Genomikai ismeretek

Genomok szerveződése és variabilitása

Genomtartalom, génszám, szerveződési komplexitás, genetikai variablitás

ELTE TTK Genetikai Tanszék

2021.10.11.

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

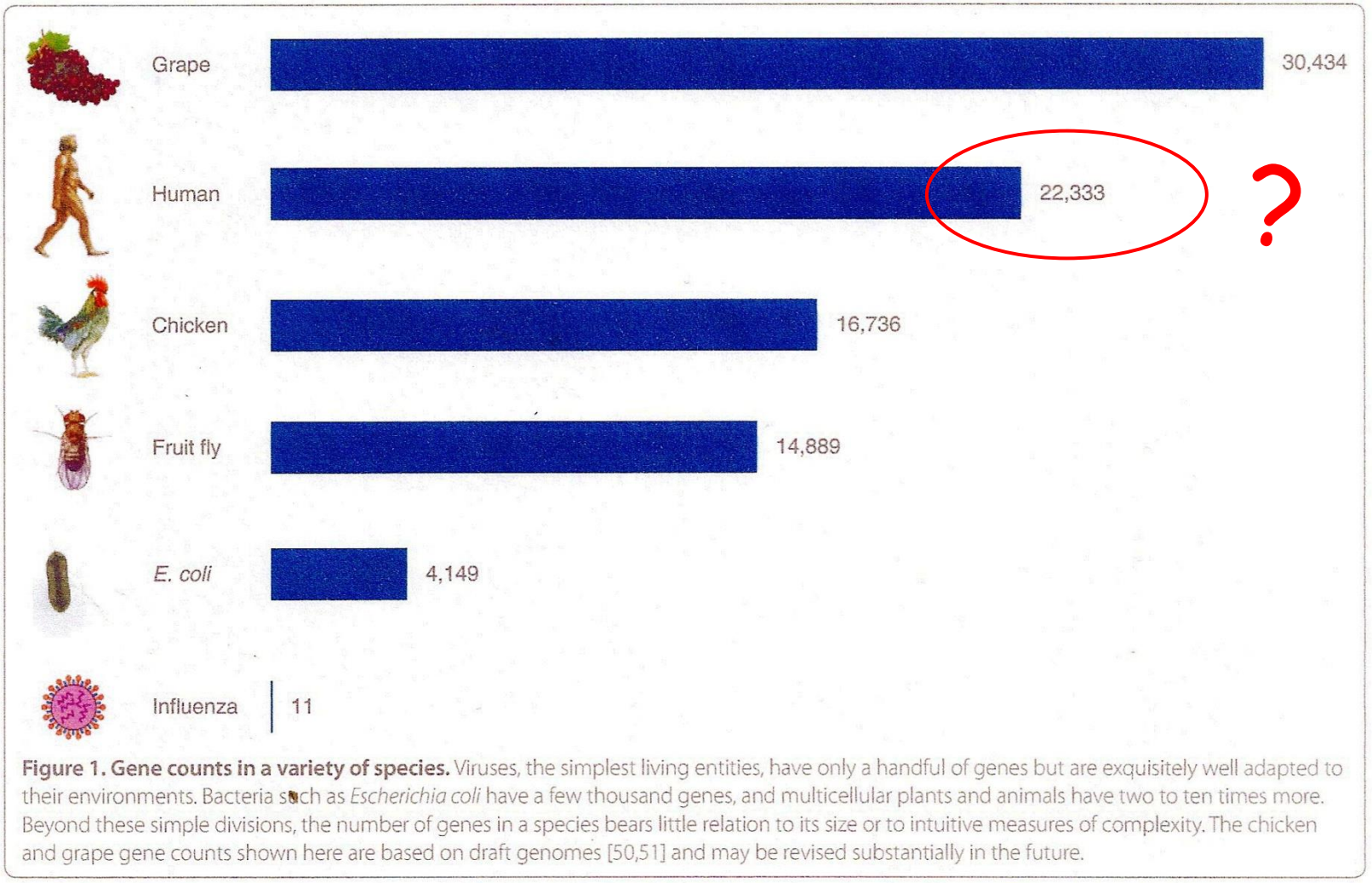


Figure 1. Gene counts in a variety of species. Viruses, the simplest living entities, have only a handful of genes but are exquisitely well adapted to their environments. Bacteria such as *Escherichia coli* have a few thousand genes, and multicellular plants and animals have two to ten times more. Beyond 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

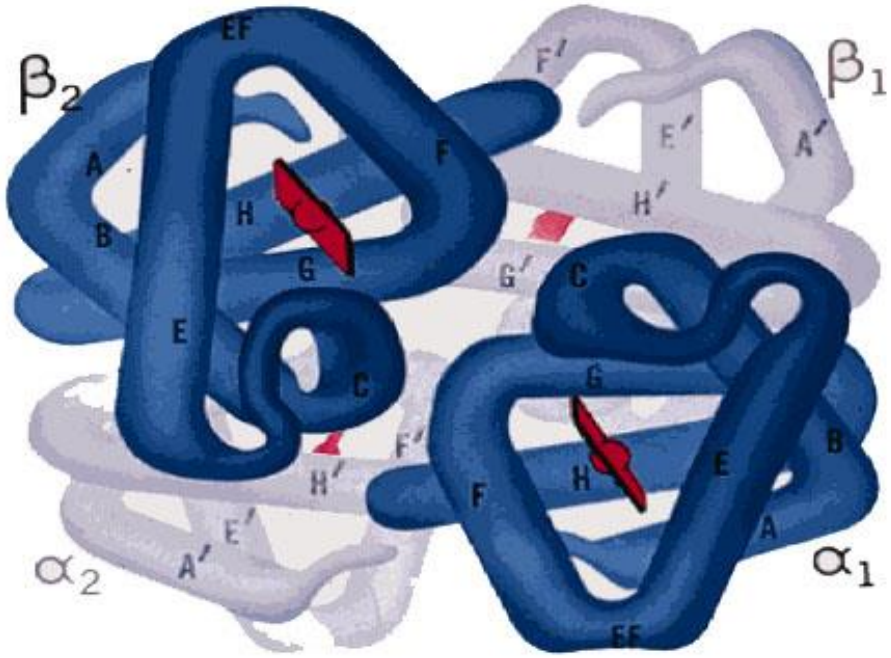
1964: F. Vogel (Heidelberg)

-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

A génfogalom fejlődése: mit nevezünk ma egy génnek?

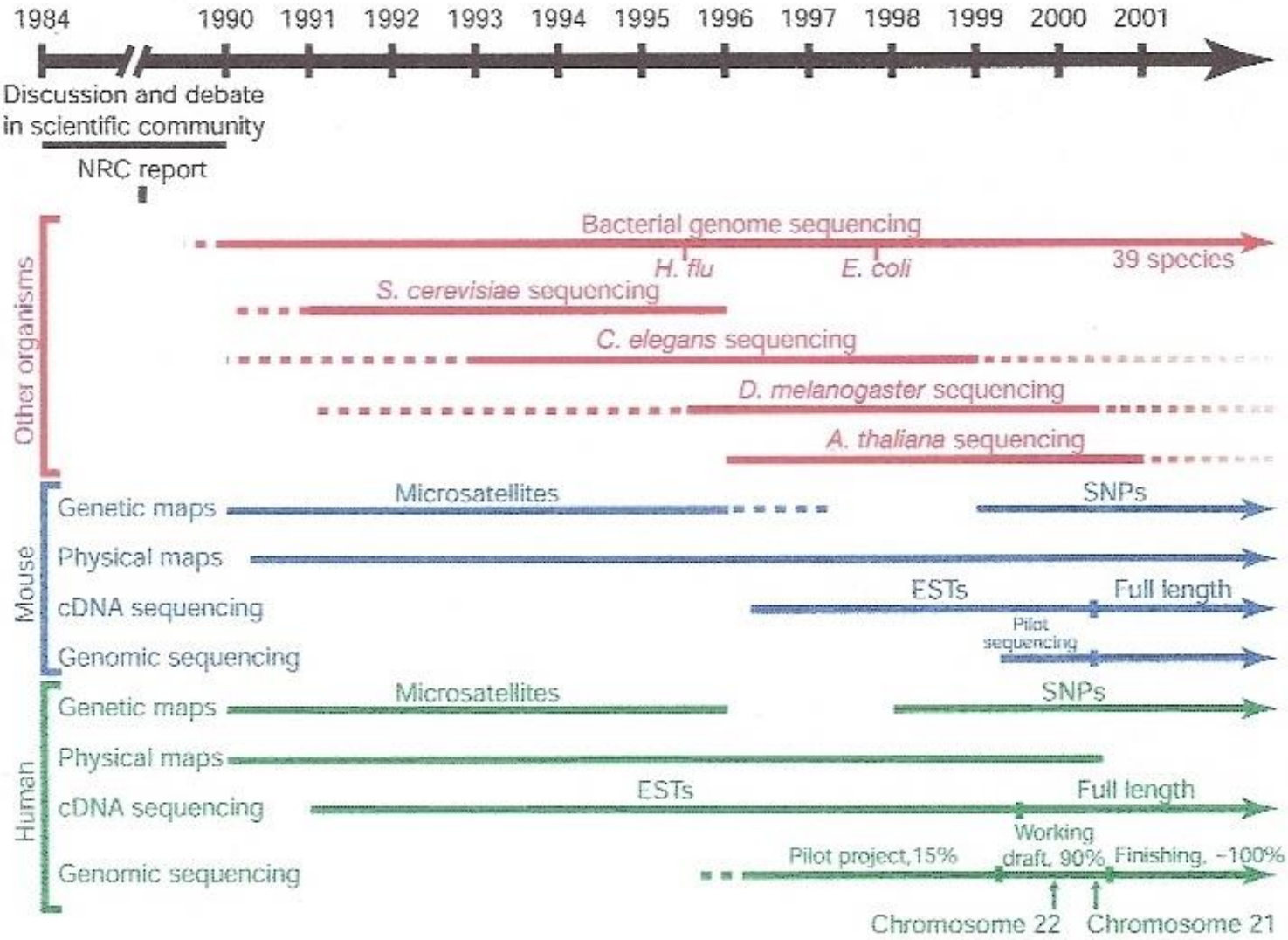
A Gén fogalmának jelentős átalakulása az elmúlt száz esztendőben: protein/RNS kódolás, intron/exon fogalom, szabályozó funkciók, stb.

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 Projekt

- **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

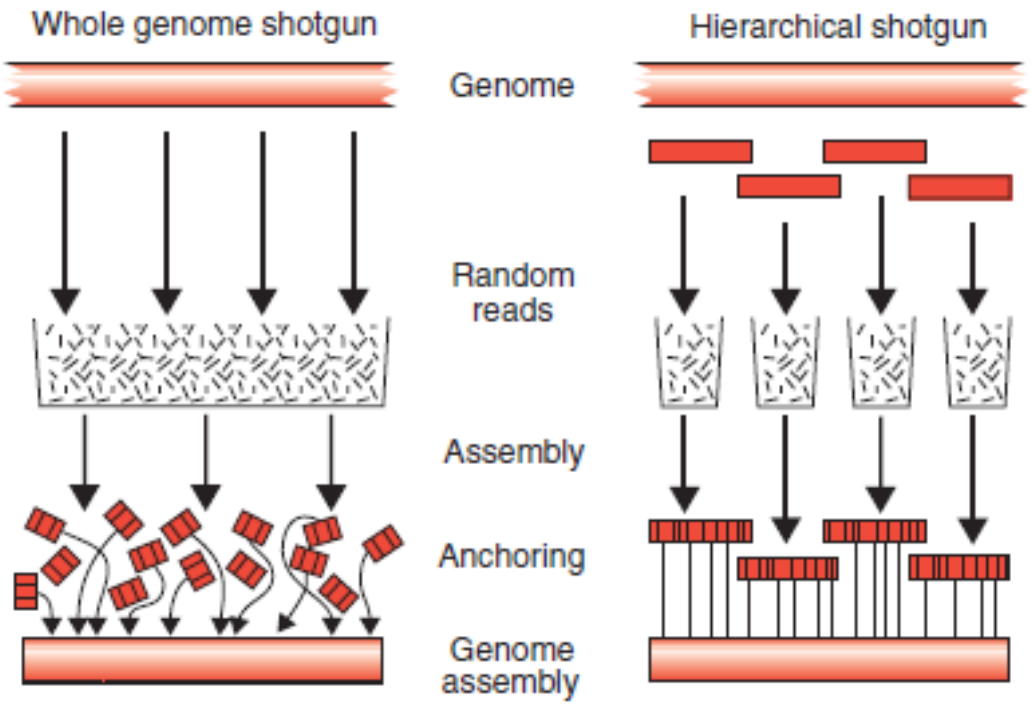


Figure 9.11. Assembling genomic data using the hierarchical and whole genome shotgun approaches. Adapted from Waterston, Lander and Sulston (2002), with permission

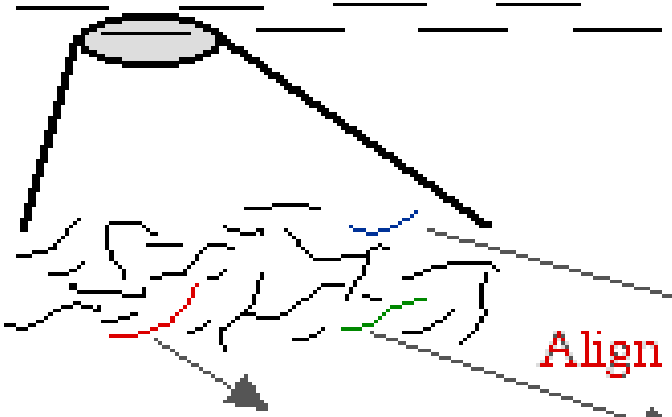
RJ Reece: Analysis of Genes and Genomes, 2004

Hierarchical Shotgun Sequencing Method

Genomic DNA

BAC Library

Create Contig Map



Sequence Each Contig
with Shotgun Approach

Align Contiguous Sequences

GCATTTCGAGTTACCTGGACAACCAGTG

CCAGTGGTACTGAGGACGCAAGAGGCTTGA

GCTTGATTGGCCAATAATAGTATAT

GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT

Generate Finished Sequence

Common gene components

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

- ***Hin*dIII and *Bam*HI**: 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.

