GENOMICS course

Organismal complexity and genome content

Eötvös Loránd University, Faculty of Science, Department of Genetics

From The Sunday Times

June 13, 2010

Genetics to solve why Ozzy Osbourne is still alive

Jack Grimston

EXPLORE HEALTH NEWS

SWINE FLU

THE mystery of why Ozzy Osbourne is still alive after decades of drug and alcohol abuse may finally be solved.

The 61-year-old former Black Sabbath lead singer — who this week begins his health advice column in The Sunday Times Magazine — is to become one of only a few people in the world to have his full genome sequenced.

In addition to giving Osbourne information that could help prevent diseases, it is hoped the results will provide insights into the way drugs are absorbed into the body.

The first full genome was sequenced in 2003 after 13 years of work. Today, analysing a genome takes three months and costs about £27,000.

COFACTOR genomics

Request a Quote | Blog | Vide

Search

Main Menu

Projects
Ultra-barcoding...
Whole-Genome Se...
The newest disc...
Sequencing Ozzy...

Sequencing Ozzy Osbourne

Cofactor Genomics LLC., in conjunct
constructed genomic DNA libraries a
generated approximately 39 Gb of se
on a newly installed Applied Biosyste
Carlsbad, CA while Knome, of Cambr

When the analysis and interpretation
at Knome, went to the UK to present
comparing Ozzys genome sequence t
Library of Medicine and human refere
discovered that Ozzy has several fam
Haplotypegroup-T and Haplotypegrou
Colbert and Henry "Skip" Gates. Ozzy

times more Neaderthal DNA than Ozzy.

Other interesting comparisons showed Ozzy is 6 times more likely than the average person to have a dependency to alcohol while showing a lower than average predilection to heroine and nicotine addiction (cigarettes were the fist thing he gave up several years ago when he went clean). Based on these results, it is no surprise that he drank several bottles of cognac a day for years. Interestingly, how he was able to handle that amount of alcohol may be explained by a mutation in the regulatory region of his ADH4 gene that metabolizes alcohol. This variation could have allowed him to process the alcohol at a faster rate than the normal person, leading to less health risks.

One of the most interesting findings was Ozzy has two version of the COMT gene (Catechol-O-methyltransferase) called "warrior" and "worrier". This is an enzyme that degrades dopamine, epinephrine, and norepinephrine. The "warrior" variant has been implicated in increased executive functions such as awareness, planning, organization, self-awareness, and potentially most important for Ozzy, self-regulation. While the "worrier" variant has been implicated in a decrease of these functions. In Ozzys own words, "I always thought it was just the booze and drugs that made me do crazy things like that, even though Ive always been a hypochondriac, and in some ways quite an anxious and insecure person. Maybe its more to do with my genes. Those two sides of my personality sum me up perfectly. Being a warrior, the crazy bat-eating Prince of Darkness, has made me famous. Being a worrier has kept me alive when some of my dearest friends never made it beyond their mid-twenties."

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

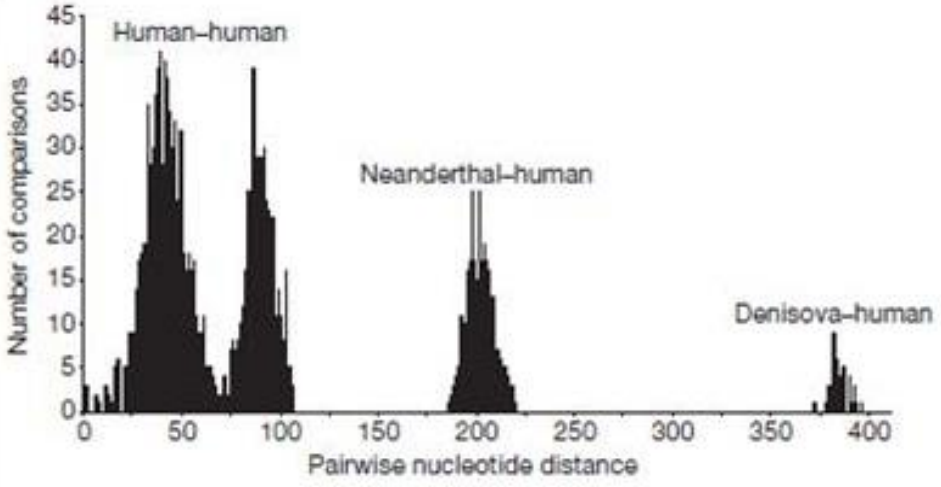


Figure 2 | Distribution of pairwise nucleotide differences. Pairwise nucleotide differences from all pairs of complete mtDNAs from 54 present-day and one Pleistocene modern human, six Neanderthals and the Denisova hominin are shown.

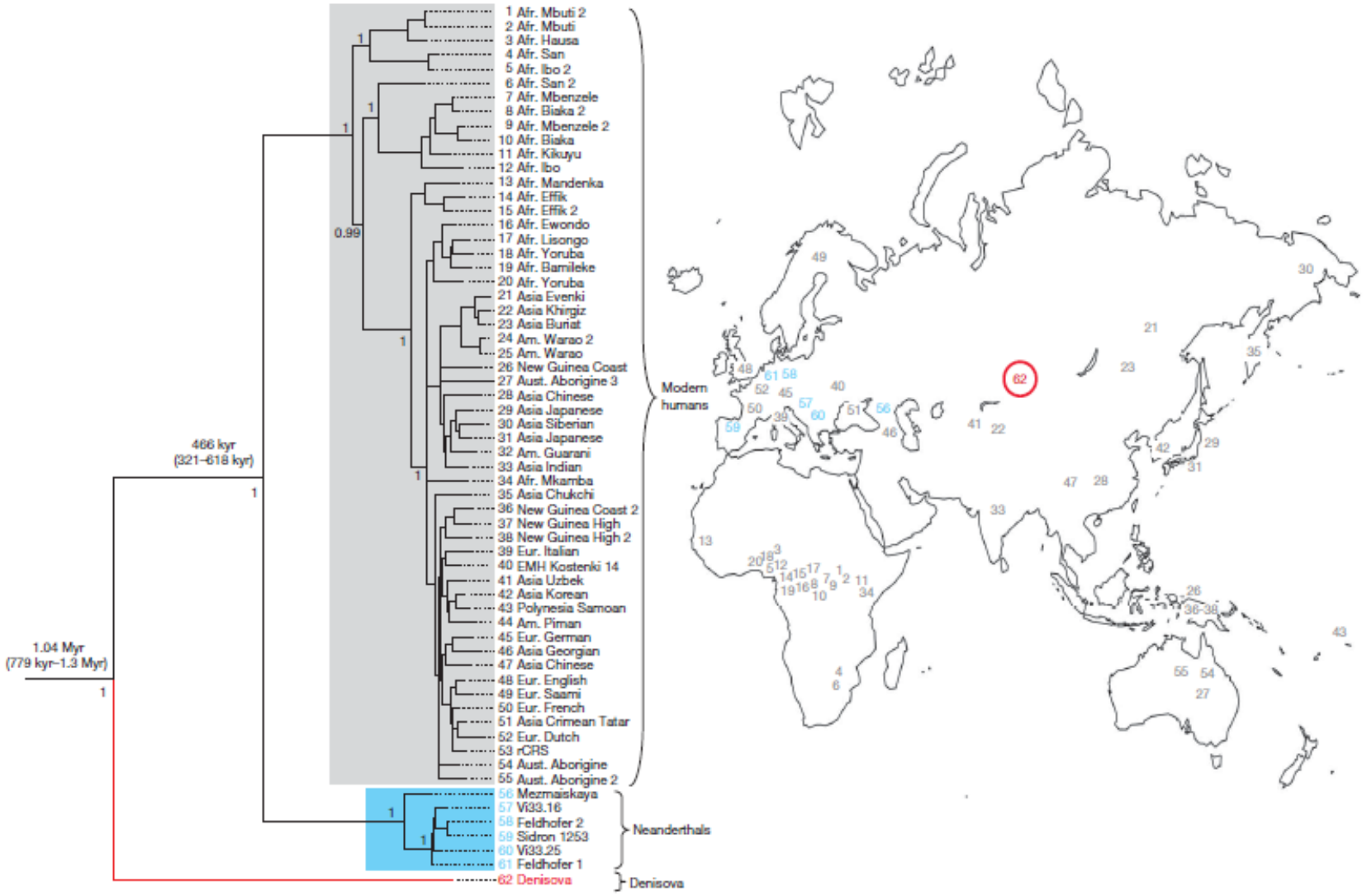
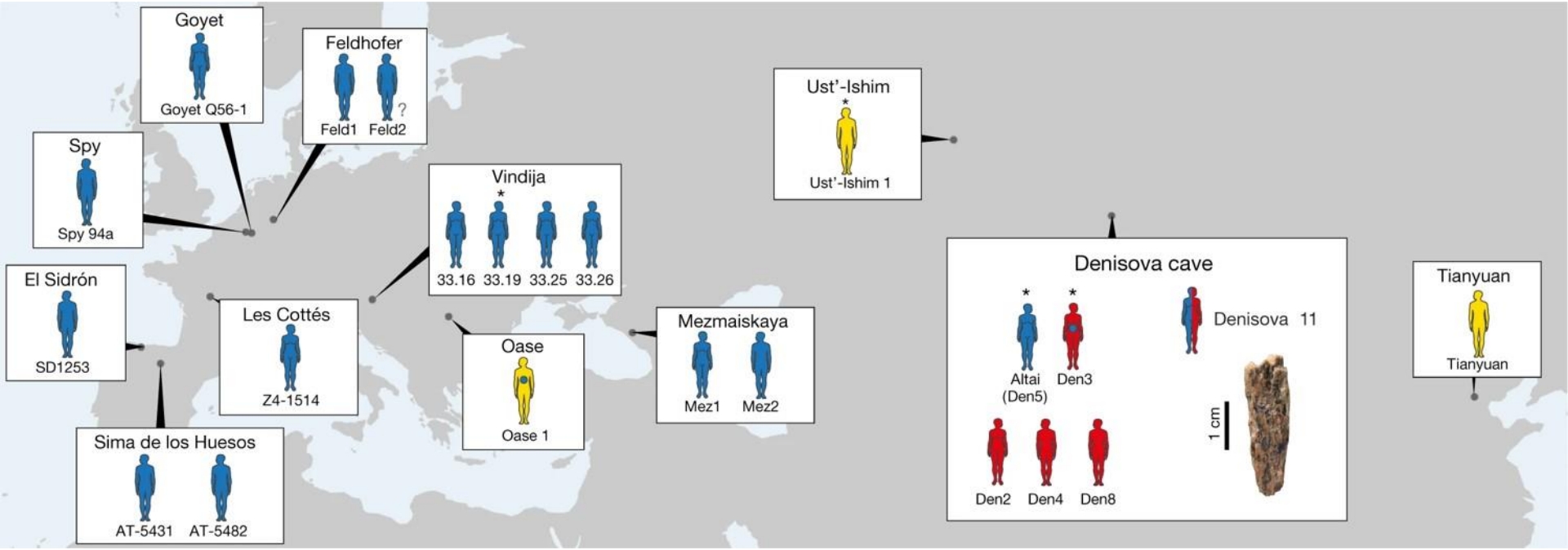


