Ch 1 – L 2.1

- More on genomes
- Comparative
- Gene structure
- Genomic browsers

Other background from Genetics you may need

Genes «families»

Similarity in «parts» of the proteins, called «domains»:

Identity, Paralogy and Orthology

Mechanisms of evolution

evolution

Sequences of the Reference Human Genome have been since continuously adjourned, revised, completed, maintained

Different versions released timely by the HGP Consortium

Sequences and annotations are conserved and available from biological **Databases**

Several organizations to maintain and run public databases. They are paid by Agencies and other public organizations

Comparative

Many other genomes sequenced completely or partially

Most of sequencing projects are publicly funded, results are open in databases

Many other are run by private funding and results are not open. They include many vegetables, bacteria, fungi.

Public **databases** :

ENSEMBL species

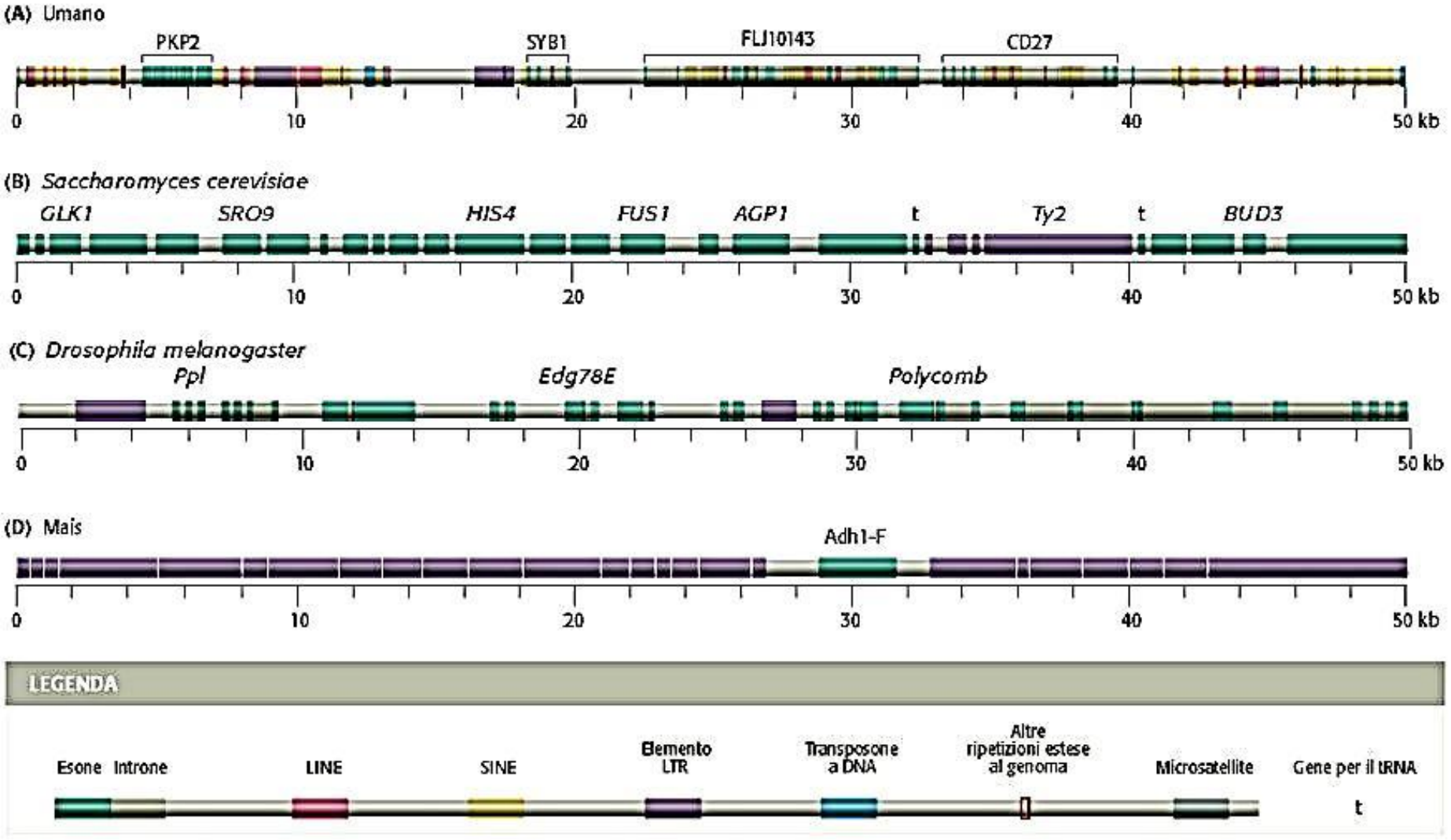
NCBI Genomes Genomic Data

T.A. Brown
Genomi, III Ed.
EdiSES

Comparative:

