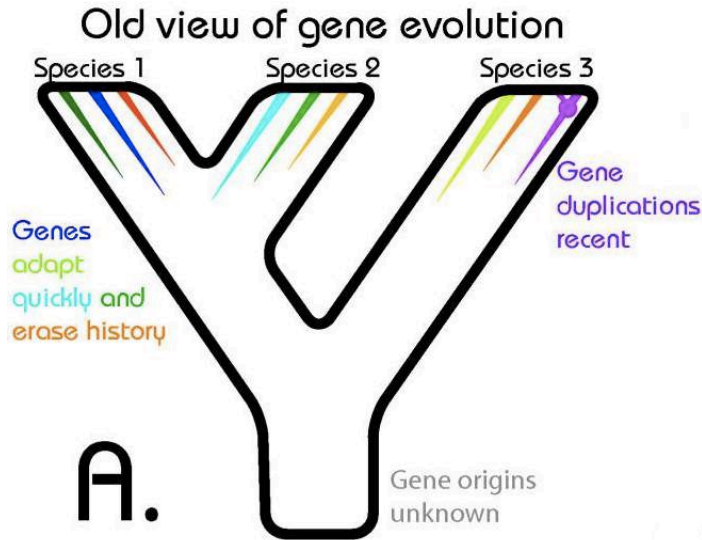


Metazoan and eukaryotic evolution through the lens of genomics



Máté Varga
(mvarga@ttk.elte.hu)

The paradigm of homologous genes

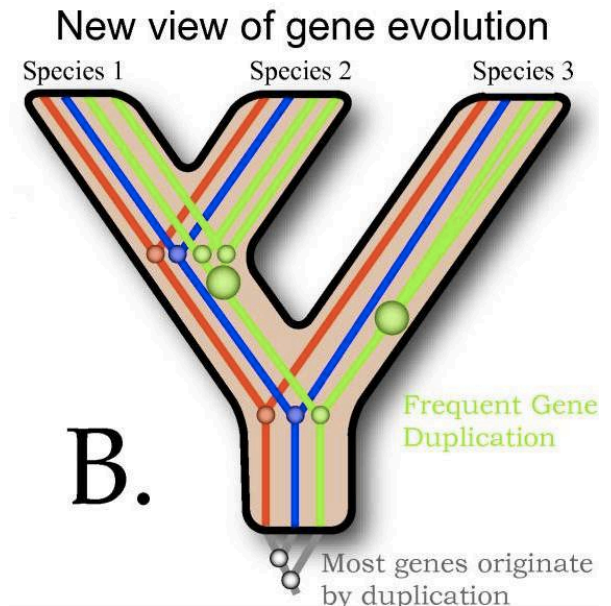


- according to the classic view of the Modern Synthesis, genes are fast changing adaptive traits of species.

"Much that has been learned about gene physiology makes it evident that the search for homologous genes is quite futile except in very close relatives". – Ernst Mayr

- after the 1960s it became obvious, that Mayr was wrong and even distantly related species share common genes (hemoglobin, cyt c).

- in the 1980s and 1990s it became also obvious that homologous genes can regulate the development of homologous organs in invertebrates and vertebrates! (*eya*, *Hox*, *tinman/nkx2.5*)



(Rose and Oakley (2007) *Biol Direct*)

The first indirect genome comparisons



Table 1. Differences in amino acid sequences of human and chimpanzee polypeptides. Lysozyme, carbonic anhydrase, albumin, and transferrin have been compared immunologically by the microcomplement fixation technique. Amino acid sequences have been determined for the other proteins. Numbers in parentheses indicate references for each protein.

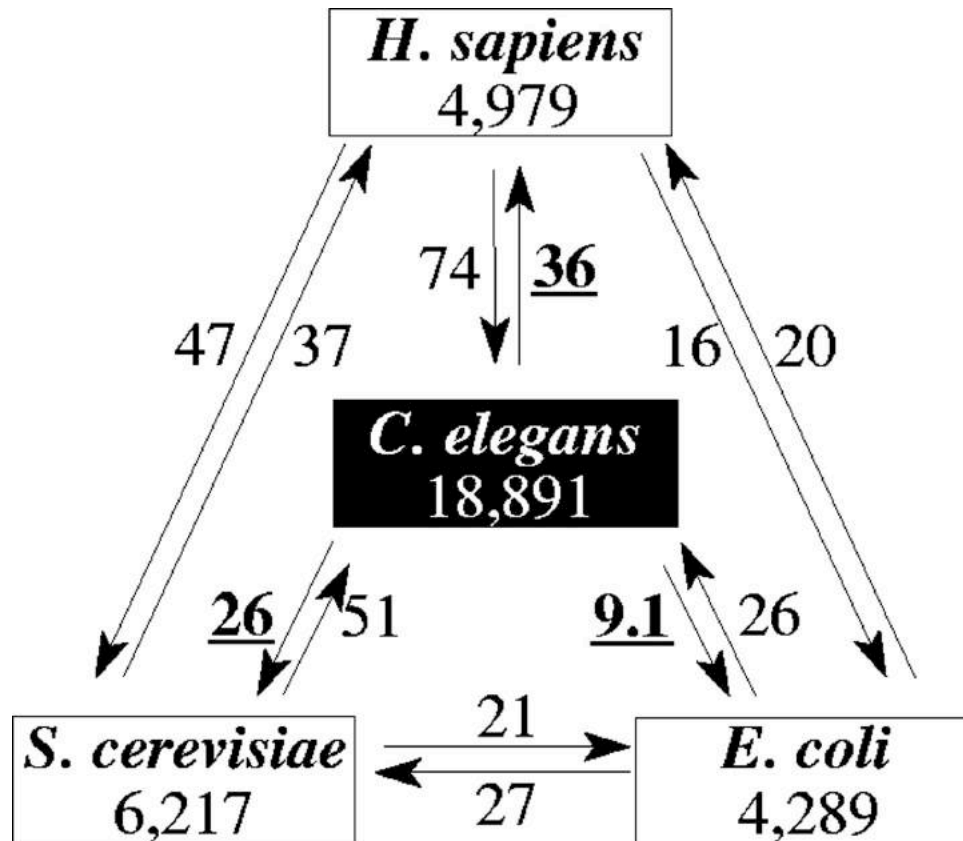
Protein	Amino acid differences	Amino acid sites
Fibrinopeptides A and B (3)	0	30
Cytochrome c (4)	0	104
Lysozyme (13)	~0	130
Hemoglobin α (4)	0	141
Hemoglobin β (4)	0	146
Hemoglobin $^A\gamma$ (5, 6)	0	146
Hemoglobin $^G\gamma$ (5, 6)	0	146
Hemoglobin δ (5, 8)	1	146
Myoglobin (7)	1	153
Carbonic anhydrase (4, 12)	~3	264
Serum albumin (10)	~6	580
Transferrin (11)	~8	647
Total	~19	2633

“A relatively small number of genetic changes in systems controlling the expression of genes may account for the major organismal differences between humans and chimpanzees.”

(King and Wilson (1975) *Science*)

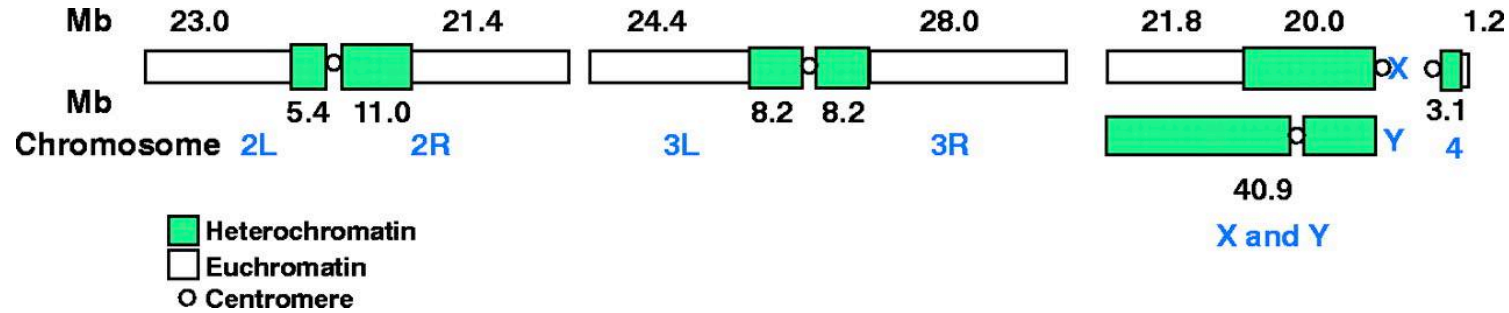
Caenorhabditis elegans

- Six chromosome pairs, 97 Mb DNA, 20,100 predicted protein coding genes (plus ~16,000 ncRNAs, but we found out only lately about these – their role is still being established/debated)



(The *C. elegans* Sequencing Consortium (1998) *Science*)

Drosophila melanogaster



- ~180Mb genome size, ~13,600 (protein coding) genes

- thus the shorter *C. elegans* genome has more genes! => no direct relation between organismal complexity and the number of genes

However: in *Drosophila* there are numerous alternative transcripts!!

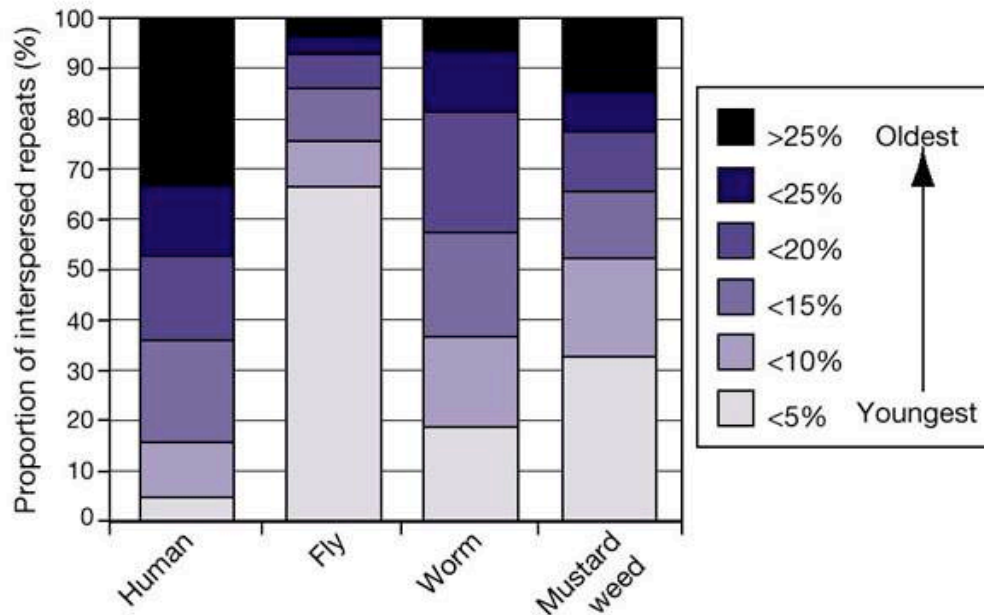
The human genome is rich in ancient transposable elements (TEs)



Table 12 Number and nature of interspersed repeats in eukaryotic genomes

	Human		Fly		Worm		Mustard weed	
	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families
LINE/SINE	33.40%	6	0.70%	20	0.40%	10	0.50%	10
LTR	8.10%	100	1.50%	50	0.00%	4	4.80%	70
DNA	2.80%	60	0.70%	20	5.30%	80	5.10%	80
Total	44.40%	170	3.10%	90	6.50%	90	10.50%	160

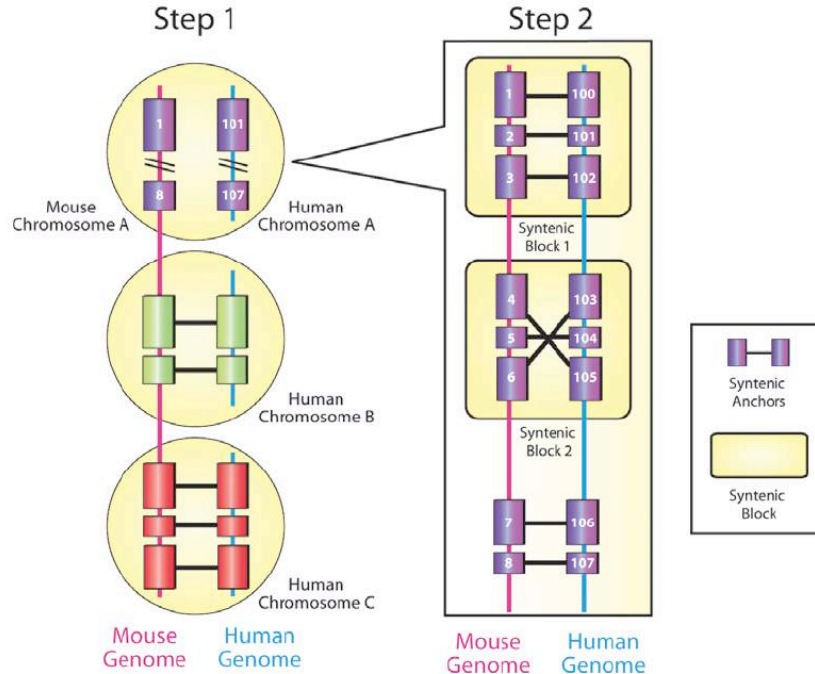
The complete genomes of fly, worm, and chromosomes 2 and 4 of mustard weed (as deposited at ncbi.nlm.nih.gov/genbank/genomes) were screened against the repeats in RepBase Update 5.02 (September 2000) with RepeatMasker at sensitive settings.



(Lander et al. (2001) *Nature*)

The mouse genome

- the 2.5 Gb mouse genome is 14% smaller than the human genome, thanks to the higher deletion rate found in mice
- they encode about the same number of genes (~20,000 protein coding), and 80% of these are homologous
- 90% of the two genomes is syntenic

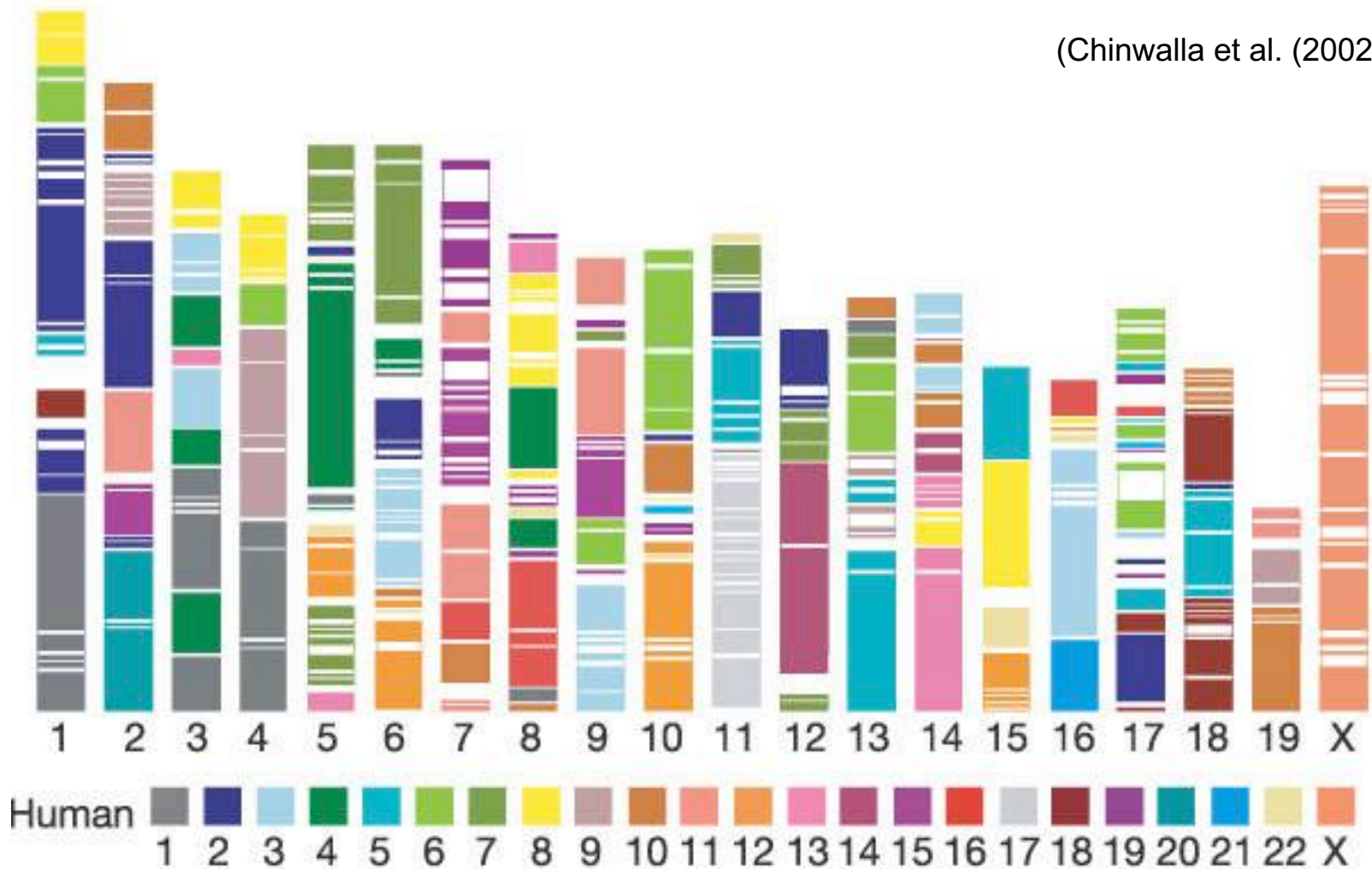


(Chinwalla et al. (2002) *Nature*)

Synthetic blocks in the mouse and human genomes



(Chinwalla et al. (2002) *Nature*)

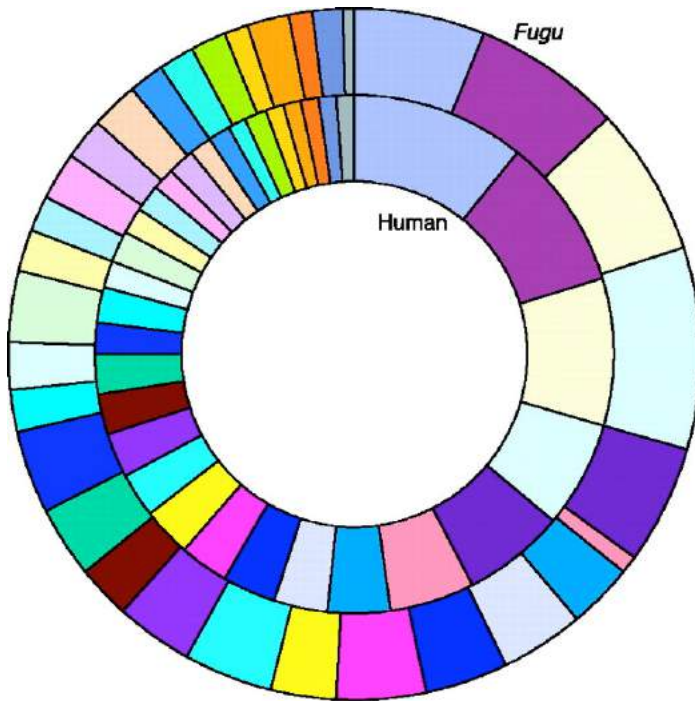


- Every band corresponds to a >300 kb synthetic block

The pufferfish (*Takifugu rubripres*)



- a 393.3 Mb genome encodes ~19,000 genes
- this is pretty much the same we can find in the human genome, however that is almost 9 times larger (2.9 Gb)

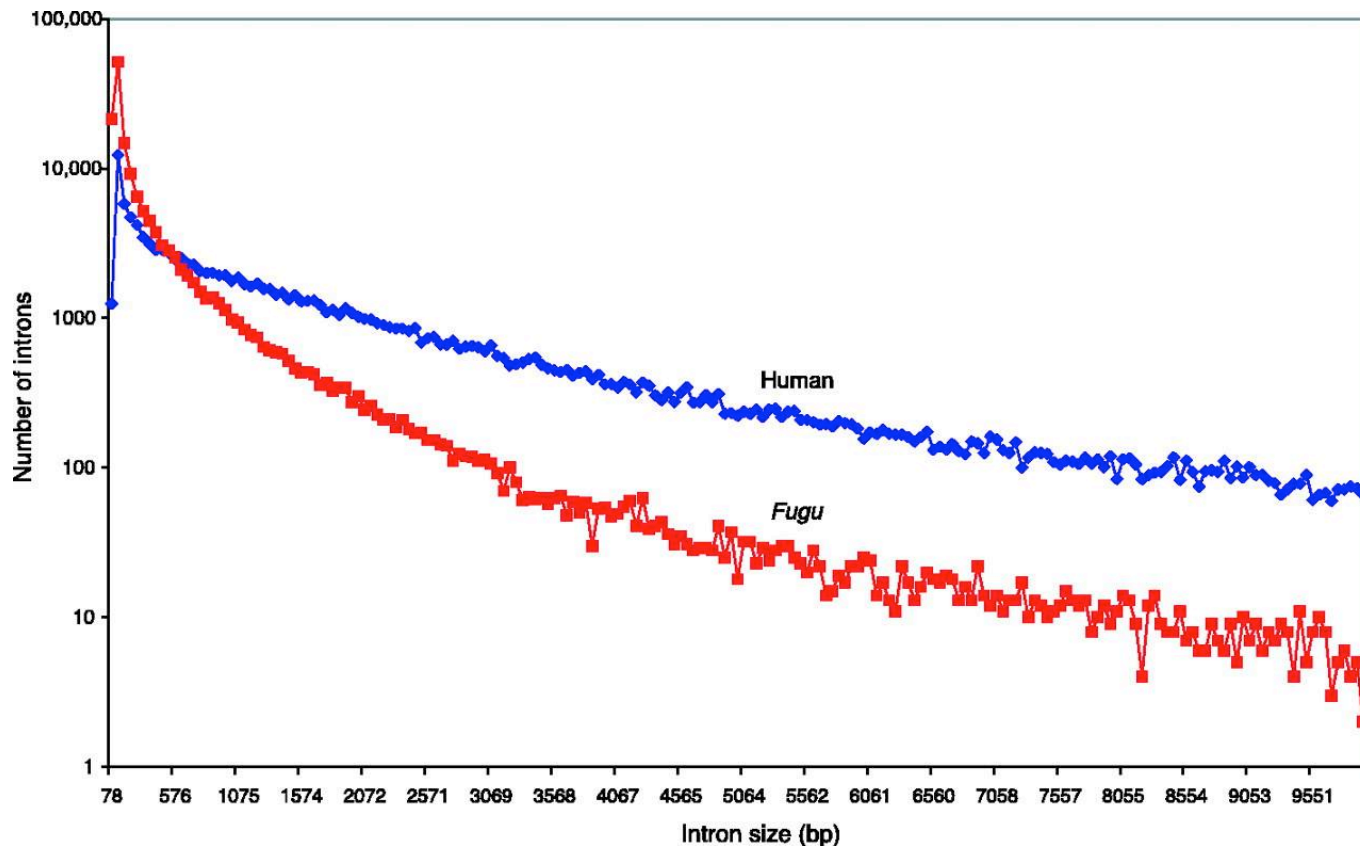


- most of the protein families are the same size, except some K⁺-channel components and Zn-finger TFs (the former are more abundant in the *Fugu* genome, whereas the latter in humans).

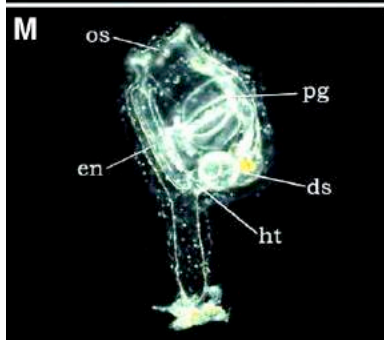
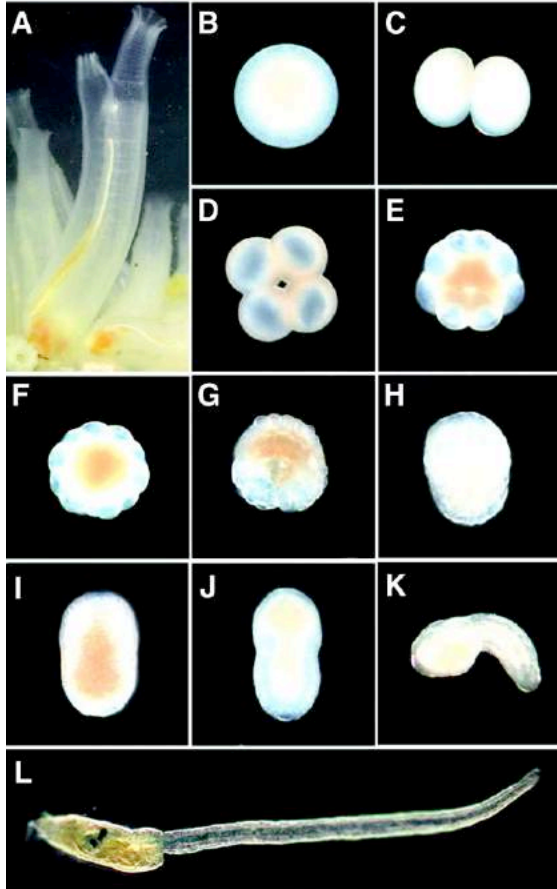
A compact vertebrate genome: less repetitive elements, shorter introns



- Only 2.7% of the genome is repetitive, which is *significantly* shorter than in mammals (35-45%)



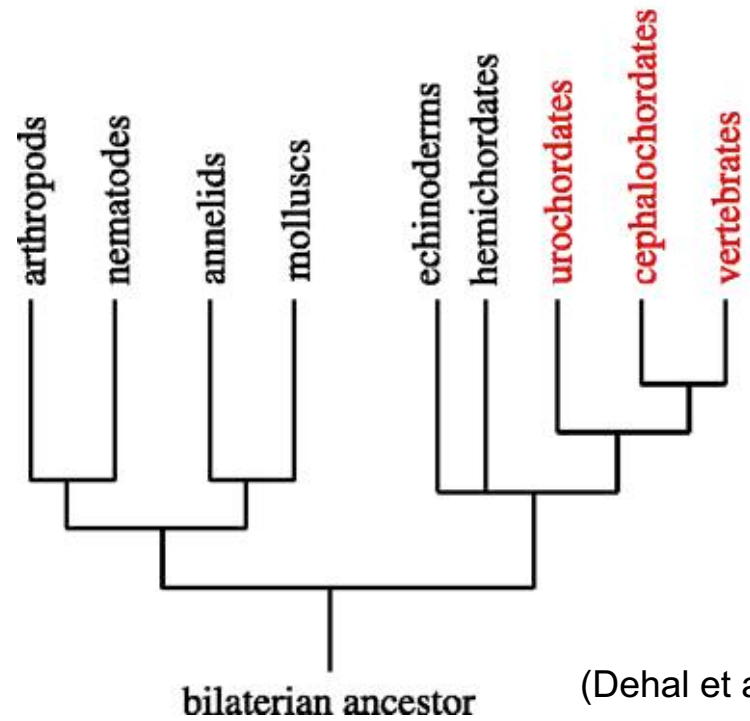
Tunicates (*Ciona intestinalis*)



Protostomes

Deuterostomes

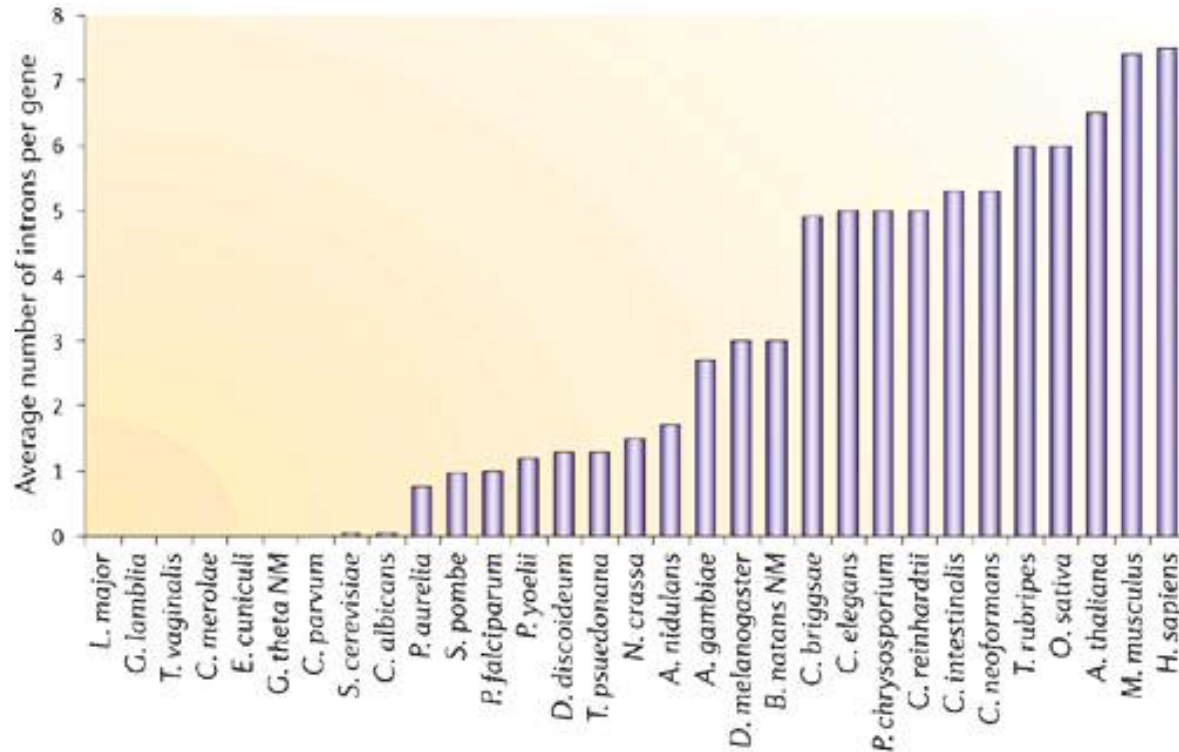
Chordates



(Dehal et al. (2002) *Science*)

- very compact: ~115 Mb, ~17 000 genes
- 60% of the genes have Protostome homologues, 20% however are unique (most likely reflecting the unique life history)
- Ciona* genes have in average 6.8 exons (vs. 5 in *Drosophila* and 8.8 in humans) => intron poor genome

Early or late introns?

