

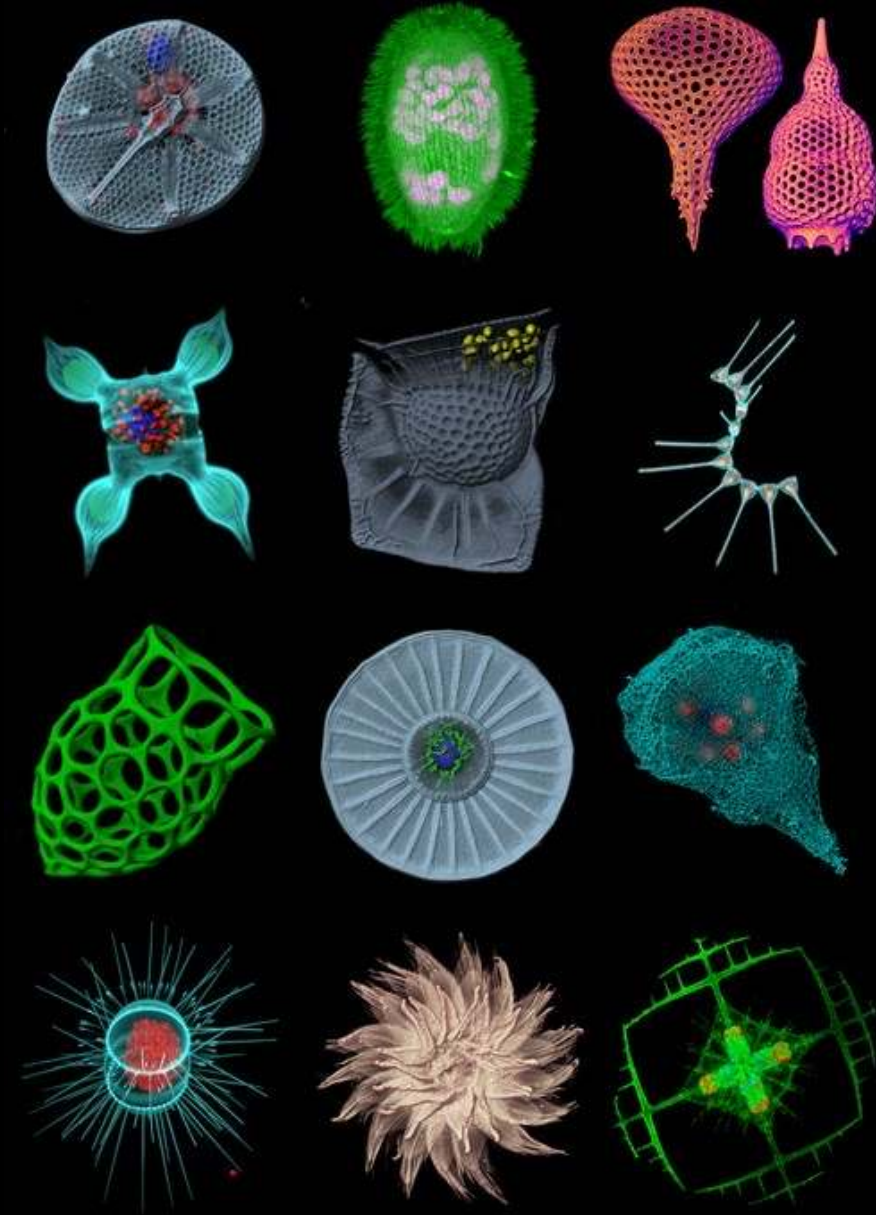
Unravelling complexity: Exploring the evolution of microbial eukaryotes through phylogenomics and phylogenetics

Anna Karnkowska
Institute of Evolutionary Biology
University of Warsaw



Institute
of
Evolutionary
Biology

© S. Colin, Max Planck Institute for Biology, Tübingen, Germany



Courtesy of Sebastien Colin

about me

Phd
2011

post-doc
2013

post-doc
2016

PI
2017

sabbatical
2022



about me

- Taxonomy & Phylogenetics of Euglenophyta



My first (published) tree ever in 2005!

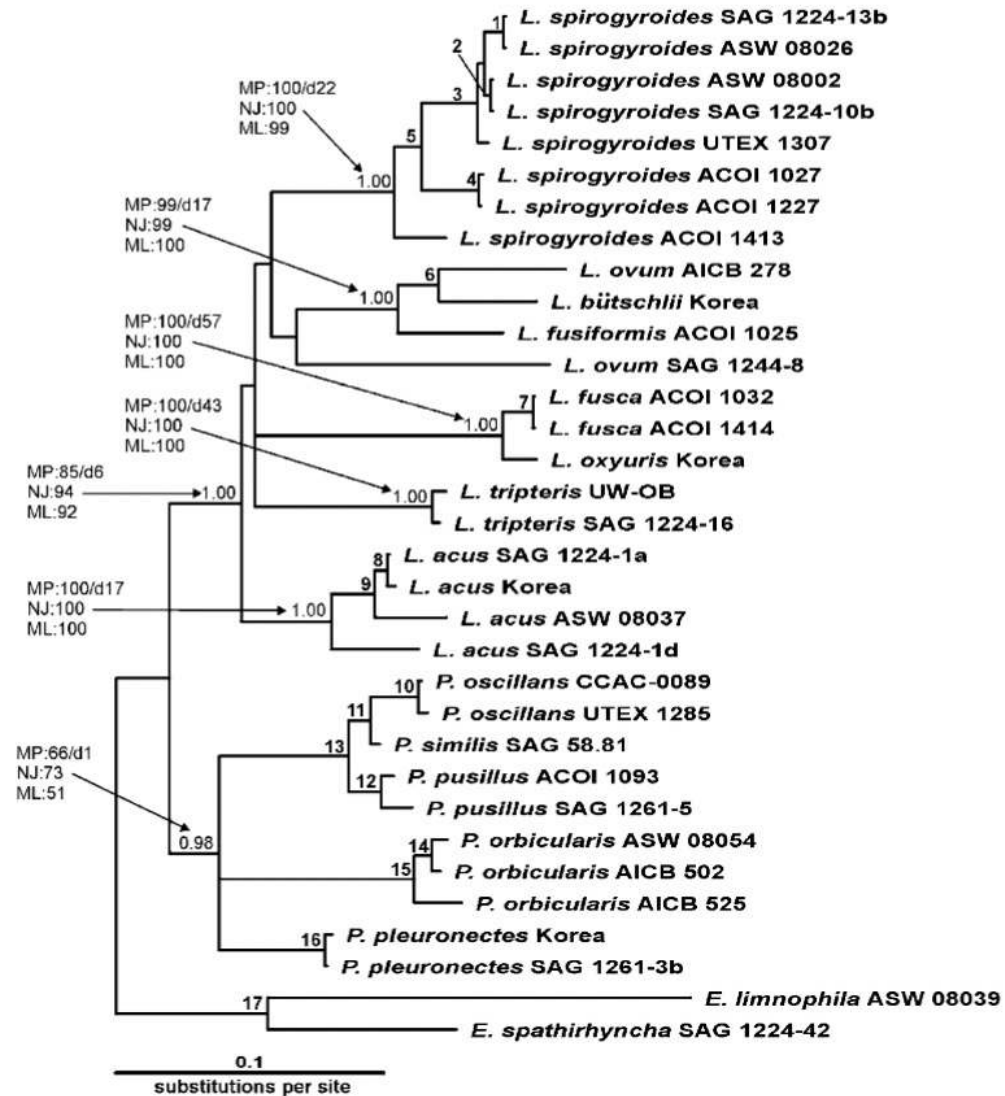
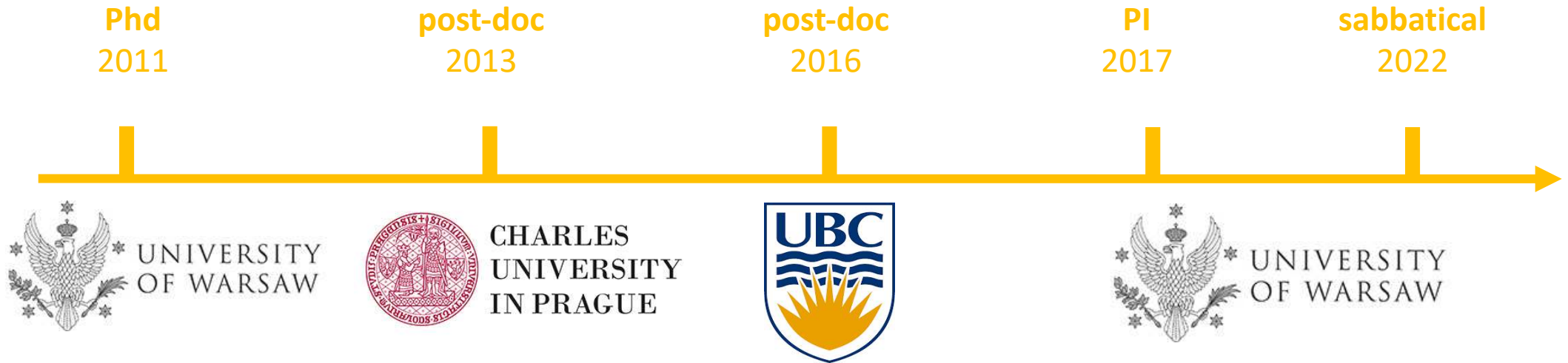
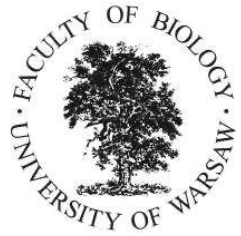


FIG. 2. The phylogenetic tree of the 18S rDNA sequence obtained by Bayesian inference (model GTR + G + I). Numbers at the essential nodes show posterior probabilities of the tree bipartitions as well as the bootstrap values/decay indices obtained for the main clades by MP analysis and bootstrap values obtained by NJ and ML analysis (model GTR + I + G). The support for the remaining nodes (numbered) is listed in Table 5. Branches leading to nodes with support of less than 50% are collapsed, and those without substantial support (pp < 0.95, bs < 0.75) are not listed.

about me

- Taxonomy & Phylogenetics of Euglenophyta
- Evolution of mitochondria
- Reductive evolution
- Eukaryotic cell evolution
- Phylogenomics & comparative genomics of microbial euks.
- Evolution of organelles
- Eukaryotic cell evolution
- Reductive evolution
- Eukaryotic rhodopsins diversity & evolution
- Freshwater protists diversity and interactions
- Phylogenomics & comparative genomics





Centrum Nauk
Biologiczno-Chemicznych
Uniwersytetu Warszawskiego

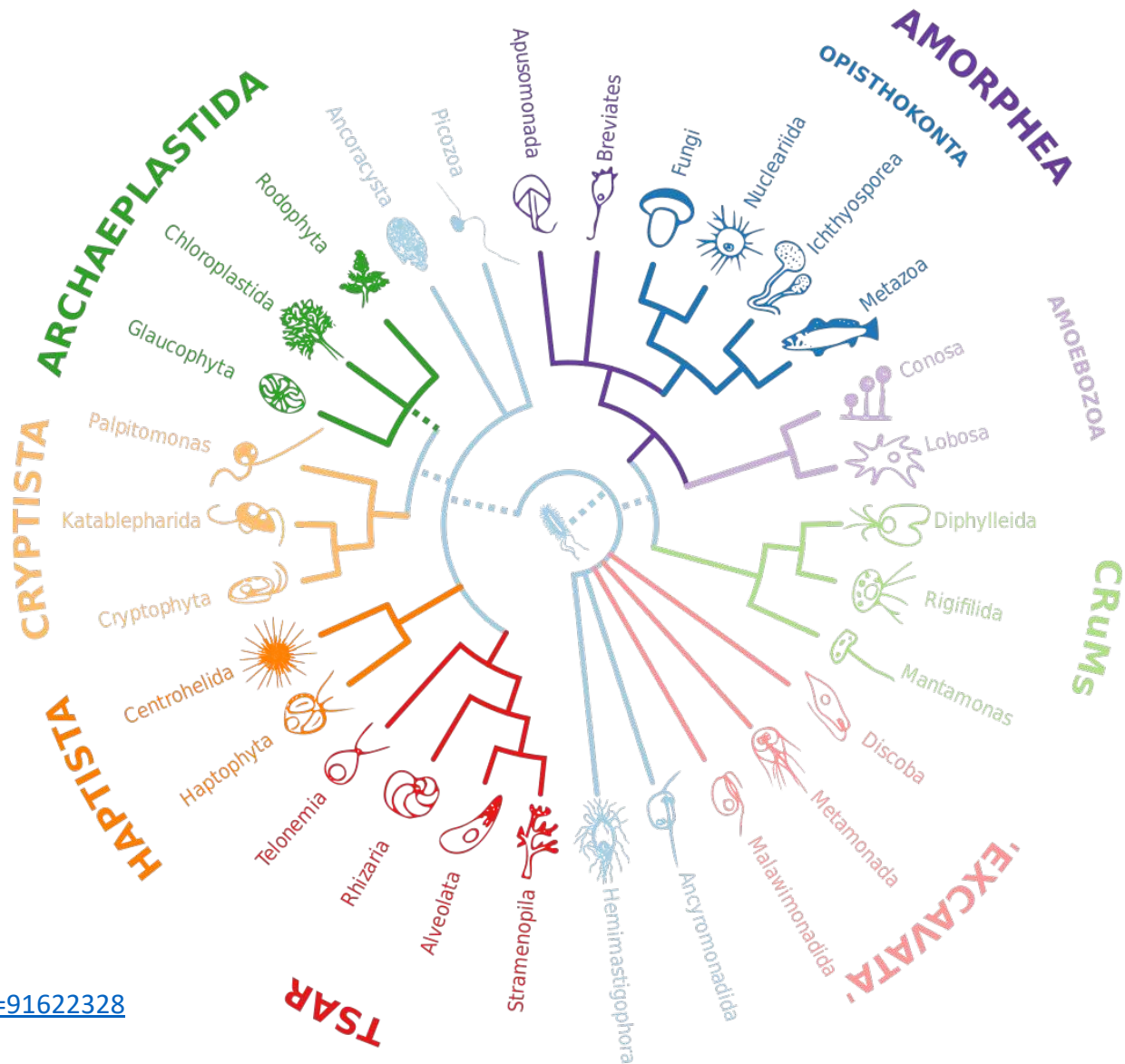


Institute
of
Evolutionary
Biology

Evolution & genomics of microbial eukaryotes



microbial eukaryotes (protists) constitute majority
of lineages on the **eukaryotic tree of life**

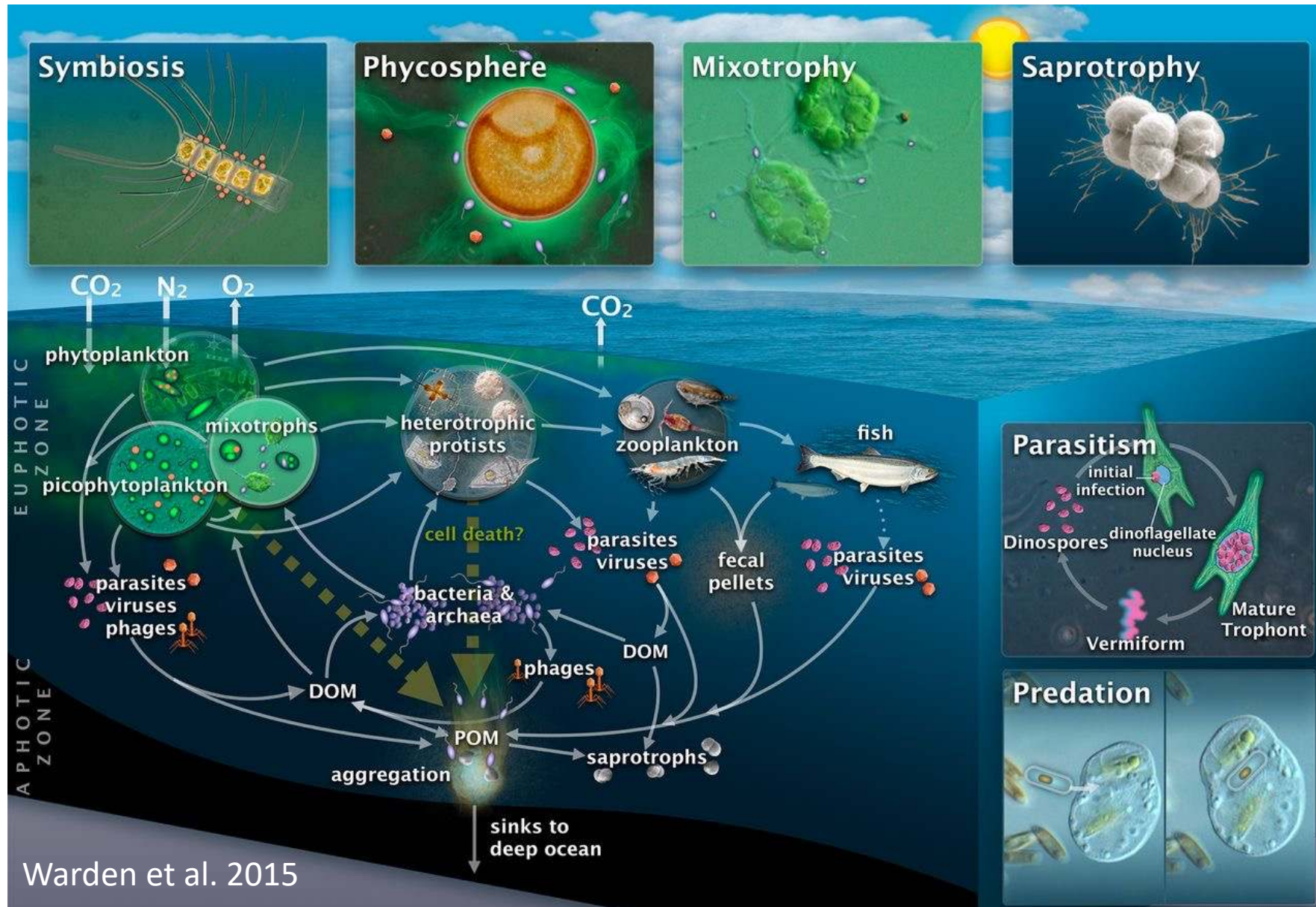


morphological diversity of protist



<https://oceans.taraexpeditions.org>

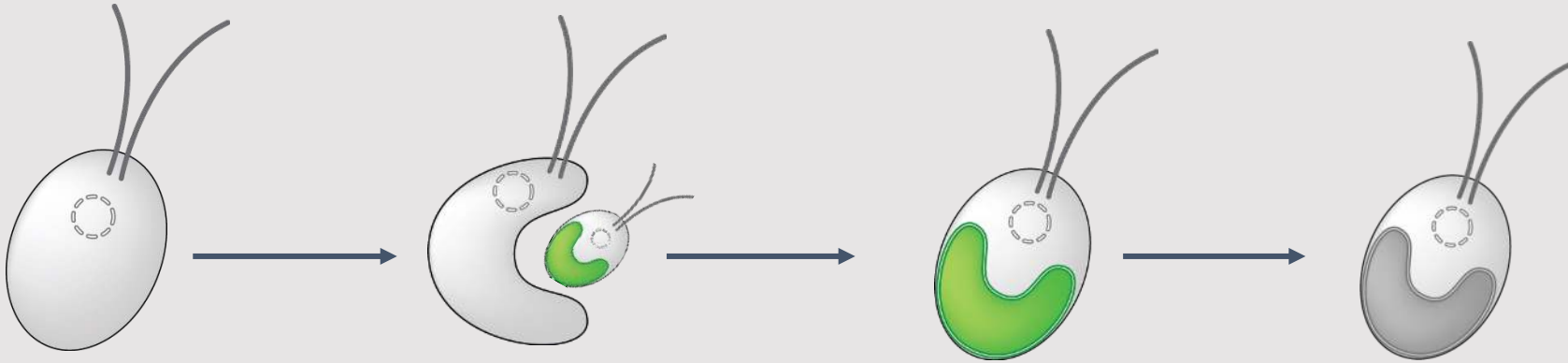
functional diversity of microbial eukaryotes



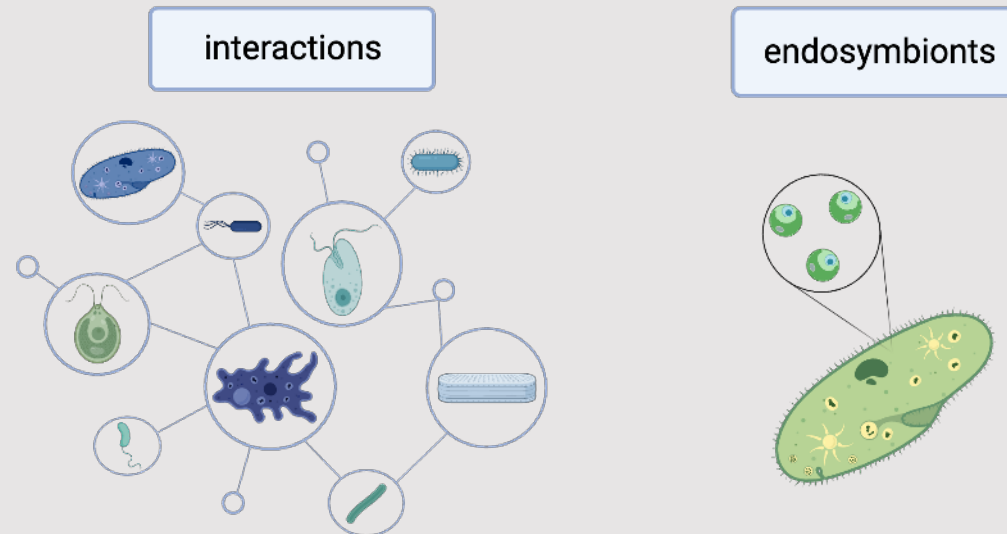
How the eukaryotic cell complexity evolved?
What drives diversification of eukaryotes?

Endosymbiosis

Origin, evolution and fate of endosymbiotic organelles



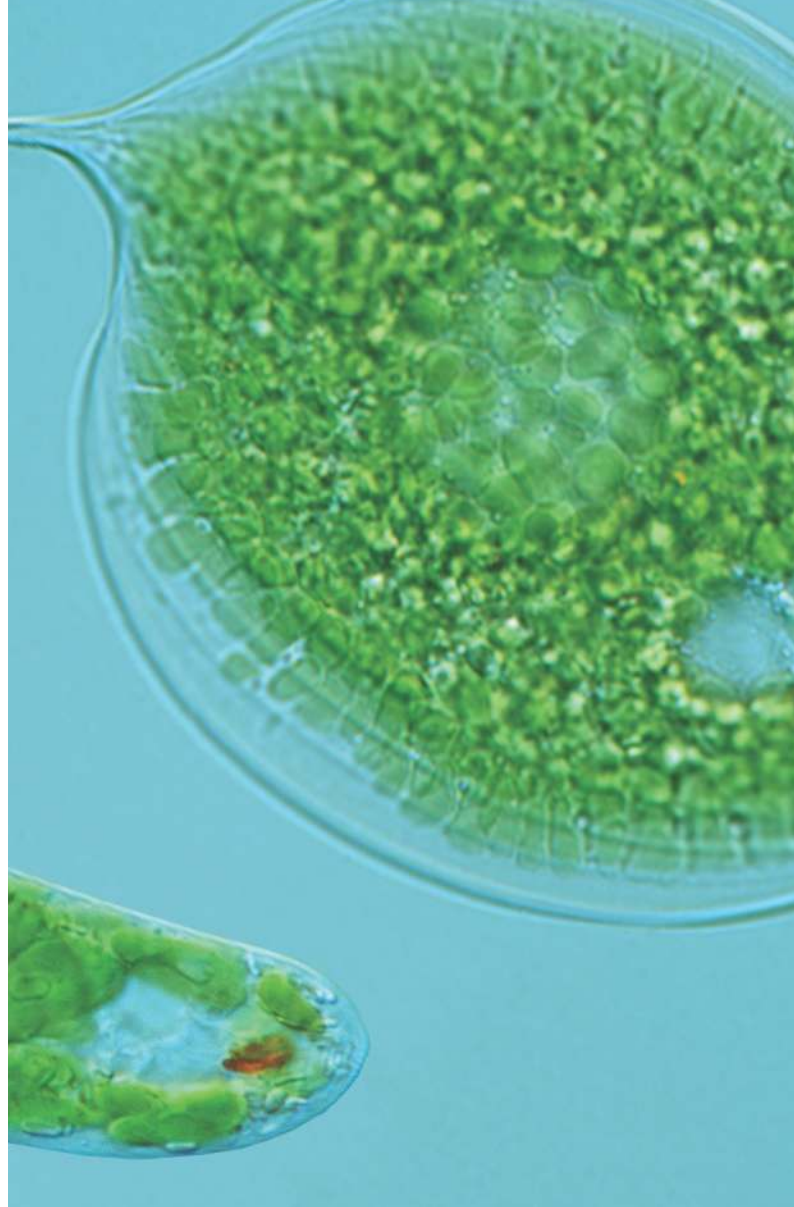
Interactions of microbial eukaryotes in the changing environment



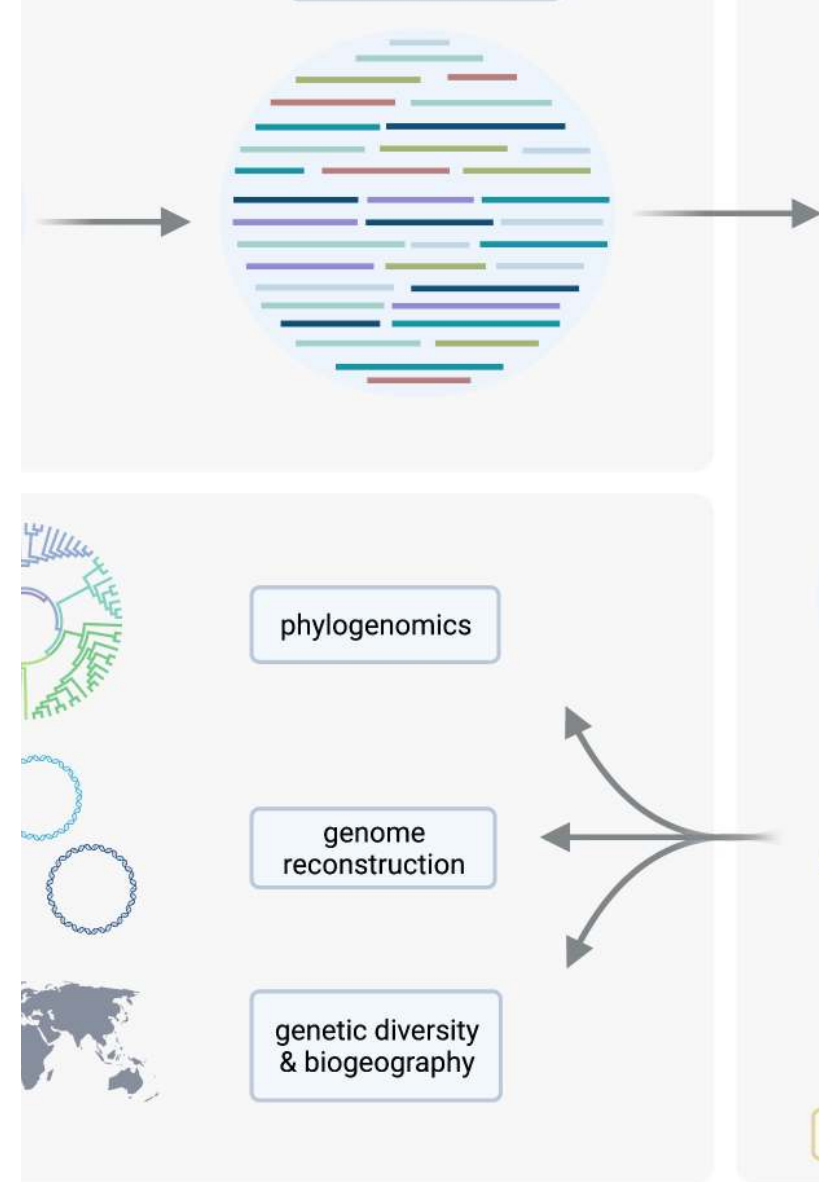
SAMPLING



EXPERIMENTS



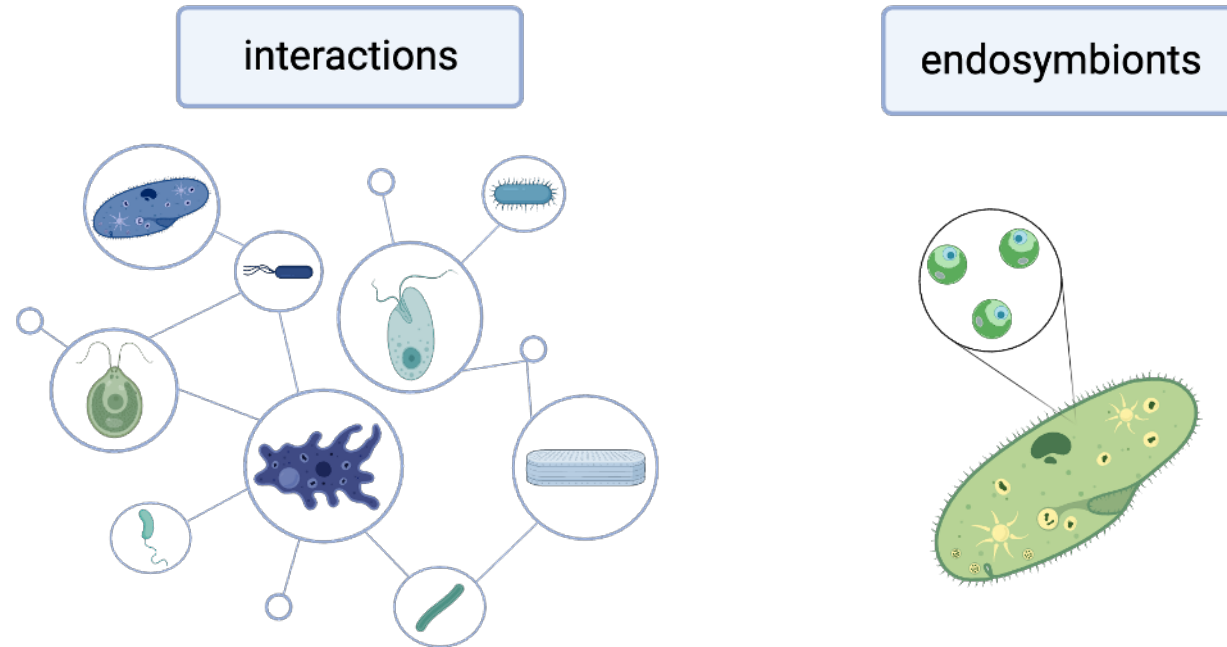
BIOINFORMATICS







Diversity & interactions of microbial eukaryotes



Who is there? What are the interactions?

