Bioinformatics and Evolutionary Genomics

Evolution of Genomes, Proteomes, Networks and Complexes

Berend Snel
Theoretical Biology and Bioinformatics
Department of Biology
Science Faculty
Utrecht University

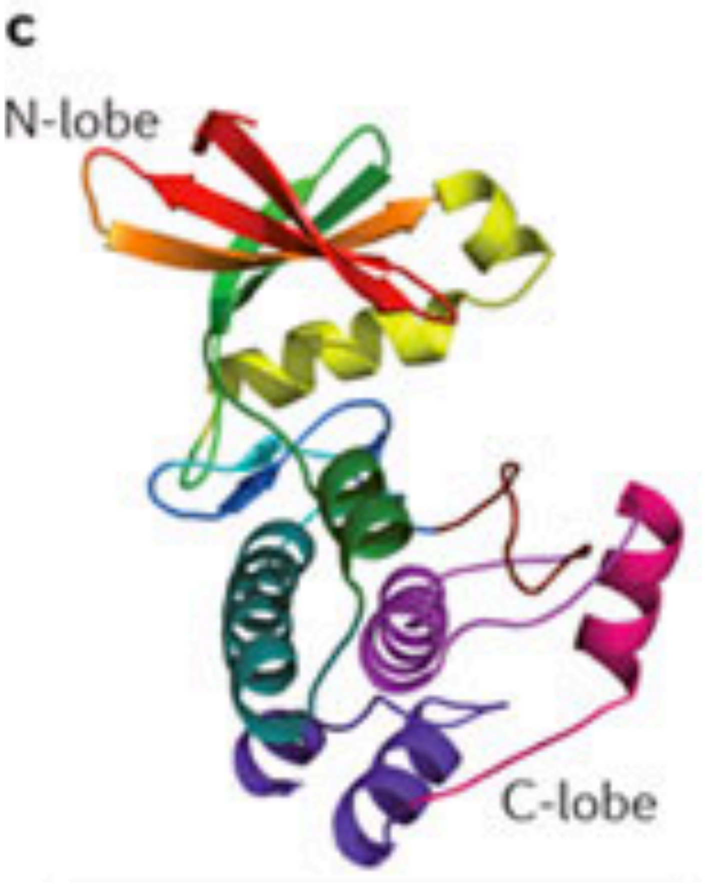
Today

- Introduction on general aims of the course
- Lecture on homology and domains (see how far we get …)
- Literature discussion on zmasek and godzik
- Some procedural stuff (maybe during computer exercises)
- Mini project (maybe during computer exercises)

Requests

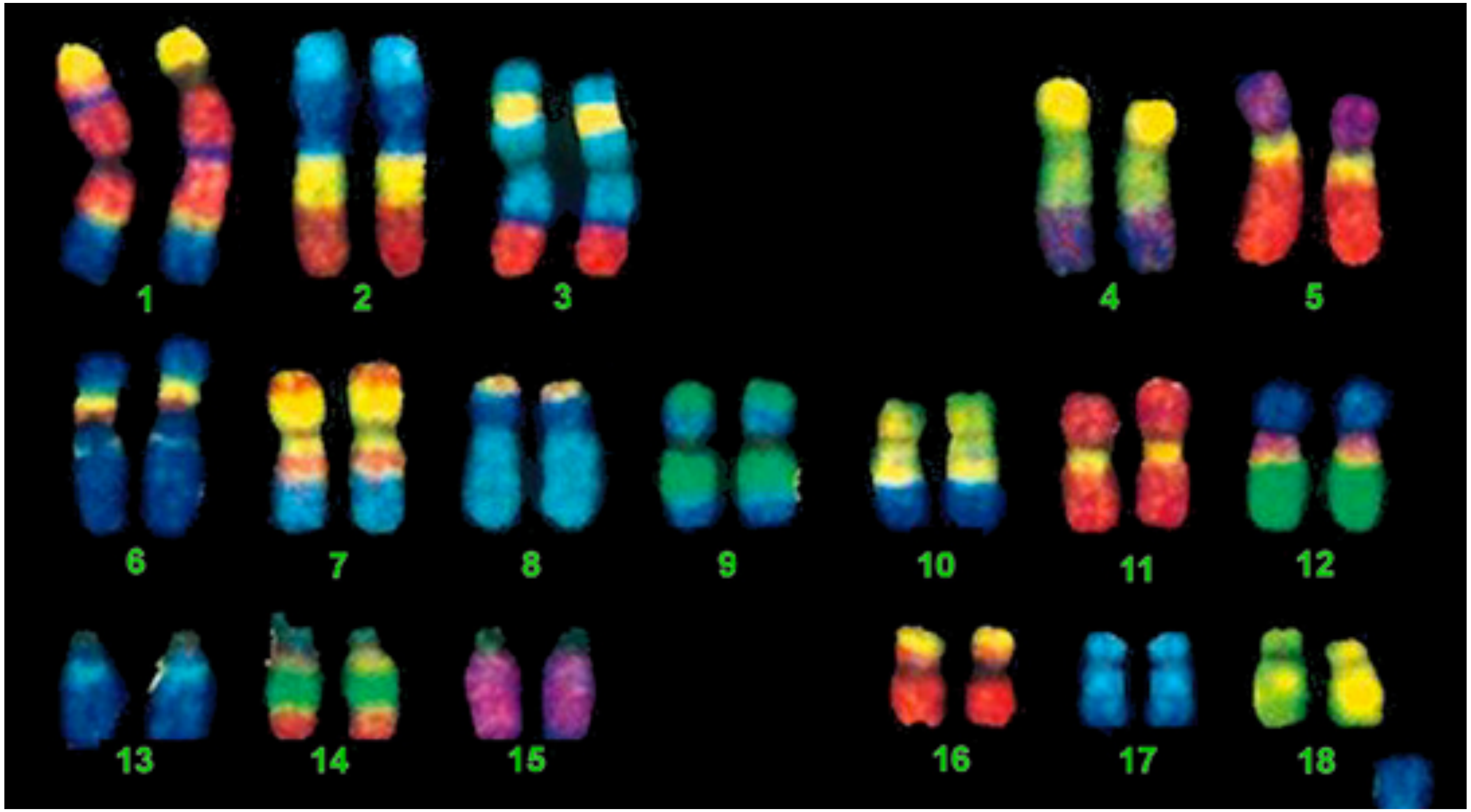
- **very** heterogeneous with respect to previous knowledge (IBMB, GB, research projects, PhD students)
- PLEASE: interrupt / ask questions when I am going to fast, when I use jargon, when I make jumps/conclusions that to me seem obvious 100% logical, but to your are erratic; please point out my implicit assumptions regarding what everybody knows
- -> Master course …

What is the evolutionary history of this protein? What happened in its evolution? Which other organisms have "**it**"? And when did it arise in evolution?

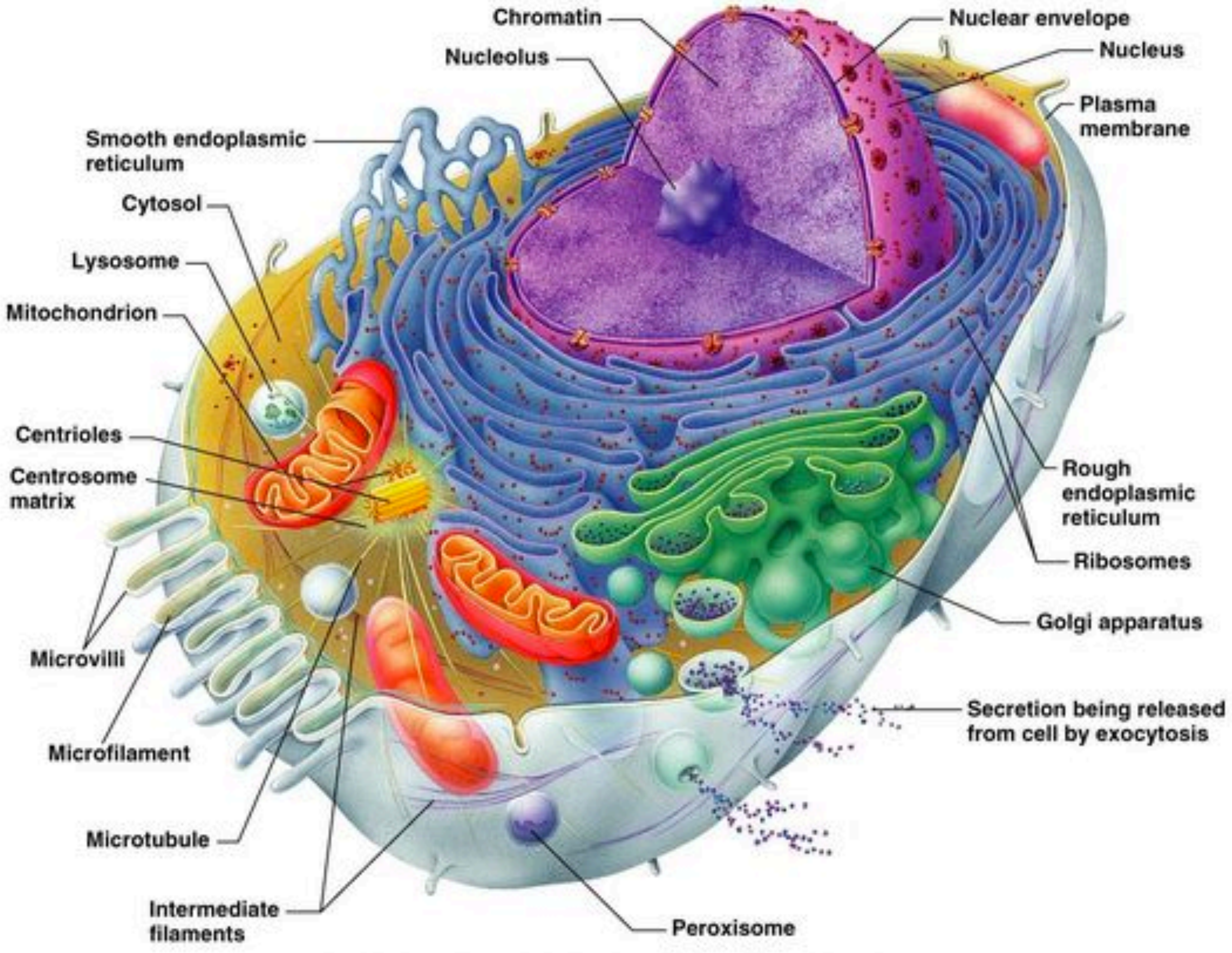


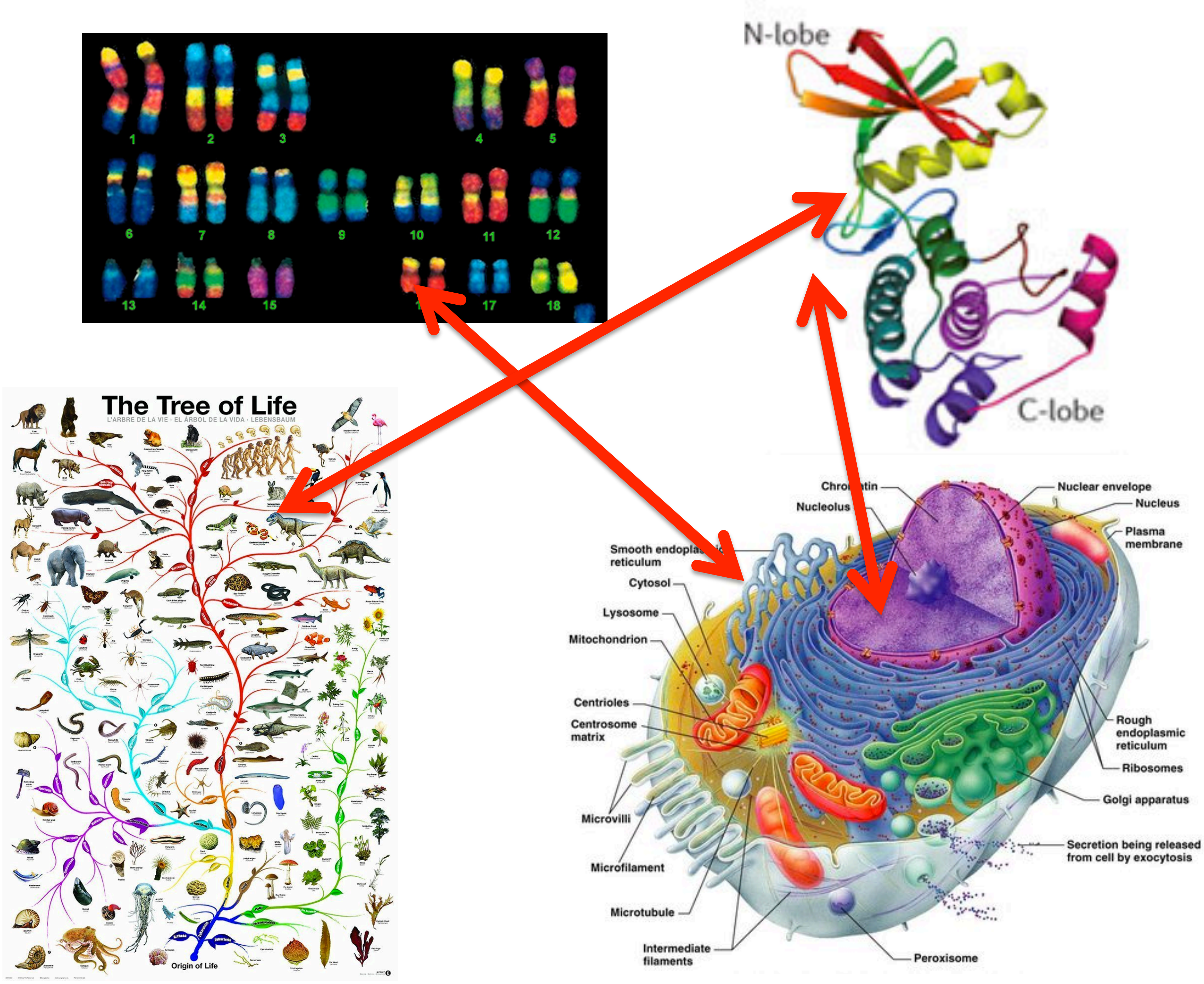
?

A bunch of genes, what is their evolutionary history?