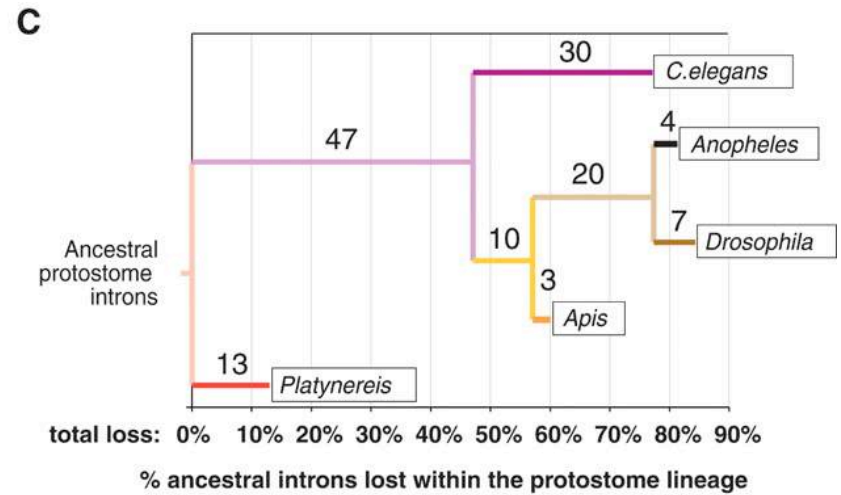
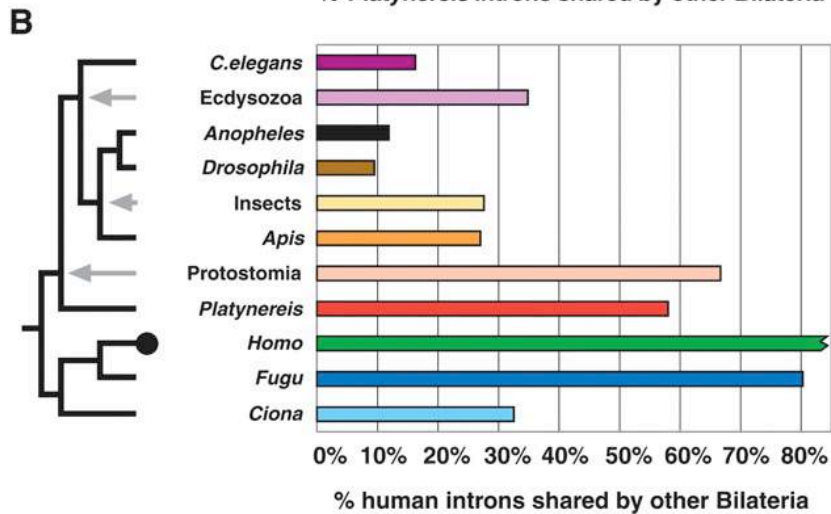
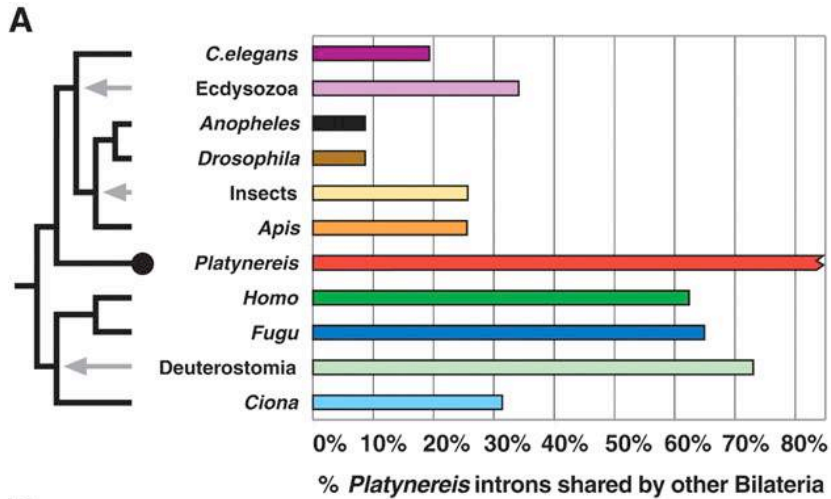


(Roy and Gilbert (2006) *Nat Rev Gen*)

Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

- Based on the first complete genome sequences it seems logical that Eumetazoan and Urbilaterian ancestors had only few introns and these became abundant only later in Deuterostomes (especially in mammals)

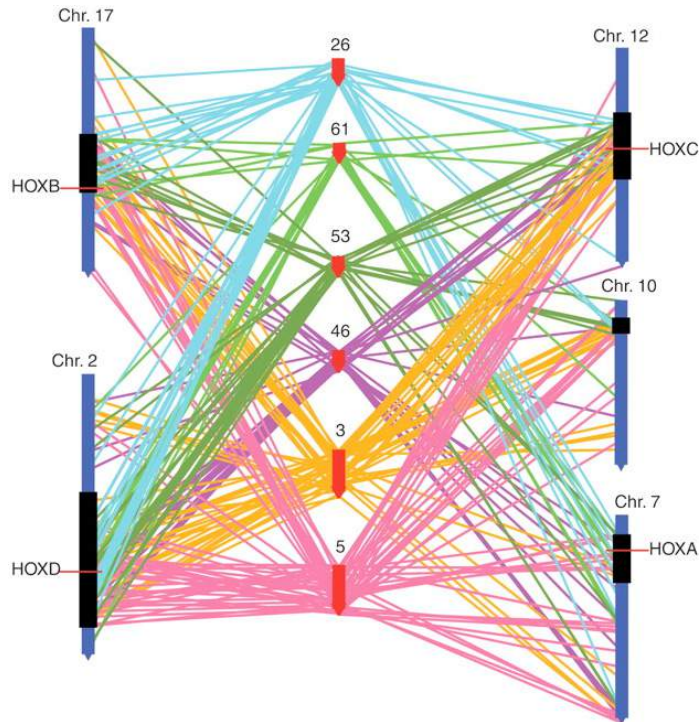
A peek into the *Platynereis* genome: the ancestral origin of histones



(Raible et al. (2005) *Science*)

The sea anemone (*Nematostella vectensis*)

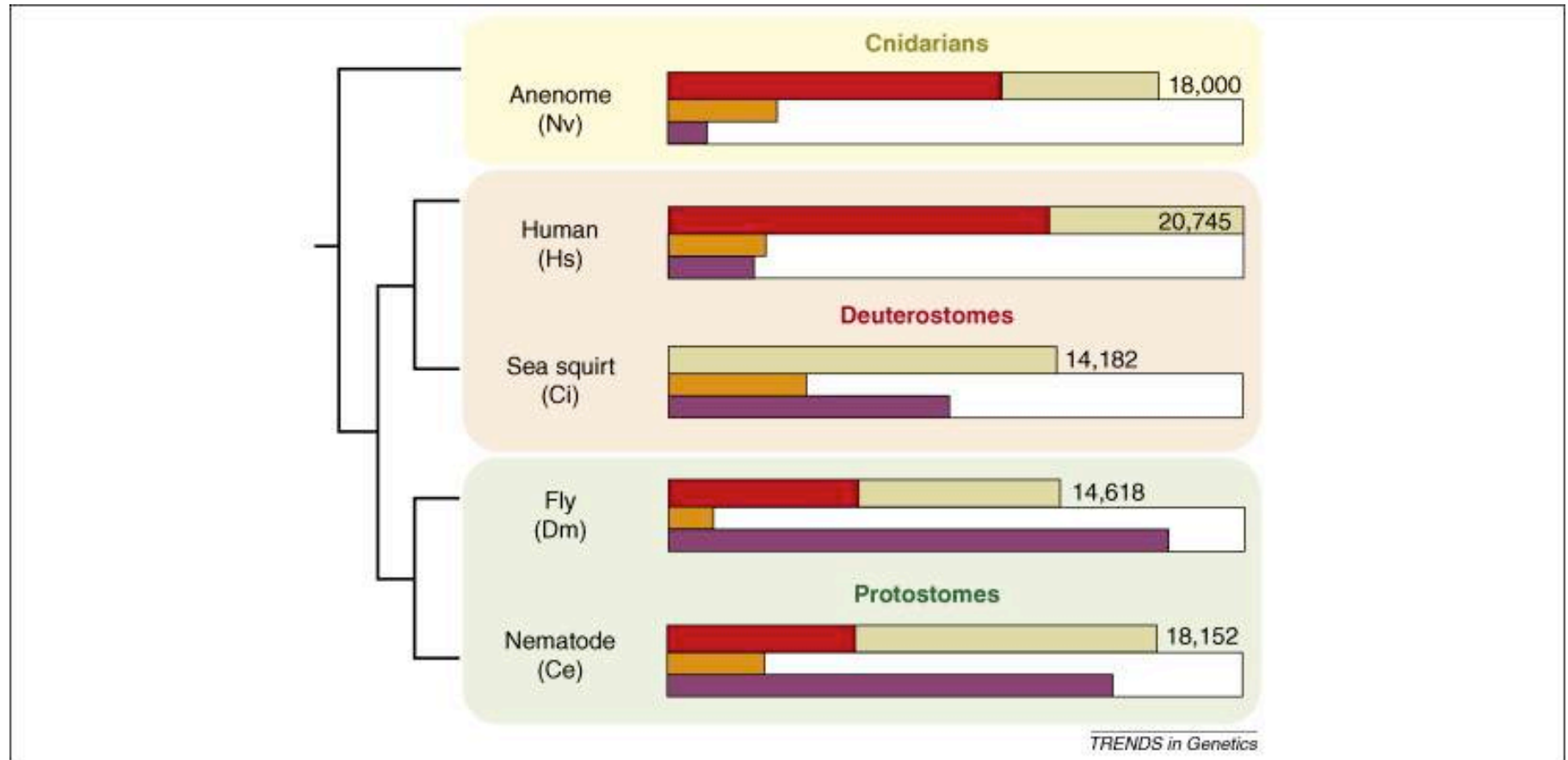
- 450Mb genome, with 18 000 protein coding genes (!!)
- a quarter of the genome is repetitive
- twice as many genes are common between a human and a sea anemone than flies/worms and humans!!!



- Despite a ~700 Mya split, we can *still* find syntenic blocks between the human and *Nematostella* genomes

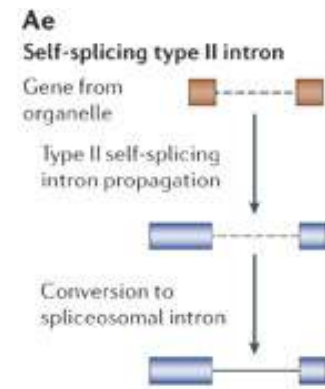
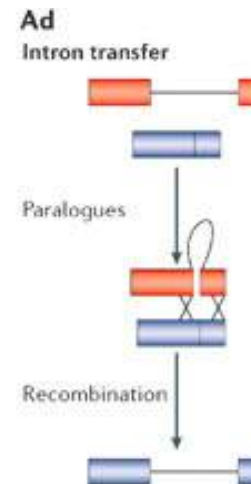
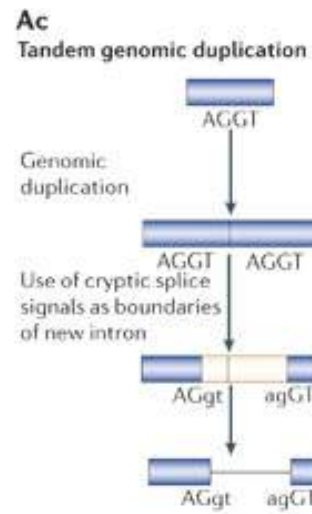
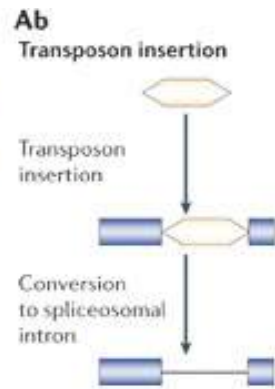
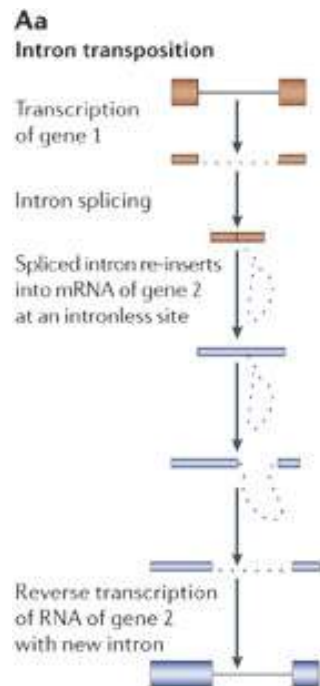
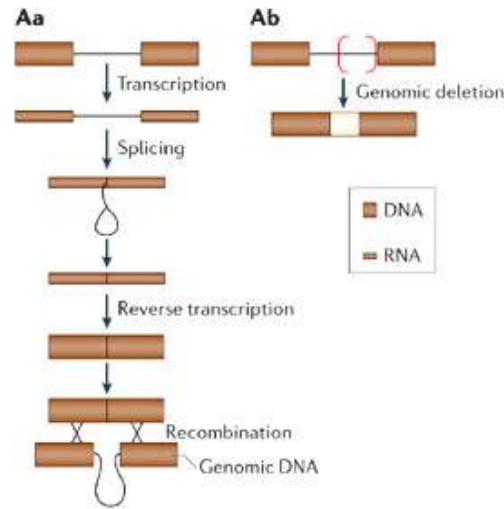
(Putnam et al (2007) *Science*)

The majority of human introns was already present in Cnidarians and Urbilateria

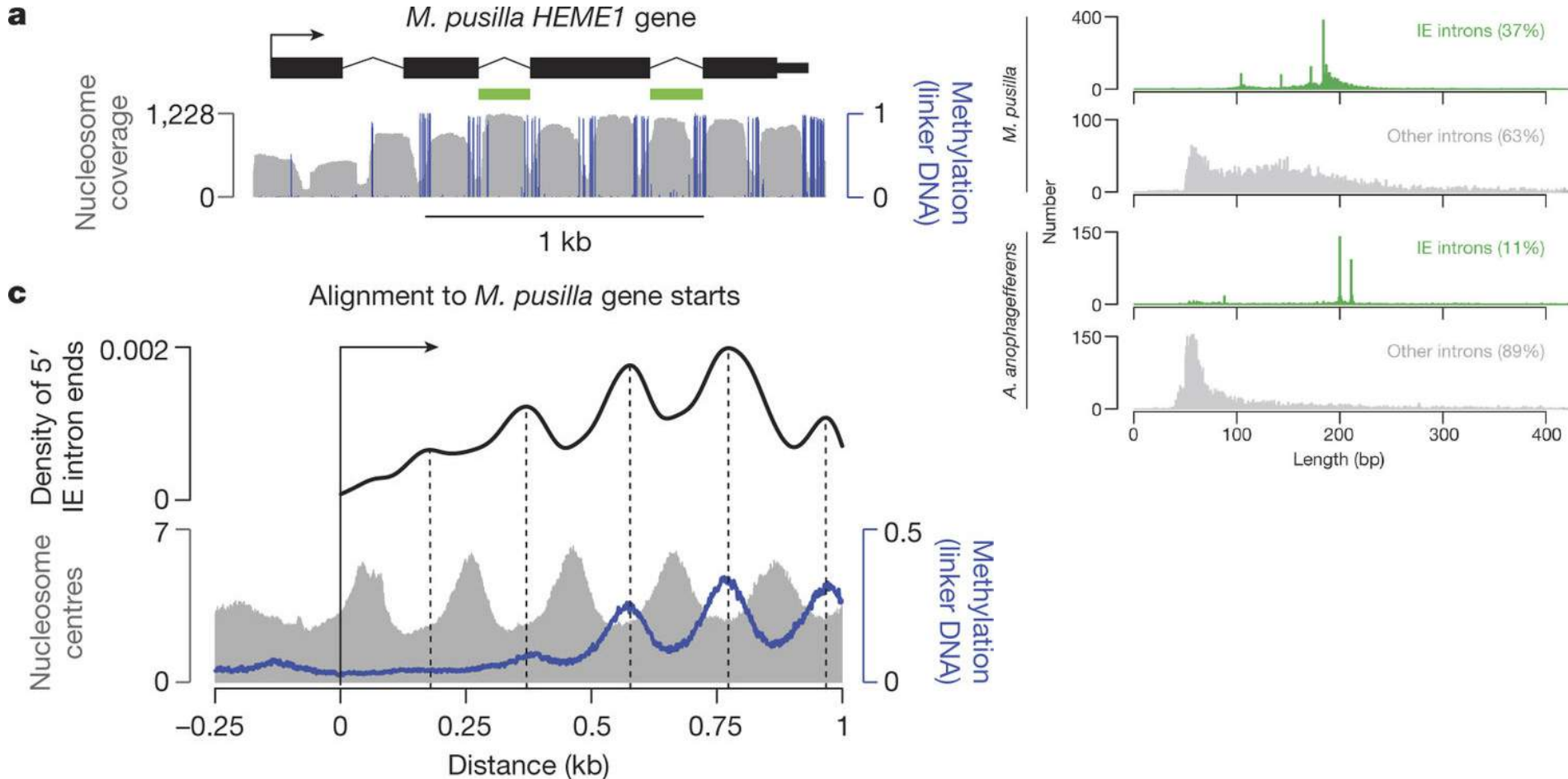


- **Upper bar:** no. of genes, the **red** portion is from the common ancestor
- **Orange bar:** the proportion of new introns vs. the total number of introns
- **Purple bar:** what proportion of ancestral introns were lost

Mechanism of intron loss and intron gain

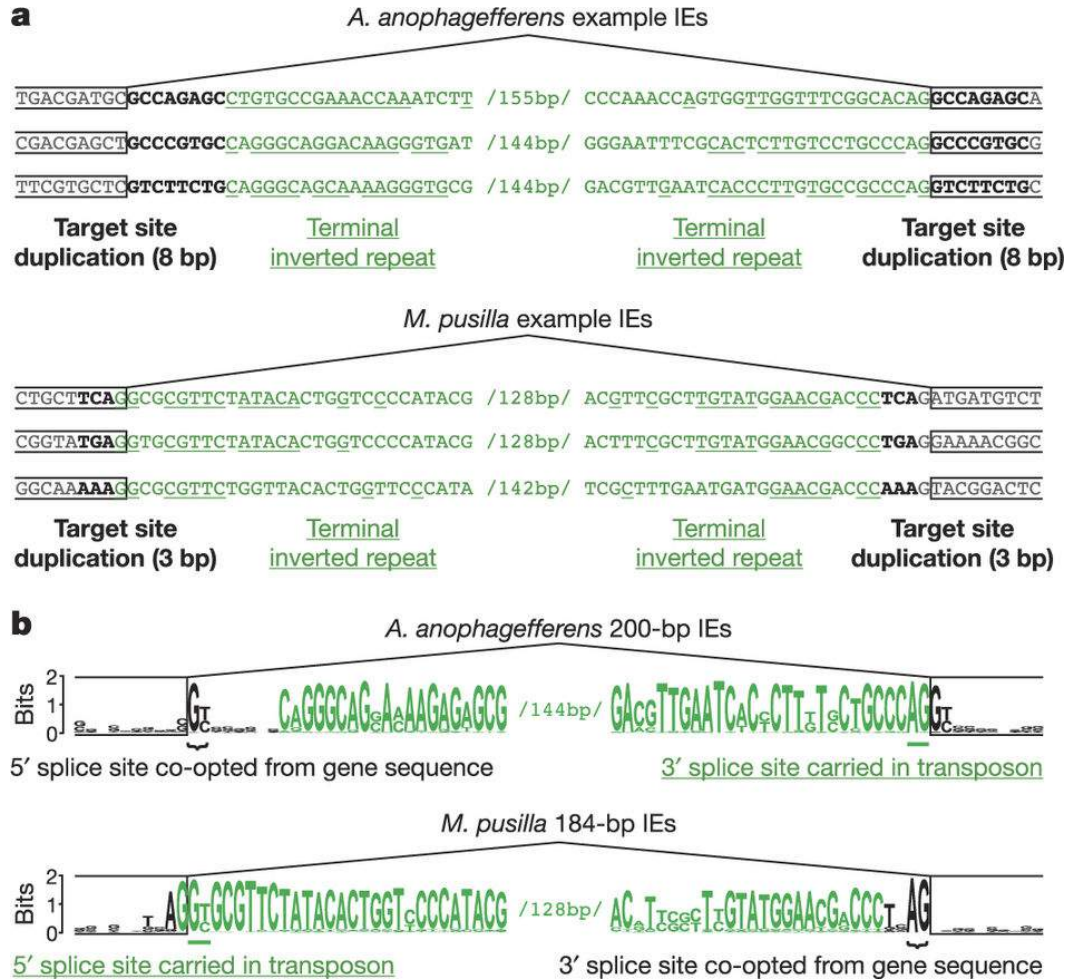


TE-derived introns in algae



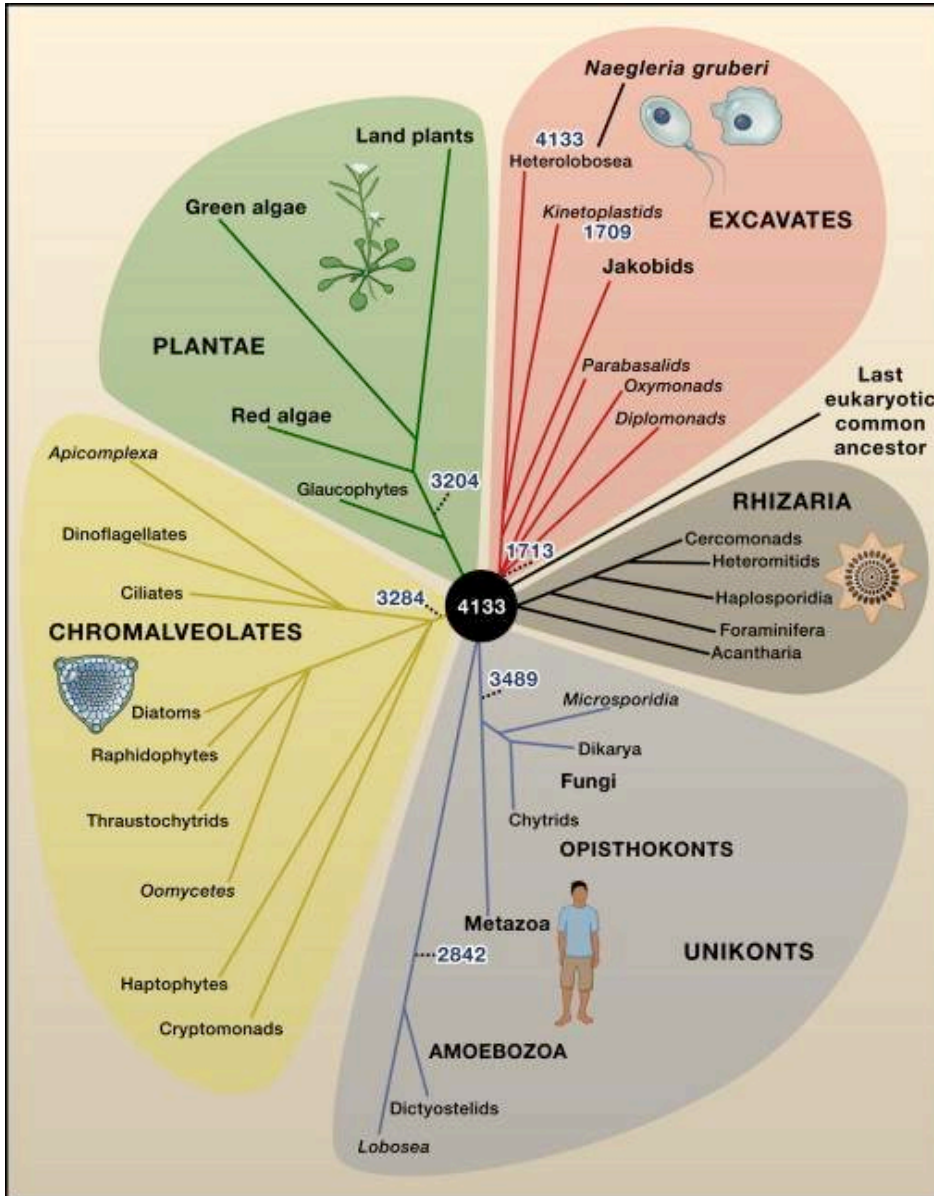
- The location of introner elements (IE) suggests that a transposon jumped into the linker region between nucleosomes (linker regions are methylated in this species)

TE-derived introns in algae

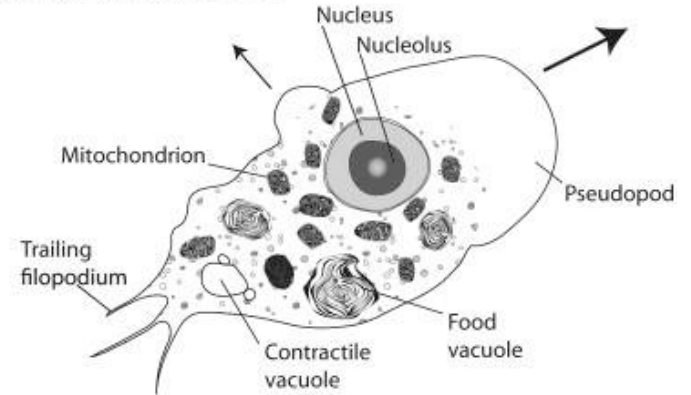


- Splice sites can evolve from the terminal region of the IE, or from target-site duplications

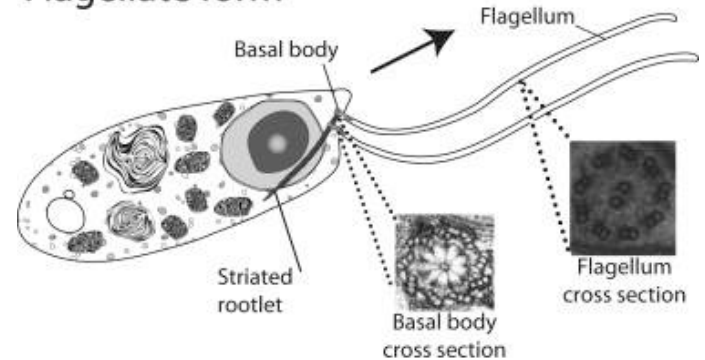
The ancestral eukaryotic cell, through the lens of an exotic protist (*Naegleria gruberi*)



Amoeboid form



Flagellate form



(Fritz-Laylin et al. (2010) *Cell*)

The ancestral eukaryotic cell, through the lens of an exotic protist (*Naegleria gruberi*)

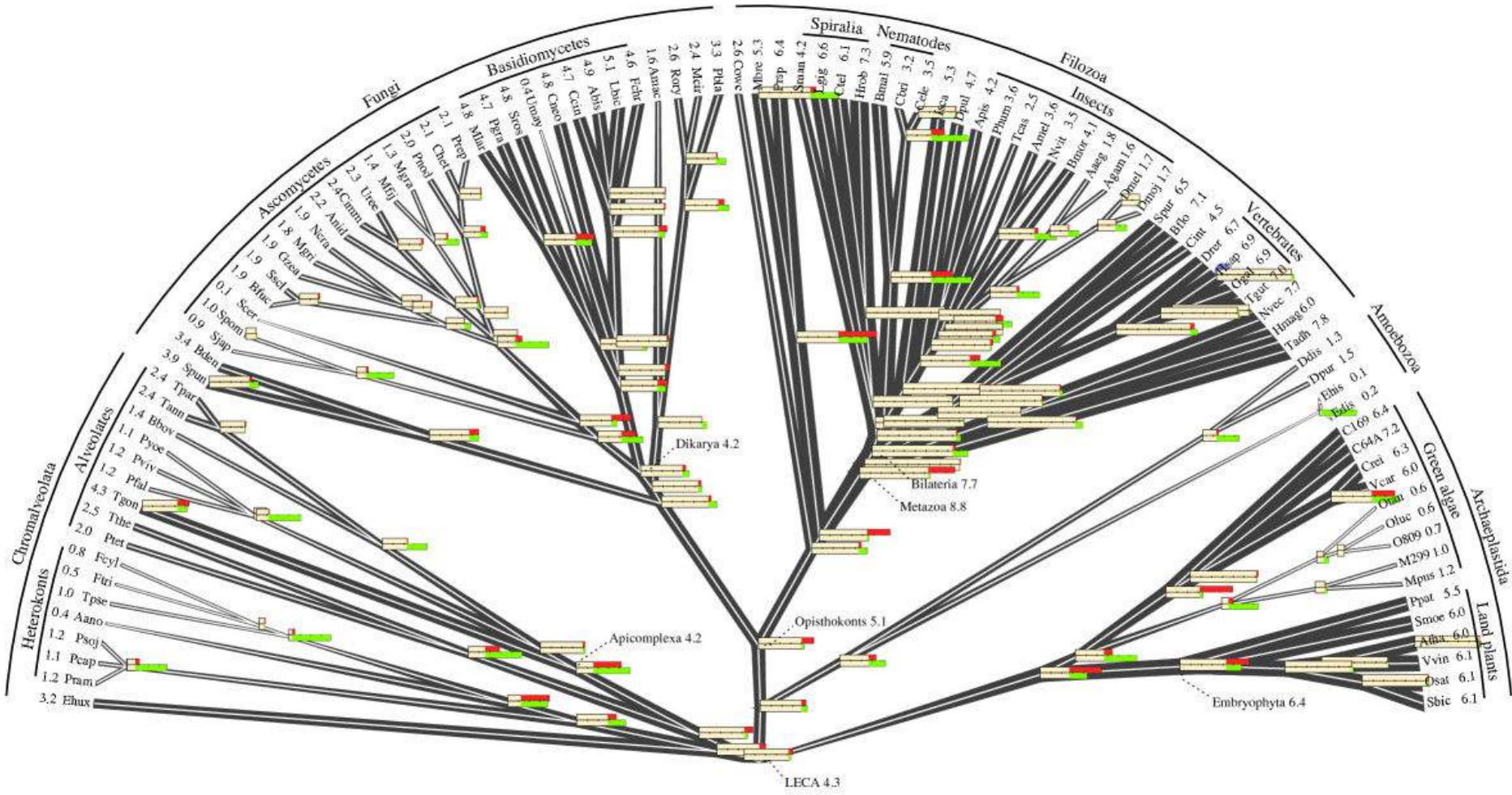


- *Naegleria* is capable of both aerobic and anaerobic metabolism
- can switch between flagellar and amoeboid forms as well