- Human
- Yeast
- Drosophila
- Mais

Figura 7.15 Confronto tra genoma umano, di lievito, del moscerino della frutta e di mais. (A) Il segmento di 50 kb del cromosoma 12 umano mostrato precedentemente, è confrontato con segmenti di 50 kb derivanti da genomi di (B) *S. cerevisiae*; (C) *Drosophila melanogaster*; (D) mais.



Textbook G

Leading Edge
Review

Cell

Looping Back to Leap Forward: Transcription Enters a New Era

Michael Levine,[1,*] Claudia Cattoglio,[1,2] and Robert Tjian[1,2,*]
[1]Department of Molecular and Cell Biology
[2]Howard Hughes Medical Institute, CIRM Center of Excellence, Li Ka Shing Center for Biomedical and Health Sciences
University of California, Berkeley, Berkeley, CA 94707, USA
*Correspondence: mlevine@berkeley.edu (M.L.), jmlim@uclink4.berkeley.edu (R.T.)
http://dx.doi.org/10.1016/j.cell.2014.02.009

Comparative genome analyses reveal that organismal complexity scales not with gene number but with gene regulation. Recent efforts indicate that the human genome likely contains hundreds of thousands of enhancers, with a typical gene embedded in a milieu of tens of enhancers. Proliferation of *cis*-regulatory DNAs is accompanied by increased complexity and functional diversification of transcriptional machineries recognizing distal enhancers and core promoters and by the high-order spatial organization of genetic elements. We review progress in unraveling one of the outstanding mysteries of modern biology: the dynamic communication of remote enhancers with target promoters in the specification of cellular identity.

Textbook G

Why is it called Textbook G ?
G stands for «General», since it concerns the entire course

Cell

Leading Edge
Review

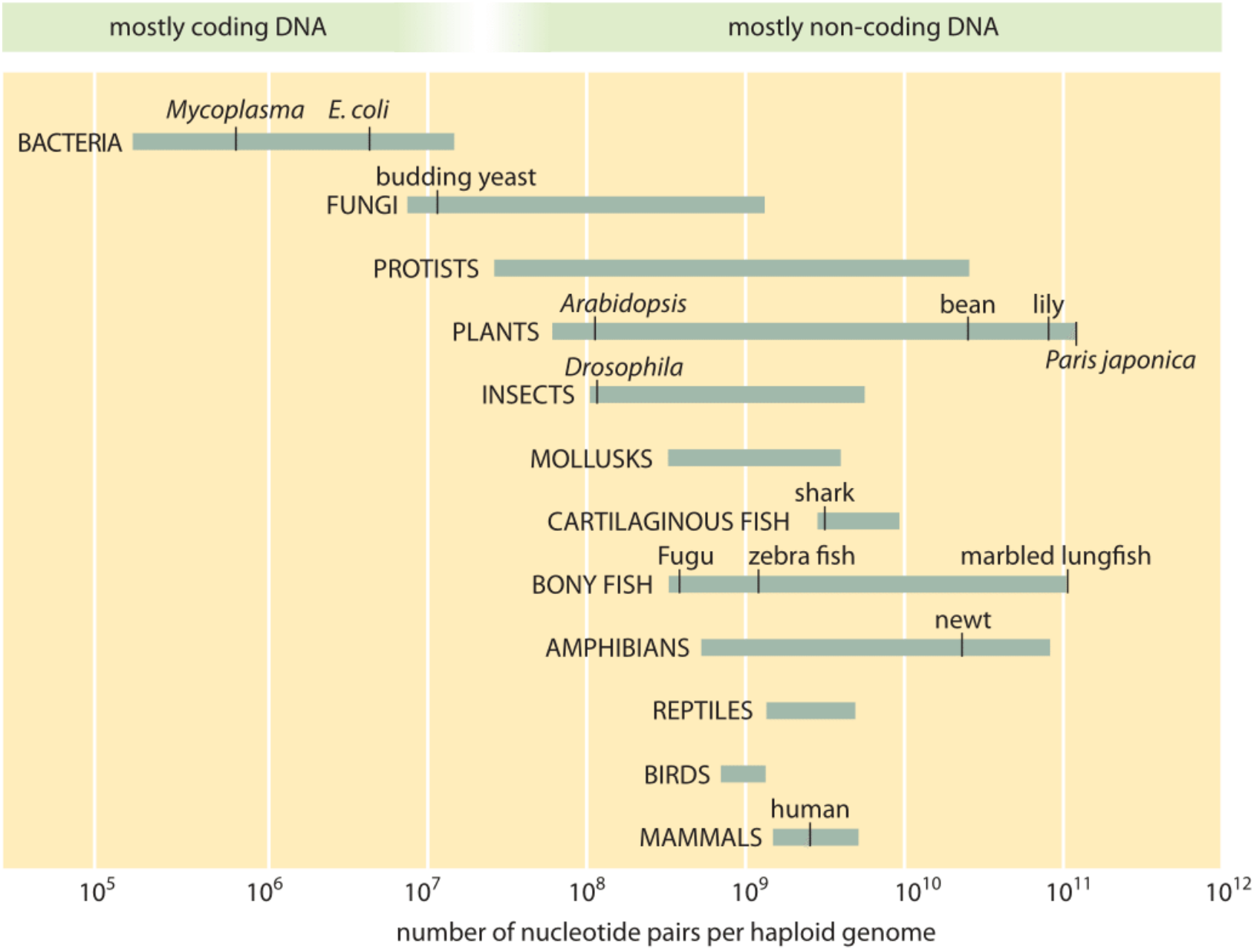
Looping Back to Leap Forward: Transcription Enters a New Era