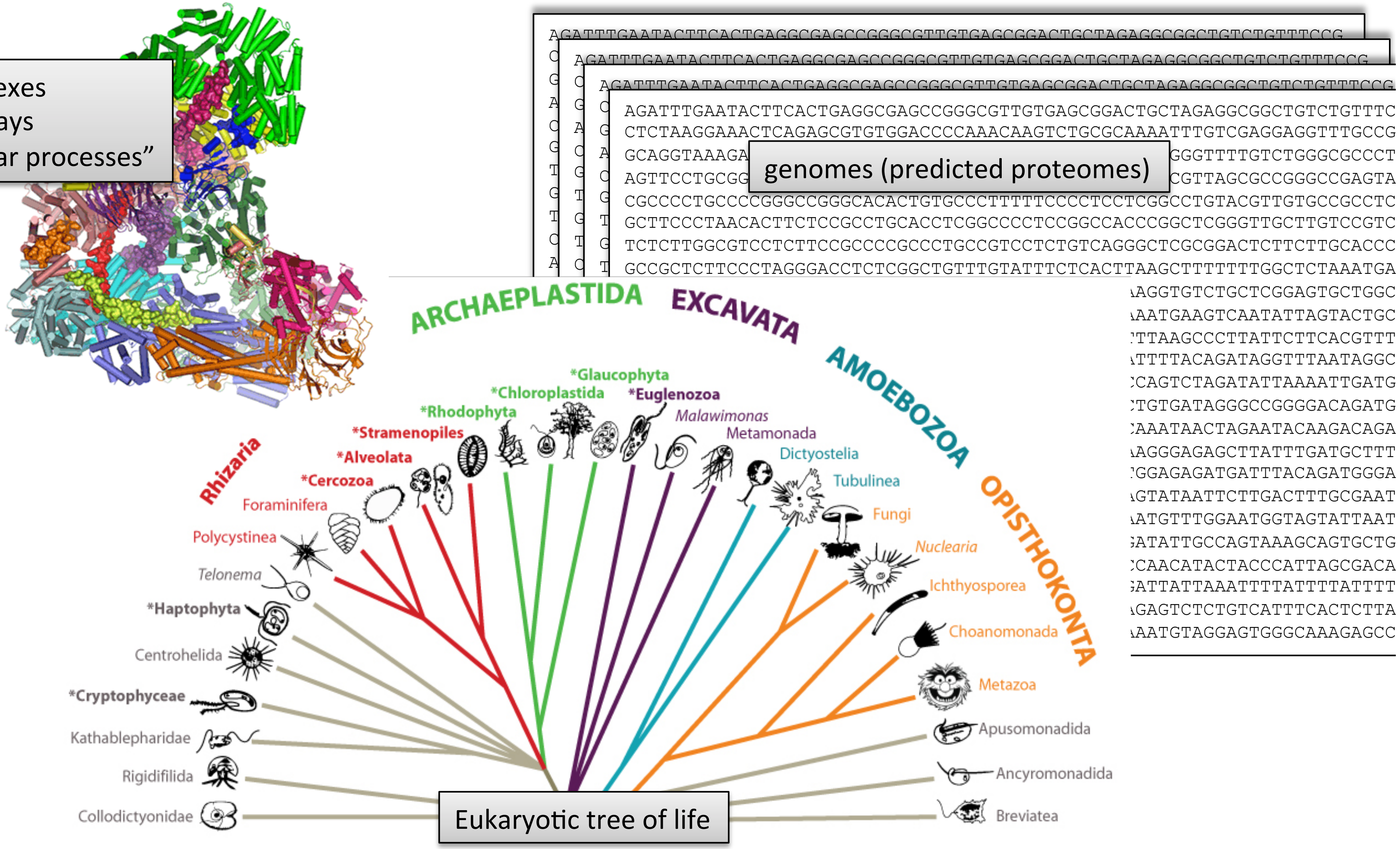


A complex cell what is the evolutionary history of this cell?




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
18
N-lobe
C-lobe
The Tree of Life
L'ARBRE DE LA VIE · EL ÁRBOL DE LA VIDA · LEBENSBAUM
Origin of Life
Chromatin
Nucleolus
Nuclear envelope
Nucleus
Plasma membrane
Smooth endoplasmic reticulum
Cytosol
Lysosome
Mitochondrion
Centrioles
Centrosome matrix
Microvilli
Microfilament
Microtubule
Intermediate filaments
Peroxisome
Rough endoplasmic reticulum
Ribosomes
Golgi apparatus
Secretion being released from cell by exocytosis

How do we do this? Find all kinds of patterns. Interpret these patterns.



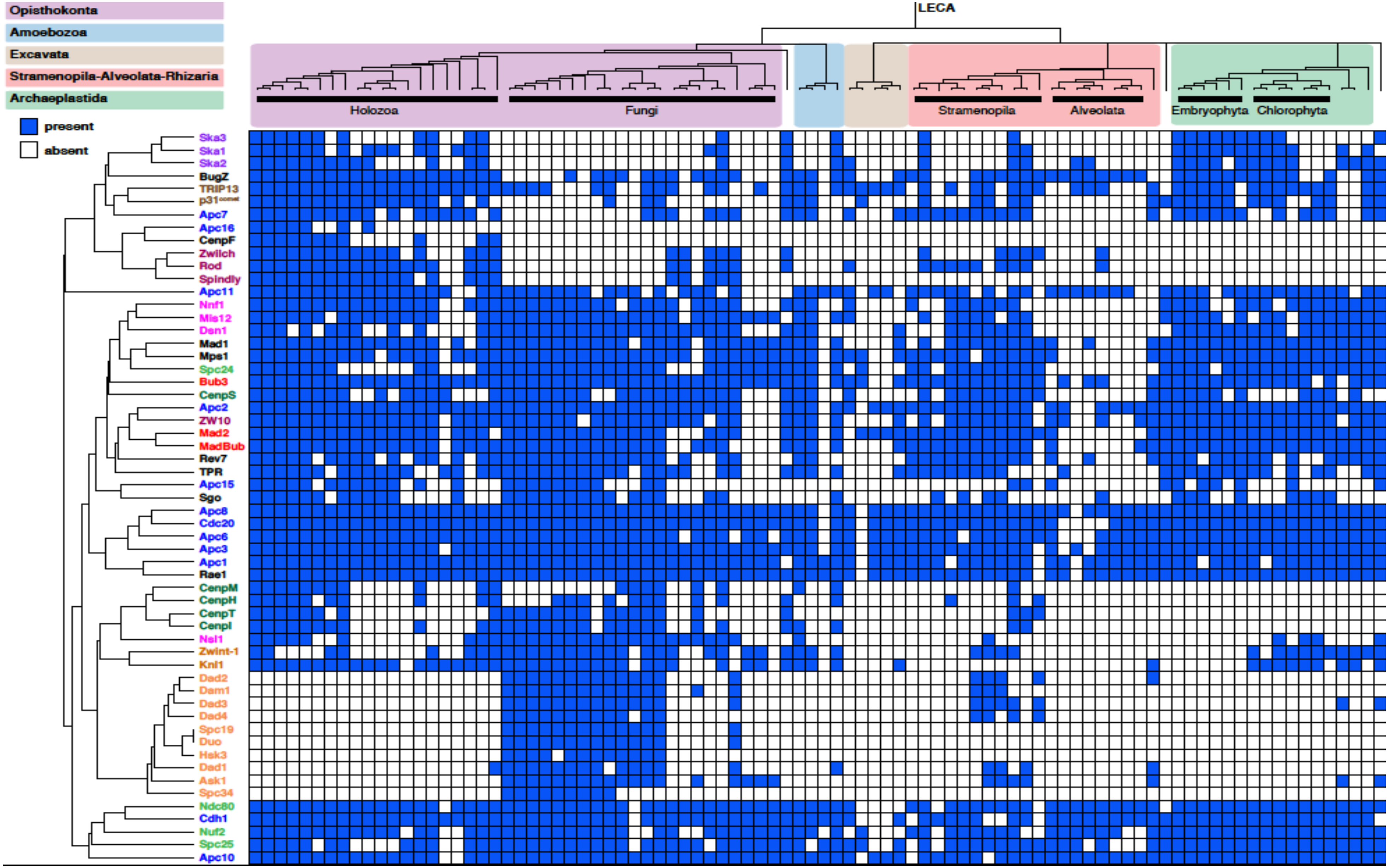
complexes
pathways
“cellular processes”

genomes (predicted proteomes)

```
AGATTTGAATACTTCACTGAGGCGAGCCGGGCGTTGTGAGCGGACTGCTAGAGGCGGCTGTCTGTTTC
CTCTAAGGAAACTCAGAGCGTGTGGACCCCAAACAAGTCTGCGCAAAATTTGTCGAGGAGGTTTGCCG
GCAGGTAAAGA                                         GGGTTTTGTCTGGGCGCCCT
AGTTCCTGCGG                                         CGTTAGCGCCGGGCCGAGTA
CGCCCCTGCCCCGGGCCGGGCACACTGTGCCCTTTTTCCCCTCCTCGGCCTGTACGTTGTGCCGCCTC
GCTTCCCTAACACTTCTCCGCCTGCACCTCGGCCCCTCCGGCCACCCGGCTCGGGTTGCTTGTCCGTC
TCTCTTGGCGTCCTCTTCCGCCCCGCCCTGCCGTCCTCTGTCAGGGCTCGCGGACTCTTCTTGCACCC
GCCGCTCTTCCCTAGGGACCTCTCGGCTGTTTGTATTTCTCACTTAAGCTTTTTTGGCTCTAAATGA
                                              AGGTGTCTGCTCGGAGTGCTGGC
                                              AATGAAGTCAATATTAGTACTGC
                                              TTAAGCCCTTATTCTTCACGTTT
                                              ATTTTACAGATAGGTTTAATAGGC
                                              CAGTCTAGATATTAAAATTGATG
                                              TGTGATAGGGCCGGGGACAGATG
                                              CAAATAACTAGAATACAAGACAGA
                                              AGGGAGAGCTTATTTGATGCTTT
                                              GGAGAGATGATTTACAGATGGGA
                                              GTATAATTCTTGACTTTGCGAAT
                                              ATGTTTGGAATGGTAGTATTAAT
                                              ATATTGCCAGTAAAGCAGTGCTG
                                              CAACATACTACCCATTAGCGACA
                                              ATTATTAAATTTTATTTTATTTT
                                              GAGTCTCTGTCATTTCACTCTTA
                                              AATGTAGGAGTGGGCAAAGAGCC
```

Eukaryotic tree of life

Complex machine in LECA and recurrent loss



Van Hooff *et al.* EMBO reports 2017

The basic proteome repertoire shaping evolutionary operators in genome evolution (the events that together compose the evolutionary history of a gene & create the patterns we see in the data)

- Classical: sequence evolution
  - (nearly)neutral
  - change in function

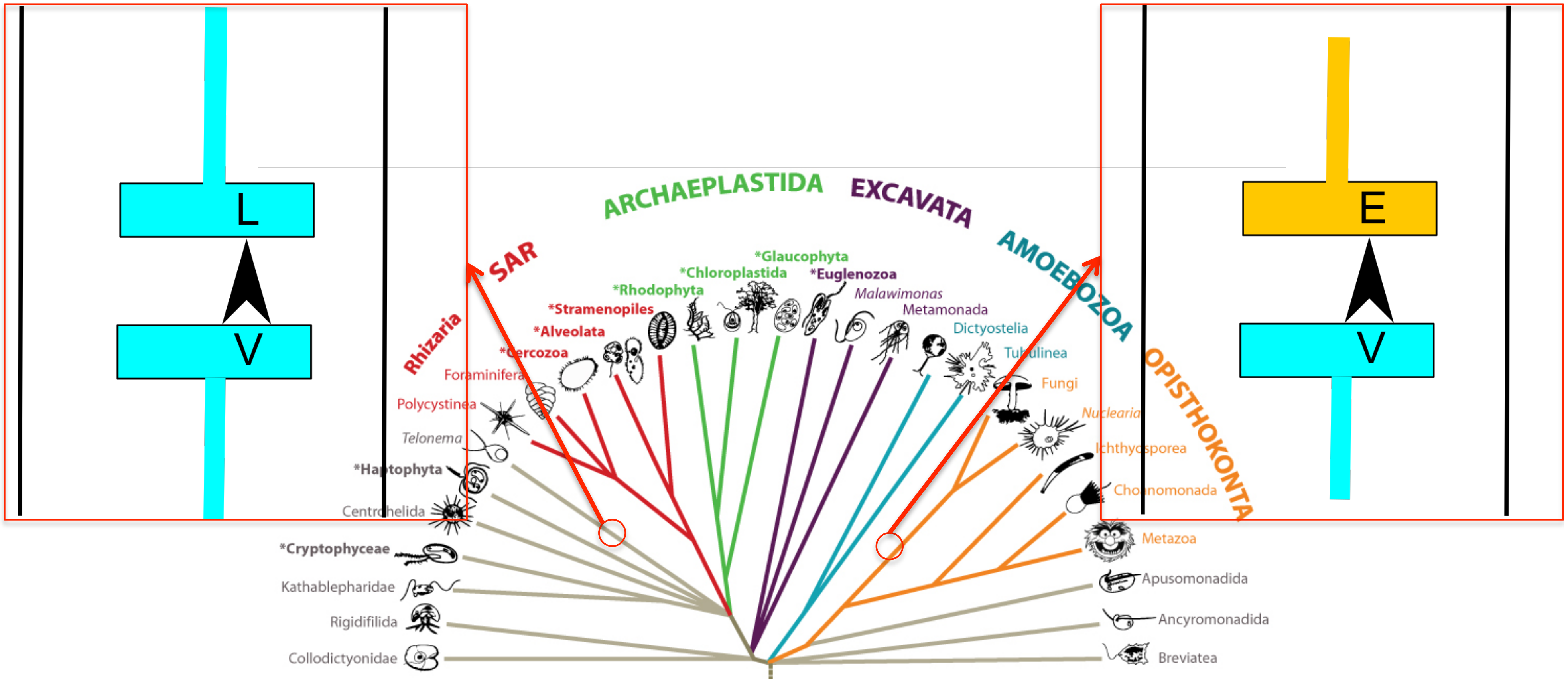
- Fusion
- (Fission)

