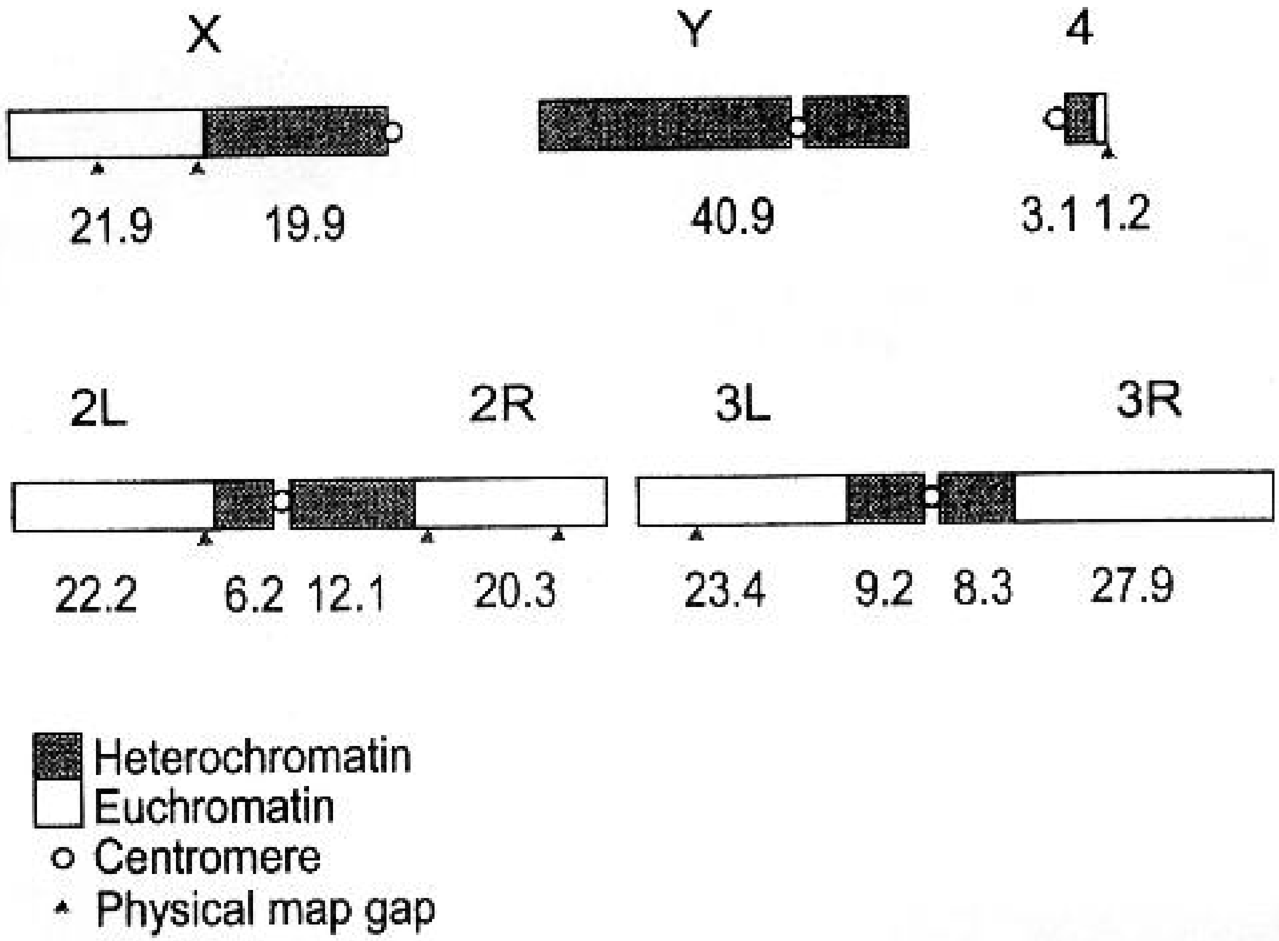
Геном эукариот

Н.Н. Колесников

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia


X
Y
4
21.9
19.9
40.9
3.1 1.2
2L
2R
3L
3R
22.2
6.2 12.1
20.3
23.4
9.2
8.3
27.9
▨ Heterochromatin
□ Euchromatin
○ Centromere
▲ Physical map gap

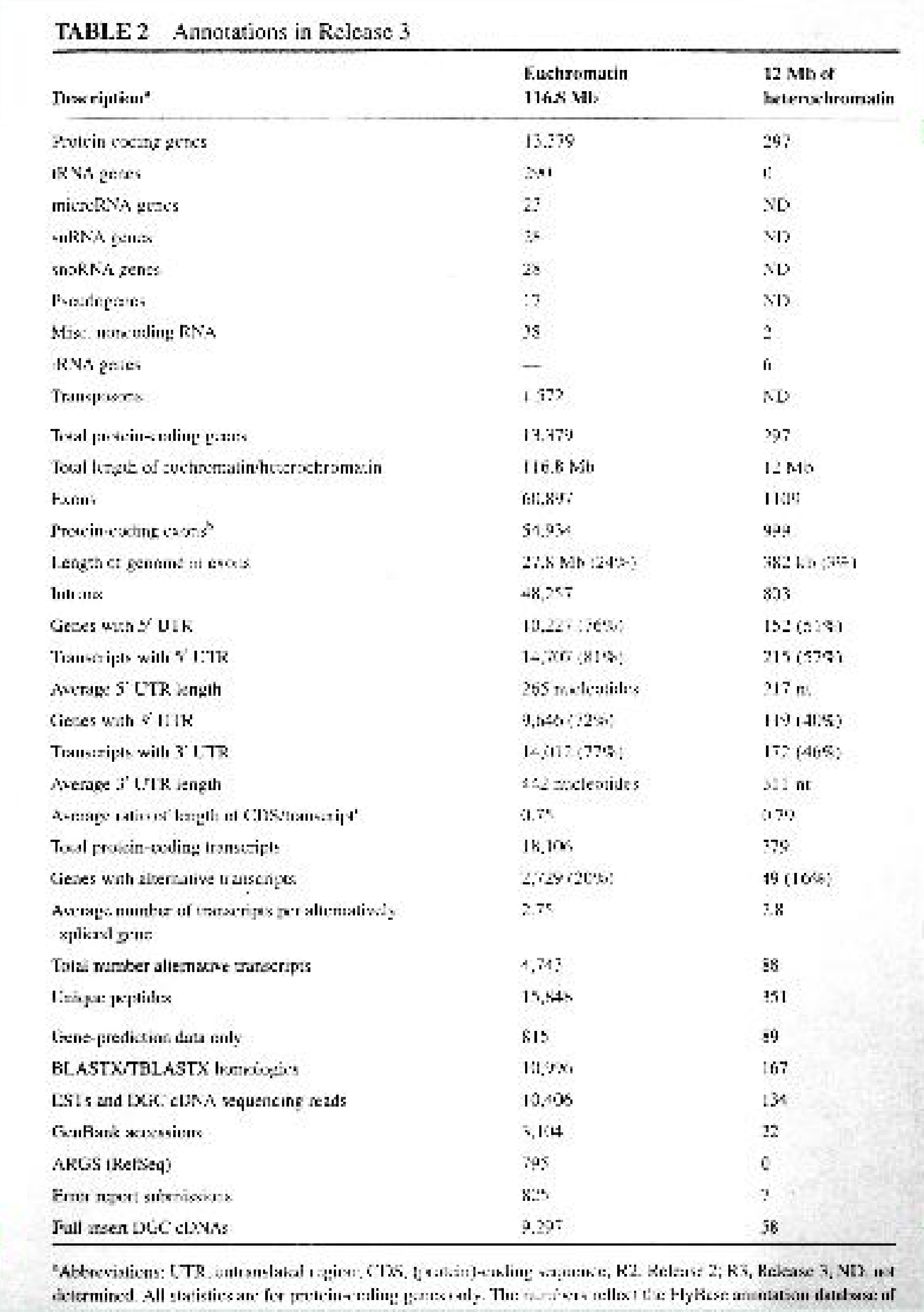


TABLE 2 Annotations in Release 3

Description[a]	Euchromatin 116.8 Mb	12 Mb of heterochromatin
Protein coding genes	13,379	297
tRNA genes	290	0
microRNA genes	27	ND
snRNA genes	35	ND
snoRNA genes	28	ND
Pseudogenes	17	ND
Misc. noncoding RNA	28	2
rRNA genes	—	6
Transposons	1,572	ND
Total protein-coding genes	13,379	297
Total length of euchromatin/heterochromatin	116.8 Mb	12 Mb
Exons	60,897	1109
Protein-coding exons[b]	54,934	999
Length of genome in exons	27.8 Mb (24%)	382 kb (3%)
Introns	48,257	803
Genes with 5′ UTR	10,227 (76%)	152 (51%)
Transcripts with 5′ UTR	14,707 (81%)	215 (57%)
Average 5′ UTR length	265 nucleotides	217 nt
Genes with 3′ UTR	9,646 (72%)	119 (40%)
Transcripts with 3′ UTR	14,017 (77%)	172 (46%)
Average 3′ UTR length	442 nucleotides	511 nt
Average ratio of length of CDS/transcript[c]	0.75	0.79
Total protein-coding transcripts	18,106	379
Genes with alternative transcripts	2,729 (20%)	49 (16%)
Average number of transcripts per alternatively spliced gene	2.75	3.8
Total number alternative transcripts	4,742	88
Unique peptides	15,848	351
Gene-prediction data only	815	89
BLASTX/TBLASTX homologies	10,986	167
ESTs and DGC cDNA sequencing reads	10,406	134
GenBank accessions	5,104	22
ARGS (RefSeq)	795	0
Error report submissions	825	2
Full insert DGC cDNAs	9,297	58

[a]Abbreviations: UTR, untranslated region; CDS, (protein)-coding sequence; R2, Release 2; R3, Release 3; ND, not determined. All statistics are for protein-coding genes only. The numbers reflect the FlyBase annotation database of