Which environmental factors influence interactions?

Do all cells of the same species have the same (endo)symbionts?

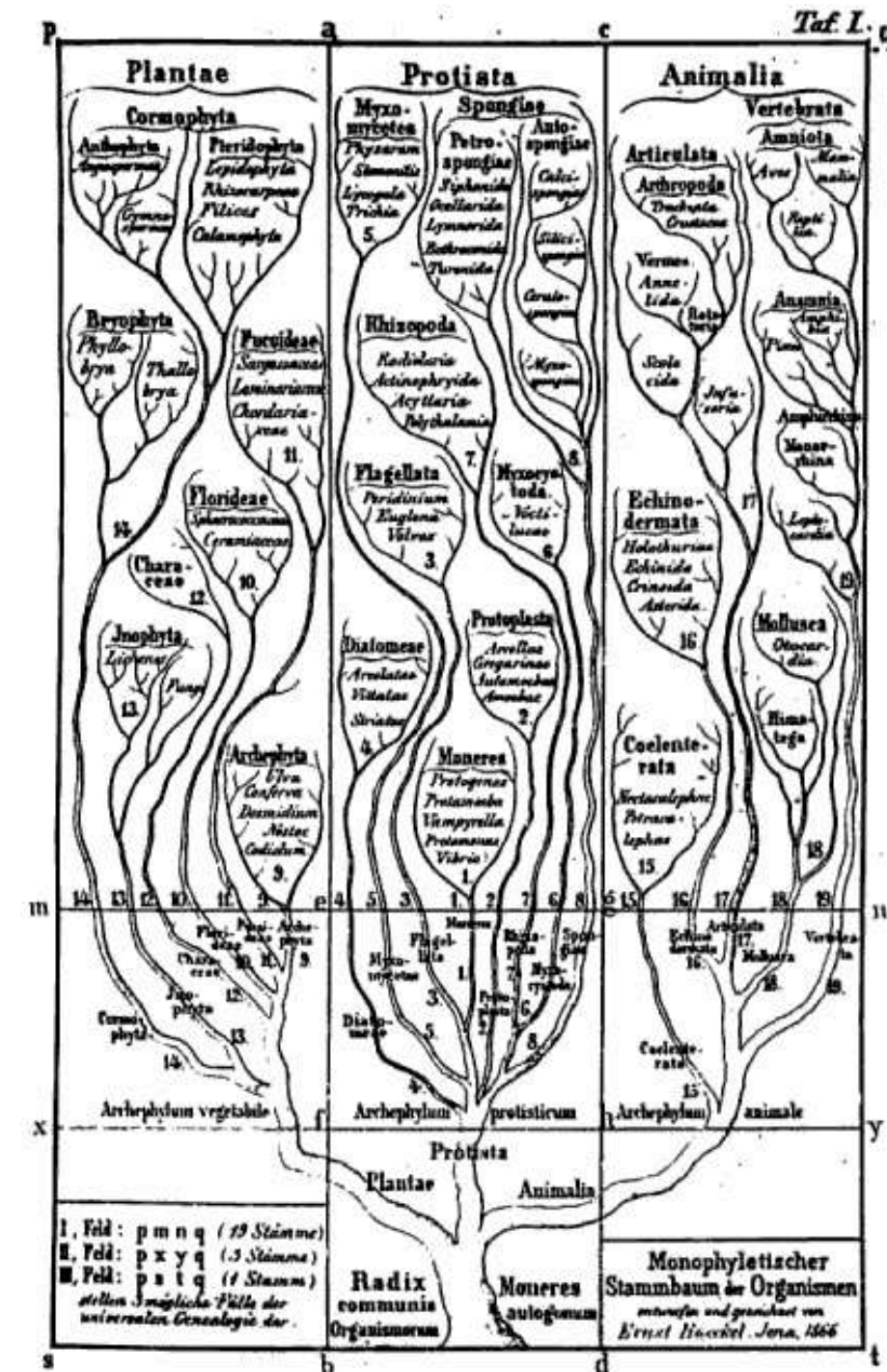
Overview

- Tree of eukaryotes – history
- Phylogenomics and the tree of eukaryotes
- Why do we need to resolve the tree of euakryotes?
- Origin of eukaryotes
- Metagenomics in phylogenomics



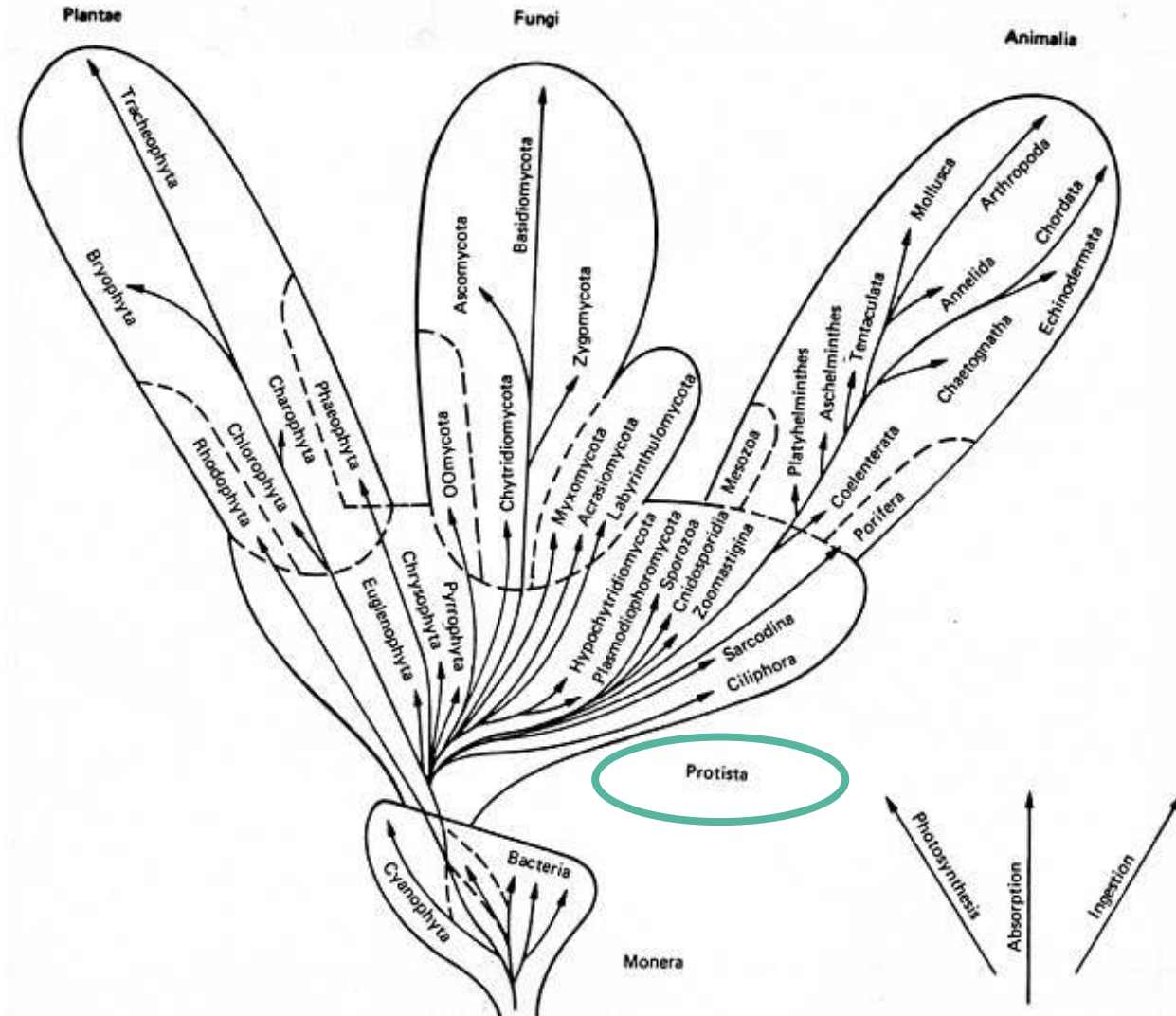
- Phylogenomics as a tool to address (my) evolutionary questions

Ernst Heckl - 1866



New kingdom of Fungi

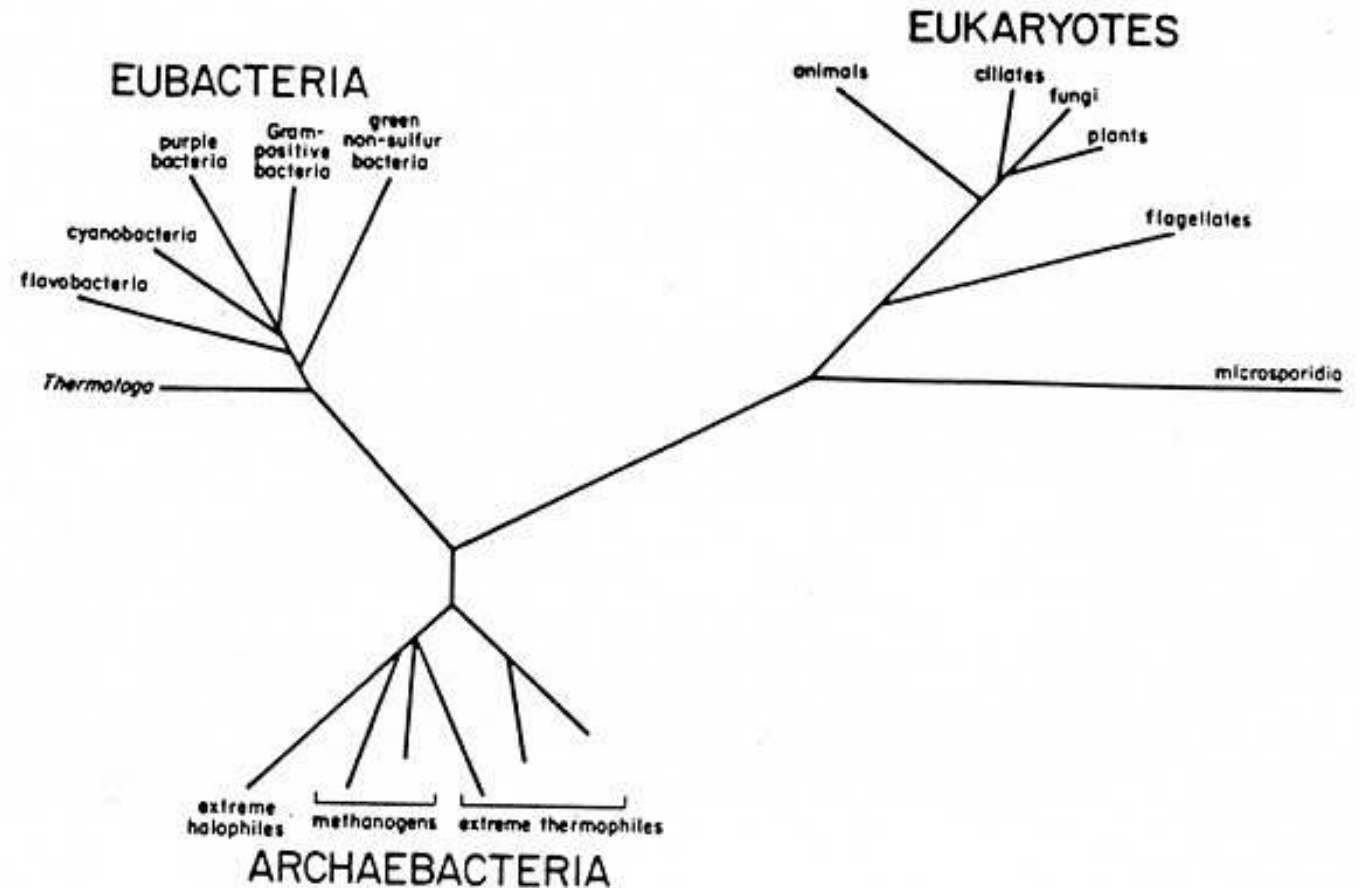
Whittaker 1967



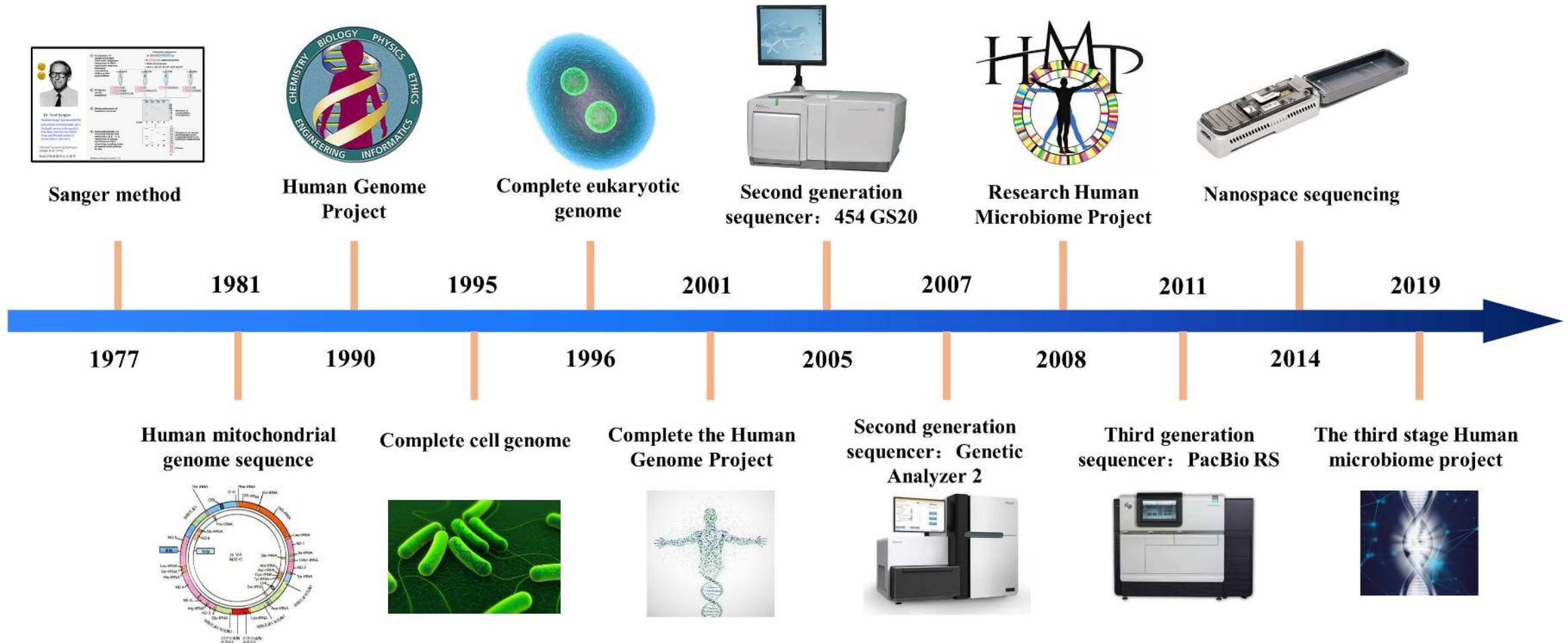
Molecular phylogeny

Carl Woese 1988

- only one phylogenetic marker - rDNA
- three domains of life
- new relationships have been detected



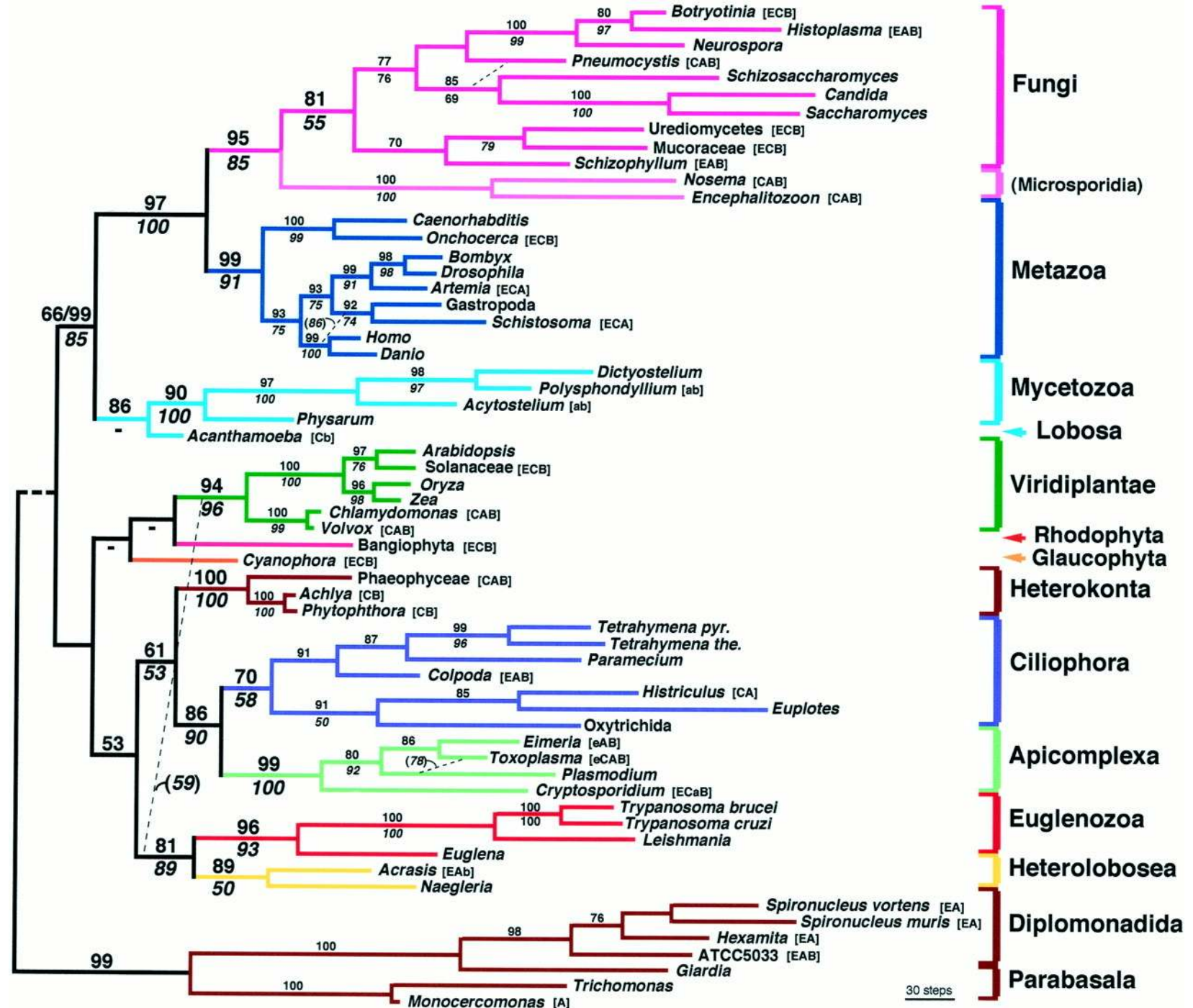
The Next Generation Sequencing Revolution



More molecular data

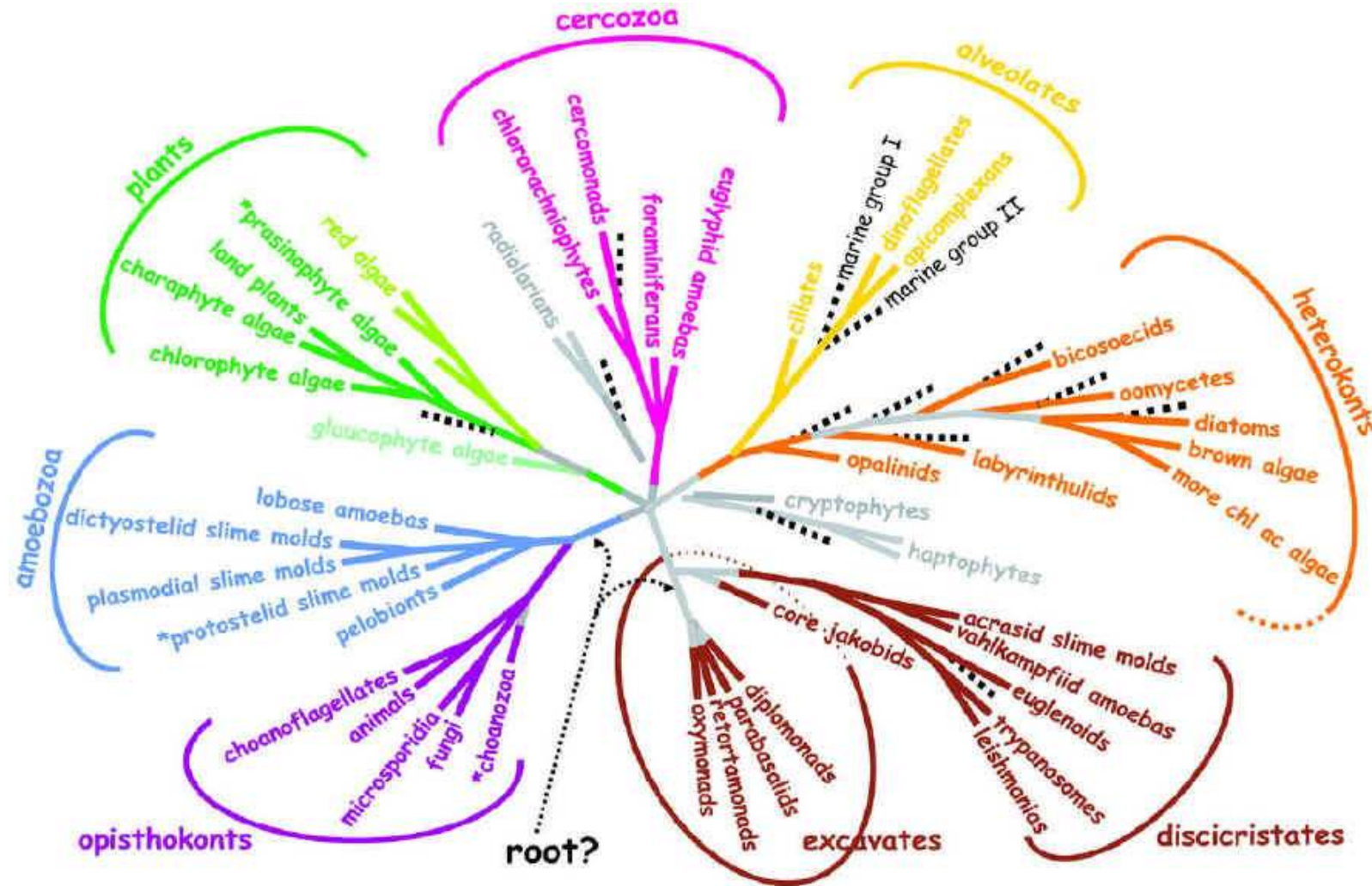
The tree shown is one of two shortest trees found by parsimony analysis of concatenated four protein-encoding genes EF-1 α , actin, α -tubulin, and β -tubulin amino acid sequence