Figure 3 | Phylogenetic tree of complete mtDNAs. The phylogeny was estimated with a Bayesian approach under a GTR+I+Γ model using 54 present-day and one Pleistocene modern human mtDNA (grey), 6 Neanderthals (blue) and the Denisova hominin (red). The tree is rooted with a chimpanzee and a bonobo mtDNA. 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)

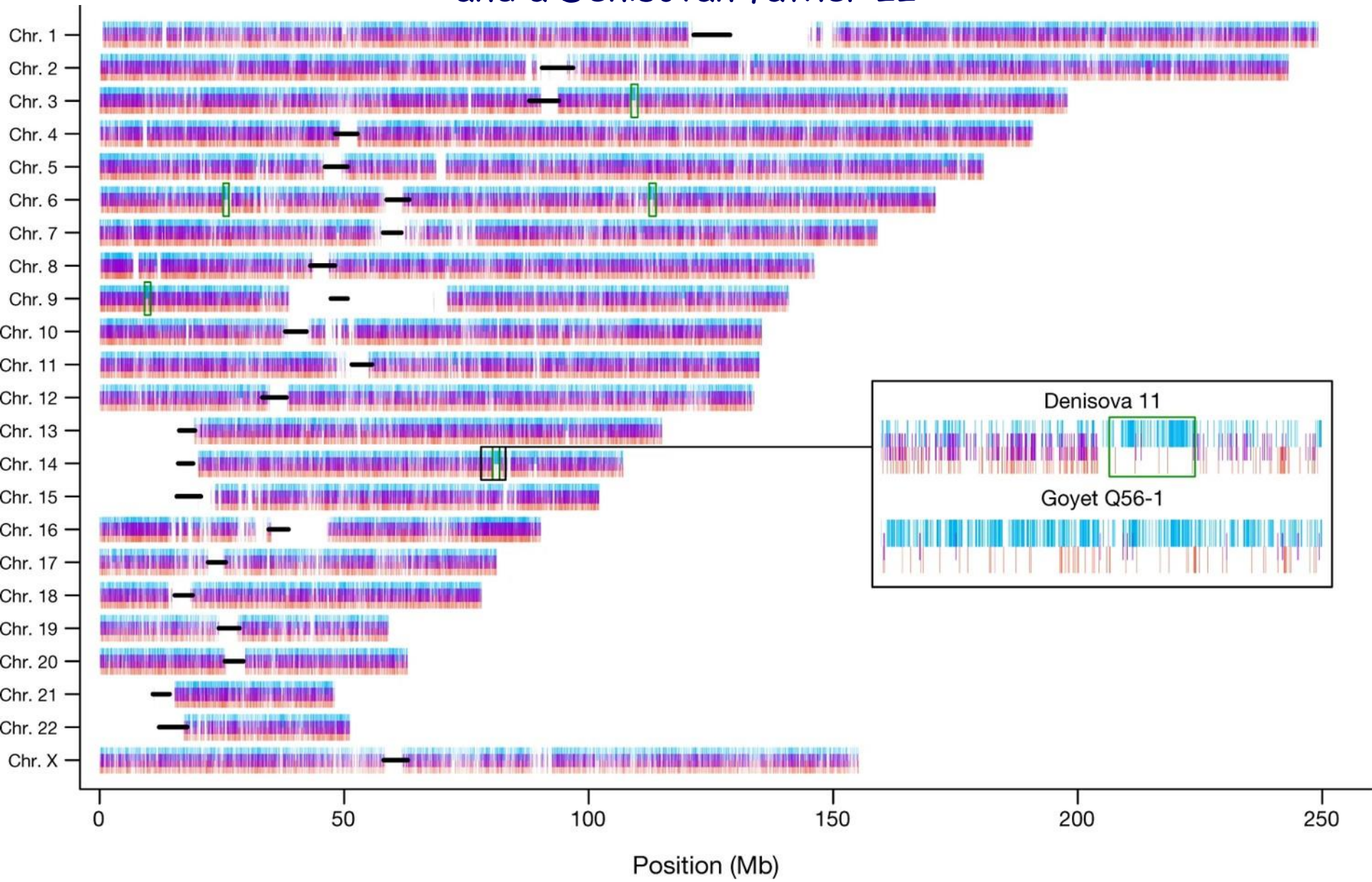
The genome of the offspring of a Neanderthal mother and a Denisovan father I



… 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)

GENOMICS course – syllabus

1. **09/09.** How did change our thinking about the genom content? Organismal complexity and gene number. (Egyed B)
2. **16/09.** Transcription regulation. Transcription site recognition at genomic level. (Varga M)
3. **23/09.** Epigenetics. (Varga M)
4. **30/09.** The Human Genome Project. Genome sequencing strategies and next generation sequencing. (Egyed B)
5. **07/10.** Structure and organization of the human genom. Genes, regulatory and mobile genetic elements, pseudogenes. (Egyed B)
6. **14/10.** Genetic variability and phenotype. Variations in the genome: DNA fingerprinting. Association studies. (Egyed B)
7. **21/10.** Animal genomes: Metazoa evolution and genomic aspects. (Varga M)

GENOMICS course – syllabus

8. **04/11.** Sex chromosomes: origin and diversity. Y chromosome degeneration. X chromosome rearrangement. (Varga M)
9. **11/11.** Plant genomics. (Kaló P/Egyed B, MBK-Gödöllő)
10. **18/11.** Prokaryote and virus genomes. (Varga M)
11. **25/11.** Gene expression studies. Transcriptomics.

    (Puskas L/Egyed B, SZBK)

8. **02/12.** Phylogenetics and rare genomic changes. (Egyed B)
9. **09/12. Consulting lecture**

16-18/12. 10.00 WRITTEN EXAM??? Which day?