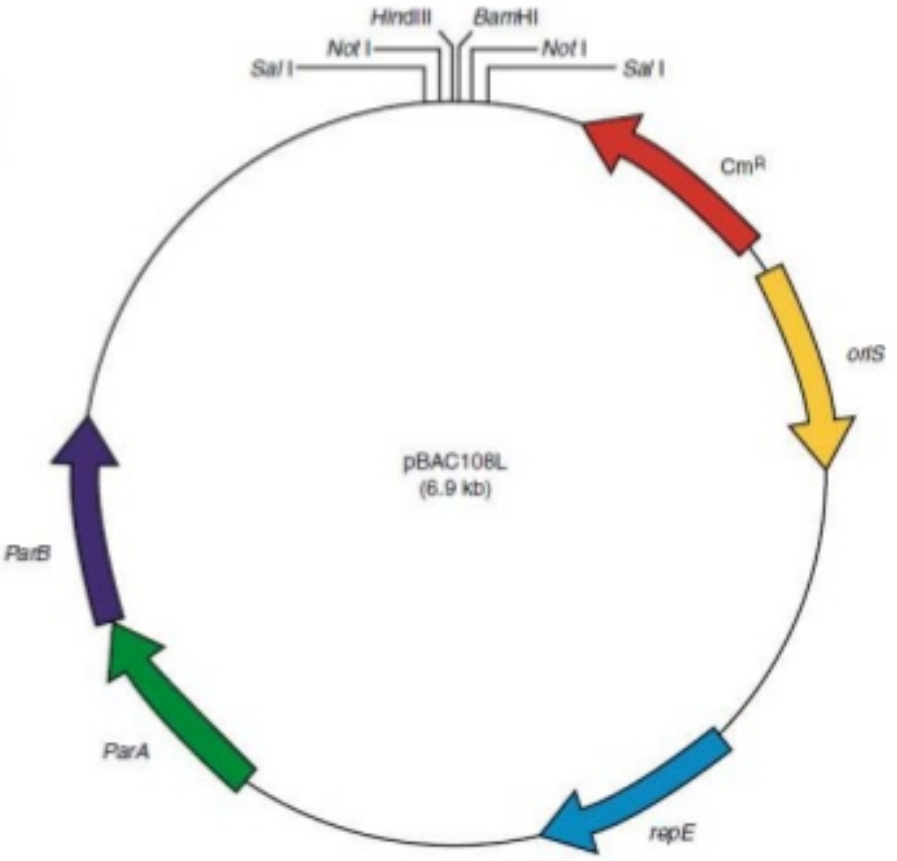


Fig: Map of the BAC vector, pBAC108L

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.

Genomok fizikai térképezése

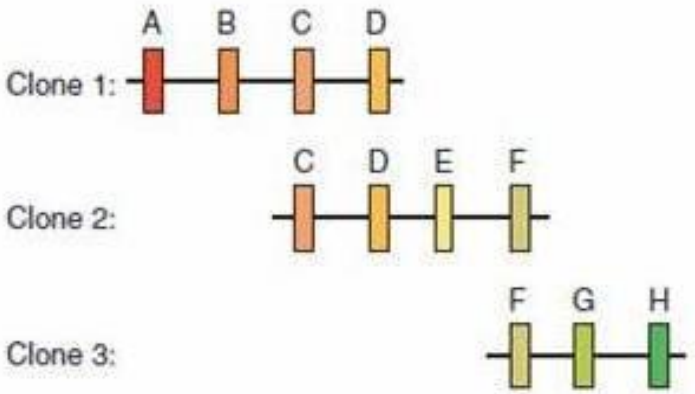


Figure 9.5. Aligning clones by STS mapping. Each clone contains several STSs. Clone 1 has four (A, B, C and D). Clone 2 also contains STSs C and D. Therefore clones 1 and 2 overlap with each other

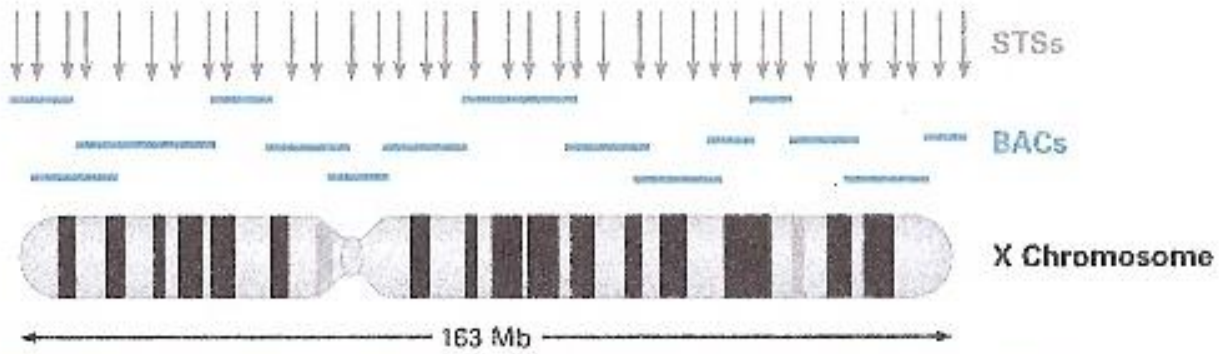


FIGURE 1.3 • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout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

Teljes genom összeszerelés

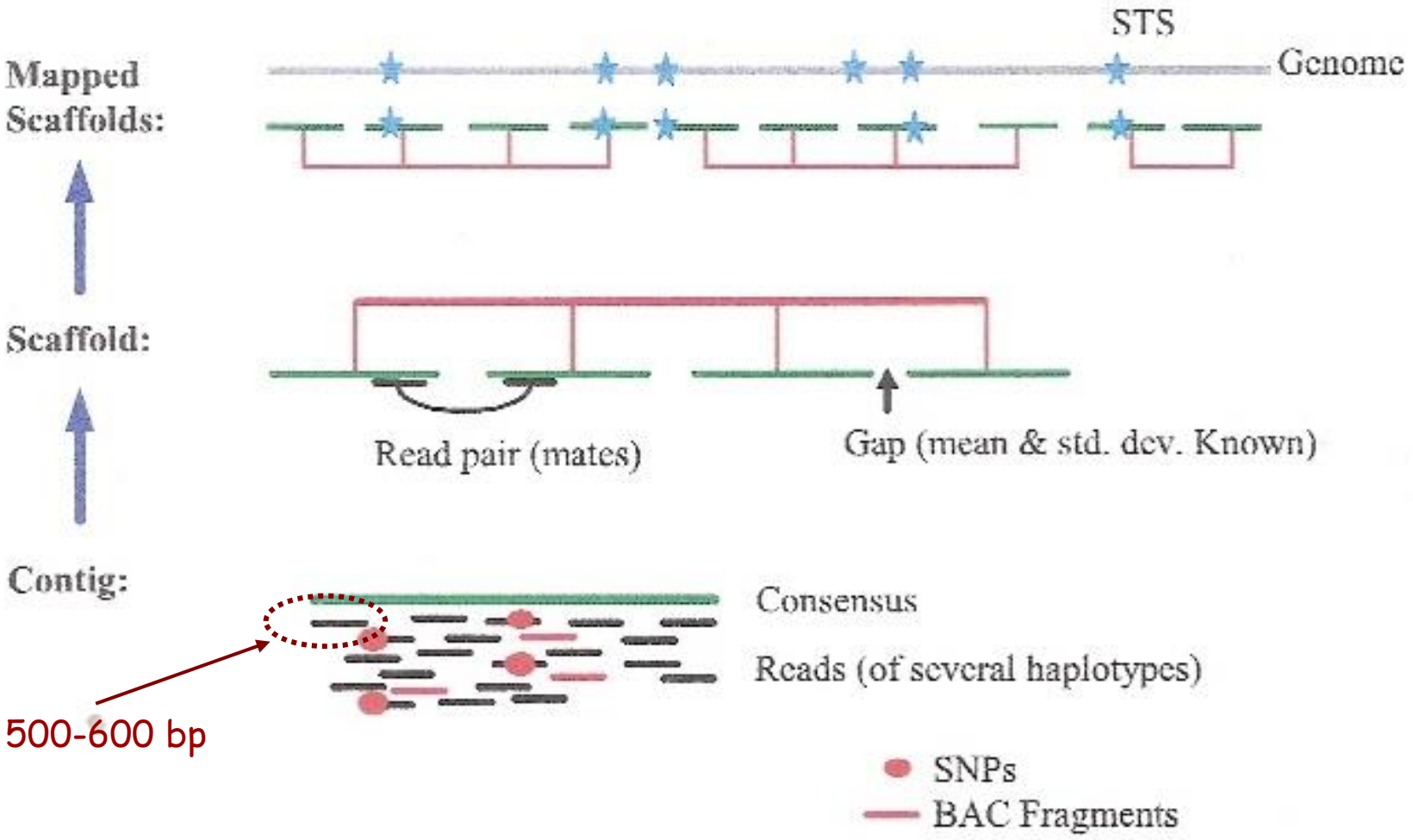
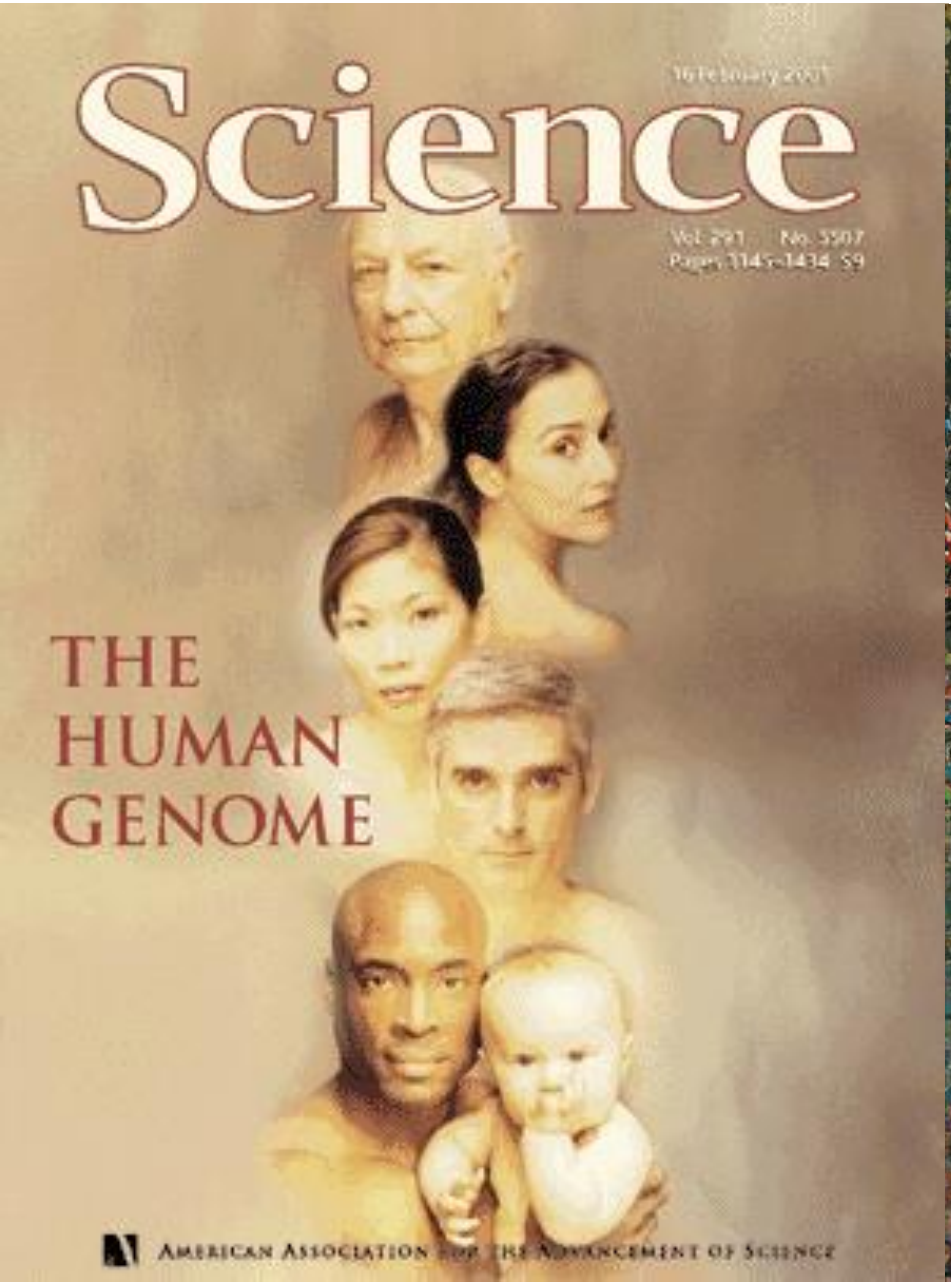
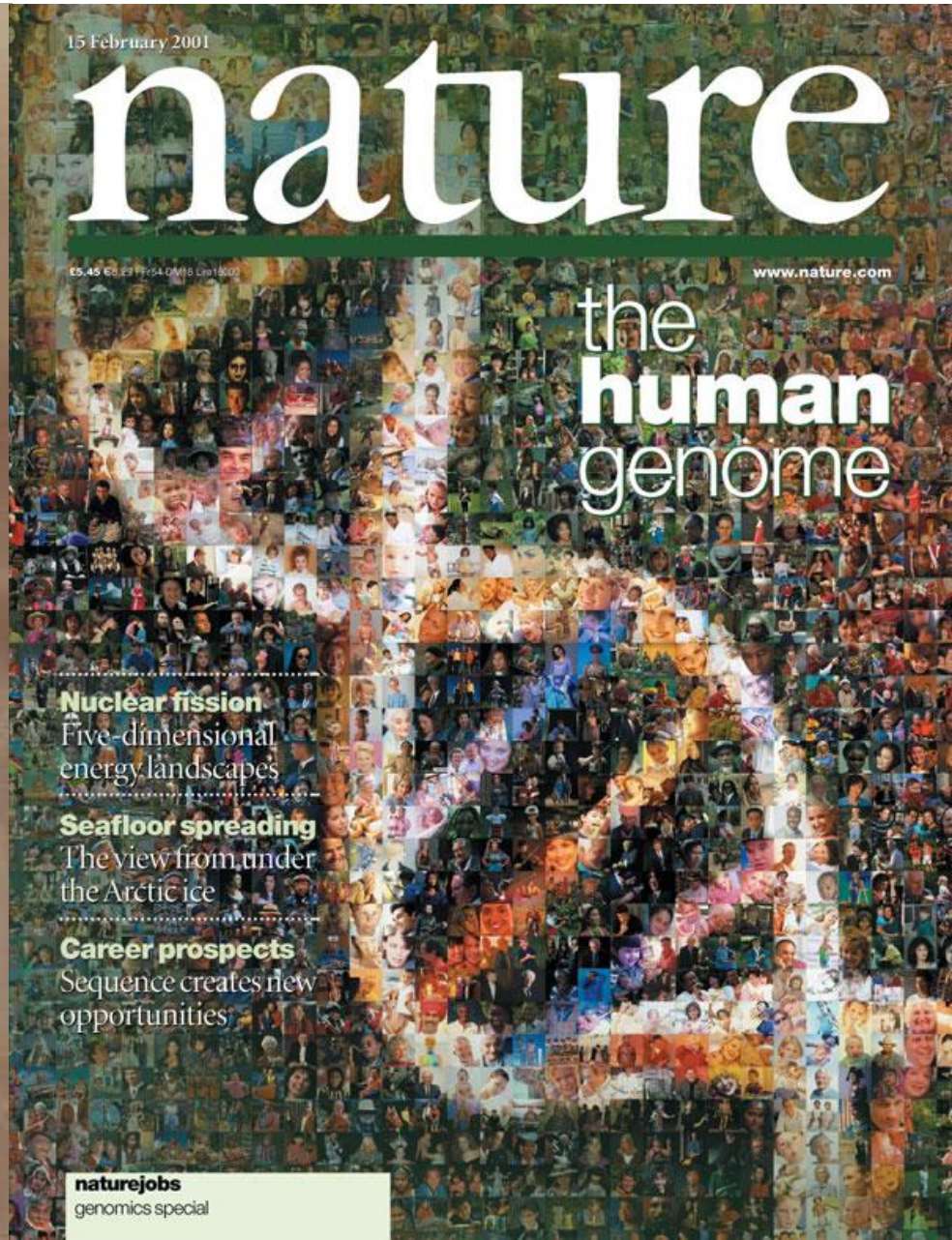