Baldauf et al. 2000, Science

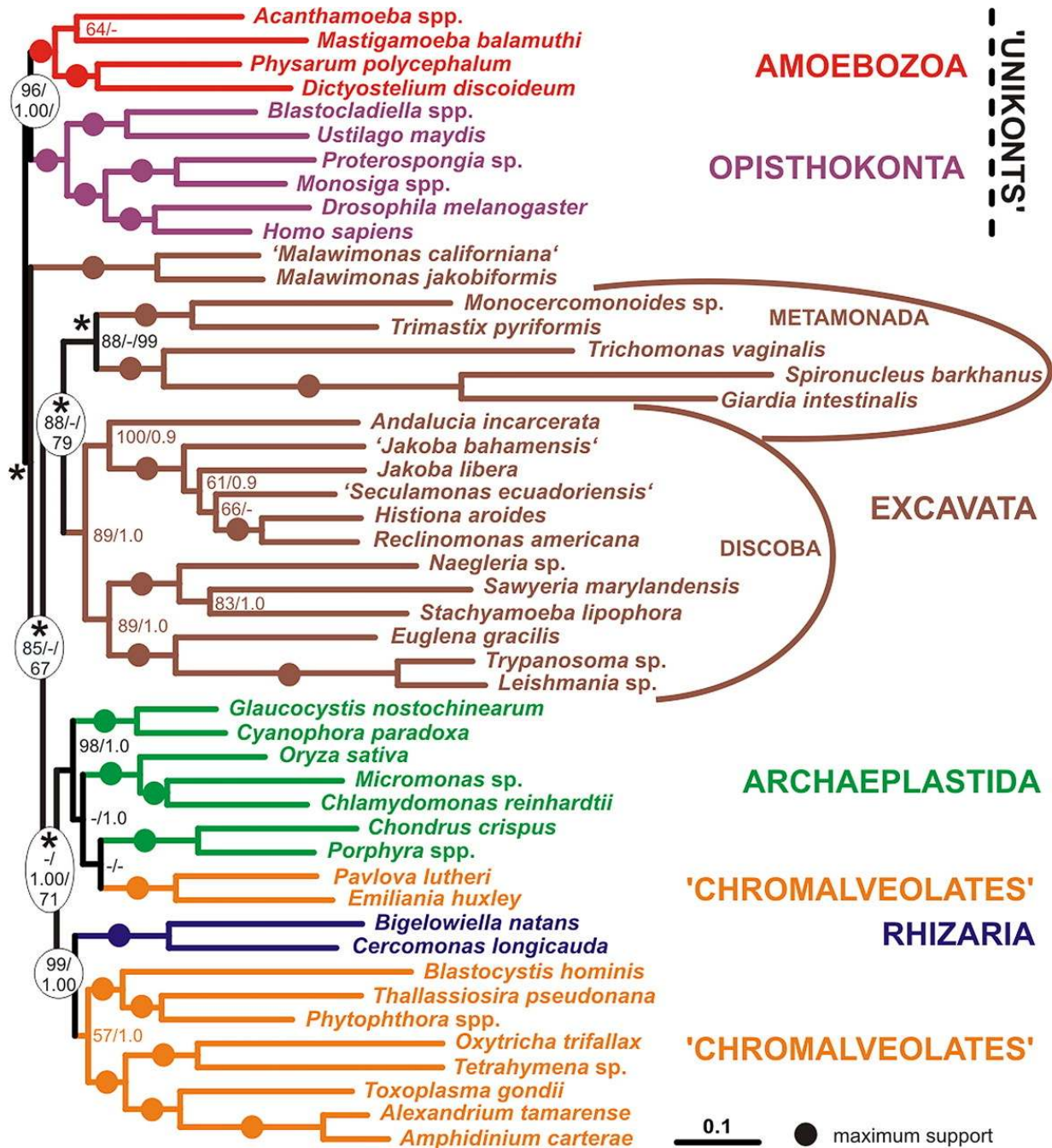


More molecular data

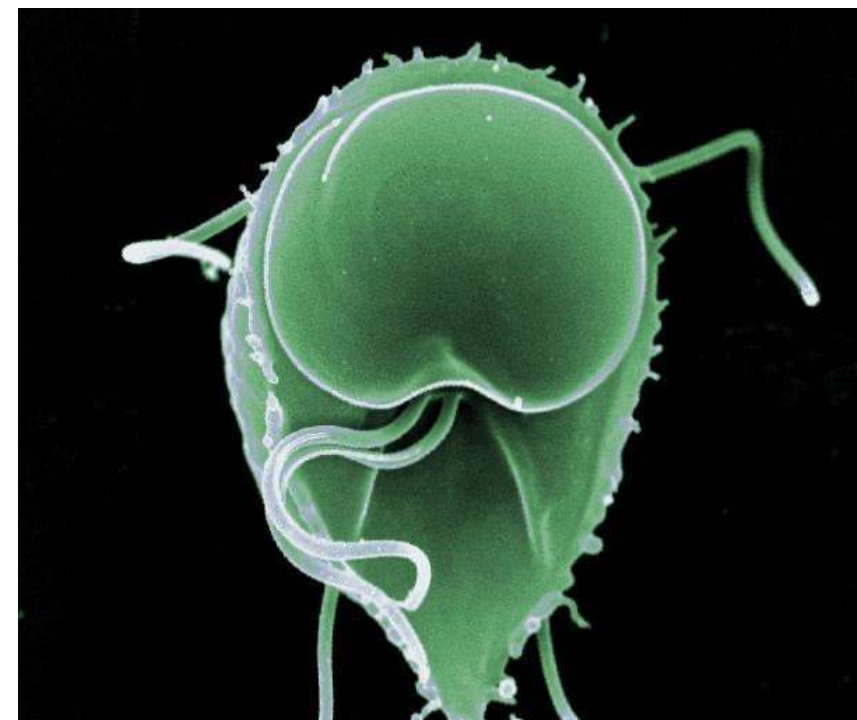
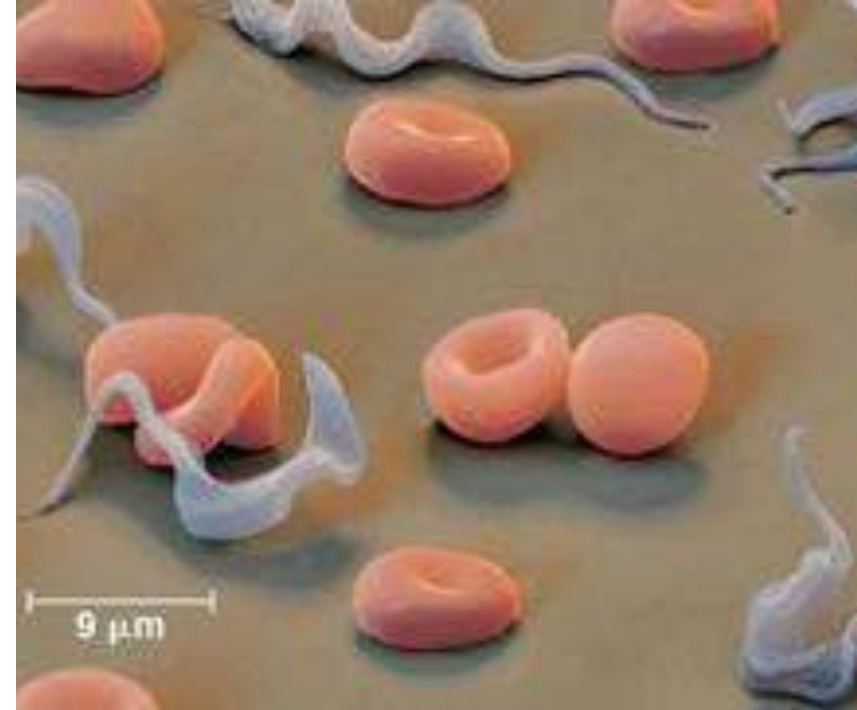
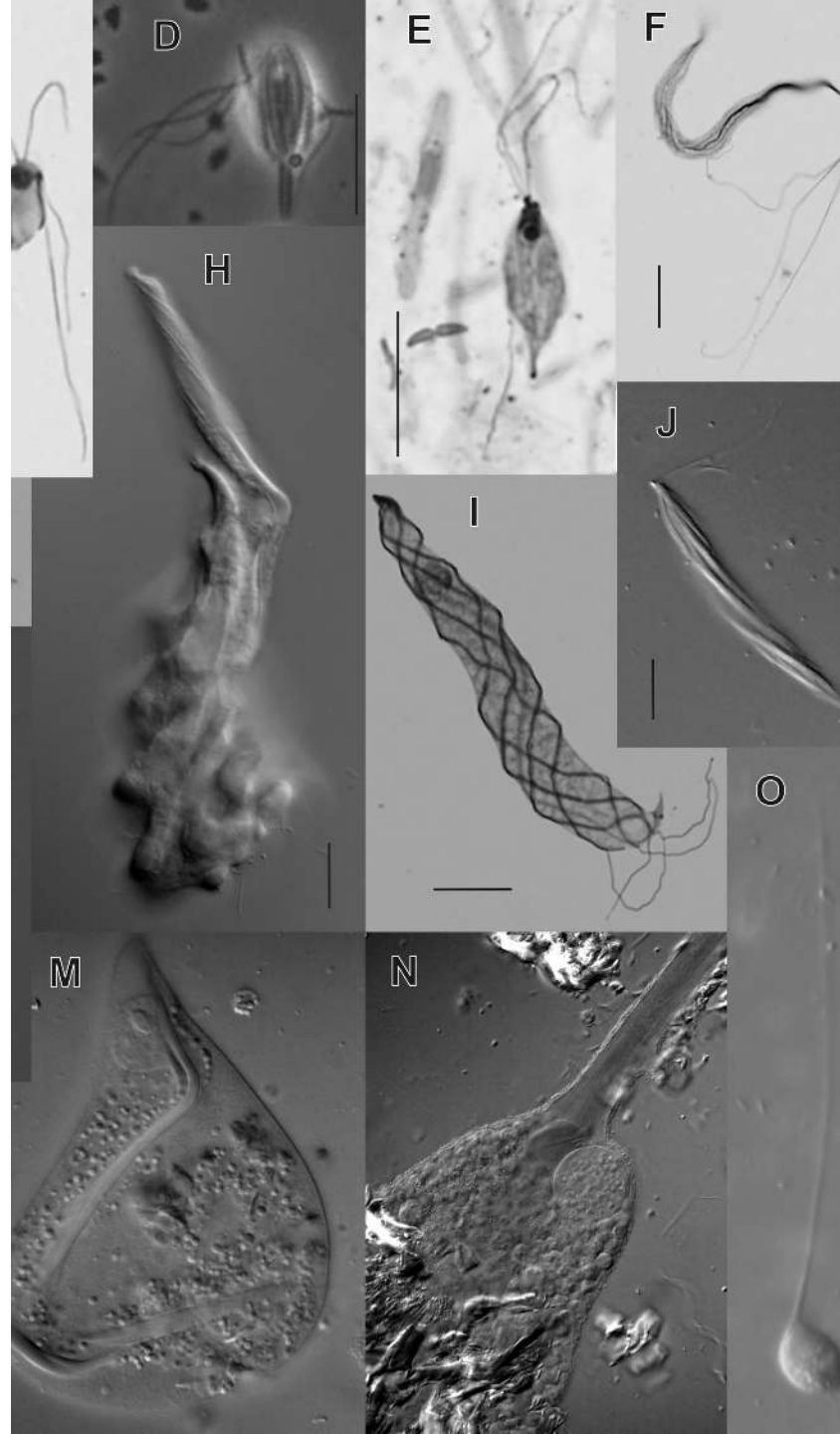
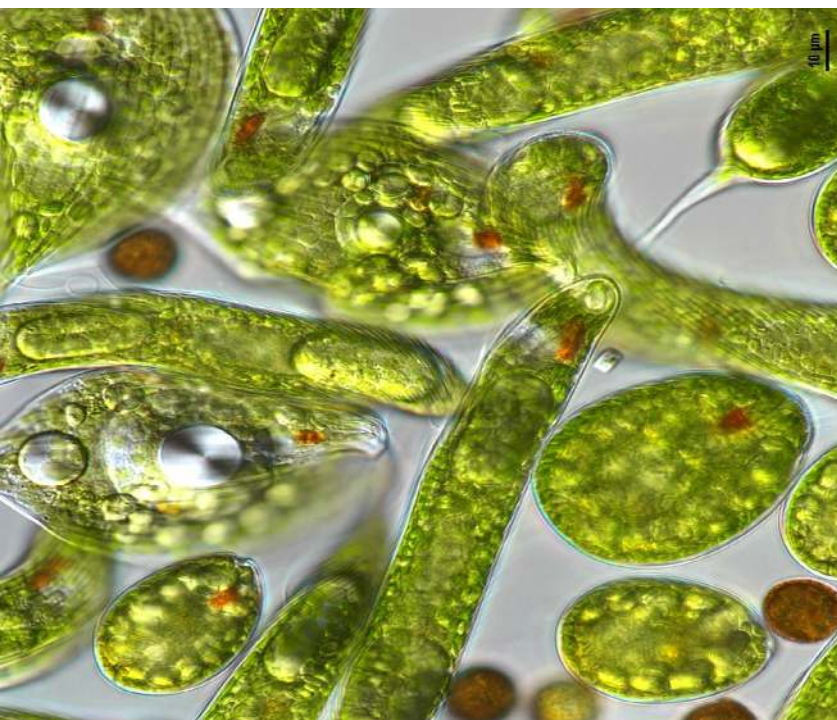
Eukaryotes subdivided into supergroups



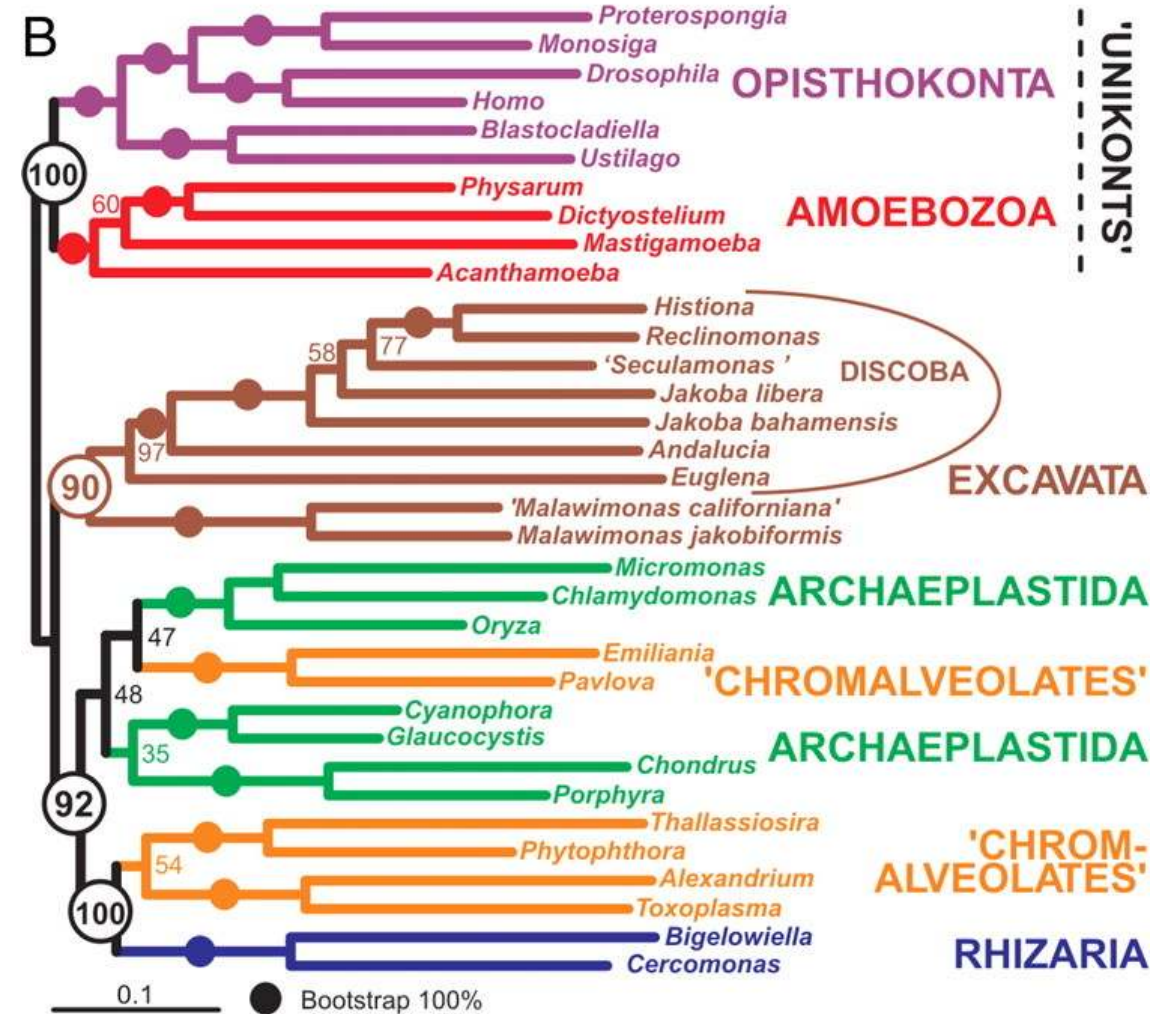
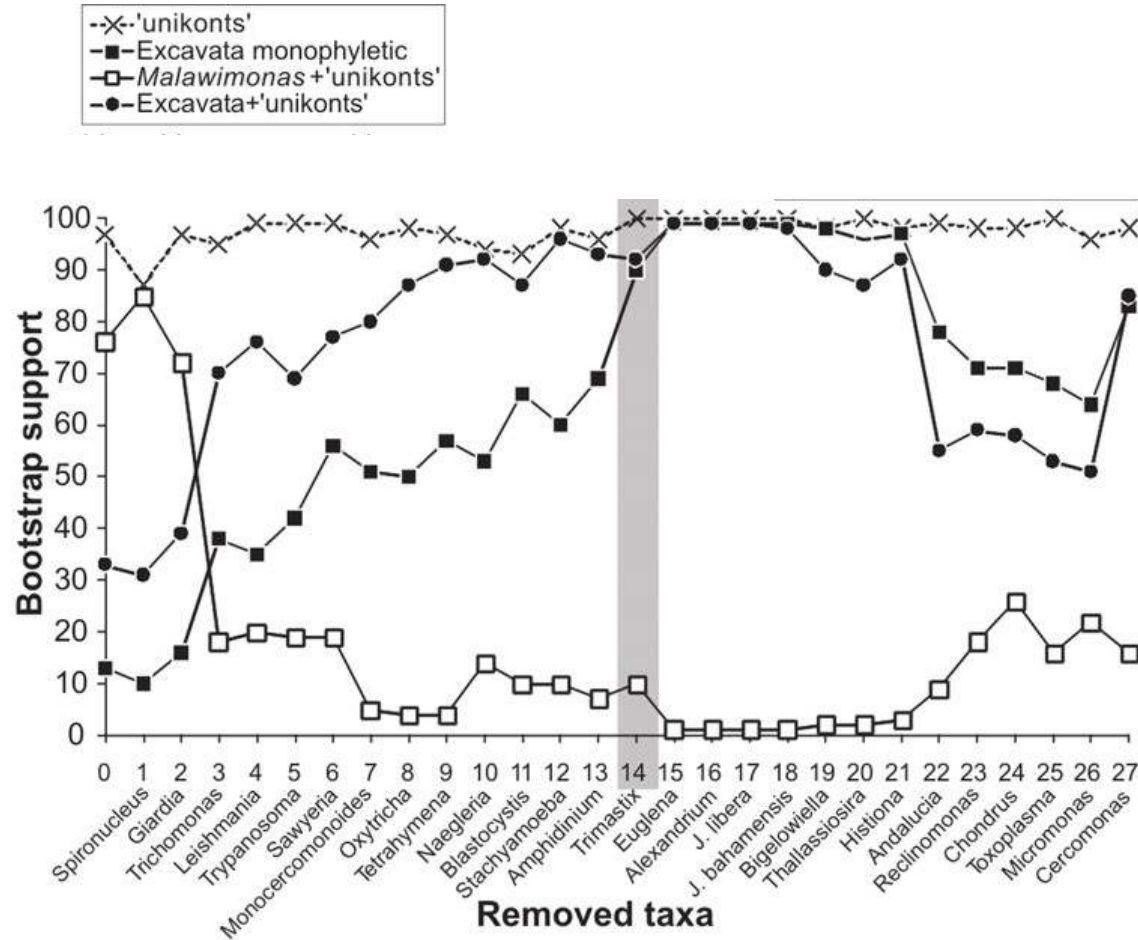
Holy Grail of "Excavata"



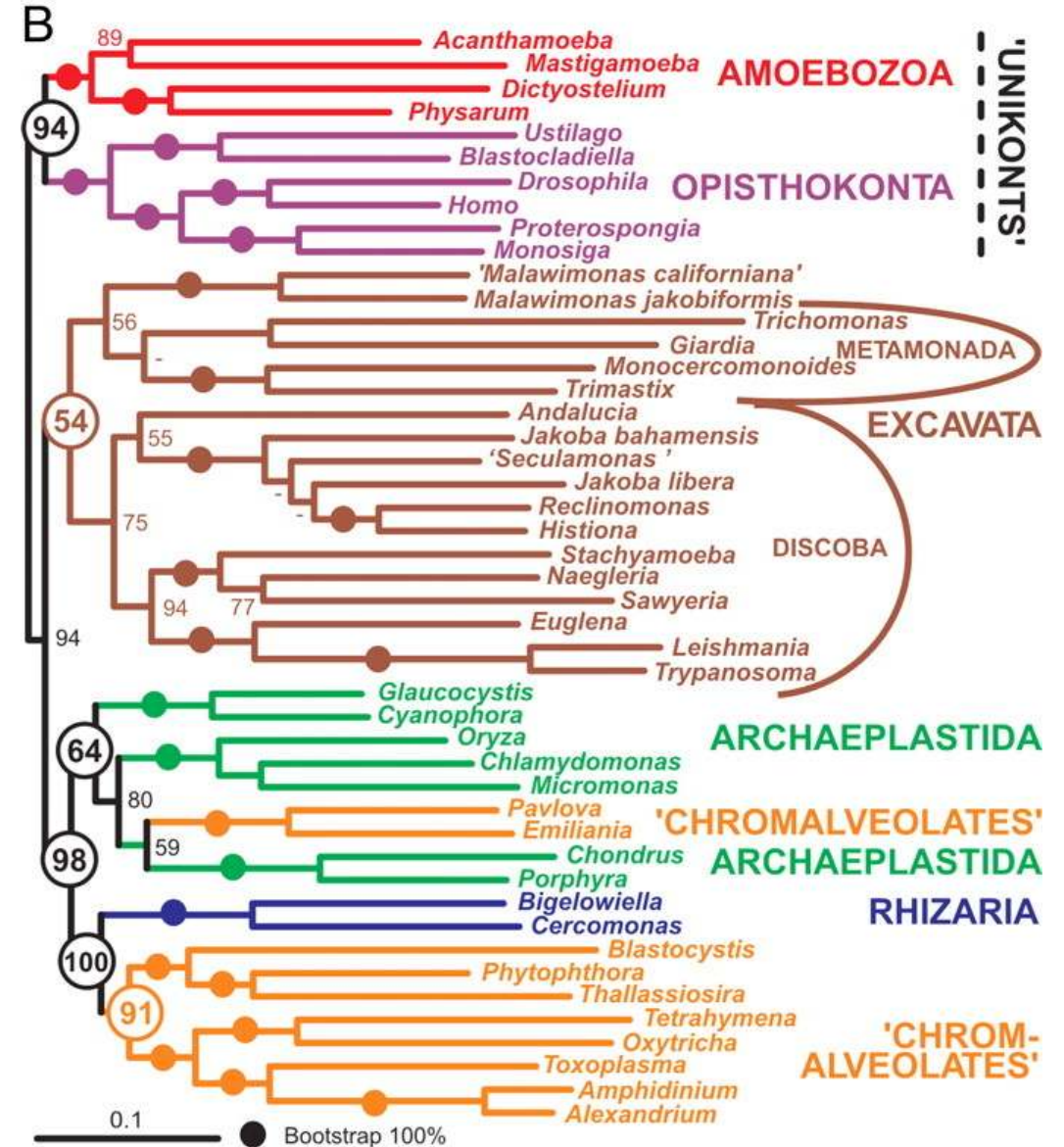
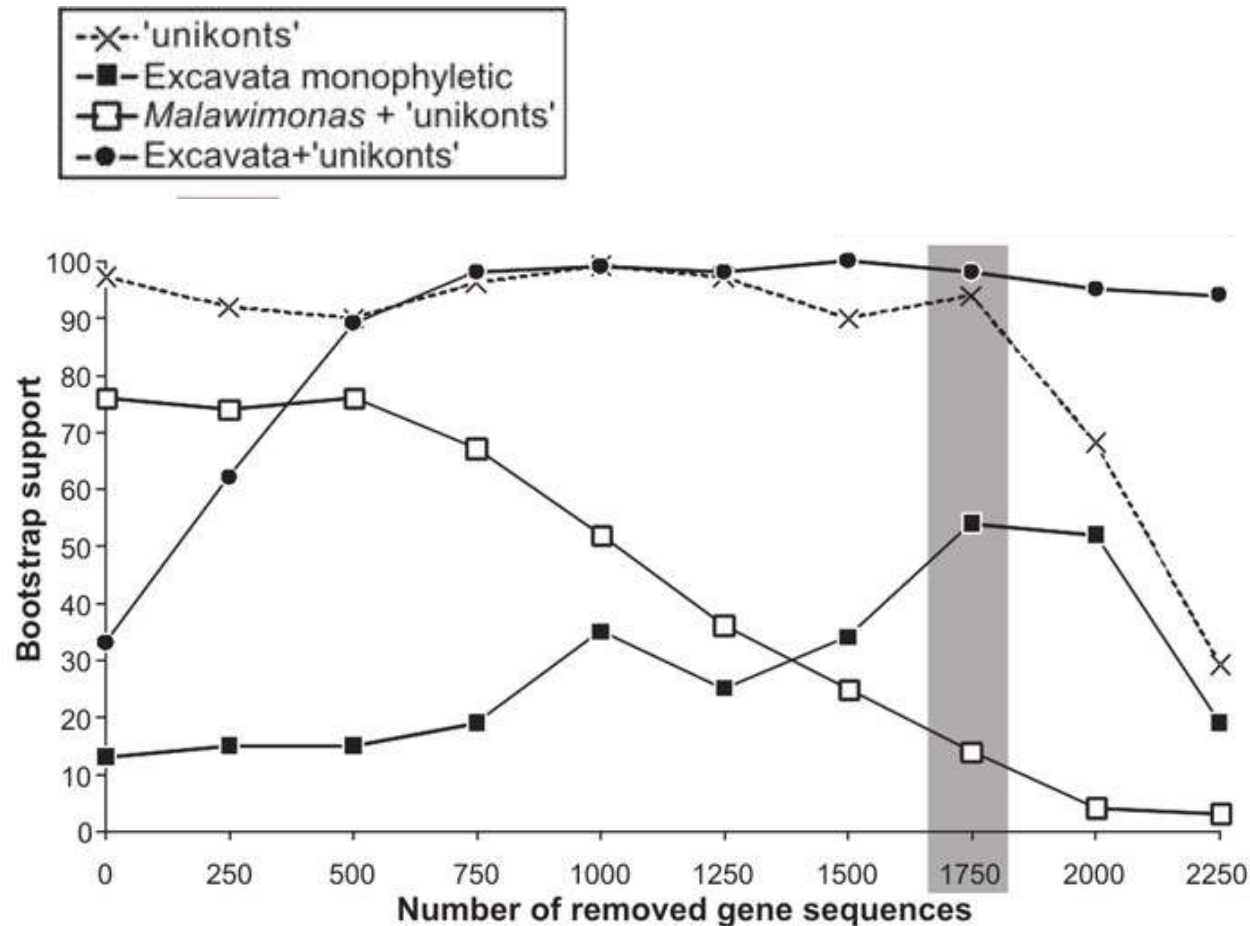
- 143 gene supermatrix
- maximum likelihood method, using RAxML under WAG+ Γ model
- Bayesian analyses, implemented in PhyloBayes version 2.1c employing the CAT+ Γ model



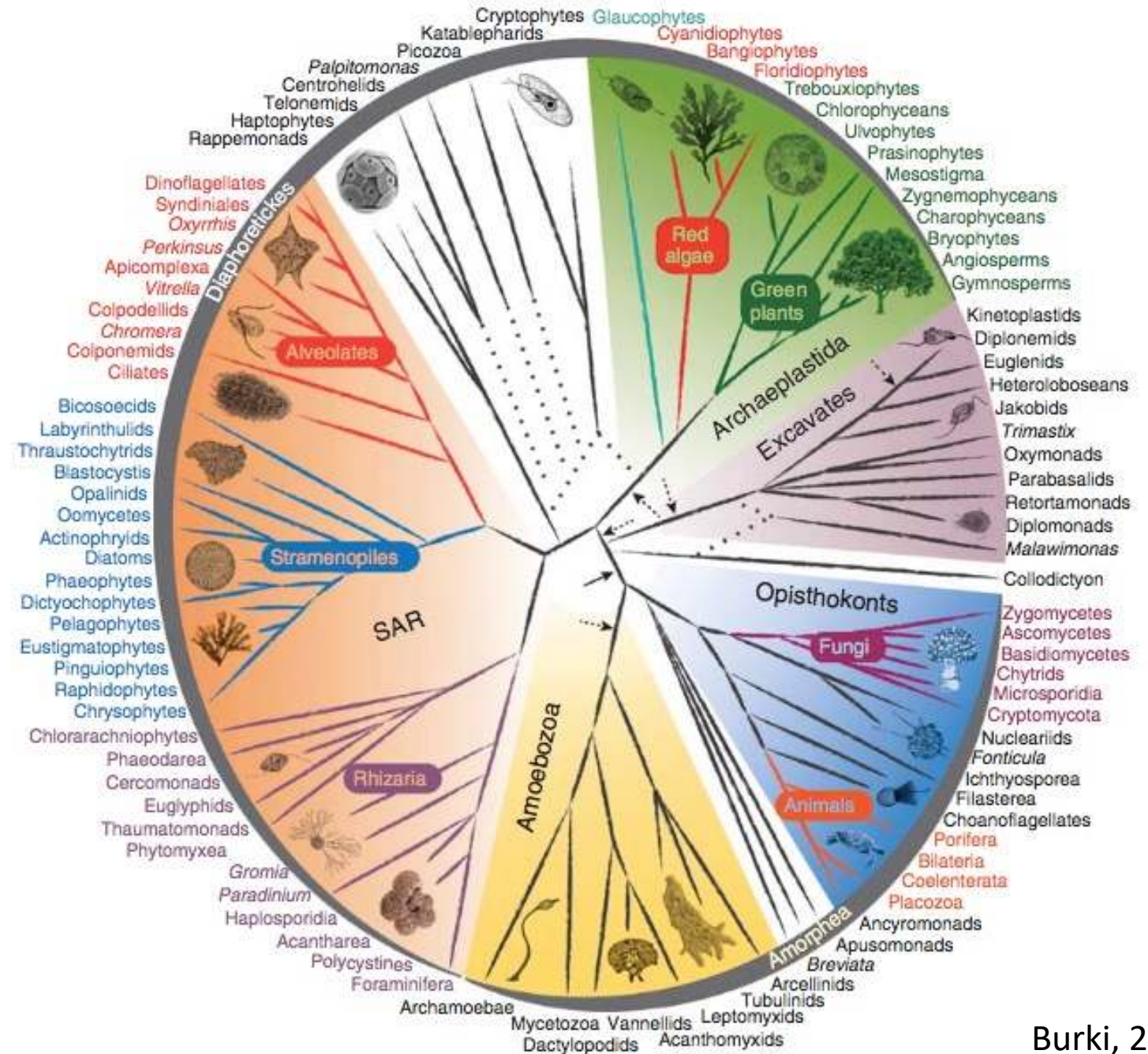
Removal of long-branch taxa



Removal of long-branch gene sequences



- Many of the general biases that apply to single-gene phylogenetics are exacerbated in a phylogenomic context
- The goal of a phylogenetic inference is to minimize the noise and maximize the true phylogenetic signal



More data & errors

- Saturated positions in the data set (stochastic errors) - arise when the number of positions in an alignment is small and the random background noise will have a neutralizing effect on the positions that contain the phylogenetic signal.

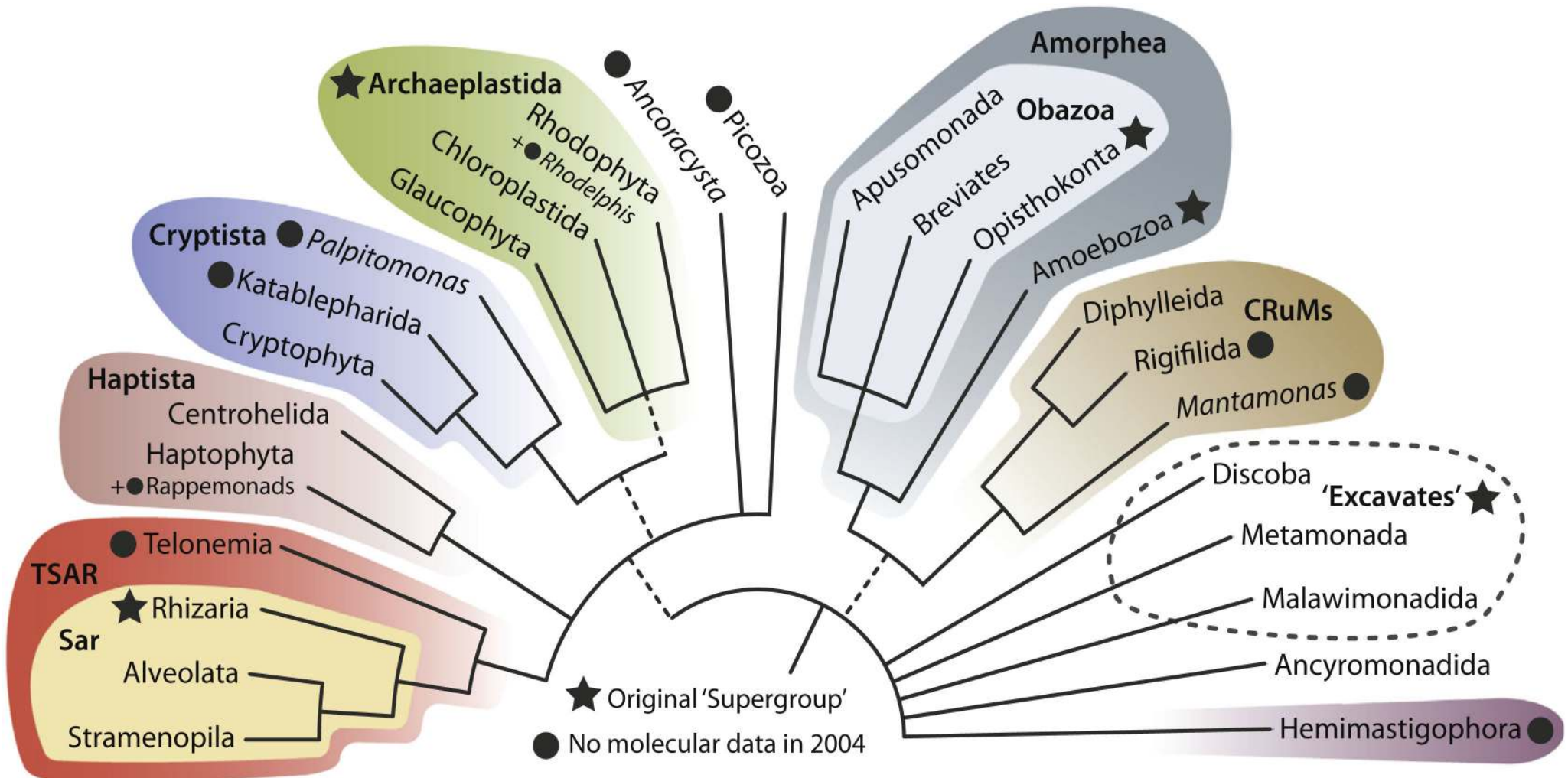
More data should have **more phylogenetic signal**

- Model misspecifications (systematic errors) - heterogeneity of nucleotide or amino acid composition, which tends to incorrectly cluster together species sharing the same composition, or the heterogeneity of the evolutionary rates, which can result in the LBA artefact.

More data **do not solve systematic errors**, yet since more data are available strategies to diminish the known sources of systematic errors can be applied

With more data there is a risk of amplification of errors!