References, text books, curricula…

The Origins of Genome Architecture
author: Michael Lynch
publisher: Sinauer Associates, Inc. Publishers, 2006

A Primer of Human Genetics
author: G. Gibson
publisher: Sinauer Associates, Inc. Publishers, 2015

The Evolution of the Genome
editor: T. Ryan Gregory
publisher: Elsevier Academic Press, 2005

ELTE Dept. Genet.: https://genetics.elte.hu

user: **genetika2019**
pw: **genetika2019**

Terminal exams: written in December, oral exams in January

How did change our imagination from genom content, about the relationship oforganismal comlexity and gene number?

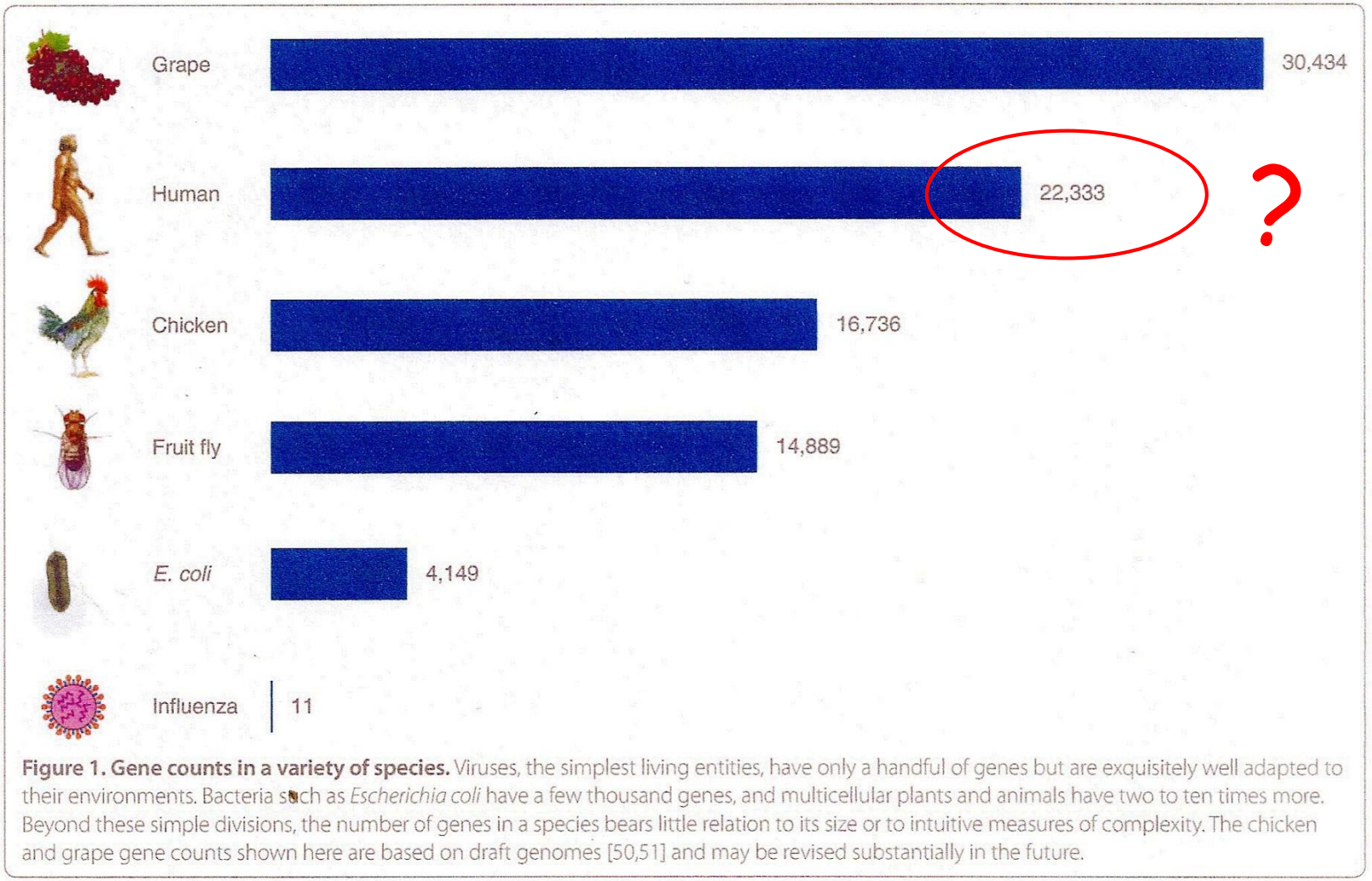


Figure 1. Gene counts in a variety of species. Viruses, the simplest living entities, have only a handful of genes but are exquisitely well adapted to their environments. Bacteria such as *Escherichia coli* have a few thousand genes, and multicellular plants and animals have two to ten times more. Beyond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

„ … somewhere between chicken and grape" (Pertea & Salzberg, Genome Biology 2010, 11:206)

First estimation about human genome size and gene number

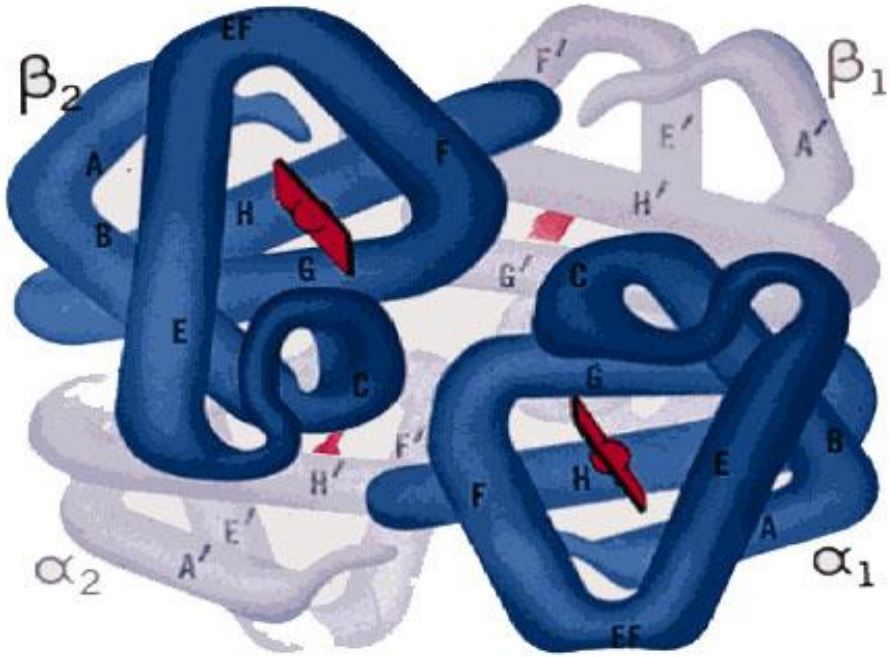
1964: F. Vogel (Heidelberg)

-Hemoglobin α and β chains

-Simplified presumption

-Human genome: 3×10^9 bp

-Gene number: 6,7 million !!!



1990: NIH/DOE report on Human Genome Project

-estimation: 100.000 genes based on average human gene size (30 000 bp)

-2001, Human Genome Project: decreasing gene no., increasing uncertainty

What do we call a gene, how can it be defined?

The „Gene" definition changed remarkably in the last one hundred years.

- protein/RNA coding, terms of intron/exon, regulatory function, etc.

- Distinction in the function

Recent definition (what we use during the lecture):

„A gene is a region of the genome that is transcribed into messenger RNA and translated into one or more proteins." (i.e. alternative splicing)

How do we call?

- i.e. non-protein coding RNA genes (pl. lncRNAs, miRNAs, snRNAs, piwiRNA)

Automated DNA sequencing and „Computer Biology"

ESTs: mRNA poly(A)3' ends → RT-PCR → cDNS library ('90-)

300 cDNA library from 37 different tissue samples: ~ 87.983 sequences

Adams MD, et al., Nature (1995): → ca. 100.000 gene (NIH/DOE)

Based on ESTs gene number at the end of 90': 35 000 – 57 000 (CpG islands)

Ho can we determine a gene ? – Based on Bioinformatics issue:

- protein coding sequences, based on sequence homology.
- based on de novo predictor signals (i.e. Genscan: 45.000 genes)
- comparative study of conserved sequences (i.e. Twinscan: 25.600 genes)
- statistical modelling (GH Markov Model, CRF: conditional random fields)
- failed *de novo* predictions, false positives: pseudogenes
- JIGSAW, Gnomon (NCBI, Ensembl): integrative metodics (2005-)