Michael Levine,[1,*] Claudia Cattoglio,[1,2] and Robert Tjian[1,2,*]
[1]Department of Molecular and Cell Biology
[2]Howard Hughes Medical Institute, CIRM Center of Excellence, Li Ka Shing Center for Biomedical and Health Sciences
University of California, Berkeley, Berkeley, CA 94707, USA
*Correspondence: mlevine@berkeley.edu (M.L.), jmlim@uclink4.berkeley.edu (R.T.)
http://dx.doi.org/10.1016/j.cell.2014.02.009

Comparative genome analyses reveal that organismal complexity scales not with gene number but with gene regulation. Recent efforts indicate that the human genome likely contains hundreds of thousands of enhancers, with a typical gene embedded in a milieu of tens of enhancers. Proliferation of *cis*-regulatory DNAs is accompanied by increased complexity and functional diversification of transcriptional machineries recognizing distal enhancers and core promoters and by the high-order spatial organization of genetic elements. We review progress in unraveling one of the outstanding mysteries of modern biology: the dynamic communication of remote enhancers with target promoters in the specification of cellular identity.

Comparative genome analyses reveal that organismal complexity scales not with gene number but with gene regulation. Recent efforts indicate that the human genome likely contains hundreds of thousands of enhancers, with a typical gene embedded in a milieu of tens of enhancers. Proliferation of cis-regulatory DNAs is accompanied by increased complexity and functional diversification of transcriptional machineries recognizing distal enhancers and core promoters and by the high-order spatial organization of genetic elements. We review progress in unraveling one of the outstanding mysteries of modern biology: the dynamic communication of remote enhancers with target promoters in the specification of cellular identity.

organismal complexity scales not with gene number but with gene regulation


mostly coding DNA
mostly non-coding DNA
Mycoplasma
E. coli
BACTERIA
budding yeast
FUNGI
PROTISTS
Arabidopsis
bean
lily
PLANTS
Paris japonica
Drosophila
INSECTS
MOLLUSKS
shark
CARTILAGINOUS FISH
Fugu
zebra fish
marbled lungfish
BONY FISH
newt
AMPHIBIANS
REPTILES
BIRDS
human
MAMMALS
10^5
10^6
10^7
10^8
10^9
10^10
10^11
10^12
number of nucleotide pairs per haploid genome

... and what about the number of genes ?