Eukaryotic phylogeny based on genomic and transcriptomic data

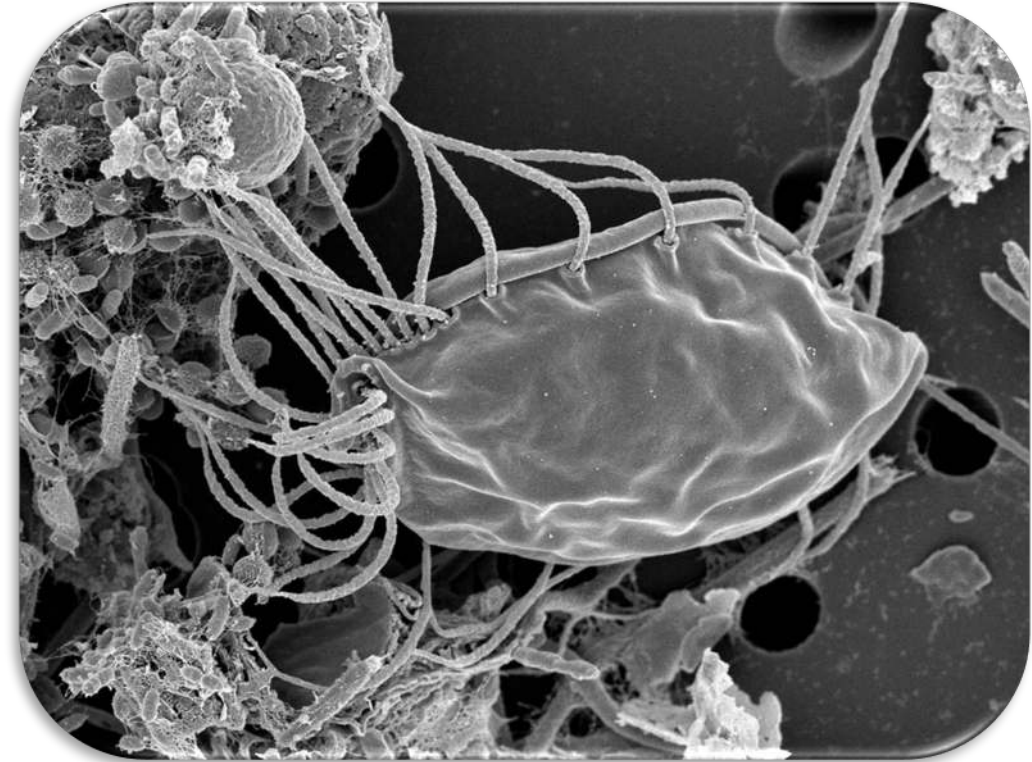
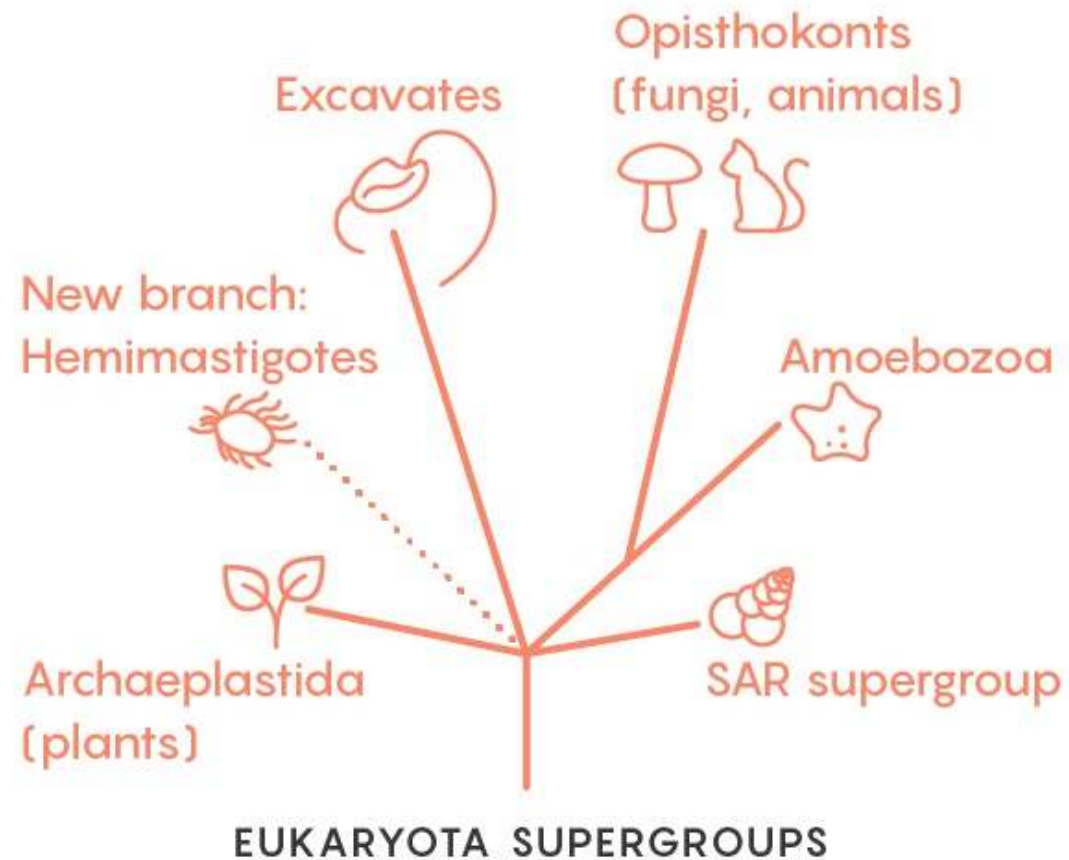
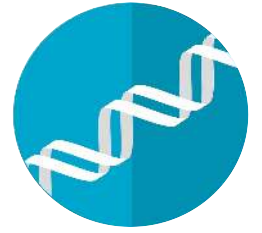


Trends in Ecology & Evolution

Typical pipeline

1. Inferred amino acid sequences of proteins retrieved from genomic or transcriptomic data
2. Adding sequences from new taxa to the well-curated sets of orthologous genes
3. Single-gene trees analysis to identify lateral gene transfers, incorrect paralog selections, and various contaminants (still mainly manual).
4. A subset of taxa is selected for the actual analysis to evenly cover the relevant phylogenetic breadth while excluding problematic species (e.g., those with limited data, extreme evolutionary rates in many genes, etc.).
5. Genes are concatenated into a phylogenomic 'supermatrix'.
6. Usually, both ML and Bayesian analyses are conducted. Various evolutionary models are employed (site-heterogeneous models), with choice often constrained by computational logistics.
7. Testing whether results are robust to perturbations of the data, especially excluding data most likely to foster incorrect phylogenetic inference (e.g., the fastest-evolving species, sites, or genes).

Discoveries of new evolutionary lineages

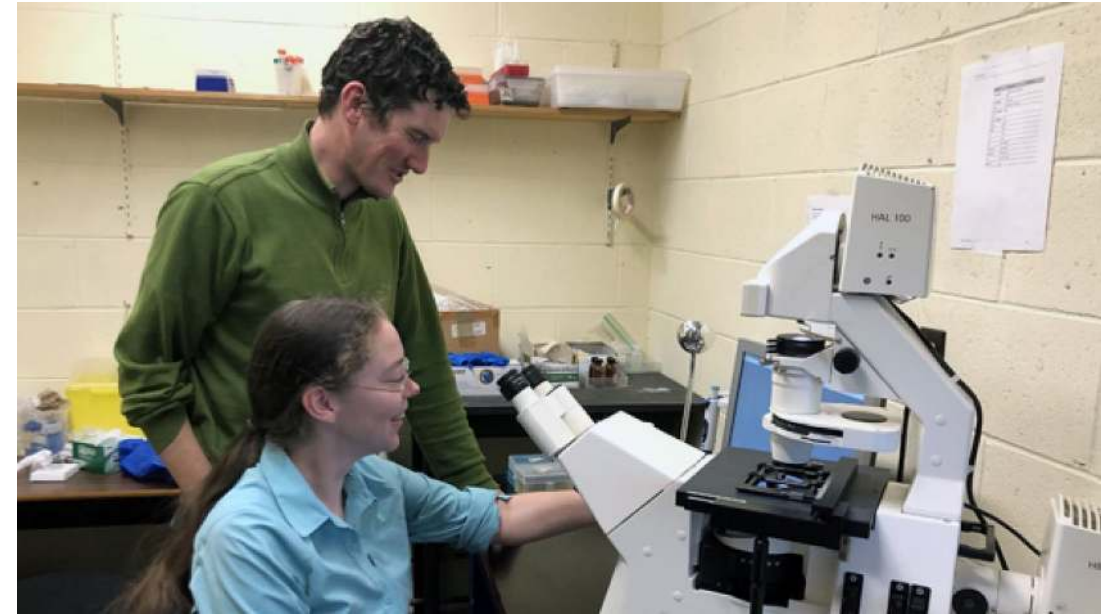


Hemimastix kukwesjijk

Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes

[Gordon Lax](#), [Yana Eglit](#), [Laura Eme](#), [Erin M. Bertrand](#), [Andrew J. Roger](#) & [Alastair G. B. Simpson](#) 

[Nature](#) **564**, 410–414 (2018) | [Cite this article](#)

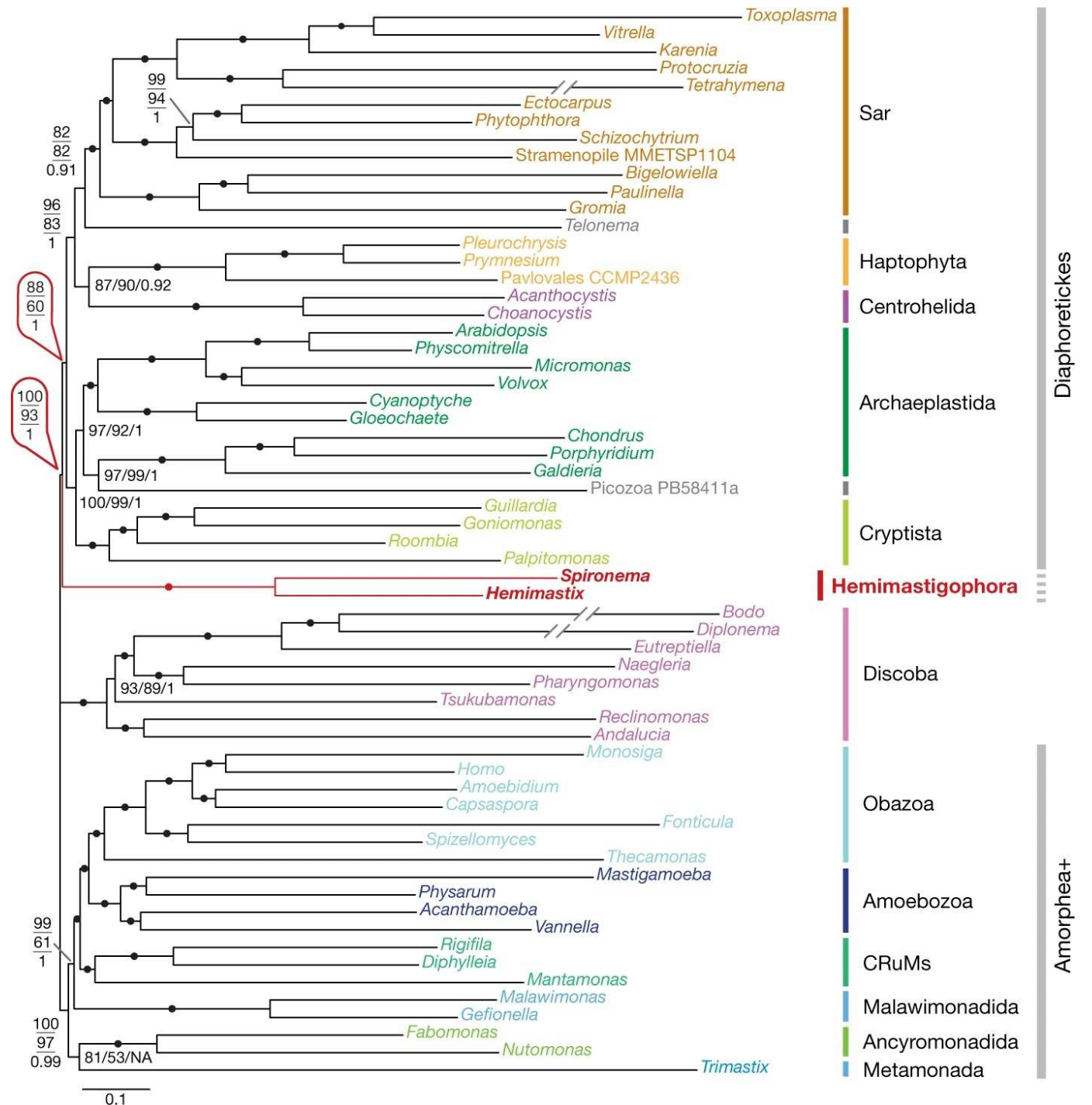


*It was a cold April day on a local trail in Nova Scotia during a hike. **Sporadic sampling is a known professional hazard in protistology**, since all we really need for an explosion of interesting things to look at is a few millilitres of water or dirt.*

*Yana soaked the soil in distilled water for about a month, periodically observing what was living there'. One evening, small, ~20 μm long, **lanceolate cells showed up that behaved not quite like any 'common' flagellate.***

Hemimastigophora is placed outside of all established eukaryote supergroups

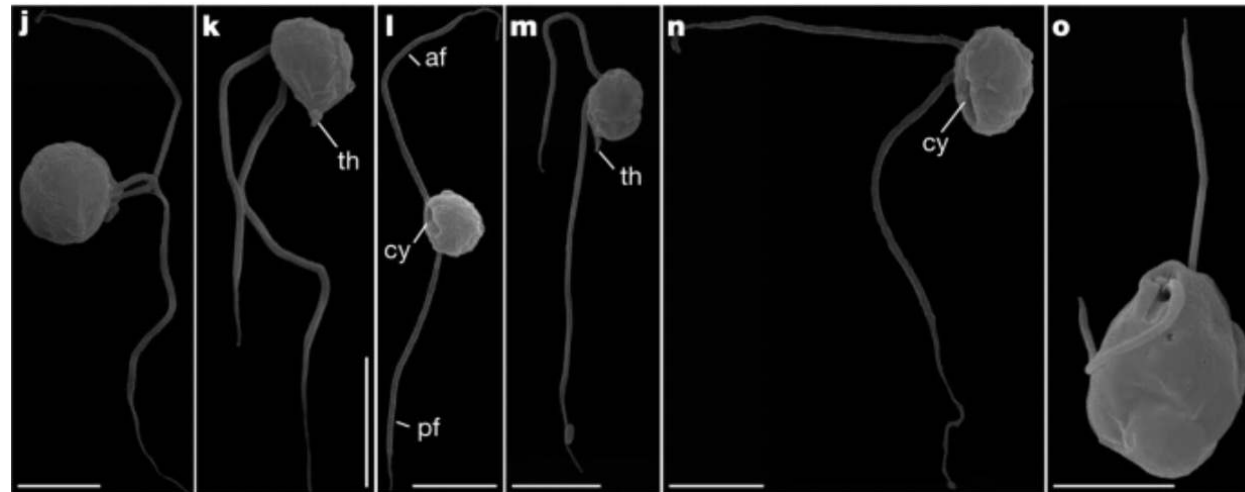
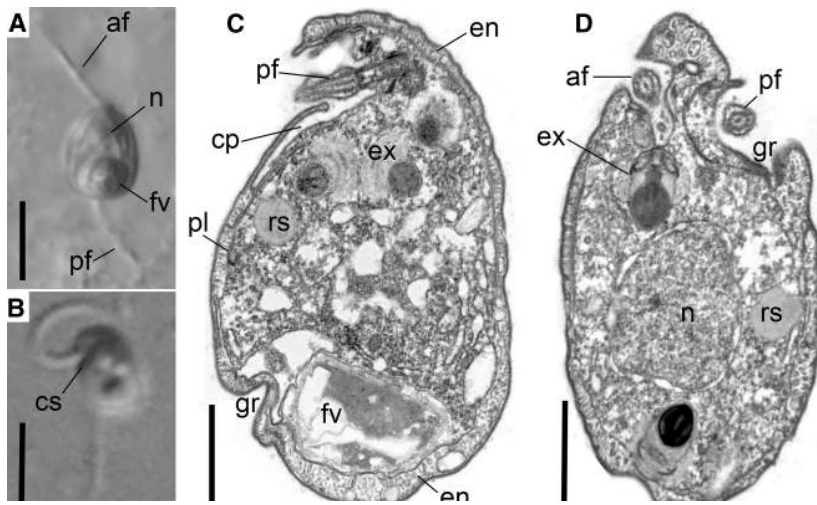
Unrooted phylogeny inferred from 351 genes and 61 taxa. Most of data comes from transcriptomes, *Hemimastix* and *Spironema* data from single cells!



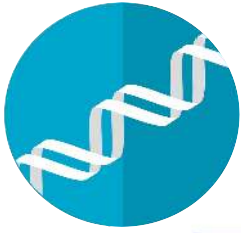
Microbial predators form a new supergroup of eukaryotes

[Denis V. Tikhonenkov](#) , [Kirill V. Mikhailov](#), [Ryan M. R. Gawryluk](#), [Artem O. Belyaev](#), [Varsha Mathur](#), [Sergey A. Karpov](#), [Dmitry G. Zagumyonnyi](#), [Anastasia S. Borodina](#), [Kristina I. Prokina](#), [Alexander P. Mylnikov](#), [Vladimir V. Aleoshin](#) & [Patrick J. Keeling](#)

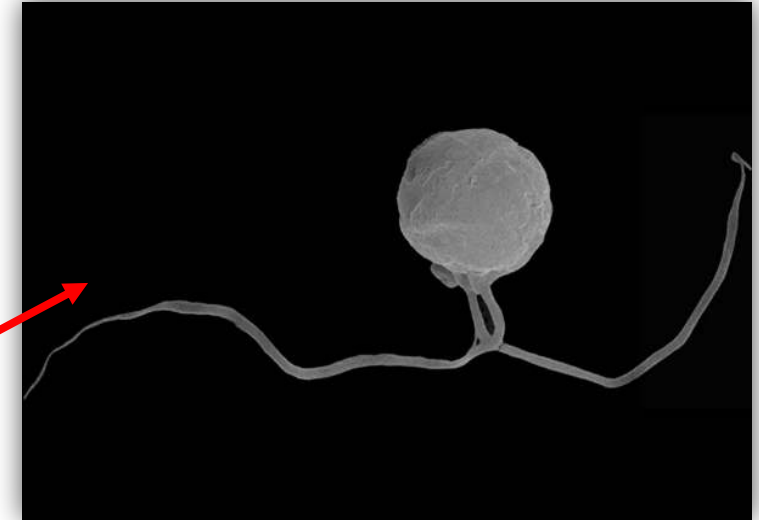
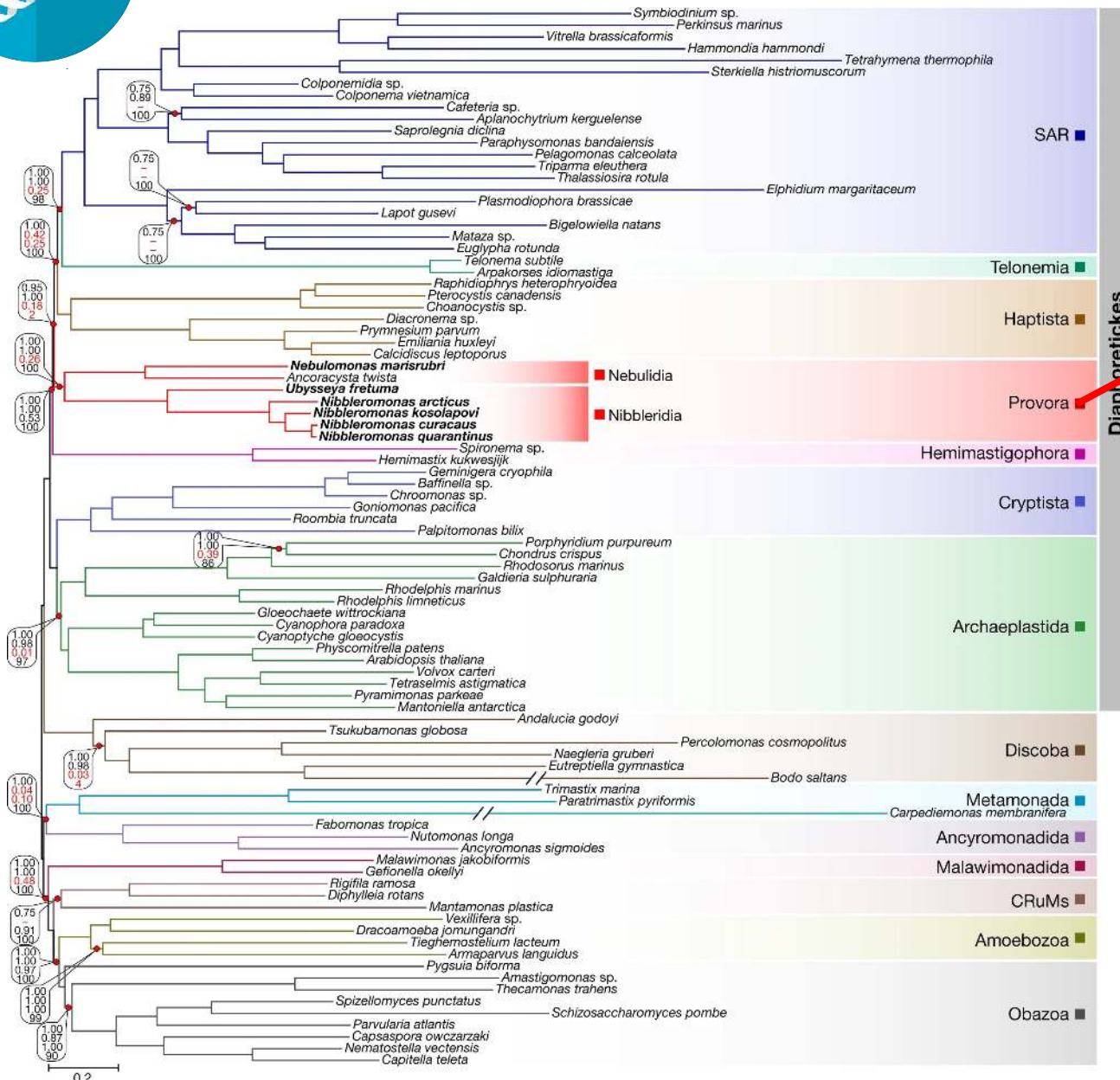
[Nature](#) **612**, 714–719 (2022) | [Cite this article](#)



“They have different behaviours. One group eats its prey. It literally just swallows cells whole. The other has a tooth — they’ll gang up in a swarm and nibble it to death” – I prefer to call them nibblers and lions



Provora at the tree of eukaryotes



320 orthologous gene groups for 69 species

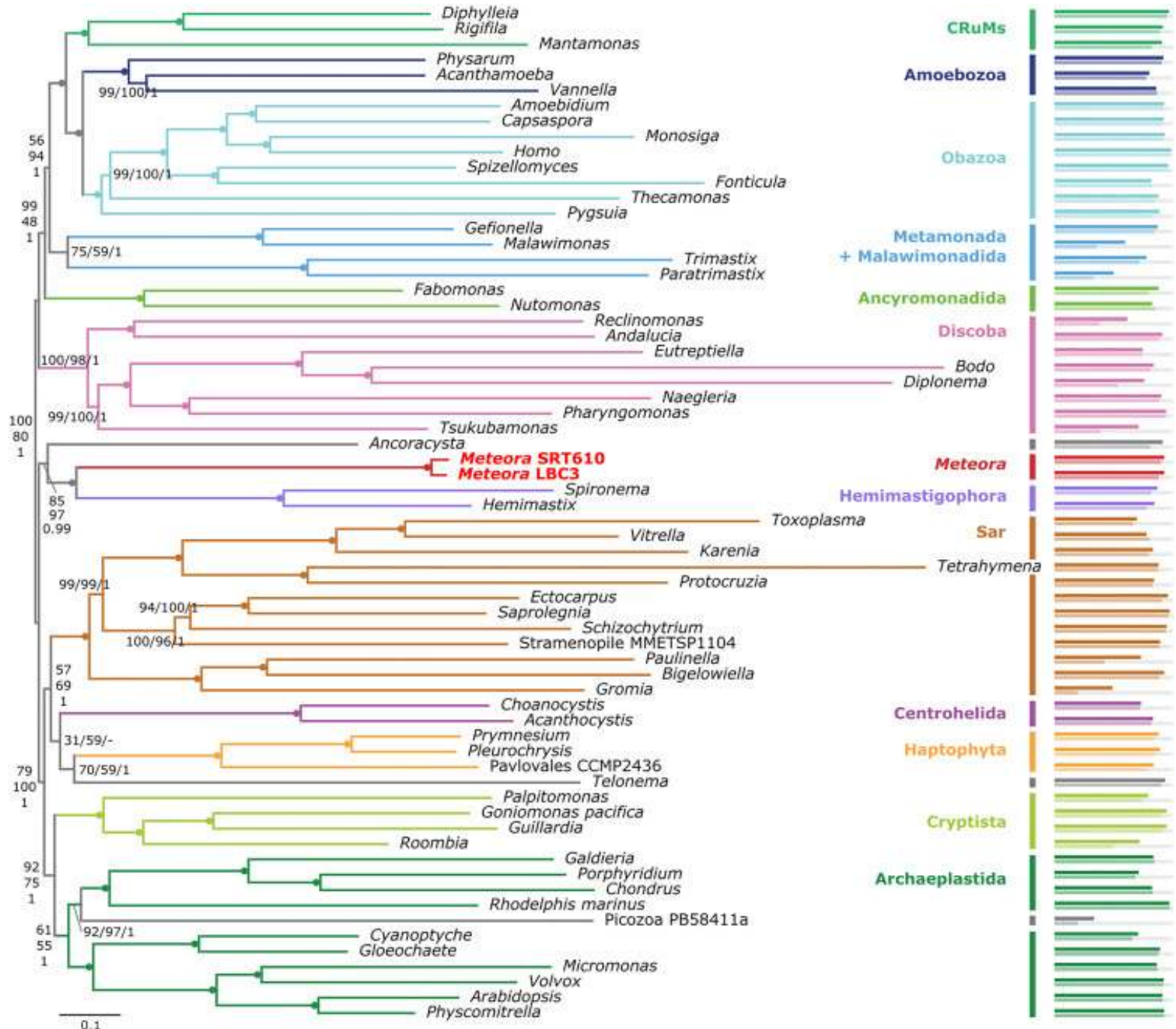
Bayesian inference approach implemented in PhyloBayes and the maximum-likelihood approach of IQ-TREE.

Anything new in 2024?

Meteora!

Maximum likelihood phylogeny
70,471 sites across 254 genes over
66 taxa LG + MAM60 + Γ model

Bars on the right indicate % coverage
by gene (above) and by site (below).



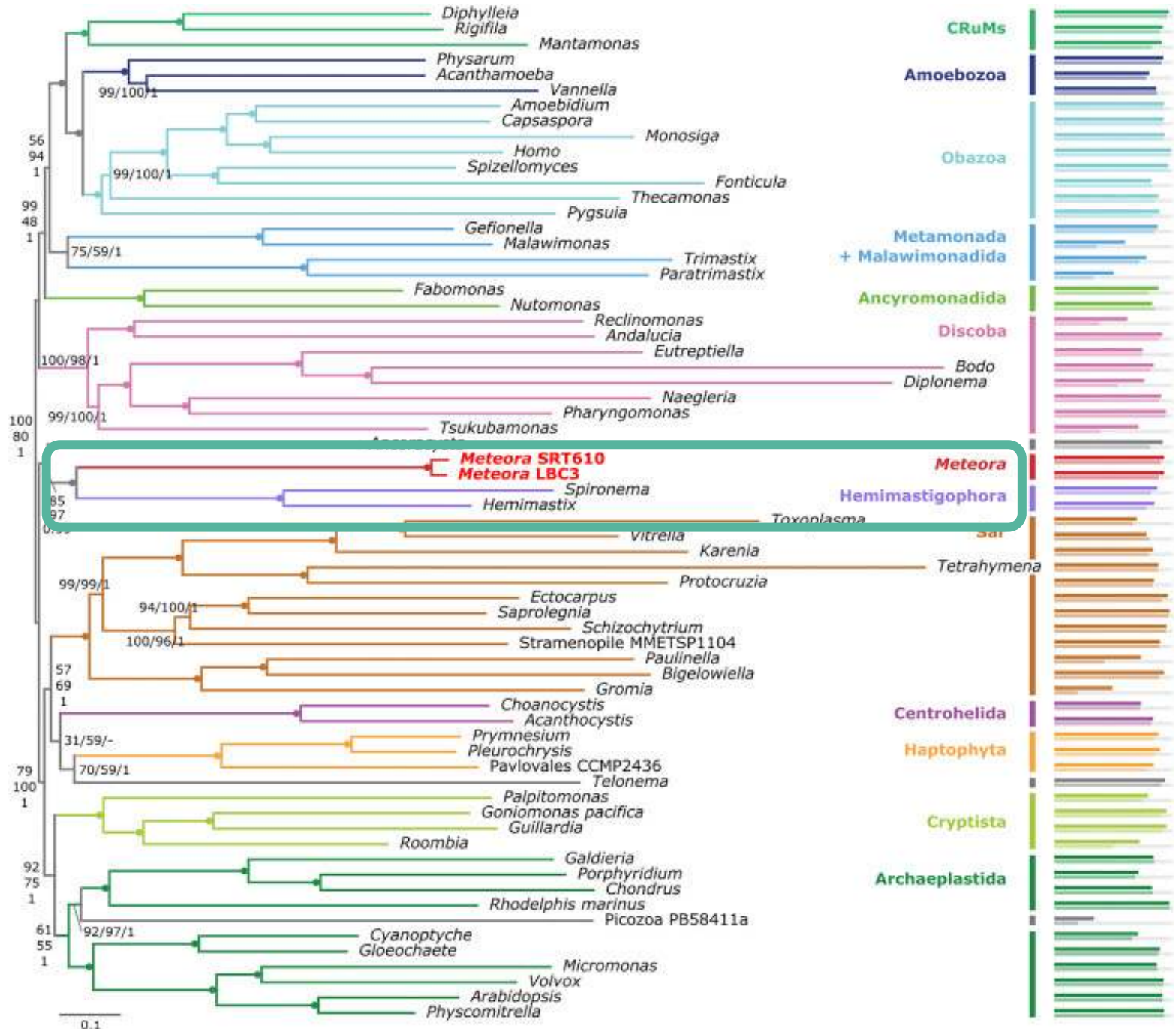


Anything new in 2024?

Meteora!

Maximum likelihood phylogeny
70,471 sites across 254 genes over
66 taxa LG + MAM60 + Γ model

Bars on the right indicate % coverage
by gene (above) and by site (below).