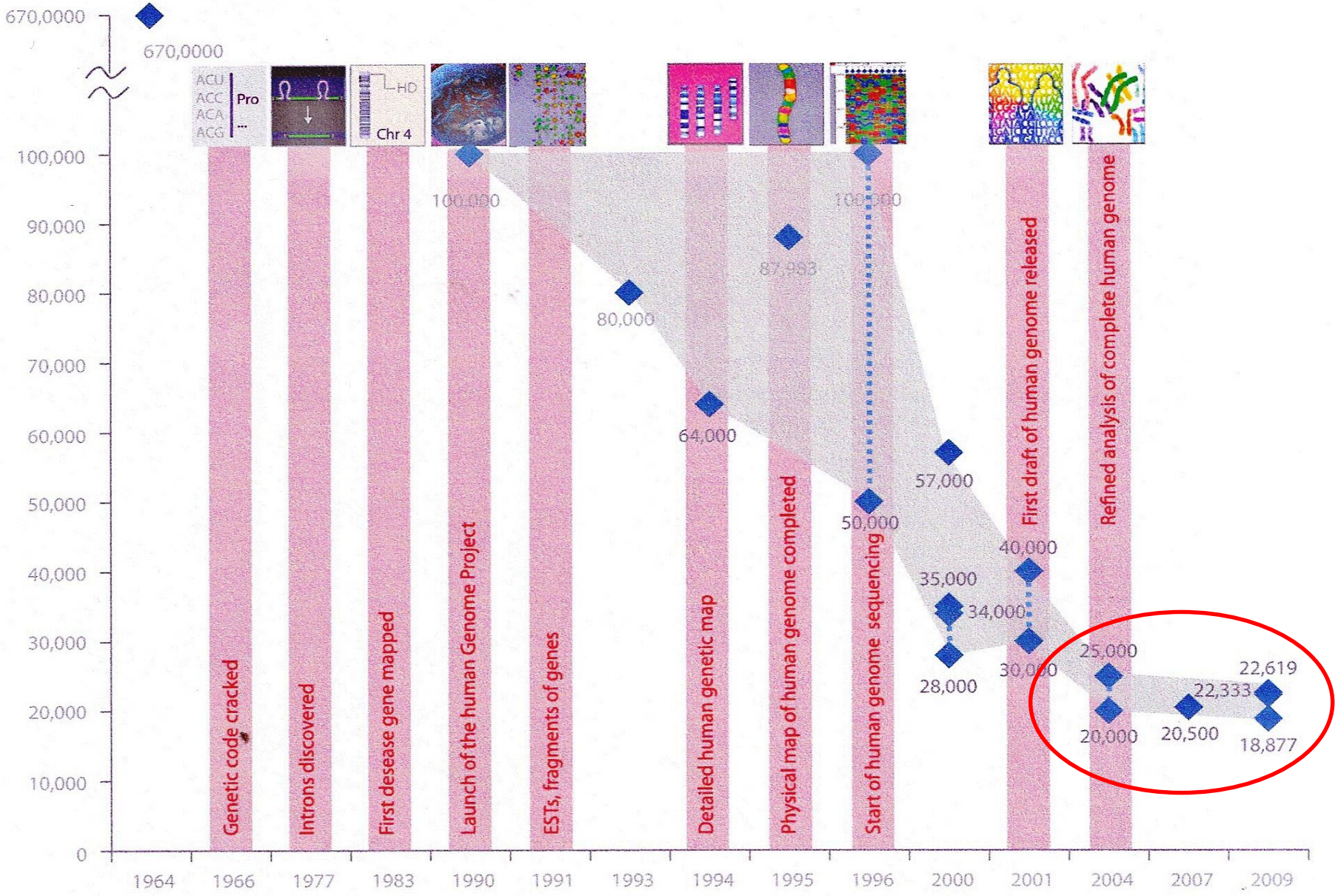


Figure 2. The trend of human gene number counts together with human genome-related milestones. Individual estimates of the human gene count are shown as blue diamonds. The range of estimates at different times is shown by the two vertical blue dotted lines. Note how this range has narrowed in recent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

Where do we are now?

2001, Human Genome Consortium: 30 000 – 40 000 protein coding genes

Celera Consortium: 26 500 „strong" + 12 000 „weak" evidence

2004, Human Genome Consortium: 20 000 – 25 000 genes

- less than Arabidopsis → **organizmal complexity?**

2010, Ensembl: 22 619 / NCBI: 22 333 protein coding genes

CCDS: 18 173 (http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi)

fals pozitives: retrotransposons, pseudogenes, „orphan" DNA

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

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

Novel genes

- CGH analyses: less differences between related species
- *de novo* gene : duplication and neofunctionalization
- gene no. differences between individuals: segmental duplications
- large-scale copy number polymorphisms (CNVs > 1000 regions)
- human „pangenom": variation between races and groups

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

- ca. 40 Mb new sequences, + 1,3 %
- *de novo* origin: non-coding sequences, ca. 18 new homo gene?

(Knowles and McLysaght, 2009, Genome Res)

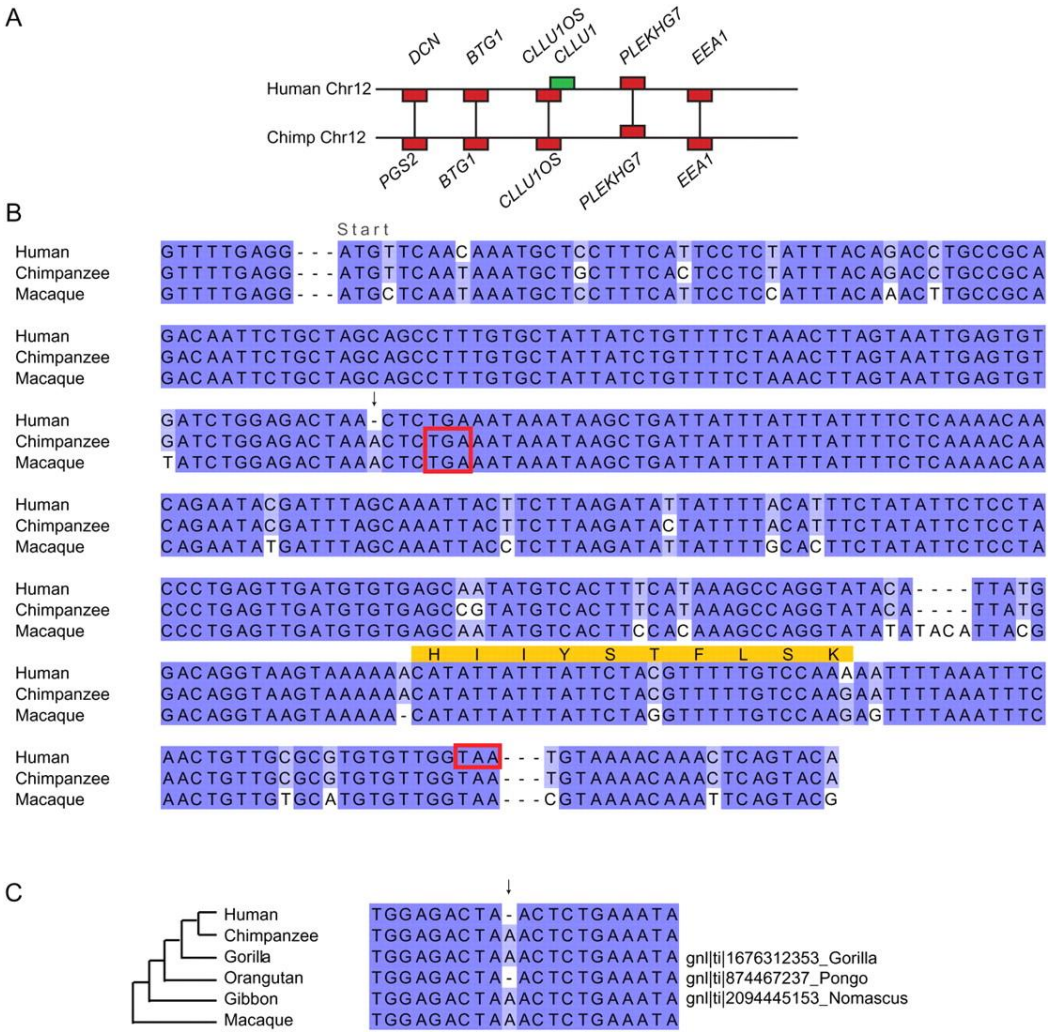
Table 1. Novel human protein-coding genes and supporting evidence.

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.