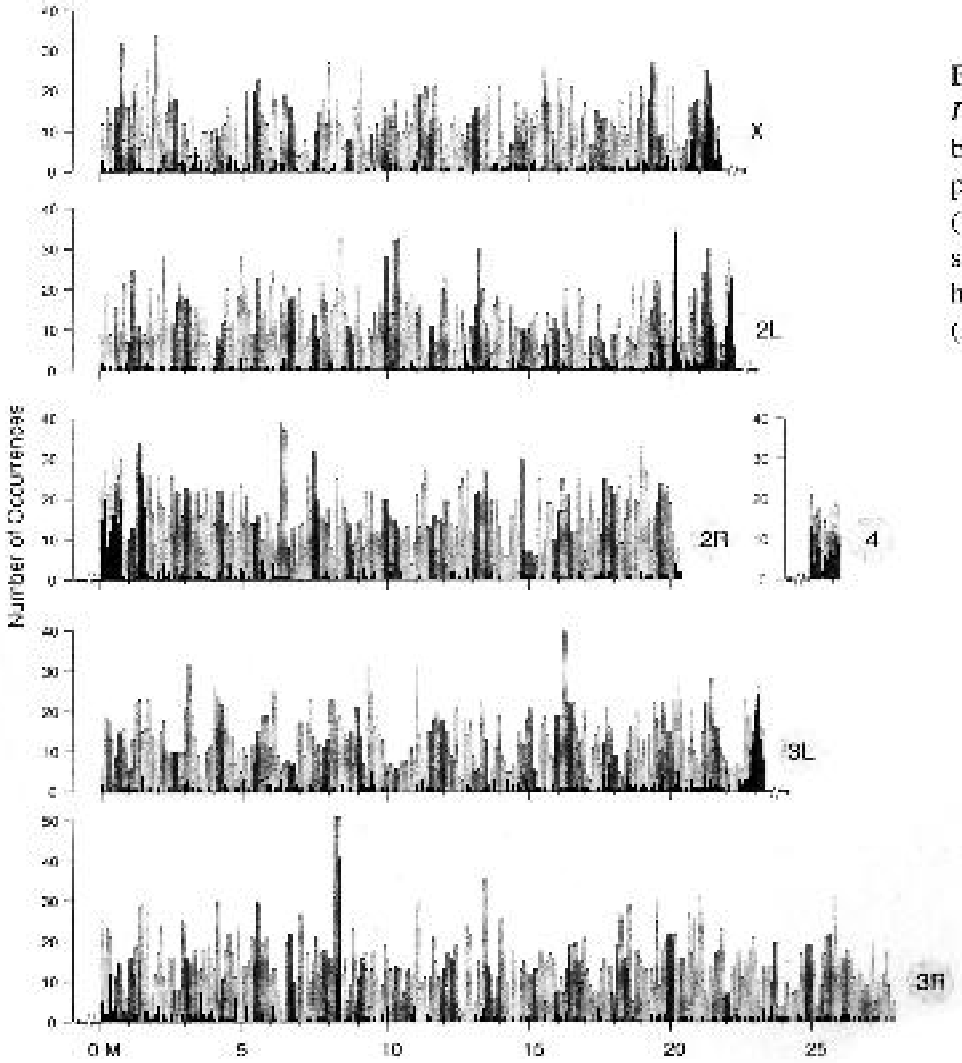
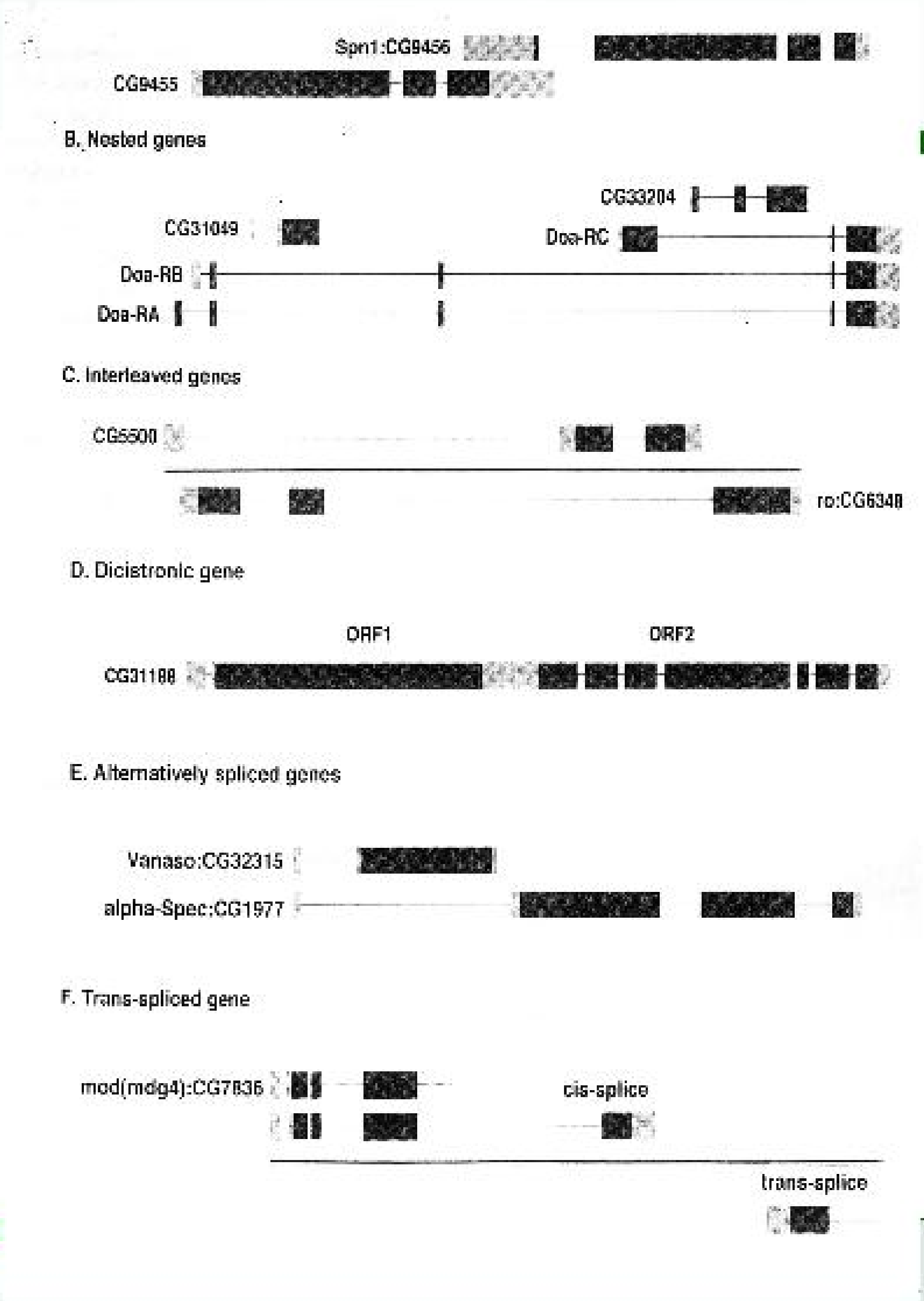


Figure 2 Distribution of protein-coding genes and transposable elements in the *Drosophila melanogaster* euchromatic genome. Each chromosome arm is represented by a black horizontal line with a circle indicating its centromere. The number of transposable elements (*black*) and protein-coding genes (*gray*) is shown for 100-kilobase (kb) windows along each chromosome arm. Although there is local variation, gene density is rather uniform along the chromosome arms, whereas transposable elements are highly enriched at the centromere-proximal regions of each arm. A scale in megabases (Mbs) is shown at the bottom of the figure.