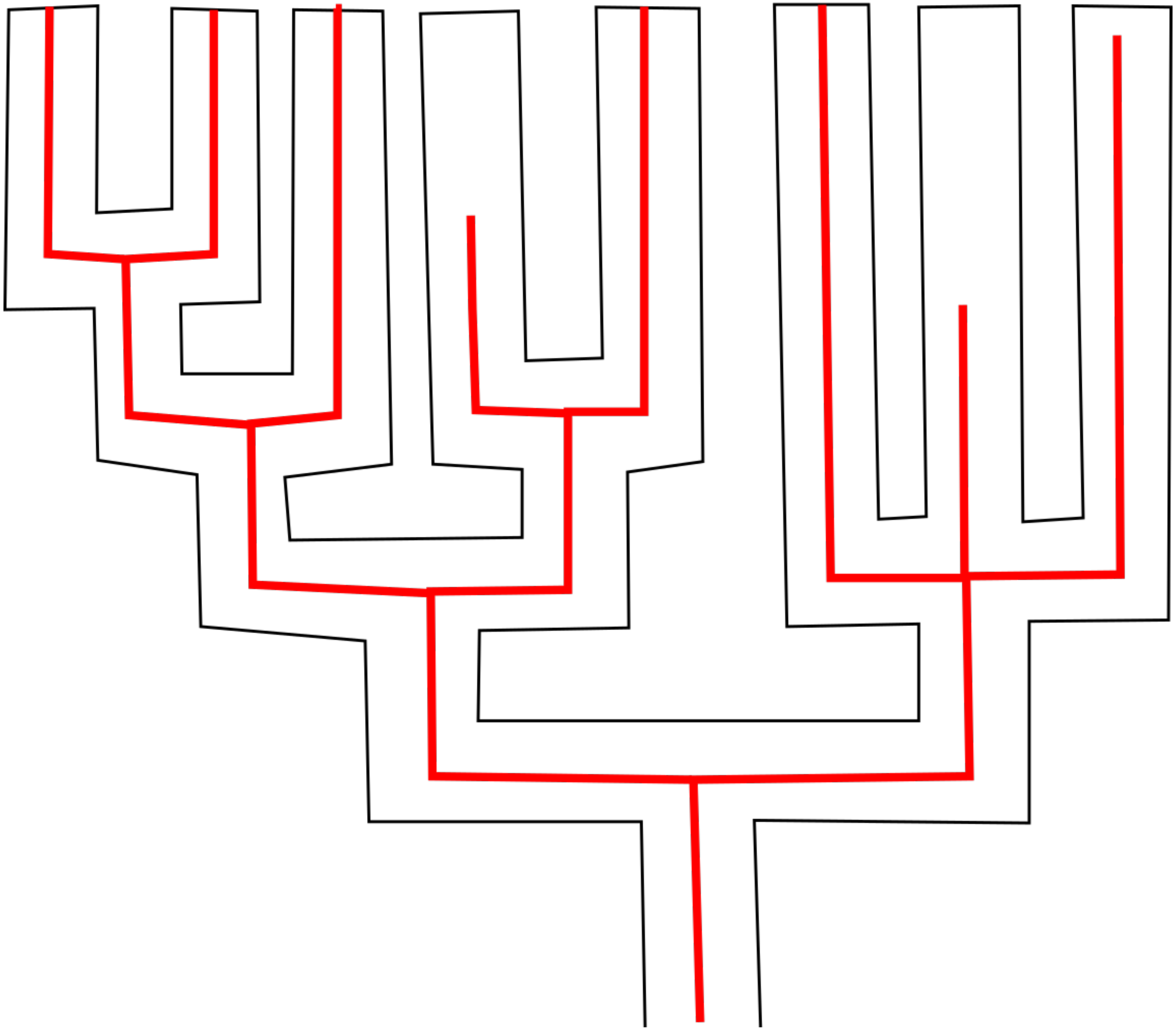
- Genome evolution:
  - Duplication
  - Loss (deletion)
  - Origin (invention) (!)
  - Horizontal Gene Transfer

Proteome evolution

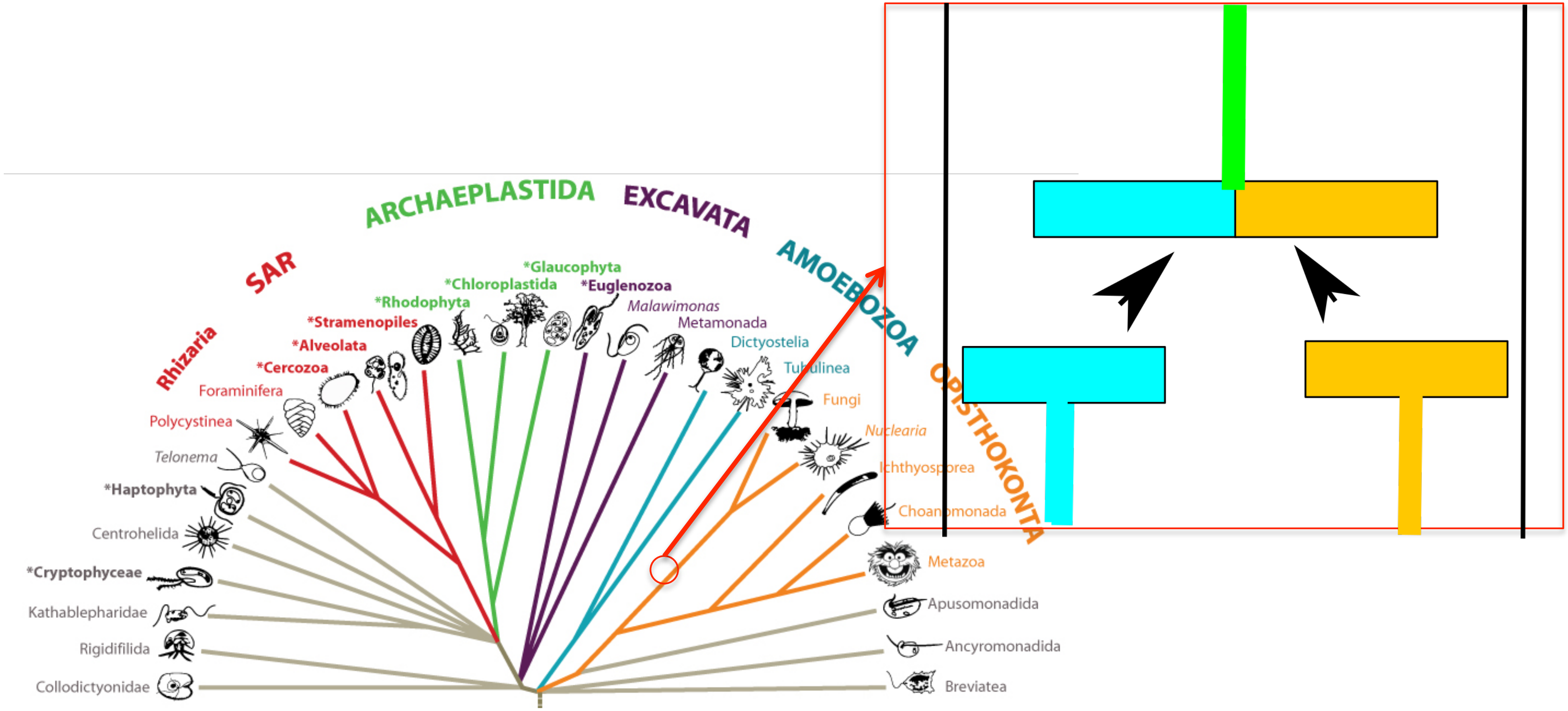
(nearly) neutral substitution

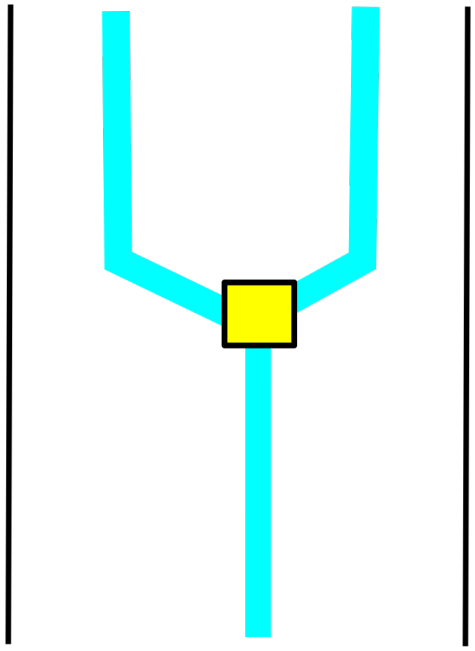
change of (partial) function substitution: loss or gain (adaptive)



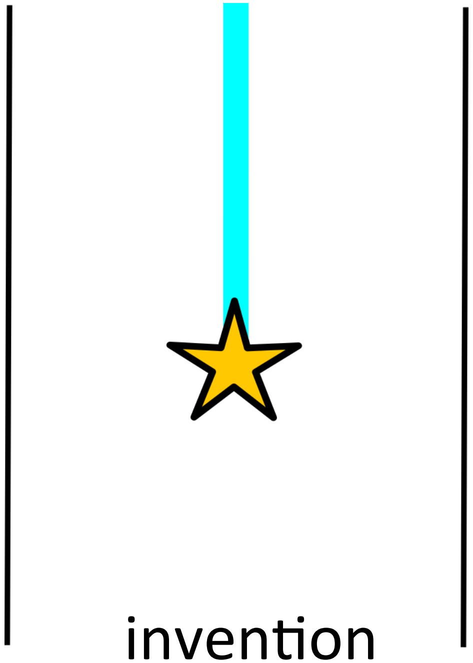


Gene fusion

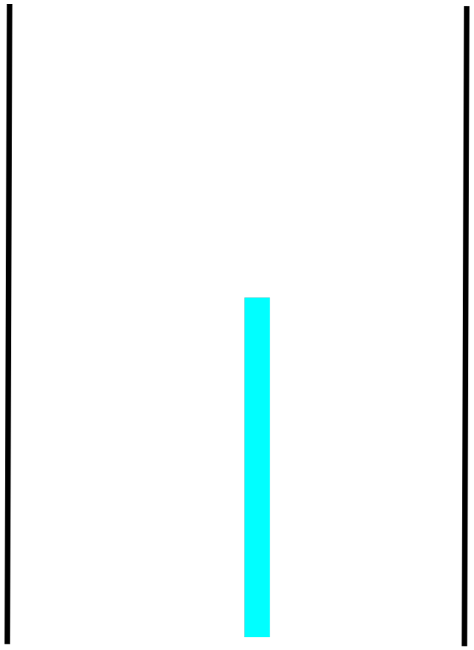

SAR
ARCHAEPLASTIDA
EXCAVATA
AMOEBOZOA
OPISTHOKONTA
Rhizaria
*Stramenopiles
*Alveolata
*Cercozoa
Foraminifera
Polycystinea
*Rhodophyta
*Chloroplastida
*Glaucophyta
*Euglenozoa
Malawimonas
Metamonada
Dictyostelia
Tubulinea
Fungi
Nuclearia
Ichthyosporea
Choanomonada
Metazoa
Apusomonadida
Ancyromonadida
Breviatea
Telonema
*Haptophyta
Centrohelida
*Cryptophyceae
Kathablepharidae
Rigidifilida
Collodictyonidae