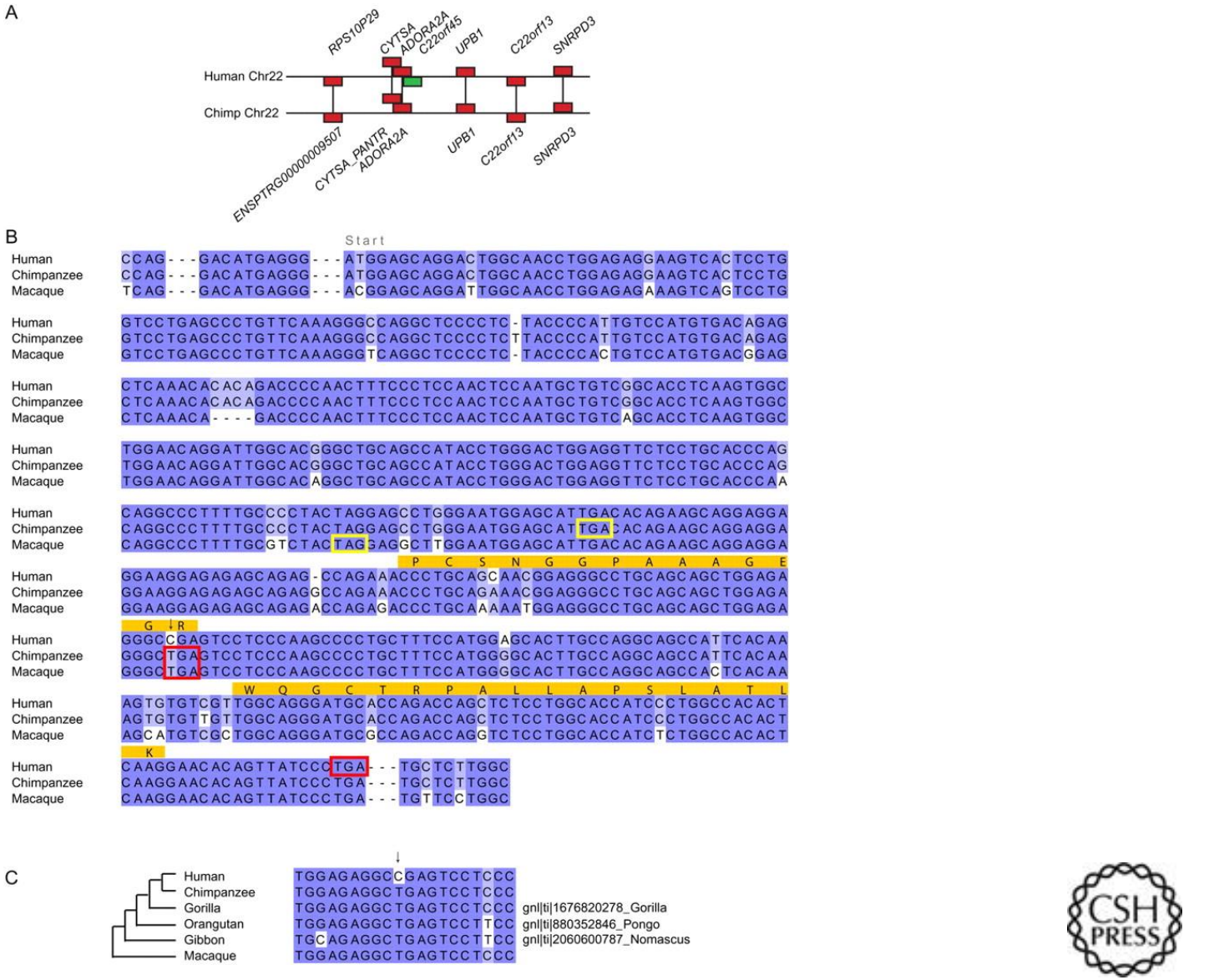
Knowles D G , McLysaght A Genome Res. 2009;19:1752-1759

Sequence changes in the origin of CLLU1 from noncoding DNA



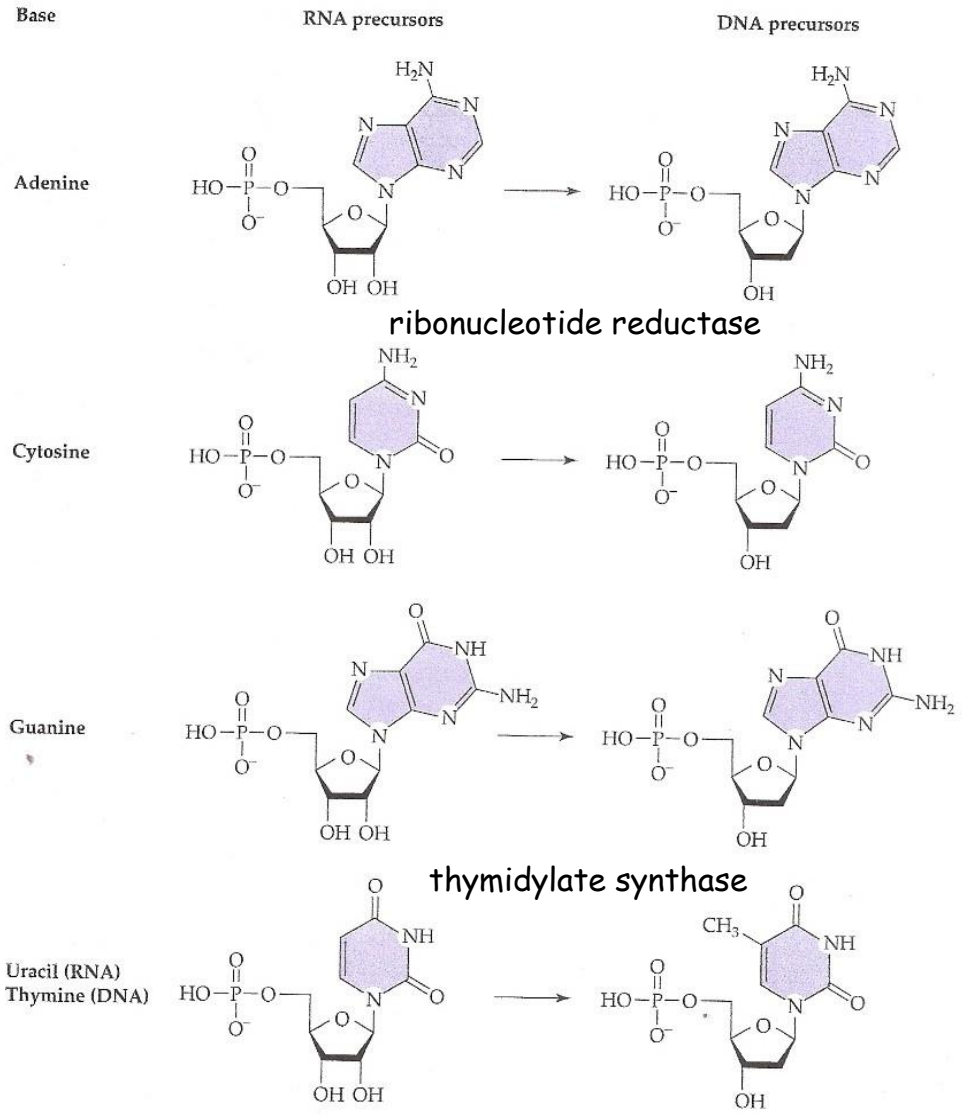
Knowles D G , McLysaght A Genome Res. 2009;19:1752-1759

Sequence changes in the origin of C22orf45 from noncoding DNA

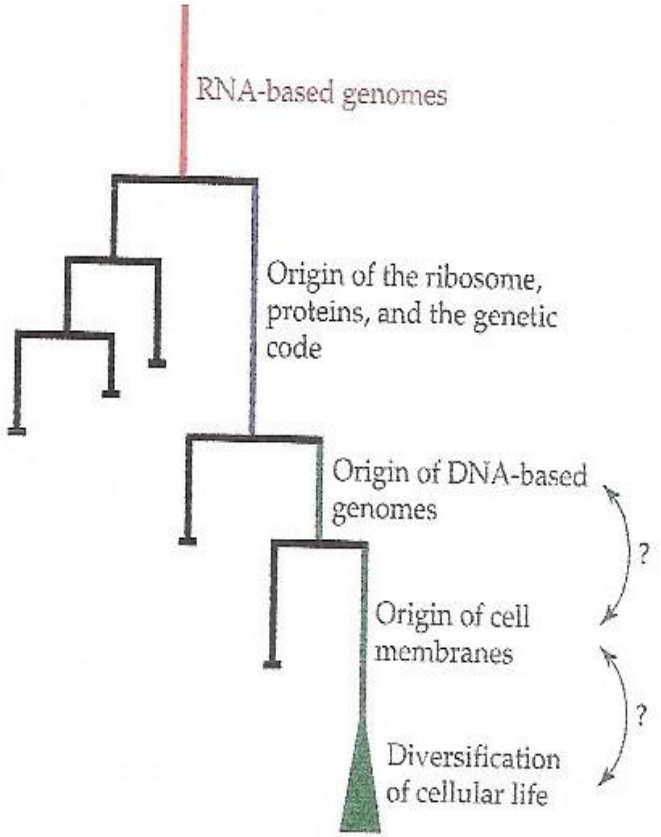


Knowles D G , McLysaght A Genome Res. 2009;19:1752-1759

Origin of the eukaryote genom: RNA world



2' –OH instability / Mutation: C » U



Different pathways in membrane lipid sinthesis regarding archeas and eubacteria:

