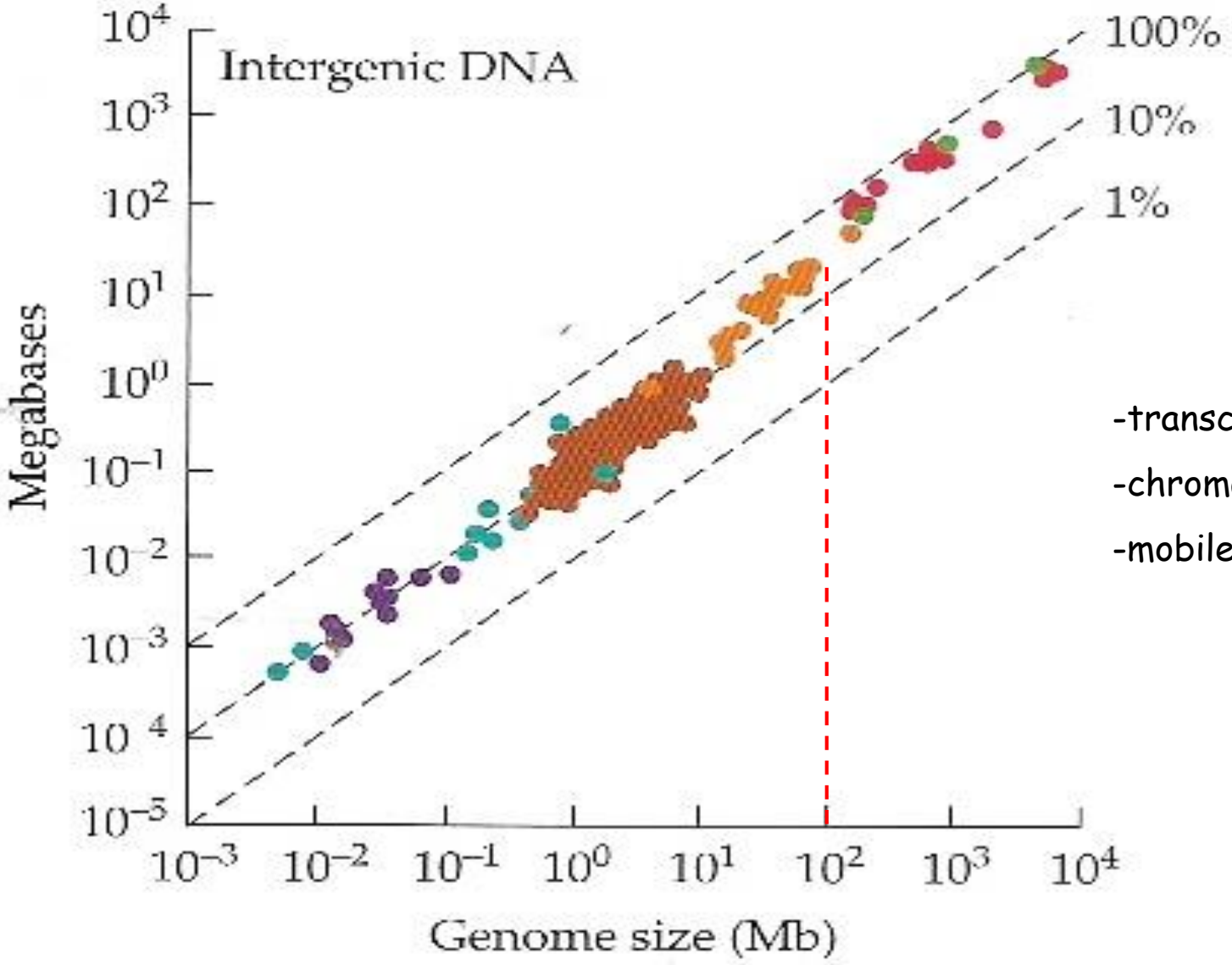
Genom méret vs. kódoló szekvenciák



Genom méret vs. intronok



Genom méret vs. intergénikus DNS



-transcription sites

-chromatin package

-mobile elements

Genom méret és szervezeti komplexitás

- The C-value Paradox: haploid genome size/cell
- Prokarióta: 350-8000 gén, 0.5 – 9 Mb genom
- Multicelluláris Eukarióta: > 13.000 gén, > 100 Mb genom
- Noncoding DNA expanzió (intronok, mobilis elemek, pseudogének)
- Organizmus mérete vs. sejttípusok száma – pozitív korreláció
- Génszám / genom méret vs. multicellularitás / szervezeti komplexitás

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