Species	Genome Size (Mbp)	No. Chromosomes	%GC	Protein-Coding Loci	% Coding	% Genes w/ Introns	Introns per Gene	Median Intron Length (bp)
<i>Naegleria</i>	41	> = 12	33	15, 727	57.8	36	0.7	60
Human	2851	23	41	23, 328	1.2	83	7.8	20, 383
<i>Neurospora</i>	40	7	54	10, 107	36.4	80	1.7	72
<i>Dictyostelium</i>	34	6	22	13, 574	62.2	68	1.3	236
<i>Arabidopsis</i>	140.1	5	36	26, 541	23.7	80	4.4	55
<i>Chlamydomonas</i>	121	17	64	14, 516	16.3	91	7.4	174
<i>T. brucei</i>	26.1	>100	46	9152	52.6	~0 (1 total)	ND	ND
<i>Giardia</i>	11.7	5	49	6480	71.4	~0 (4 total)	ND	ND

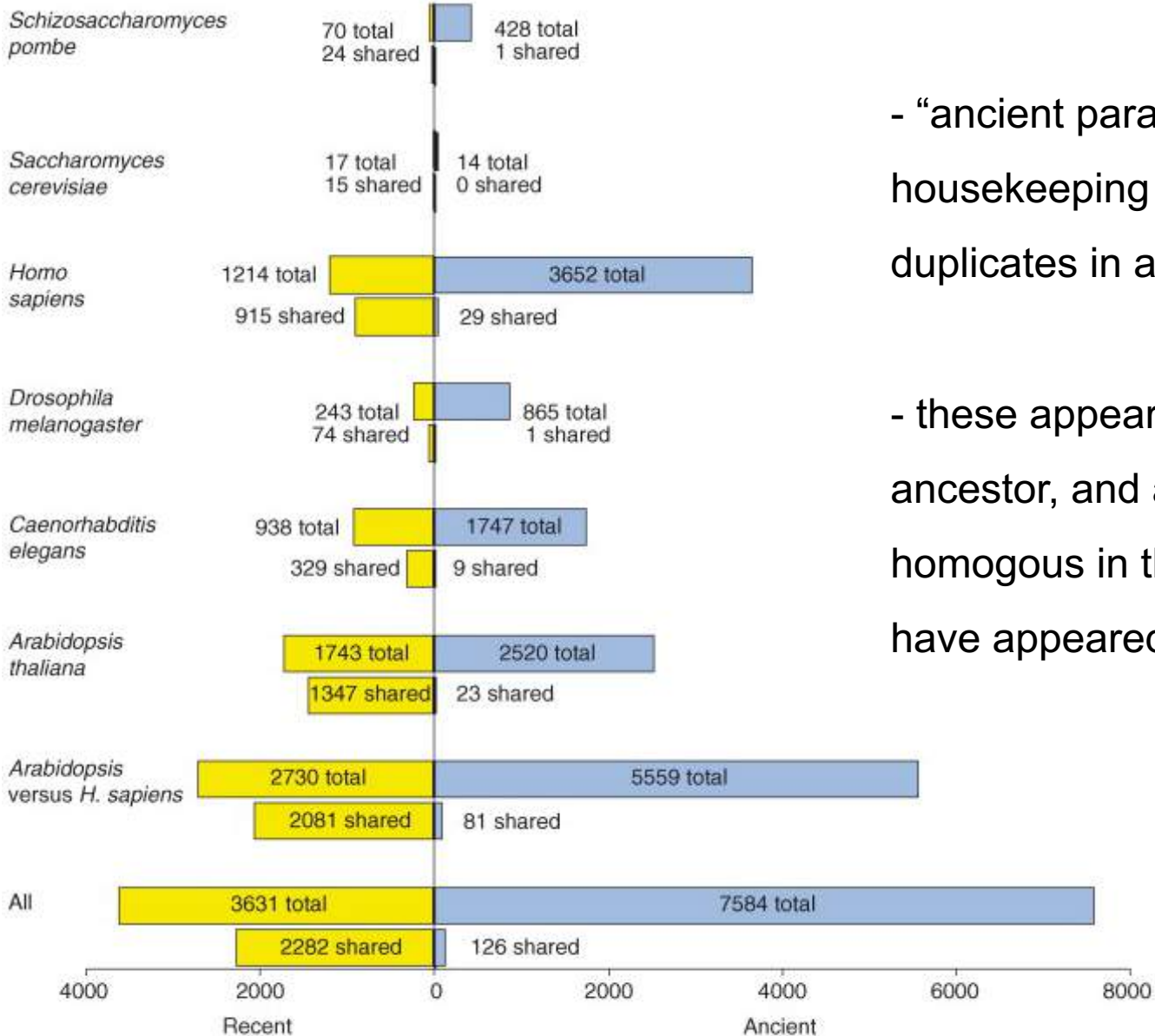
- many of the *Naegleria* introns are in orthologous position, compared to other eukaryotic introns!!
- Only 5.1% of the genome is repetitive sequence

Intron-evolution in eukaryotes



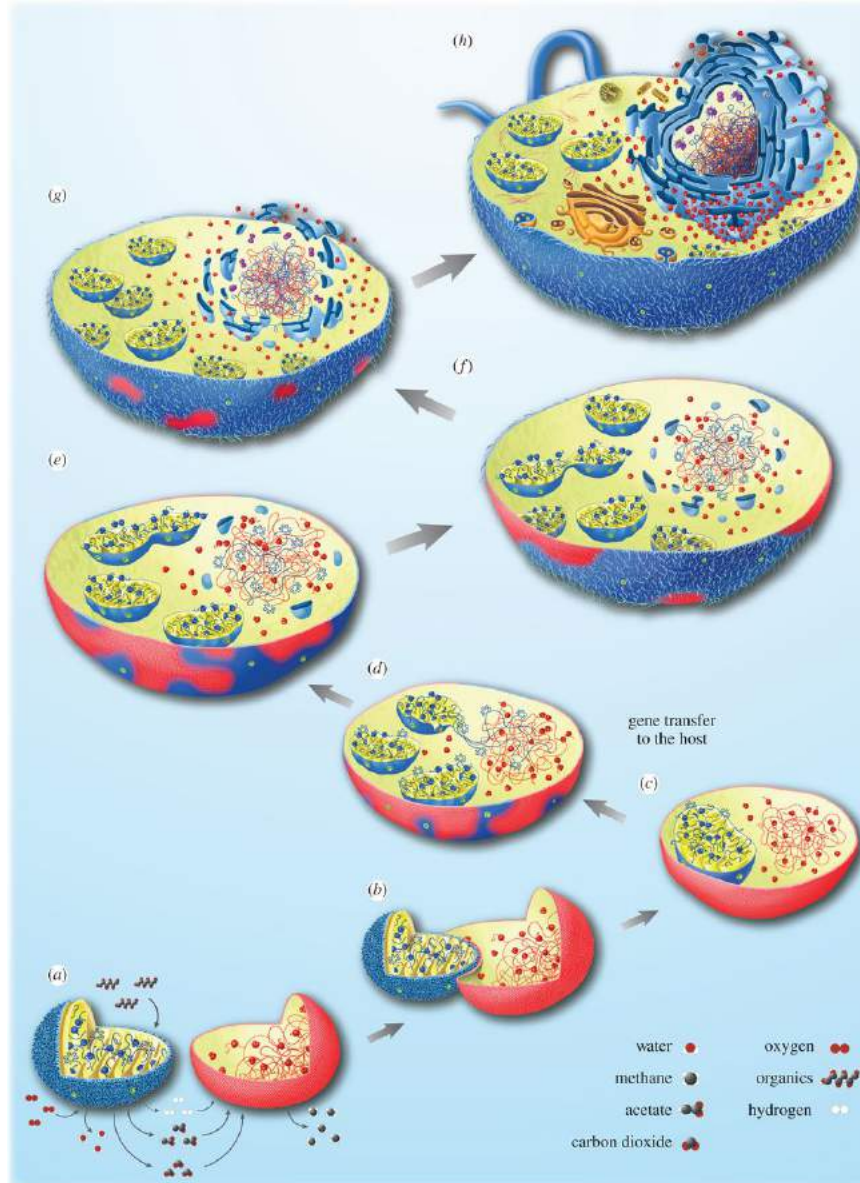
- the width of the lines is proportional to the amount of introns

In general introns are conserved, but not is the most ancient eukaryotic genes

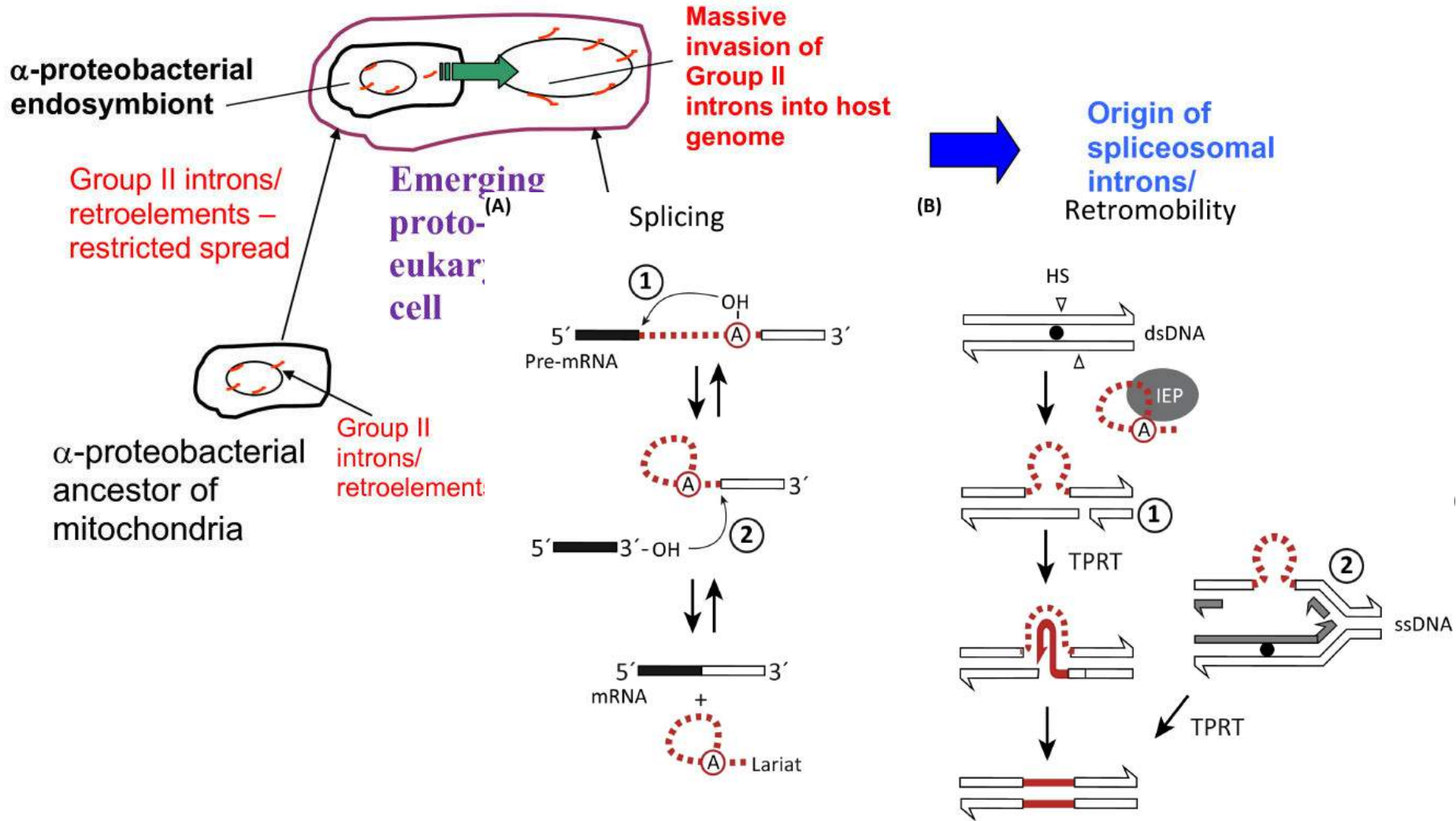


- “ancient paralogs” – mostly housekeeping genes, present in duplicates in all eukaryotes
- these appeared in the common ancestor, and as introns are not homogenous in the paralogs, they must have appeared about the same time

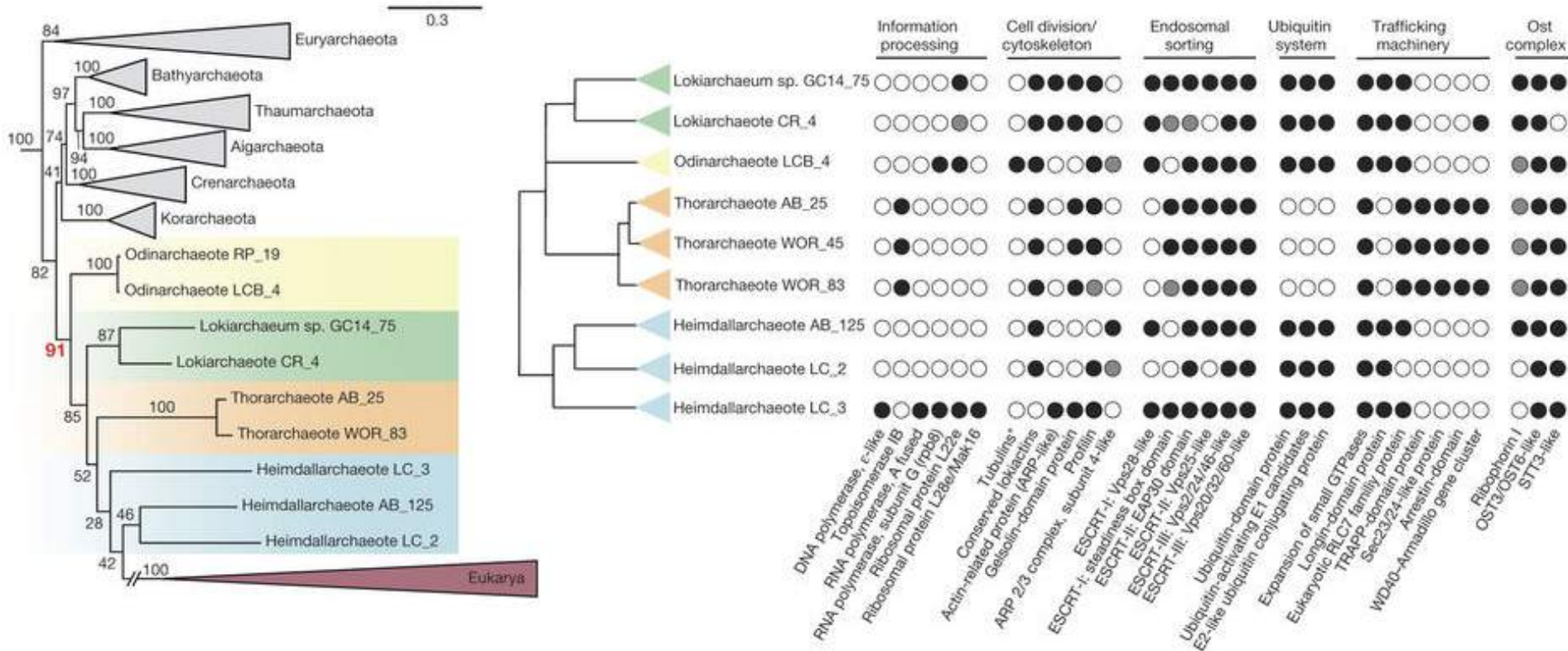
The endosymbiont-origin of eukaryotic cells



The endosymbiont-origin of eukaryotic cells



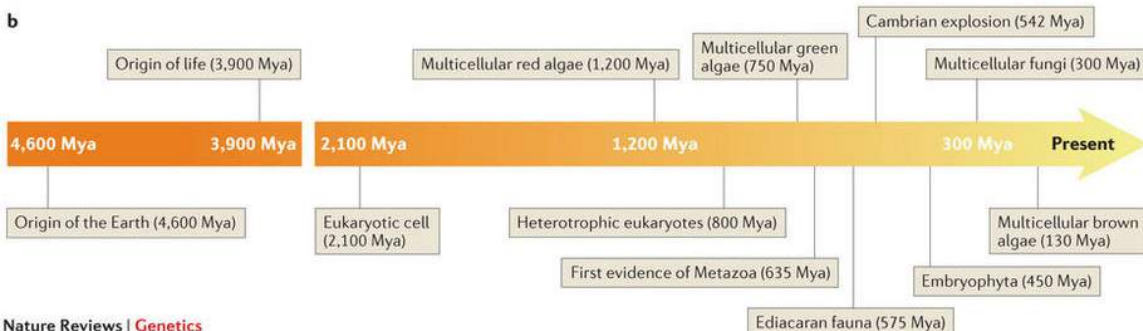
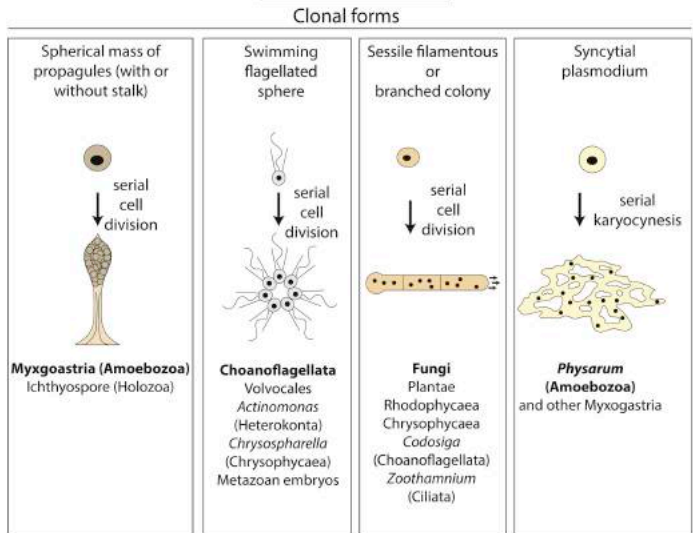
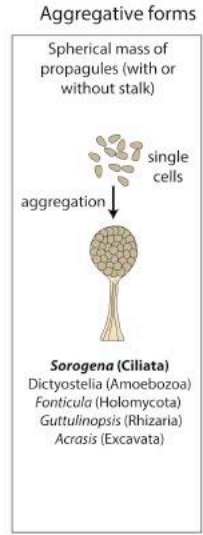
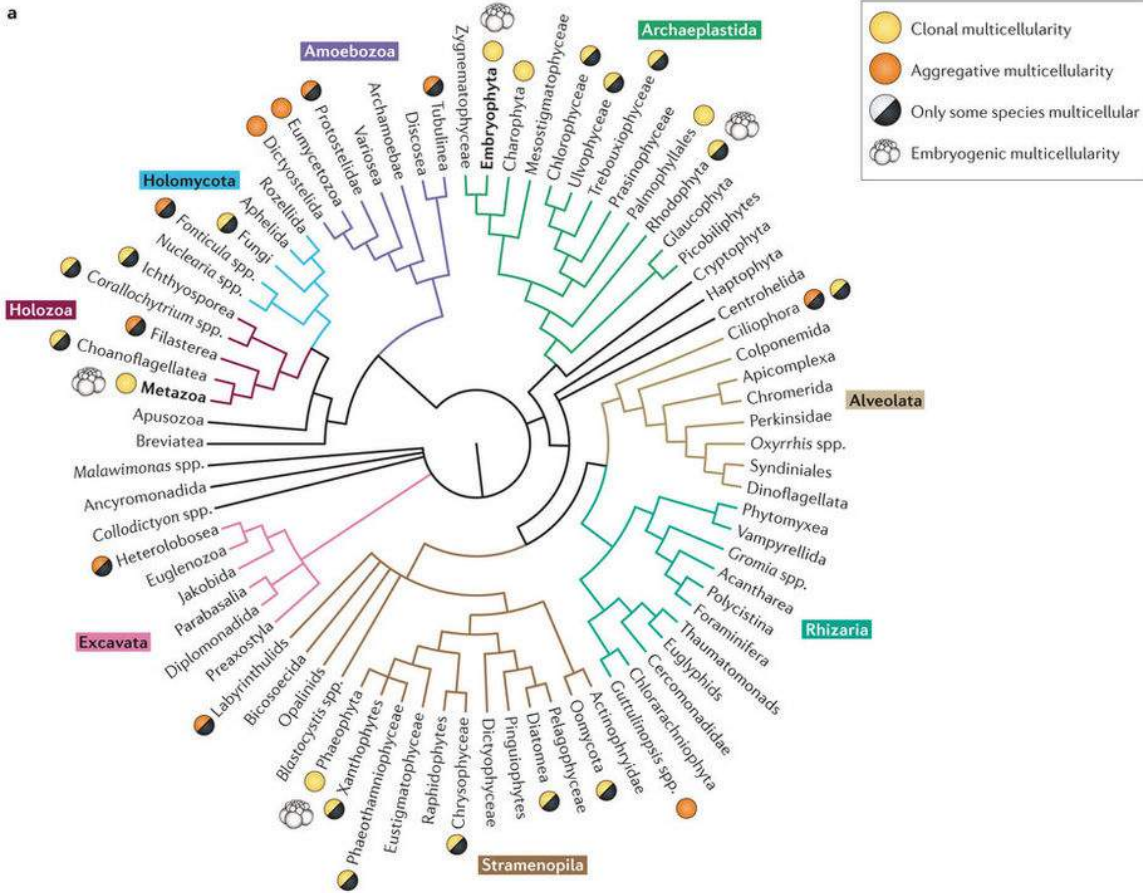
The majority of essential eukaryotic genes can be found in Archea



- in the Asgard archeal clade a sophisticated cytoskeleton is present (actin and tubulin, with members of the ARP complex)

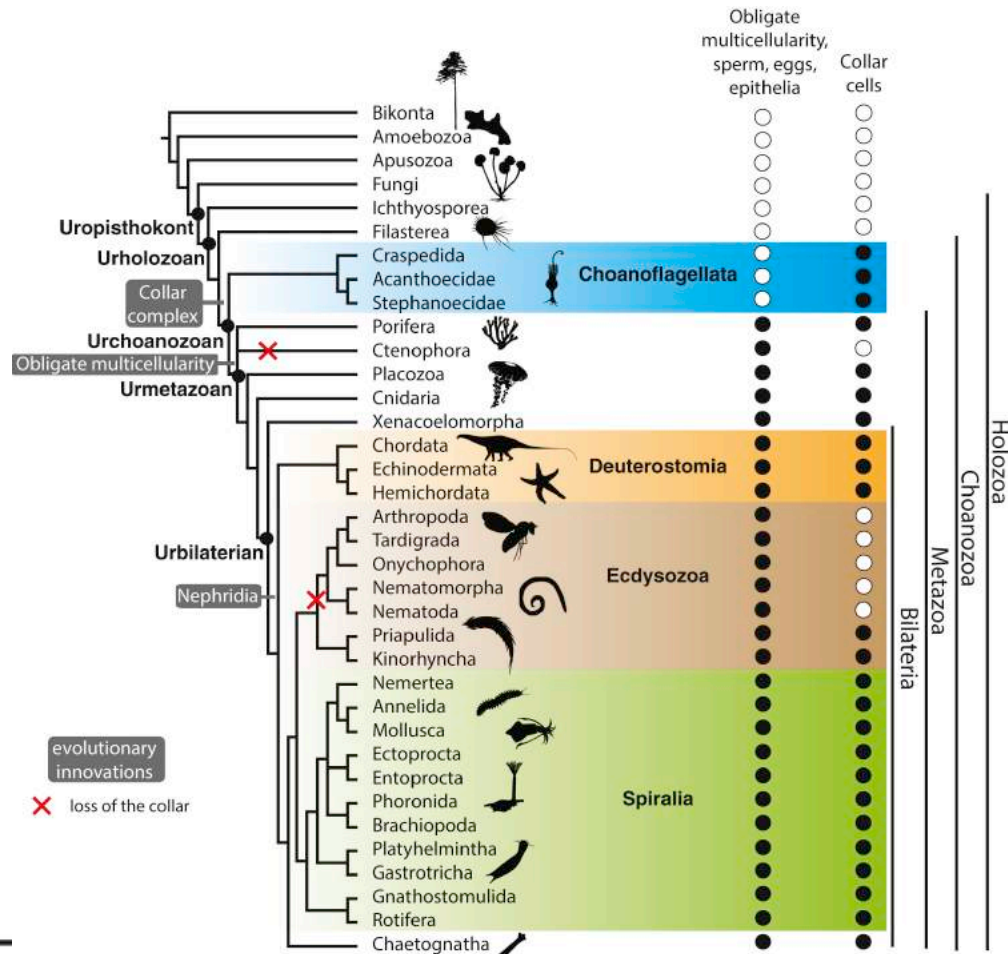
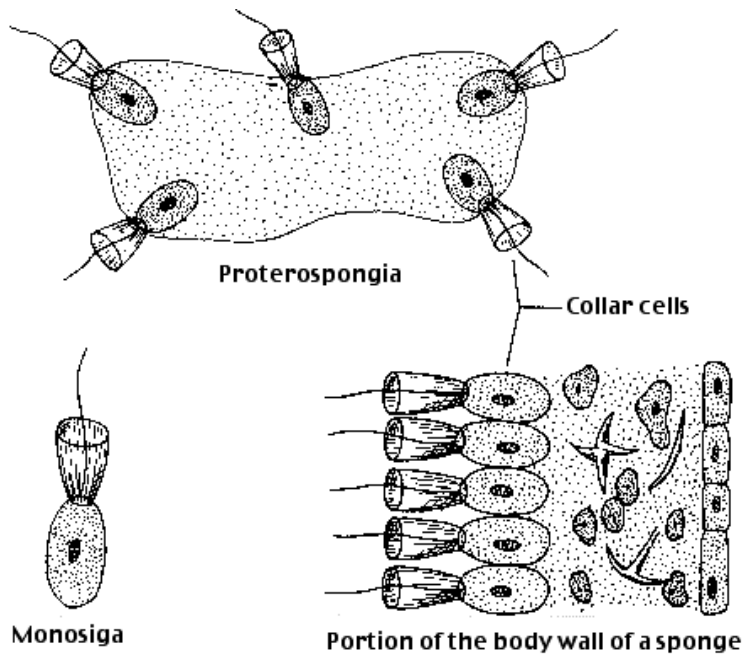
- complex membrane dynamics (ESCRT complex)

The multiple origin of multicellularity



(Sebé-Pedros et al. 2017 *Nat Rev Gen*, Brunet and King 2017 *Dev Cell*)

The origins of multicellularity through the genome of choanoflagellate (*Monosiga brevicollis*)



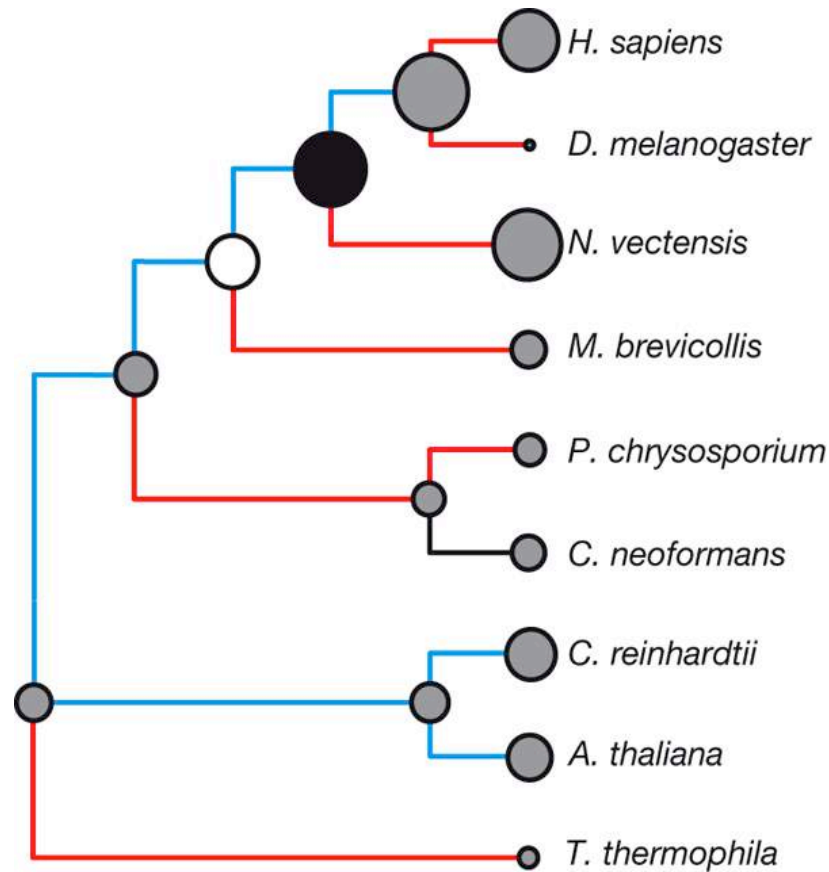
(King et al. 2008 *Nature*, Brunet and King 2017 *Dev Cell*)