izopren ether vs. Lipid-ester

Genome evolution based on rRNA sequences

Woese and Fox, 1977

Woese et al., 1990

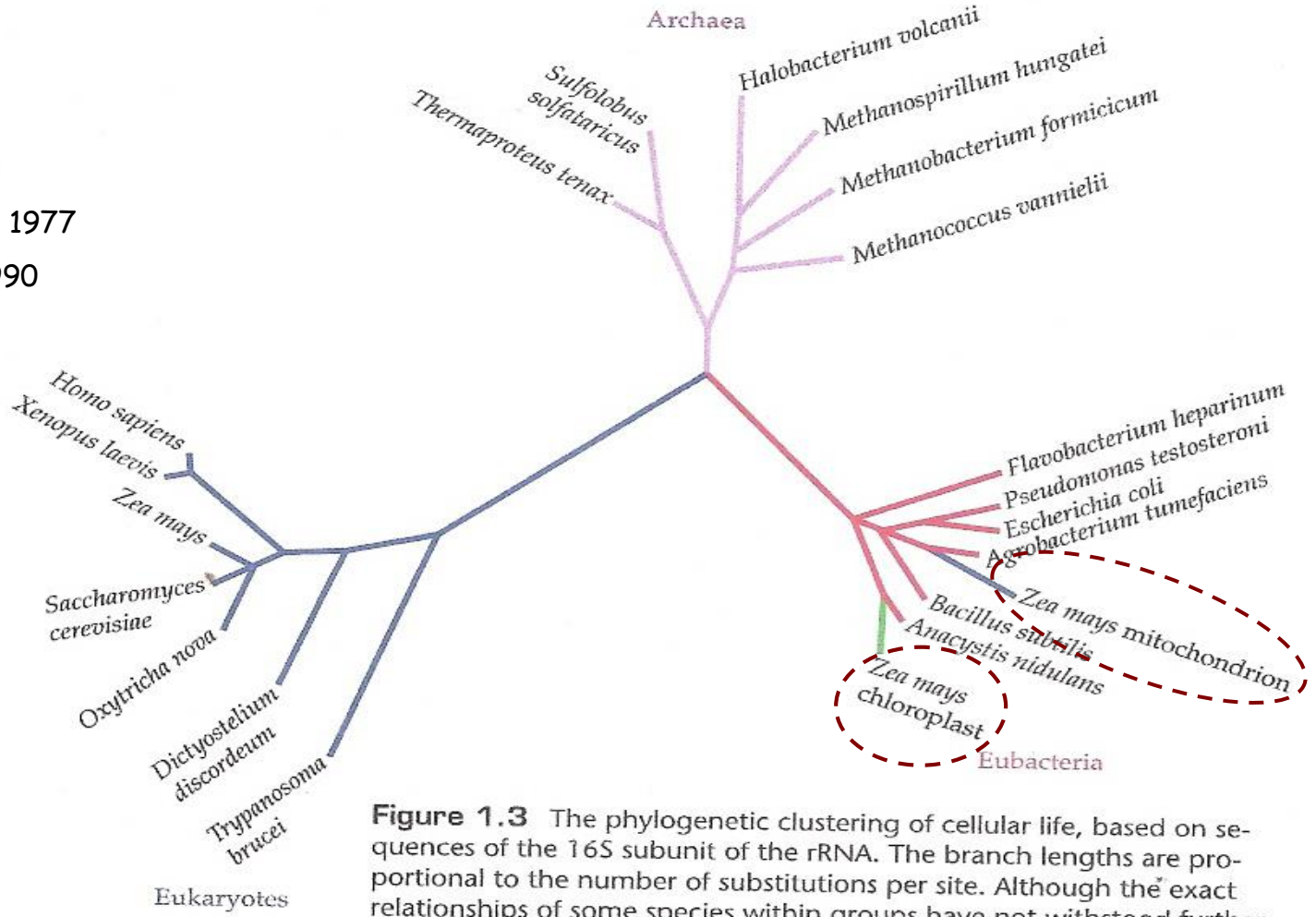
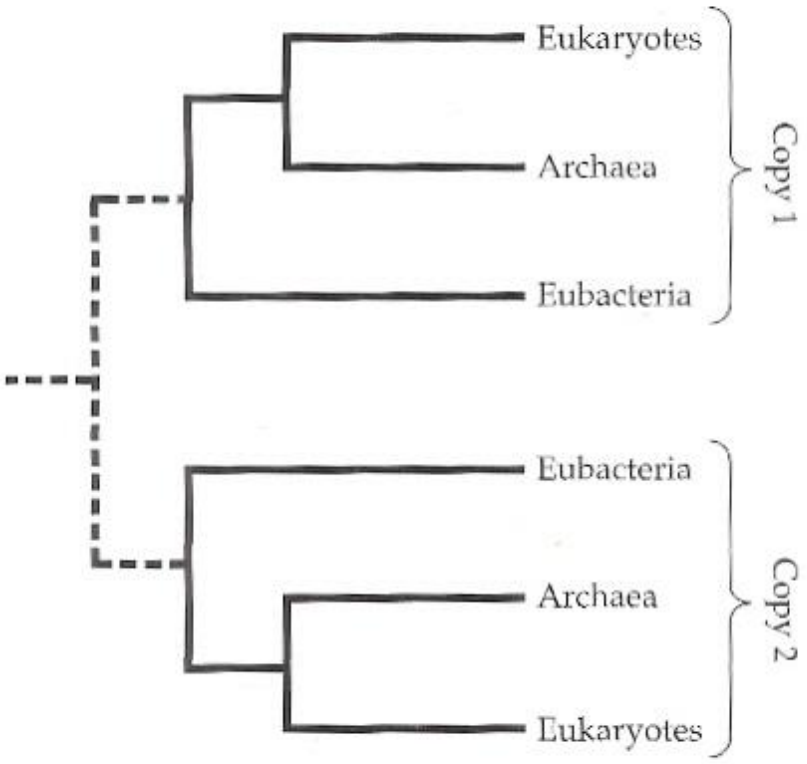


Figure 1.3 The phylogenetic clustering of cellular life, based on sequences of the 16S subunit of the rRNA. The branch lengths are proportional to the number of substitutions per site. Although the exact relationships of some species within groups have not withstood further scrutiny, the distinct nature of the three major domains is well accepted. The presence of mitochondrial and chloroplast sequences in the eubacterial lineage provides compelling evidence for the eubacterial ancestry of these organelles. The tree is unrooted, as the position of the most recent common ancestor of the three major groups is not identified. (Modified from Pace et al. 1986.)

Genome evolution based on gene duplication

ATPase membrane
duplicated subunits:



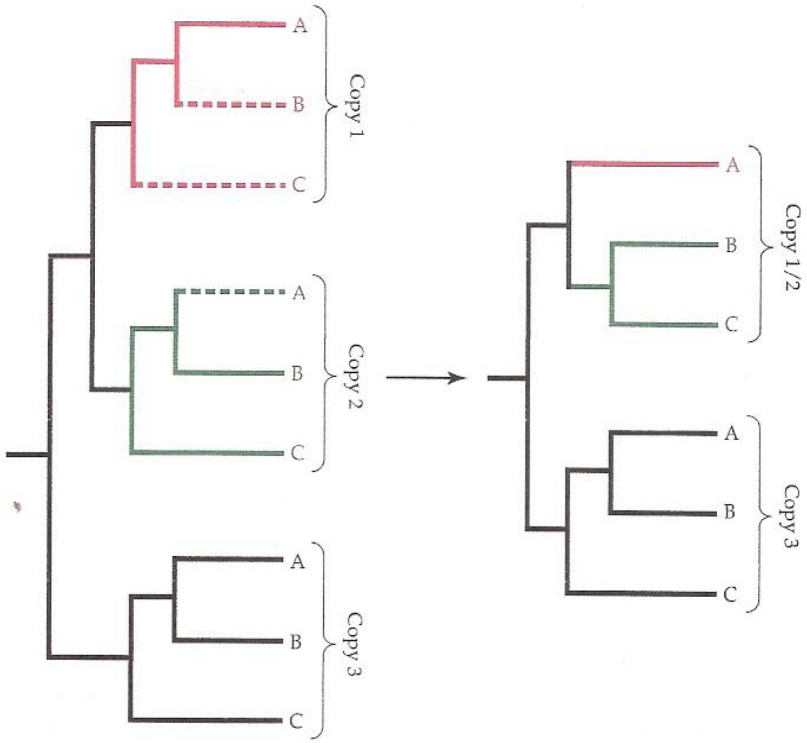


Figure 1.5 *Left*: Here the ancestral species had three gene copies (the first duplication led to copy 3 and the ancestor to 1 and 2; and the second duplication created copies 1 and 2). Copy 1 was lost in species B and C, and copy 2 was lost in species A (dashed lines). *Right*: Using just the surviving gene copies, a conflicting signal is obtained on the phylogenetic relationships of species A, B, and C, even though the overall topology of extant gene relationships is correct. The top cluster incorrectly implies a phylogeny in which species B and C are grouped together (as a consequence of an incorrect mixture of copy 1 and 2 genes), whereas the bottom cluster correctly groups A and B.