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair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

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

Humán Genom Projekt





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: 19 093 genes > 1 CCDS ID: 7 869

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)

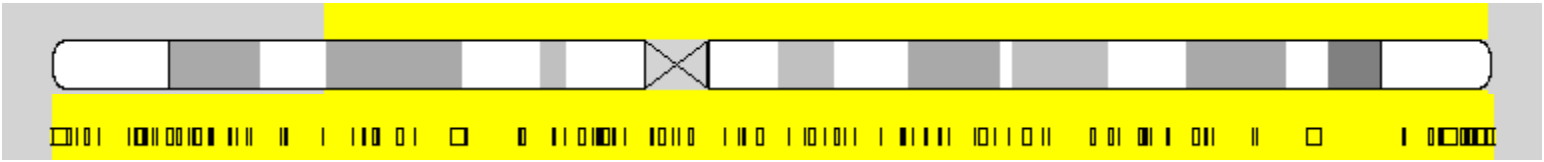
Copy Number Variation (CNV)

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 keletkezése

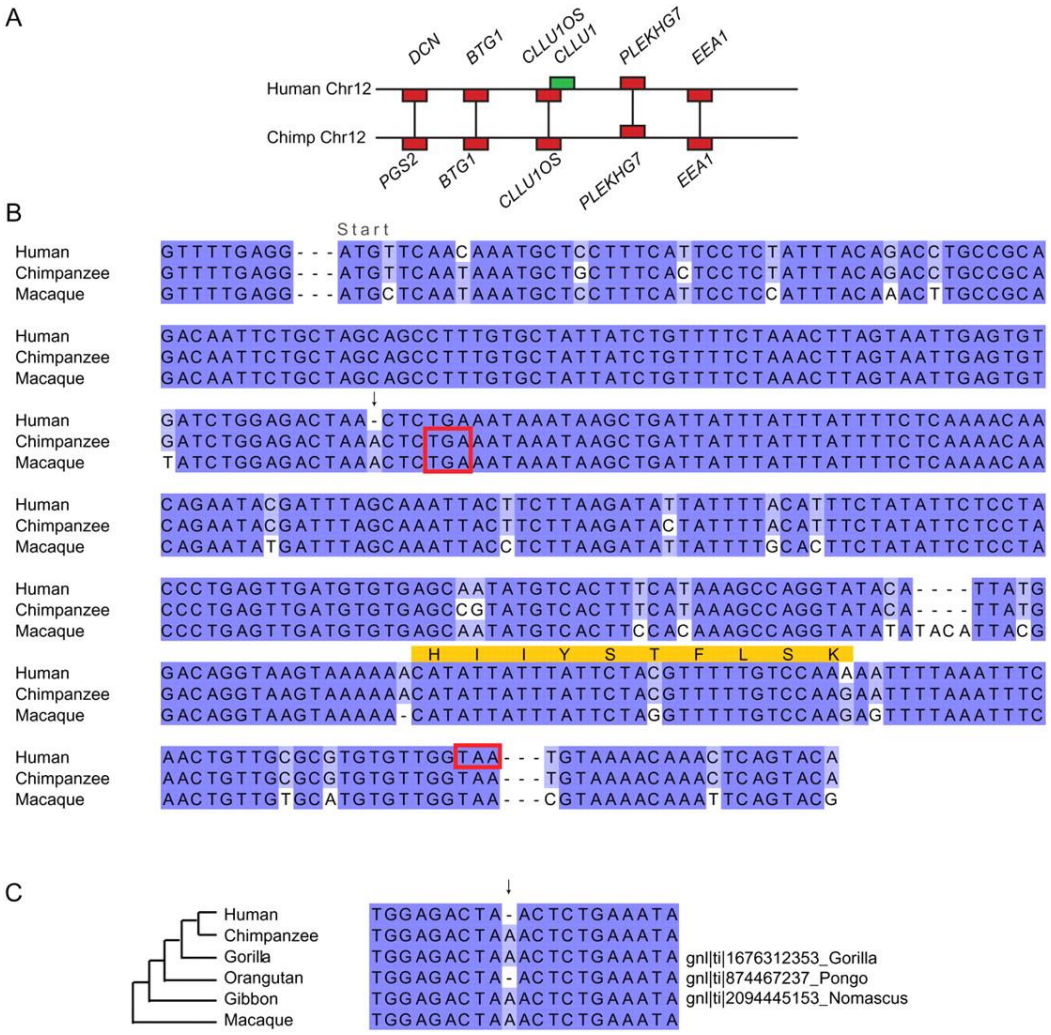
Table 1. Novel human protein-coding genes and supporting evidence.

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

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.

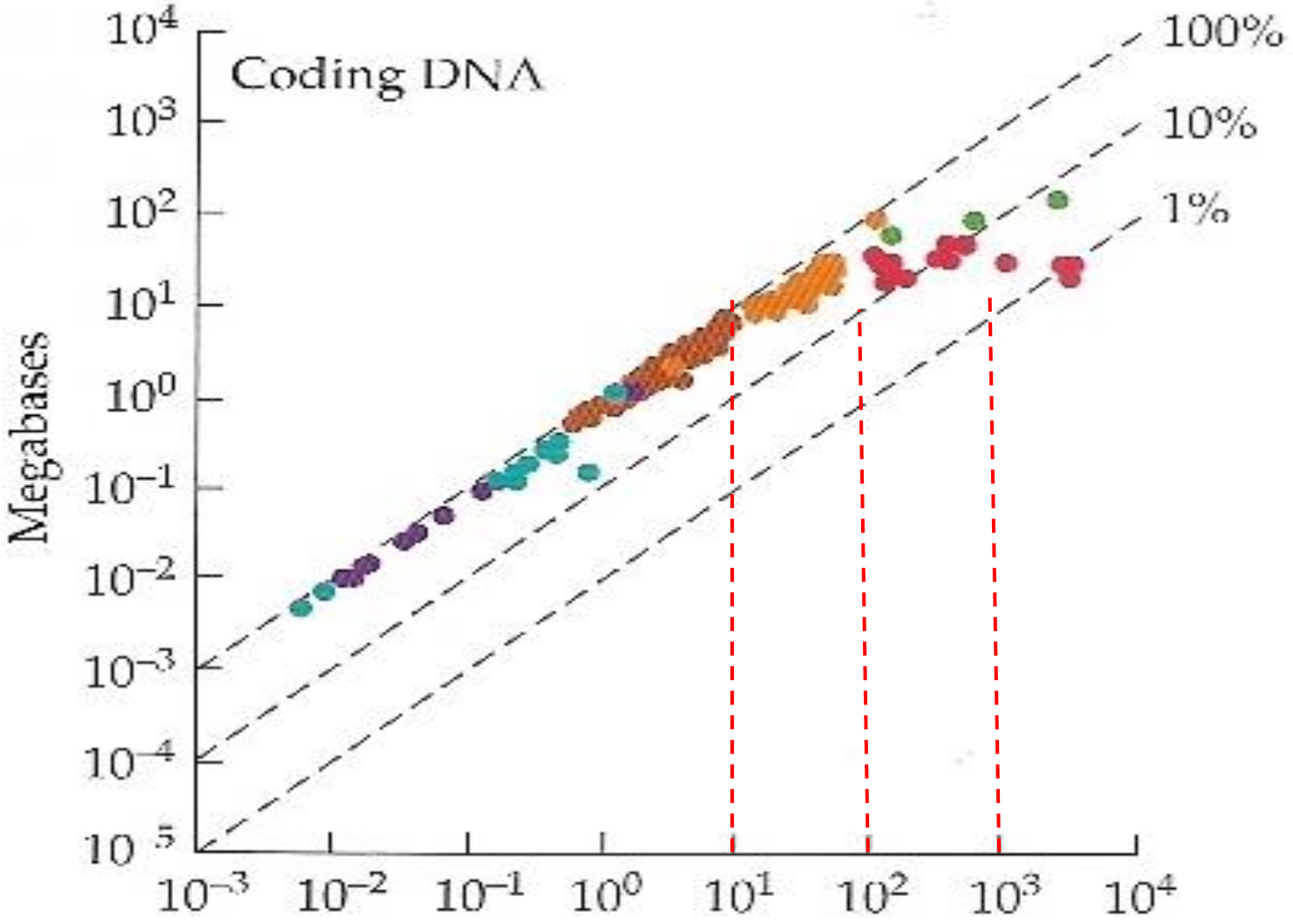
Kódoló szekvenciák hossza

Genom méret

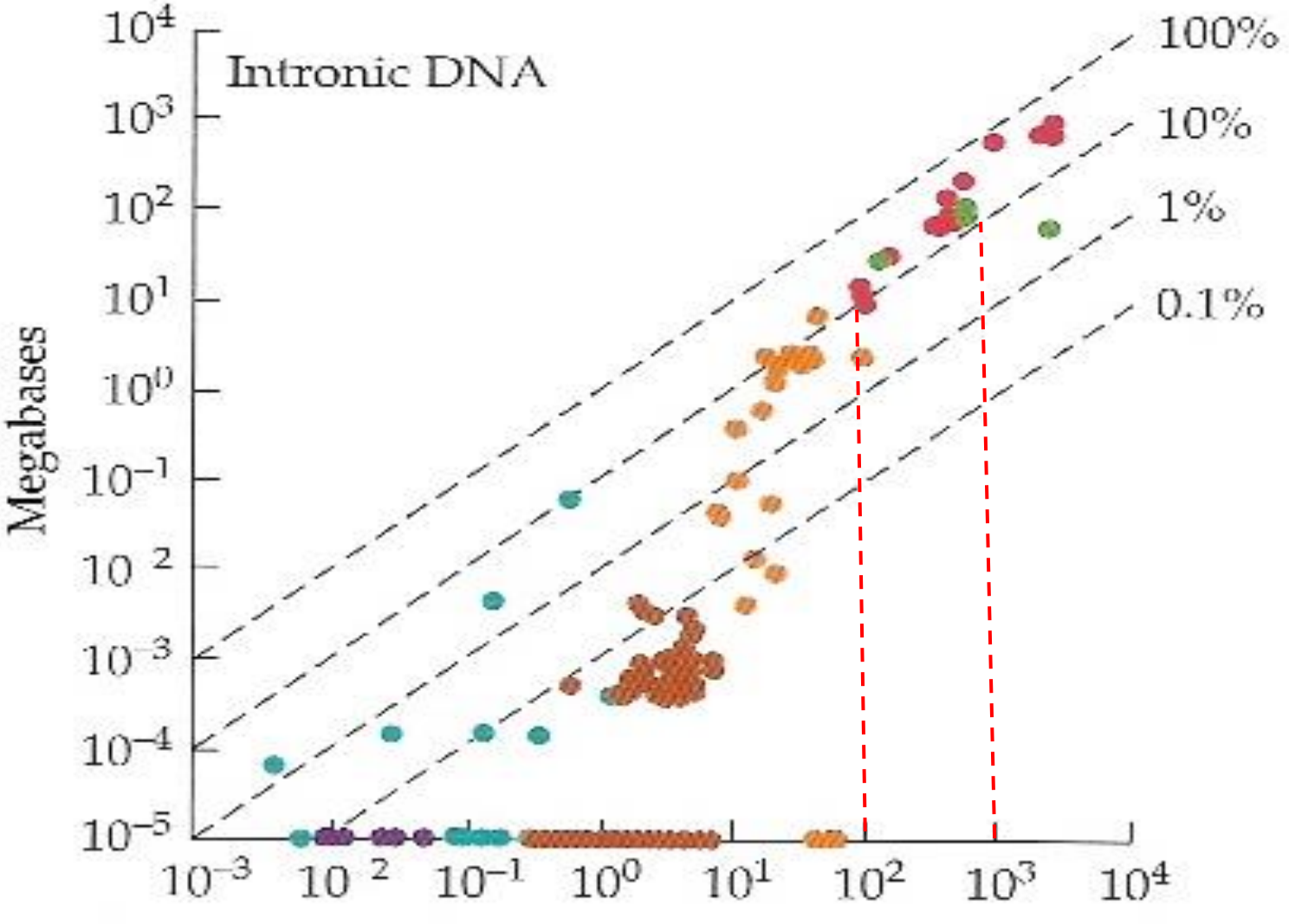
vs.