B. Nested genes

C. Interleaved genes

D. Dicistronic gene

E. Alternatively spliced genes

F. Trans-spliced gene

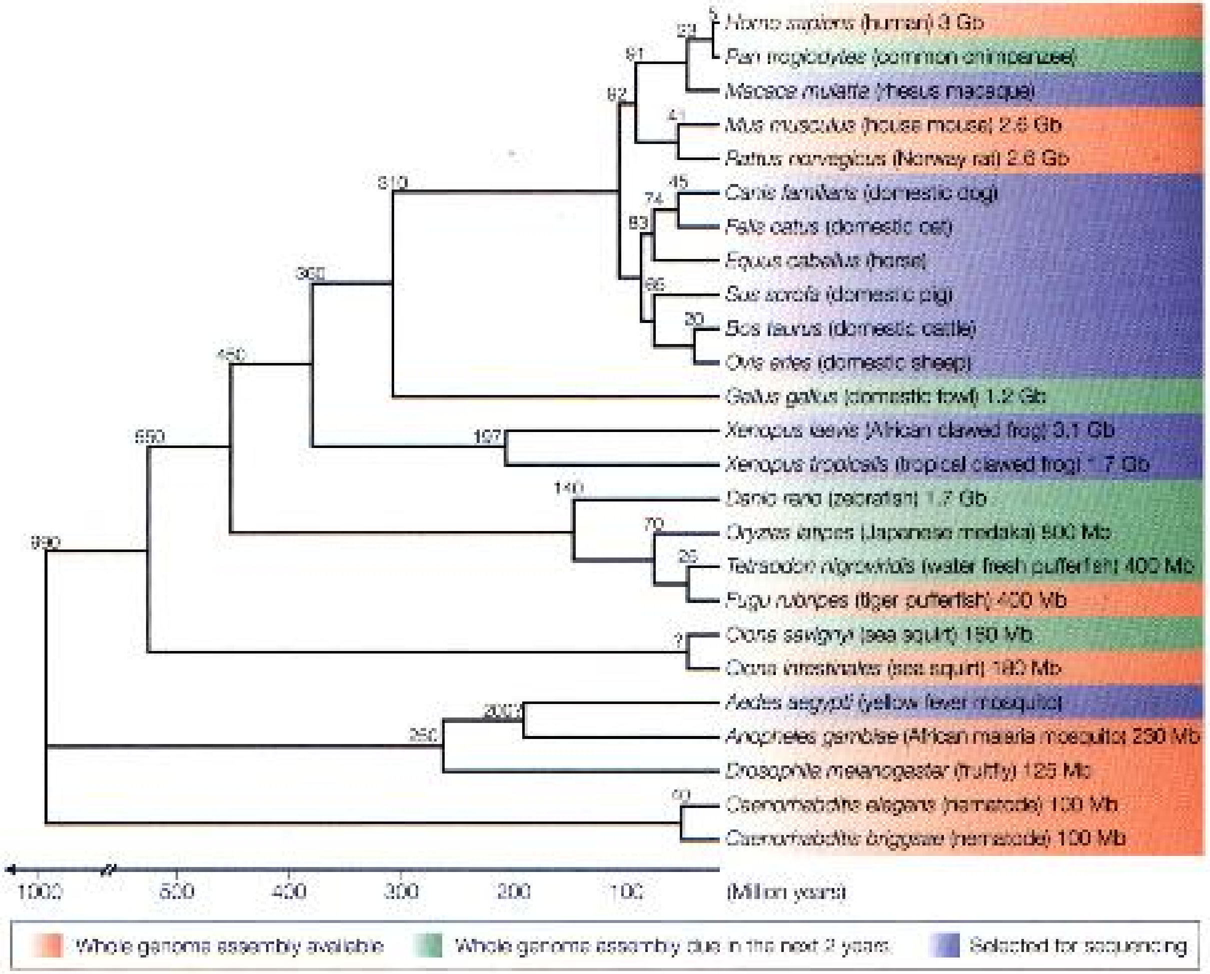

5
23
91
92
41
310
45
74
83
360
88
20
450
197
550
140
70
25
990
?
200
250
50
Homo sapiens (human) 3 Gb
Pan troglodytes (common chimpanzee)
Macaca mulatta (rhesus macaque)
Mus musculus (house mouse) 2.6 Gb
Rattus norvegicus (Norway rat) 2.8 Gb
Canis familiaris (domestic dog)
Felis catus (domestic cat)
Equus caballus (horse)
Sus scrofa (domestic pig)
Bos taurus (domestic cattle)
Ovis aries (domestic sheep)
Gallus gallus (domestic fowl) 1.2 Gb
Xenopus laevis (African clawed frog) 3.1 Gb
Xenopus tropicalis (tropical clawed frog) 1.7 Gb
Danio rerio (zebrafish) 1.7 Gb
Oryzias latipes (Japanese medaka) 800 Mb
Tetraodon nigroviridis (water fresh pufferfish) 400 Mb
Fugu rubripes (tiger pufferfish) 400 Mb
Ciona savignyi (sea squirt) 180 Mb
Ciona intestinalis (sea squirt) 160 Mb
Aedes aegypti (yellow fever mosquito)
Anopheles gambiae (African malaria mosquito) 230 Mb
Drosophila melanogaster (fruitfly) 125 Mb
Caenorhabditis elegans (nematode) 100 Mb
Caenorhabditis briggsae (nematode) 100 Mb
1000
500
400
300
200
100
(Million years)
🟥 Whole genome assembly available
🟩 Whole genome assembly due in the next 2 years
🟦 Selected for sequencing

Figure 1 | **Evolutionary relationship between metazoans that are sequenced or due for sequencing.** The simplified phylogenetic relationships between the metazoans for which the complete, or nearly complete, genome sequences are available or will be available soon. Evolutionary distances (in million years) and genome sizes are based on REFS 8,10,13,98–100.

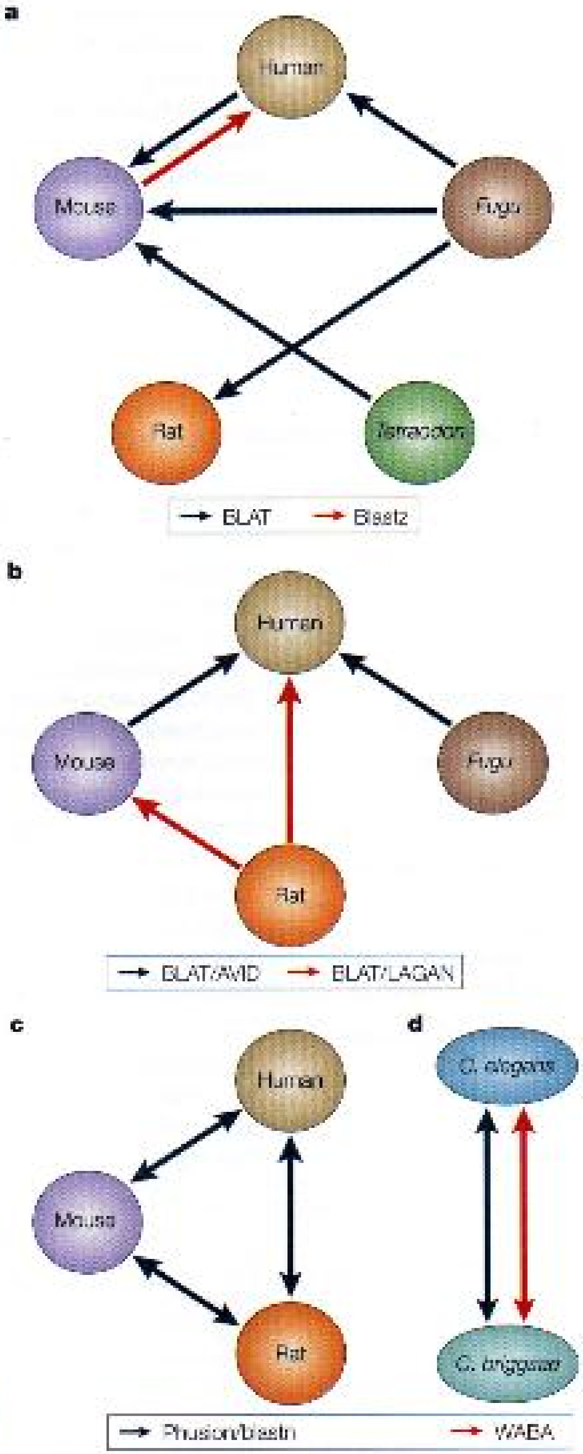
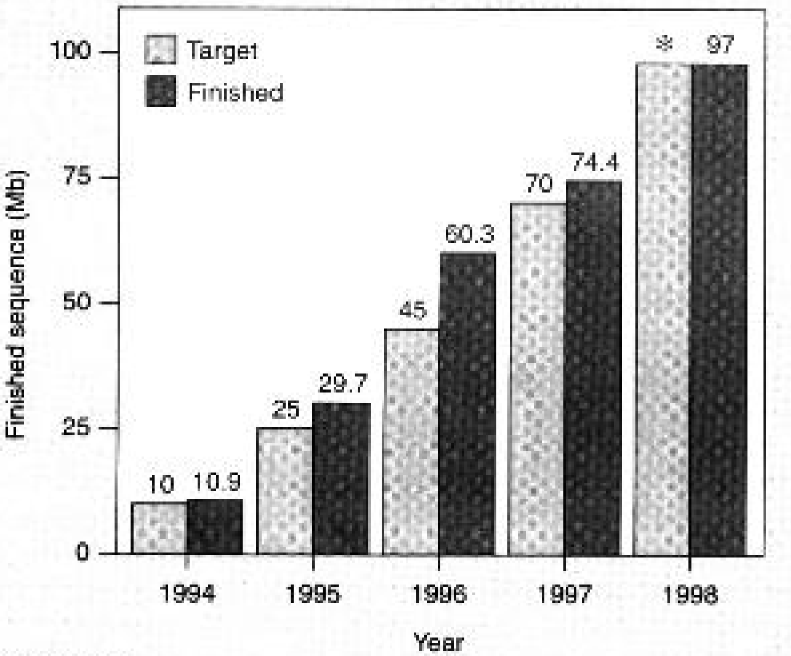


Figure 2 | **Whole-genome alignments available online.** Whole-genome precomputed alignments are available from the web sites of several different groups. Each group uses a different alignment method. The figure gives an overview of what is provided by each group and the methods used. The direction of the arrows indicate the species that are used as the reference for the display. For example, in the rat–mouse pair in panel **b**, the arrow pointing to mouse indicates that rat alignments can be seen on the mouse genome (but not the opposite), and in the human–mouse pair in panel **c**, the arrows pointing to both species indicate that human alignments can been seen on the mouse genome, and vice versa. Alignments can be found at the University of California, Santa Cruz UCSC Genome Browser and the Penn State Bioinformatics Group (**a**), the Berkeley Genome Pipeline and the ECR Browser (**b**), the Ensembl Genome Browser (**c** and **d**) and WormBase (**d**). The figure represents a snapshot of the alignments that are available at the time of writing.

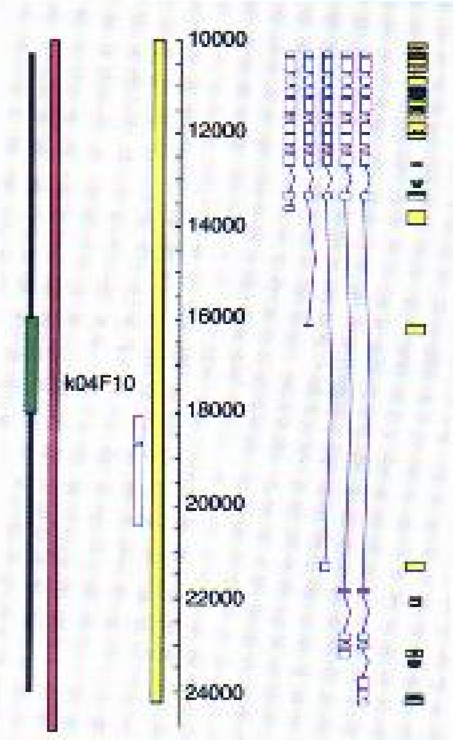
FIGURE 1. Sequencing project



Trends in Genetics

Sequencing targets and progress in the *C. elegans* genome sequencing project. Asterisk indicates target completion date.

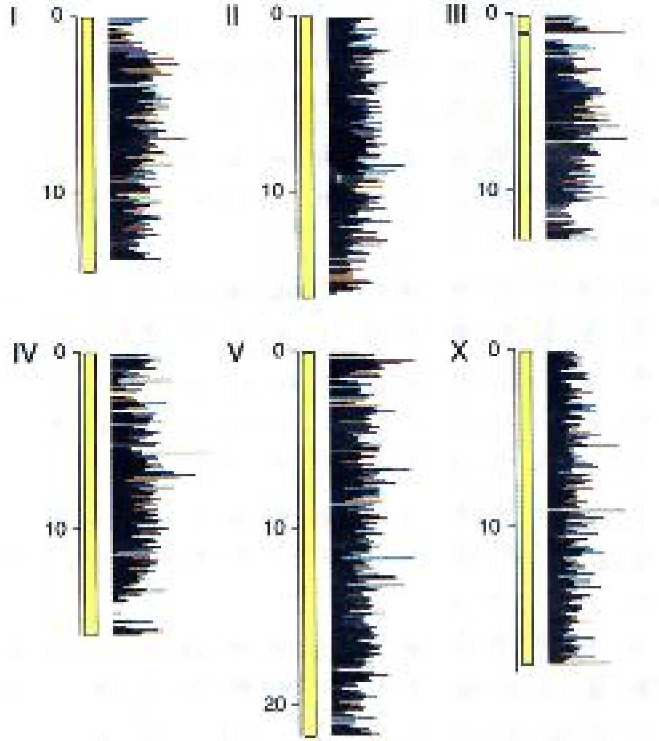
FIGURE 4. Alternatively spliced genes



Trends in Genetics

An ACeDB sequence display of a portion of the *C. elegans* cosmid K04F10 shows the multiple, alternatively spliced forms of the gene *bli-4*. From left are bli-4E, bli-4A, bli-4B, bli-4C and bli-4D. All the alternatively spliced forms are confirmed either by EST matches or experimentally[25]. The displayed sequence features are described in the legend for Fig. 1.