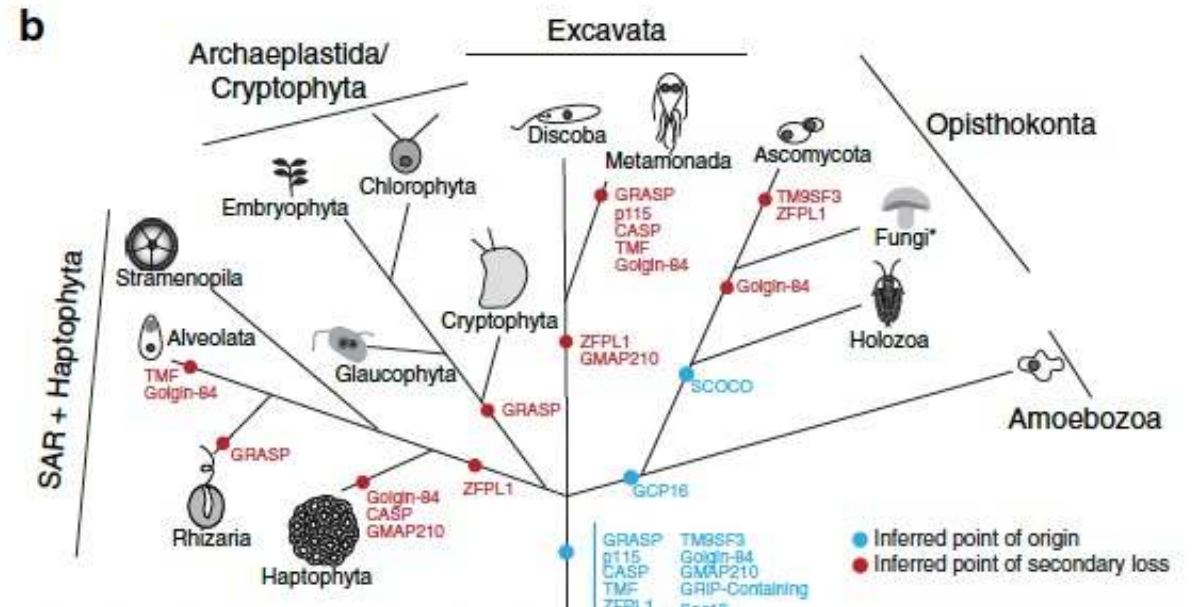
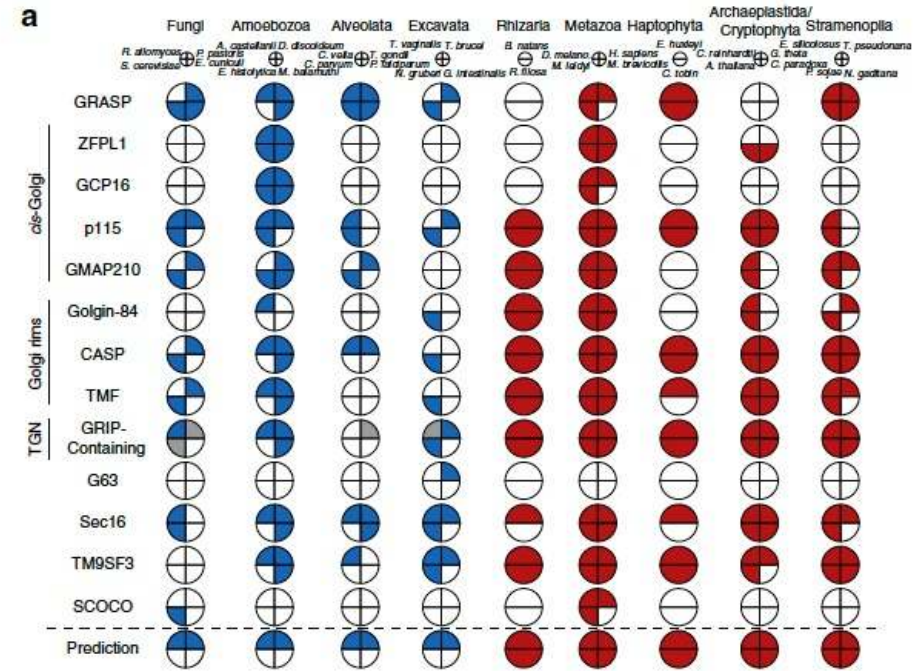
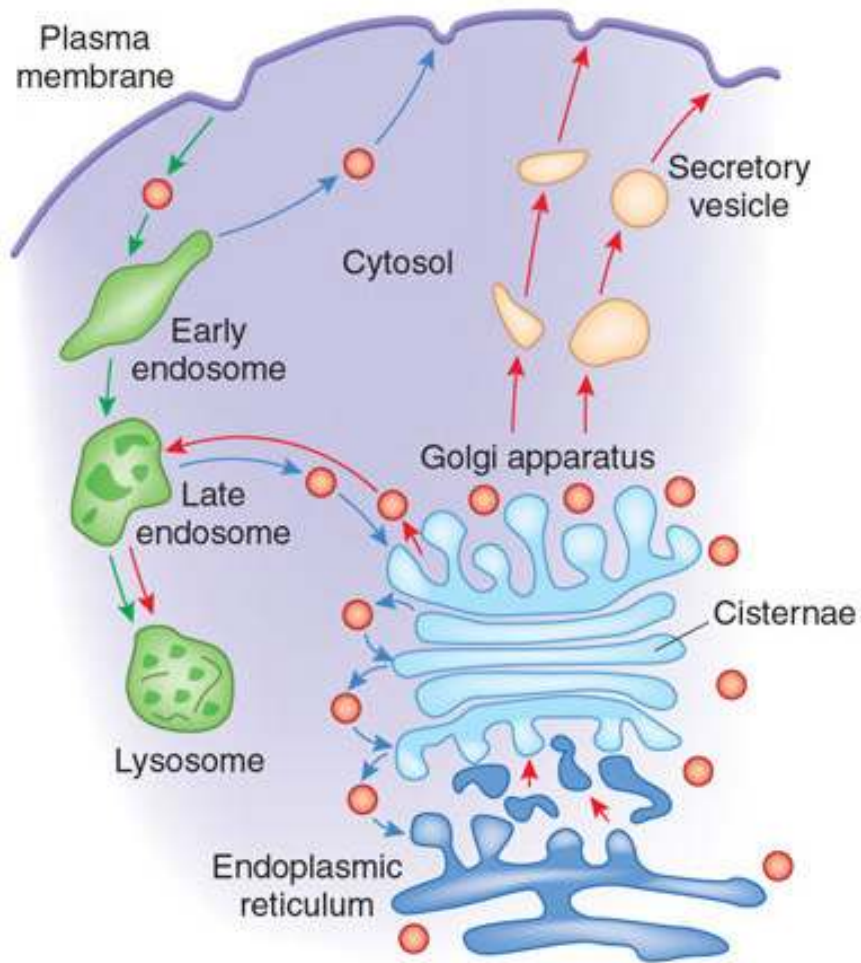
Validating robustness of the *Meteora* + Hemimastigophora clade

1. An analysis that excluded three taxa identified as long-branching outliers -> maximal support for the *Meteora* + Hemimastigophora clade
2. Recoding the amino acid data into a reduced alphabet of 4 classes -> robustly supported *Meteora* + Hemimastigophora
3. Removal of the fastest evolving sites in 10% increments (fast-site removal analysis) -> *Meteora* + Hemimastigophora as maximally supported until 30% sites remaining
4. Random subsampling of 50% of the genes -> robust support for *Meteora* + Hemimastigophora. The phylogenetic signal for the *Meteora* + Hemimastigophora relationship is broadly distributed across genes.

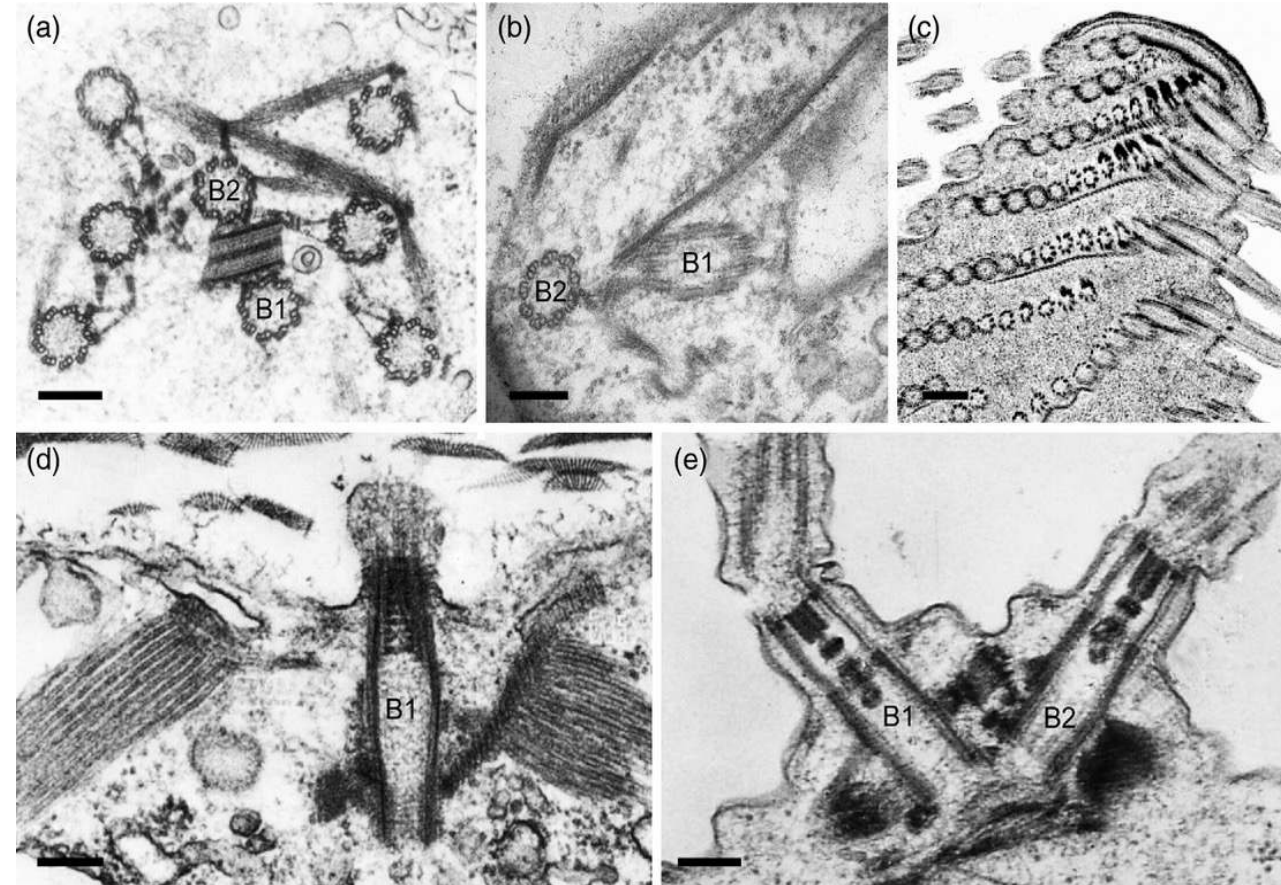
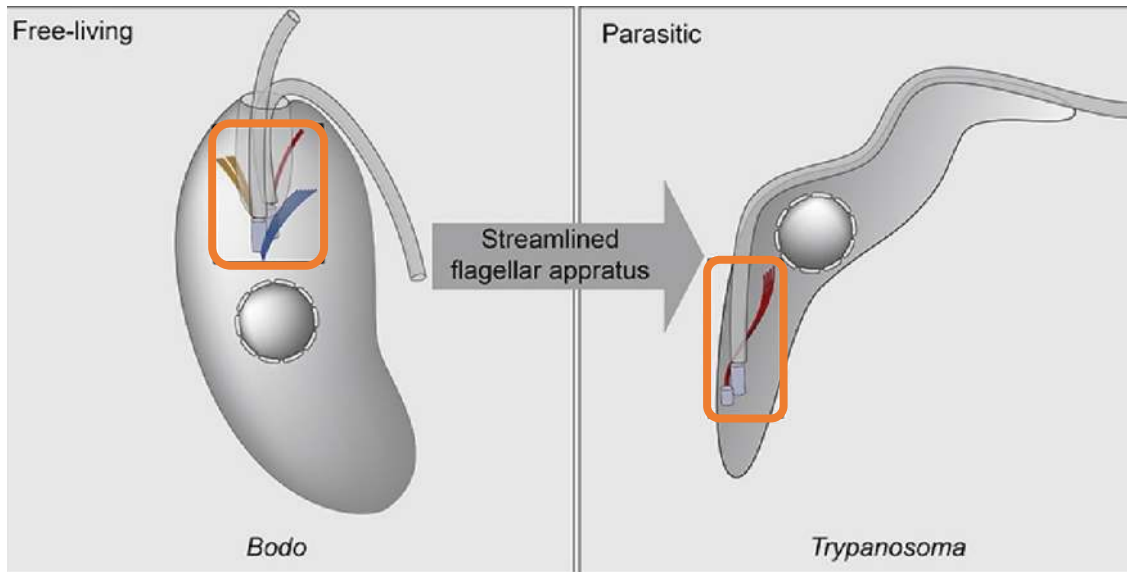
Why do we need well resolved tree of eukaryotes?

Evolution of cellular systems

Intracellular transport

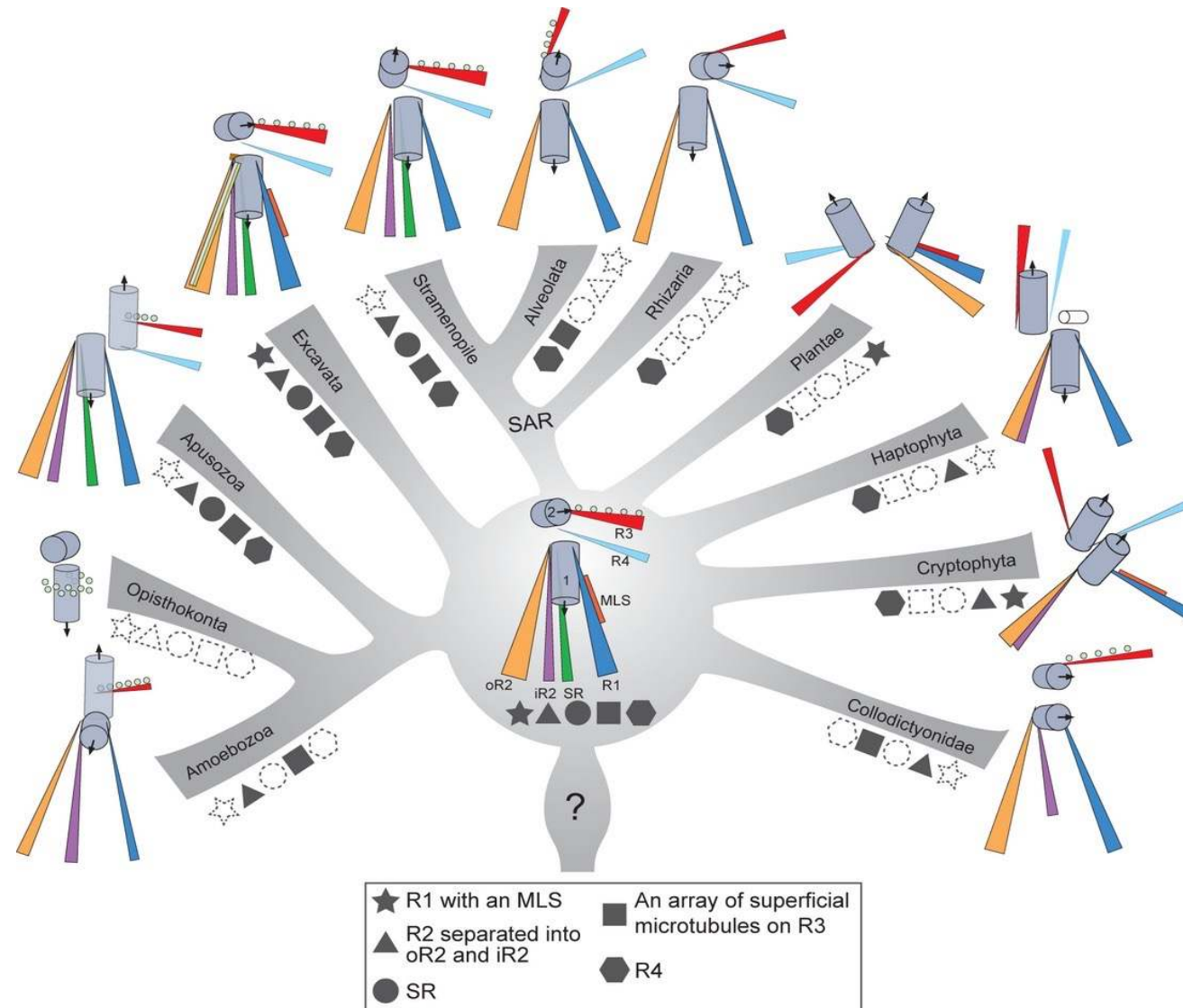
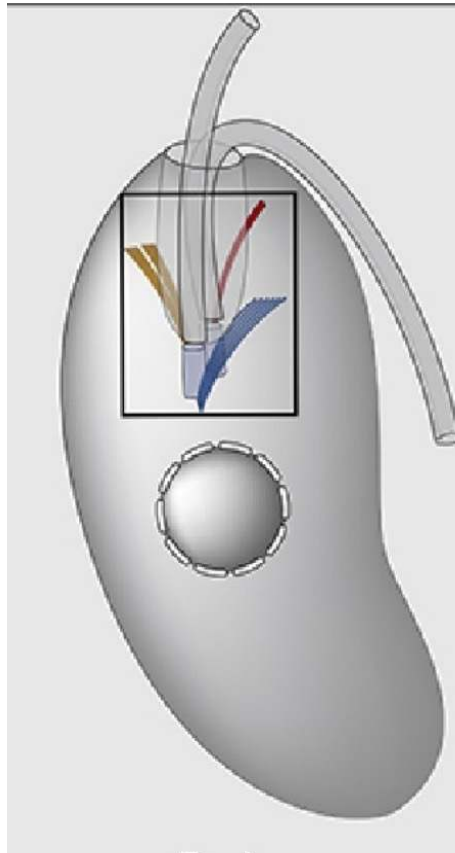


Flagellar apparatus



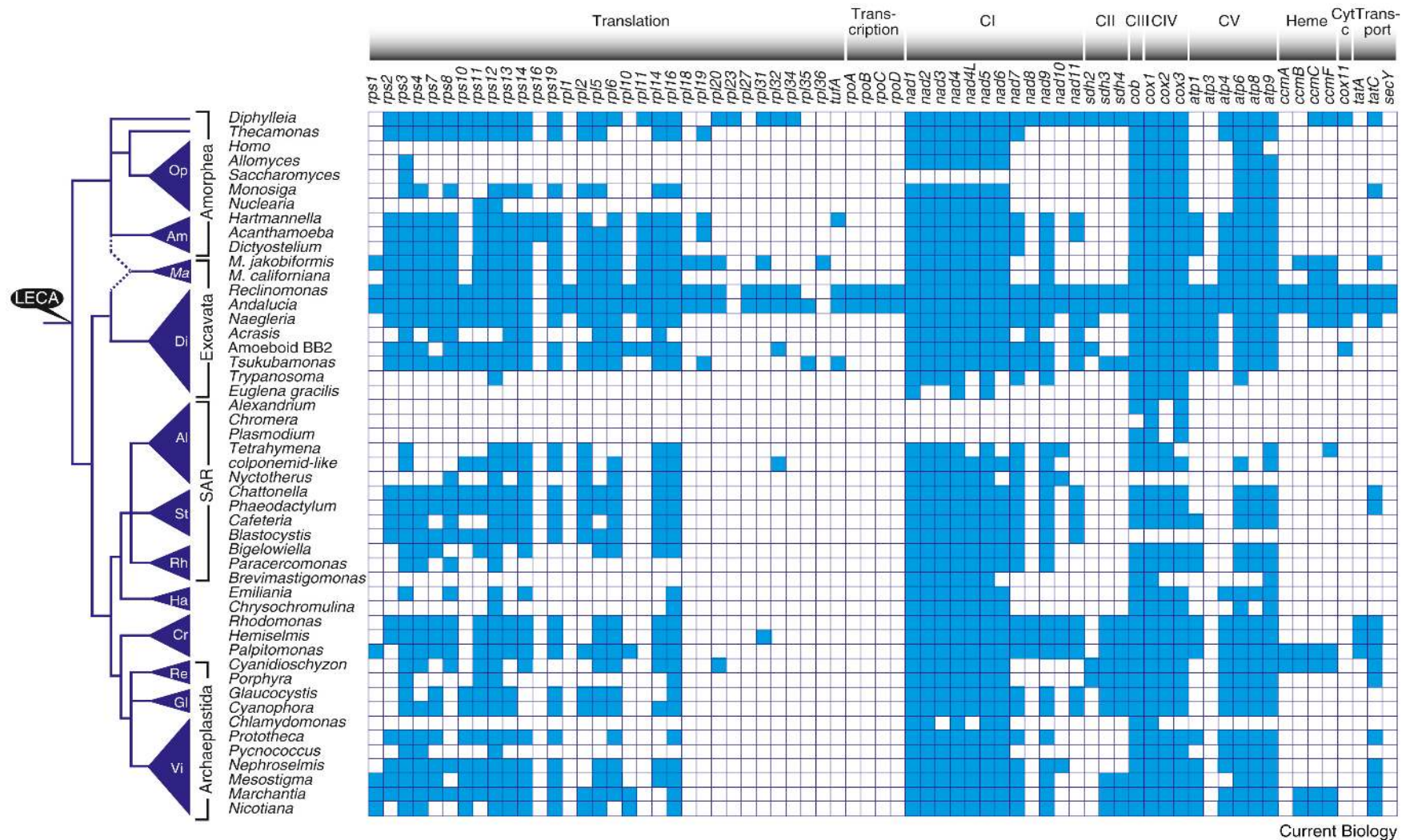
Transmission electron microscope micrograph

Reconstruction of the cytoskeleton supporting the flagellar apparatus



Evolution of organelles

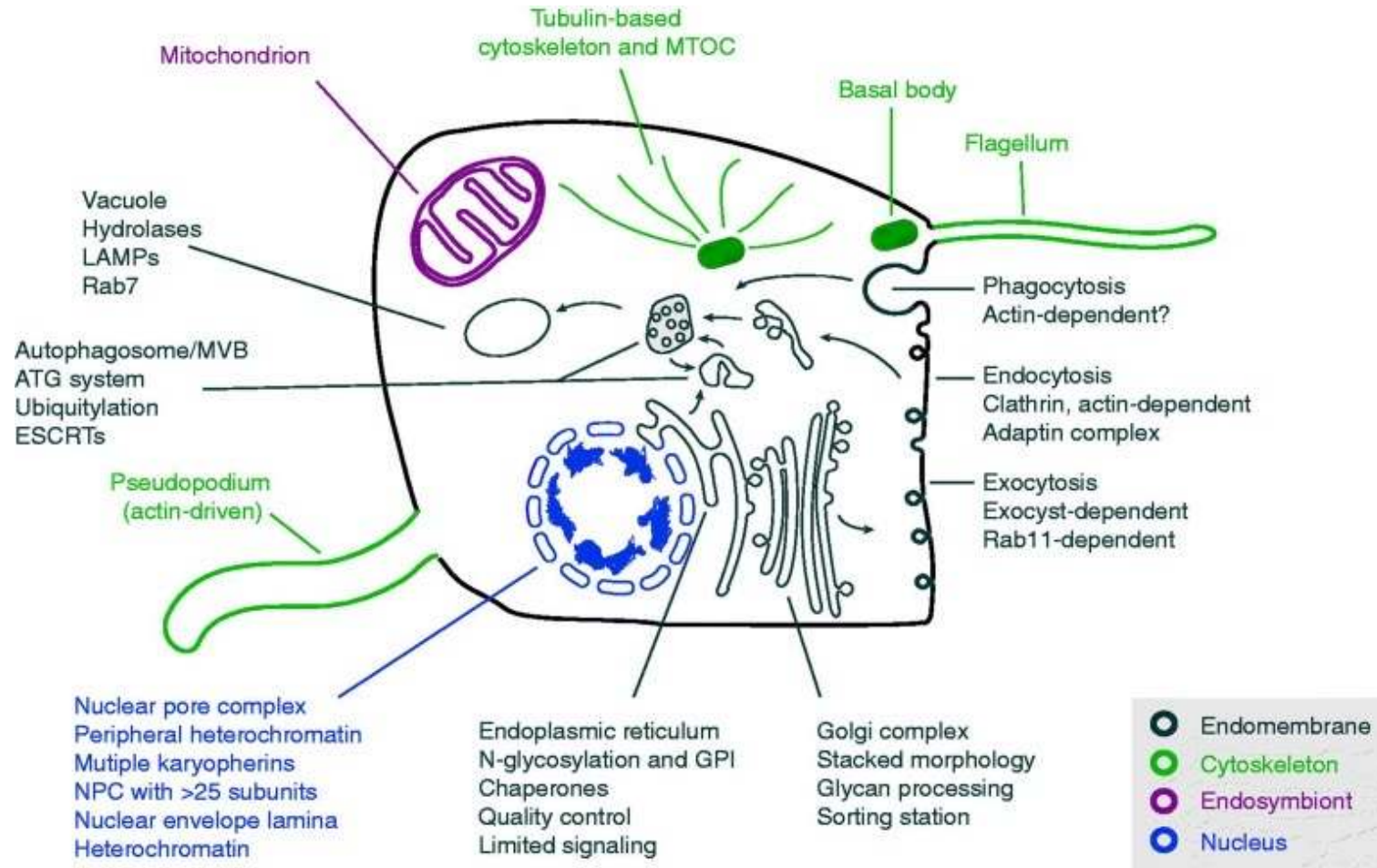
gene content of mitochondrial genome



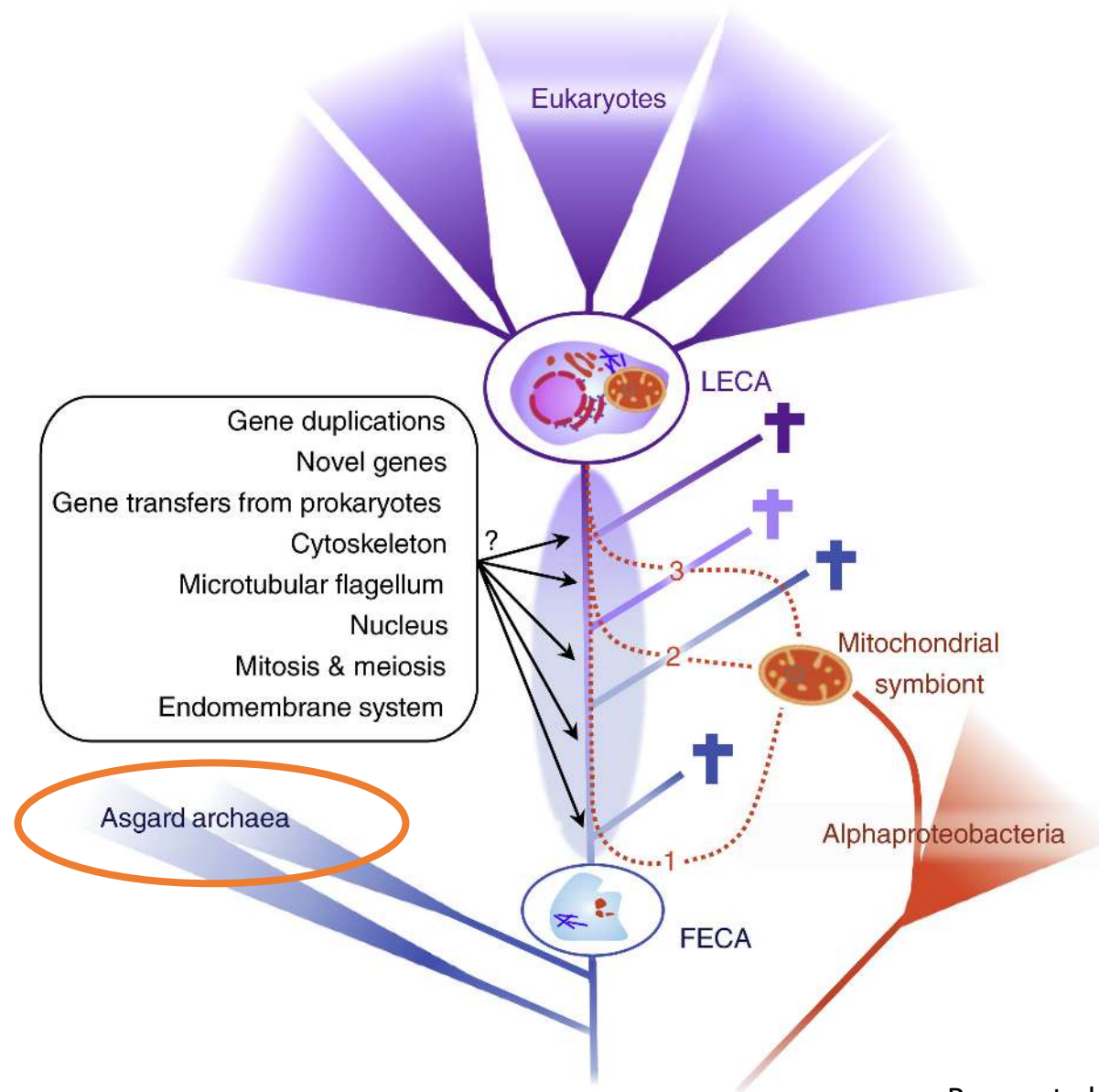
Current Biology

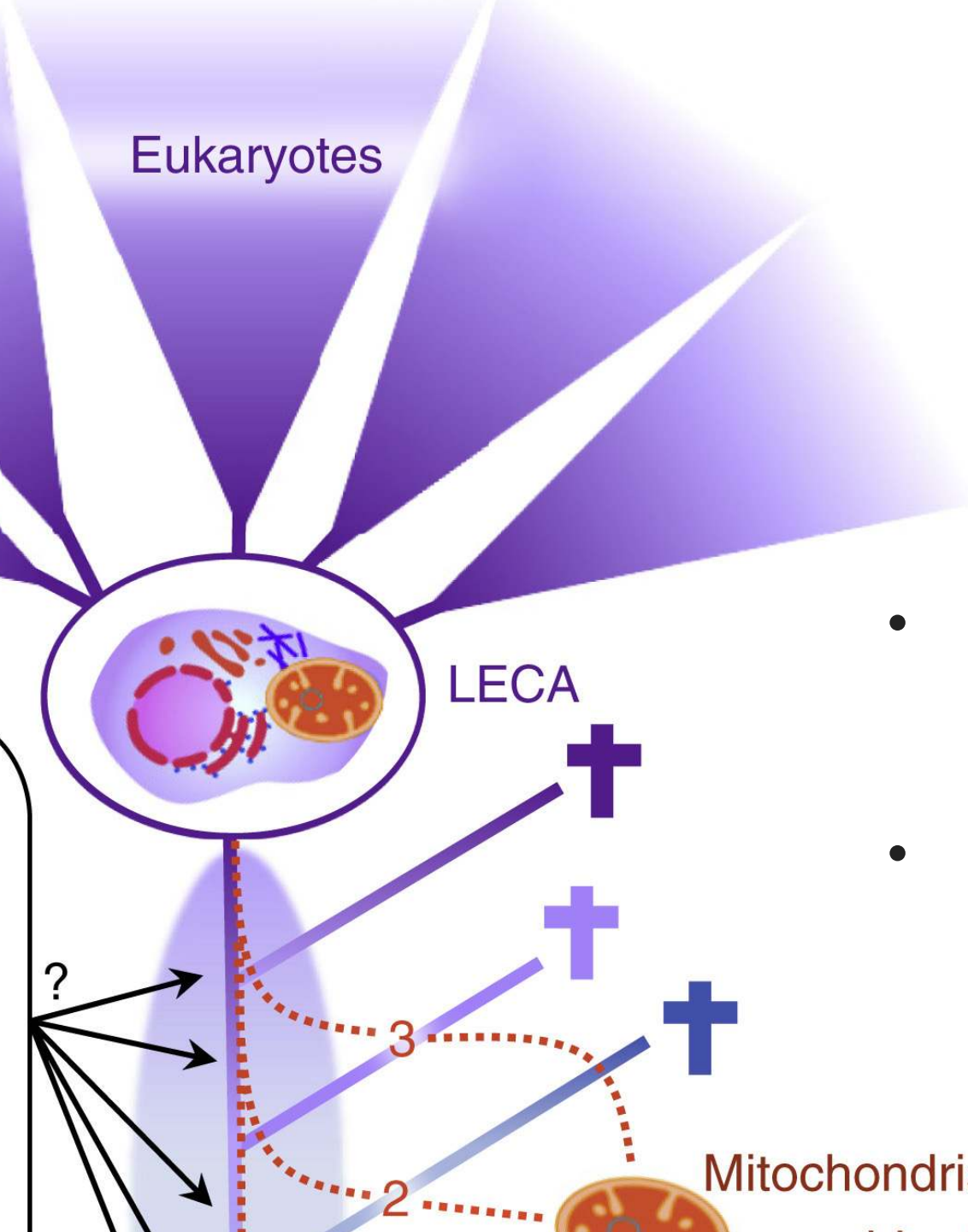
Presence or absence of genes in mitochondrial genomes of various eukaryotes is shown by closed and open boxes, respectively.