Nem-kódoló szekvenciák hossza

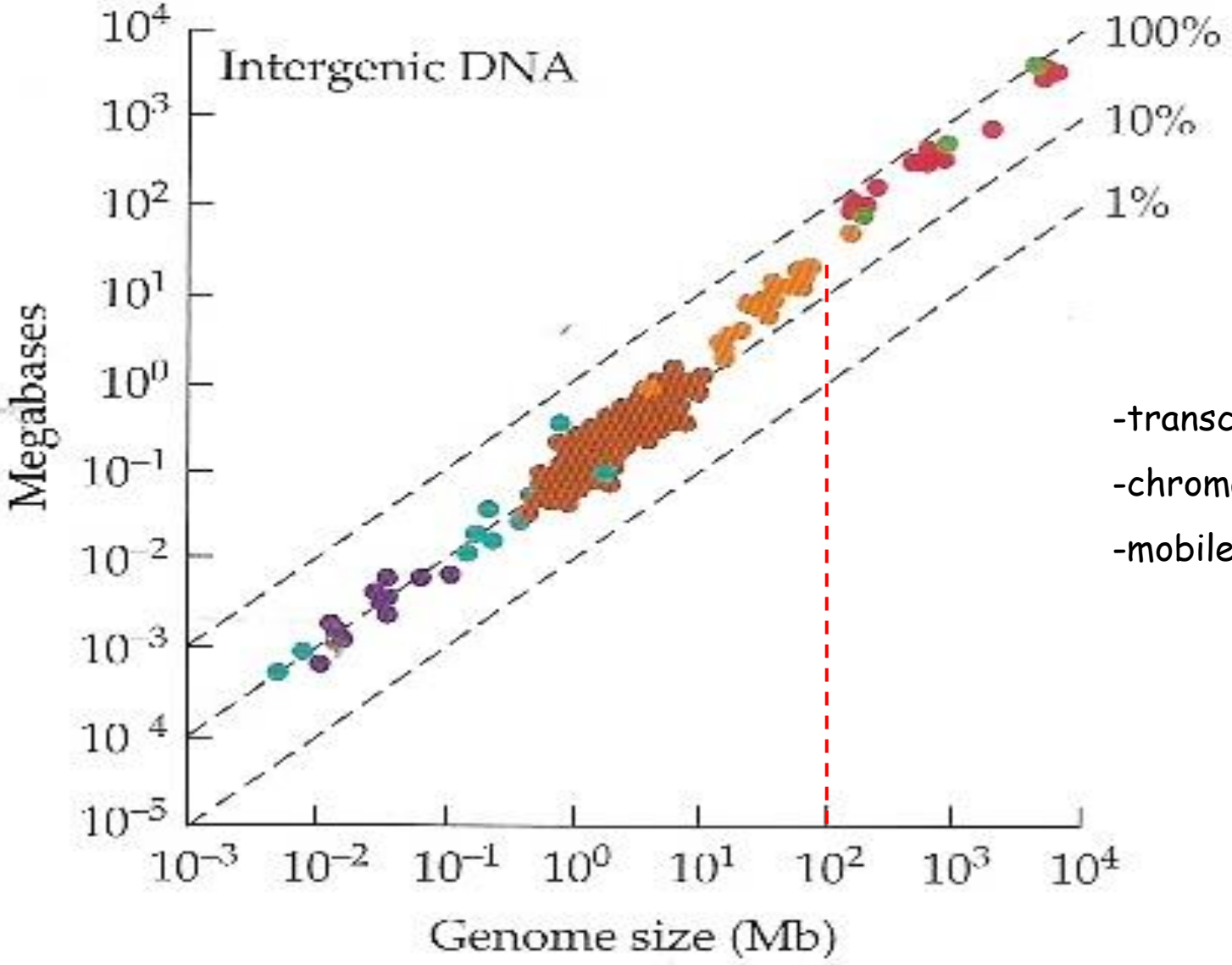
Genom méret vs. kódoló szekvenciák



Genom méret vs. intronok



Genom méret vs. intergénikus DNS



-transcription sites

-chromatin package

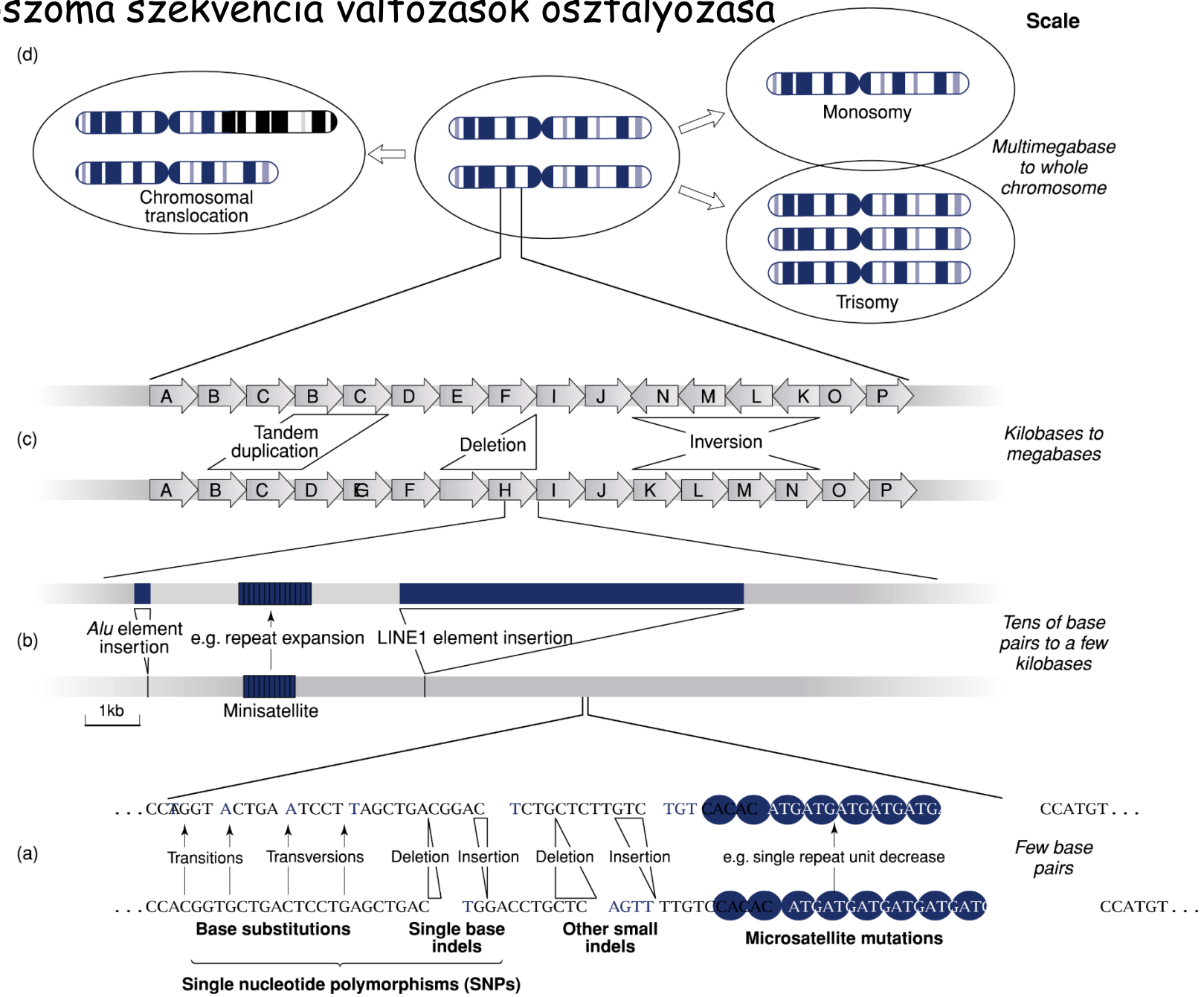
-mobile elements

Genom méret és szervezeti komplexitás

- **C-value paradoxon:** haploid genom méret/sejt – nincs összefüggés a szervezeti 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 / szervezeti komplexitás

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

Kromoszóma szekvencia változások osztályozása



Human Evolutionary Genetics, Jobling, 2004

DNS-molekula szekvencia-variabilitás: polimorfizmusok

(A) Pontmutációk: szubsztitúció, inzerció/deléció → szekvencia polimorfizmus → single nucleotide polymorphism

--------AGACTAGACATT--------

--------AGATTAG-CATT---------

SNP- 1 bp-t érintő változás

(B) Szekvencia ismétlődések: szatellit DNS, repetitív szekvenciák → hosszpolimorfizmus → short tandem repeats

-----(AATG)(AATG)(AATG)-----

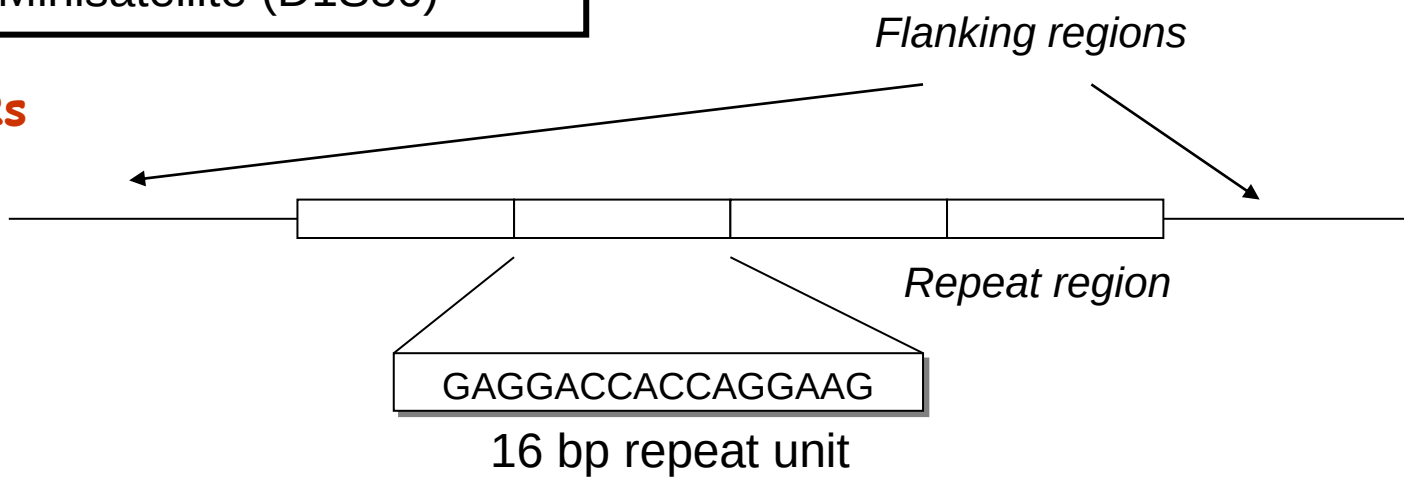
-----(AATG)(AATG)----------

STR/VNTR

Mikro- és miniszatelliták

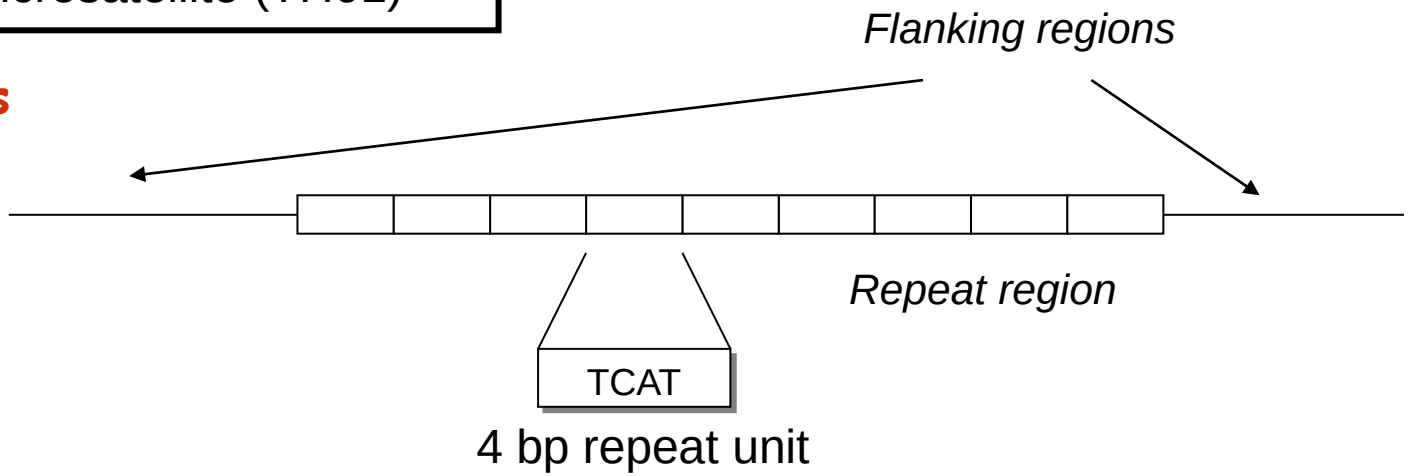
Minisatellite (D1S80)