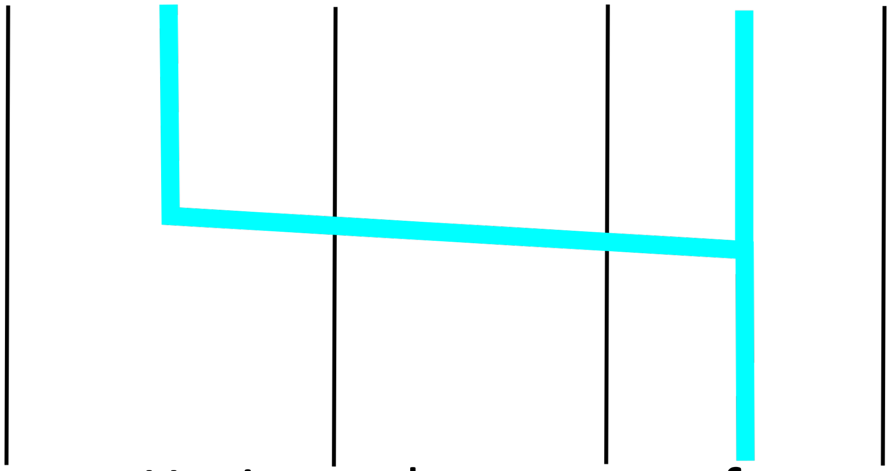
duplication



invention

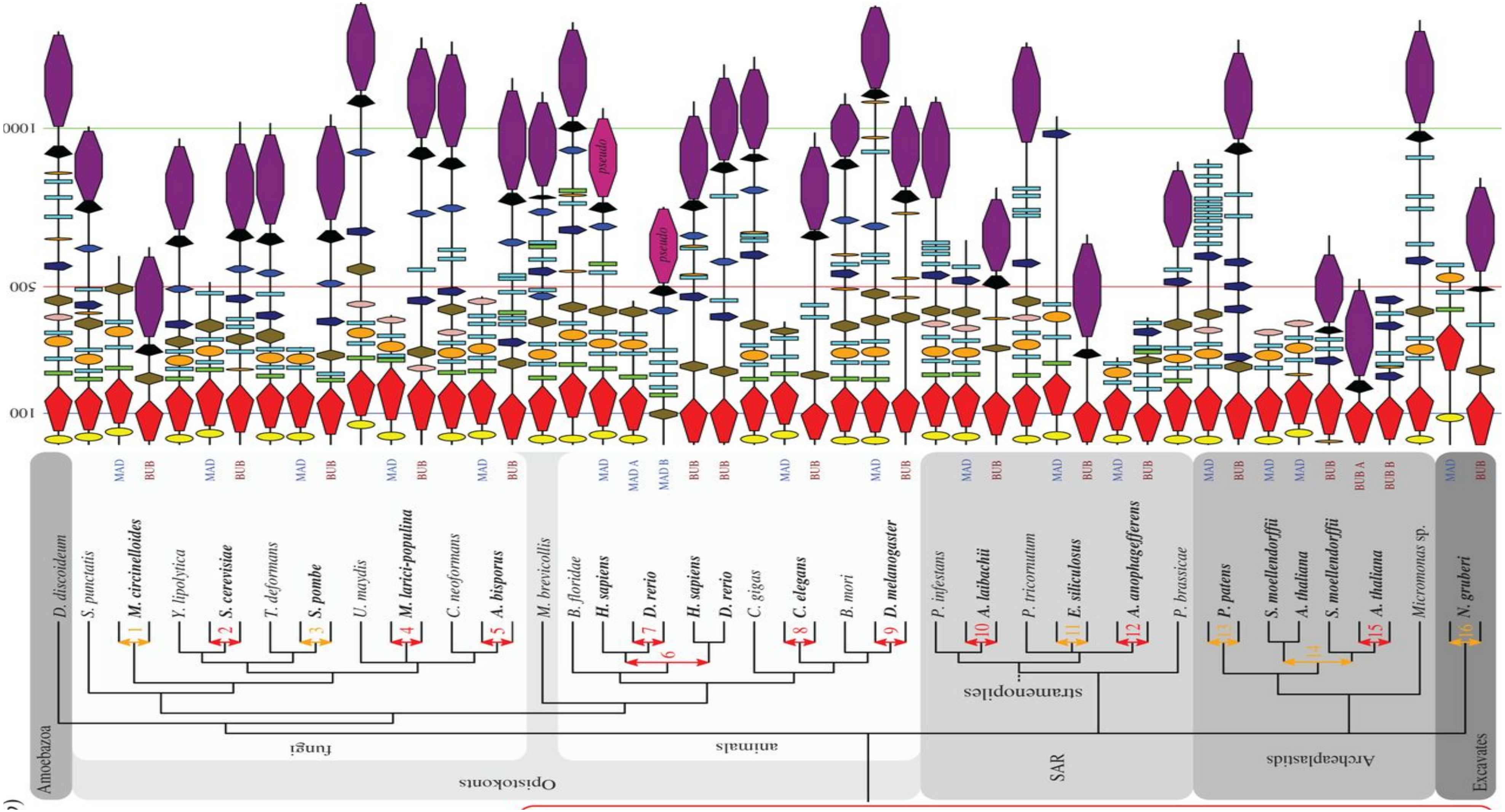


Loss / deletion



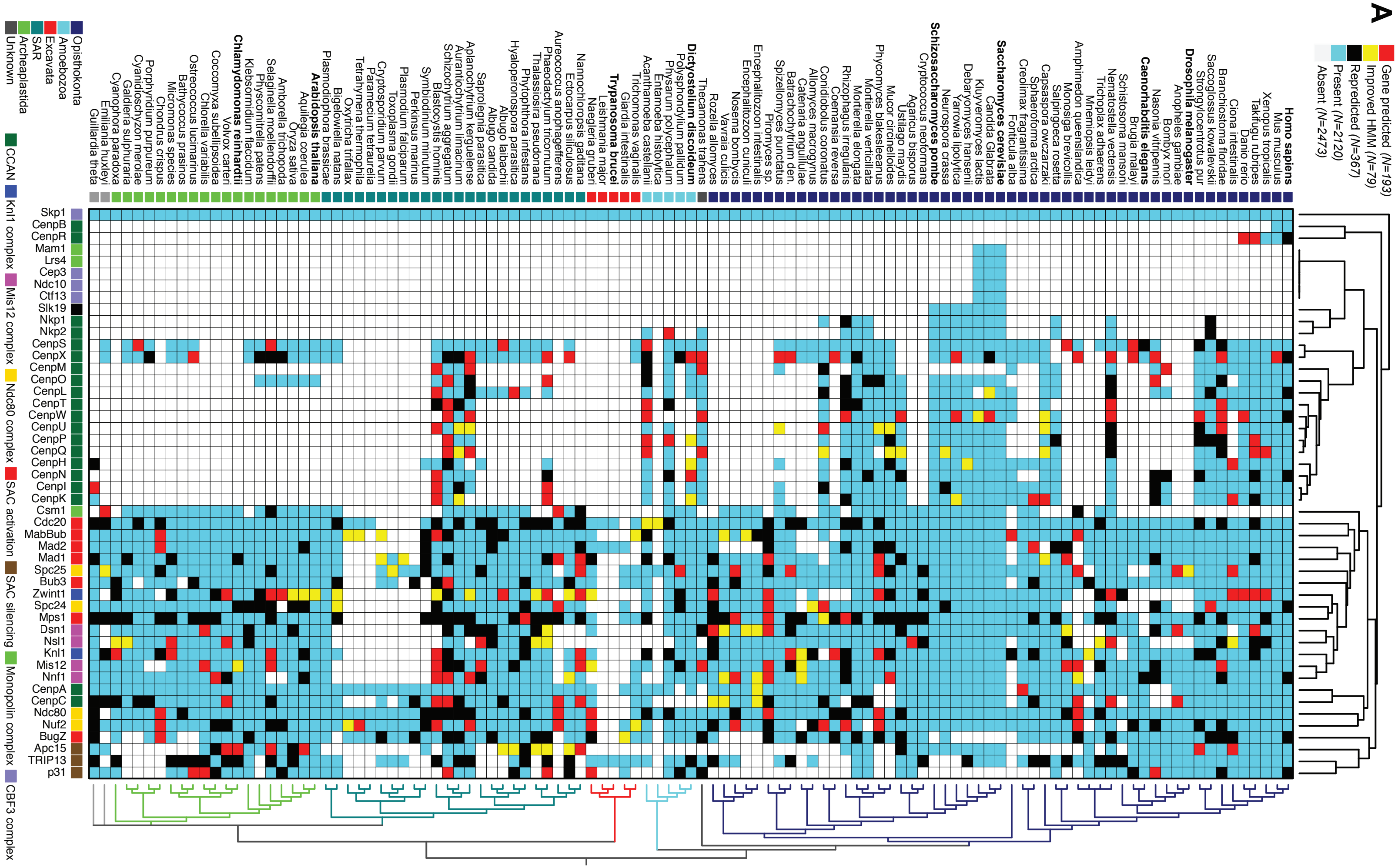
Horizontal gene transfer

These operators cooperate with each other to make genomes and life complicated: recurrent **duplication** + loss of function **substitutions**

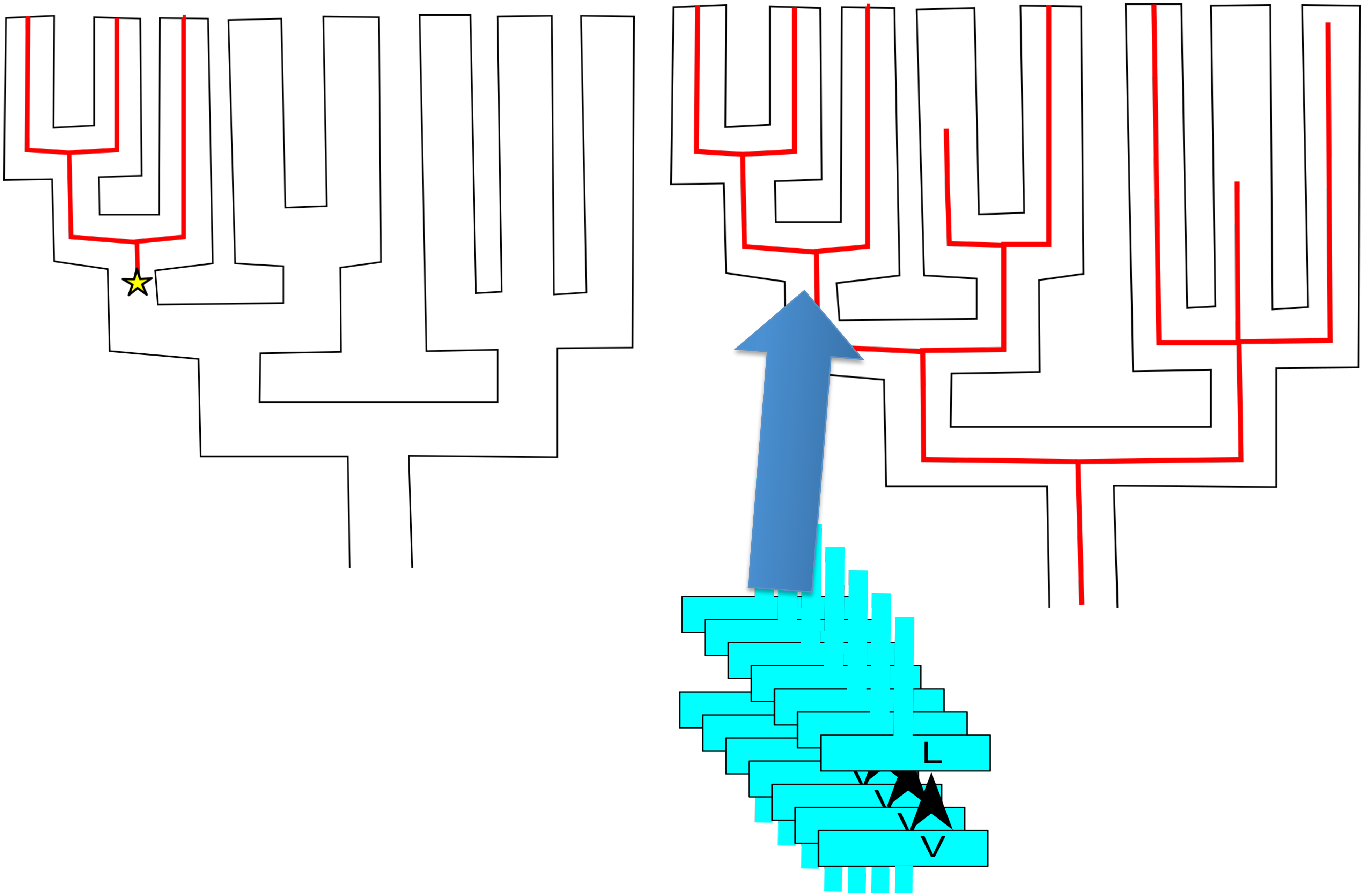


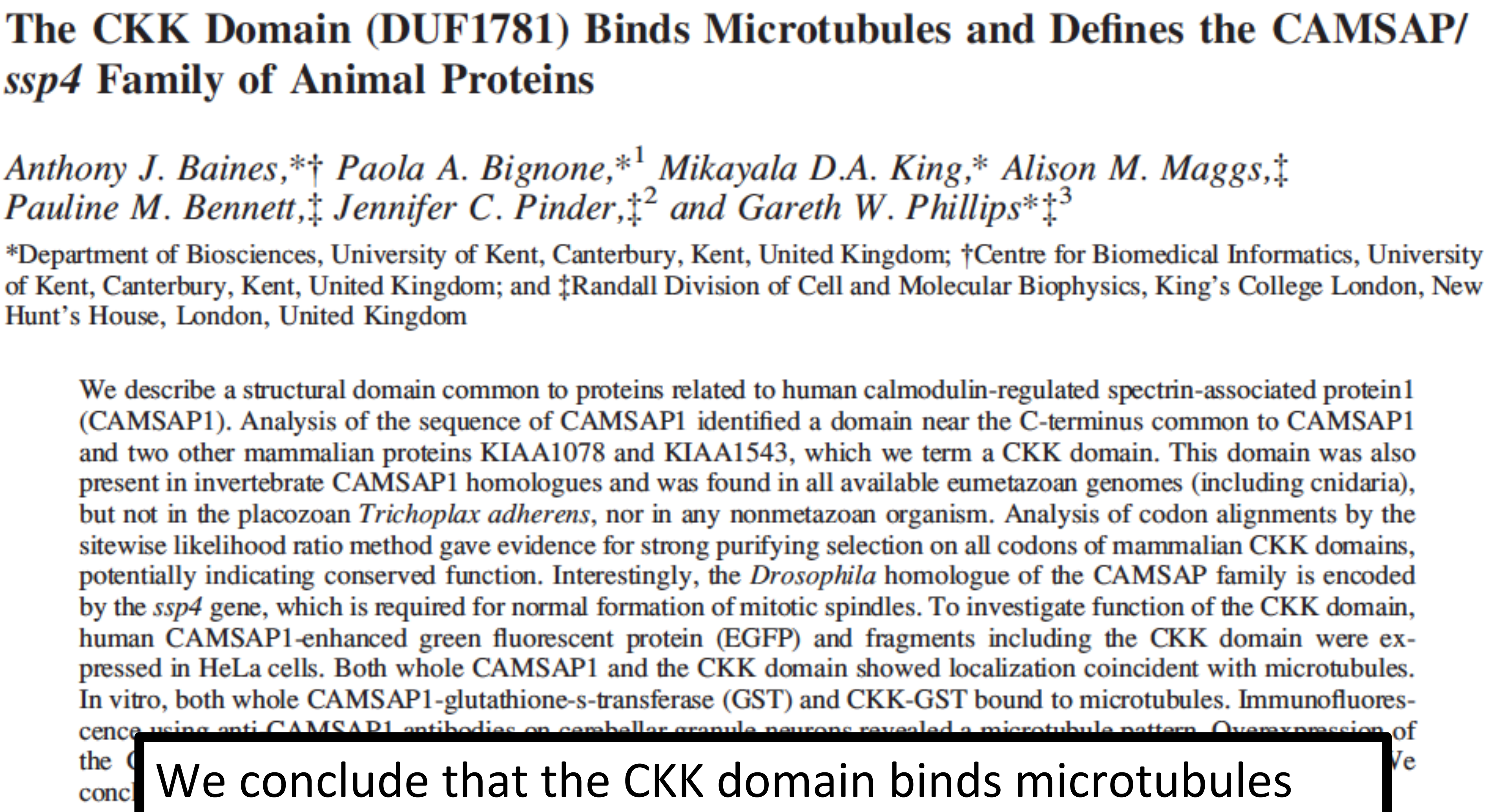
And these events conspire with a host of technical problems to make them difficult to detect

- E.g. gene loss or a problem with our predicted proteome



e.g. invention or neutral evolution at speed beyond ability of blast to detect homologs?