Table 1 | *M. brevicollis* genome properties in a phylogenetic context

	Metazoa								
	<i>Hsap</i>	<i>Cint</i>	<i>Dmel</i>	<i>Nvec</i>	<i>Mbre</i>	<i>Ccin</i>	<i>Ncra</i>	<i>Ddis</i>	<i>Atha</i>
Genome size (Mb)	2,900	160	180	357	42	38	39	34	125
Total number of genes	23,224	14,182	14,601	18,000	9,196	13,544	9,826	13,607	27,273
Mean gene size (bp)	27,000	4,585	5,247	6,264	3,004	1,679	1,528	1,756	2,287
Mean intron density (introns per gene)	7.7	6.8	4.9	5.8	6.6	4.4	1.8	1.9	4.4
Mean intron length (bp)	3,365	477	1,192	903	174	75	136	146	164
Gene density (kb per gene)	127.9	11.9	13.2	19.8	4.5	2.7	4.0	2.5	4.5

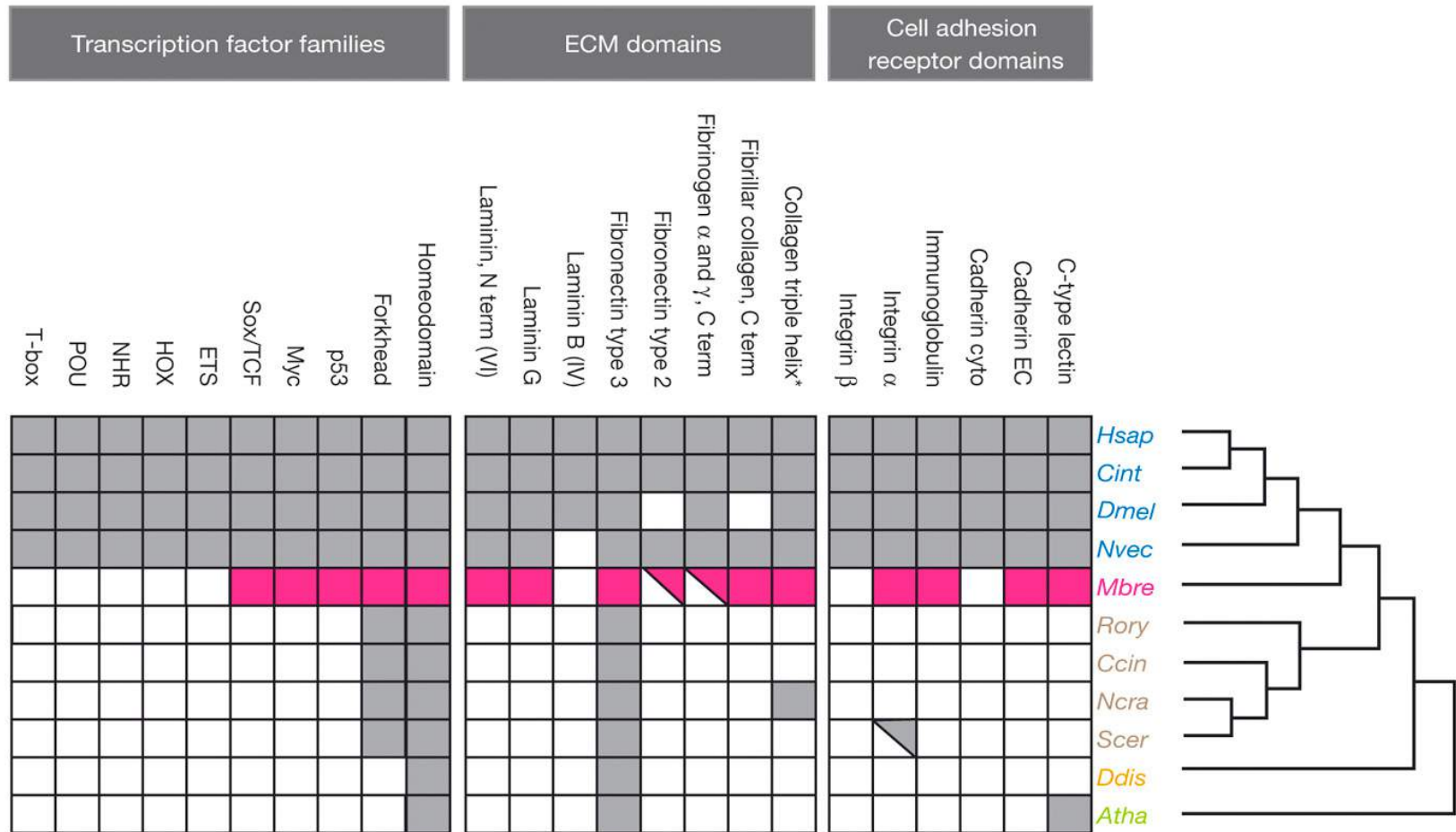
Species names follow the four-letter convention from Fig. 1.

Intron-evolution



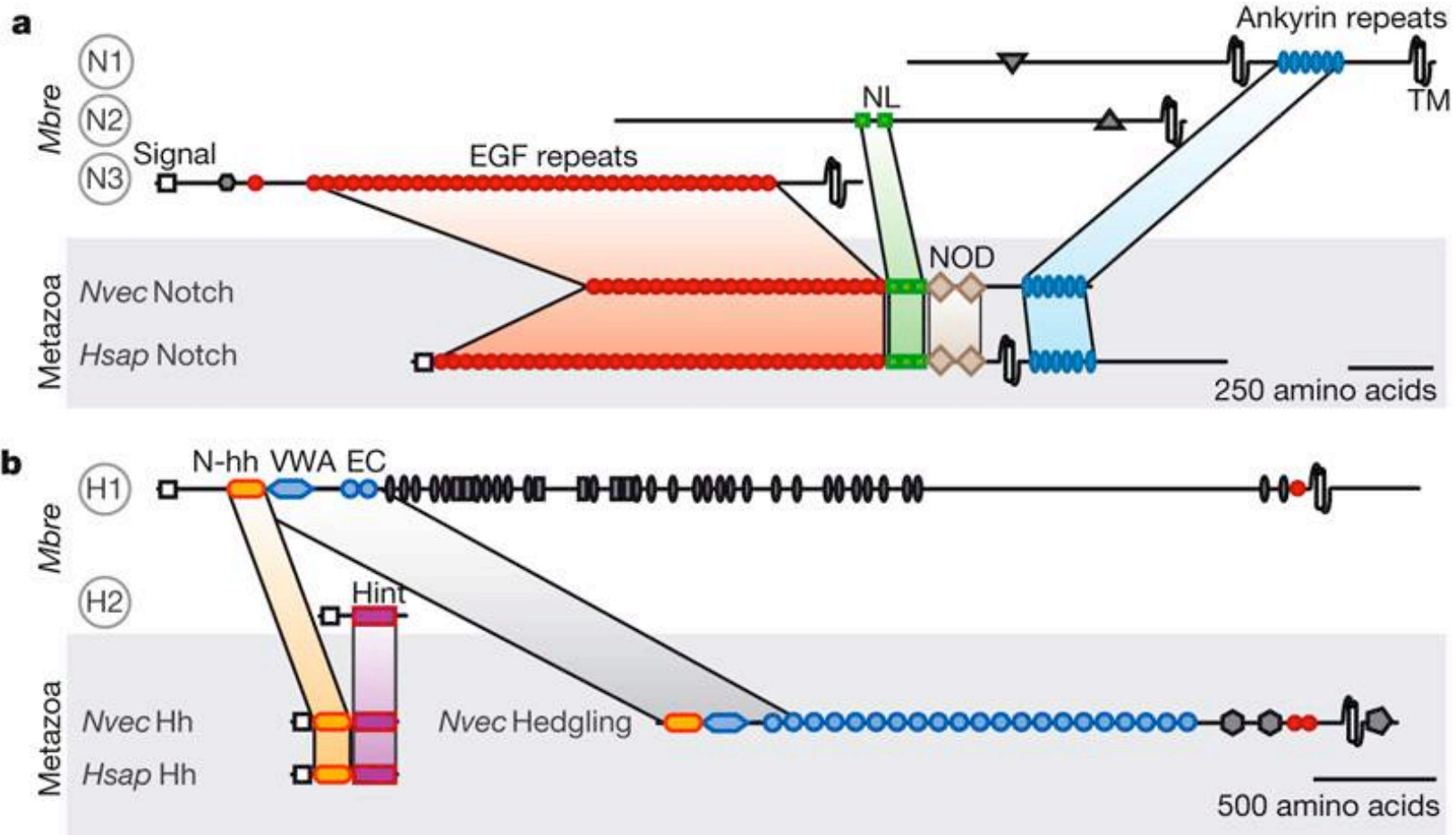
- the diameter of the circles is proportional with the number of introns
- **blue** denotes lineages with overall intron-gain, **red** those with overall intron-loss

Genes important for multicellularity/ cell-cell adhesion were present in the common ancestor



- Some of the cell adhesion and ECM domains are present in unique combinations in the *Monosiga* genome.

Origin of the Eumetazoan signaling pathways: domain-shuffling

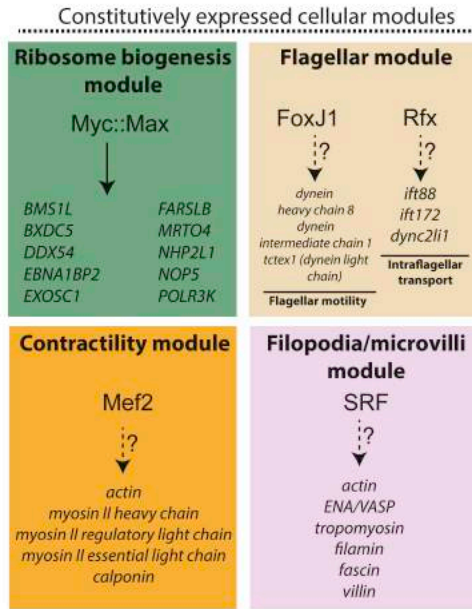
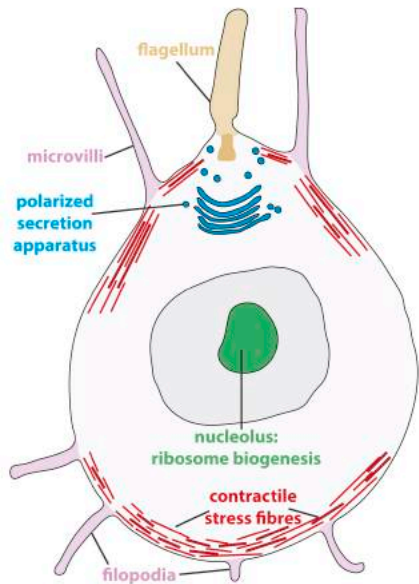


- WNT, TGF β , JAK/STAT pathways have no trace in *Monosiga*

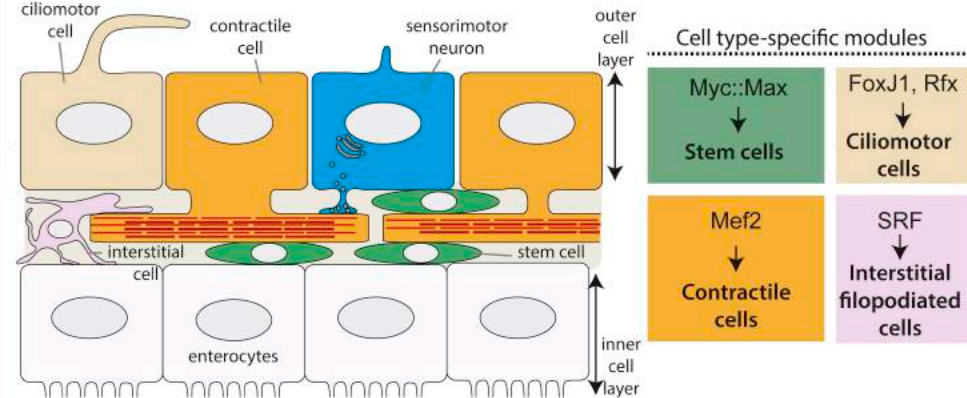
The division of labor hypothesis of multicellularity



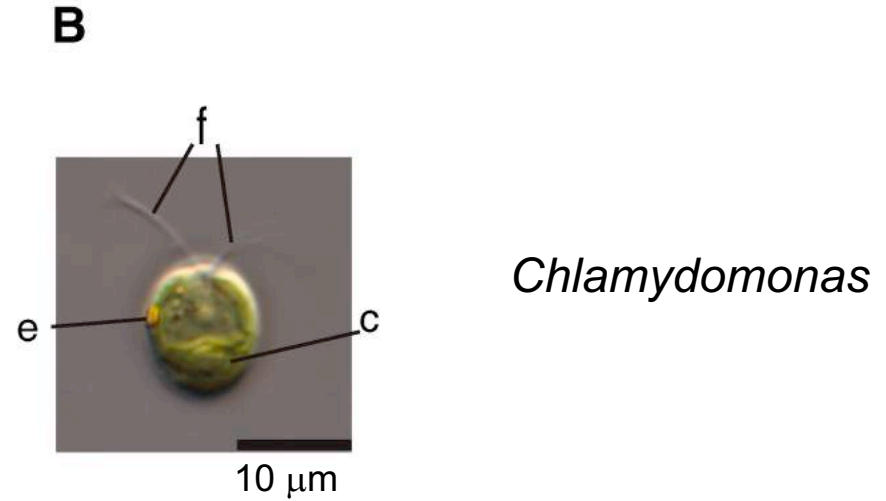
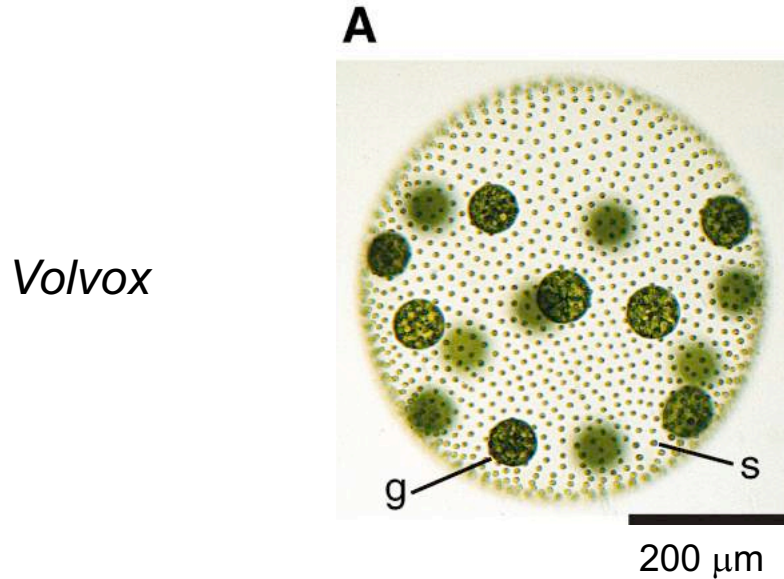
A Choanoflagellate: no division of labor



B Cnidarian-bilaterian ancestor: division of labor



The origins of multicellularity 2. – the *Volvox* genome



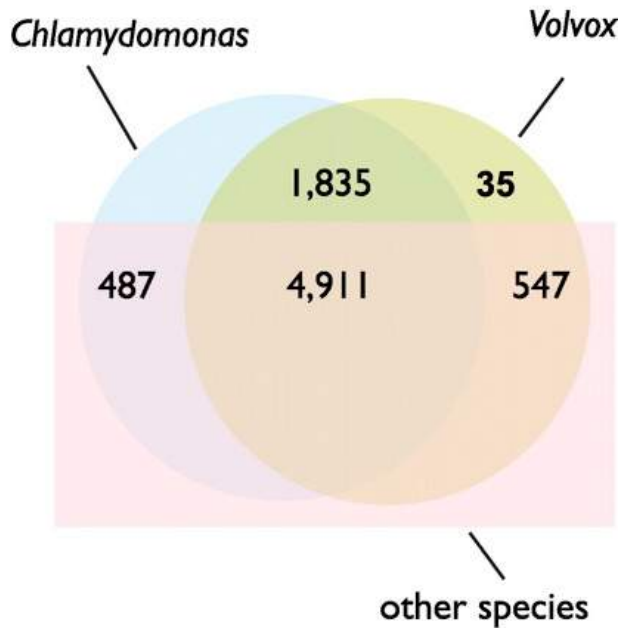
Species	Genome size (Mbp)	Number of chromosomes	% G and C	Protein-coding loci	% coding	% of genes with introns	Introns per gene	Median intron length (bp)
<i>V. carteri</i>	138	14*	56	14,520	18.0	92	7.05	358
<i>C. reinhardtii</i>	118	17	64	14,516	16.3	91	7.4	174

- the 17% increase in genome-size is due to the increase in the numbers of TE

The origins of multicellularity 2. – the *Volvox* genome

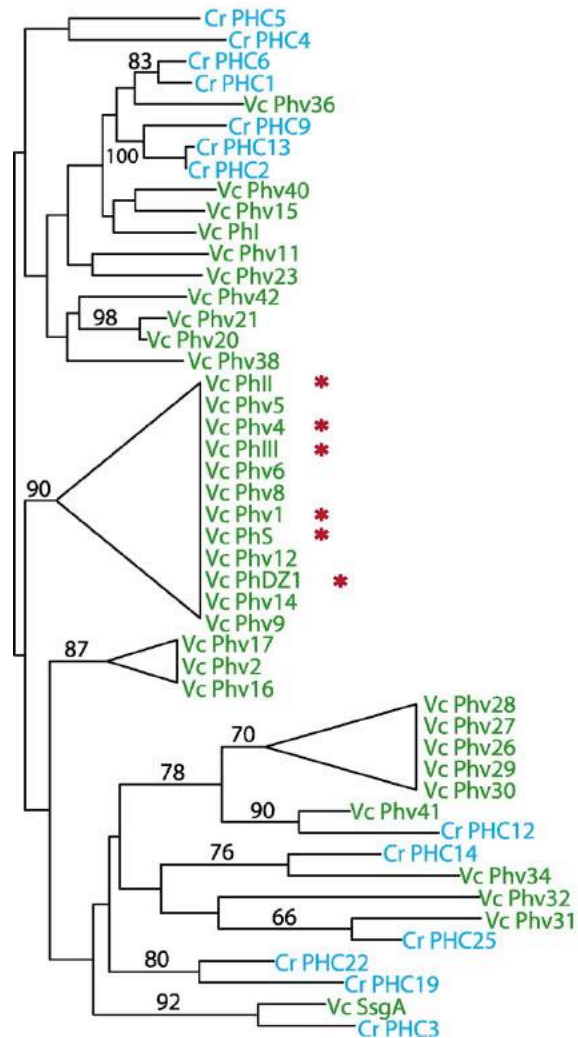


- Protein families

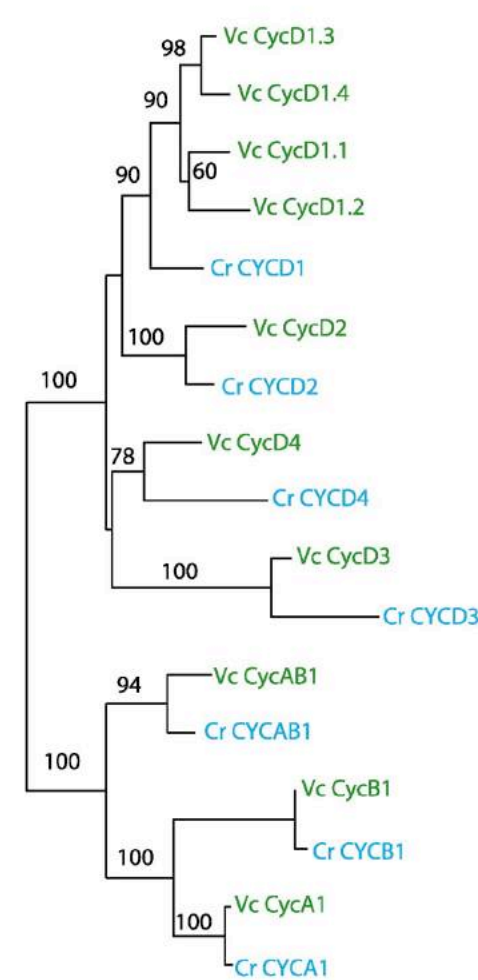


=> The most important changes are most likely in the regulatory regions!

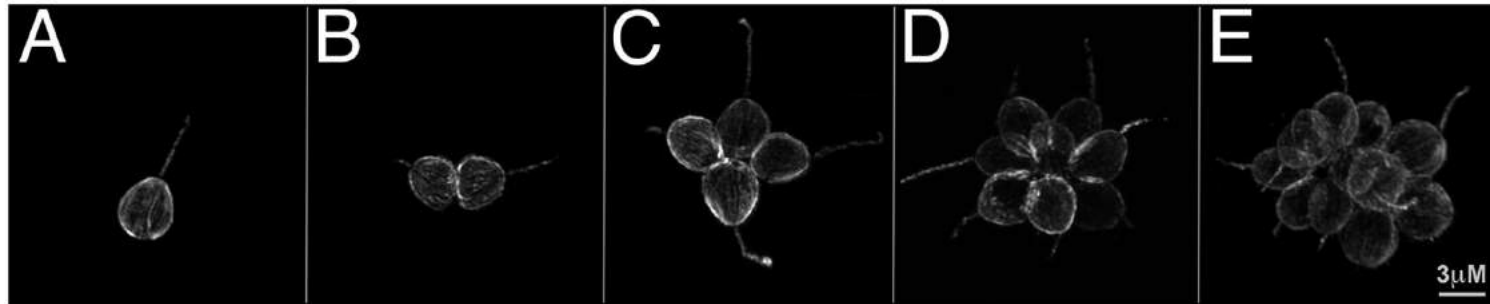
A ECM (perophorins)



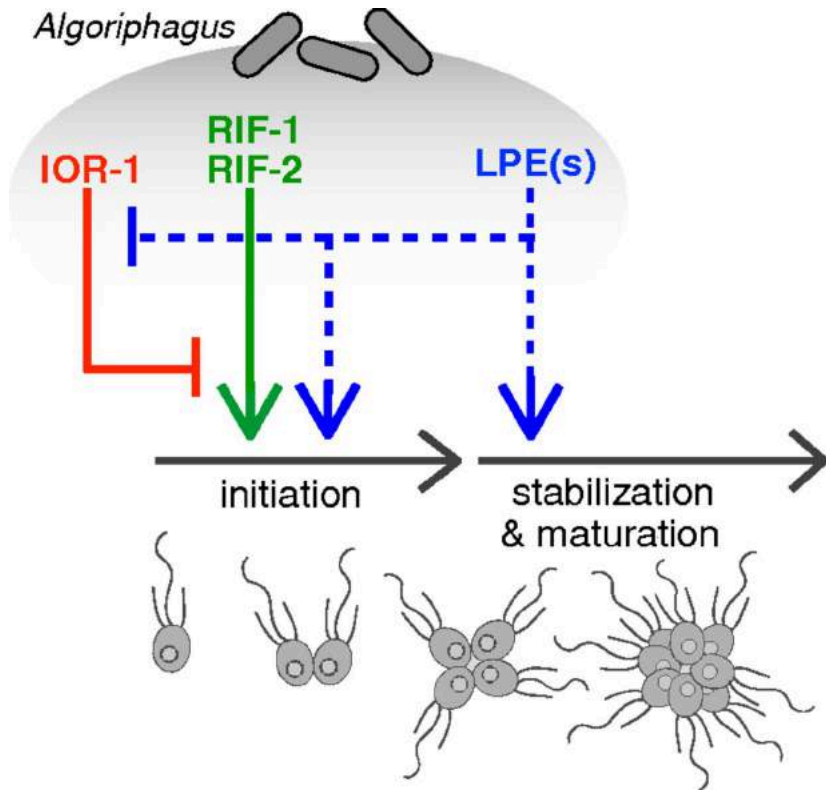
B sejtciklus



The origins of multicellularity – a role for the microbiome?

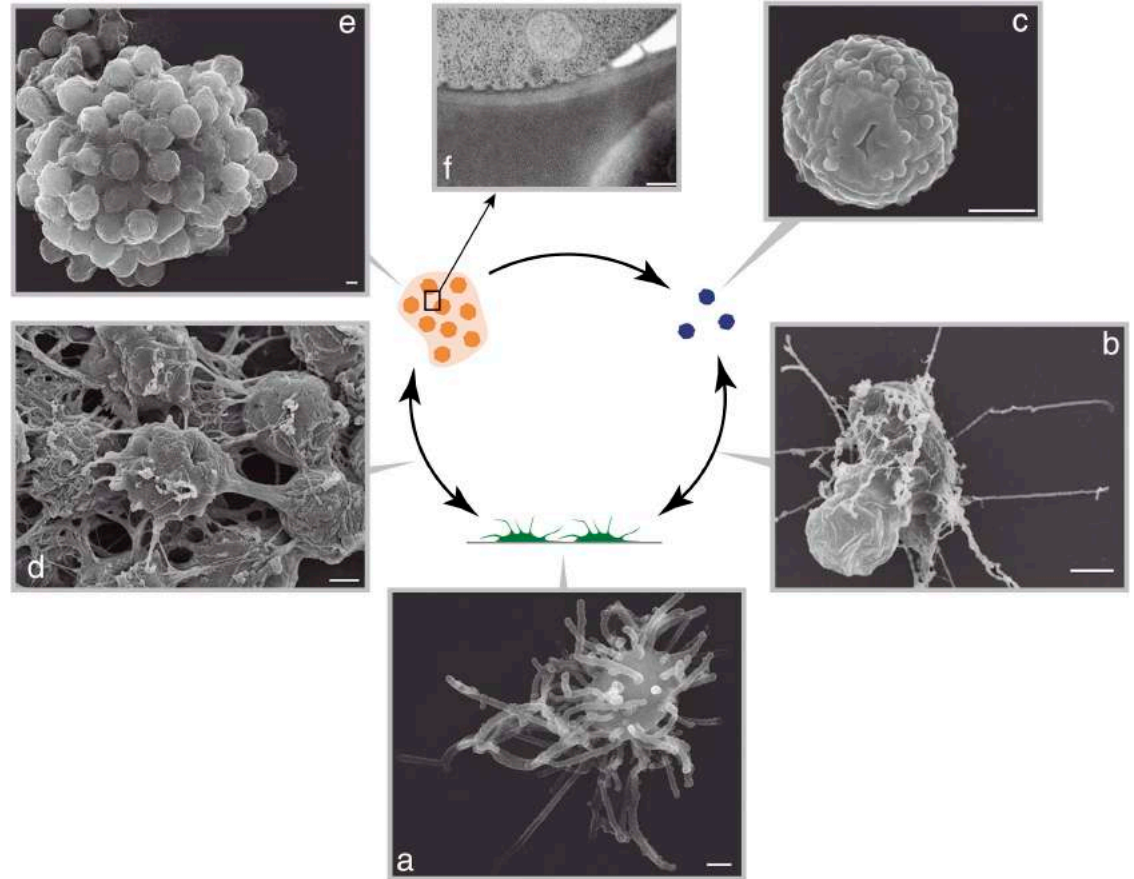
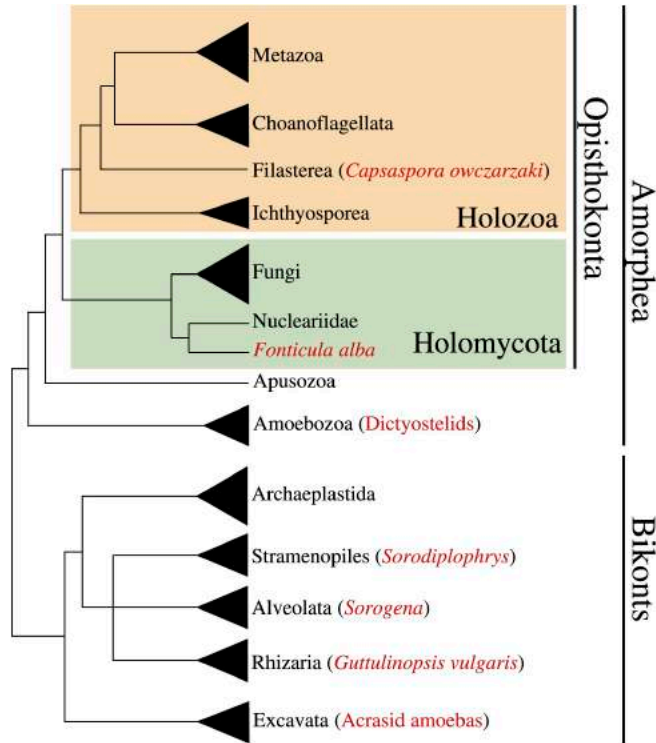


- In some Choanoflagellates the daughter cells stay together and form rosettes



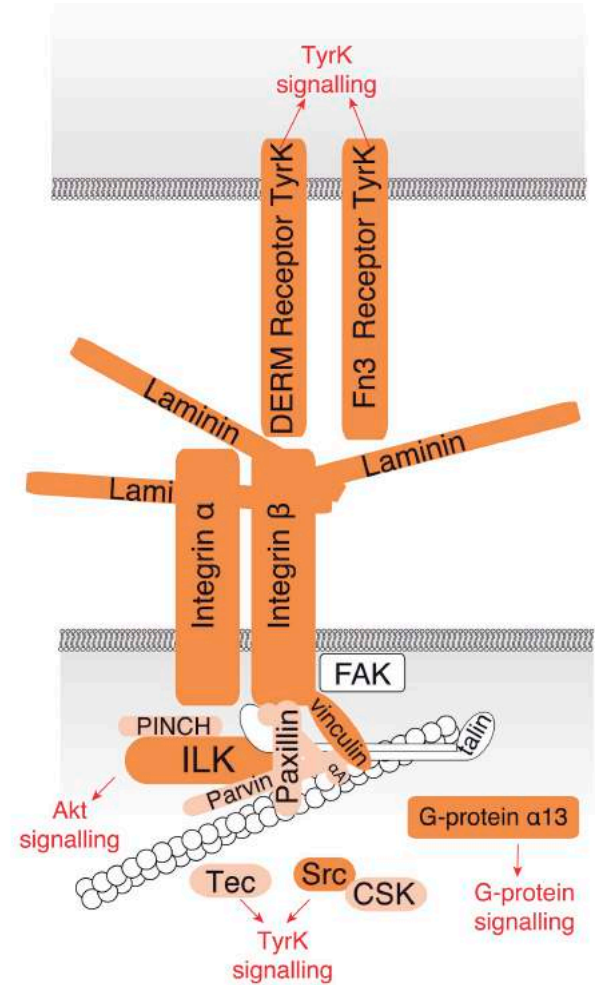
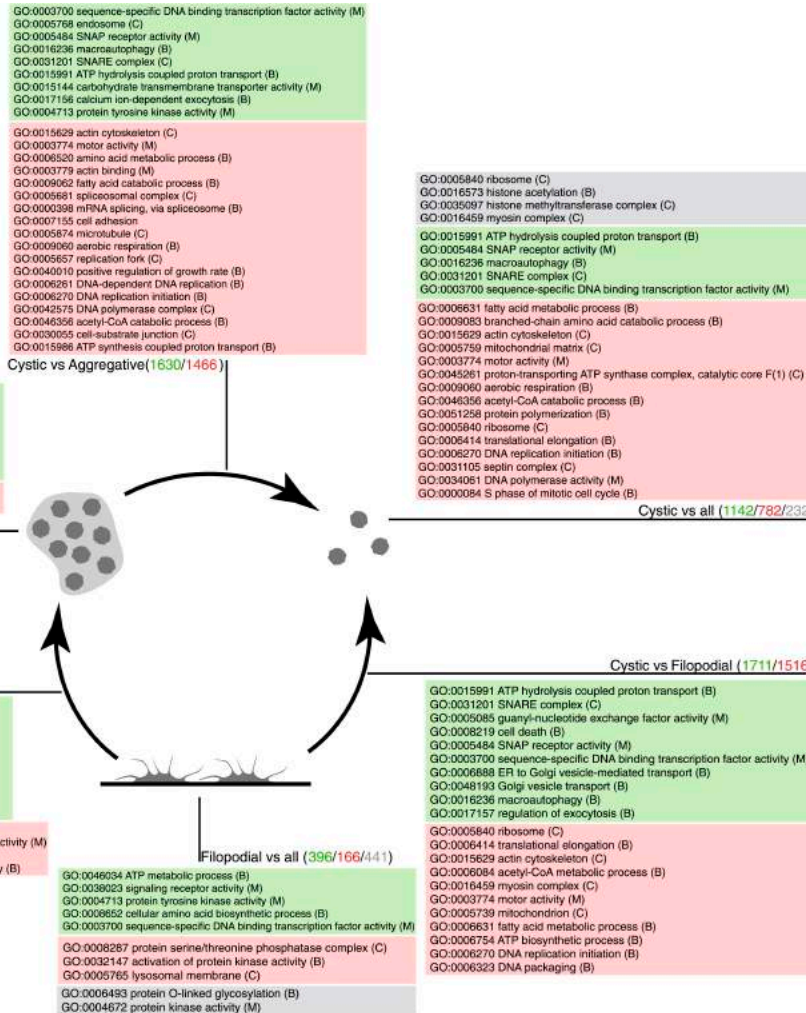
- To form rosettes it is necessary to have in the environment bacteria that secrete bioactive lipids.