VNTRs

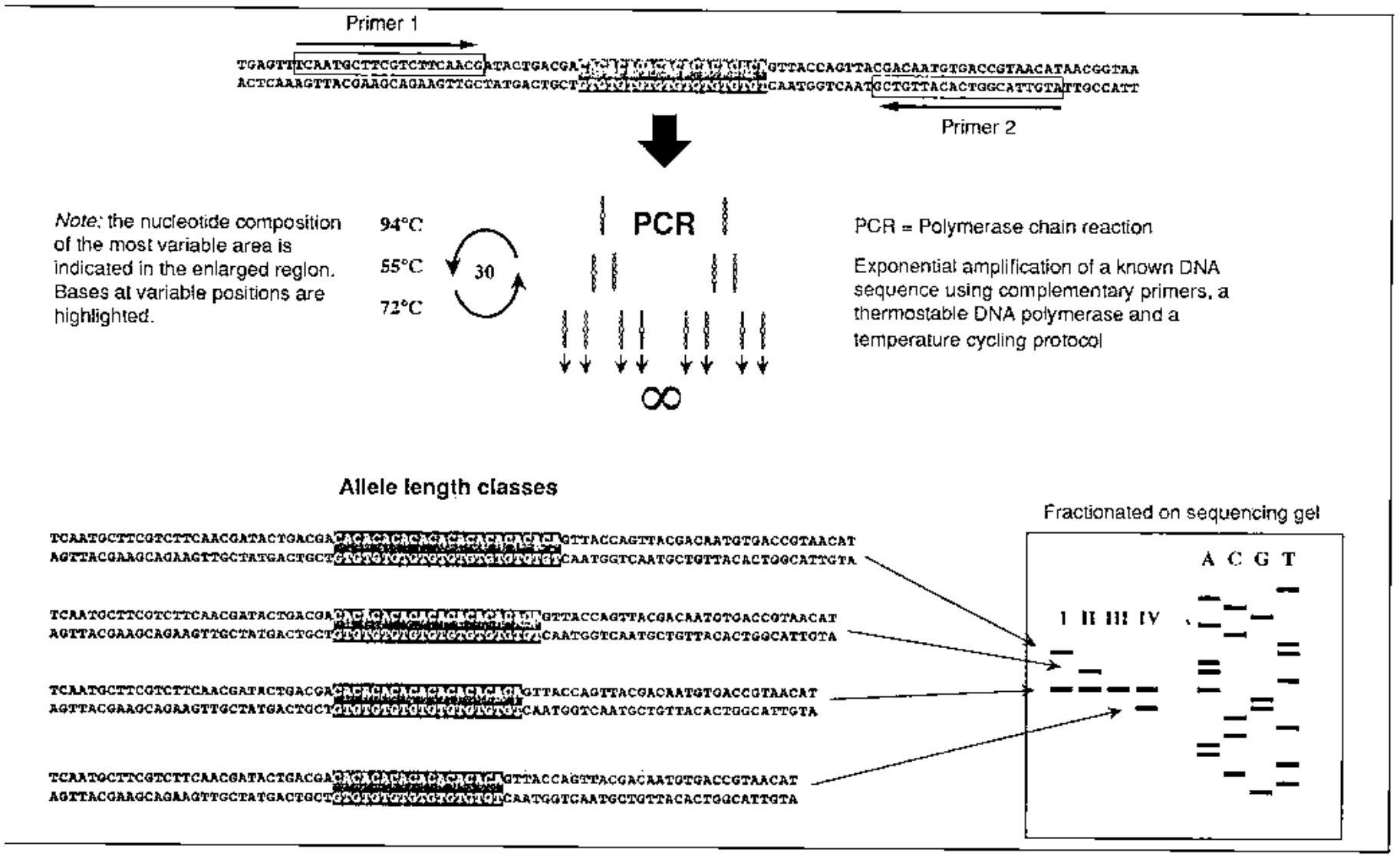


Microsatellite (TH01)