With more data and well resolved tree of eukaryotes we've learned that the last common ancestor was complex eukaryotic cell



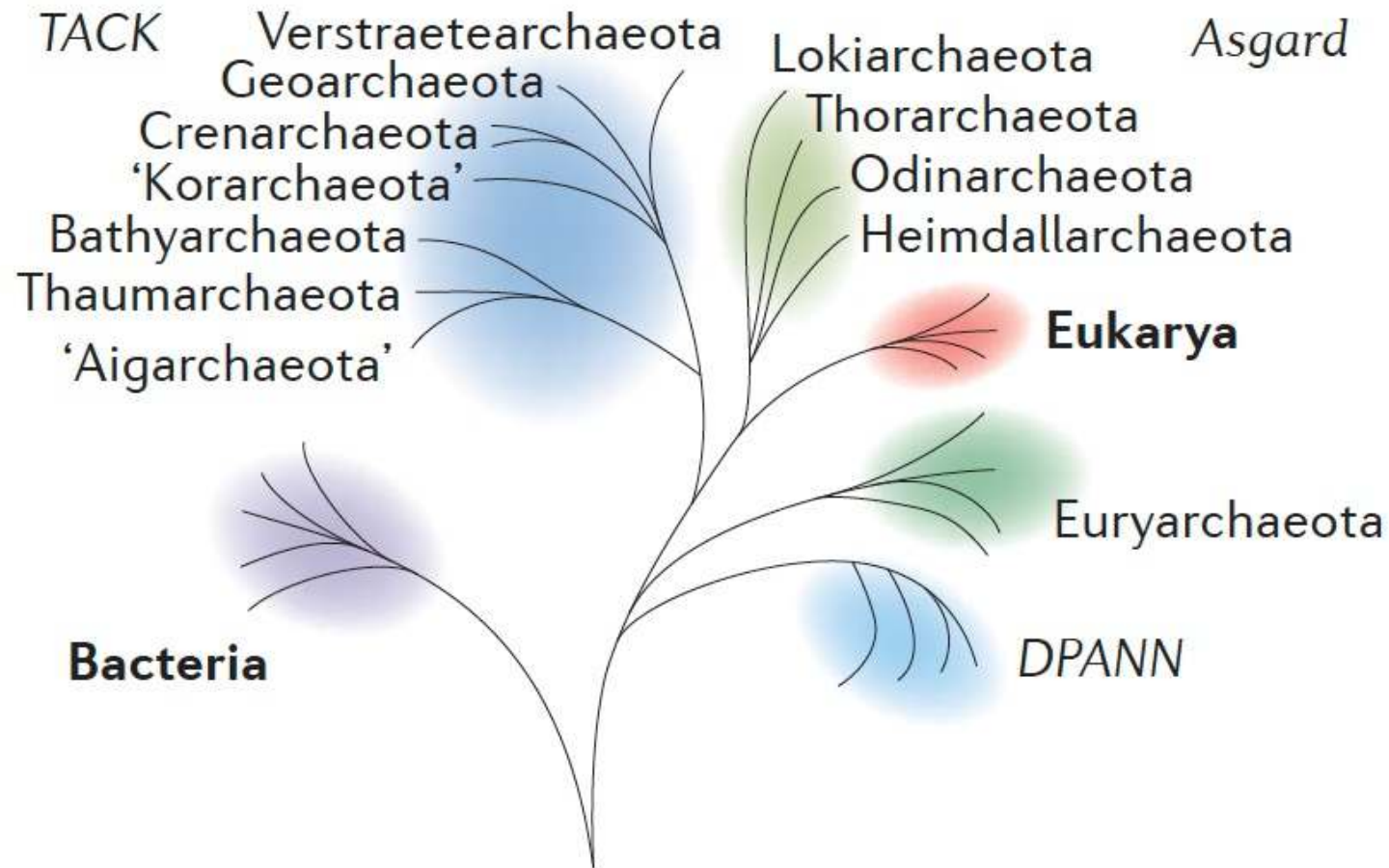
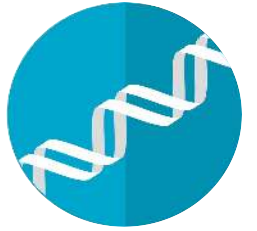
The long march of eukaryotes



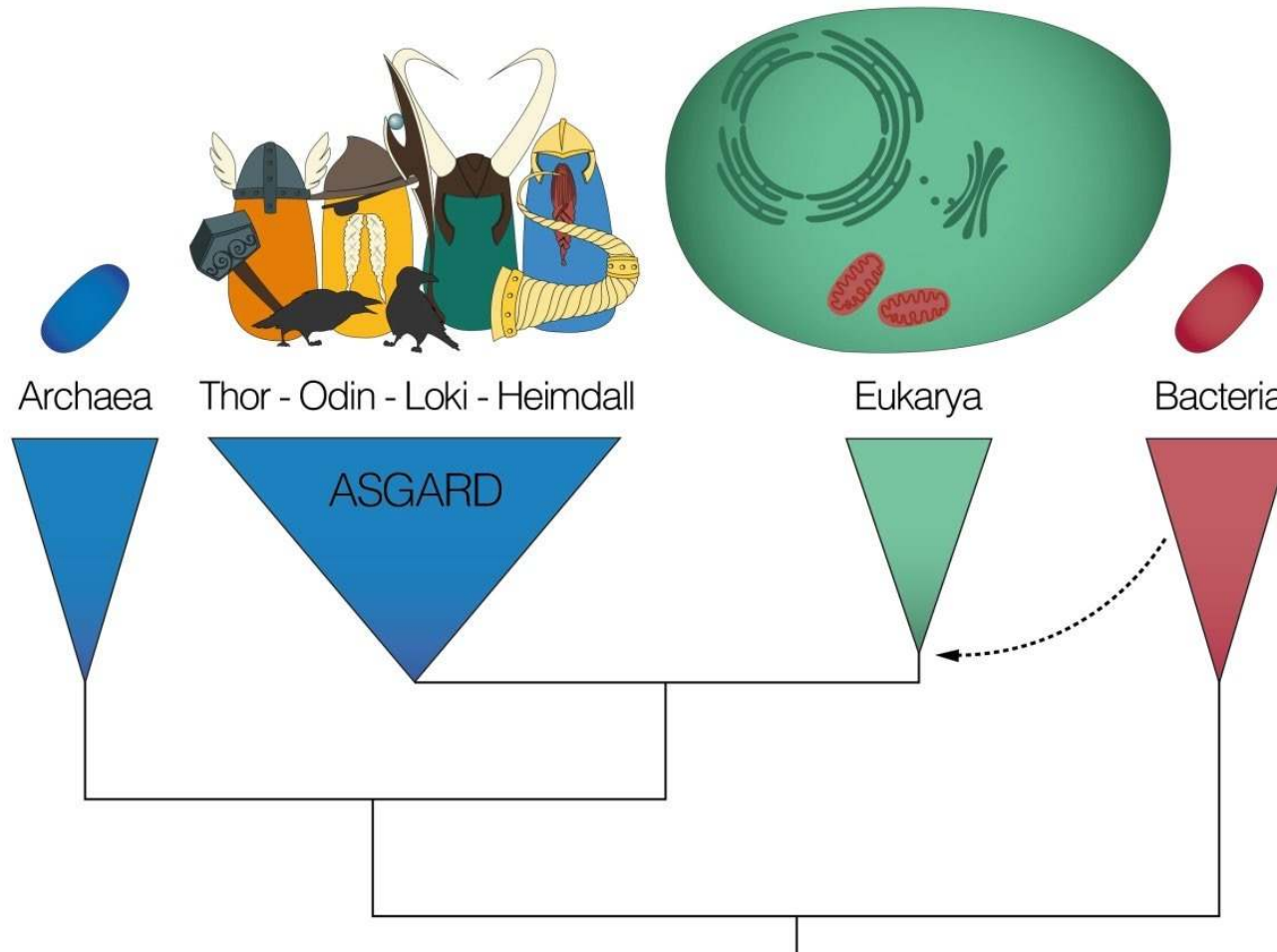


- Comparative genomic analyses have reconstructed a complex last eukaryotic common ancestor.
- But how and in which order these complex eukaryotic features evolved in an Asgard archaea?

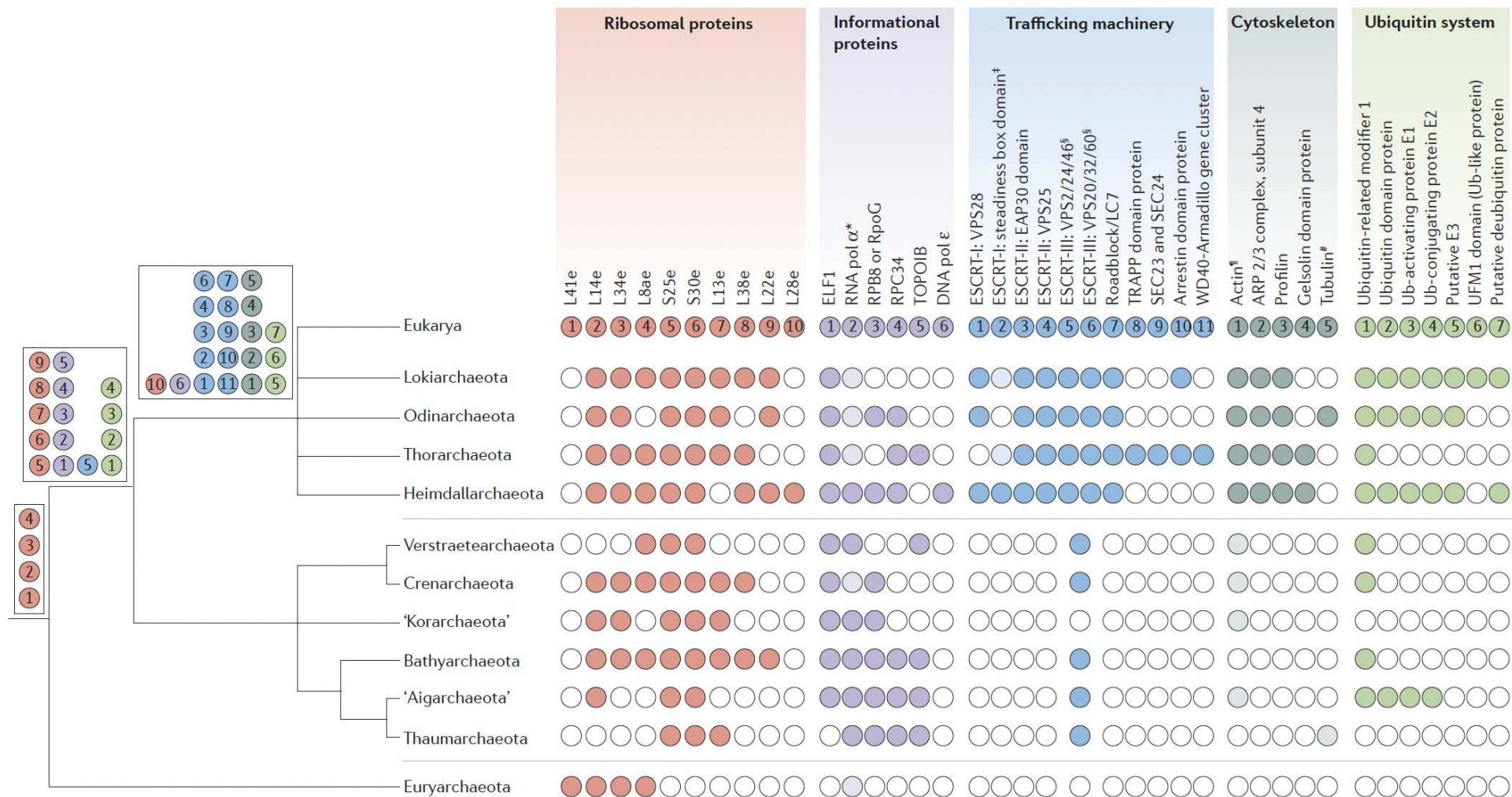
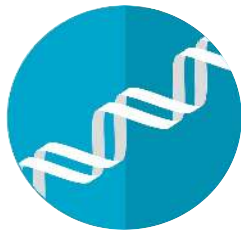
Metagenomics has revealed the vast diversity of Archaea, including the recently described Asgard superphylum



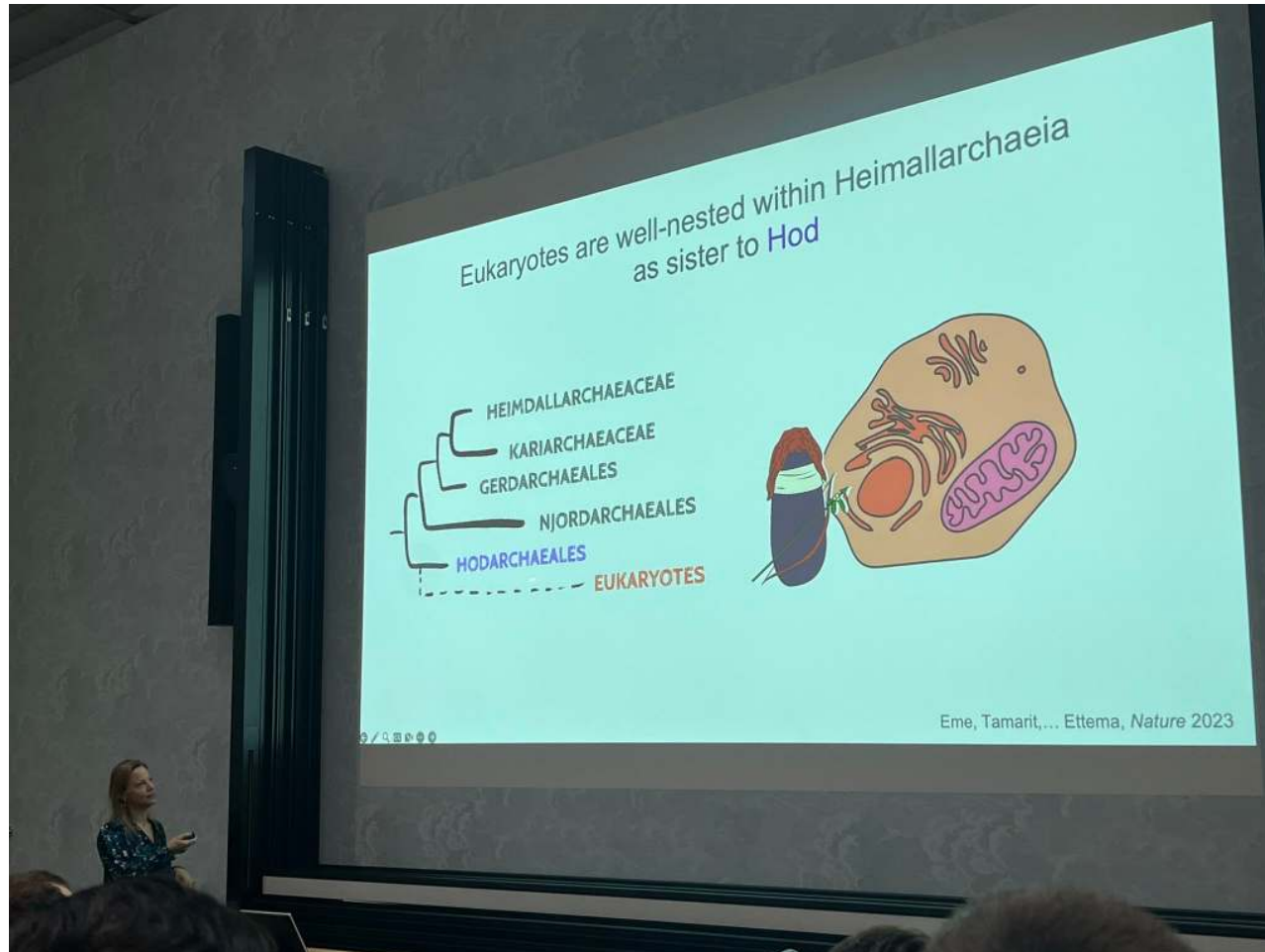
Phylogenomic analyses have placed the Asgard archaea as the closest prokaryotic relatives of eukaryotes



Genomic investigation of Asgard archaea showed that they carry several genes believed to be eukaryotic specific



Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes



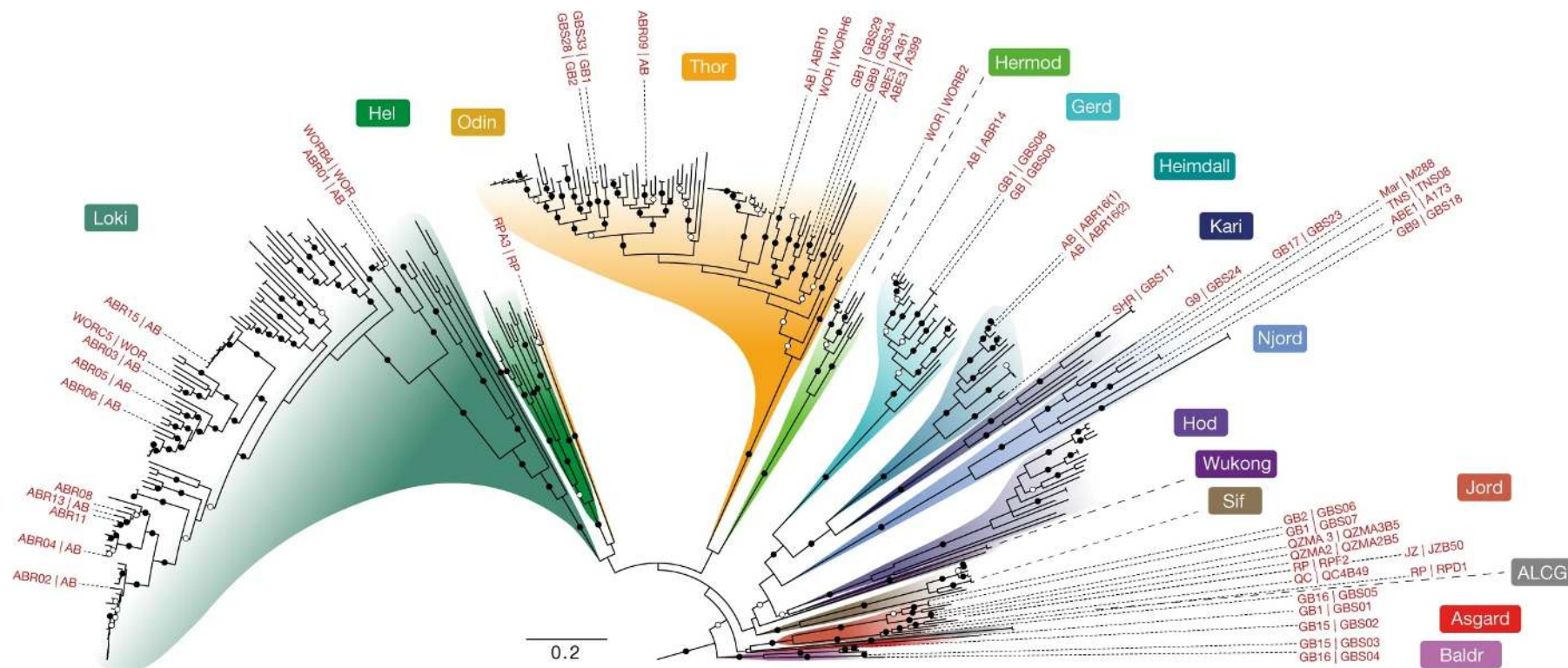
Hod in Norse mythology, is a blind god, associated with night and darkness.

Received: 23 April 2021

Accepted: 10 May 2023

Expanded genomic diversity of Asgard archaea with metagenomics

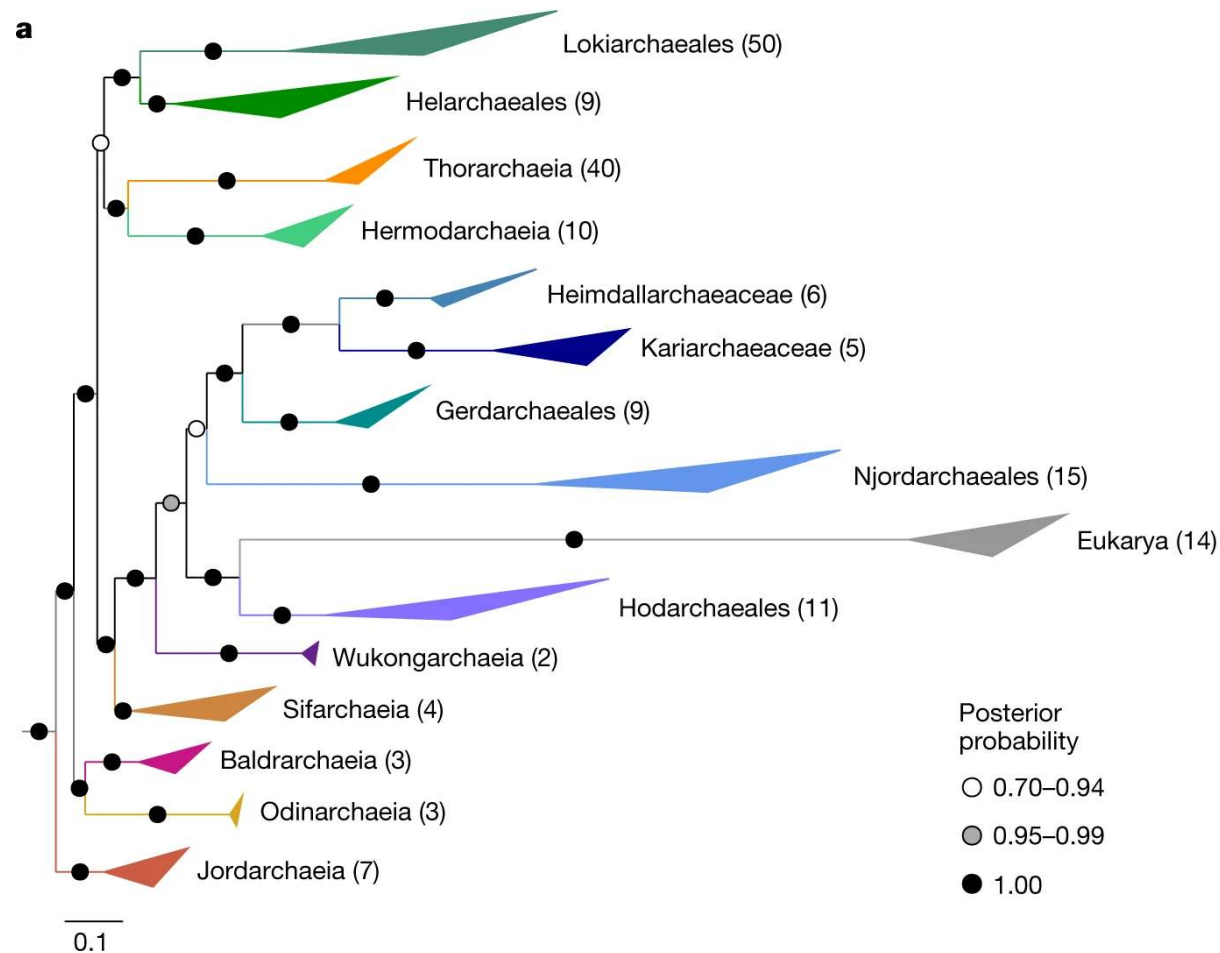
63 new Asgard archaeal metagenome-assembled genomes (MAGs) from samples obtained from 11 locations around the world



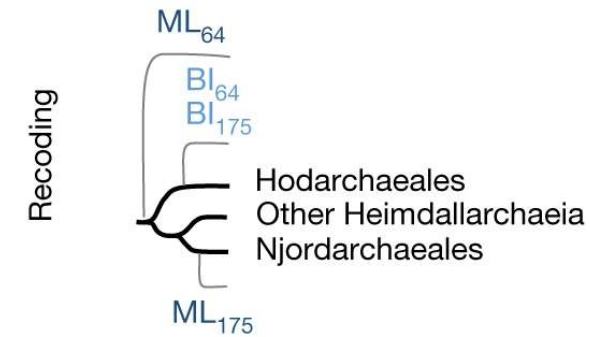
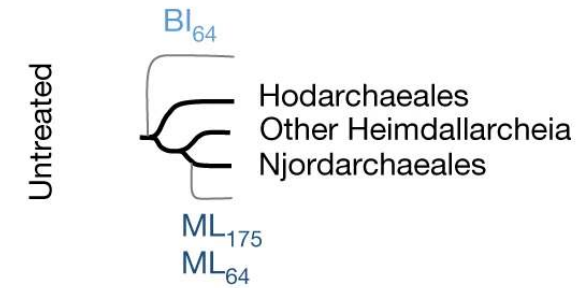
Exploring new phylogenomic markers

- typically ribosomal proteins have been used in phylogenomic analyses of archaea and eukaryotes (RP56), however they might co-evolve and therefore might contribute to phylogenetic artefacts
- independent new marker dataset comprising 57 proteins of archaeal origin in eukaryotes
- the NM57 proteins are mostly involved in informational, metabolic and cellular processes, but do not include ribosomal proteins
- they are longer and therefore putatively more phylogenetically informative compared with the RP56 markers
- broader functional distribution of NM57 markers is less likely to cause phylogenetic reconstruction artefacts

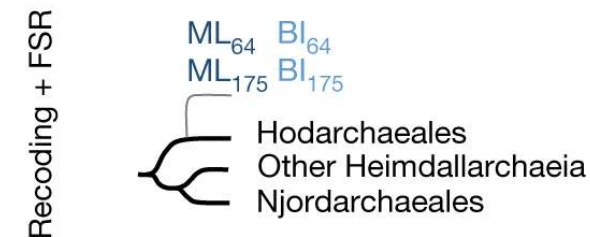
Eukarya as sister to Hod



BI based on 278 archaeal taxa, 15,733 AA positions
The concatenation was SR4-recoded and analysed using the CAT+GTR model



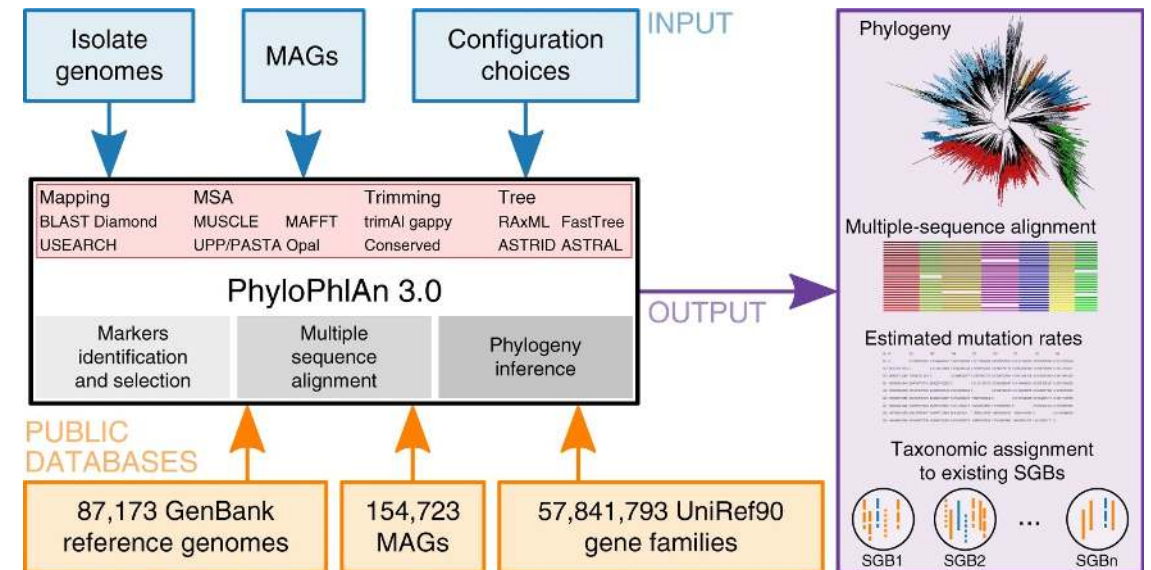
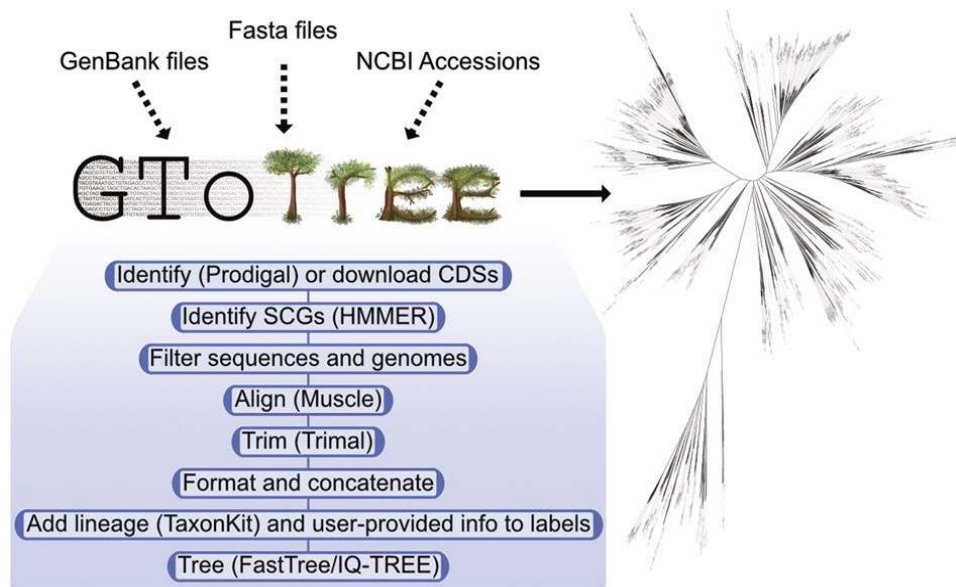
The indices 175 and 64 refer to phylogenomic datasets containing 175 and 64 Asgard archaea, respectively.