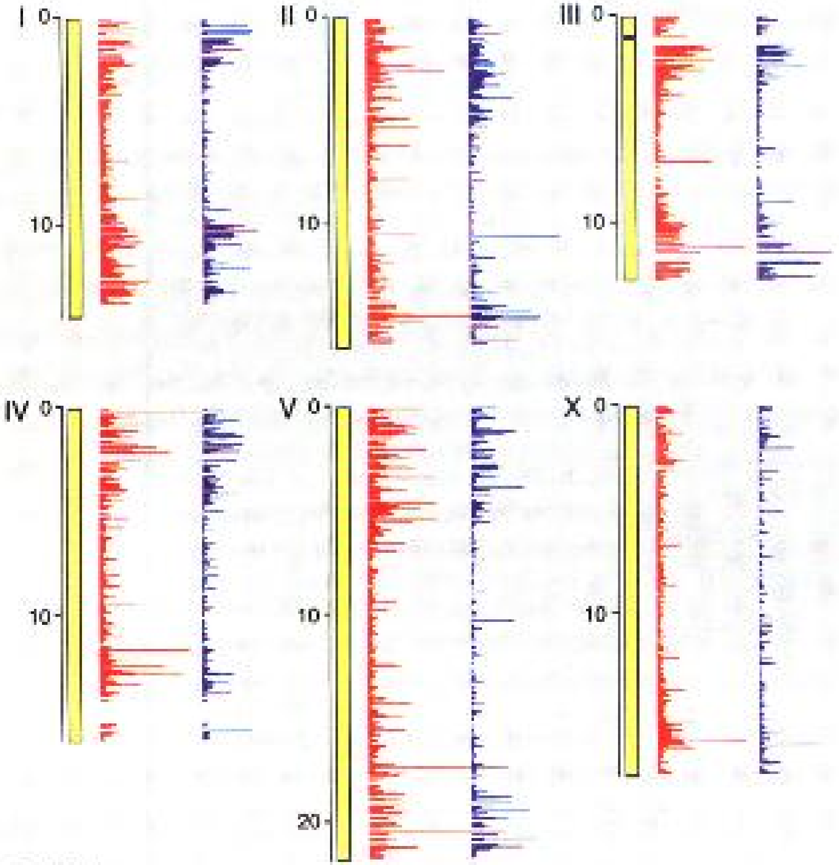
FIGURE 5. Gene distribution



Trends in Genetics

The distribution of predicted genes is plotted along each chromsome. The vertical yellow bars represent the clonal physical map of the genome (in Mb).

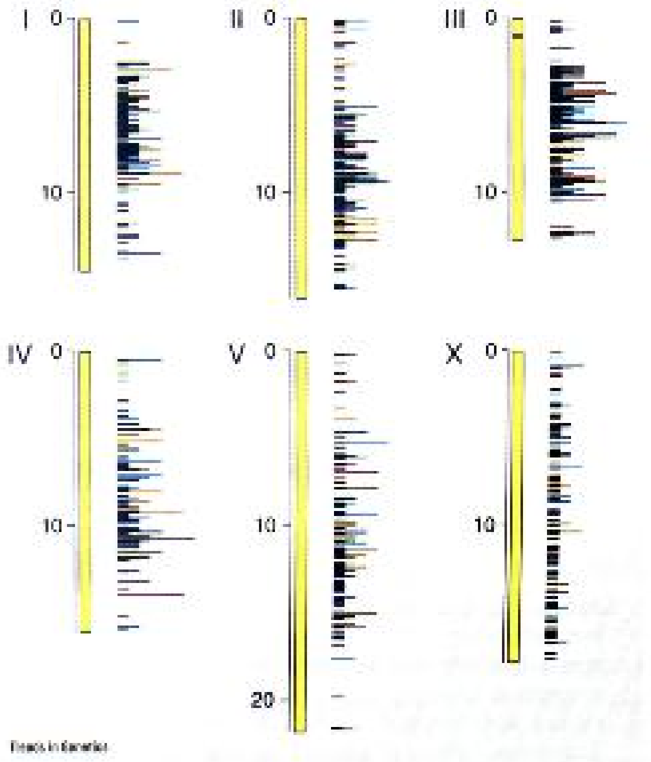
FIGURE 6. Tandem and inverted repeats



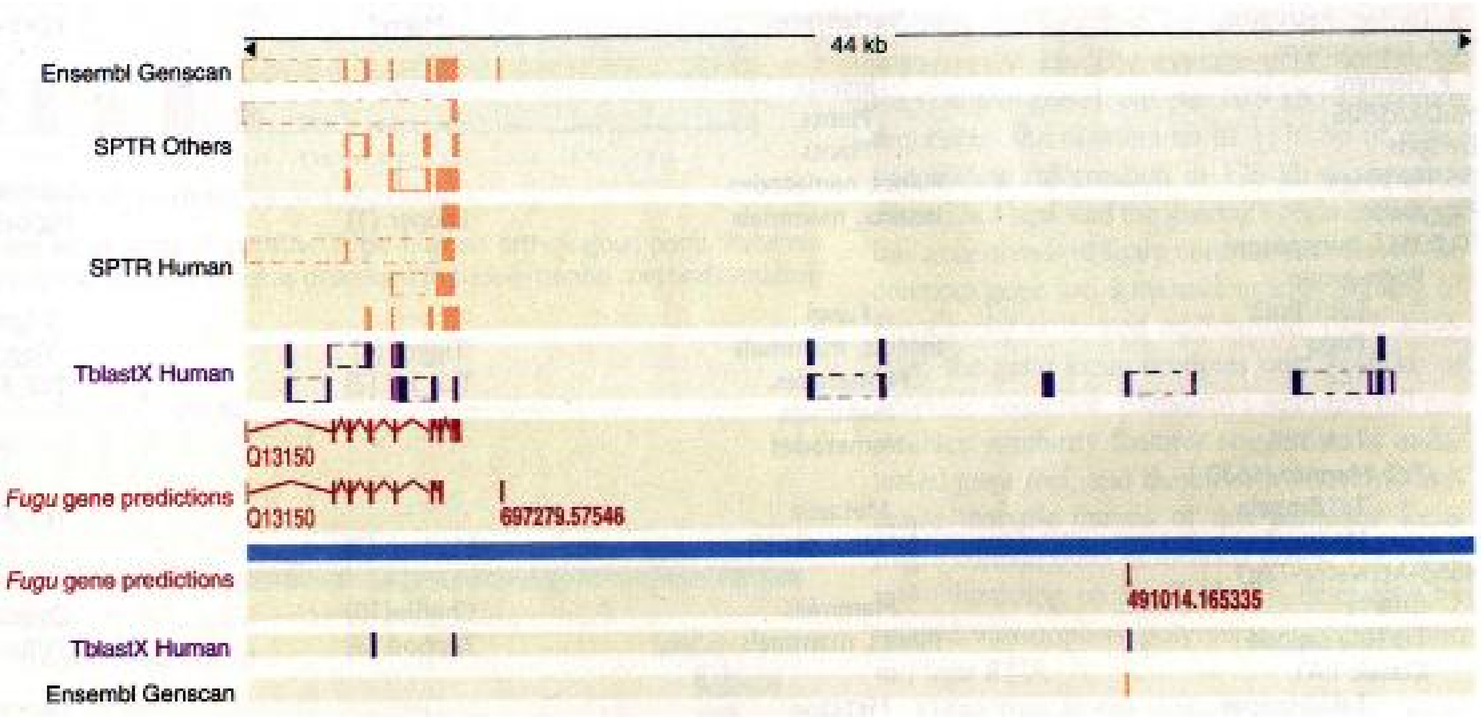
Trends in Genetics

Distribution of local tandem and inverted repeats along each of the chromosomes. Inverted repeats are shown in red while tandem repeats are blue. Both kinds of repeats are more frequent on the arms of the autosomes than in the central gene-rich regions, while they appear more uniformly distributed on the X chromosome.

FIGURE 8. Comparing *C. elegans* with *S. cerevisiae*



Distribution along each chromosome of the genes that are conserved between *S. cerevisiae* and *C. elegans*. These genes are clustered and coincide with the locations of genes with EST matches.


44 kb
Ensembl Genscan
SPTR Others
SPTR Human
TblastX Human
Q13150
Fugu gene predictions
Q13150
697279.57546
Fugu gene predictions
491014.165335
TblastX Human
Ensembl Genscan

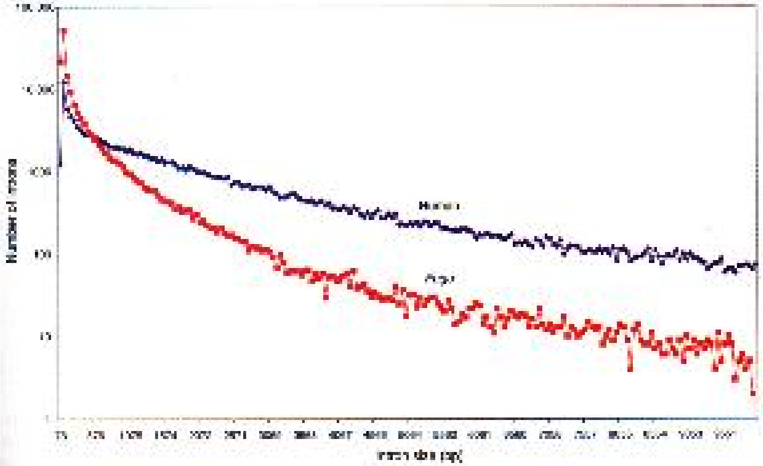


Fig. 2. Comparative frequency distribution of intron sizes in *Fugu* and human.