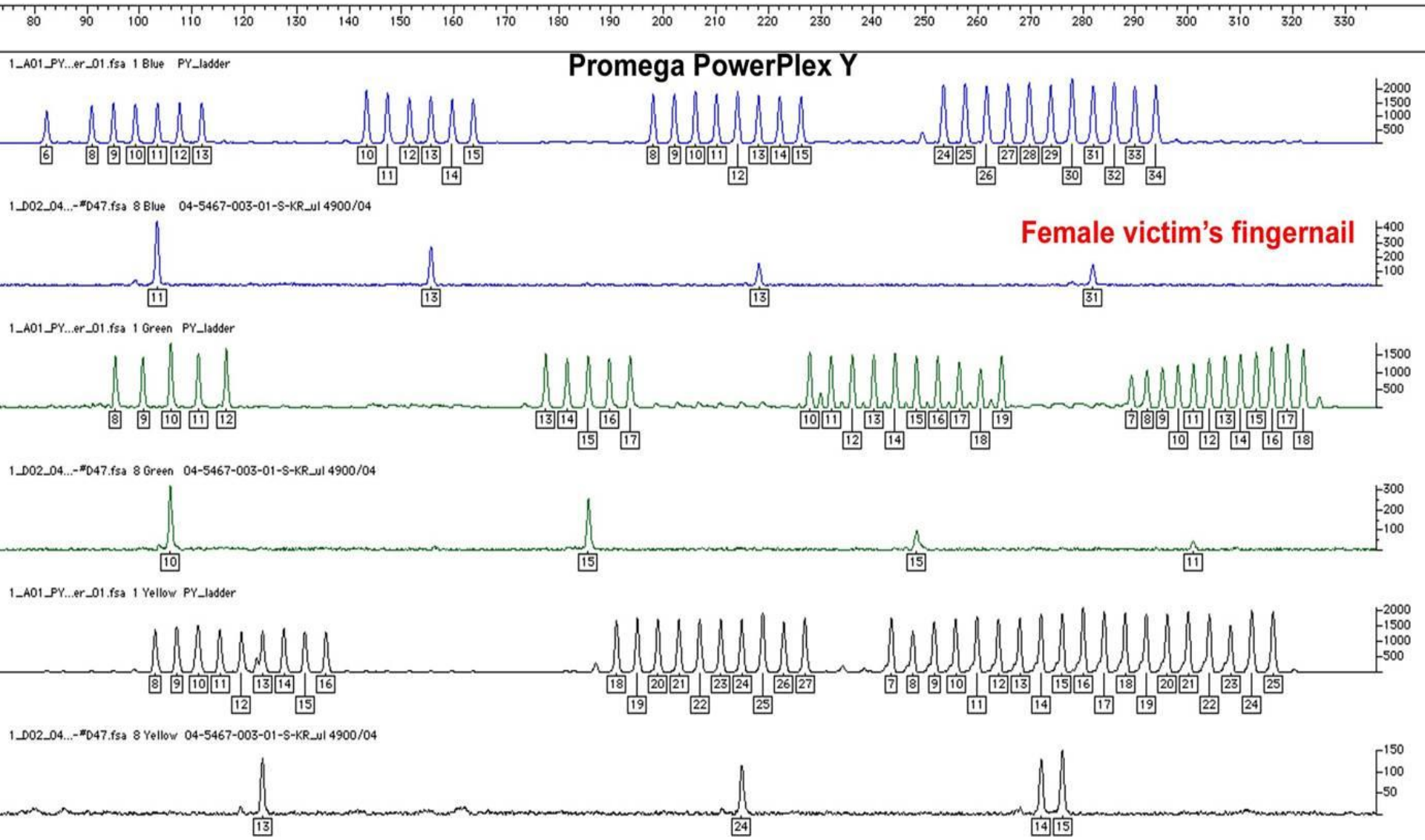
STRs



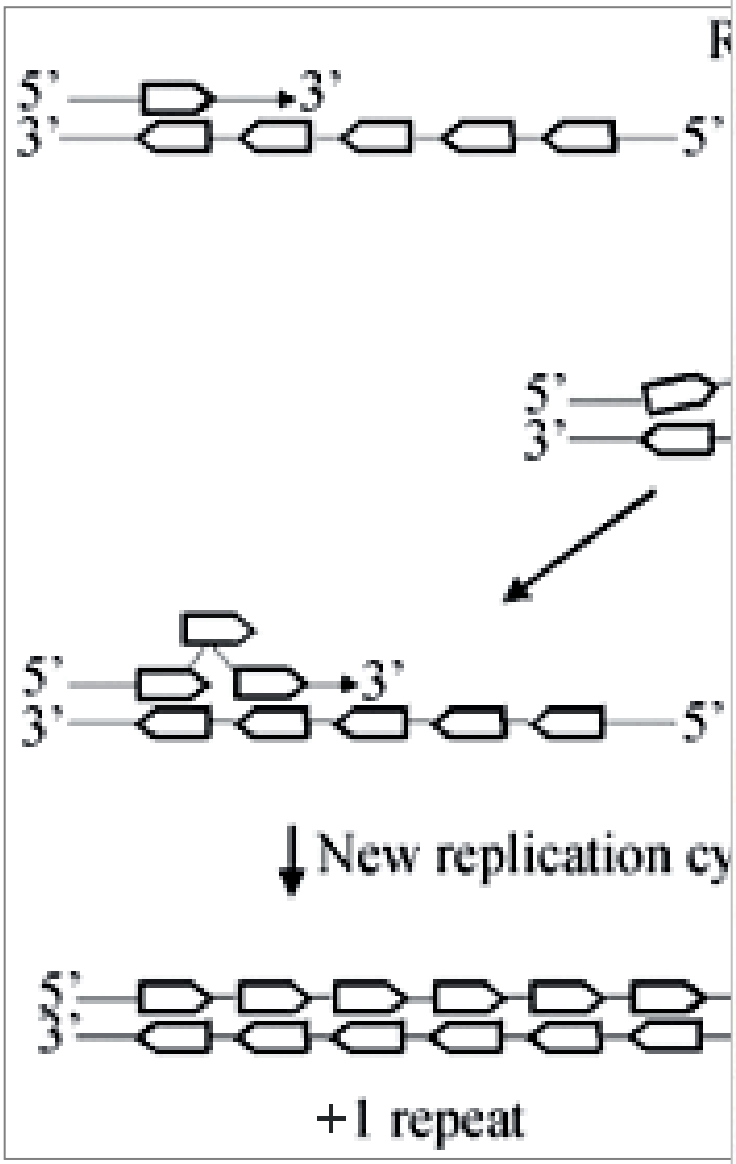
Mikroszatellita genotipizálás PCR-rel

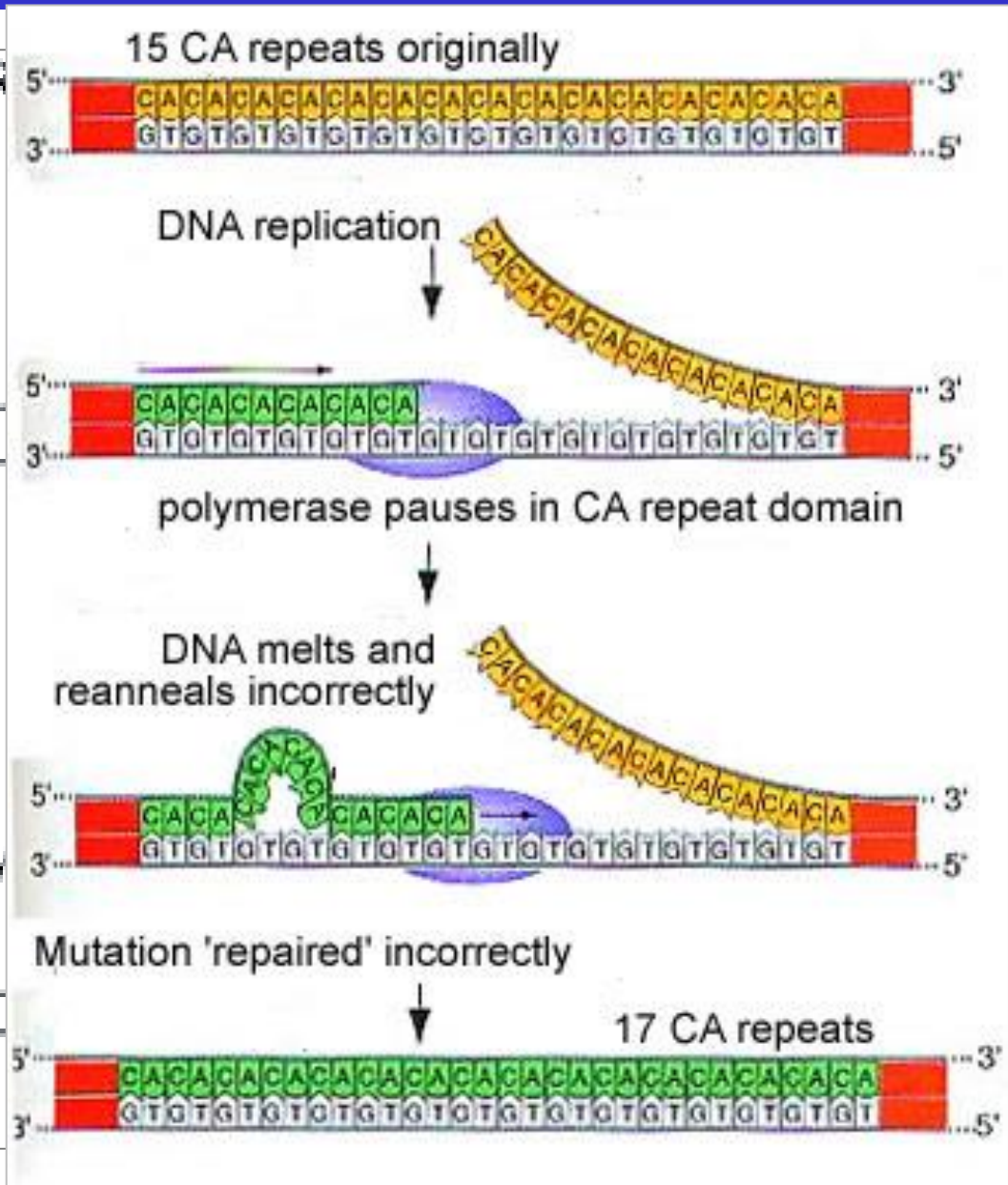


Y kromoszómás mikroszatellita genetikai profil

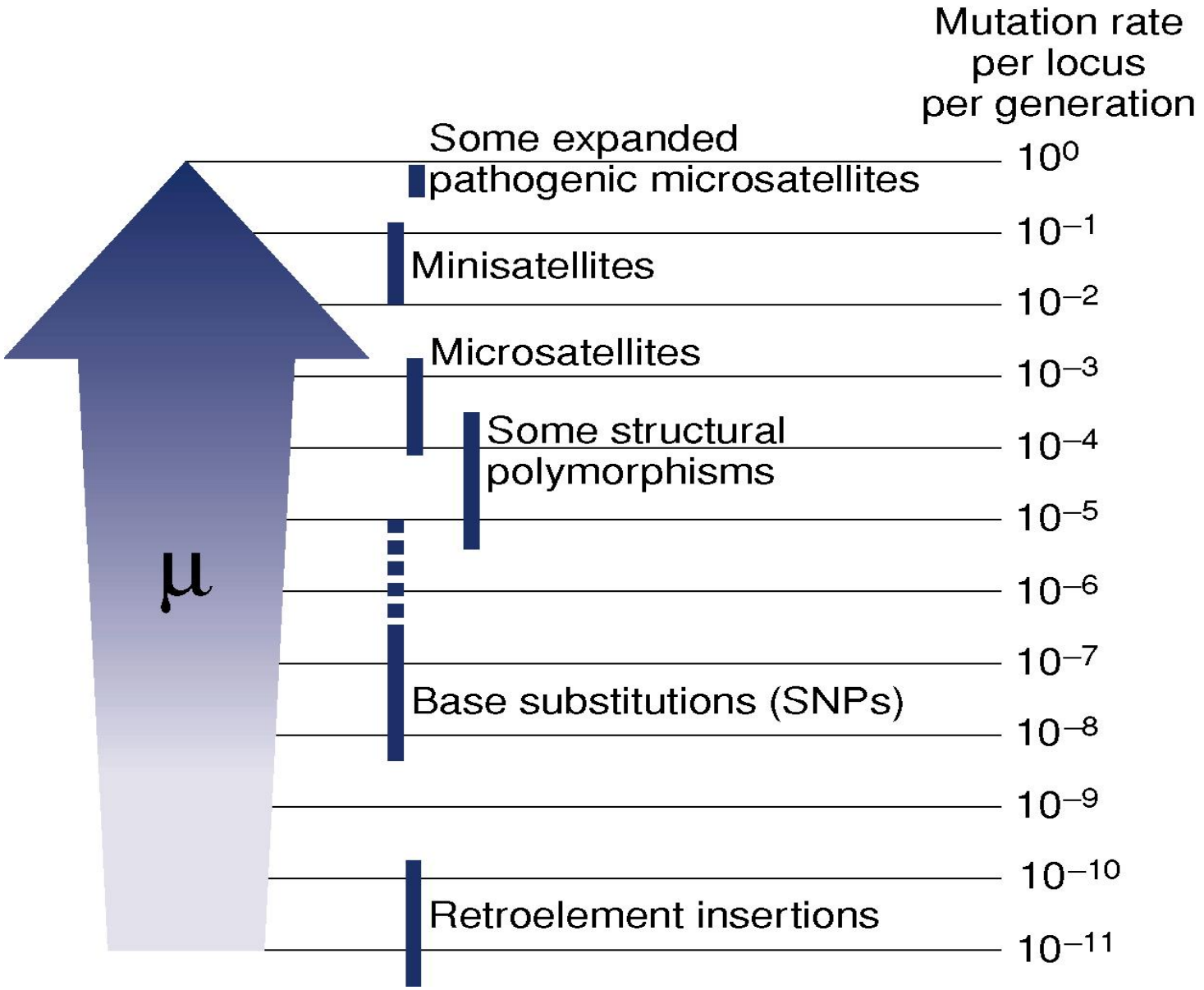


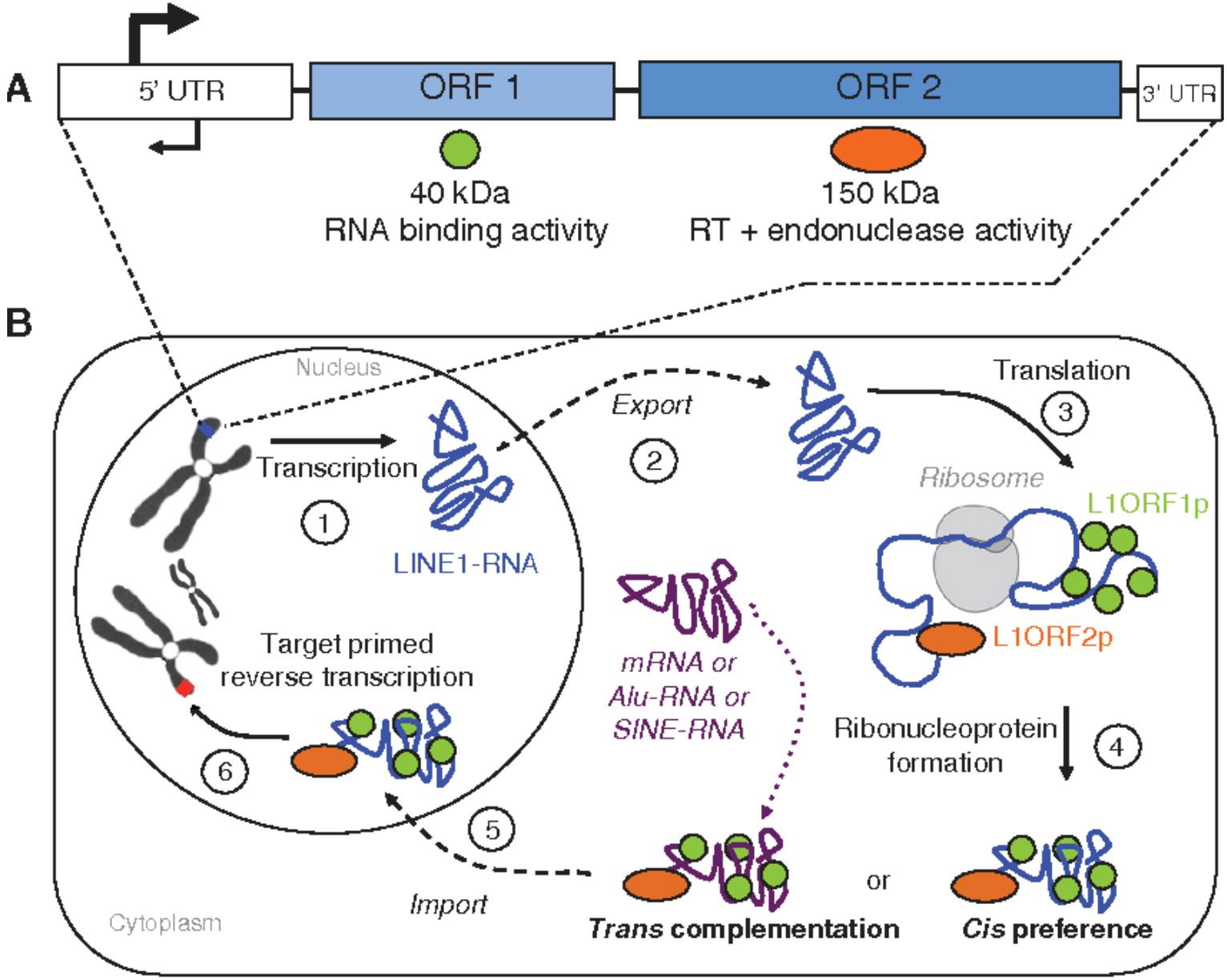
Mikroszatellita evolúció





Polimorf genomi szekvenciák mutációs rátája (μ)

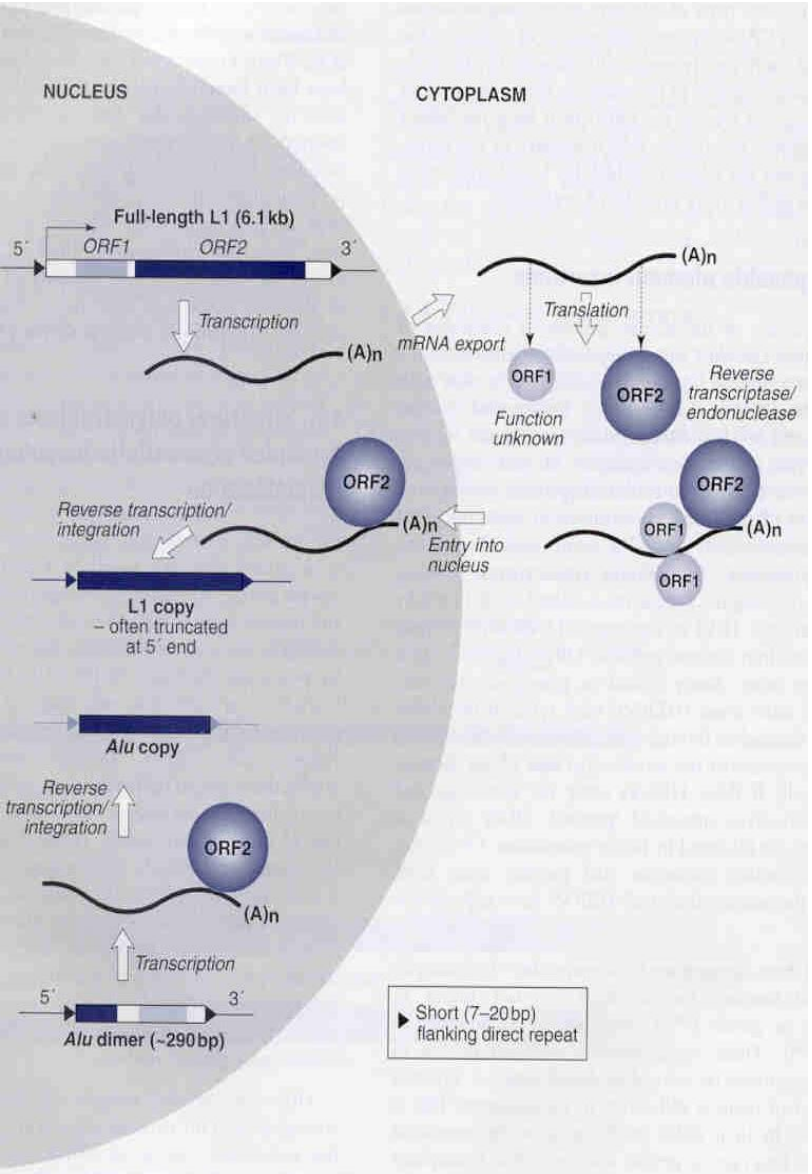




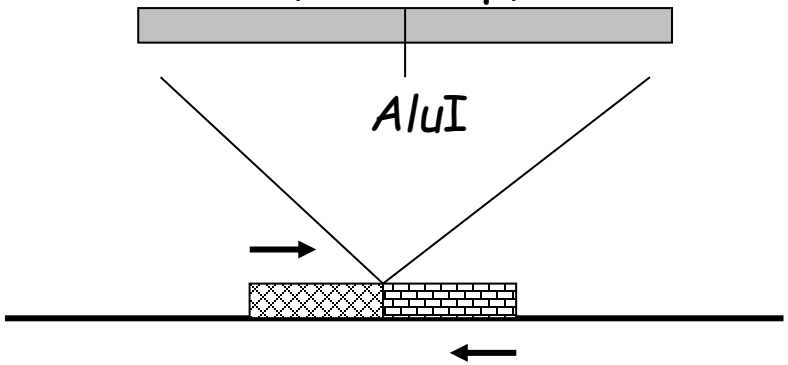
Regulation of LINE-1 in mammals
Maxime Bodak / Jian Yu / Constance Ciaudo
Published Online: 2014-09-25 | DOI: https://doi.org/10.1515/bmc-2014-0018

Mobilis elemek: biallélikus polimorfizmus a DNS-ben

LINE és SINE elemek: long ill. short beépült szekvenciák



Human *Alu* Repeat (~300 bp)