Origin of eukaryote genome: an archea-eubacteria chimera?

Transcription and translation: **Archea**

housekeeping functions: **Eubacteria**

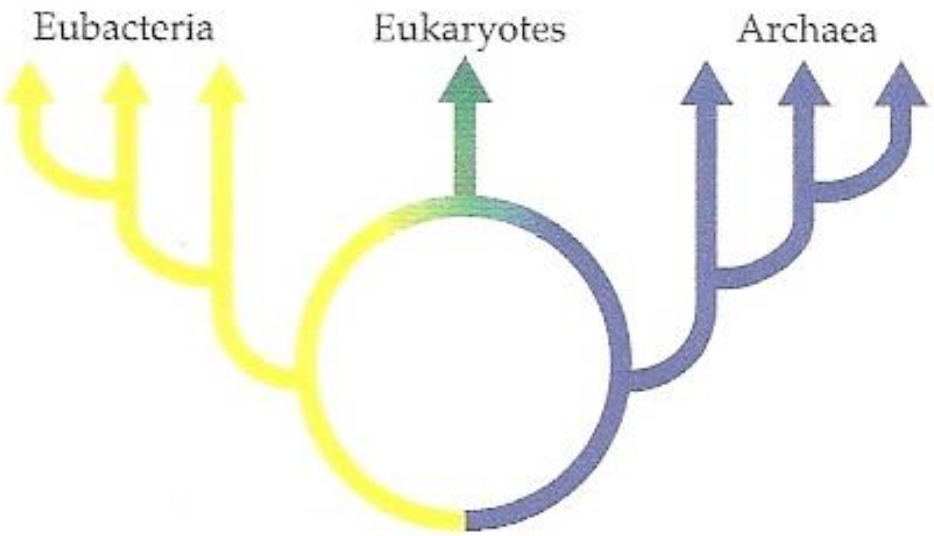


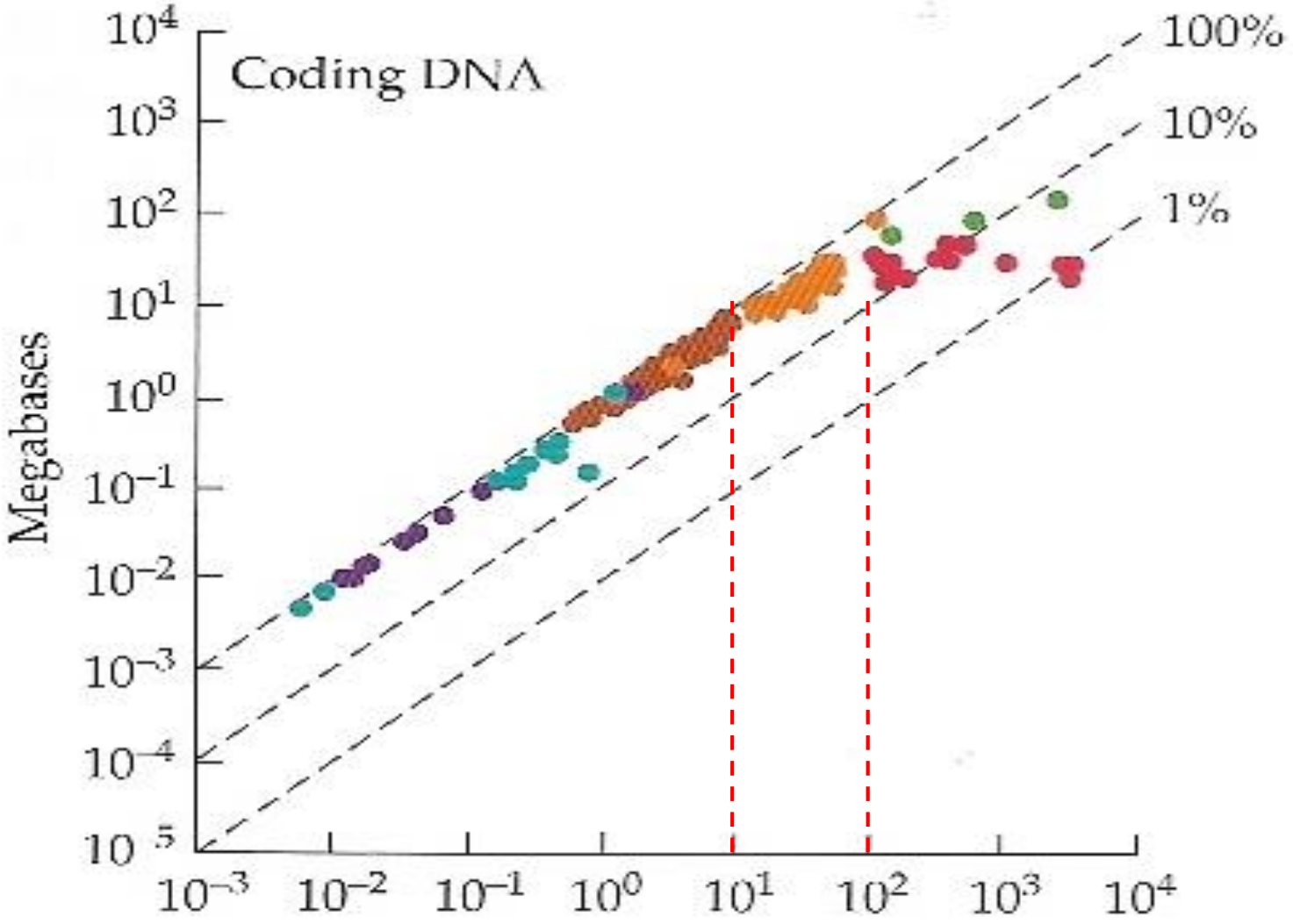
Figure 1.7 The "ring-of-life" hypothesis for the origin of eukaryotes. Yellow and blue lineages denote branches in the phylogenetic trees for eubacteria and archaea, respectively. Members of two such lineages fused to form the eukaryotic domain (green). (Modified from Rivera and Lake 2004.)