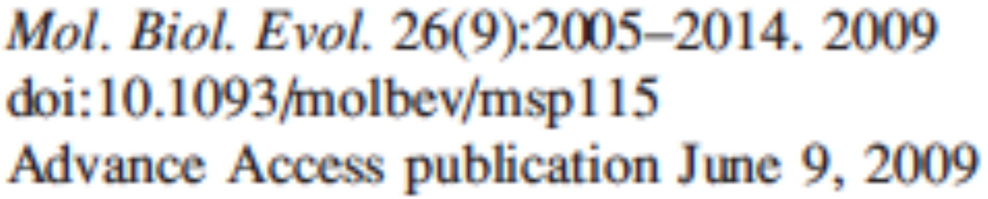
The CKK Domain (DUF1781) Binds Microtubules and Defines the CAMSAP/ *ssp4* Family of Animal Proteins

Anthony J. Baines,† Paola A. Bignone,*[1] Mikayala D.A. King,* Alison M. Maggs,‡ Pauline M. Bennett,‡ Jennifer C. Pinder,‡[2] and Gareth W. Phillips*‡[3]*

*Department of Biosciences, University of Kent, Canterbury, Kent, United Kingdom; †Centre for Biomedical Informatics, University of Kent, Canterbury, Kent, United Kingdom; and ‡Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, London, United Kingdom

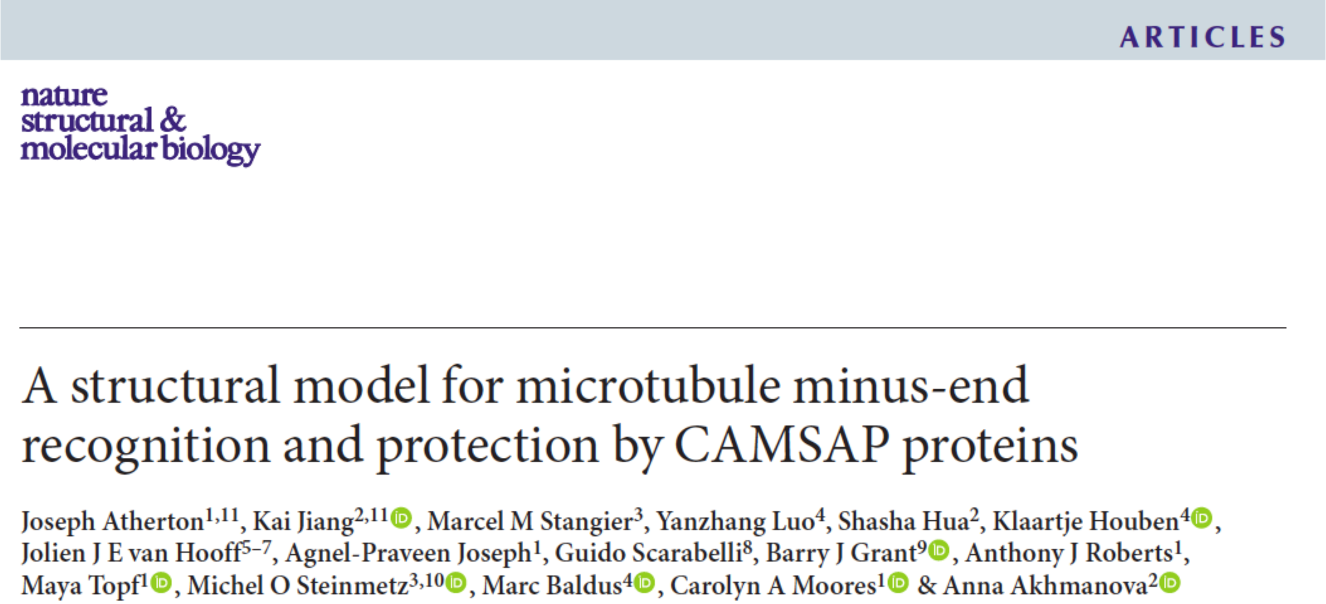
We describe a structural domain common to proteins related to human calmodulin-regulated spectrin-associated protein1 (CAMSAP1). Analysis of the sequence of CAMSAP1 identified a domain near the C-terminus common to CAMSAP1 and two other mammalian proteins KIAA1078 and KIAA1543, which we term a CKK domain. This domain was also present in invertebrate CAMSAP1 homologues and was found in all available eumetazoan genomes (including cnidaria), but not in the placozoan *Trichoplax adherens*, nor in any nonmetazoan organism. Analysis of codon alignments by the sitewise likelihood ratio method gave evidence for strong purifying selection on all codons of mammalian CKK domains, potentially indicating conserved function. Interestingly, the *Drosophila* homologue of the CAMSAP family is encoded by the *ssp4* gene, which is required for normal formation of mitotic spindles. To investigate function of the CKK domain, human CAMSAP1-enhanced green fluorescent protein (EGFP) and fragments including the CKK domain were expressed in HeLa cells. Both whole CAMSAP1 and the CKK domain showed localization coincident with microtubules. In vitro, both whole CAMSAP1-glutathione-s-transferase (GST) and CKK-GST bound to microtubules. Immunofluorescence using anti-CAMSAP1 antibodies on cerebellar granule neurons revealed a microtubule pattern. Overexpression of the C … We conc…

We conclude that the CKK domain binds microtubules and represents a domain that evolved with the metazoa.



Mol. Biol. Evol. 26(9):2005–2014. 2009
doi:10.1093/molbev/msp115
Advance Access publication June 9, 2009

But:

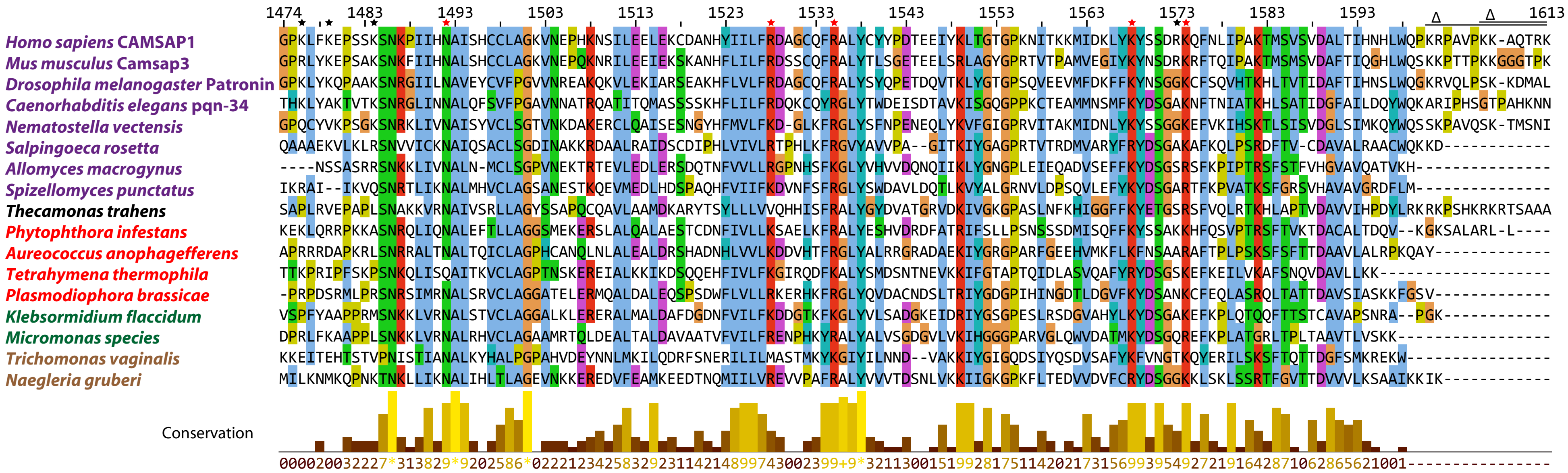


ARTICLES

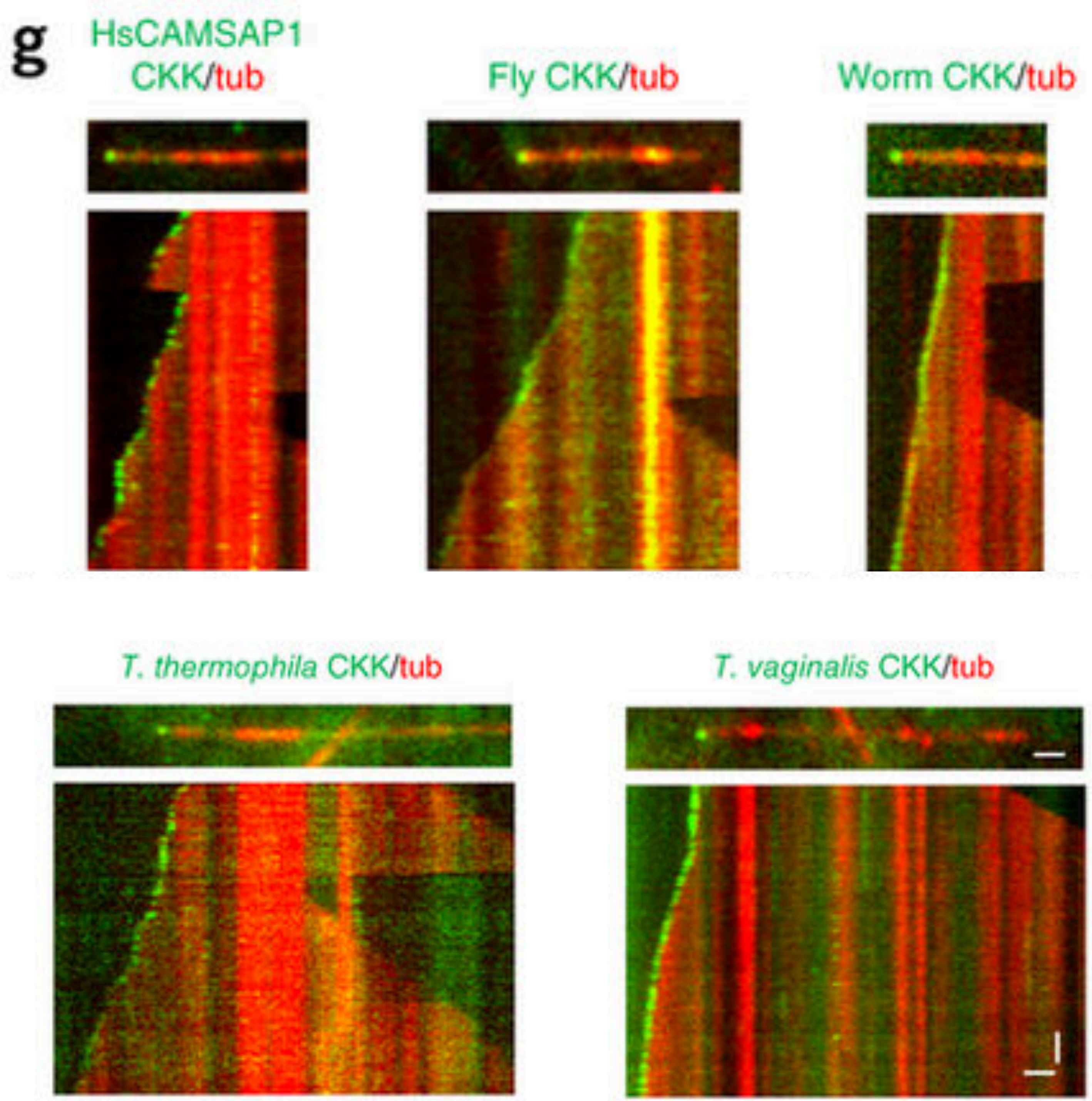
nature structural & molecular biology

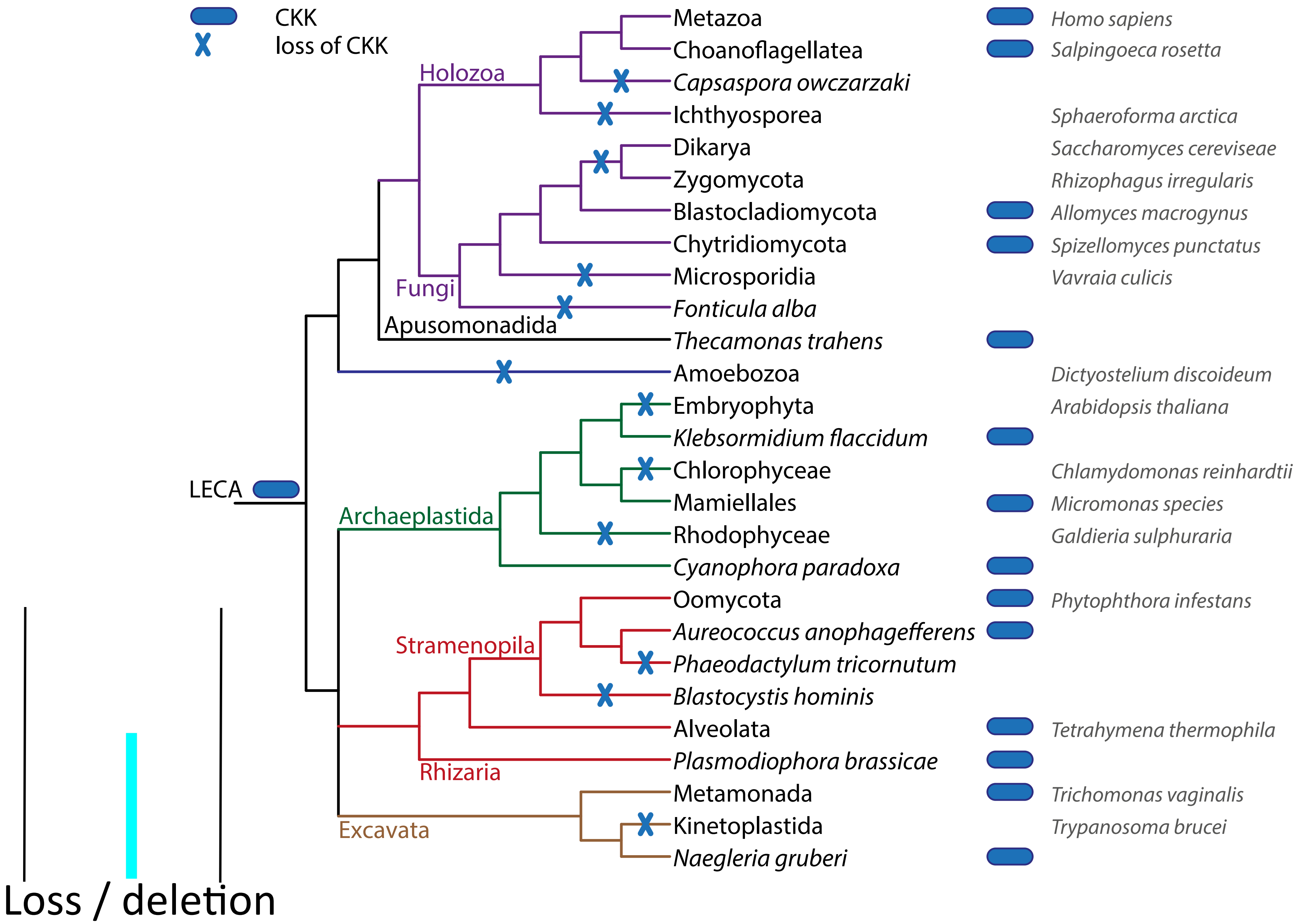
A structural model for microtubule minus-end recognition and protection by CAMSAP proteins

Joseph Atherton[1,11], Kai Jiang[2,11], Marcel M Stangier[3], Yanzhang Luo[4], Shasha Hua[2], Klaartje Houben[4], Jolien J E van Hooff[5–7], Agnel-Praveen Joseph[1], Guido Scarabelli[8], Barry J Grant[9], Anthony J Roberts[1], Maya Topf[1], Michel O Steinmetz[3,10], Marc Baldus[4], Carolyn A Moores[1] & Anna Akhmanova[2]