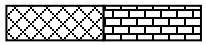
Restrikciós endonukleáz

Kétféle alléltípus

"long" (+) allél

400 bp

"short" (-) allél



100 bp

Next Generation Sequencing –

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

-Process millions of sequence reads in parallel

-Library preparation

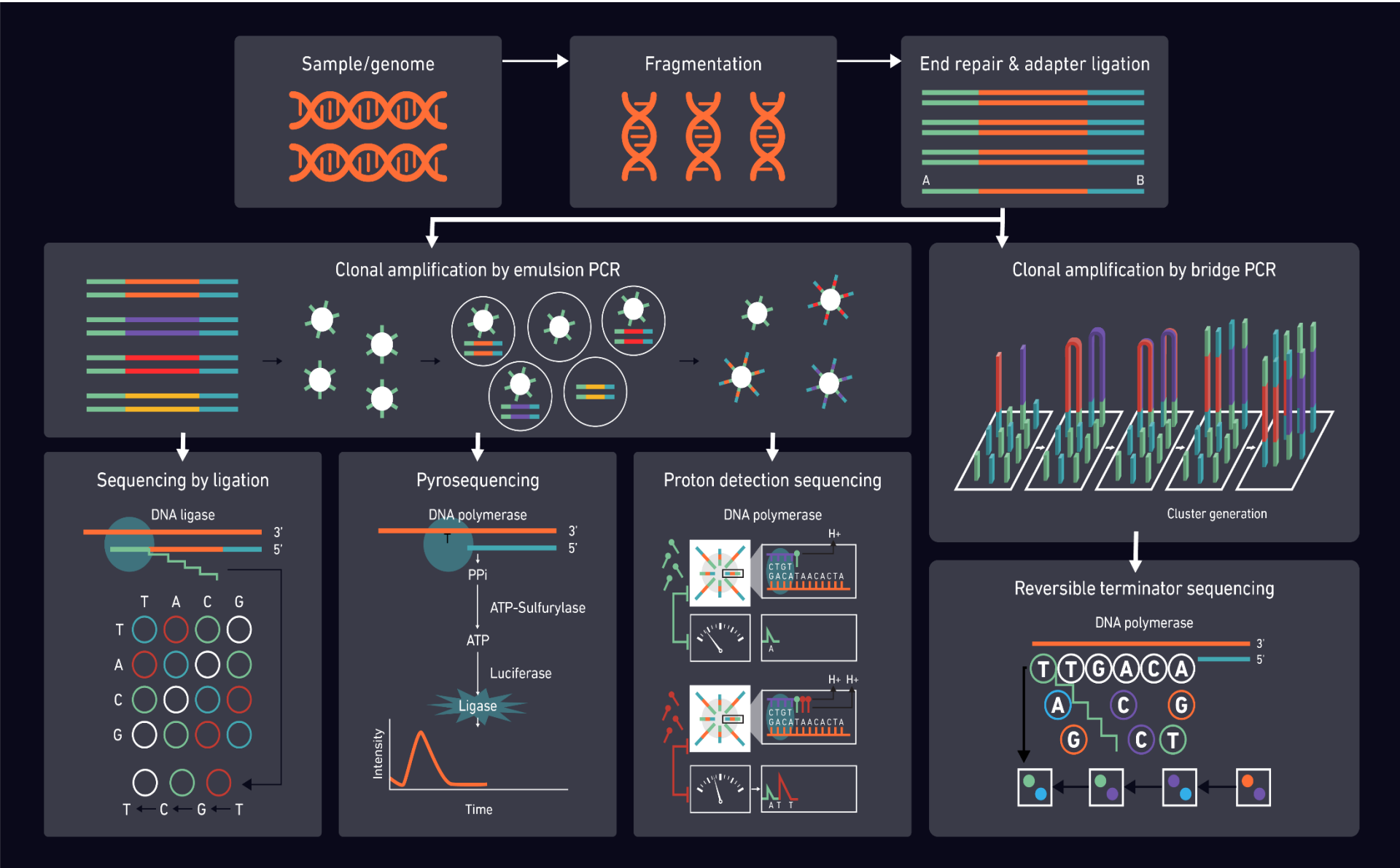
-Specific adaptor oligos

-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



ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

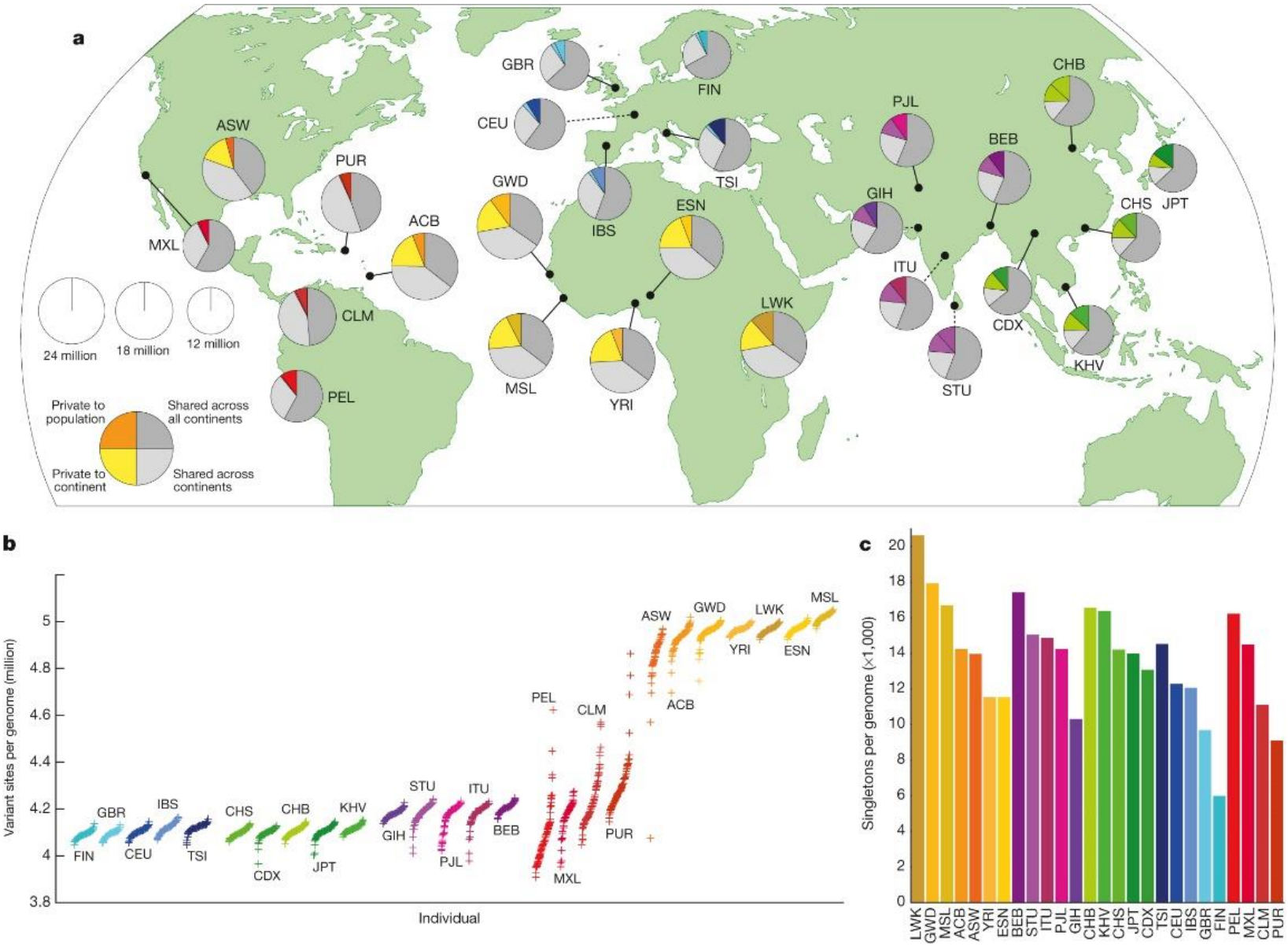
The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 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



nature

	Autosomes	Exome target regions**	chrX***	chrY***	Totals
Samples	2,504	2,504	2,504	1,233	-
Total Raw Bases (Gb)	85,426	18,273	3,213	291	-
Mean Mapped Depth (X)*	8.45	75.25	6.20	2.60	-
Total Variant Sites	84,801,880	1,416,049	3,468,093	62,042	88,332,015
Biallelic SNPs	81,102,777	1,383,927	3,223,927	60,505	84,387,209
Indels	3,196,364	19,832	212,196	1,427	3,409,987
Mean Indel Length (bp)	2.94	3.46	2.64	2.00	-
Multiallelic sites	444,026	6,153	30,996	-	475,022
Multiallelic SNPs	274,425	4,706	15,055	-	289,480
Multiallelic Indels	169,601	1,447	15,941	-	185,542
Structural Variants	58,713	6,137	974	110	59,797
ALU Insertion	12,491	52	-	-	12,491
LINE1 Insertion	2,910	10	-	-	2,910
Large Deletion	33,336	2,684	974	-	34,310
Duplication	5,896	2,513	-	-	5,896
SVA Insertion	822	5	-	-	822
Other Insertion	165	1	-	-	165
Inversion	100	8	-	-	100
CNV	2,993	864	-	110	3,103

Supplementary Information Table 3: Integrated callset summary. *Assuming 2.84Gb as the genome size. The mapping of exome sequence to targeted pull down regions was calculated by Picard function *calculateHsMetrics*. **The exome targeted regions were exome pulldown targets derived from CCDS (NimbleGen EZ Exome v1 and Agilent SureSelect v2). These variant totals are included in the other columns. ***chrX and chrY statistics are for the entire chromosomes.

- 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 affect more bases:
- typical genome contains an estimated 2,100 to 2,500 structural variants (1,000 large deletions, 160 copy-number variants, 915 Alu insertions, 128 L1 insertions, 51 SVA insertions, 4 NUMTs and 10 inversions) affecting 20 million bases of sequence.

nature Vol 464|8 April 2010|doi:10.1038/nature08976

LETTERS

The complete mitochondrial DNA genome of an unknown hominin from southern Siberia

Johannes Krause[1], Qiaomei Fu[1], Jeffrey M. Good[2], Bence Viola[1,3], Michael V. Shunkov[4], Anatoli P. Derevianko[4] & Svante Pääbo[1]

With the exception of Neanderthals, from which DNA sequences of numerous individuals have now been determined[1], the number and genetic relationships of other hominin lineages are largely unknown. Here we report a complete mitochondrial (mt) DNA sequence retrieved from a bone excavated in 2008 in Denisova Cave in the Altai Mountains in southern Siberia. It represents a hitherto unknown type of hominin mtDNA that shares a common ancestor with anatomically modern human and Neanderthal mtDNAs about 1.0 million years ago. This indicates that it derives from a hominin migration out of Africa distinct from that of the ancestors of Neanderthals and of modern humans. The stratigraphy of the cave where the bone was found suggests that the Denisova hominin lived close in time and space with Neanderthals as well as with modern humans[2–4].

The first hominin group to leave Africa was *Homo erectus* about 1.9 million years (Myr) ago[5]. Archaeological as well as genetic data indicate that at least two groups of hominins left Africa after this event: first, the ancestors of the Neanderthals between 500,000 and 300,000 years ago (500 and 300 kyr ago, respectively), presumably *Homo heidelbergensis* or *Homo rhodesiensis*[6–9]; and, second, anatomically

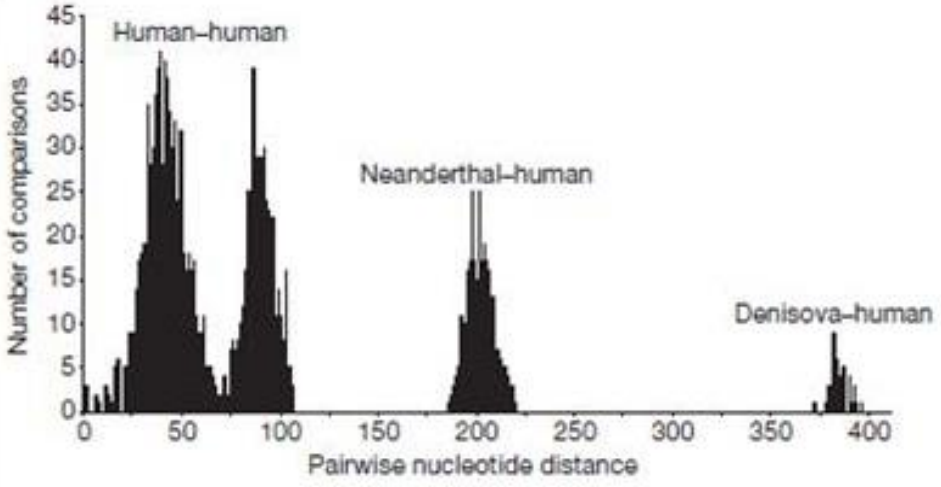


Figure 2 | Distribution of pairwise nucleotide differences. Pairwise nucleotide differences from all pairs of complete mtDNAs from 54 present-day and one Pleistocene modern human, six Neanderthals and the Denisova hominin are shown.