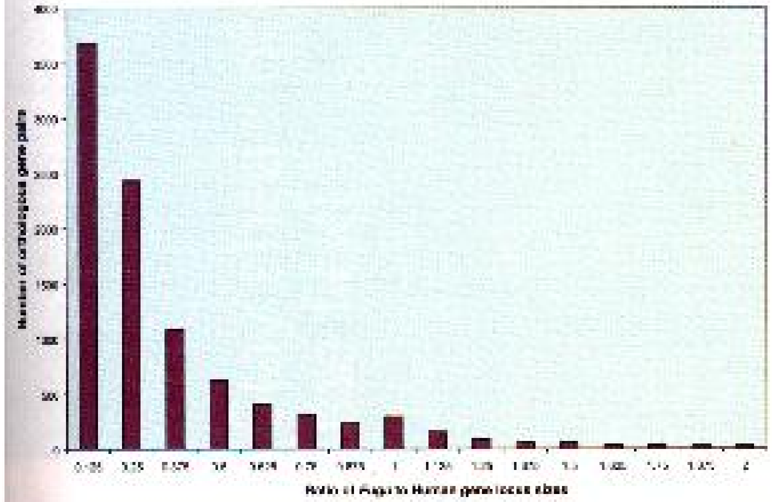


Fig. 3. Distribution of ratios for gene locus sizes of putative *Fugu*-human orthologous pairs. Putative *Fugu*-human orthologous gene pairings were determined as described in supplemental methods relating to conservation of synteny.

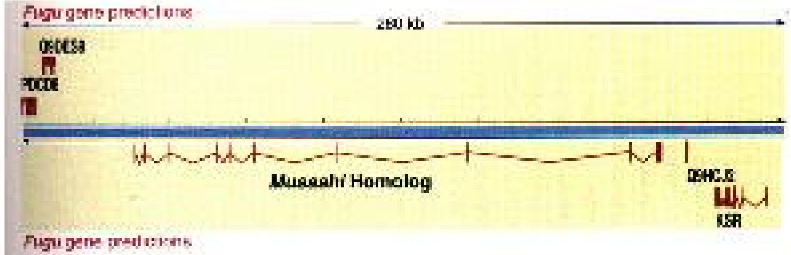
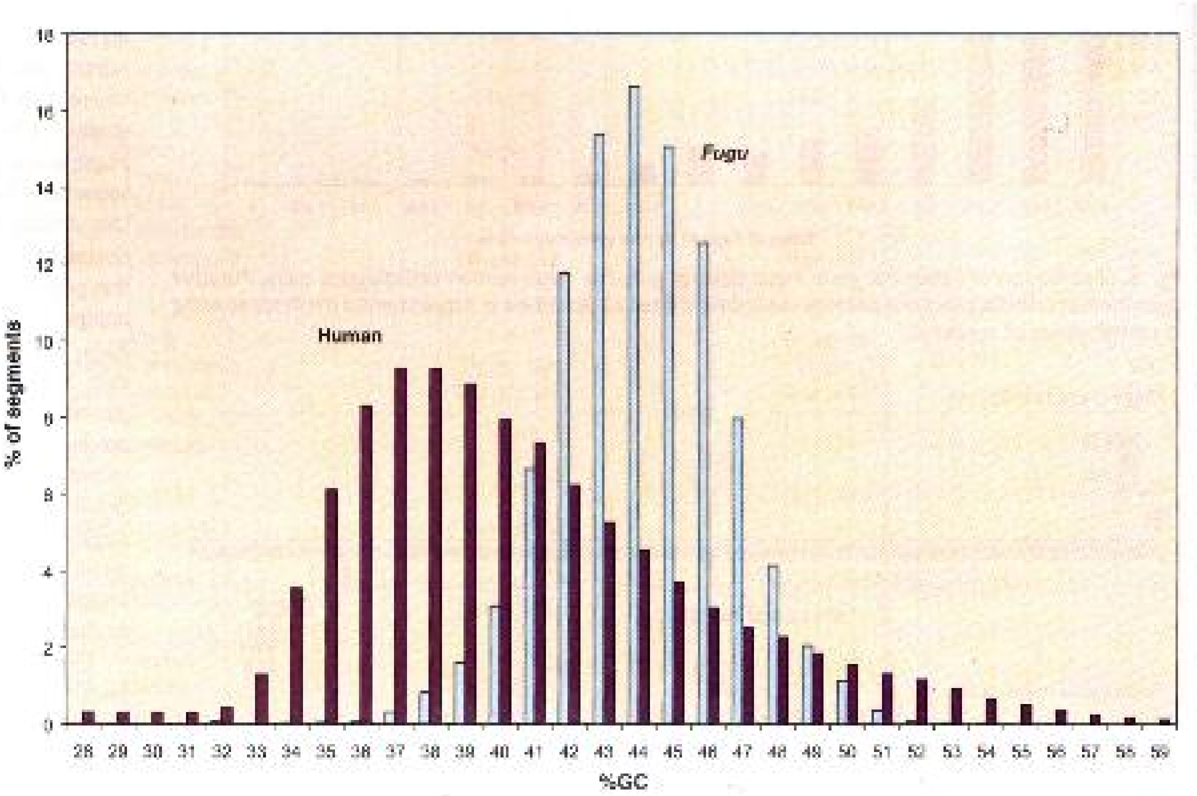


Fig. 4. Organization of a giant gene locus in the compact genome of *Fugu*. The schematic shows a region of scaffold_1 with *Fugu* gene predictions shown in brown. Exons are represented by vertical brown lines. Introns are show as v-shaped brown lines between exons.

Fig. 5. Distribution of GC content in the *Fugu* and human genomes. Sliding windows of 50 kb were used; similar conclusions were derived with windows of 25 and 100 kb (not shown).



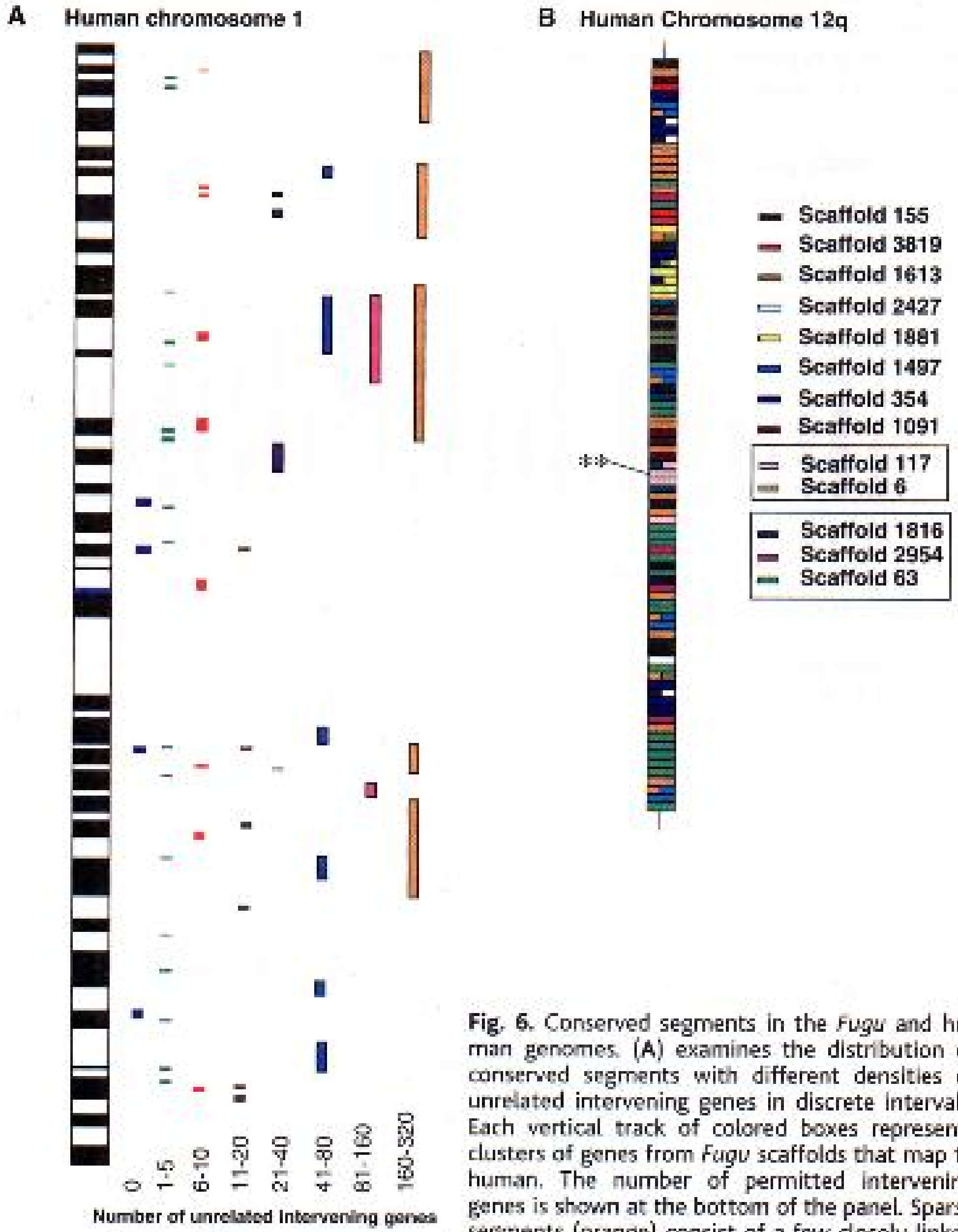


Fig. 6. Conserved segments in the *Fugu* and human genomes. (**A**) examines the distribution of conserved segments with different densities of unrelated intervening genes in discrete intervals. Each vertical track of colored boxes represents clusters of genes from *Fugu* scaffolds that map to human. The number of permitted intervening genes is shown at the bottom of the panel. Sparse segments (orange) consist of a few closely linked *Fugu* genes whose orthologs are spread over large chromosomal distances in human. Very sparse segments (>320 intervening genes) are not shown on this panel. (**B**) A detailed view of human chromosome 12q to illustrate shuffling gene order. Colored boxes represent individual genes from *Fugu* scaffolds whose orthologs on human chromosome 12q were determined through alignment by hand. The order of the orthologs along the human chromsome is shown, with the corresponding *Fugu* scaffold of origin in the key on the right. The scaffolds shown grouped together in boxes in the key are known to be linked in *Fugu*. ** indicates the position of the *Hox-c* complex on this chromosome, represented by scaffolds 117, 1327, and 1458 (the latter two are not shown in the key). Where a human gene has equally matching (co-orthologous) *Fugu* genes, this is shown as a double- or triple-colored box.