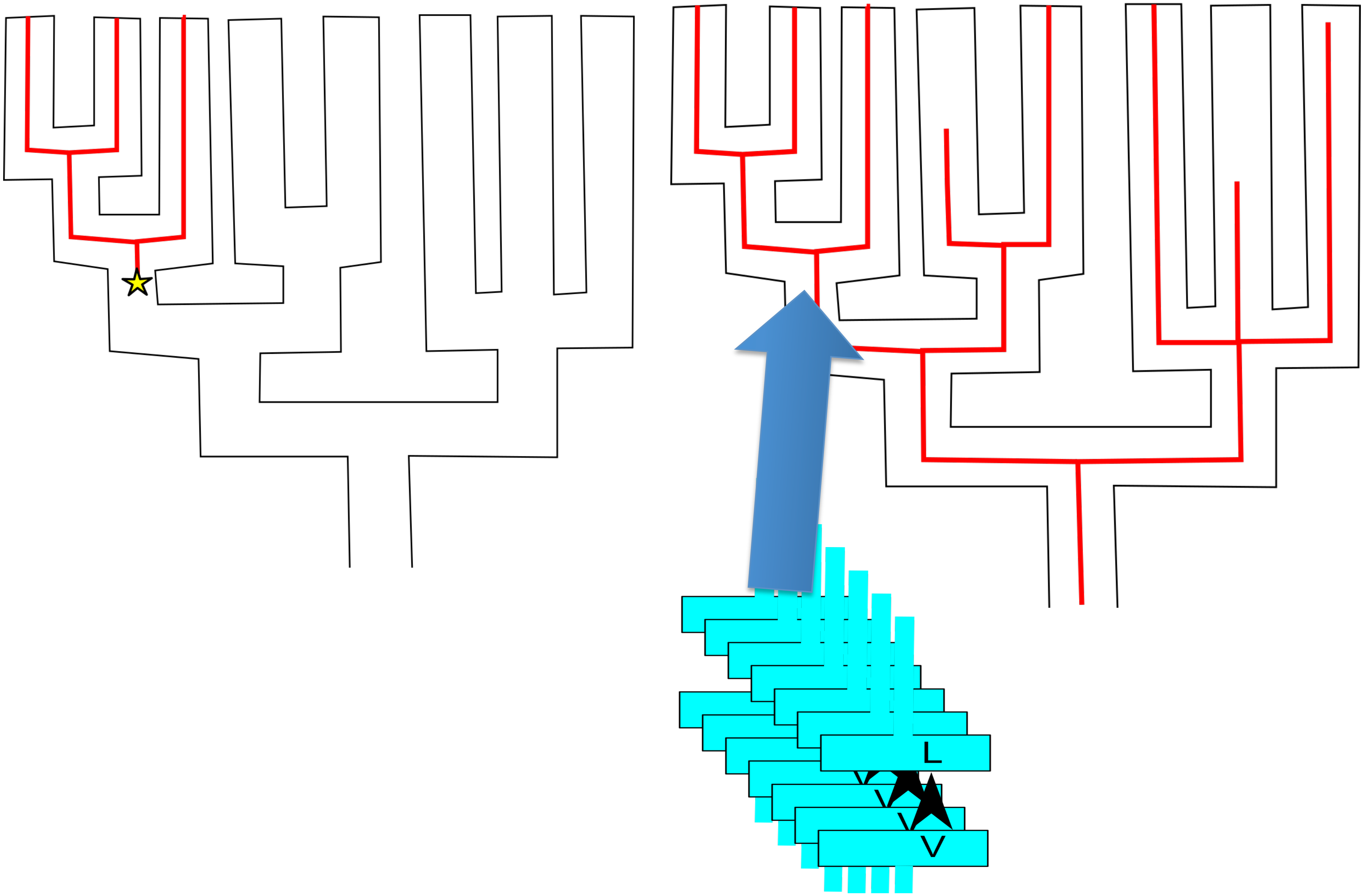


```
                                   1474     1483      1493      1503      1513      1523      1533      1543      1553      1563      1573      1583      1593      Δ     Δ   1613
Homo sapiens CAMSAP1               GPKLFKEPSSKSNKPIIHNAISHCCLAGKVNEPHKNSILEELEKCDANHYIILFRDAGCQFRALYCYYPDTEEIYKLTGTGPKNITKKMIDKLYKYSSDRKQFNLIPAKTMSVSVDALTIHNHLWQPKRPAVPKK-AQTRK
Mus musculus Camsap3               GPRLYKEPSAKSNKFIIHNALSHCCLAGKVNEPQKNRILEEIEKSKANHFLILFRDSSCQFRALYTLSGETEELSRLAGYGPRTVTPAMVEGIYKYNSDRKRFTQIPAKTMSMSVDAFTIQGHLWQSKKPTTPKKGGGTPK
Drosophila melanogaster Patronin   GPKLYKQPAAKSNRGIILNAVEYCVFPGVVNREAKQKVLEKIARSEAKHFLVLFRDAGCQFRALYSYQPETDQVTKLYGTGPSQVEEVMFDKFFKYNSGGKCFSQVHTKHLTVTIDAFTIHNSLWQGKRVQLPSK-KDMAL
Caenorhabditis elegans pqn-34      THKLYAKTVTKSNRGLINNALQFSVFPGAVNNATRQATITQMASSSSKHFLILFRDQKCQYRGLYTWDEISDTAVKISGQGPPKCTEAMMNSMFKYDSGAKNFTNIATKHLSATIDGFAILDQYWQKARIPHSGTPAHKNN
Nematostella vectensis             GPQCYVKPSGKSNRKLIVNAISYVCLSGTVNKDAKERCLQAISESNGYHFMVLFKD-GLKFRGLYSFNPENEQLYKVFGIGPRVITAKMIDNLYKYSSGGKEFVKIHSKTLSISVDGLSIMKQYWQSSKPAVQSK-TMSNI
Salpingoeca rosetta                QAAAEKVLKLRSNVVICKNAIQSACLSGDINAKKRDAALRAIDSCDIPHLVIVLRTPHLKFRGVYAVVPA--GITKIYGAGPRTVTRDMVARYFRYDSGAKAFKQLPSRDFTV-CDAVALRAACWQKKD------------
Allomyces macrogynus               ----NSSASRRSNKKLIVNALN-MCLSGPVNEKTRTEVLEDLERSDQTNFVVLLRGPNHSFKGLYVHVVDQNQIIKLYGNGPLEIEQADVSEFFKYDSGSRSFKPIPTRSFSTFVHGVAVQATVKH---------------
Spizellomyces punctatus            IKRAI--IKVQSNRTLIKNALMHVCLAGSANESTKQEVMEDLHDSPAQHFVIIFKDVNFSFRGLYSWDAVLDQTLKVYALGRNVLDPSQVLEFYKYDSGARTFKPVATKSFGRSVHAVAVGRDFLM---------------
Thecamonas trahens                 SAPLRVEPAPLSNAKKVRNAIVSRLLAGYSSAPQCQAVLAAMDKARYTSYLLLVVQHHISFRALYGYDVATGRVDKIVGKGPASLNFKHIGGFFKYETGSRSFVQLRTKHLAPTVDAVVIHPDYLRKRKPSHKRKRTSAAA
Phytophthora infestans             KEKLQRRPKKASNRQLIQNALEFTLLAGGSMEKERSLALQALAESTCDNFIVLLKSAELKFRALYESHVDRDFATRIFSLLPSNSSSDMISQFFKYSSAKKHFQSVPTRSFTVKTDACALTDQV--KGKSALARL-L----
Aureococcus anophagefferens        APRRRDAPKRLSNRRALTNALTQICLAGPHCANQLNLALEALDRSHADNHLVVLKDDVHTFRGLYALRRGRADAEKIYGRGPARFGEEHVMKFLKFNSAARAFTPLPSKSFSFTTDAAVLALRPKQAY-------------
Tetrahymena thermophila            TTKPRIPFSKPSNKQLISQAITKVCLAGPTNSKEREIALKKIKDSQQEHFIVLFKGIRQDFKALYSMDSNTNEVKKIFGTAPTQIDLASVQAFYRYDSGSKEFKEILVKAFSNQVDAVLLKK-------------------
Plasmodiophora brassicae           -PRPDSRMLPRSNRSIMRNALSRVCLAGGATELERMQALDALEQSPSDWFLVLLRKERHKFRGLYQVDACNDSLTRIYGDGPIHINGDTLDGVFKYDSANKCFEQLASRQLTATTDAVSIASKKFGSV-------------
Klebsormidium flaccidum            VSPFYAAPPRMSNKKLVRNALSTVCLAGGALKLERERALMALDAFDGDNFVILFKDDGTKFKGLYVLSADGKEIDRIYGSGPESLRSDGVAHYLKYDSGAKEFKPLQTQQFTTSTCAVAPSNRA--PGK------------
Micromonas species                 DPRLFKAAPPLSNKKLVRNALRHVCLAGAAMRTQLDEALTALDAVAATVFVILFRENPHKYRALYALVSGDGVLVKIHGGGPARVGLQWVDATMKYDSGQREFKPLATGRLTPLTAAVTLVSKK-----------------
Trichomonas vaginalis              KKEITEHTSTVPNISTIANALKYHALPGPAHVDEYNNLMKILQDRFSNERILILMASTMKYKGIYILNND--VAKKIYGIGQDSIYQSDVSAFYKFVNGTKQYERILSKSFTQTTDGFSMKREKW----------------
Naegleria gruberi                  MILKNMKQPNKTNKLLIKNALIHLTLAGEVNKKEREDVFEAMKEEDTNQMIILVREVVPAFRALYVVVTDSNLVKKIIGKGPKFLTEDVVDVFCRYDSGGKKLSKLSSRTFGVTTDVVVLKSAAIKKIK------------

Conservation

                                   000020032227*313829*9202586*02221234258329231142148997430023 99+9*321130015199281751142021731569939549272191642871062865621001----------------
```


g
HsCAMSAP1 CKK/tub
Fly CKK/tub
Worm CKK/tub
T. thermophila CKK/tub
T. vaginalis CKK/tub


CKK
loss of CKK
Holozoa
Fungi
Apusomonadida
LECA
Archaeplastida
Stramenopila
Rhizaria
Excavata
Metazoa
Choanoflagellatea
Capsaspora owczarzaki
Ichthyosporea
Dikarya
Zygomycota
Blastocladiomycota
Chytridiomycota
Microsporidia
Fonticula alba
Thecamonas trahens
Amoebozoa
Embryophyta
Klebsormidium flaccidum
Chlorophyceae
Mamiellales
Rhodophyceae
Cyanophora paradoxa
Oomycota
Aureococcus anophagefferens
Phaeodactylum tricornutum
Blastocystis hominis
Alveolata
Plasmodiophora brassicae
Metamonada
Kinetoplastida
Naegleria gruberi
Homo sapiens
Salpingoeca rosetta
Sphaeroforma arctica
Saccharomyces cereviseae
Rhizophagus irregularis
Allomyces macrogynus
Spizellomyces punctatus
Vavraia culicis
Dictyostelium discoideum
Arabidopsis thaliana
Chlamydomonas reinhardtii
Micromonas species
Galdieria sulphuraria
Phytophthora infestans
Tetrahymena thermophila
Trichomonas vaginalis
Trypanosoma brucei
Loss / deletion

neutral evolution at speed beyond ability of blast to detect homologs



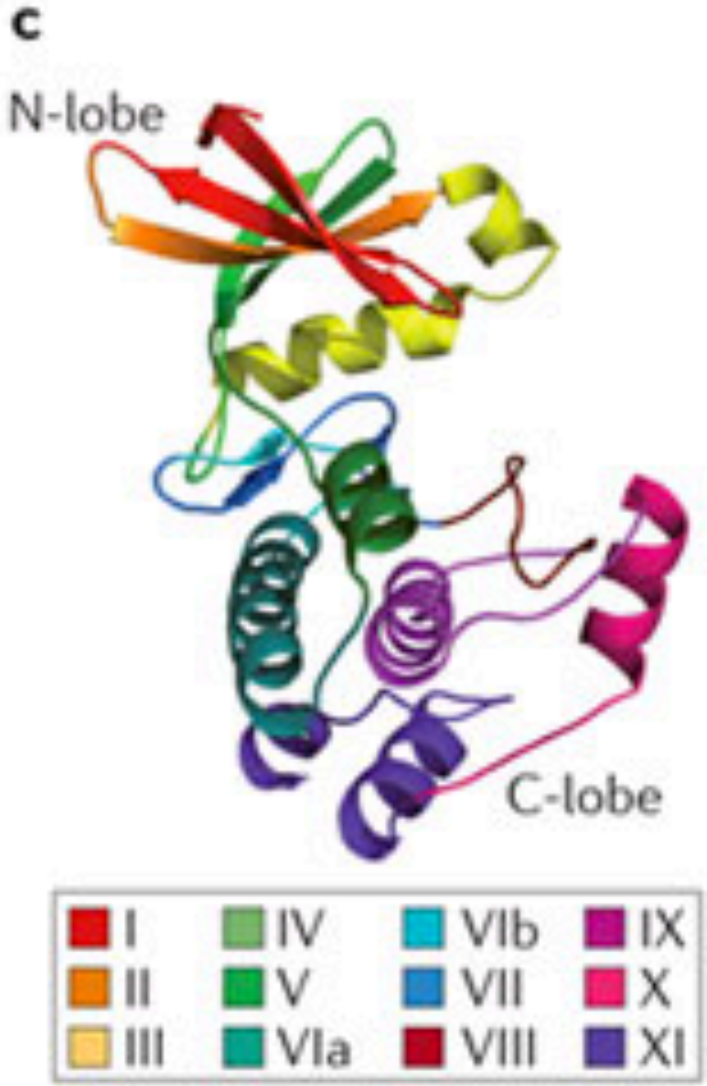
Despite / because of these issues, we want to infer the evolutonary history because it is

- challenging & thus fun (puzzle!)
- necessary to find out what is in fact happening in genome evolution
- provides an unique / complementary insight into why the cell works the way it works
- needed to describe what happened at major transitions in evolution: such as single-cell-multicellular, origin of eukaryotes, & much more

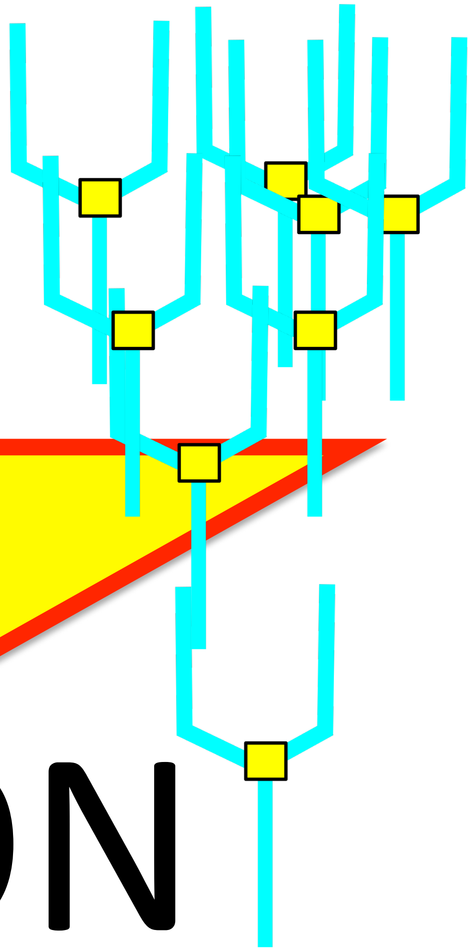
Why do I do this? Why is this research relevant? From personal to practical

- Interplay genome & network evolution Just like sequence alignment and substitution matrix has learned us a lot (hydrophobic core, motifs, important residues) and is still being used (also by us) as primary tool to understand proteins.
- Where does gills, fins or wings come from? Do comparative anatomy, same questions for molecular machine's … not so ”easy” as dissecting the body, finding fossils … instead everything via genome,
- W.r.t. to the latter, it turns out a lot complexity vis-à-vis cellular machines arose during *eukaryogenesis*, for which no intermediates and no fossils, comparative genomics is all we have ..
- Understanding / Describing what happened: how did the composition of our genome or that of yeast look like, history, phylo-stratigraphy, when did what arose, when did what was lost etc.