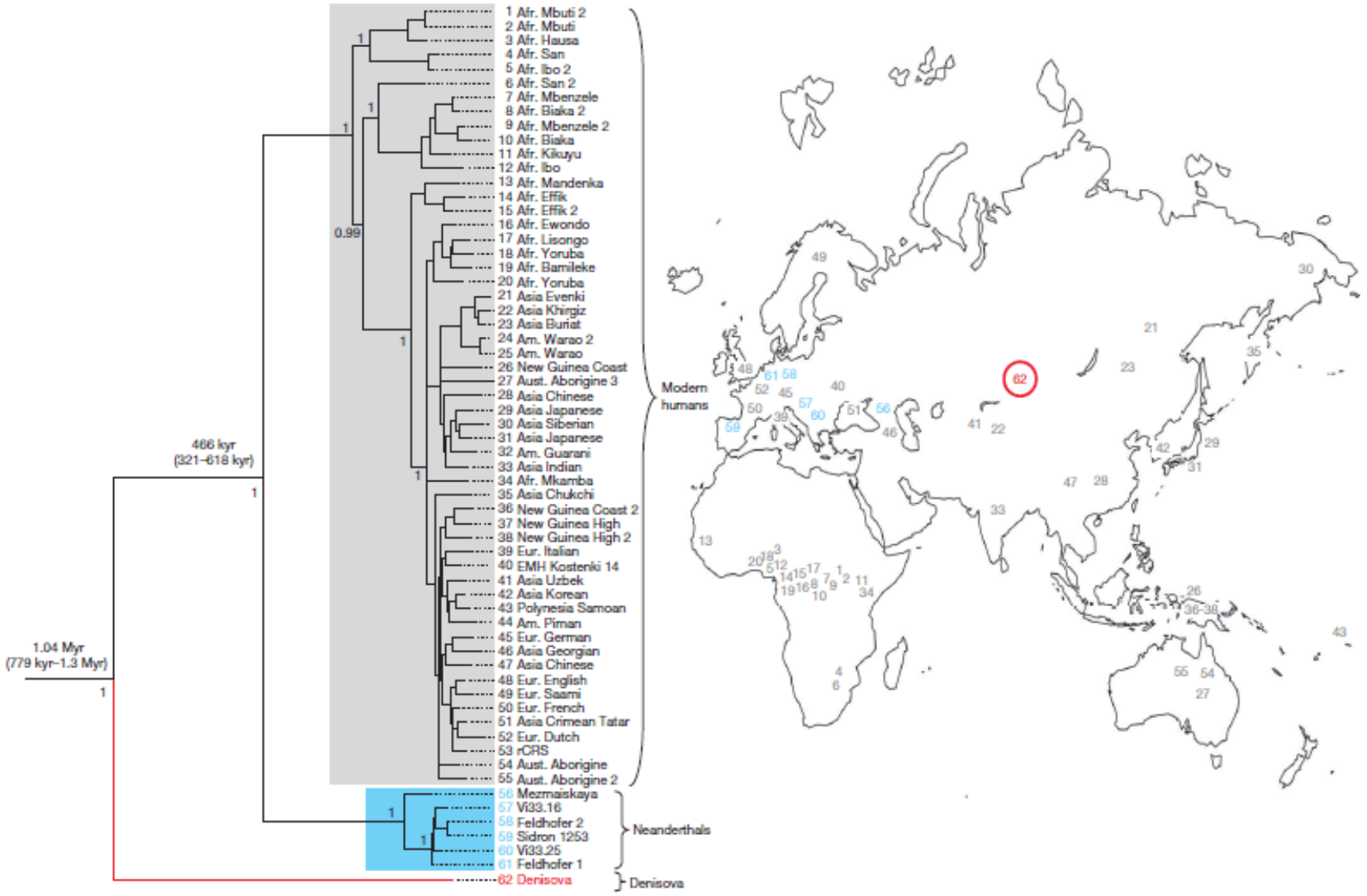
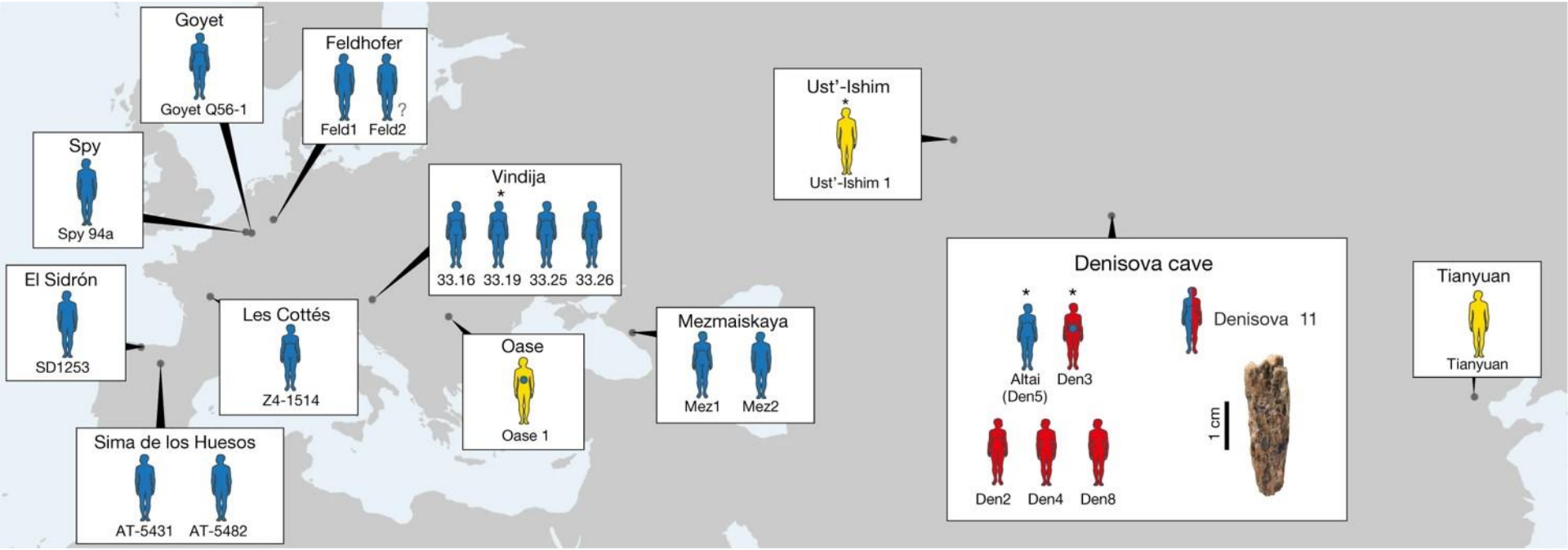


Figure 3 | Phylogenetic tree of complete mtDNAs. The phylogeny was estimated with a Bayesian approach under a GTR+I+Γ model using 54 present-day and one Pleistocene modern human mtDNA (grey), 6 Neanderthals (blue) and the Denisova hominin (red). The tree is rooted with a chimpanzee and a bonobo mtDNA. Posterior probabilities are given for each major node. The map shows the geographical origin of the mtDNAs (24, 25, 32, 44 are in the Americas). Note that two partial mtDNAs sequenced from Teshik Tash and Okladikov Cave in Central Asia fall together with the complete Neanderthal mtDNAs in phylogenies[4] (not shown).

(Krause et al., 2010)

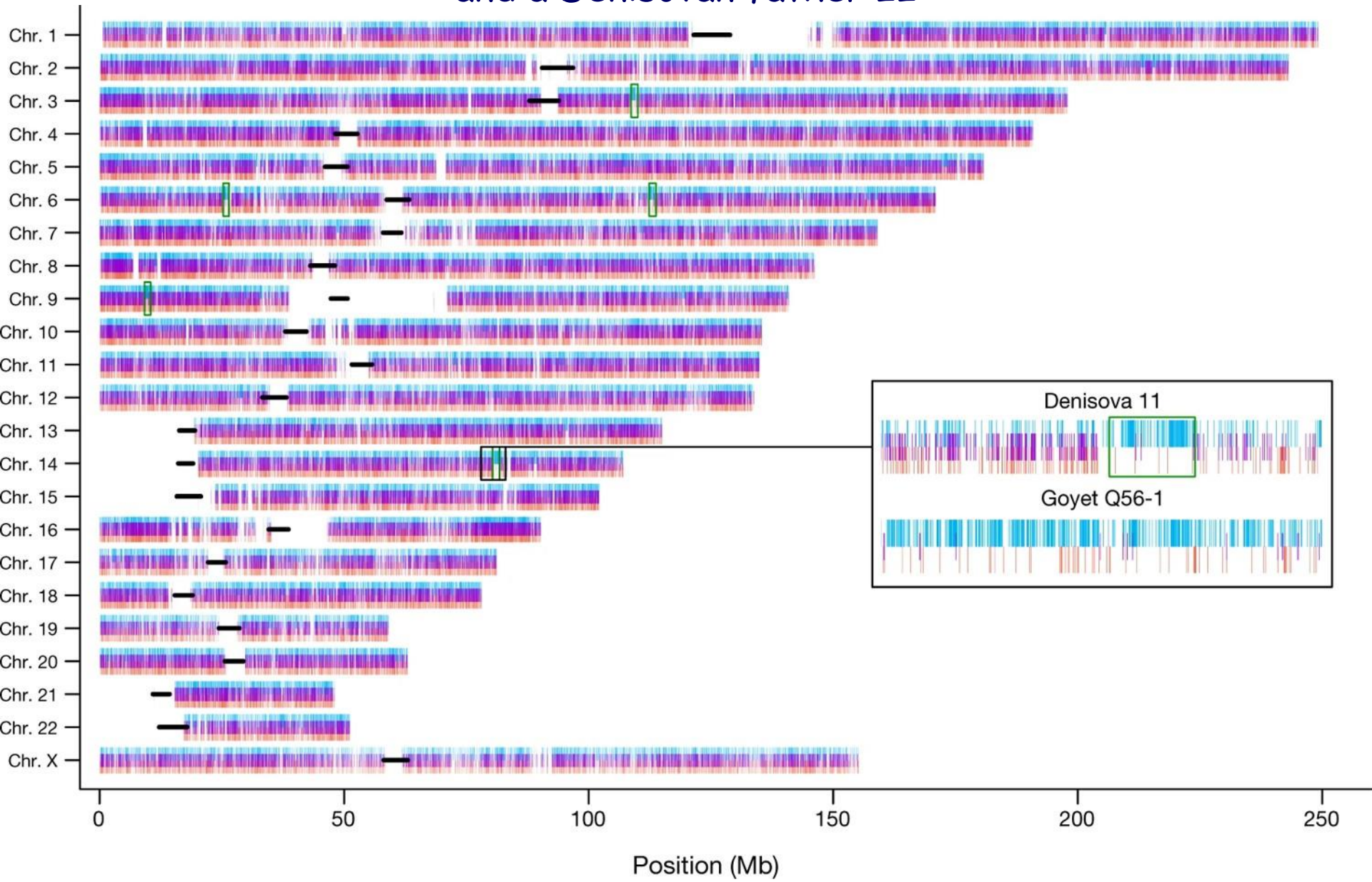
The genome of the offspring of a Neanderthal mother and a Denisovan father I



… 38.6% of fragments from Denisova 11 carried alleles matching the Neanderthal genome and 42.3% carried alleles matching the Denisovan genome.

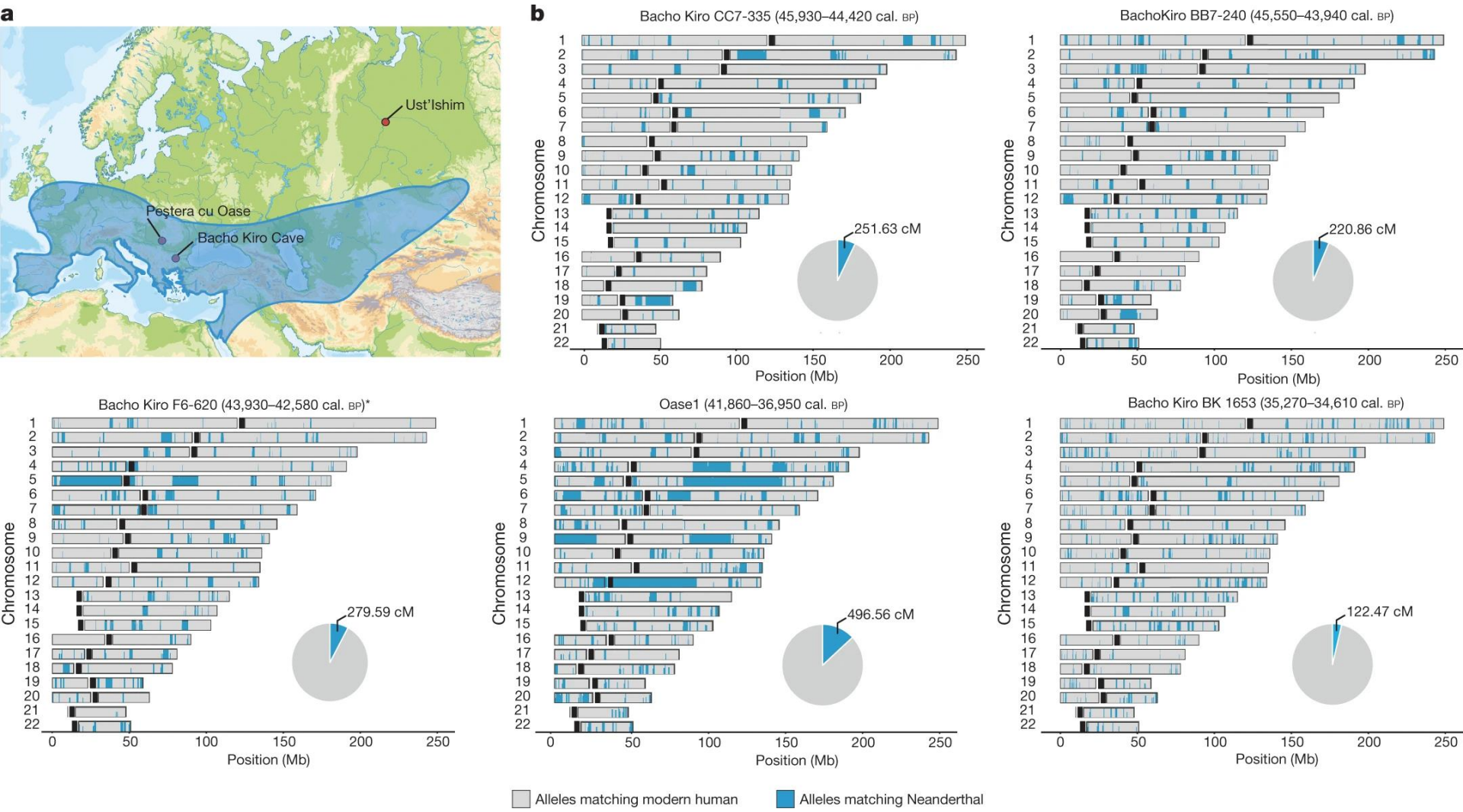
The finding of a first-generation Neanderthal–Denisovan offspring among the small number of archaic specimens sequenced to date suggests that mixing between Late Pleistocene hominin groups was common when they met (Slon et al., Nature 2018).

The genome of the offspring of a Neanderthal mother and a Denisovan father II



(Slon et al., Nature 2018)

Neandervölgyi DNS eloszlása ősi (min. 40.000 éves) modern emberi (Homo sapiens sapiens) genomokban



Hajdinjak, M., Mafessoni, F., Skov, L. *et al.* Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. *Nature* **592,** 253–257 (2021)