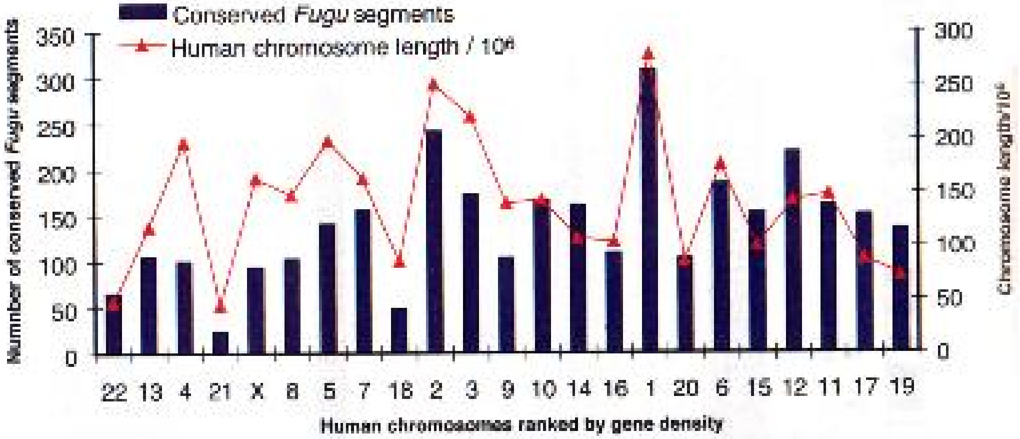


Fig. 7. Distribution of conserved segments of *Fugu* on human chromosomes ranked by gene density. The figure shows the relation between the number of conserved segments of *Fugu* on human chromosomes, the length of human chromosomes, and their gene density. Chromosome 22 is the most gene poor, chromosome 19 the most gene dense. There is no apparent relation between human chromosomal gene density and the number of segments. The distribution of conserved segments varies with human chromosomal length.

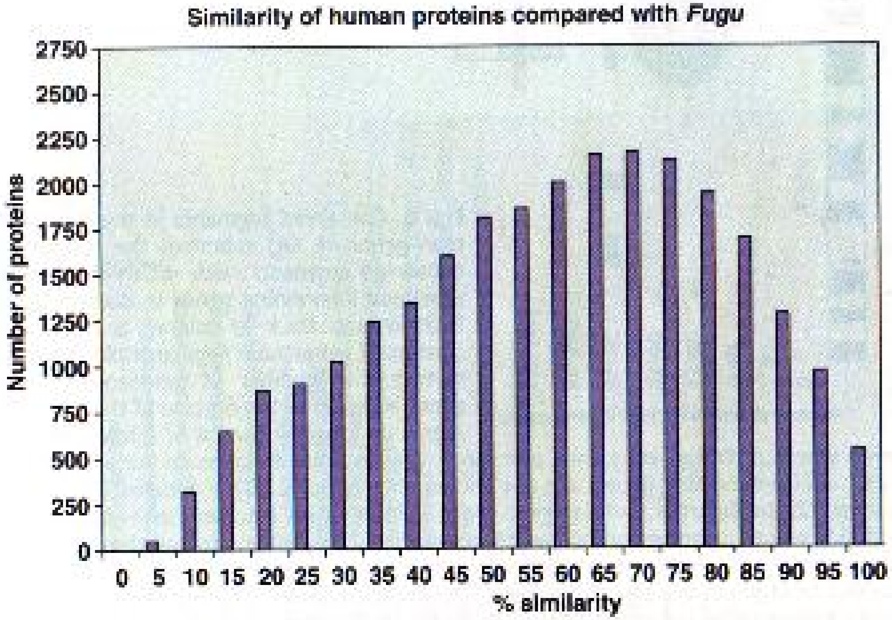


Fig. 8. Distribution of protein similarities between *Fugu* and human proteomes. Global similarities were calculated as the sum of similarities in all nonoverlapping HSPs using a BLOSUM62 matrix over the query (human) sequence length.

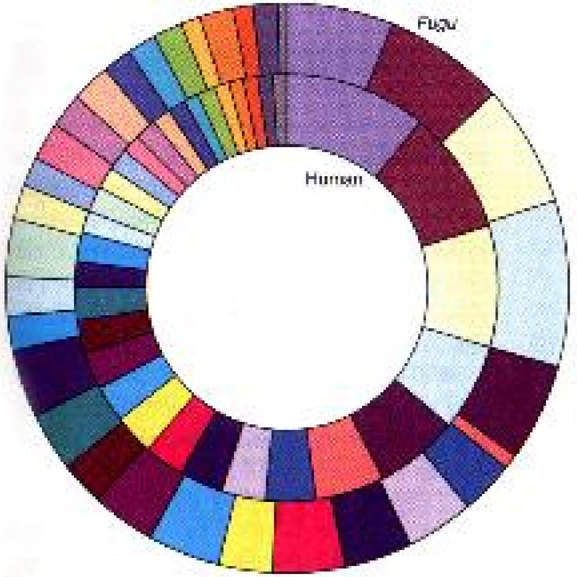


Fig. 9. Protein domains of *Fugu* and human. The schematic shows the number of gene loci with corresponding Interpro domains in *Fugu* and human for the 32 most populous families.

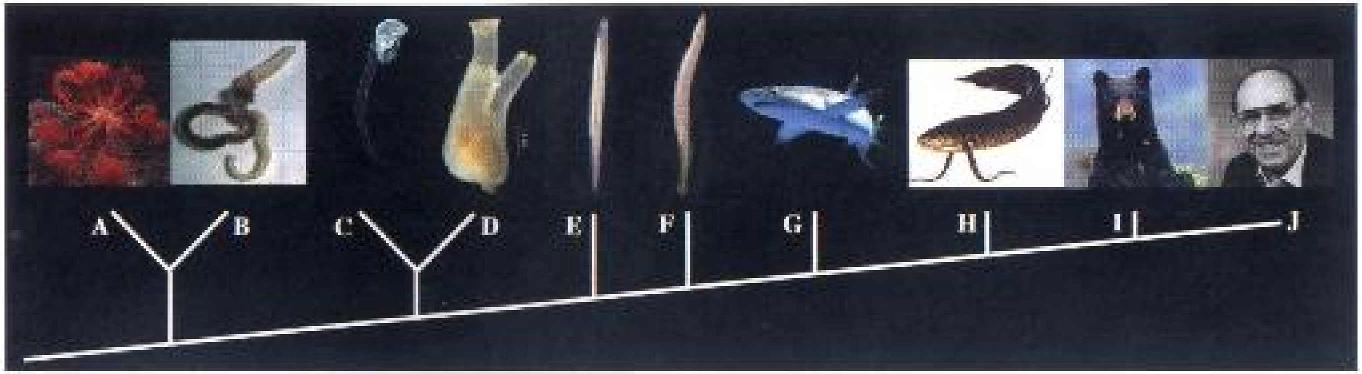
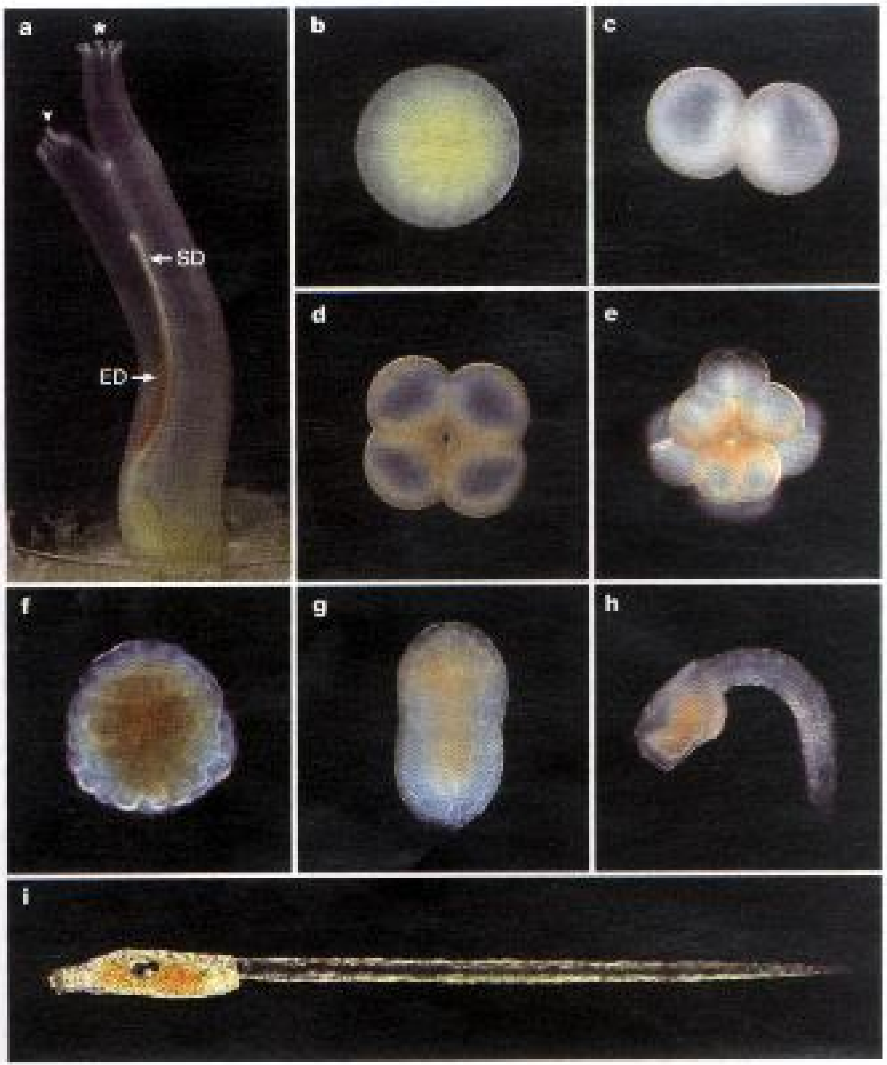


Figure 1. A deuterostome phylogeny. A, an echinoderm, a starfish. B, a hemichordate. C, a larvacean, *Oikopleura*. D, an ascidian, *Ciona*. E, a cephalochordate, amphioxus. F, a jawless fish, a lamprey. G, a gnathostome, a shark. H, a sarcopterygian fish, a lungfish. I, a mammal, a bear. J, a human, Yogi Berra.