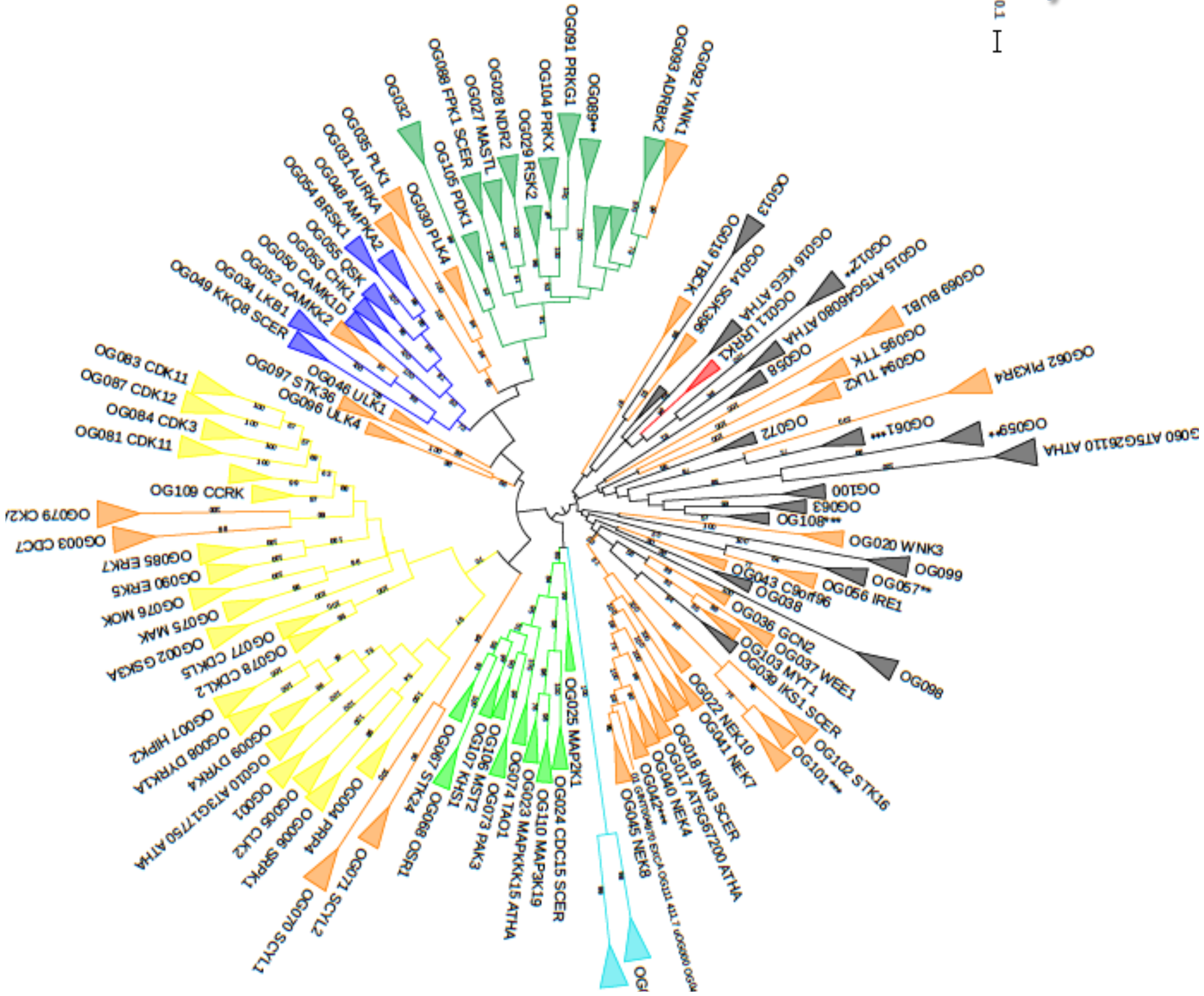
±5 protein kinases donated from prokaryotes to eukaryotes



MASSIVE
GENE
DUPICATION



Kinome of the ancestor of all eukaryotes

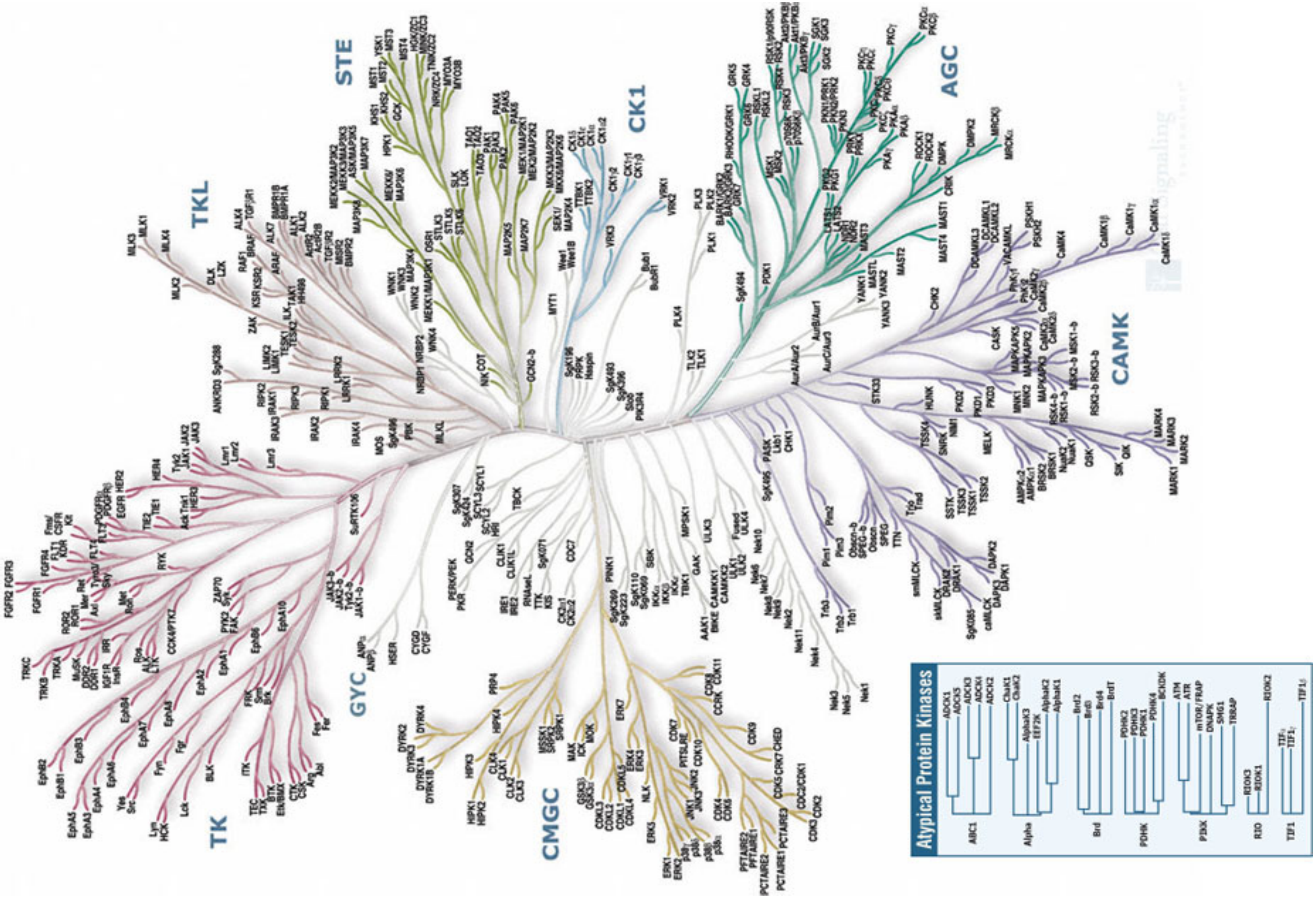


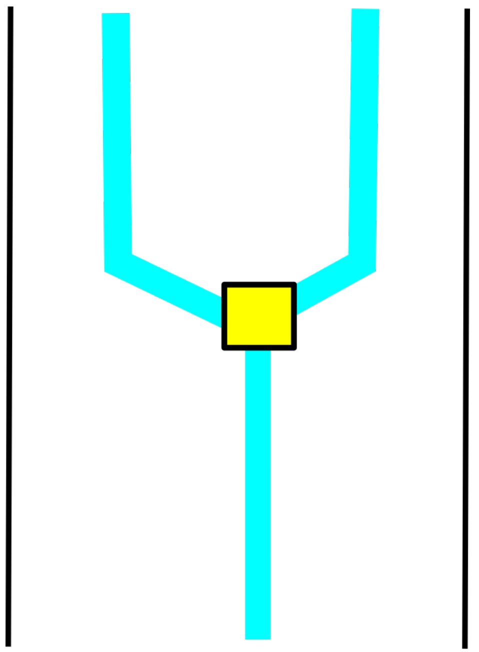
MASSIVE

GENE

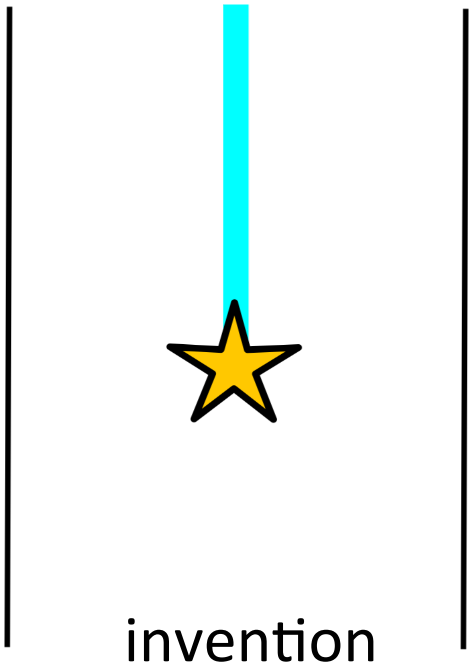
DUPICATION

Human kinome ±500 kinases

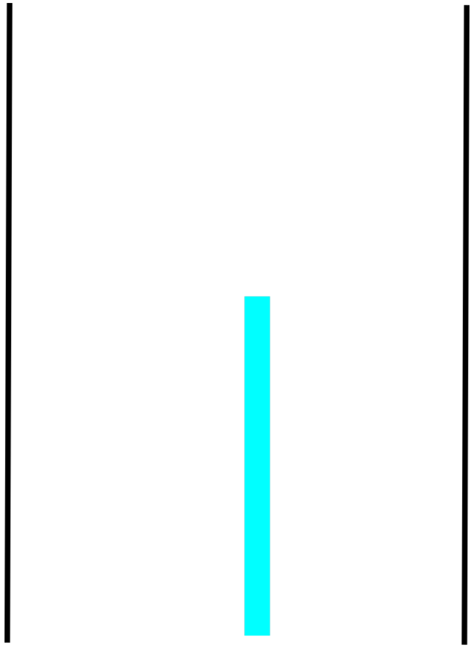




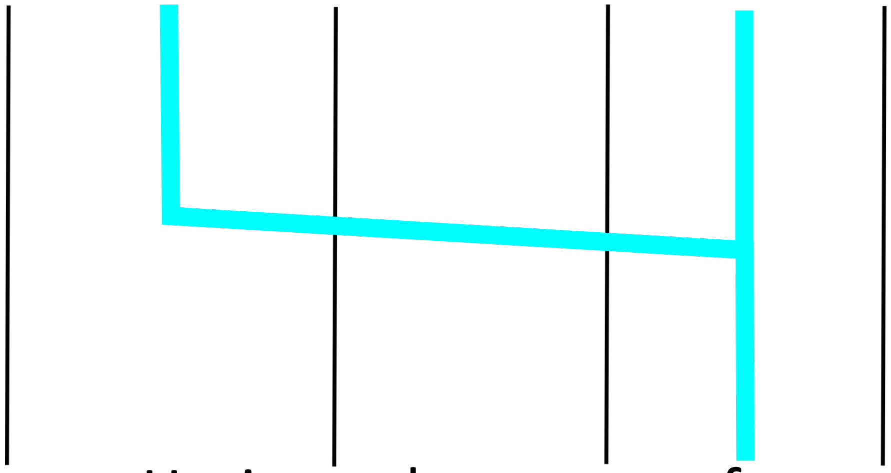
duplication



invention



Loss / deletion



Horizontal gene transfer

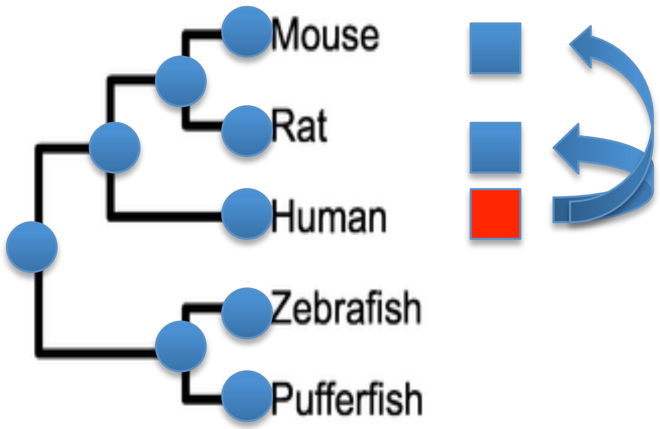
Reformulation of course goals

- How to postulate / infer the occurrence of these events (sketch a scenario of what happened)?
  - What you need to know to be able to do this
  - "By hand" using bioinformatic tools
  - Automatically by bioinformatic pipelines
  - Knowledge on common scenario's provide a prior. i.e. proteins involved in core cellular processes were present in common ancestor of eukaryotes but also subject to independent loss.