The origins of multicellularity 3. – the *Capsaspora* genome



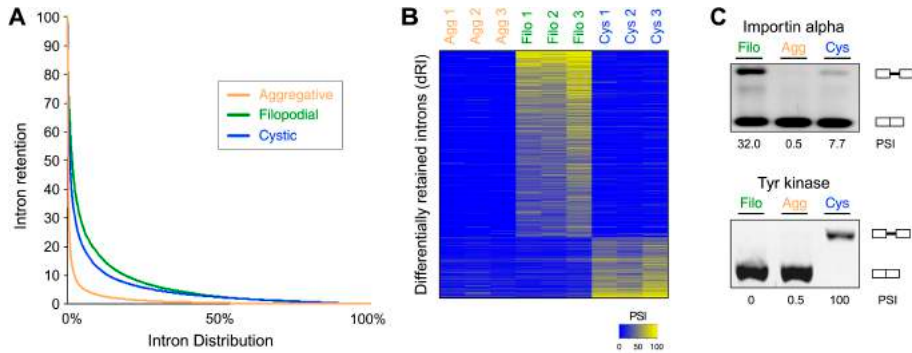
- *Capsaspora* is a unicellular amoeboid species rich in TFs
- The filopodial wandering form sometimes develop directly into growing cysts, but under certain conditions this transition happens through an aggregate that is bound together by an ECM

The origins of multicellularity 3. – the *Capsaspora* genome



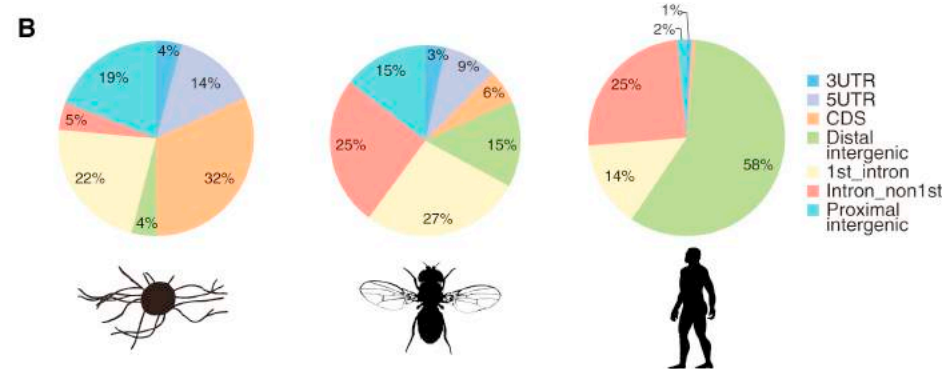
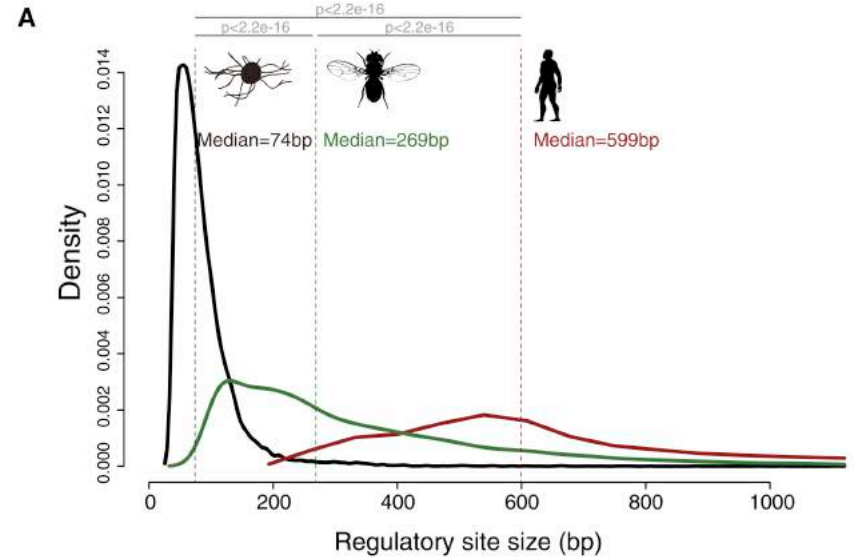
- Different phases of the life-cycle show characteristic transcriptomic signatures

The origins of multicellularity 3. – the *Capsaspora* genome



(Sebé-Pedros et al. (2013) *eLife*)

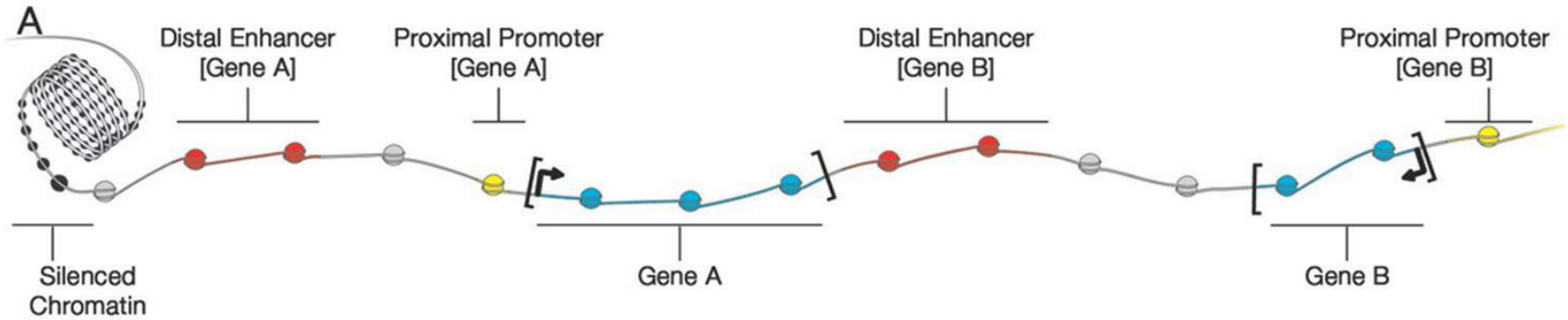
- In the case of *Capsaspora* alternative-splicing (intron retention) contributes to the increase of the transcriptome size



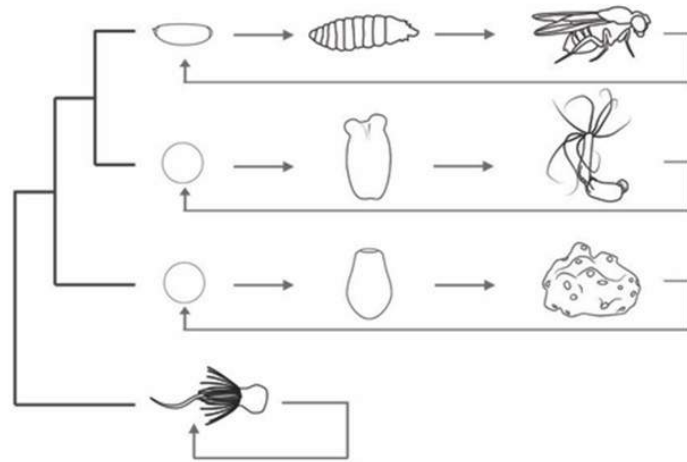
(Sebé-Pedros et al. (2016) *Cell*)

- Compared with Metazoan species, it is evident that regulatory sequences are small (only few TFs can bind), and are mostly proximal (distal enhancers, necessary for greater transcriptional complexity apparently arose later in development)

Multicellularity requires complex regulation

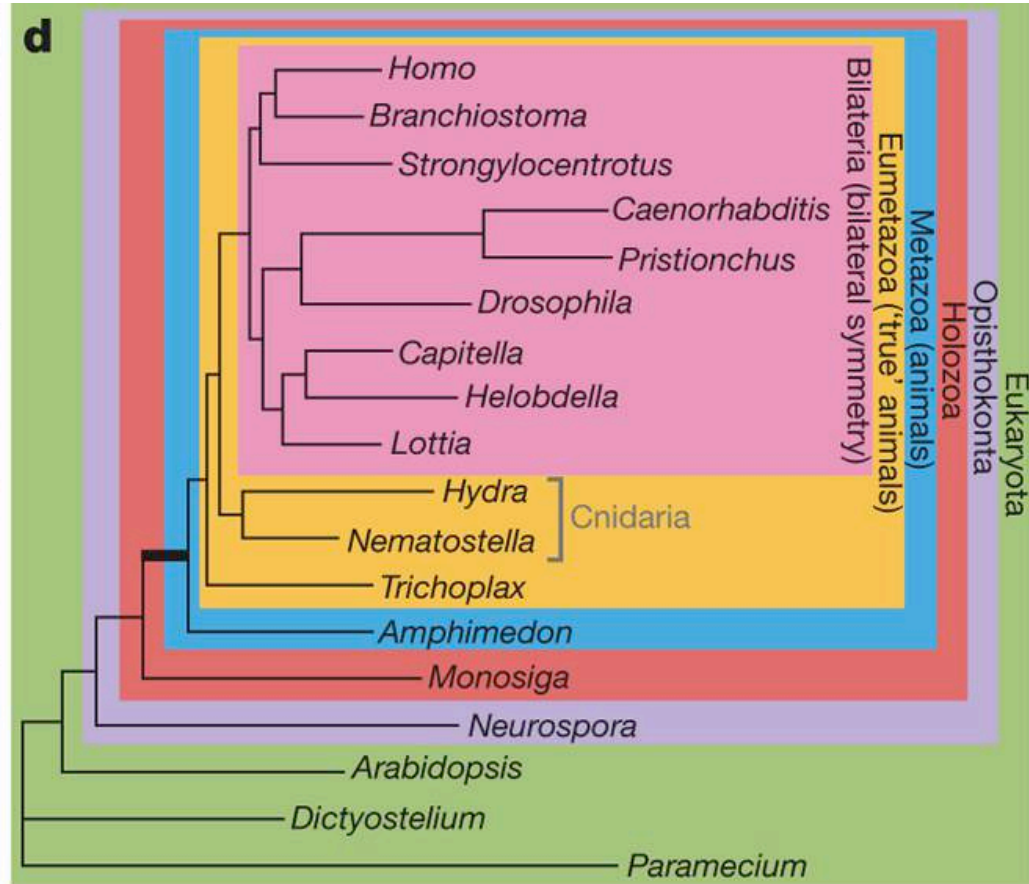


B



		Regulatory Features				
Phylogeny	Bilateria	■	■	■	■	■
	Eumetazoa	■	■	■	■	■
	Metazoa	■	■	■	■	■
	Choanoflagellate	■	■	■	■	■
		Developmental TF	Proximal cis-regulation	Distal cis-regulation	Polycomb repressive complex silencing	Higher-order chromatin structure

The sponge genome (*Amphimedon queenslandica*)

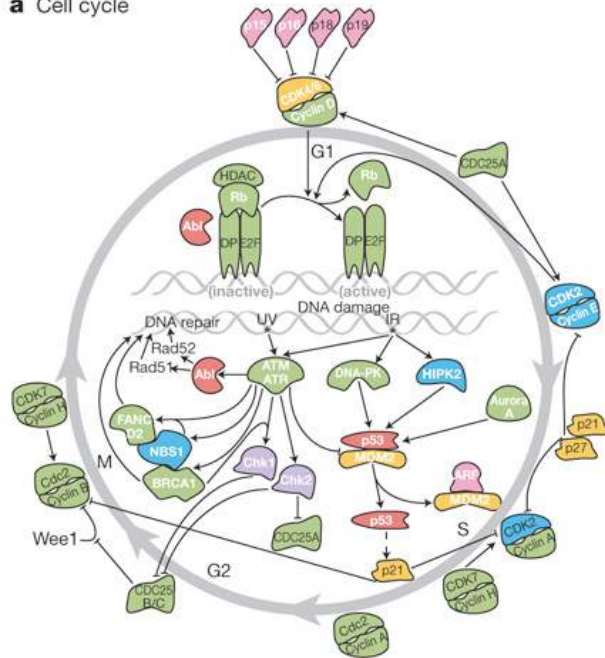


0.1 changes per site

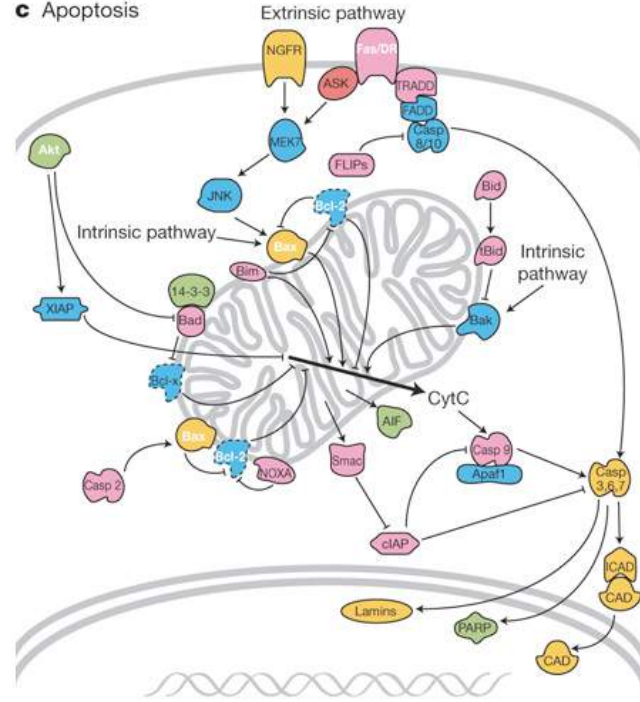
- ~30 000 protein coding genes, 63% have orthologs in other animals
- 84% of the ancient Metazoan introns are present

Ancient cell cycle genes vs. recent apoptosis genes

a Cell cycle



c Apoptosis



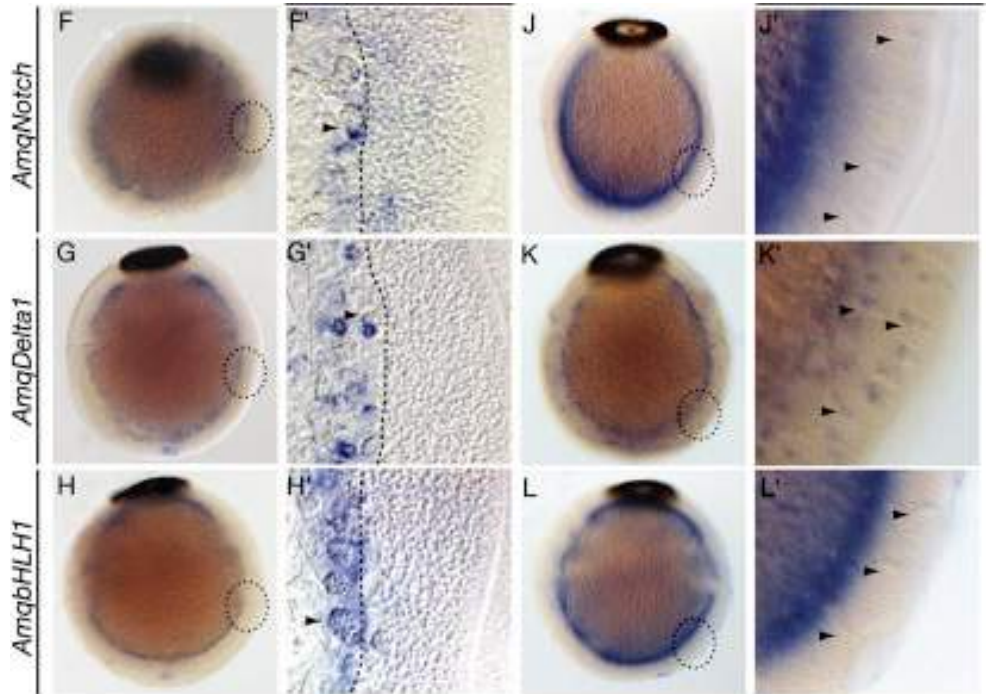
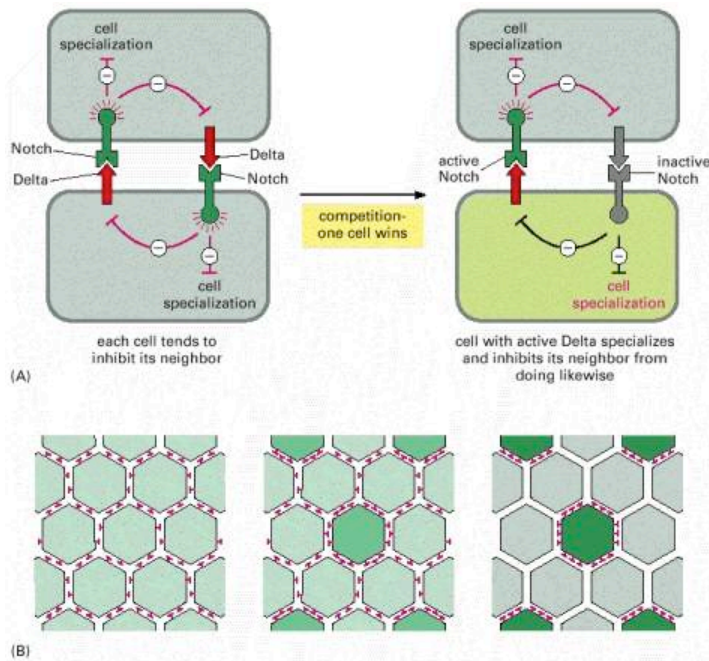
- | | | |
|--|--|--|
| ■ Ancient eukaryotic | ■ Holozoan origin | ■ Eumetazoan origin |
| ■ Opisthokont origin | ■ Animal origin | ■ Bilaterian/vertebrate origin |

- While most cell cycle genes are derived from the ancient Eukaryotic gene-set, programmed cell death is an animal invention

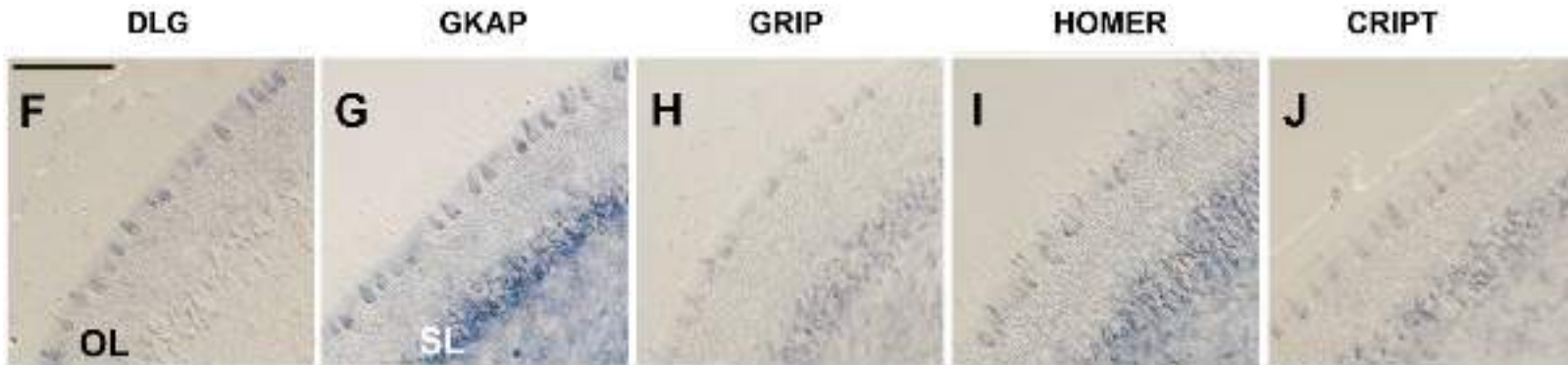
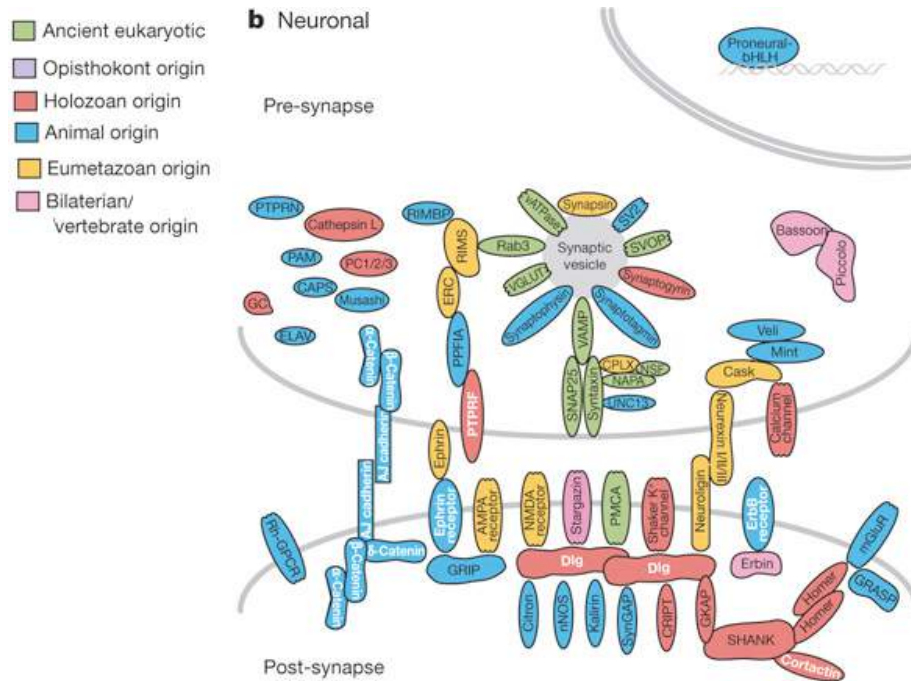
The origins of the nervous system



- Components of the Notch-Delta signaling pathway, involved in nervous system development are present in sponges and are expressed in larval cell types



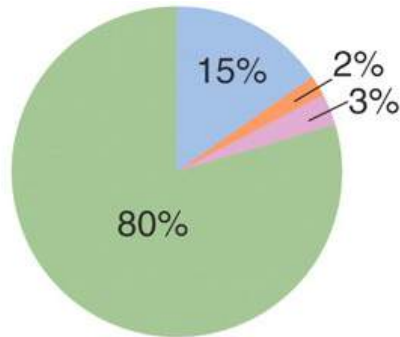
The origins of the nervous system: members of the post-synaptic complex are present in sponges



The origin of Eumetazoan genes



A



- Type I (completely novel)
- Type II (novel domain)
- Type III (novel pairing)
- Ancient

Type I Novelty: SMAD Family Proteins



Type II Novelty: Notch Proteins



Type III Novelty: Lim Homeodomain Proteins



- 80% of the ancestral Eumetazoan gene set has homologs in non-animal species

- the remaining 20% is Eumetazoan “invention”:

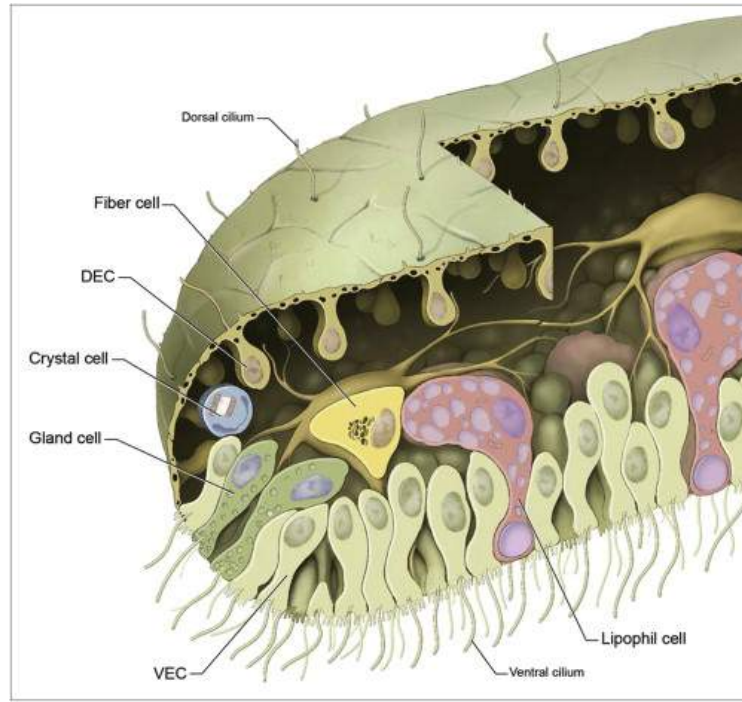
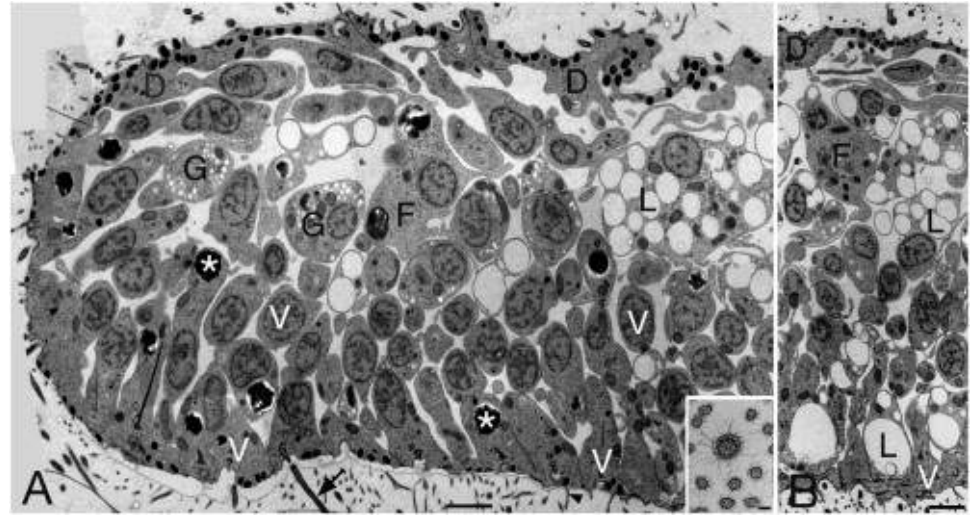
- 15% is completely novel, no sequence homology with other groups (Type I)

- 2% genes, which have some domains that already exist in other groups, but other domains are novel (Type II)

- 3% genes with domains that exist in other groups, but not in this particular combination (Type III)

(Putnam et al. (2007) *Science*)