Eukaryote versus prokaryote genomes

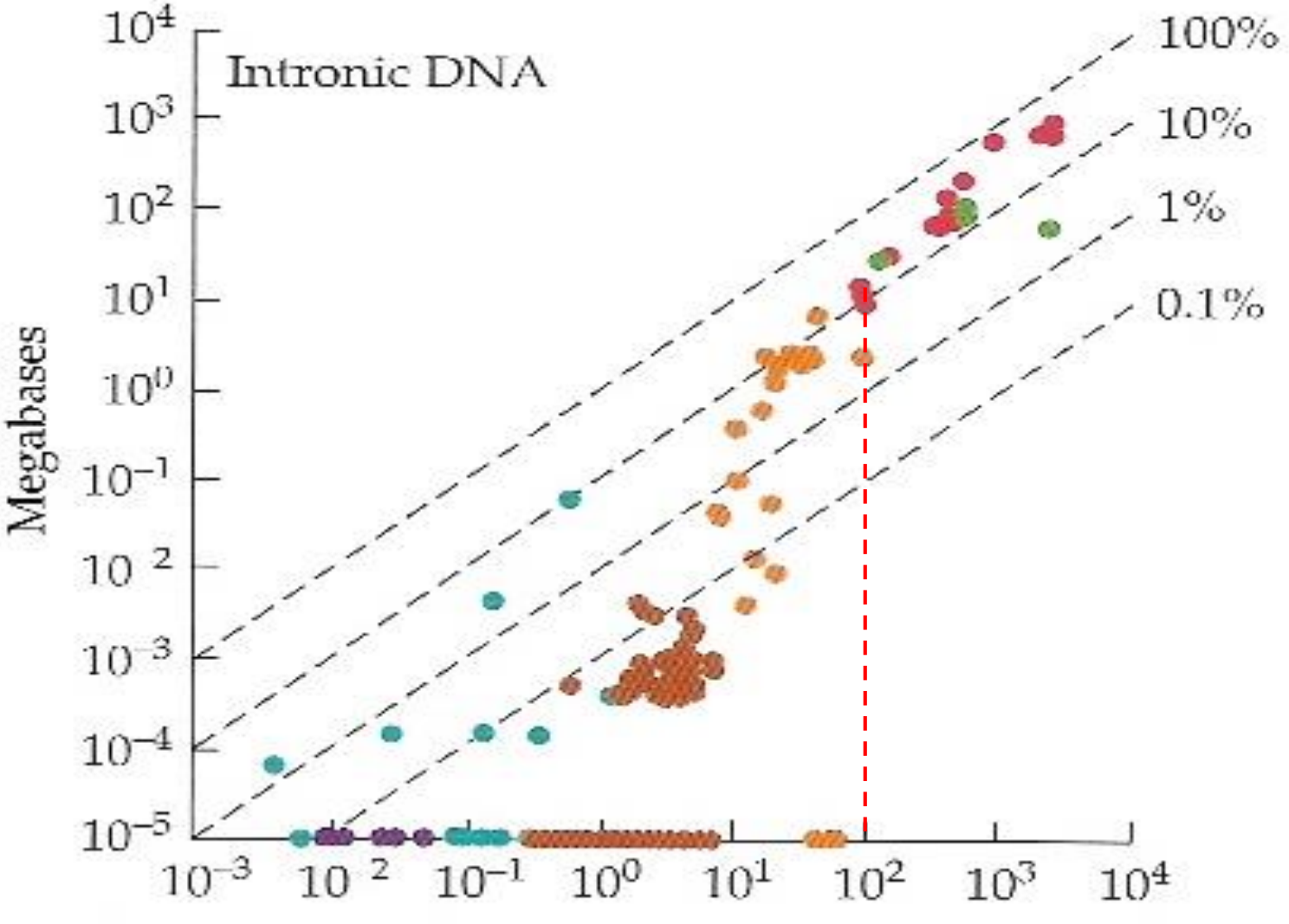
TABLE 1.1 Some of the features that set eukaryotic genomes apart from those of prokaryotes, and their exceptions

EUKARYOTES	PROKARYOTES
Presence of a nuclear membrane	Also present in the Planktomycetes
Organelles derived from endosymbionts	Also present in the β-proteobacteria
Cytoskeleton and vesicle transport machinery	Tubulin-related proteins, but not microtubules
Trans-splicing	Absent
Introns in protein-coding genes, and a complex spliceosomal apparatus for excising them	Rare self-splicing introns, but almost never in coding DNA
Expansion of the untranslated regions of transcripts	Untranslated regions are generally very short
Addition of poly(A) tails to all mRNAs	Rare and nonessential polyadenylation of transcripts
Translation initiation by scanning for start codon	Ribosome binds directly to a Shine-Dalgarno sequence
Messenger RNA surveillance	The nonsense-mediated decay pathway is absent
Multiple linear chromosomes capped with telomeres	Single linear chromosomes in a few eubacteria
Mitosis and meiosis	Absent
Expansion in gene number	The largest prokaryotic genomes contain more genes than the smallest eukaryotic genomes
Expansion of cell size and number	A few have very large cell sizes (e.g., *Thiomargarita*), and several produce multiple cell types

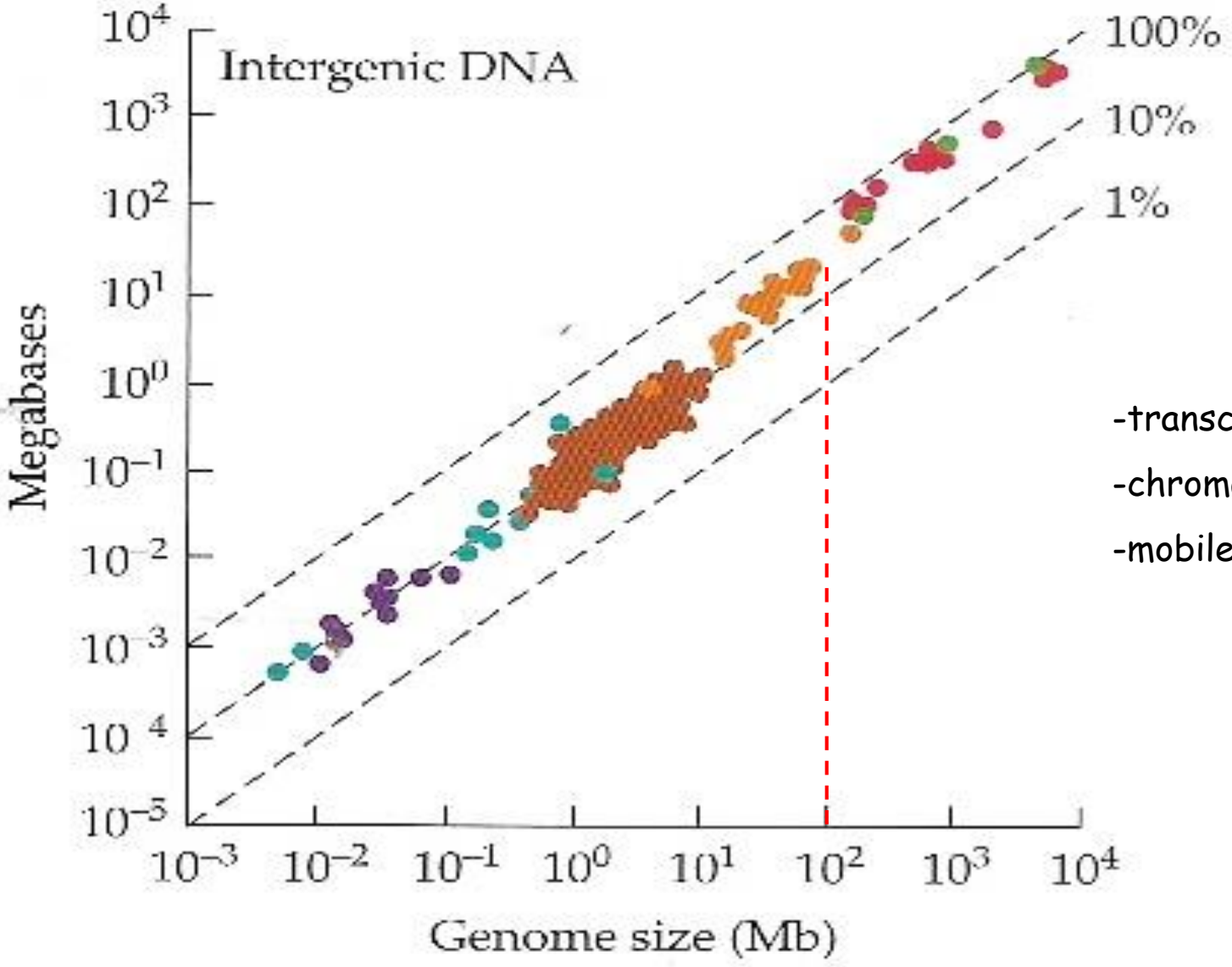
Genome size vs. coding sequences



Genome size vs. introns



Genome size vs. intergenic DNA



-transcription sites

-chromatin package

-mobile elements

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

Genome size and organizmal comlexity

- WGC: recurrent mutations comparing whole individual genomes
- Prokaryote: 350-8000 genes, 0.5 – 9 Mb genome
- Multicellulare Eukaryote: > 13.000 genes, > 100 Mb genome
- Noncoding DNA expansion (introns, mobile elements, pseudogenes)
- Organism size vs. No. of cell types – pozitive correlation
- Gene no. / genome size vs. multicellularity / organizmal komplexity

Correlation? It does not depend on gene no. and genome size but even more how they operate! (transcription regulation, alternative splicing etc.)

Genome size and complexity

• There is essentially no correlation between genome size and organismal complexity.

•Clear ranking from viruses to prokaryotes to uni- and multi-cellular eukaryotes in terms of genome size, gene no. etc.

• Despite this gradient, there are no abrupt discontinuities in the scaling of genome content with genome size (C-value paradox).

• indirect evidence that the evolution of genomic architecture are unlikely to be direct consequences of organismal differences in cell structures or physiologies.