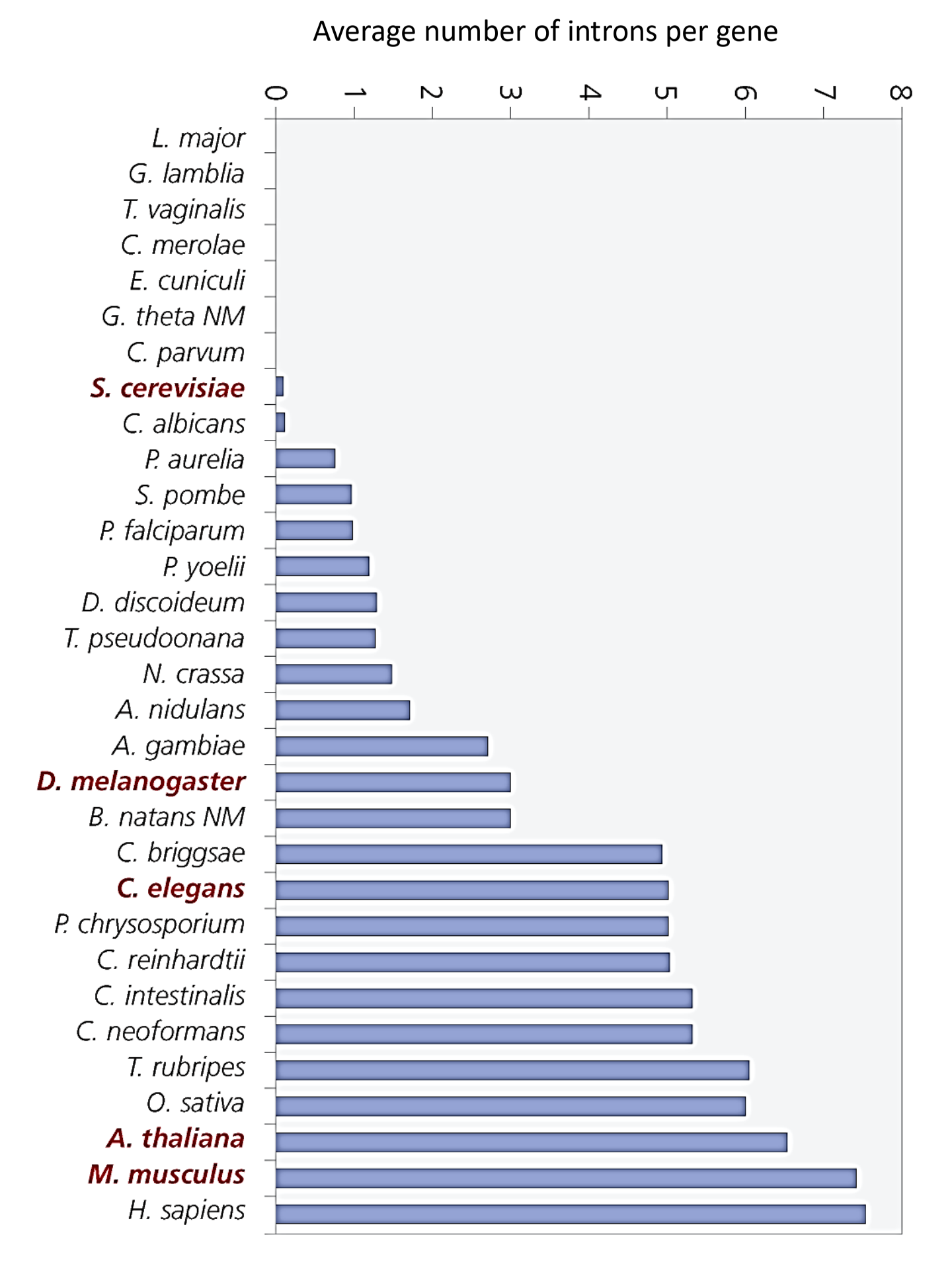
NCBI Genomes Genomic Data

Gene structure

Exon-Intron structure is present in all Eukaryotes

Hower the average number of introns, as well as the lenght of introns and central exons, varies considerably

Are introns an evolutionary feature ?


Average number of introns per gene
0
1
2
3
4
5
6
7
8
L. major
G. lamblia
T. vaginalis
C. merolae
E. cuniculi
G. theta NM
C. parvum
S. cerevisiae
C. albicans
P. aurelia
S. pombe
P. falciparum
P. yoelii
D. discoideum
T. pseudoonana
N. crassa
A. nidulans
A. gambiae
D. melanogaster
B. natans NM
C. briggsae
C. elegans
P. chrysosporium
C. reinhardtii
C. intestinalis
C. neoformans
T. rubripes
O. sativa
A. thaliana
M. musculus
H. sapiens