A placozoan (*Trichoplax*) genome



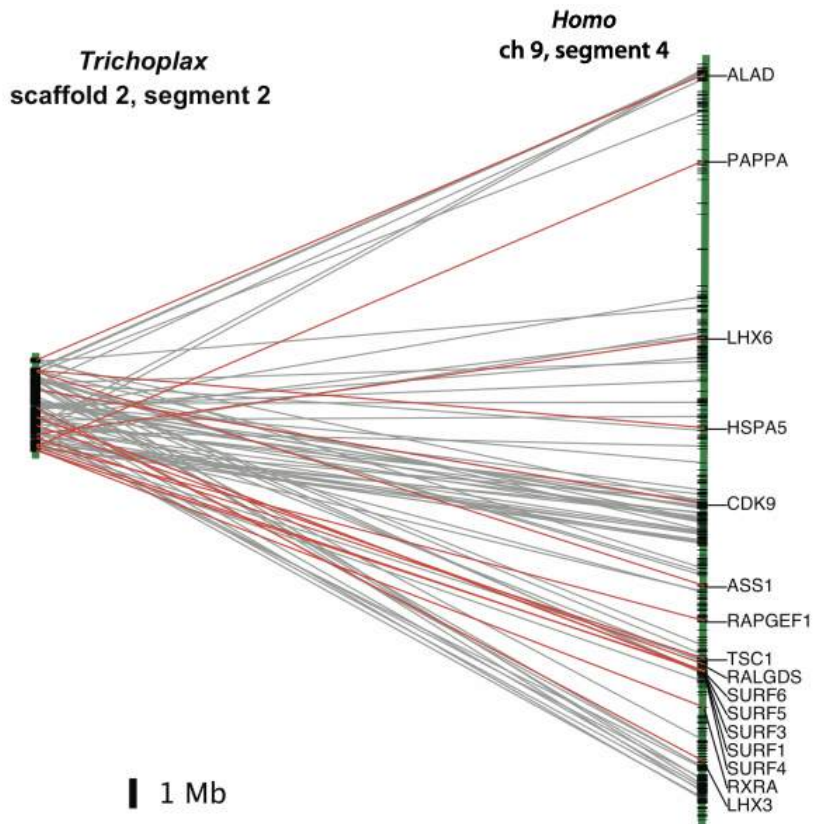
(Srivastava et al. (2008) *Nature*)

(Smith et al., (2014) *Curr Bio*)

A placozoan (*Trichoplax*) genome

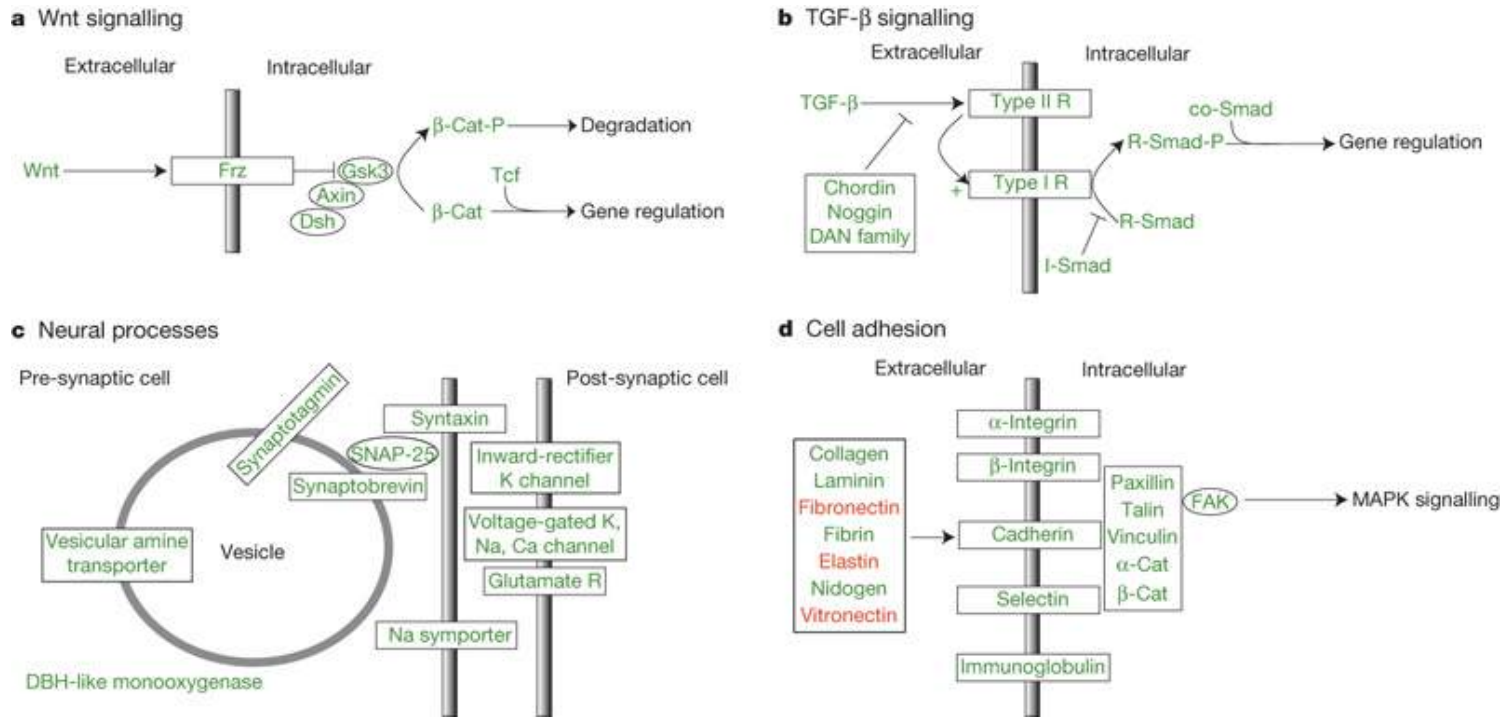
	Trichoplax	Nematostella	Drosophila	C. elegans	H. sapiens
Trichoplax	11511	5798	4319	3692	5500
Nematostella		27273	6144	4537	6977
Drosophila			14039	4523	5757
C. elegans				20074	4814
H. sapiens					22842

There are 1,127 trichoplax genes with MBH to a human genes but neither to a fly nor a worm gene. On the other hand, trichoplax has 417 genes with a MBH to either fly or worm, but not human.



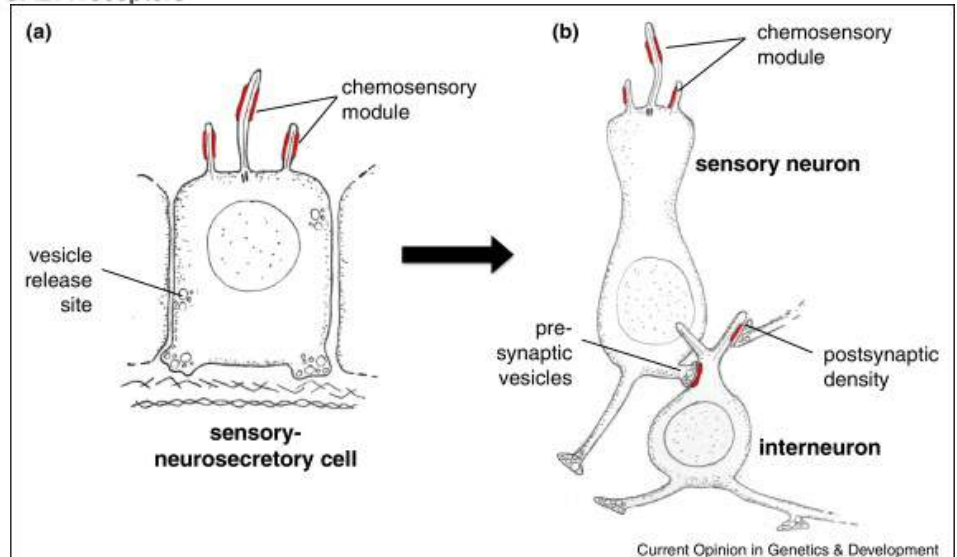
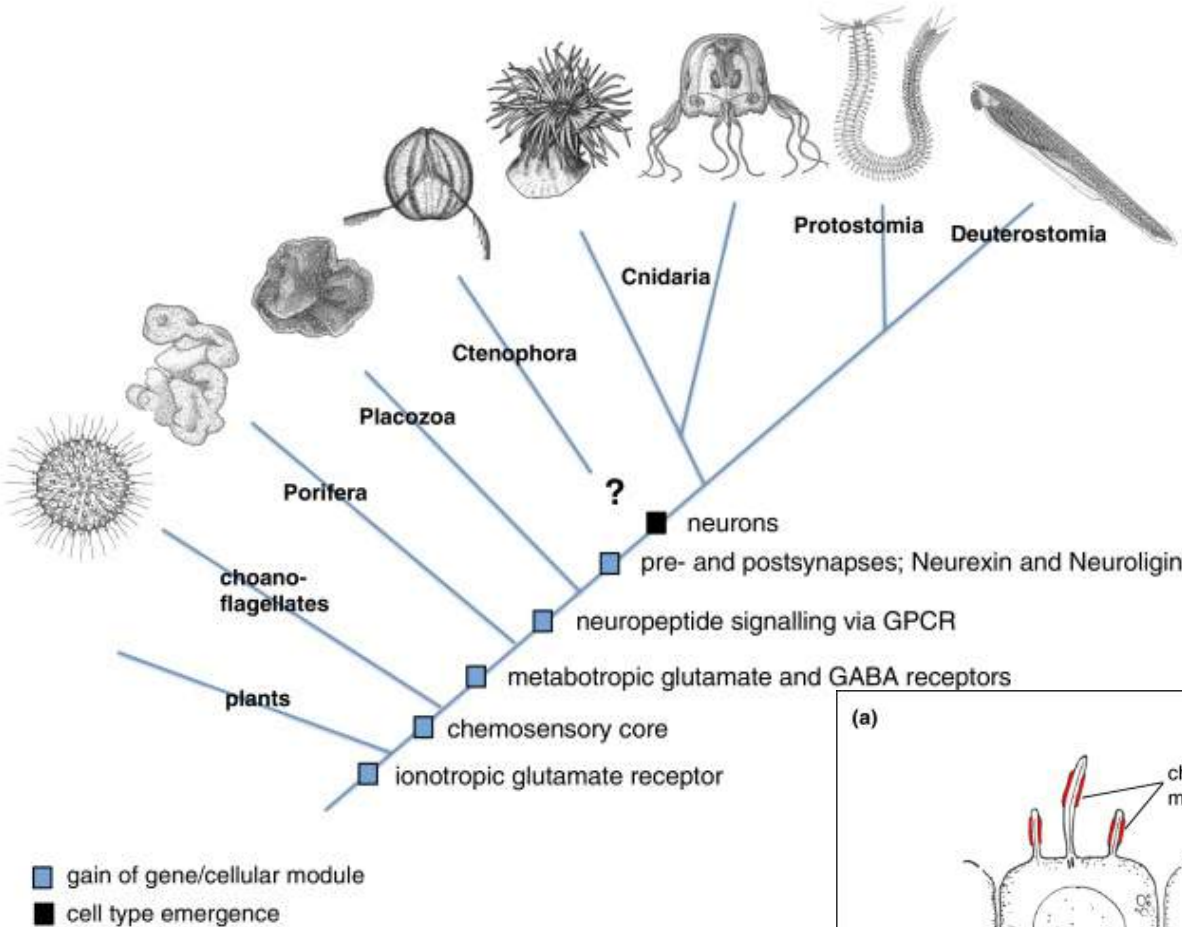
- 87% of the 11 511 genes have homologs in other animals
- In syntenic regions, 82% of human introns have a *Trichoplax* counterpart
- Large-scale synteny (higher than from flies and worms)

A placozoan (*Trichoplax*) genome



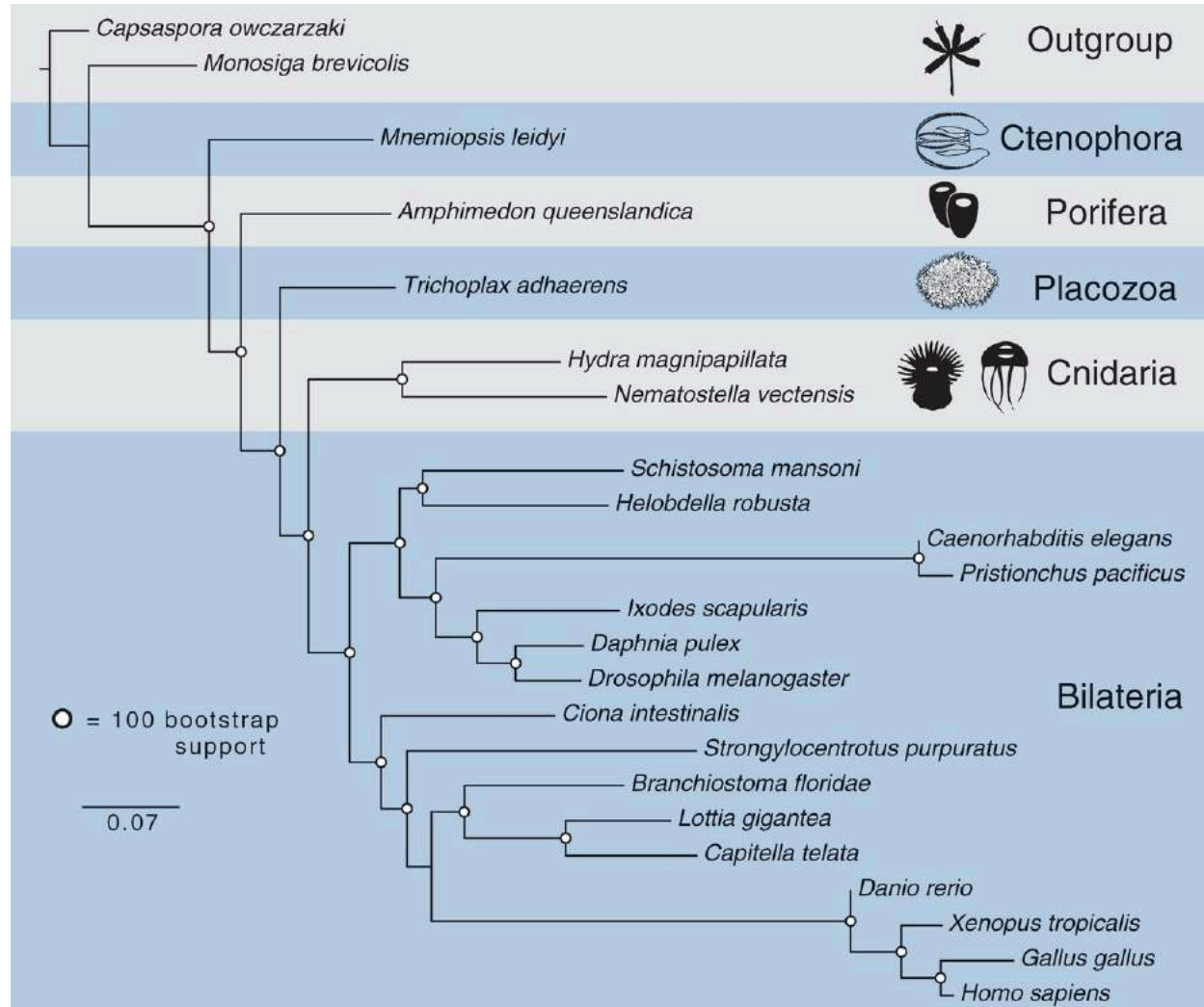
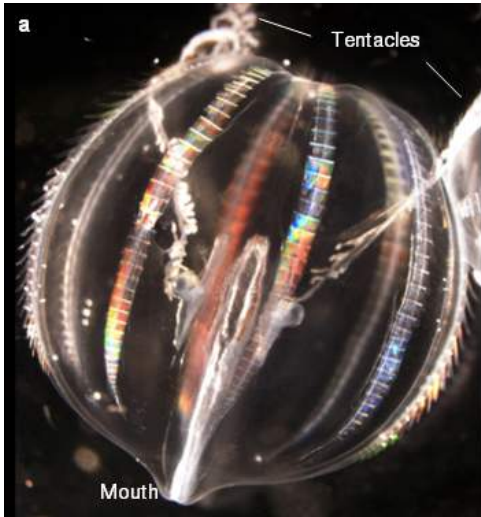
“Transmembrane proteins important in nerve conduction (multiple candidate ionotropic glutamate receptors) and in neurotransmitter release and uptake (for example, sodium neurotransmitter symporter) are encoded by the genome.”

The origins and evolution of the nervous system



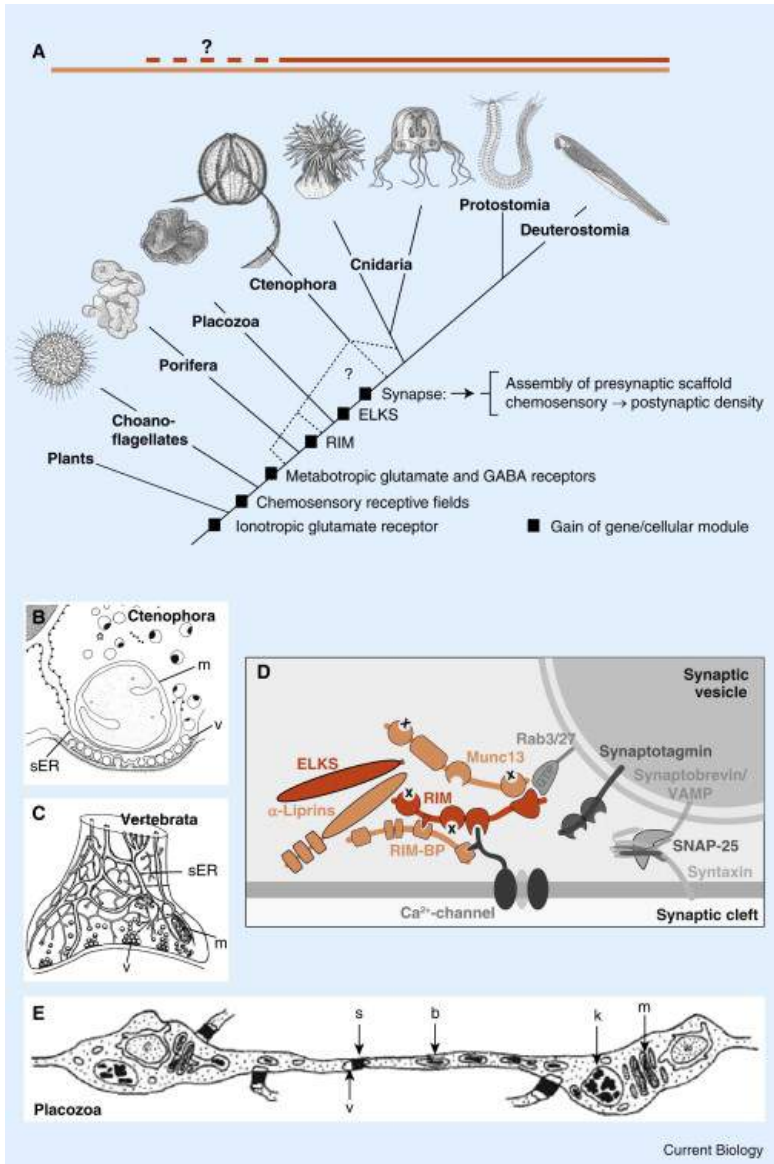
(Achim and Arendt (2014) *Curr Op Gen&Dev*)

The ctenophoran genome

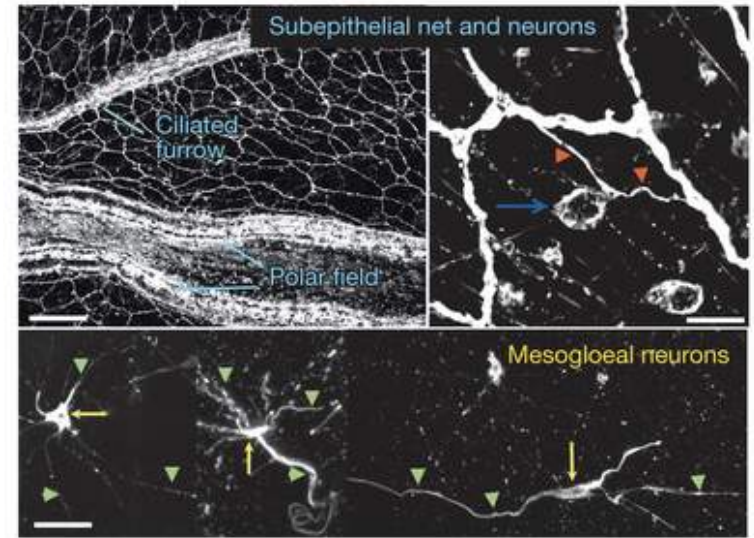


(Moroz et al. (2014) *Nature*)

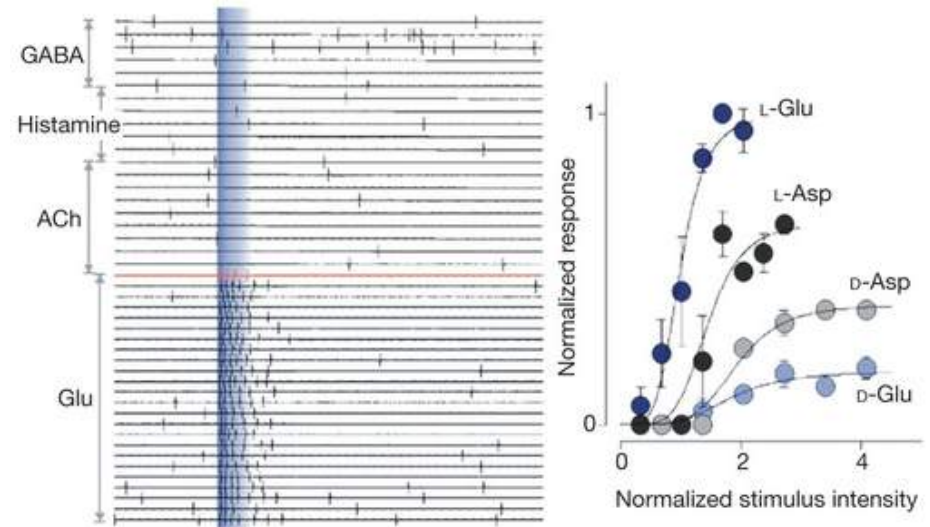
The ctenophoran genome



(Marlow and Arendt (2014) *Curr Bio*)

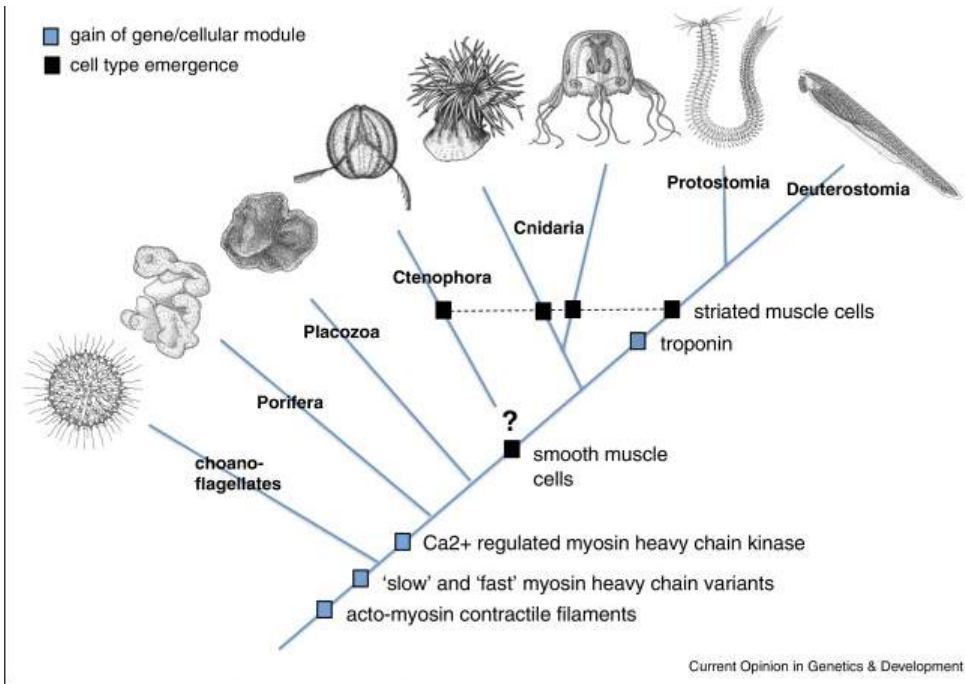


b Glu as neuromuscular transmitter in *Pleurobrachia*



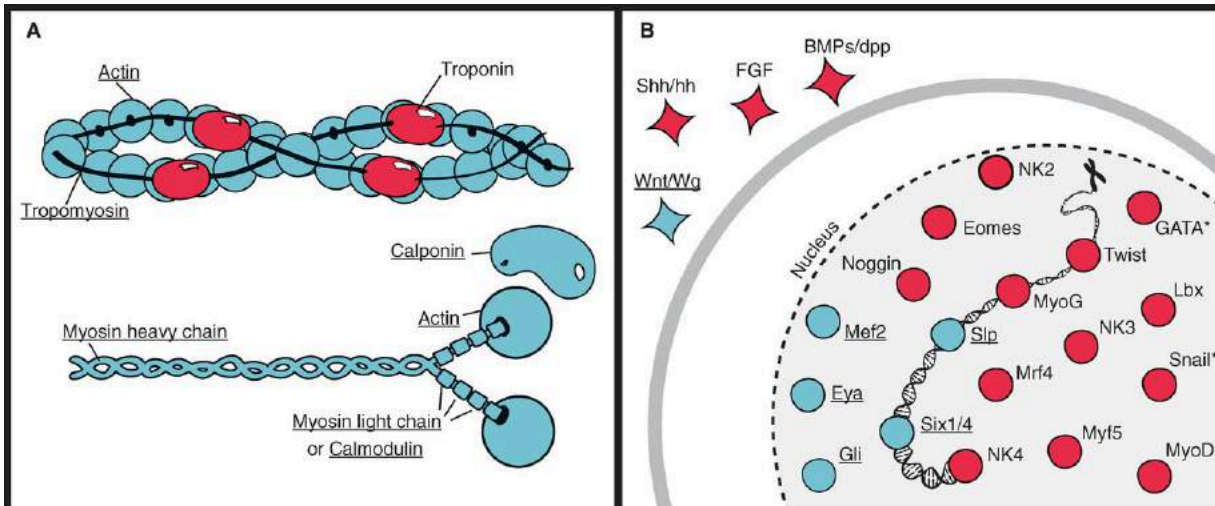
(Moroz et al. (2014) *Nature*)

The ctenophoran genome



- The proteins that form muscle fibers are present in Ctenophores, their transcriptional regulation is different to that observed in Bilateria
- Skeletal muscle might have evolved independently 2-3 times during evolution!