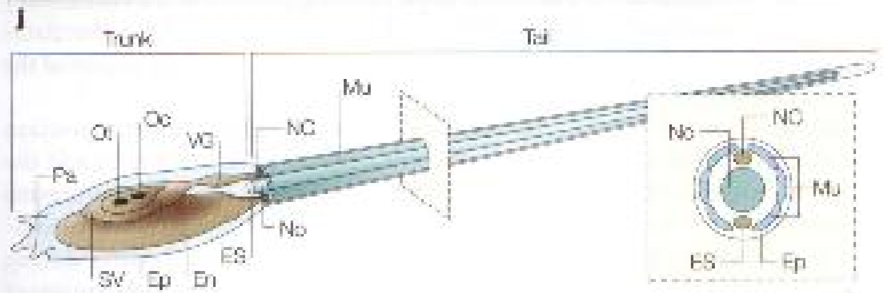


Figure 1 | **Tadpole-type larva of the ascidian *Ciona intestinalis*.** Panel **a** shows adults with incurrent siphons (arrowhead) and outcurrent siphons (asterisk) for circulating seawater from which the ascidians filter plankton and nutrient materials. The white duct is the sperm duct (SD), and the orange duct is the egg duct (ED). Panels **b–i** show the stages of embryogenesis: fertilized egg (**b**), 2-cell embryo (**c**), 4-cell embryo (**d**), 16-cell embryo (**e**), gastrula (**f**), early-tailbud embryo (**g**), mid-tailbud embryo (**h**) and tadpole larva (**i** and **j**). Embryos were dechorionated to show their outer morphology. Mesenchyme in the posterior region of the trunk is not shown. En, endoderm; Ep, epidermis; ES, endodermal strand; Mu, muscle; NC, nerve cord; No, notochord; Oc, ocellus; Ot, otolith; Pa, palps; SV, sensory vesicle; VG, visceral ganglion.

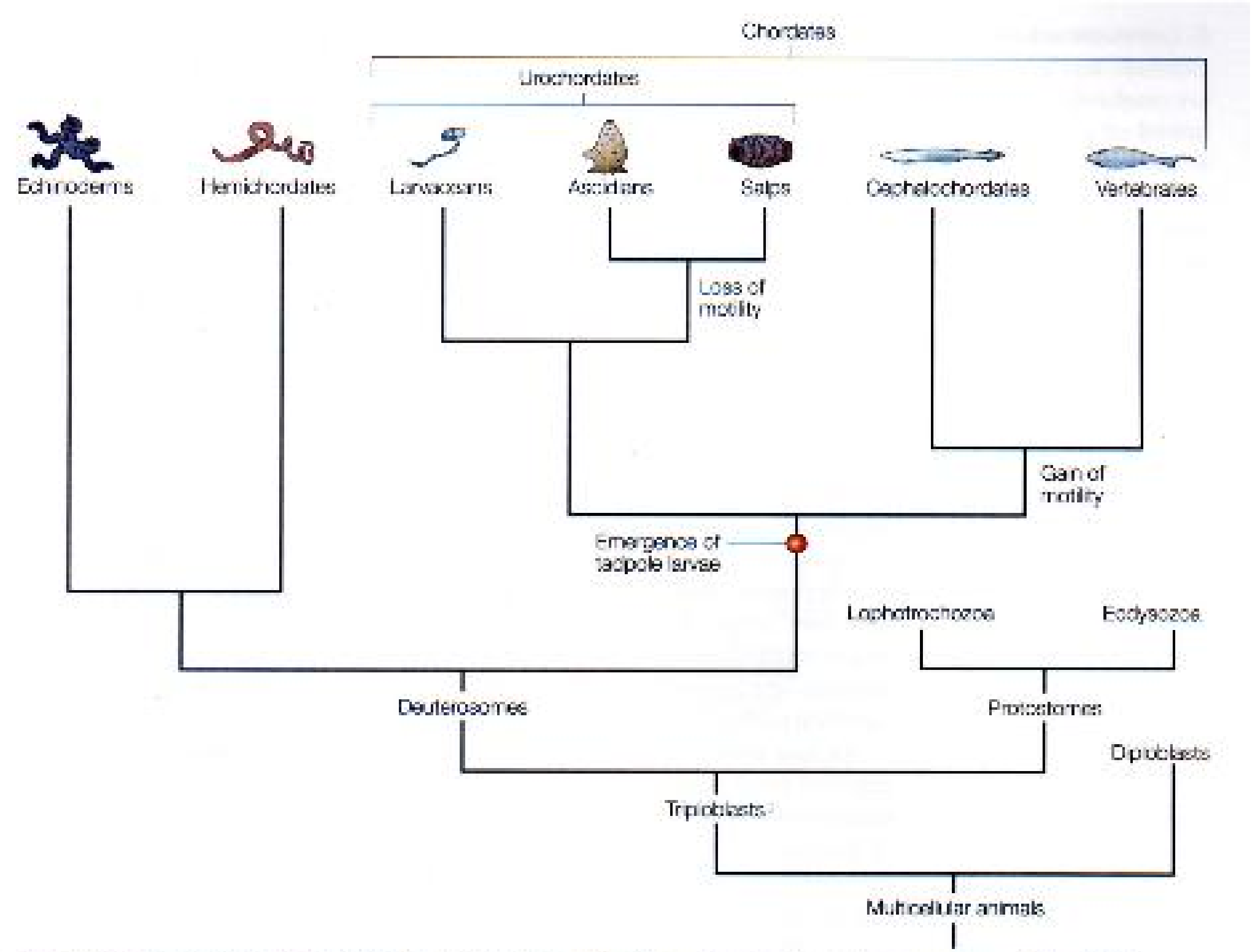


Figure 2 | **The evolution of the chordates.** The chordates comprise the urochordates, the cephalochordates and the vertebrates. They are thought to have evolved from a common ancestor shared with the non-chordate deuterostomes (the echinoderms and the hemichordates). The emergence of tadpole-type larvae was a key event in the evolution of the chordates.

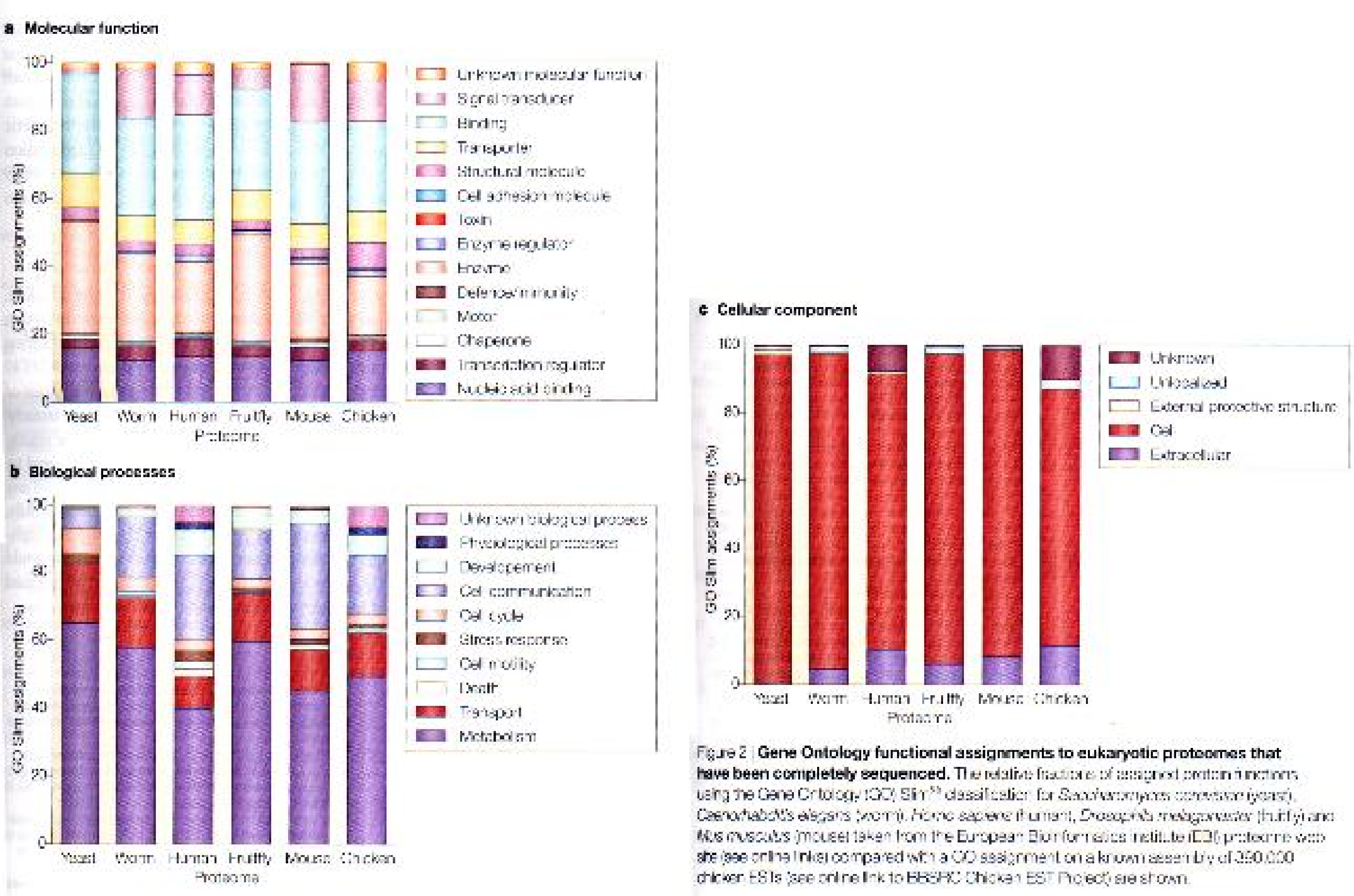


Figure 2 | **Gene Ontology functional assignments to eukaryotic proteomes that have been completely sequenced.** The relative fractions of assigned protein functions using the Gene Ontology (GO) Slim[33] classification for *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Homo sapiens* (human), *Drosophila melanogaster* (fruitfly) and *Mus musculus* (mouse) taken from the European Bioinformatics Institute (EBI) proteome web site (see online links) compared with a GO assignment on a known assembly of 390,000 chicken ESTs (see online link to BBSRC Chicken EST Project) are shown.