- What we have learned from research that performs these kinds of analyses

What happened What you see

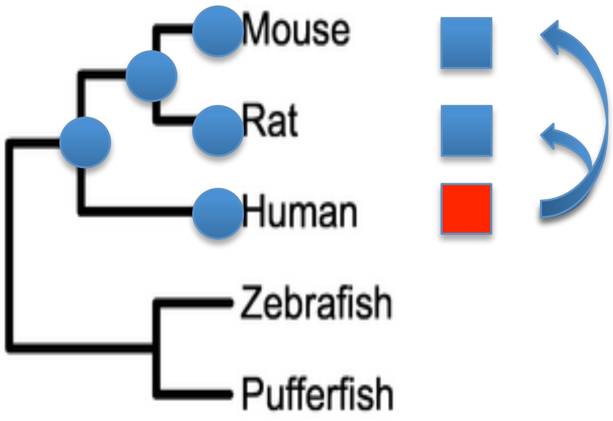
blast



Gene originates in common ancestor… but evolves rapidly (coiled coil, disordered, very short globular domain)

What happened What you see

blast

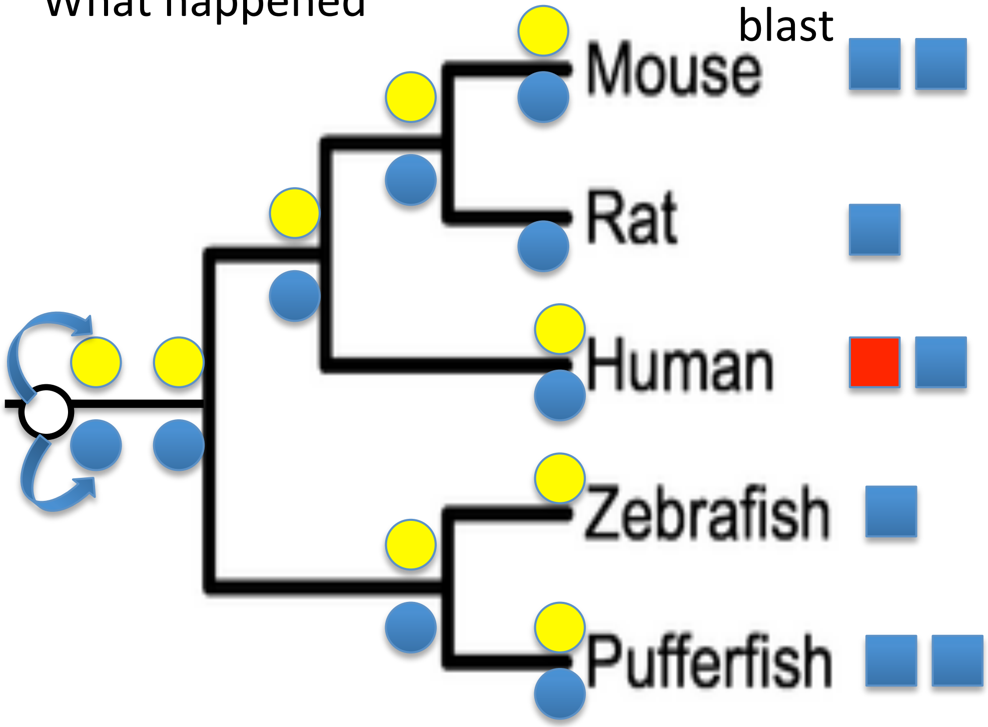


Gene originates later… evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

Problem: divergent homology = orthology

What happened

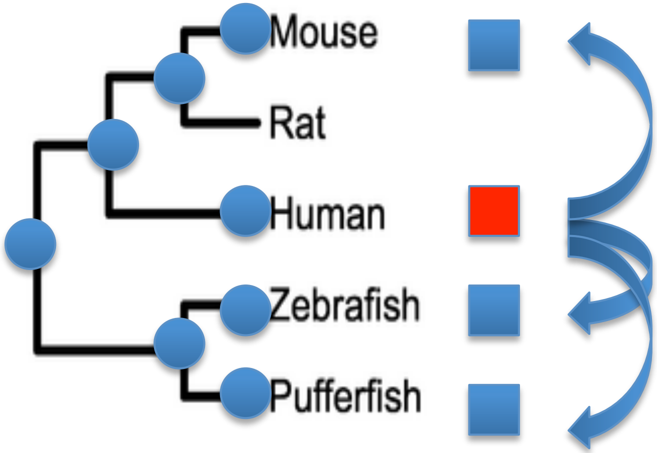
What you see/ blast



What happened

What you see

blast



Gene originates before common ancestor. Duplicates. … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

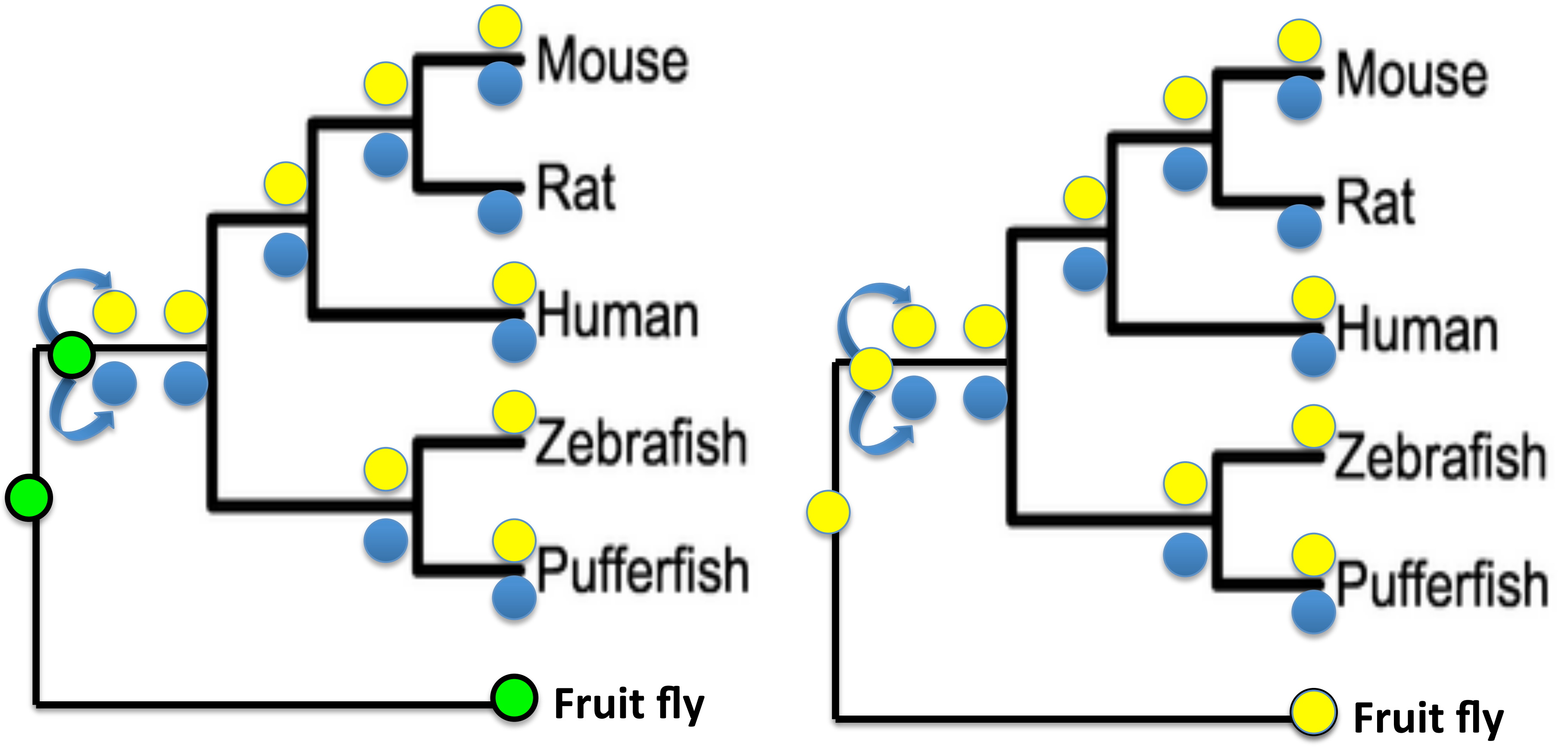
Gene originates in common ancestor … evolves normally (has decent length e.g. 200AA and globular fold). Few losses.

Problem: distinguishing paralogs from orthologs

What happened in genome evolution vs **why** it happened and what happened to **function**

- Next to “genome evolution”: I want to also discuss evolution of function /interplay of genome and network evolution, but
  - Big problem 1: we are crap at formalizing what we mean by function
  - Big problem 2: we in fact have very little data on function compared to genome sequencing data, for many reasons but for example because function is e.g. condition dependent while a genome sequence is a genome sequence
- (hence the initial focus on what happened)
- Nevertheless we will discuss sometimes the implications for function prediction and speculate on what are the functional/ phenotypic consequences of all these genome evolution events
- Some data from high-throughput experiments that measure function: “comparative interactomics”

Same scenario? Color = function, green is multifunctional



In blast & tree they are likely the same …

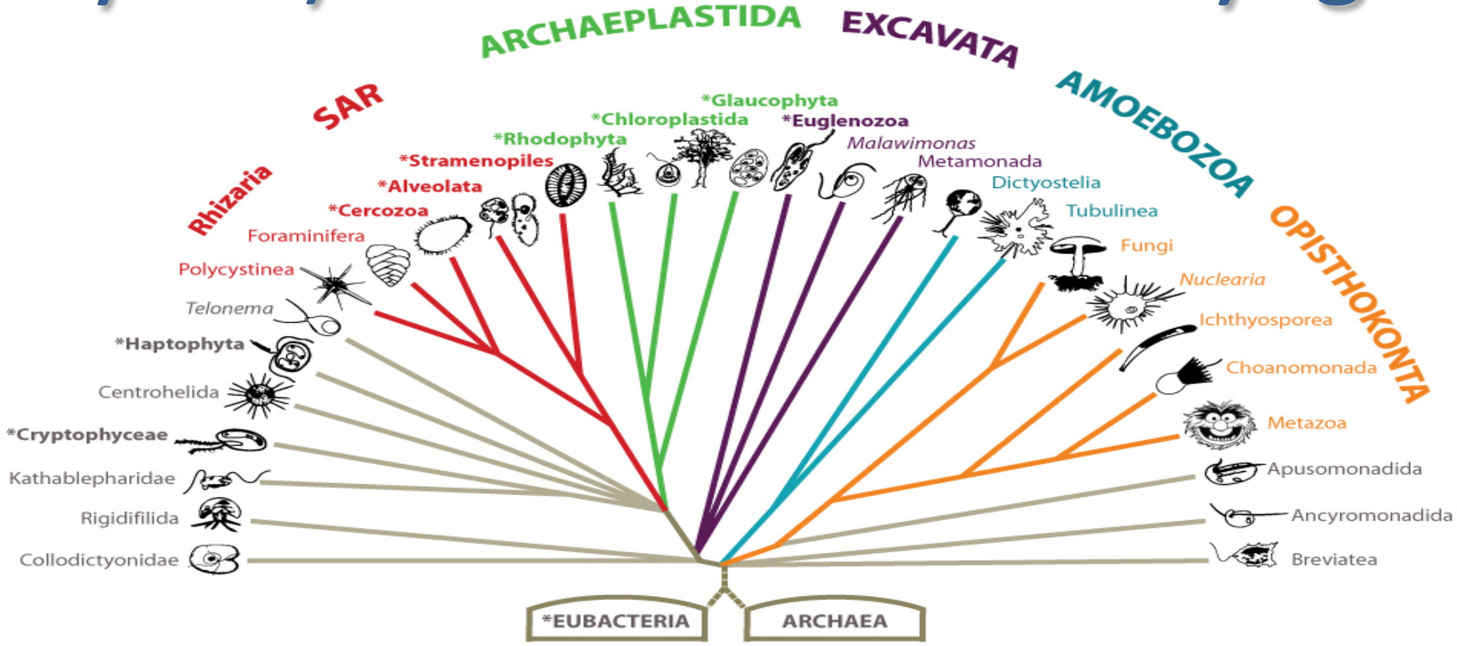
Problem: same gene evolution vs same gene function

practical/procedural: Small scale & Large Scale

- How did my protein, complex, pathway evolve? (collaborations)(COO, mini project)

- Large scale, how do genome, networks and complexes evolve (context/expectation, bioinformatics senior authorships)(paper discussions) What can we infer about eukaryogenesis?

(Eukaryotic) tree of life & eukaryogenesis



- Which genome to include. What does an absence mean?
- Essential for interpreting gene trees:
  - Knowing (at least the outline) by heart >>> having to look it up
- With regards to evolutionary signaling cell biology (kinases, smallgtpases etc.)the diversity in present day genomes is staggering and dwarfs e.g. human-fruit fly difference

More Practical stuff

- (Schedule)
- Literature discussion
  - You should have read the papers in depth before the discussion
  - I will shortly introduce and then invite people to discuss figures / pieces of the results
  - This + participation in the discussion is 10% of grade
- Lectures online, last minute
- Mini project, let me first explain some bioinformatics … than this afternoon let's discuss it & pick proteins

Computer Exercises

- Mostly use of web resources.
- Computer exercises for some topics many others more difficult (i.e. evolution of interaction networks based on HTP analysis).
- Ask help from fellow students.
- Should tie strongly into mini-projects
- (I am slightly afraid the data bases are getting unwieldy w.r.t. number of genomes … searches very slow … you need to already know the ToL to pick relevant species)

Mini project 1

- **The protein.**
- **What does my protein look like (protein topology, domains, coiled coil, disordered regions, etc.).**
- **What (if anything) has already been postulated about the evolution of your protein in the literature**
- **Size of the (super)family in the genome you're sequence is from and a few other eukaryotes**
- **Homologs across tree of life**
- **Tree of relevant sequences in diverse genomes**
- **Orthologs in genomes from your tree. (or from homology searches)**
- **Cartoon tree of species or genes depicting what happened in evolution**
- Optional: Does your potein or any of its orthologs in other species have Whole Genome Duplicates (WGD)/ Ohnologs?
- Optional: Point of invention of the eukaryotic orthologous group your protein belongs to.
- Optional: interactions of proteins in your tree according to biogrid
- Optional: orthology of interactors of your proteins according to biogrid and an automatic ortology database such as. E.g. panther.

Mini project 2

- Species tree, you really get to know the outline if you are using the ToL to describe the evolution of a protein. Similarly for e.g. smart/pfam etc.
- Students are often finished too long after the course … for your own benefit try to prevent that
- Some students get stuck on is what they find novel. It does **not** have to be novel! Just describe what you find!

Mini project / Molecular evolution is recursive / iterative 3: generalized

- To study the a evolution of a gene you need a model / framework of the evolution of the gene, but to get an idea of a proper framework / model of the evolution of a gene, you need the need to study the evolution of a gene
- Thus: heuristics & build on previous results. Start from stuff you trust (alignment of highly identical sequences), and/or only the use the general but flawed overview (e.g. guide tree). Then iterate
- Not yet so automatically solved for evolutionary history of a gene and its homologs as it is for other case …