(Achim and Arendt (2014) *Curr Op Gen&Dev*)

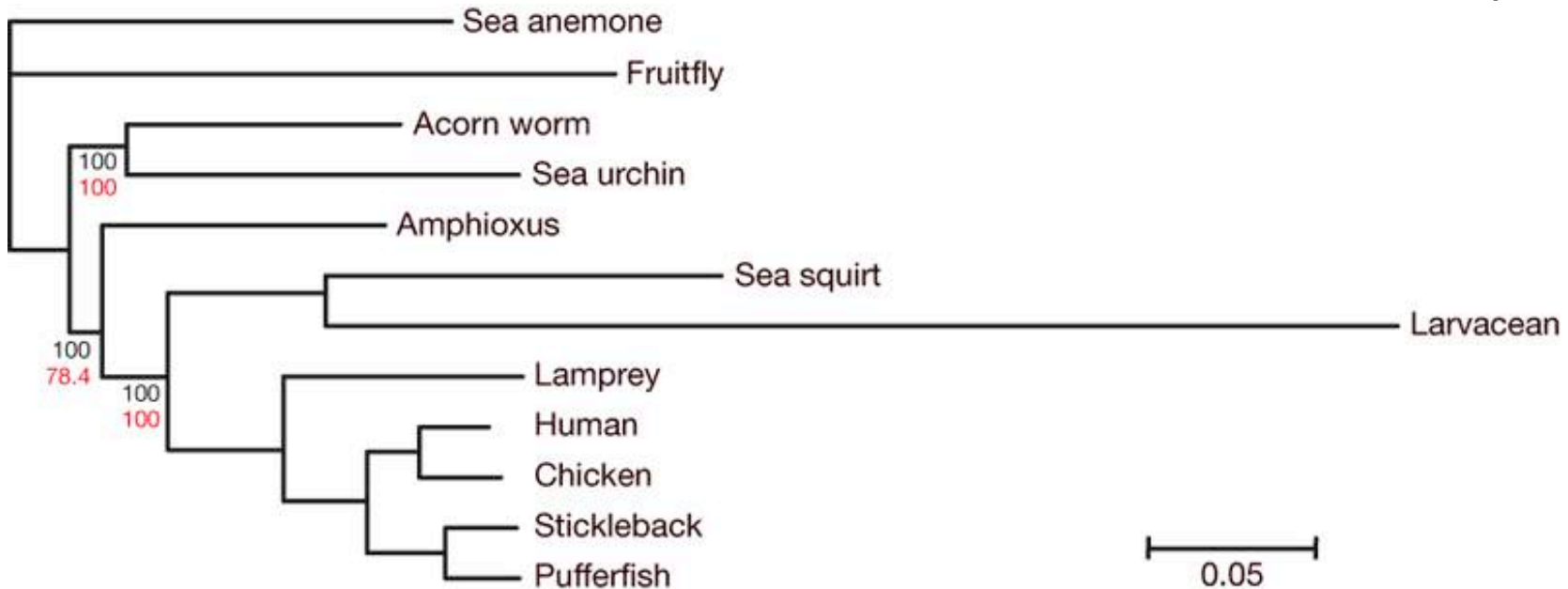


(Ryan et al. (2013) *Science*)

The *Amphioxus* genome and the origin of the vertebrate genome

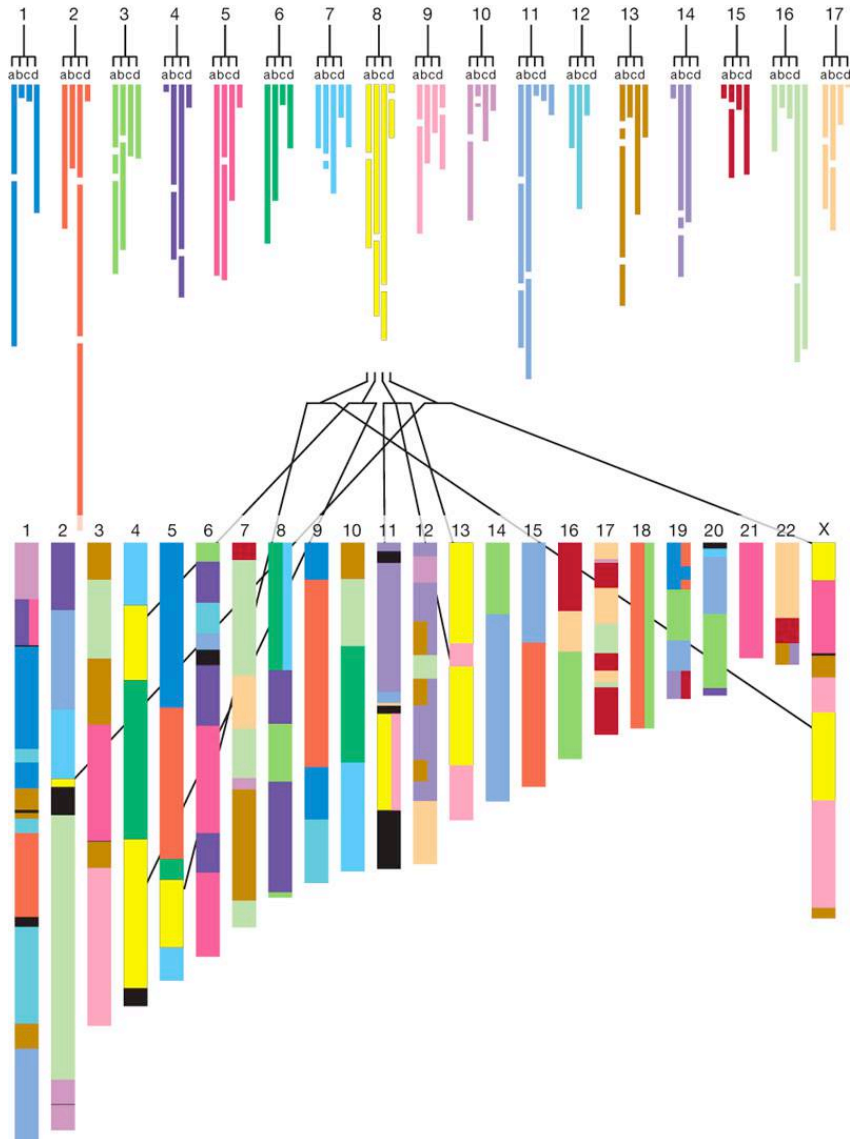


- the common ancestor live 550 Mya
- genome size is ~520 Mb, on 19 chromosomes (17 scaffolds)
- ~20 000 protein coding loci
- 30% of the genome is derived from TEs
- 85% of the introns has a human counterpart



(Nicholas et al. (2008) *Nature*)

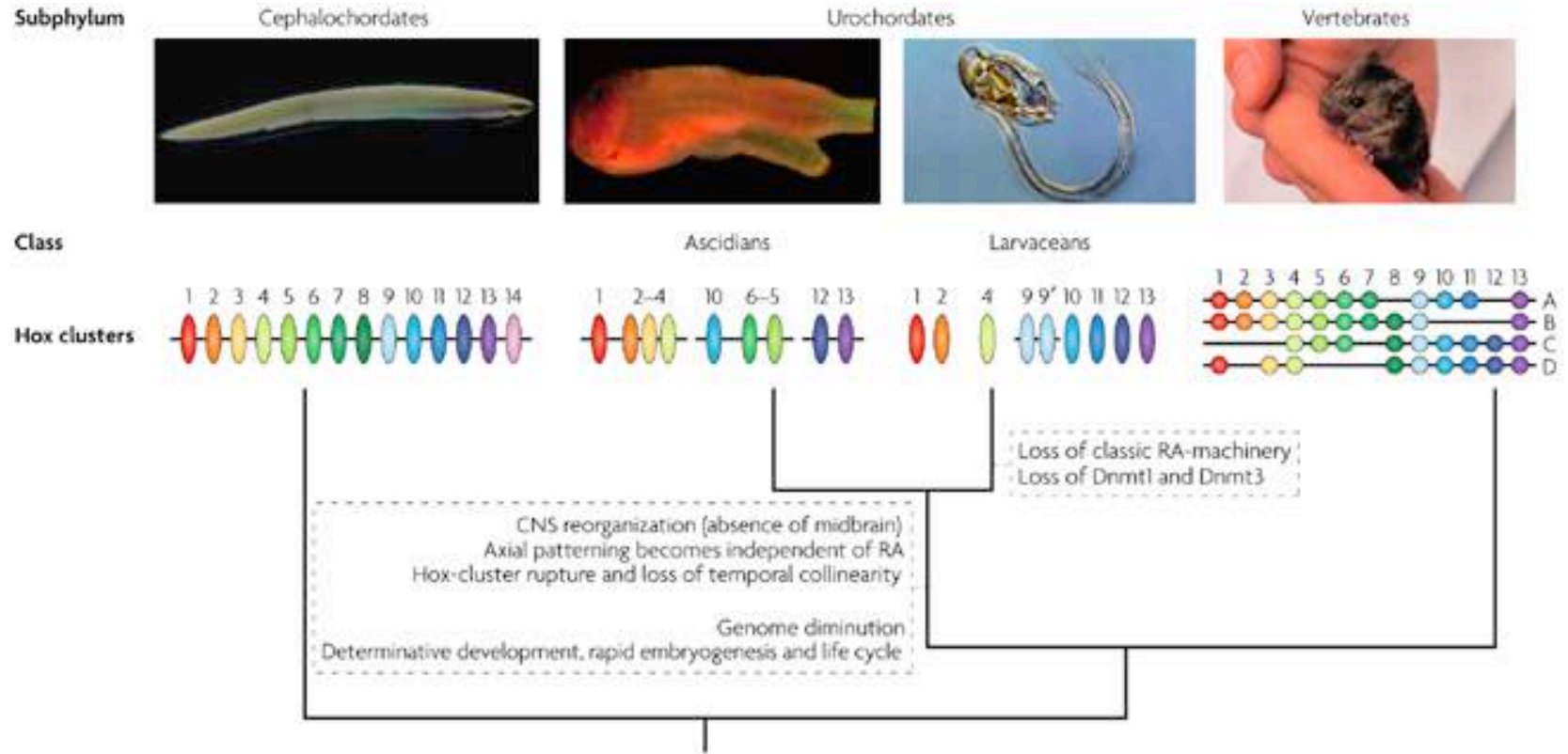
Genome duplications during the evolution of vertebrates



- the large scale synteny makes it possible to show the genome duplications during early vertebrate evolution: most *Amphioxus* genomic regions have 4 vertebrate counterparts



Genome duplications during the evolution of vertebrates: the Hox cluster

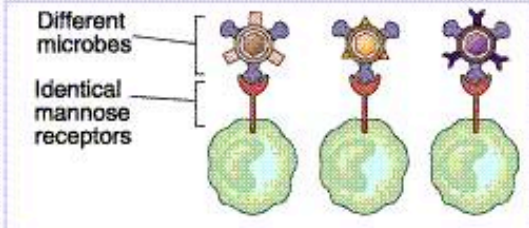
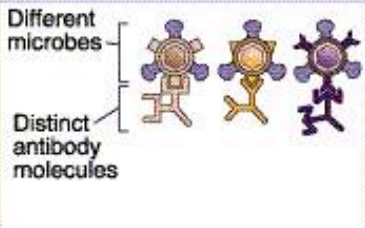
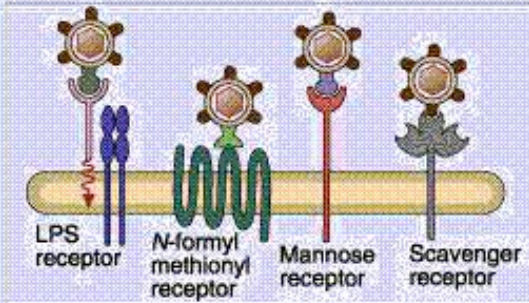
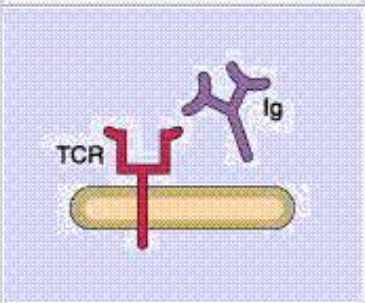


Nature Reviews | Genetics

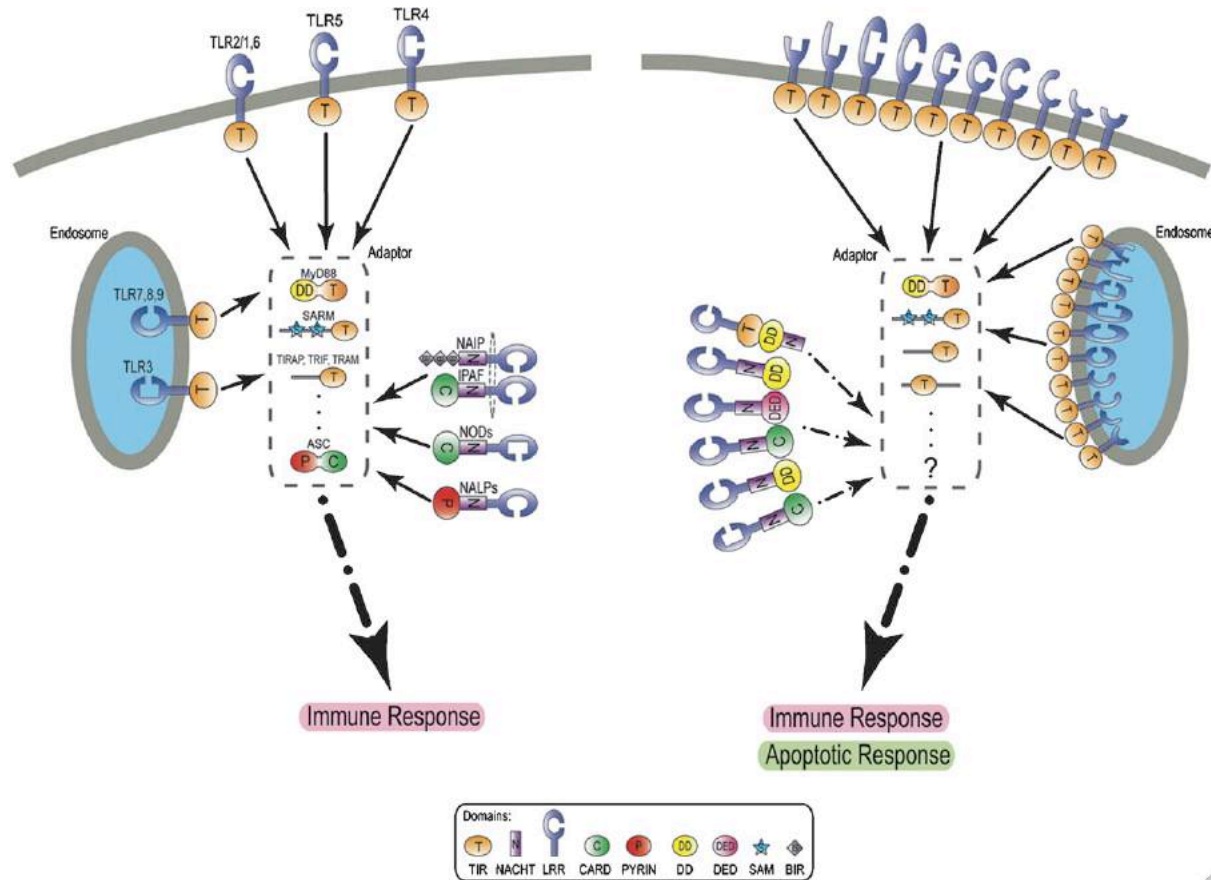
- one of the best examples for early genome duplications is the Hox-cluster
- in the *Amphioxus* genome a single, complete cluster is present
- in Urochordates the specialized life history caused the fragmentation of the cluster

(Cañestro et al. (2007) *Nat Rev Gen*; Holland et al. (2008) *Genome Res*)

The evolution of innate immunity in *Amphioxus*

	Innate immunity	Adaptive immunity
Specificity	<p>For structures shared by classes of microbes ("molecular patterns")</p>  <p>Different microbes</p> <p>Identical mannose receptors</p>	<p>For structural detail of microbial molecules (antigens); may recognize non-microbial antigens</p>  <p>Different microbes</p> <p>Distinct antibody molecules</p>
Receptors	<p>Encoded in germline; limited diversity</p>  <p>LPS receptor</p> <p>N-formyl methionyl receptor</p> <p>Mannose receptor</p> <p>Scavenger receptor</p>	<p>Encoded by genes produced by somatic recombination of gene segments; greater diversity</p>  <p>TCR</p> <p>Ig</p>
Distribution of receptors	<p>Non-clonal: identical receptors on all cells of the same lineage</p>	<p>Clonal: clones of lymphocytes with distinct specificities express different receptors</p>
Discrimination of self and non-self	<p>Yes; host cells are not recognized or they may express molecules that prevent innate immune reactions</p>	<p>Yes; based on selection against self-reactive lymphocytes; may be imperfect (giving rise to autoimmunity)</p>

The evolution of innate immunity in *Amphioxus*

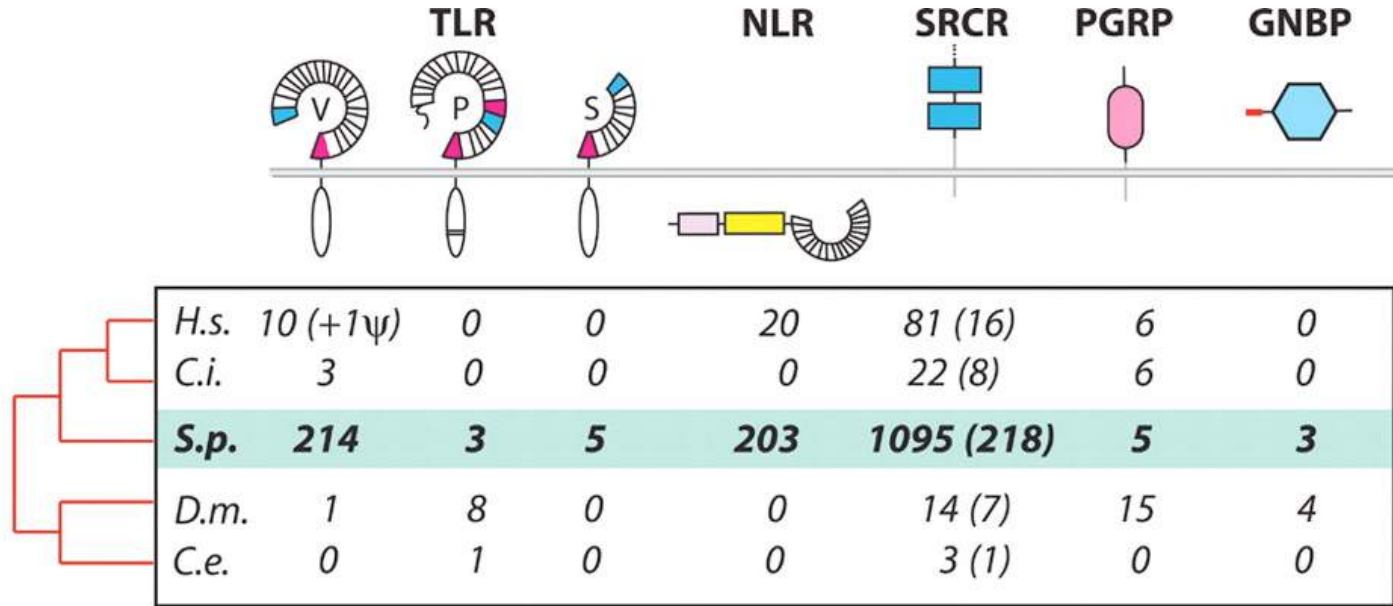


- 2-3x as many *Toll-receptor* genes than in vertebrates

- expansion of the apoptotic genes (also probably related to innate immunity)

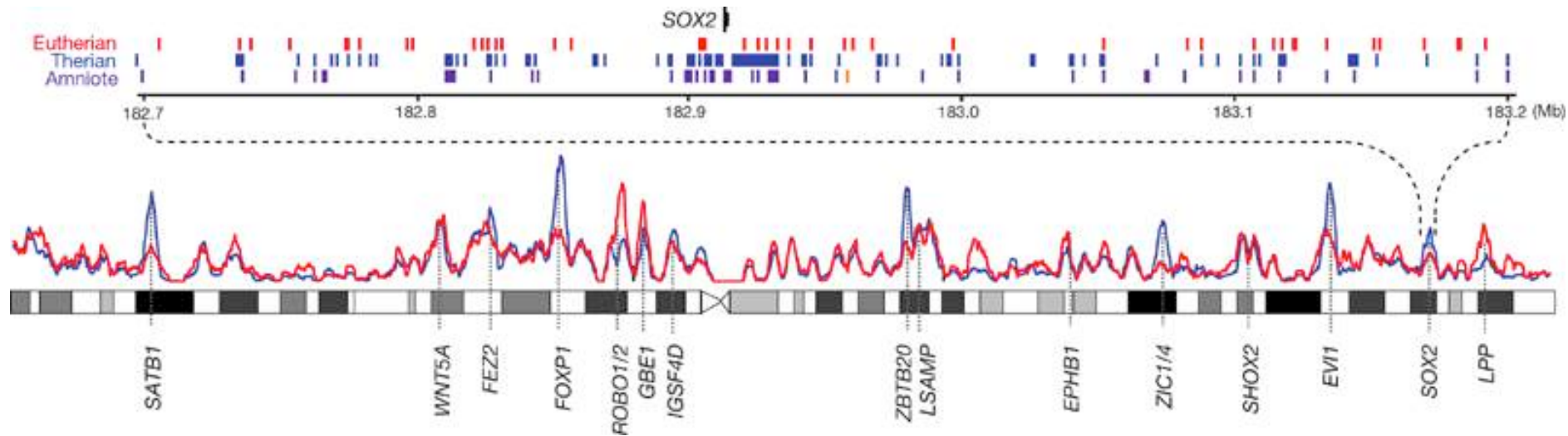
(Holland et al. (2008) *Genome Res*)

Convergent evolution of innate immunity in sea urchins (*S. purpuratus*)



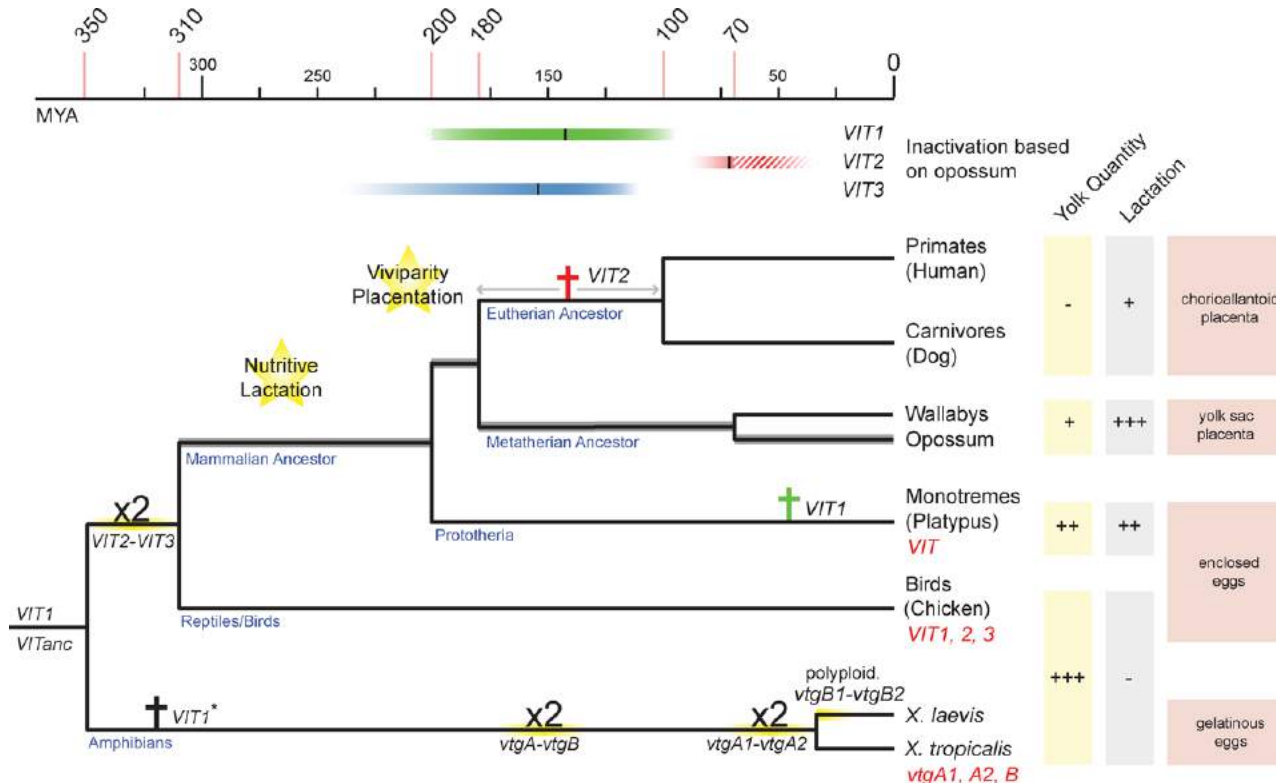
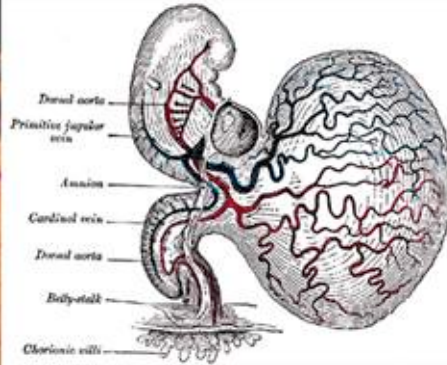
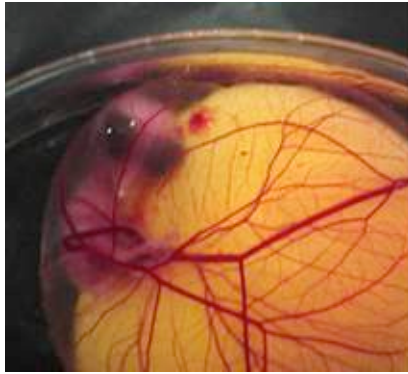
- 4-5% of all genes in the sea urchin genome are involved in innate immunity

Origins of mammalian regulatory sequences – the opossum (*Monodelphis domestica*) genome



- Half of the CNEs found in Amniotes, and 35% of those in placental mammals are organized in 204 large clusters
- These surround approx. 240, slowly evolving, essential developmental genes => the fine-tuning of pleiotropic genes is an important avenue for evolutionary change.

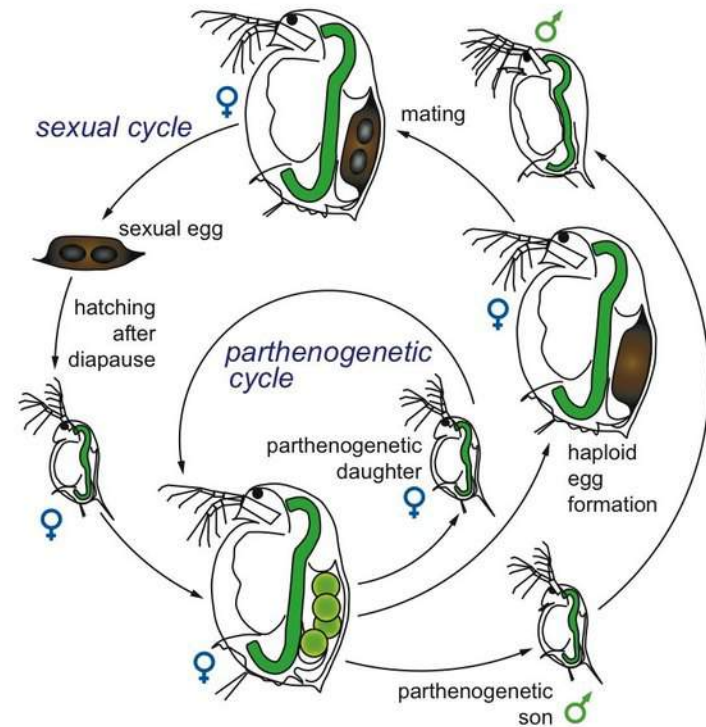
The evolution of lactation (and placenta) resulted in the decay of yolk protein coding genes



(Brawand et al. (2008) *PLoS Biol*)



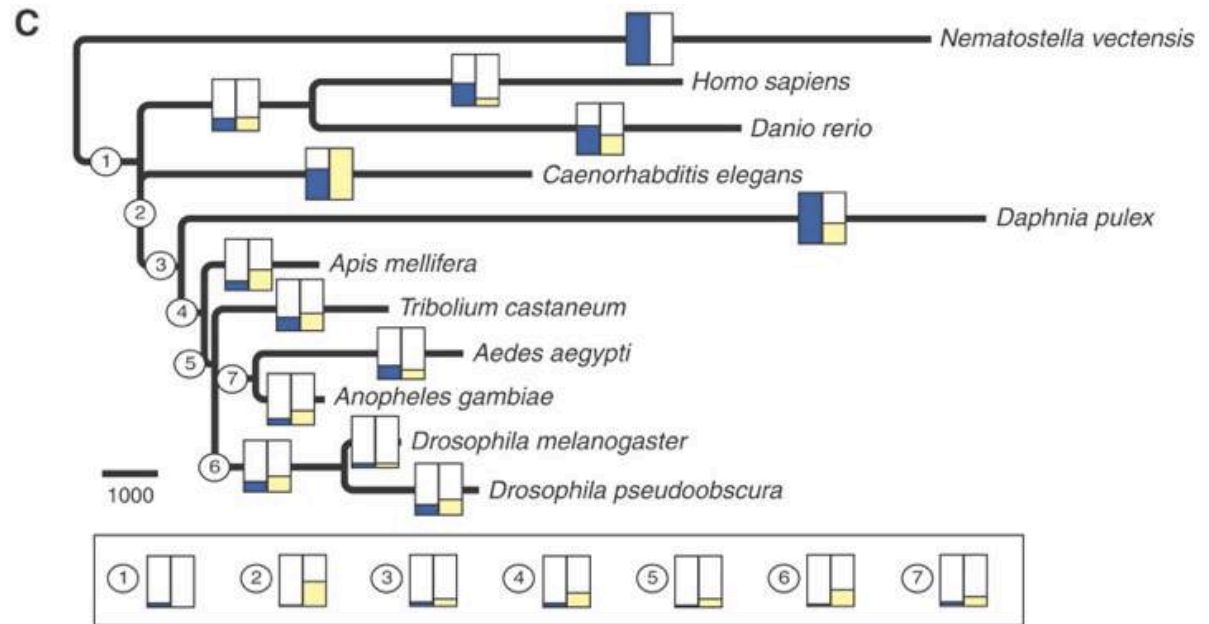
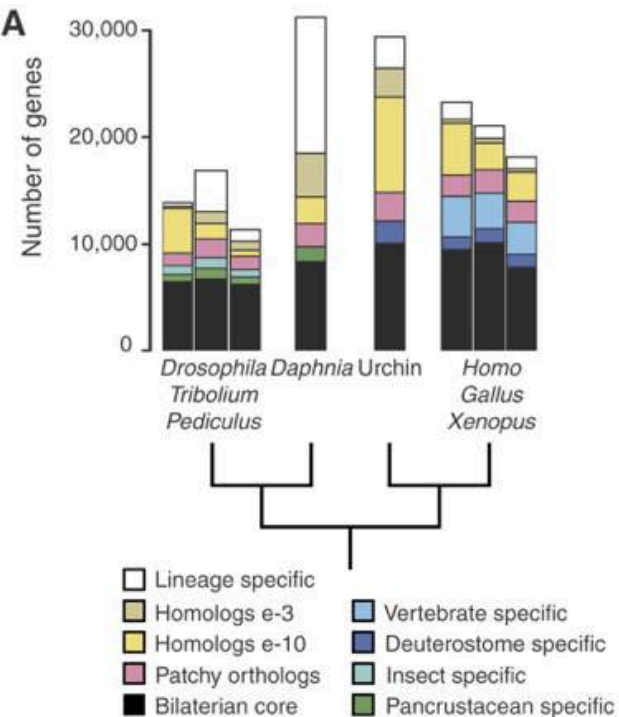
Adaptive genomes - the *Daphnia* genome



- long parthenogenetic and short sexual cycles mix during the *Daphnia* life
- this is an excellent example of an “ecoresponsive genome”, suitable for quick adaptations to the changing environment



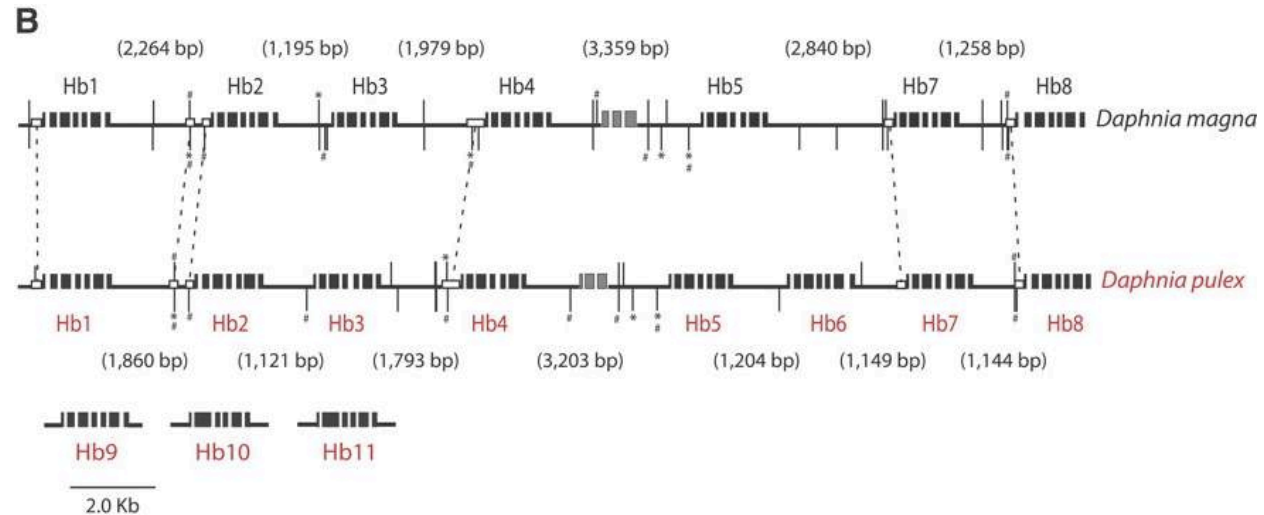
Adaptive genomes - the *Daphnia* genome



- the *Daphnia* genome is 200 Mb, but encodes 31 000 proteins (humans: ~20 000 protein encoding genes on 3,000 Mb)
- less TEs and shorter introns make the genome compact
- many genes are lineage-specific and are linked to the life-style of the animal (these are not complete novelties, just expansion of already existing gene families)



Adaptive genomes - the *Daphnia* genome

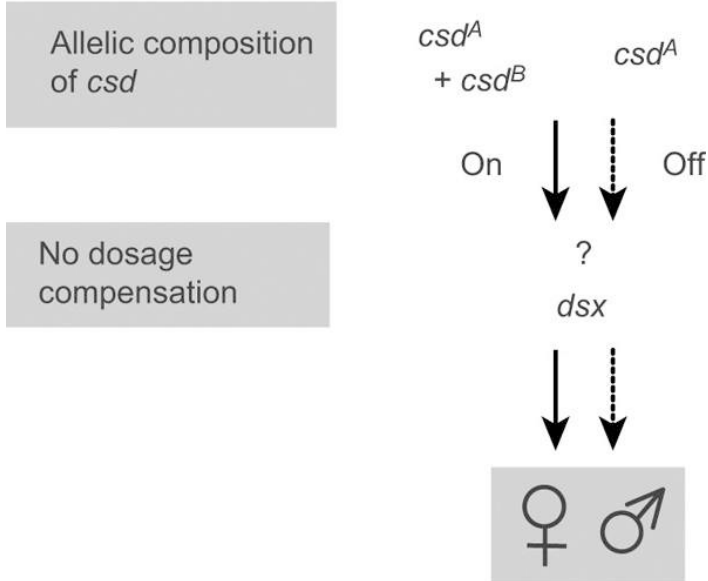


- the lack of oxygen results in the animals turning red
- this is due to the excess hemoglobin synthesis: duplicated copies of the hemoglobin gene are organized in a cluster that is regulated by hypoxia-elements

Epigenetic regulation of development: the honey bee (*Apis mellifera*) genome



Apis mellifera



- haploids and hemizygotes develop into males, diploids are females
- the genome of the queen and workers are identicalm the *only* difference is in their food: females fed with royal jelly develop into queens
- how can the same genome encode for so different phenotypes?