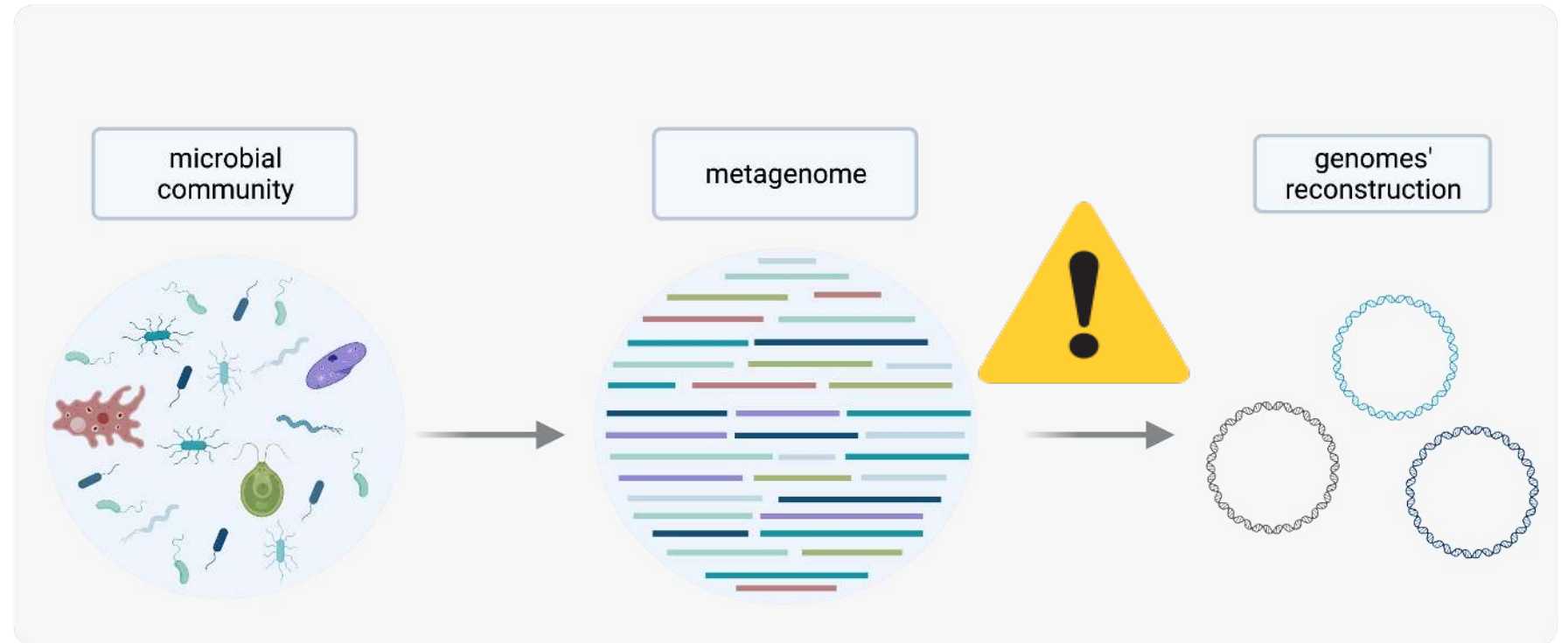
Schematic representation of the shift in the position of eukaryotes (grey branches) in ML and BI analyses under different treatments

Metagenomic data to the rescue

If MAGs can be recovered from metagenomic data, they can be used as a source of data for phylogenomics in a similar manner as genomes (with higher risk of contamination)



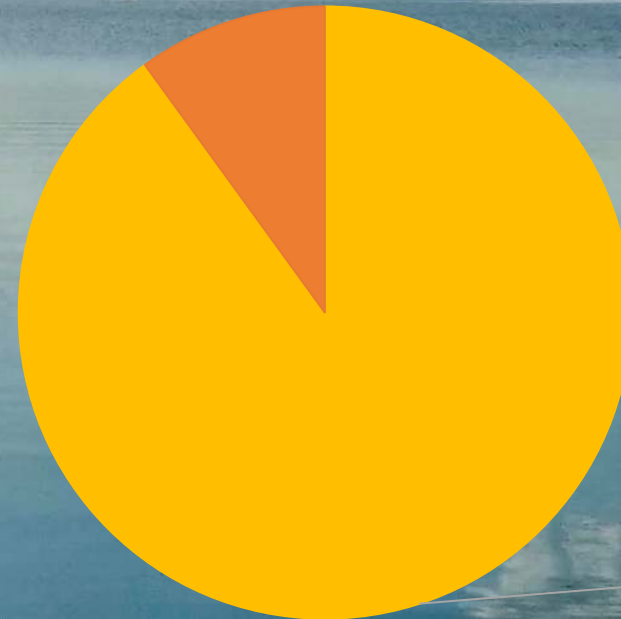
Metagenomics



The genomic analysis of microbial communities by extraction and sequencing of their DNA, which allows studying organisms directly in their natural environment

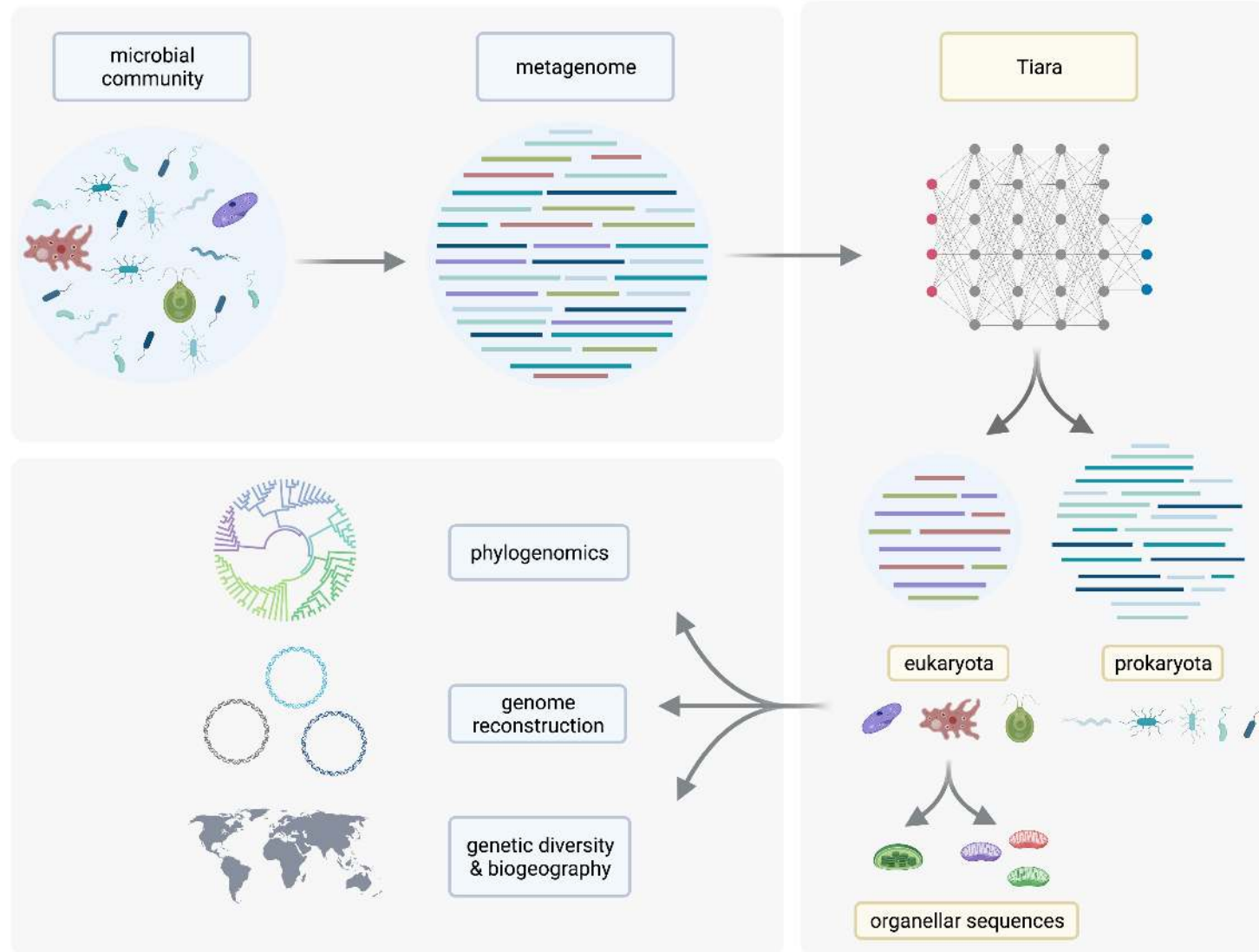
Metagenome of lake

Protists 10%



Prokaryotes 90%

How to obtain eukaryotic genomes from metagenomes?



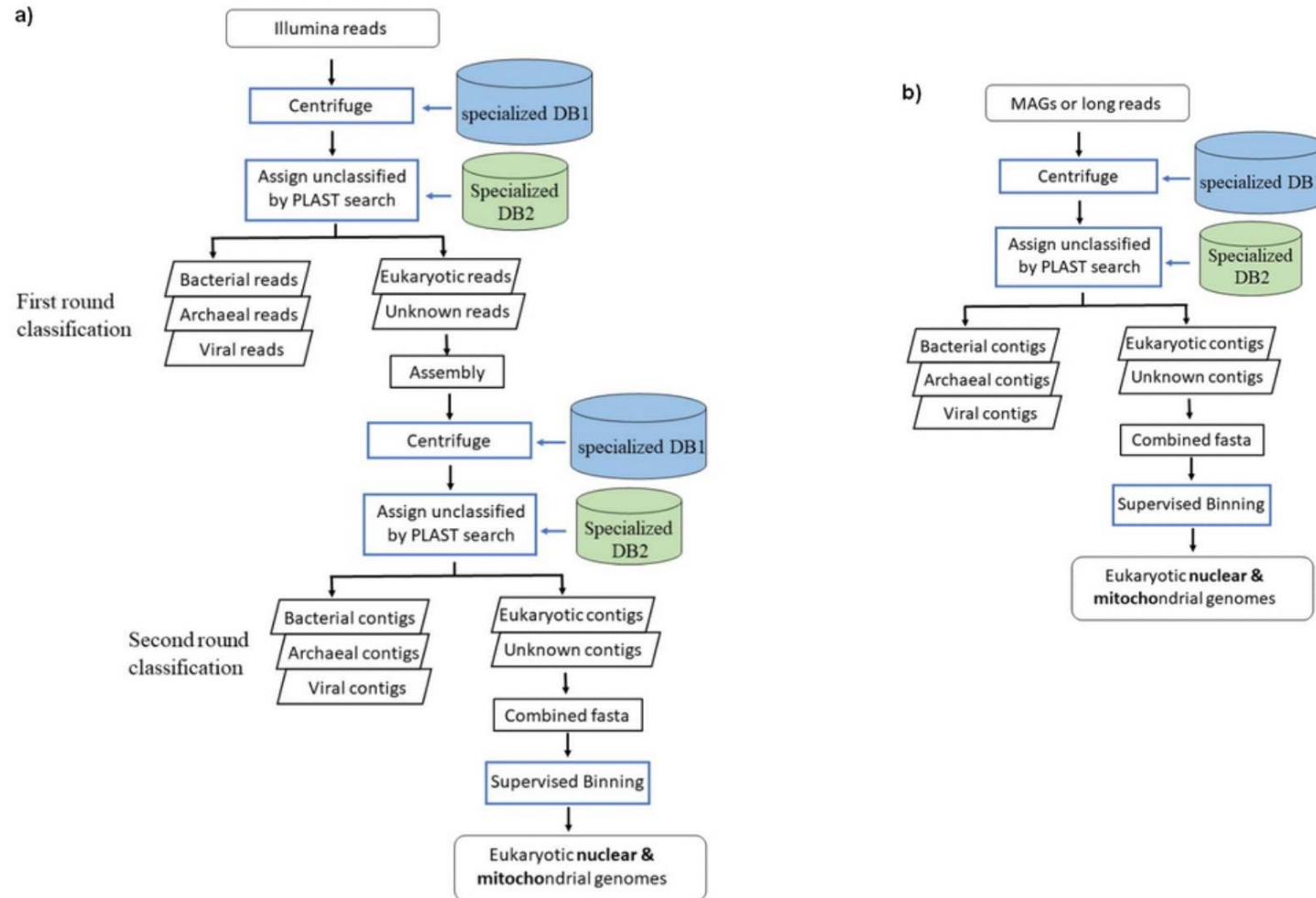
tiara

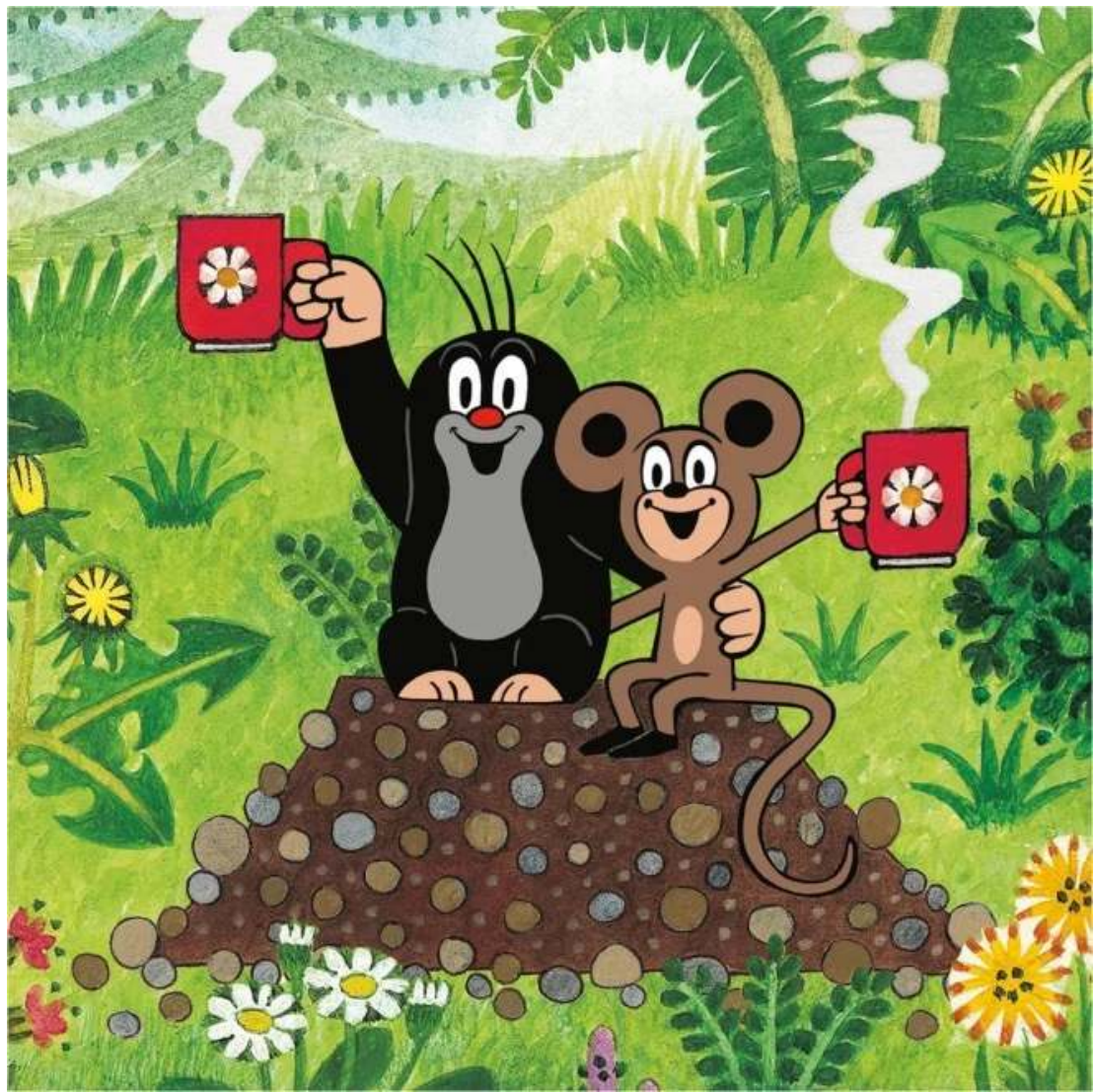
tiara – a tool for DNA sequence classification

[View on GitHub](#)

Other tools - Eukfinder

With sufficient reads, Eukfinder efficiently assembles high-quality near-complete nuclear and mitochondrial genomes from diverse Blastocystis subtypes from metagenomic data without the aid of a reference genome.

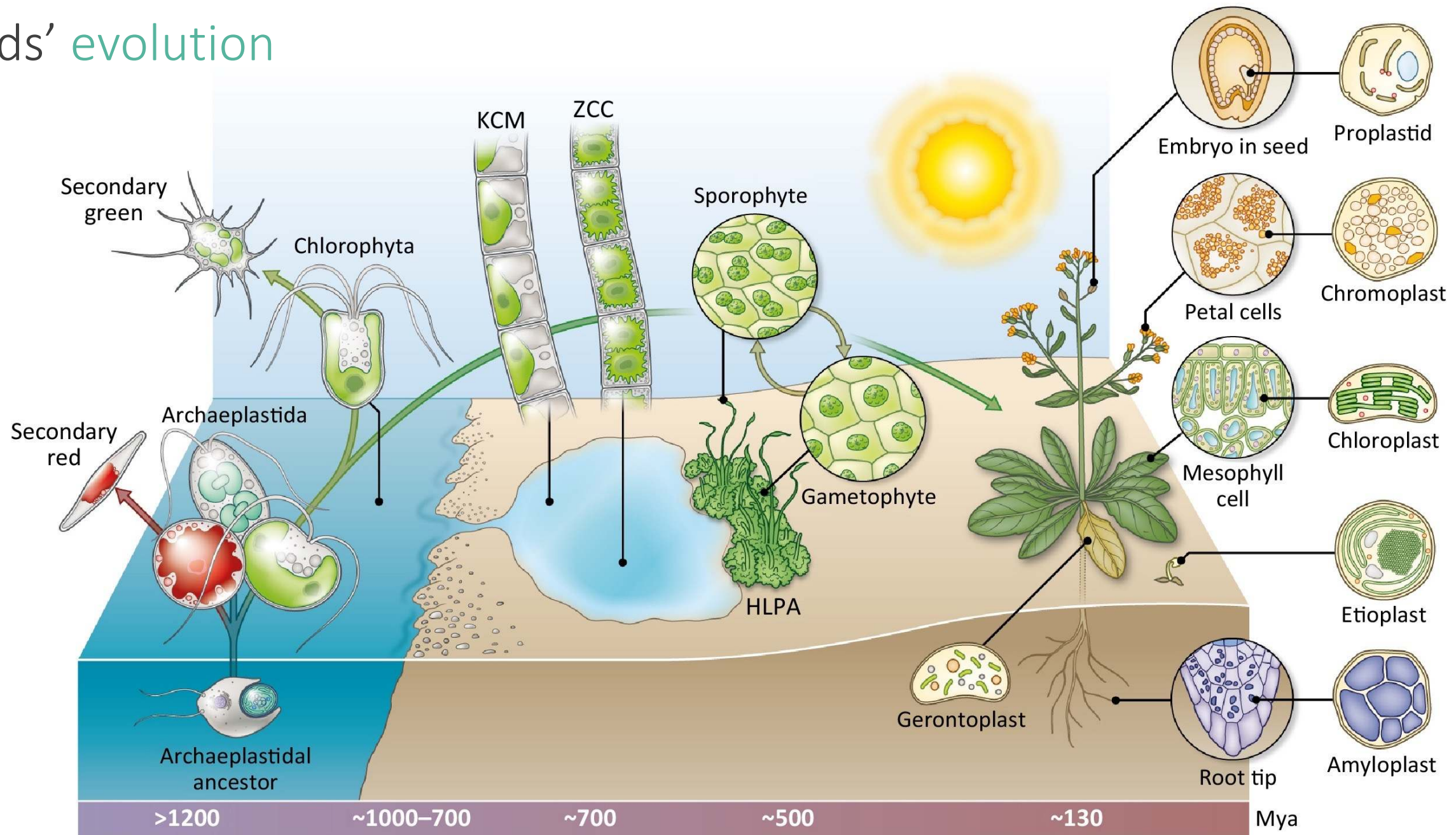


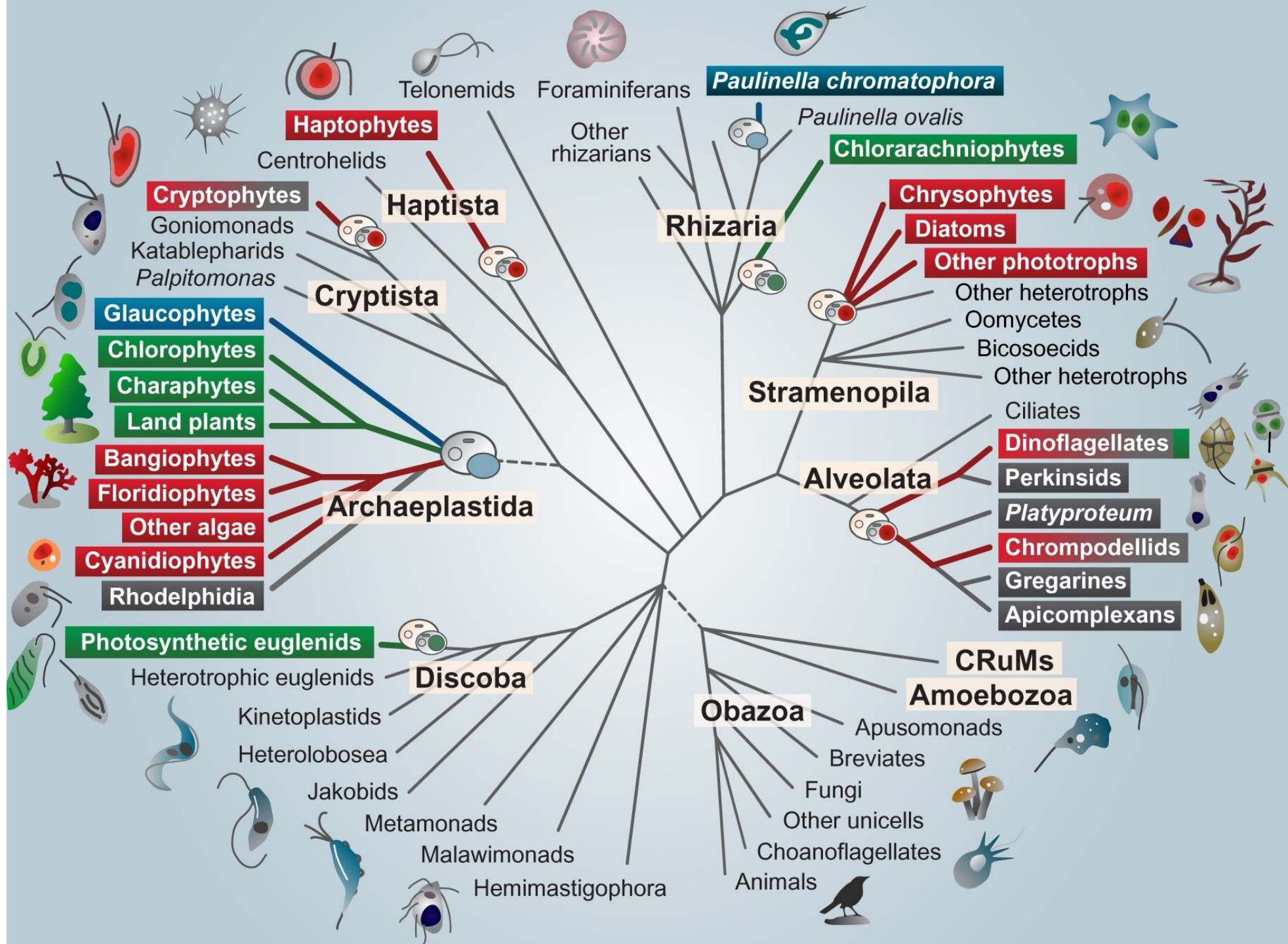


How the eukaryotic cell complexity evolved?
What drives diversification of eukaryotes?

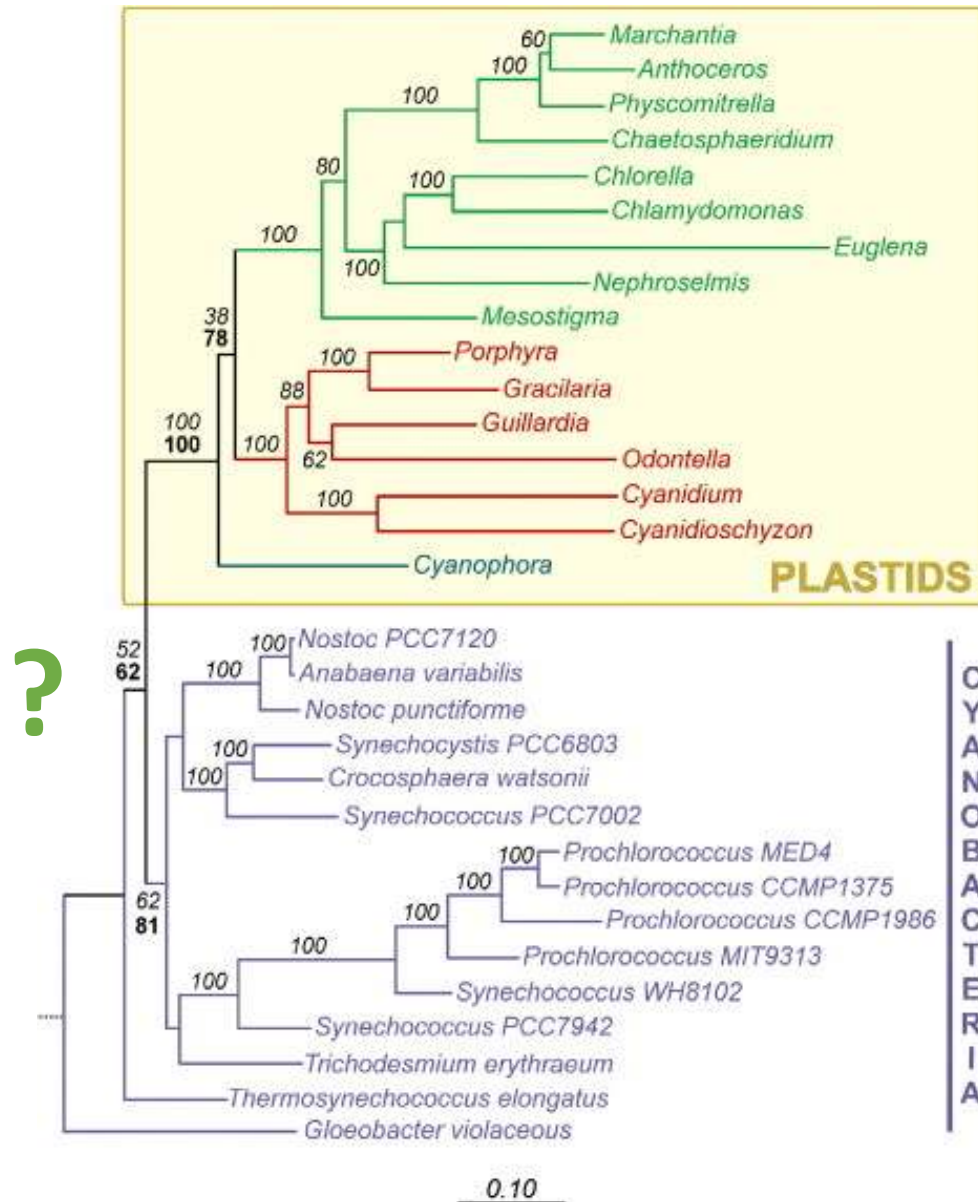
Endosymbiosis

plastids' evolution



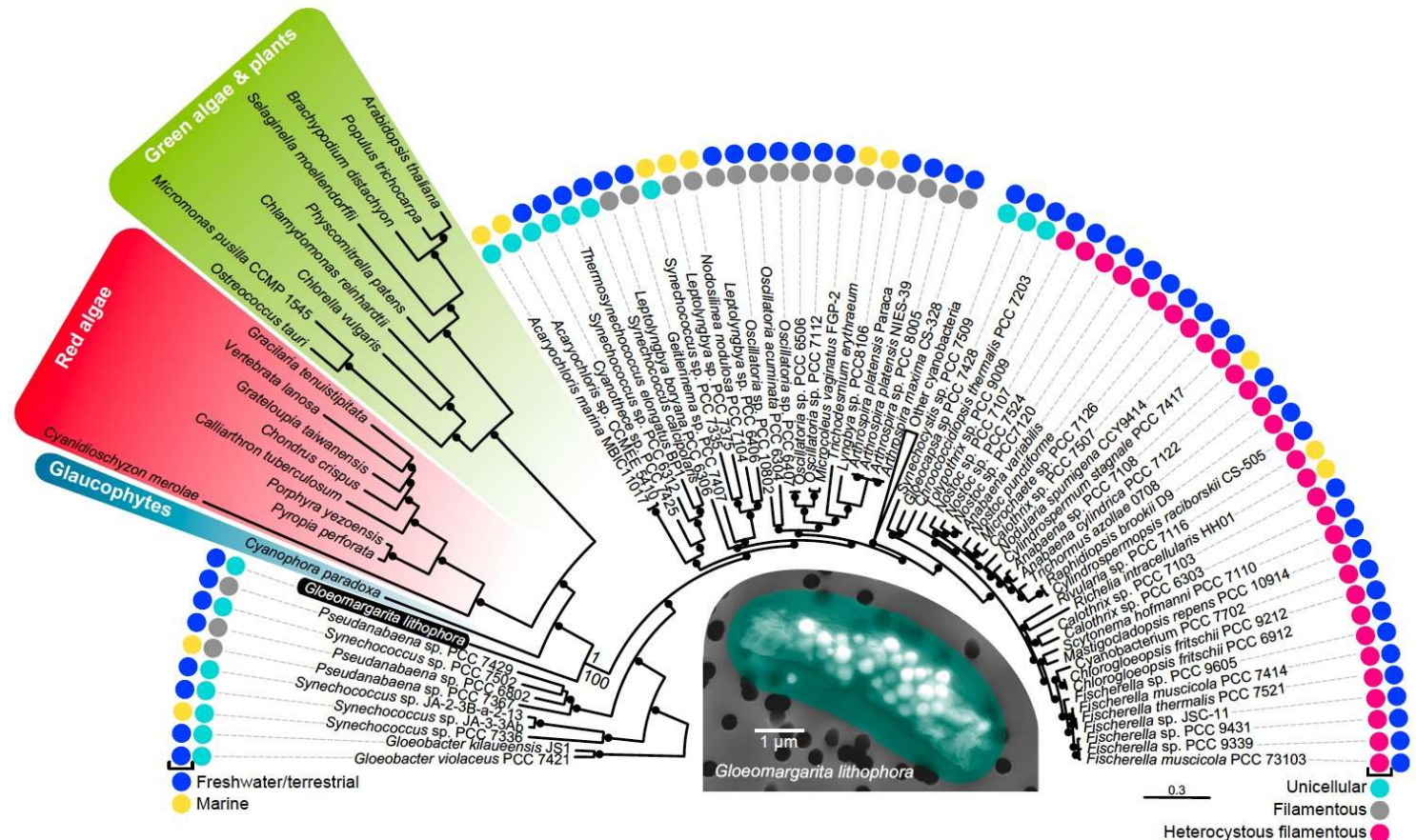


Plastids and cyanobacteria have been recovered repeatedly as a monophyletic group pointing to the cyanobacterial origin of primary plastids