Averages in Human Genome: protein coding genes

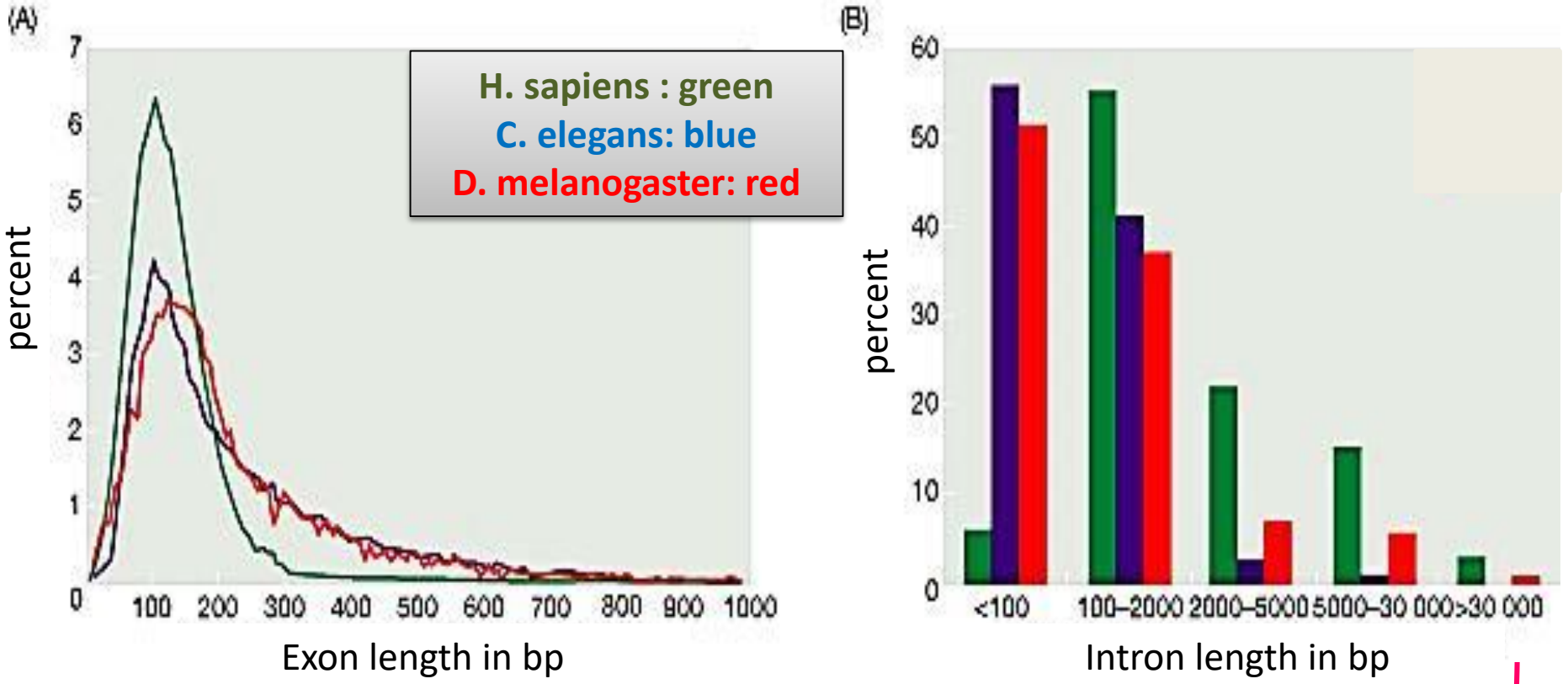
Number of exons	8.8
Exon length	170 bp (quite narrow range, 85%<200bp)
Intron length	5420 bp (large range 20bp to 100Kb)

Range:

Intron =0 (3350 single-exon genes)

Max number of Introns = 147 (NEB gene).