Methylation and cast-based societies

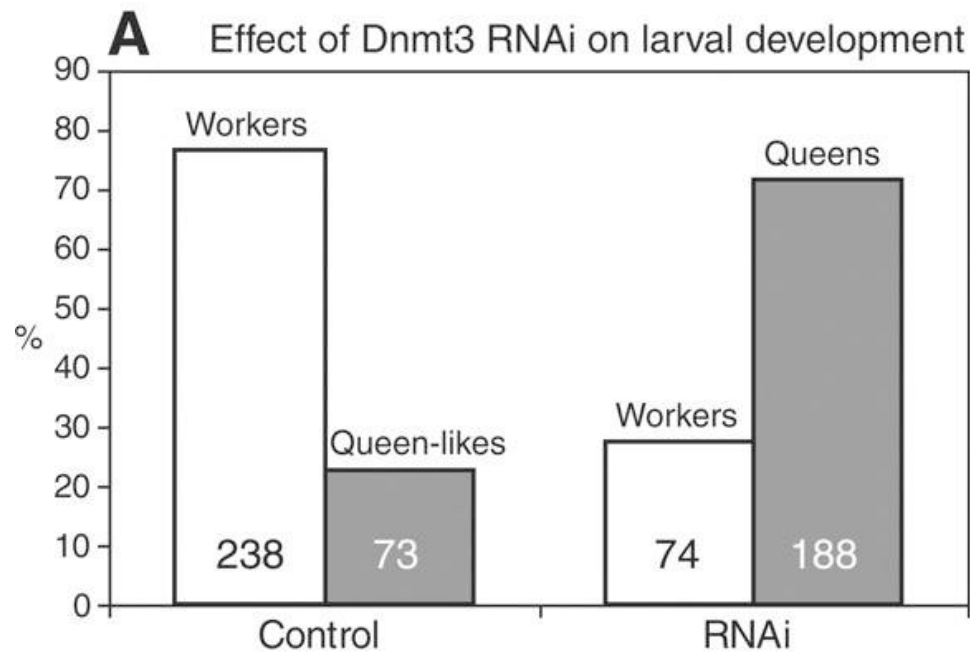


	Total	Methylated in Queens	Methylated in Workers	Methylated in Both Castes
CG	10,030,209	69,064	68,222	54,312
CHG	8,673,113	14	130	0
CHH	45,072,611	561	3,019 ^a	0

The thresholds used for methylation calls are detailed in the Methylation Assessment section.

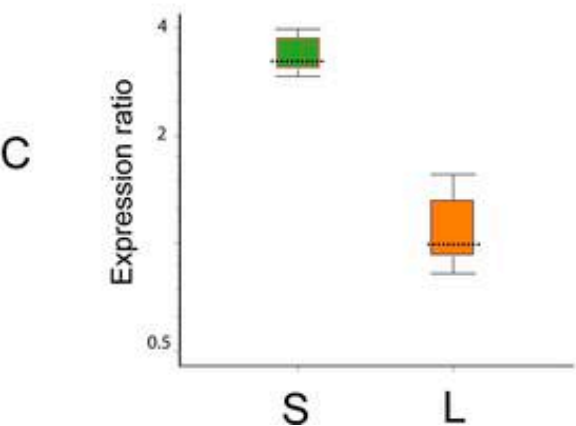
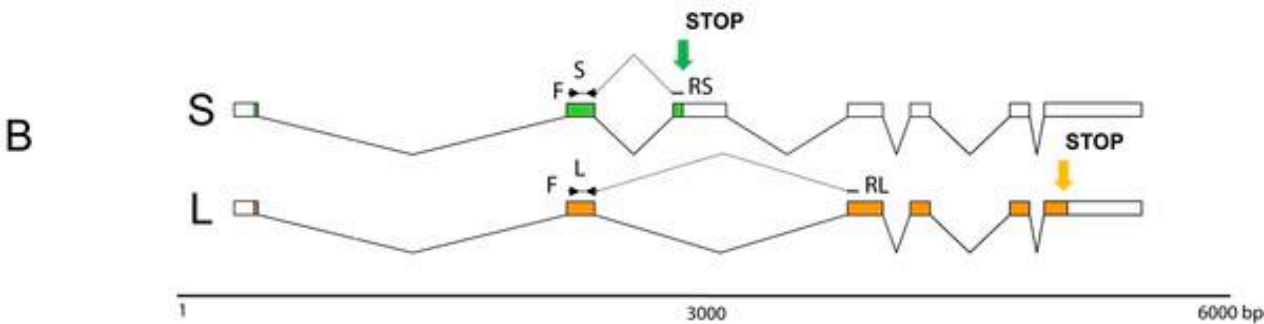
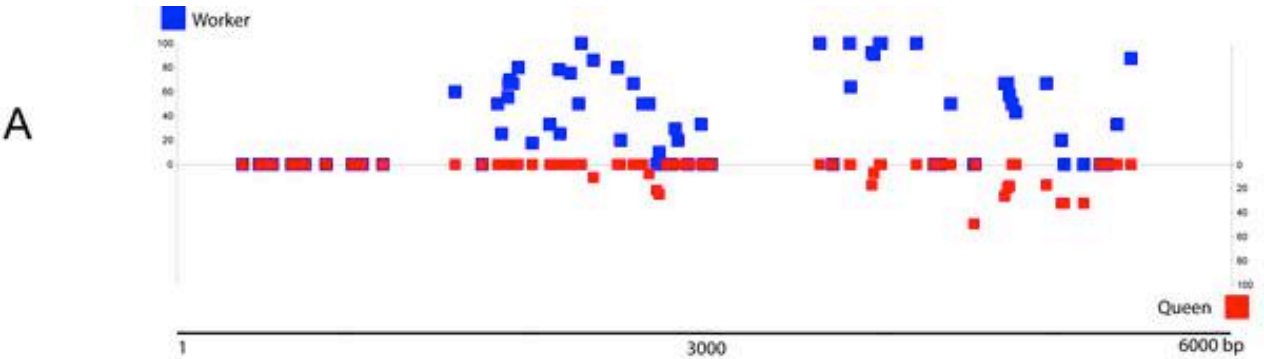
^aNearly all of the 3,019 CHH that were inferred to be methylated in worker brains on the basis of Solexa reads were found to be not methylated by an additional sequencing of selected amplicons using the 454 technology.

doi:10.1371/journal.pbio.1000506.t001



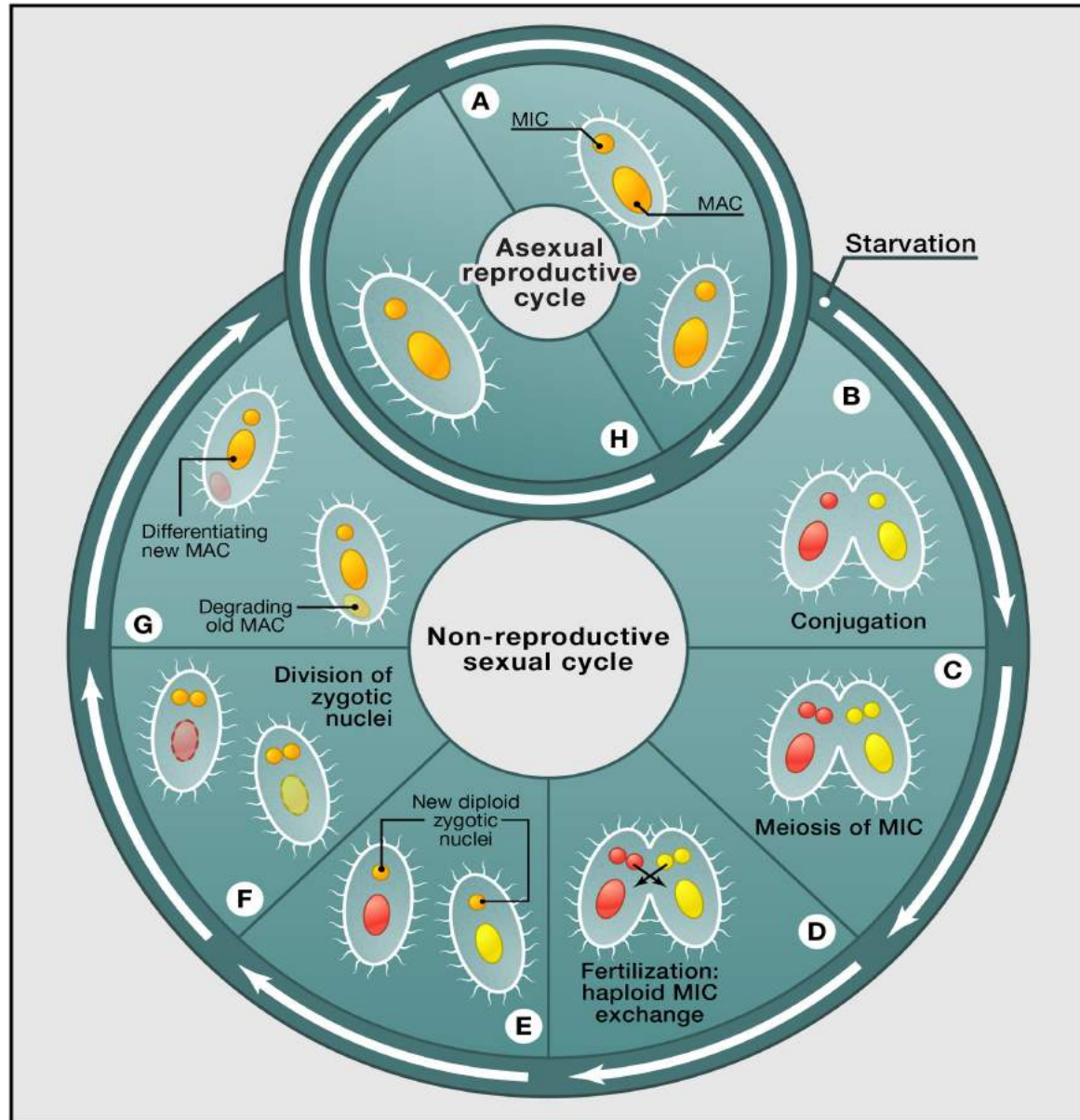
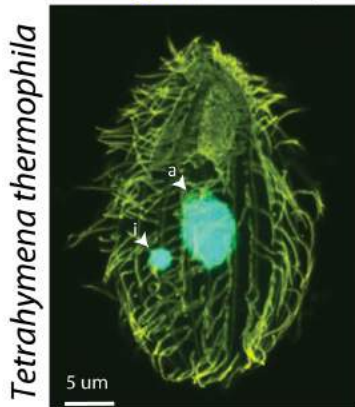
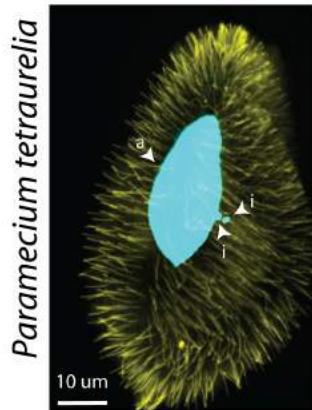
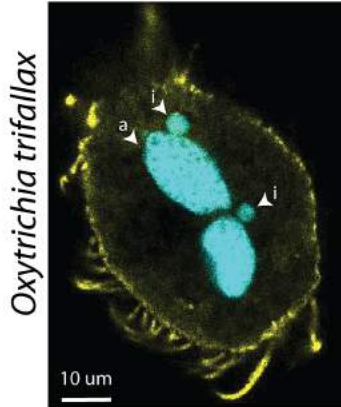
(Kucharski et al. (2008) *Science*; Lyko et al. (2010) *PLoS Biol*)

Methylation and cast-based societies



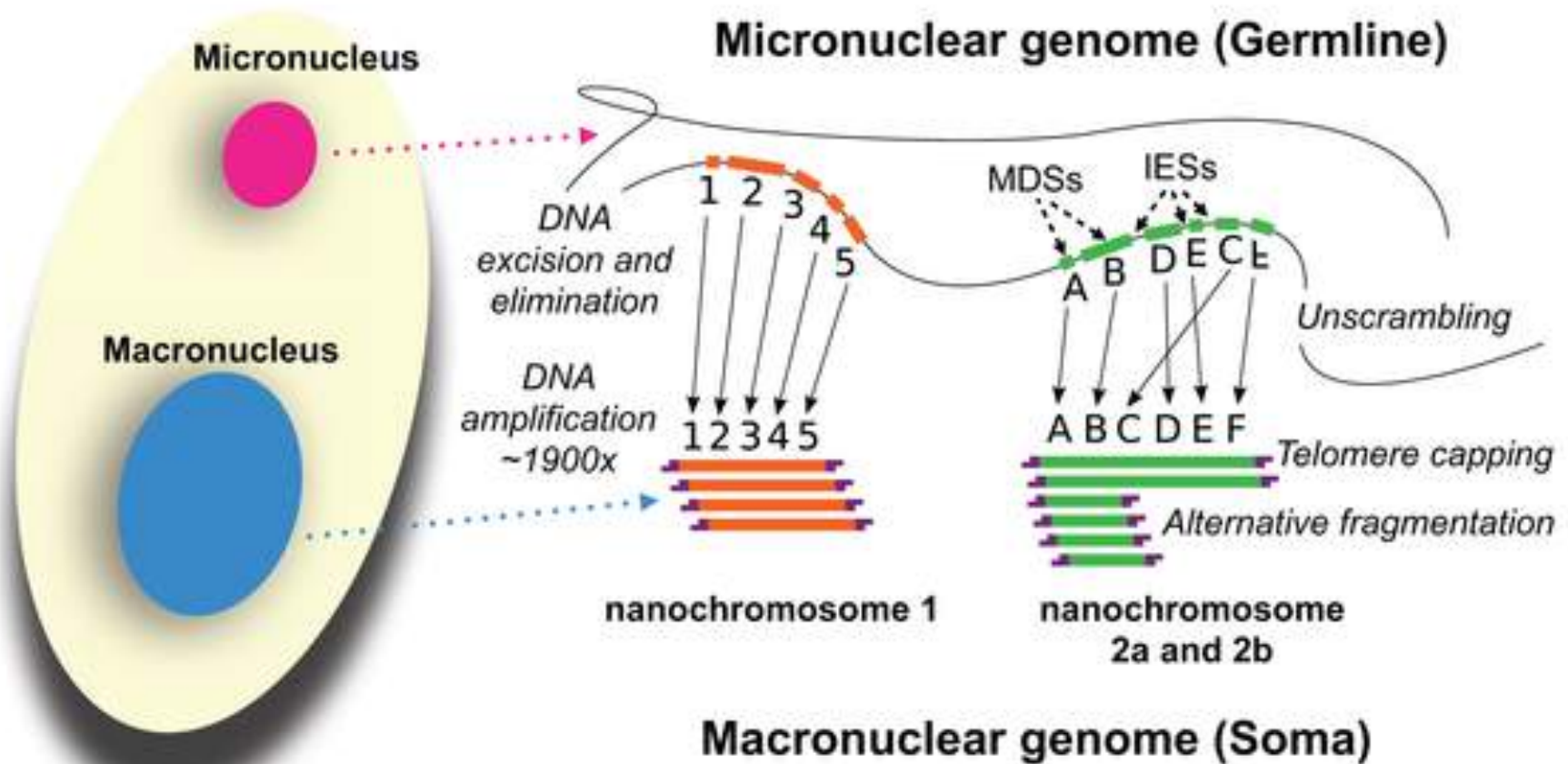
- *GB18602* – an example for methylation-based gene expression regulation: the long isoform is present in both queens and workers, but the short one only in queens

The weird genome of the ciliates



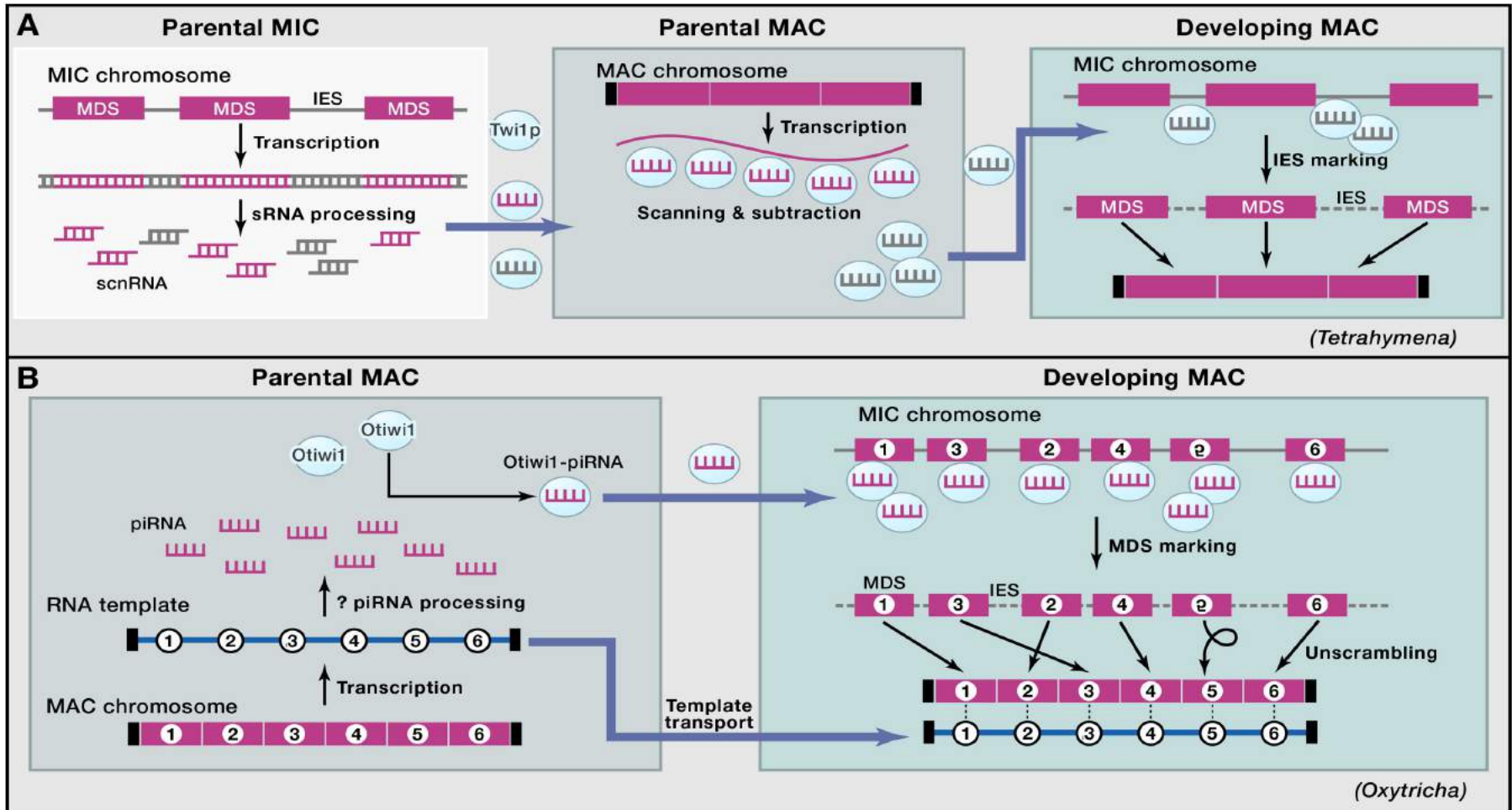
(Bracht et al. (2013) Cell)

Oxytrichia: an example for extreme genom-rearrangement



- the macronuclear (MAC) genome is formed of 16 000 intron-less minichromosomes, with high ploidity (1900n)

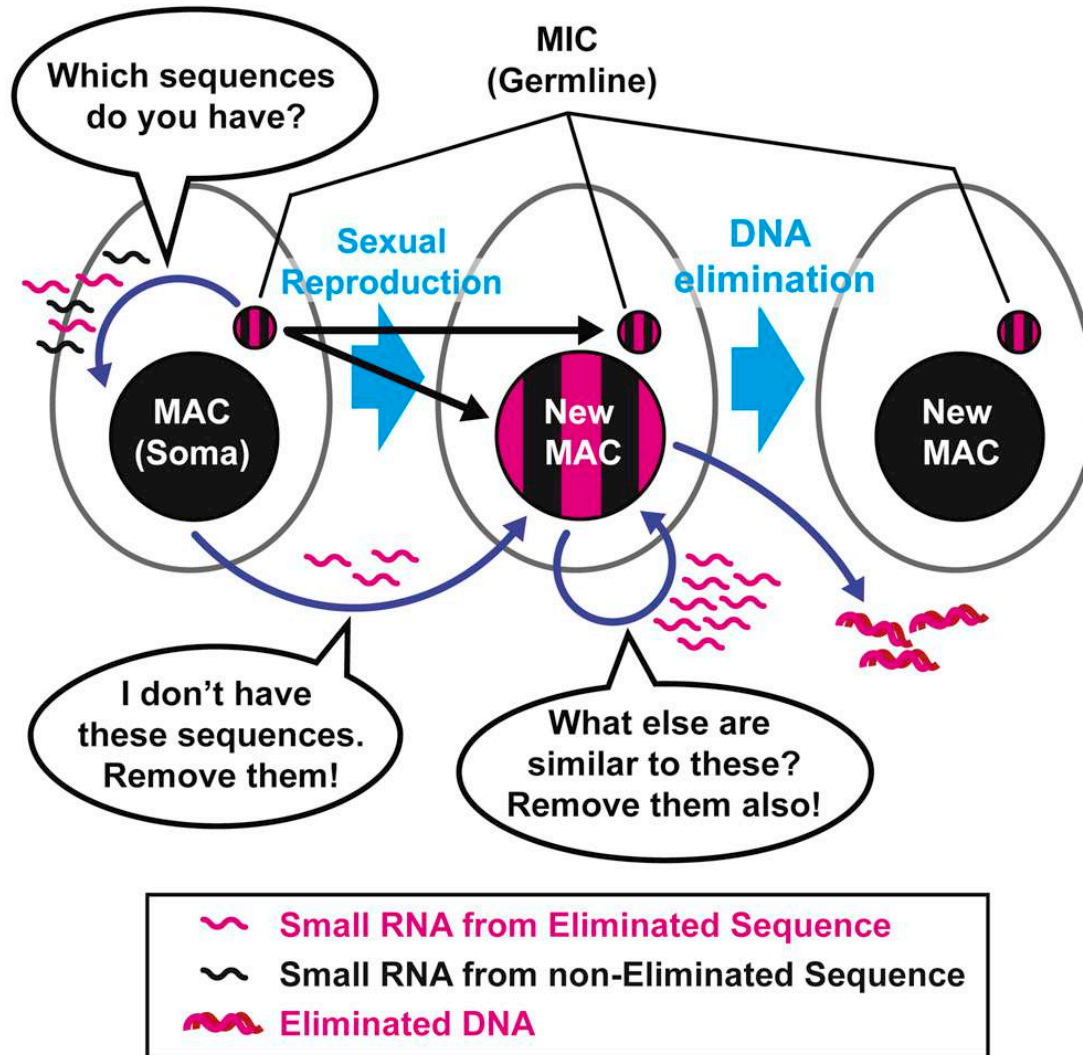
RNA mediated genome rearrangements in ciliates (two models)



- IES sequences originate from transposons

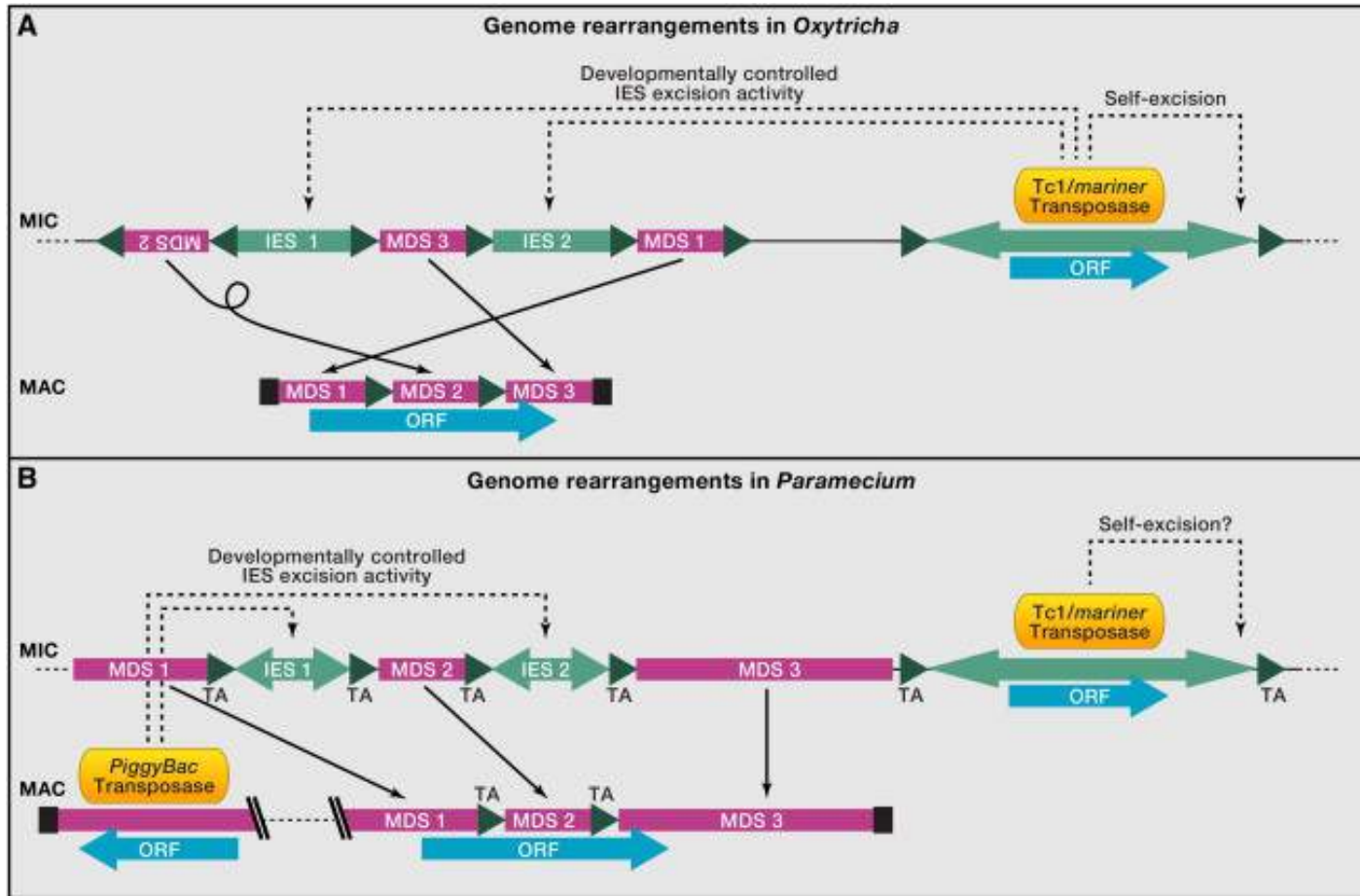
(Bracht et al. (2013) *Cell*)

RNA mediated genome rearrangements in ciliates



A model for MAC DNA elimination in *Tetrahymena*.

The role of domesticated transposons in ciliate genome rearrangements



- In *Oxytricha* several thousand Tc1/mariner transposons are present
- In *Paramecium* a domesticated PiggyBac transposon regulates editing