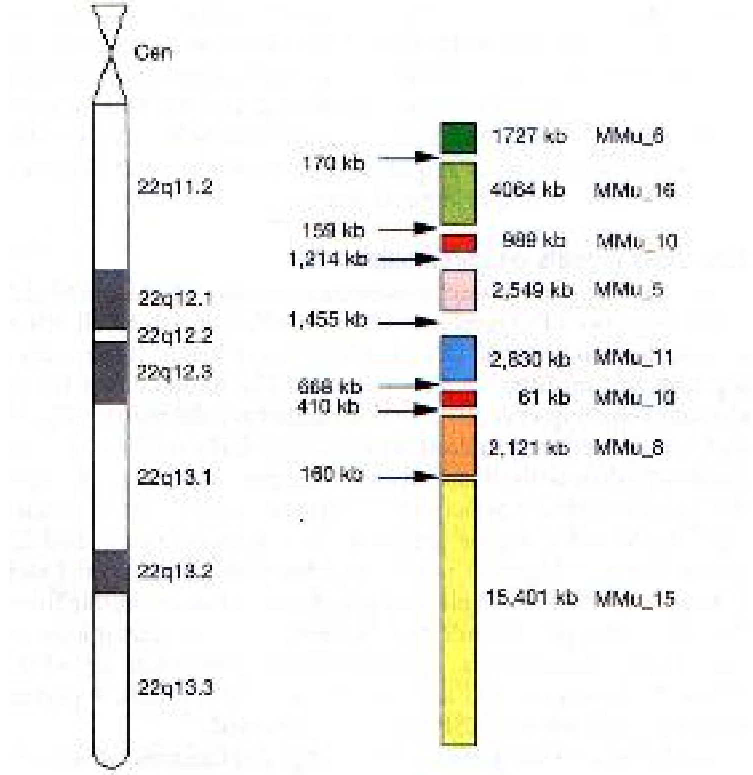
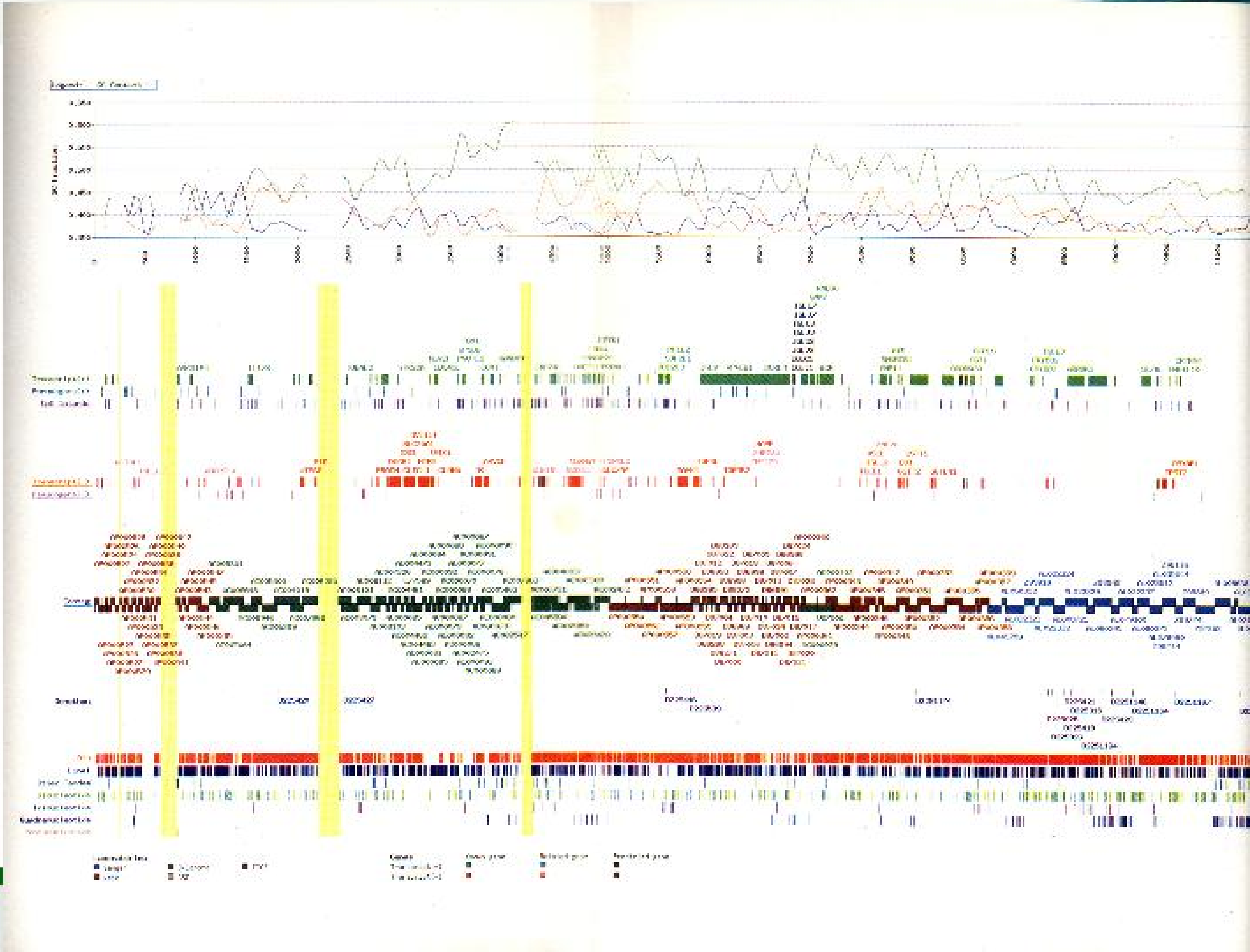


Figure 5 Regions of conserved synteny between human chromosome 22 and the mouse genome. Regions of mouse chromosomes with conserved syntany to human chromosome 22 are shown as adjacent coloured blocks, determined by the mouse map position of mouse orthologues to human chromosome 22 genes. The size of human chromosome 22 corresponding to each mouse chromosomal region is indicated in kb, as well as the size of the gap between the last orthologue in each conserved block. These data are available at http://www.sanger.ac.uk/Chr22/Mouse.



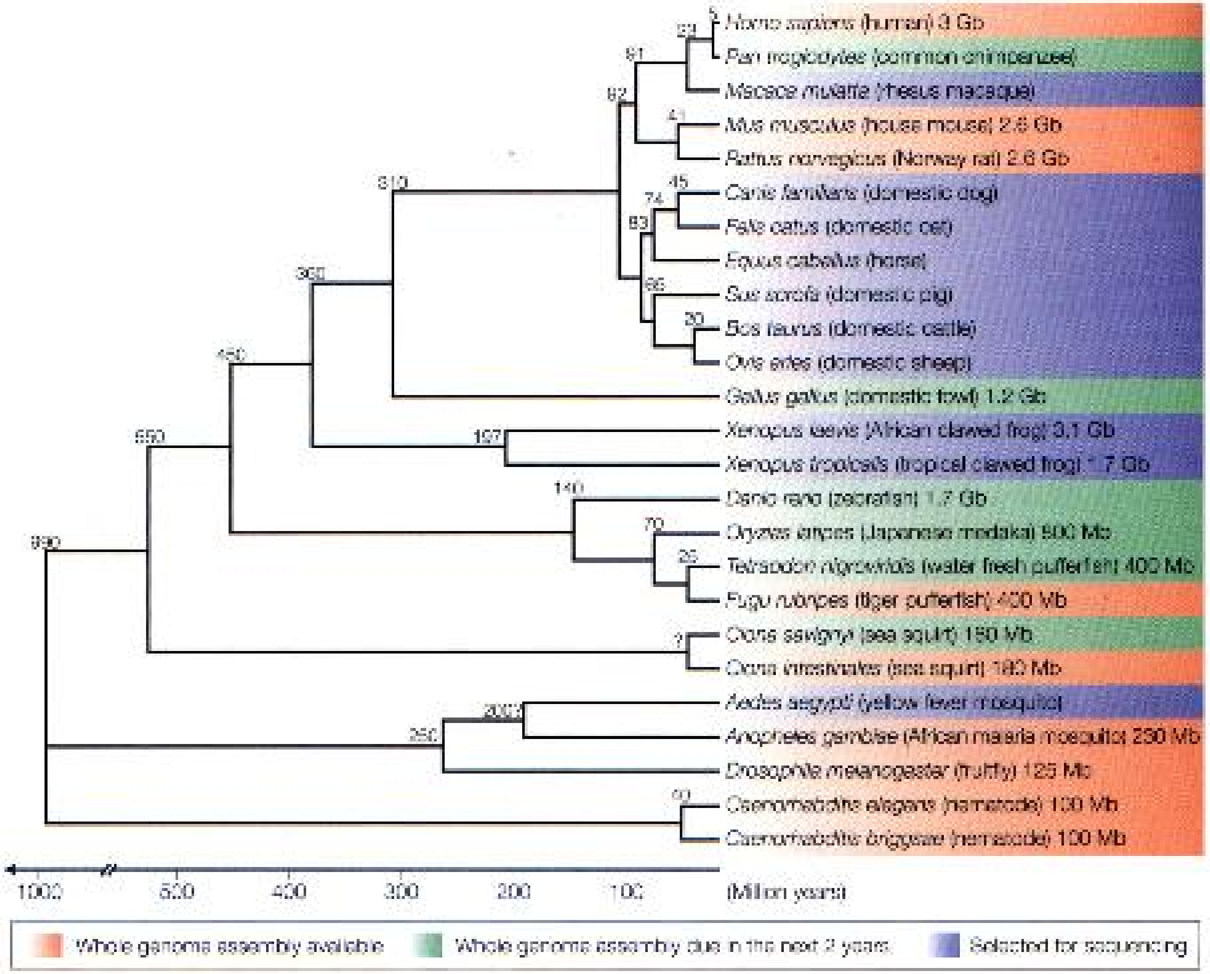

5
23
91
92
41
310
45
74
83
360
66
30
450
197
550
140
70
25
990
?
200?
250
90
Homo sapiens (human) 3 Gb
Pan troglodytes (common chimpanzee)
Macaca mulatta (rhesus macaque)
Mus musculus (house mouse) 2.6 Gb
Rattus norvegicus (Norway rat) 2.6 Gb
Canis familiaris (domestic dog)
Felis catus (domestic cat)
Equus caballus (horse)
Sus scrofa (domestic pig)
Bos taurus (domestic cattle)
Ovis aries (domestic sheep)
Gallus gallus (domestic fowl) 1.2 Gb
Xenopus laevis (African clawed frog) 3.1 Gb
Xenopus tropicalis (tropical clawed frog) 1.7 Gb
Danio rerio (zebrafish) 1.7 Gb
Oryzias latipes (Japanese medaka) 800 Mb
Tetraodon nigroviridis (water fresh pufferfish) 400 Mb
Fugu rubripes (tiger pufferfish) 400 Mb
Ciona savignyi (sea squirt) 150 Mb
Ciona intestinalis (sea squirt) 180 Mb
Aedes aegypti (yellow fever mosquito)
Anopheles gambiae (African malaria mosquito) 230 Mb
Drosophila melanogaster (fruitfly) 125 Mb
Caenorhabditis elegans (nematode) 100 Mb
Caenorhabditis briggsae (nematode) 100 Mb
1000
500
400
300
200
100
(Million years)
Whole genome assembly available
Whole genome assembly due in the next 2 years
Selected for sequencing

Figure 1 | **Evolutionary relationship between metazoans that are sequenced or due for sequencing.** The simplified phylogenetic relationships between the metazoans for which the complete, or nearly complete, genome sequences are available or will be available soon. Evolutionary distances (in million years) and genome sizes are based on REFS 8,10,13,98–100.