How exons and introns changed during evolution



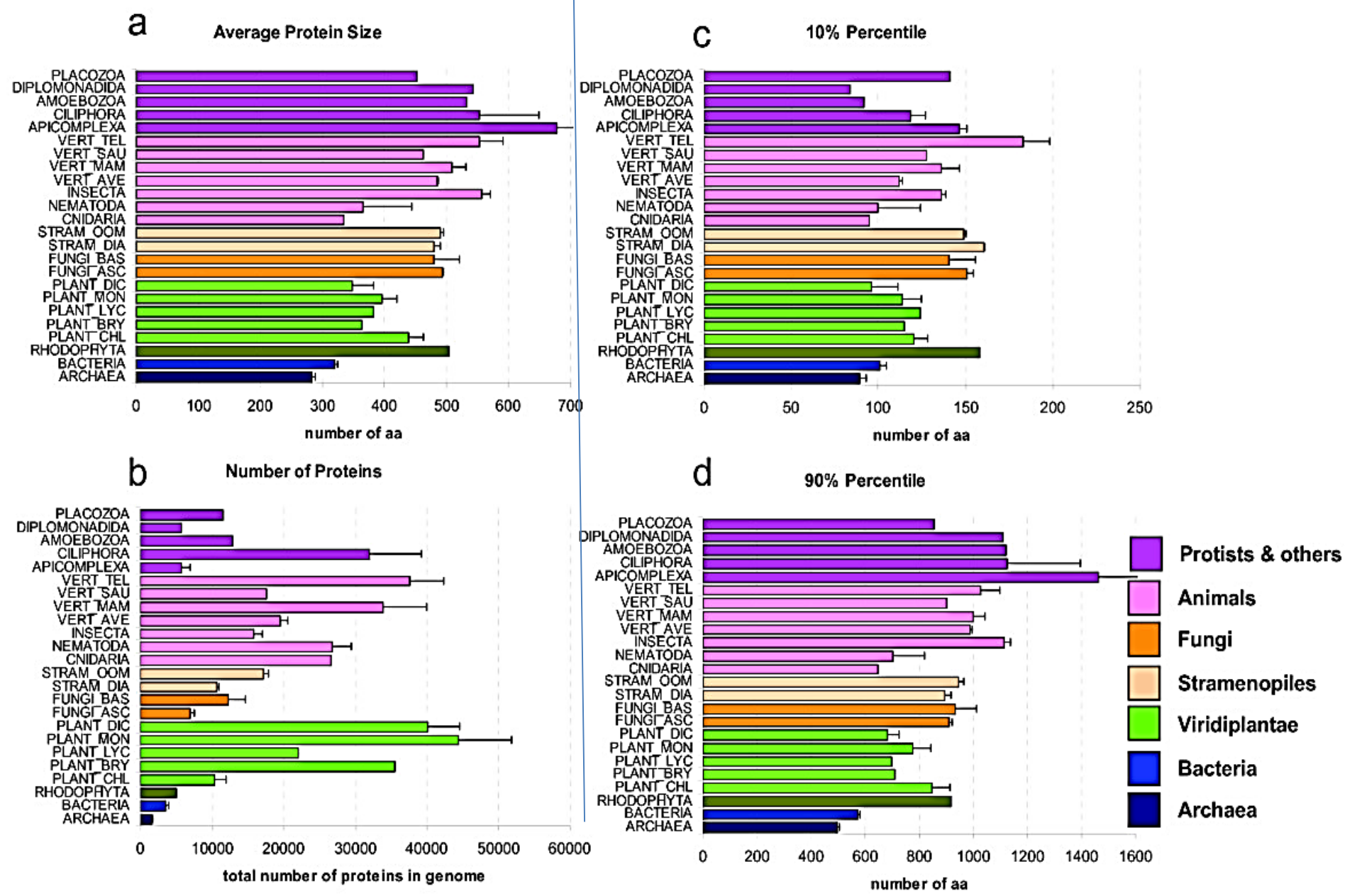
one intron in the human neurexin gene is approx. 480,000 nt !

While genes vary enormously in size from bacteria to mammals, due to intronic prevalence, **coding regions** (ORF) are quite uniform, possibly due to protein structural constraints.
Note that the absolute number of genes does not follow organism complexity.

Predicted ORF products mean size in completely sequenced organisms

Organis	size(Mb)	Mean	std	ORFs	min	Max	Tot. aa
SC	1.3	458.8	362.3	6213	25	4910	2850290
CE	97	423.3	371.6	19099	4	7829	8096713
DM	170	497.7	451.2	13695	5	7182	6816125
ATH	100	439.4	318.4	22671	8	5079	9960638
CA		479.6	333.9	6169	21	4162	2958521
HS*	3000	481.4	426.3	21724	16	6669	10484673
SP	15	456.9	353.8	3579	13	4717	1635306
PF+	100	768.9	760	421	54	4981	322400

Average a.a. ~ 128 Da in peptides: 110 Da



Summary of protein number and protein size (set 1). Comparison of the protein length attributes in species from different phylogenetic groups. Species were grouped as indicated in Table 1. a) Average protein size. b) Total number of proteins in genome. c) Average of the 10% percentiles. d) Average of the 90% percentiles. Bars indicate mean values ± standard error (SE). In panels acd the x axis indicates the number of amino acids (aa), whereas in panel b it gives the average number of proteins in those species.

Tiessen *et al. BMC Research Notes* 2012 **5**:85