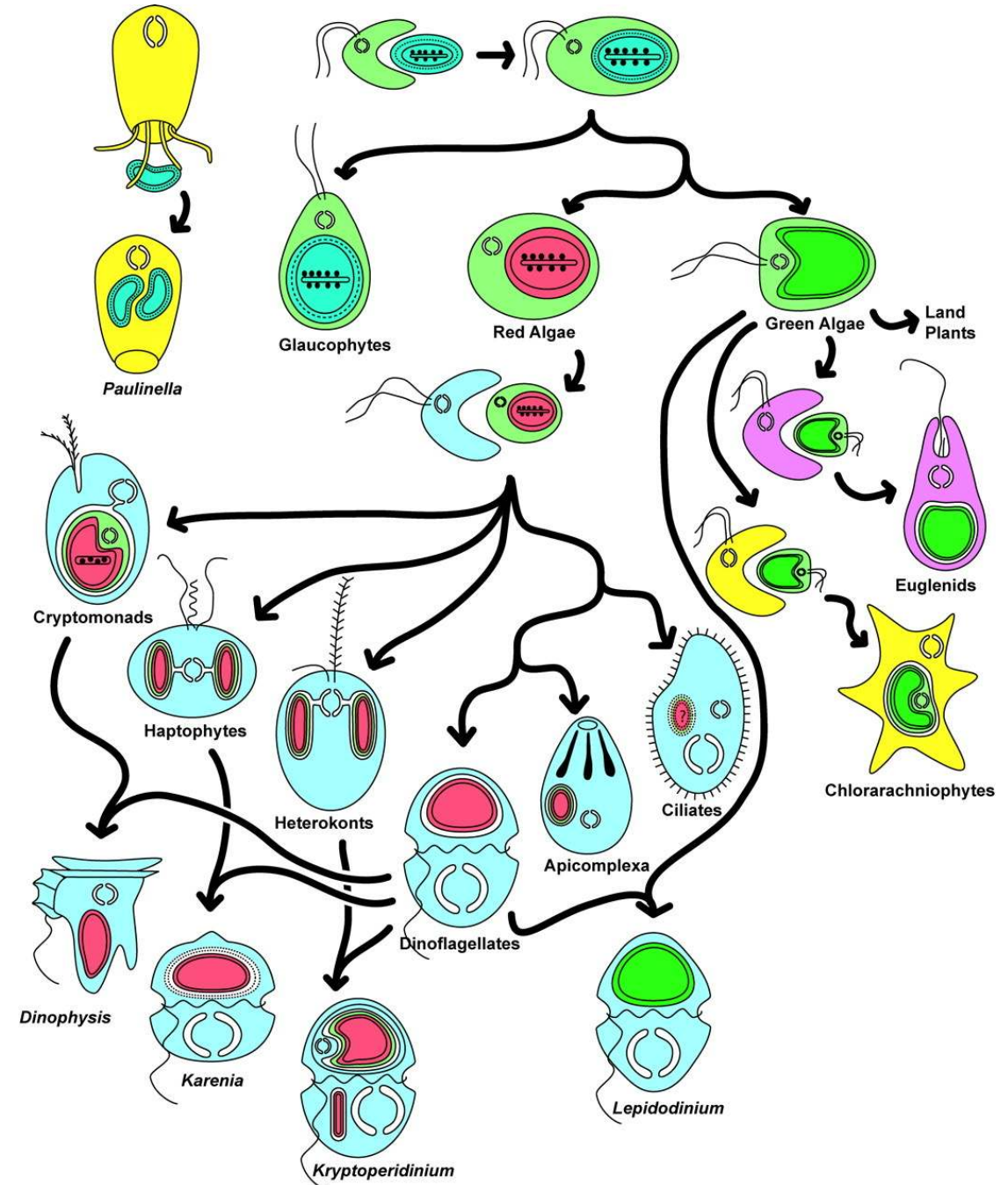
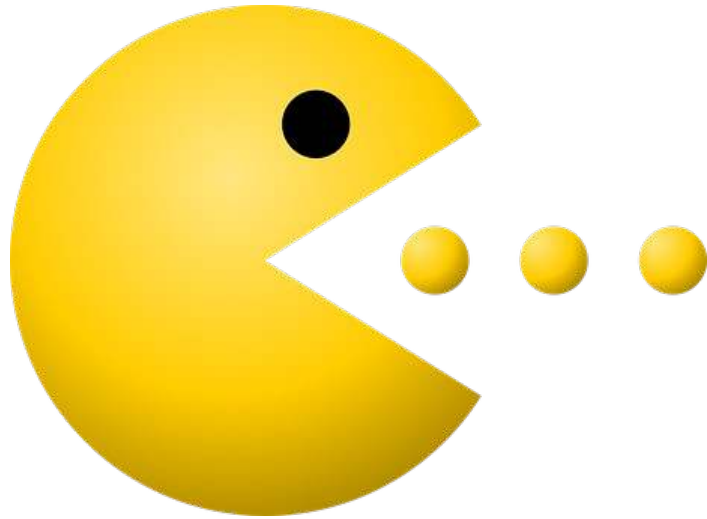


Cyanobacteria which became a chloroplast

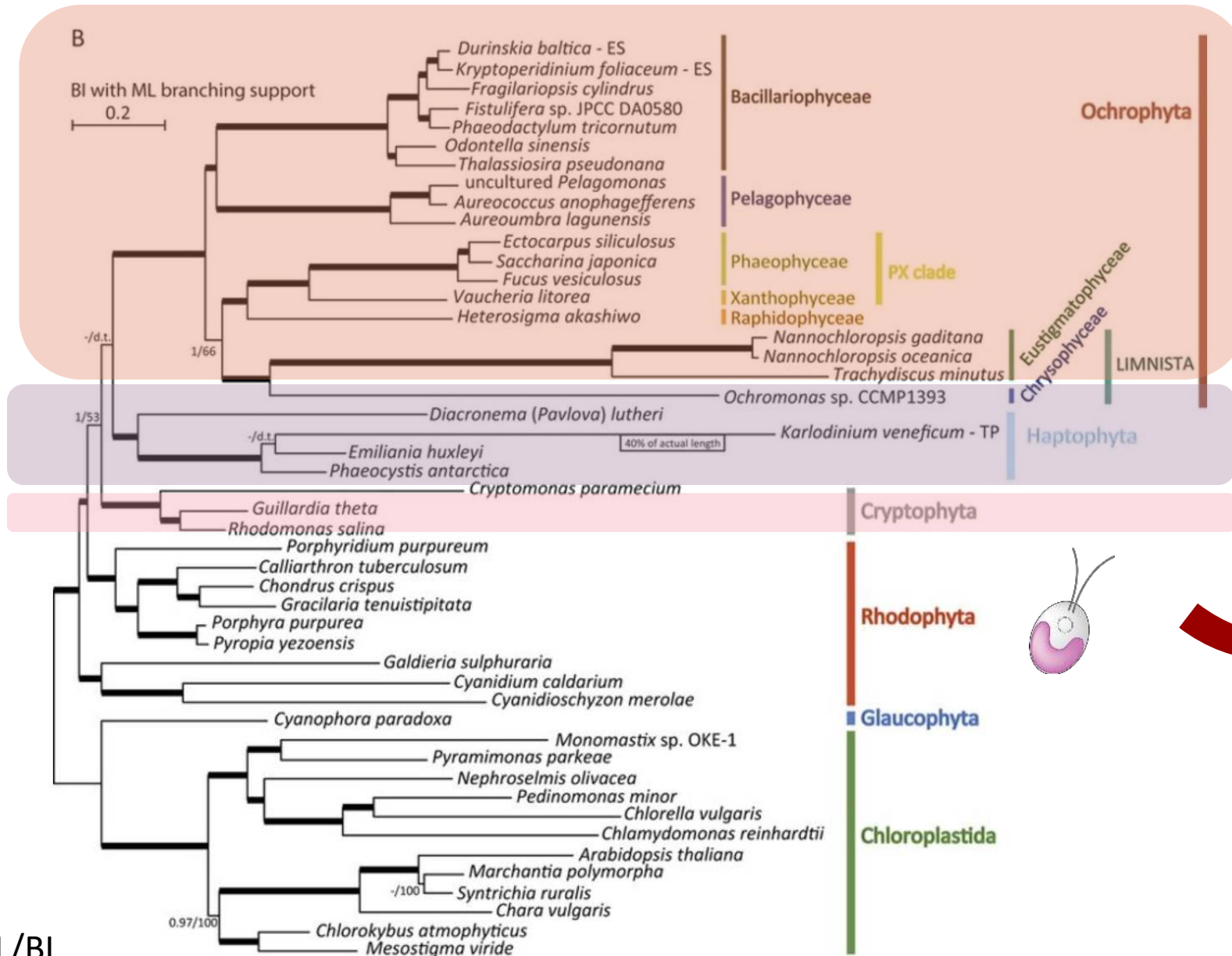
- *Gloeomargarita lithophora* is the closest extant cyanobacteria to plastids
- old cyanobacterial lineage
- freshwater cyanobacteria (!)



Secondary endosymbioses in several lineages of microbial eukaryotes



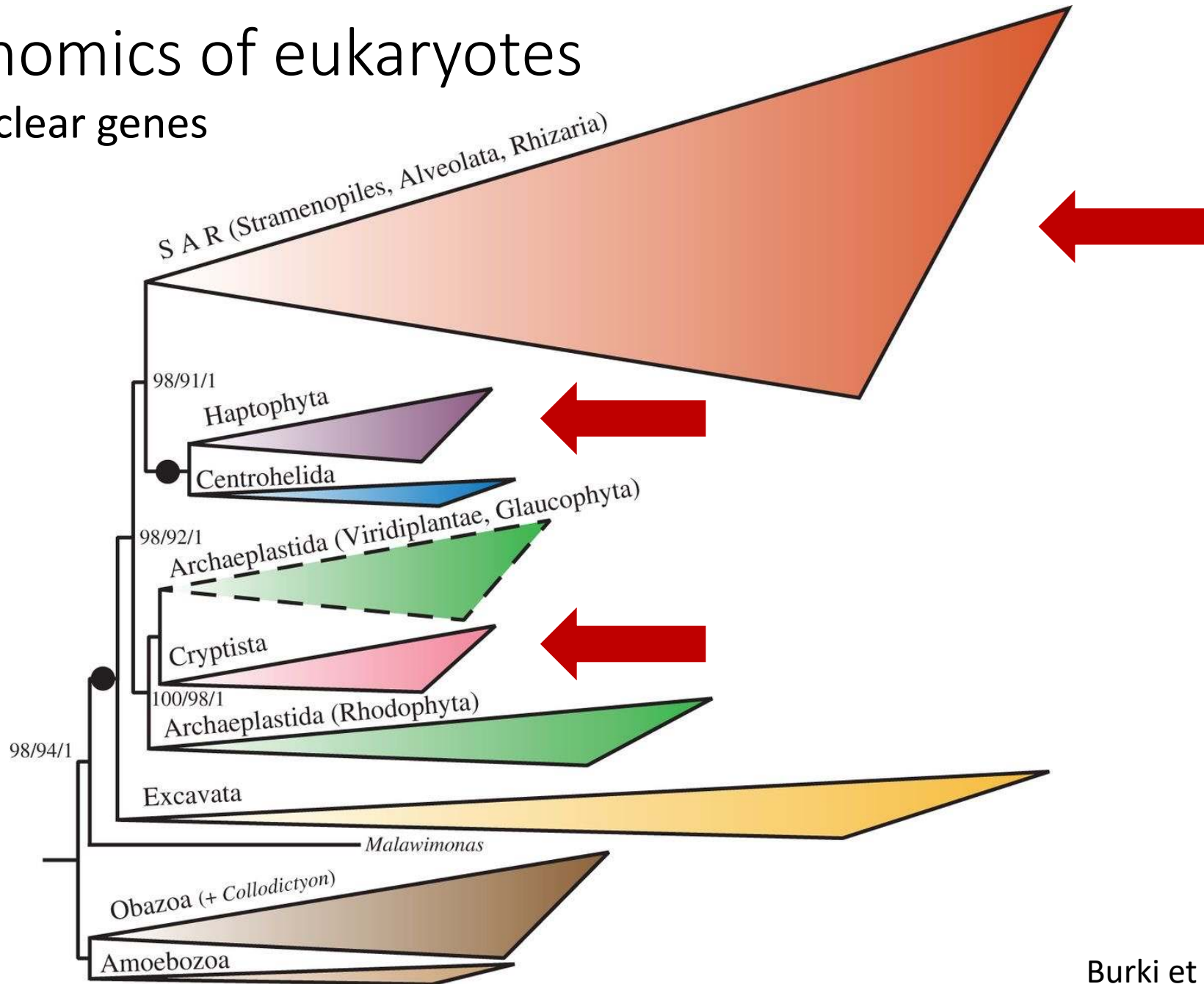
Phylogenomics of plastids



secondary „red” endosymbionts

Phylogenomics of eukaryotes

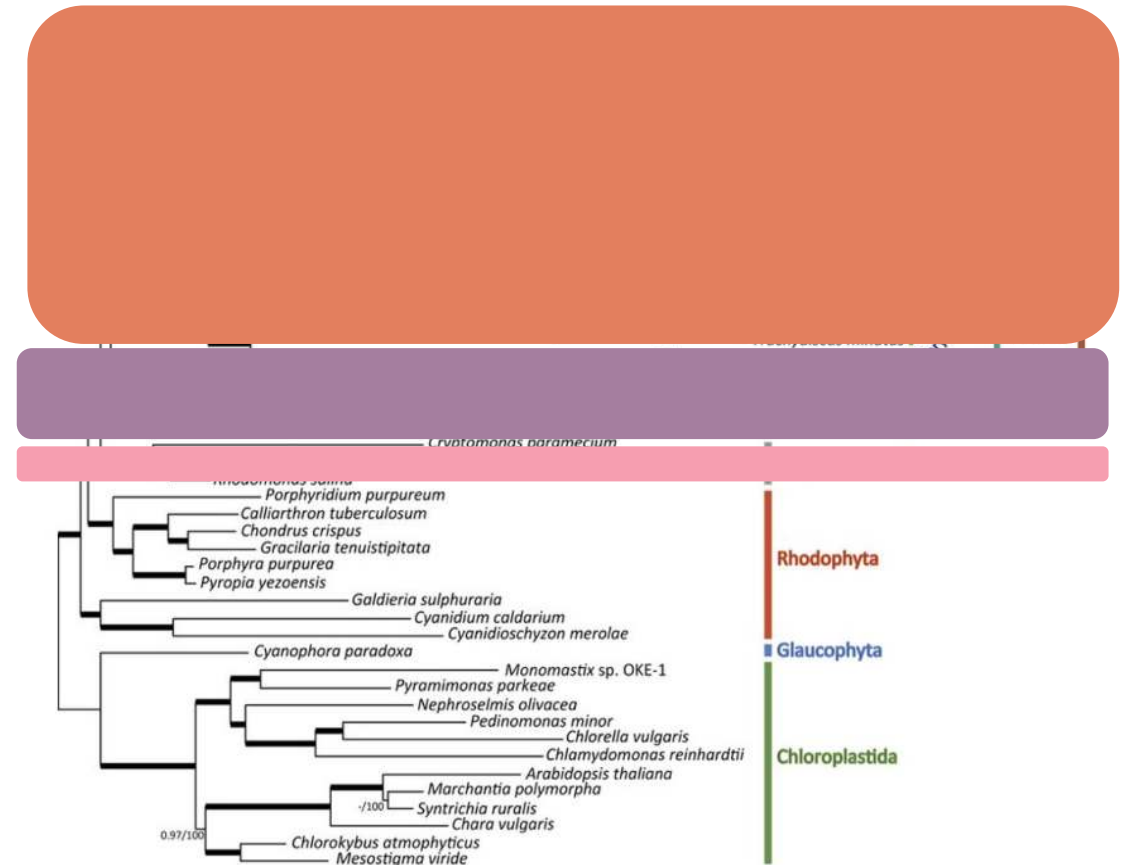
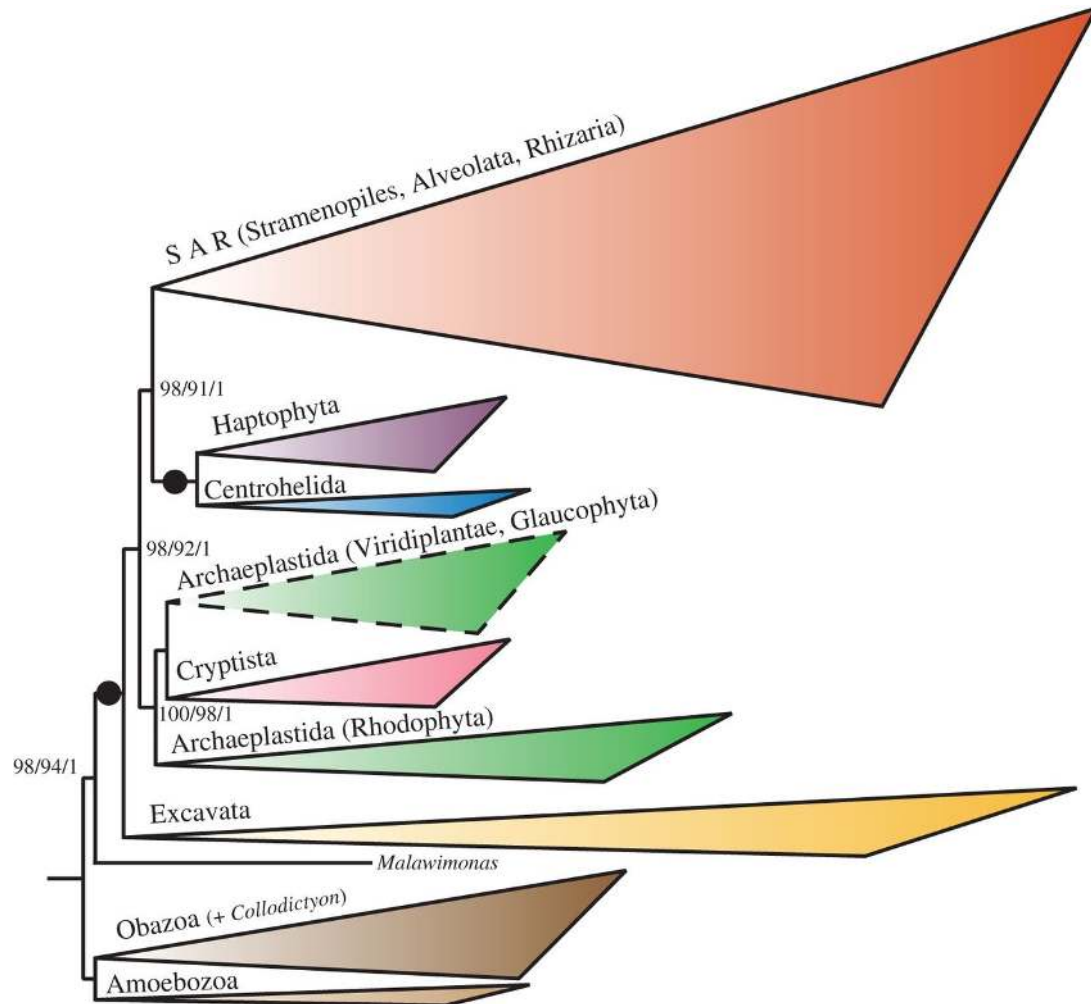
based on nuclear genes



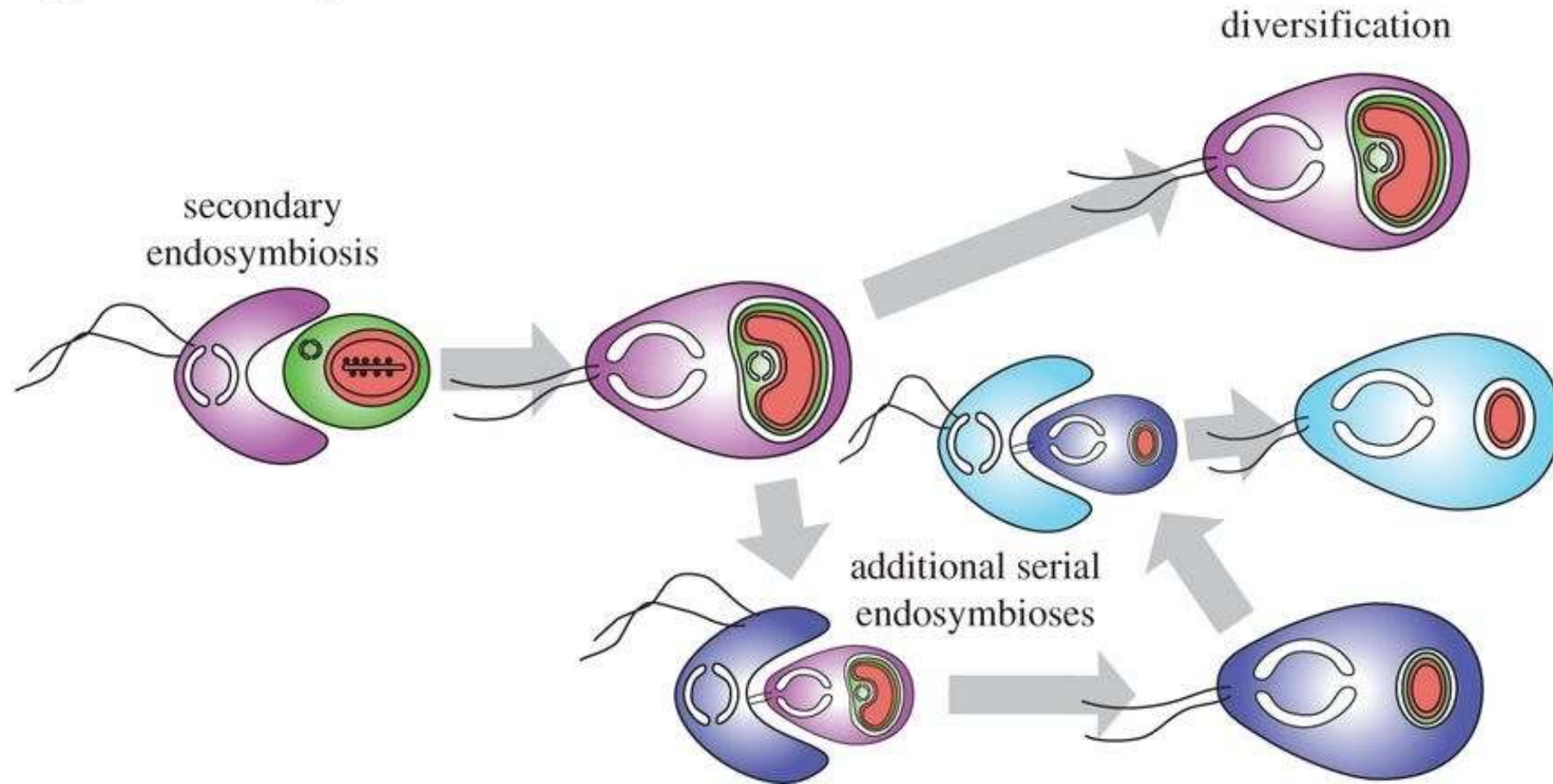
ML/BI 250 genes

Burki et al. 2016

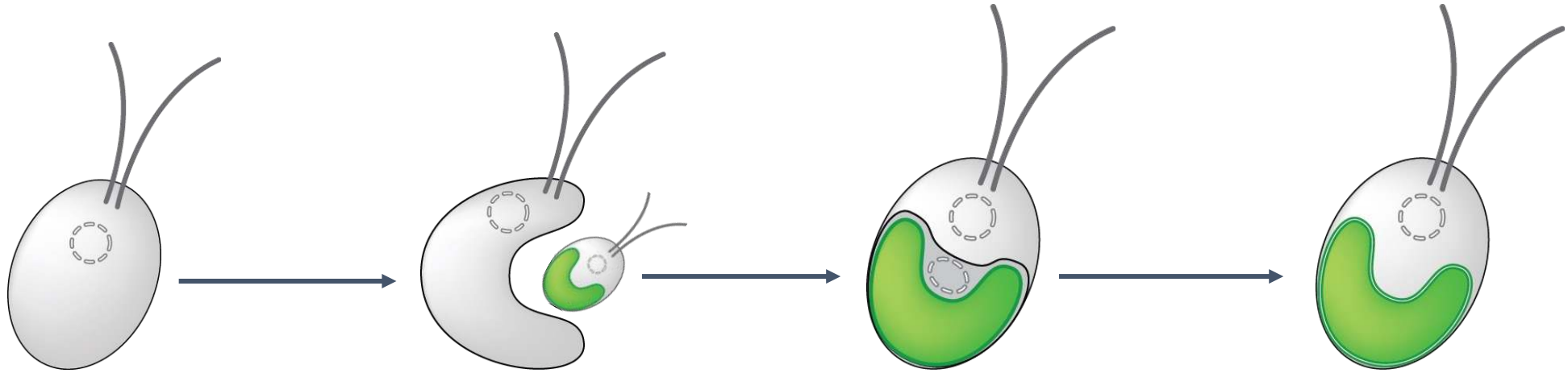
Nuclear-encoded genes vs plastid-encoded genes



Cryptic serial endosymbiosis



Plastid endosymbiosis



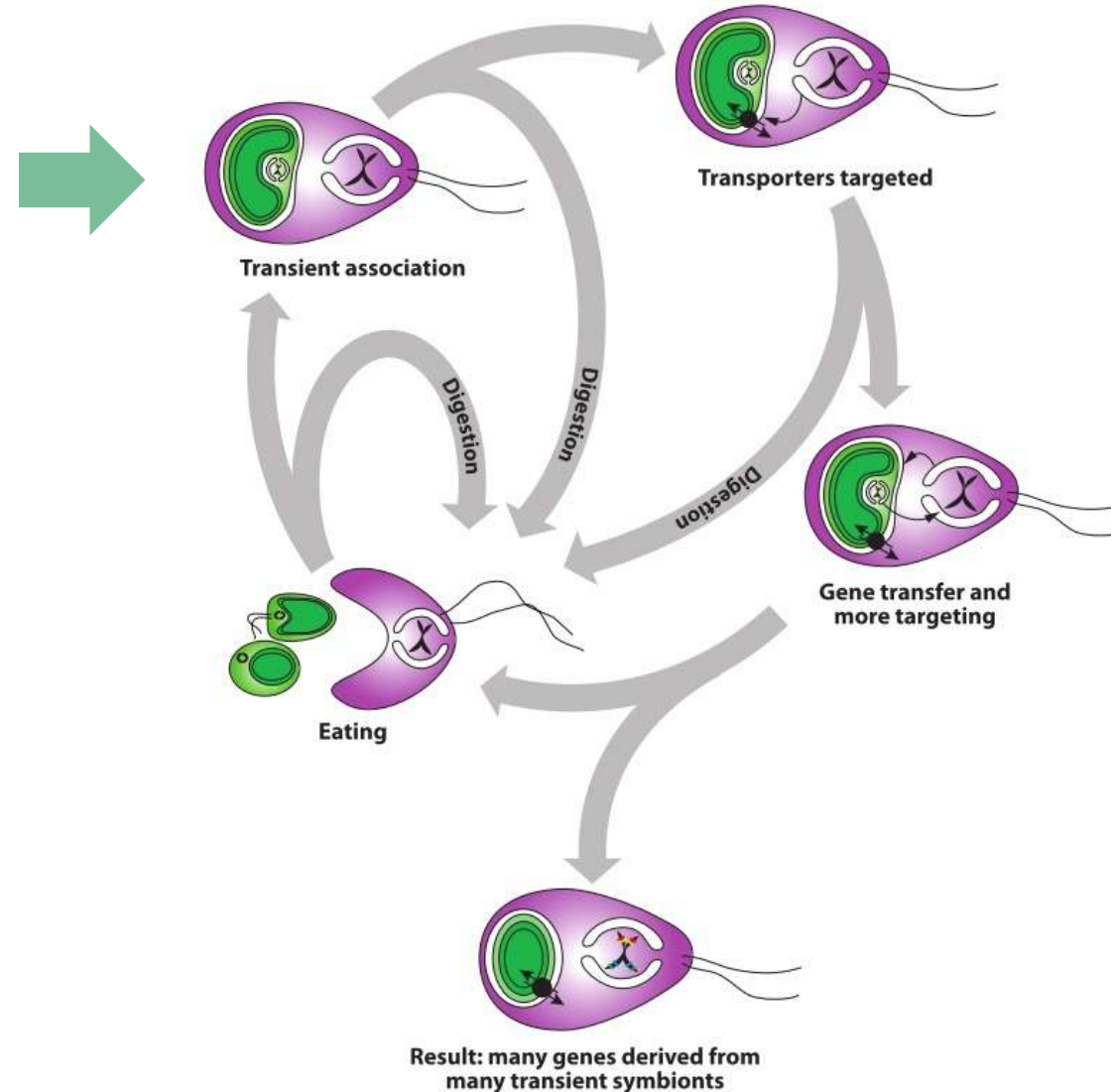
What are the initial steps of endosymbiosis?

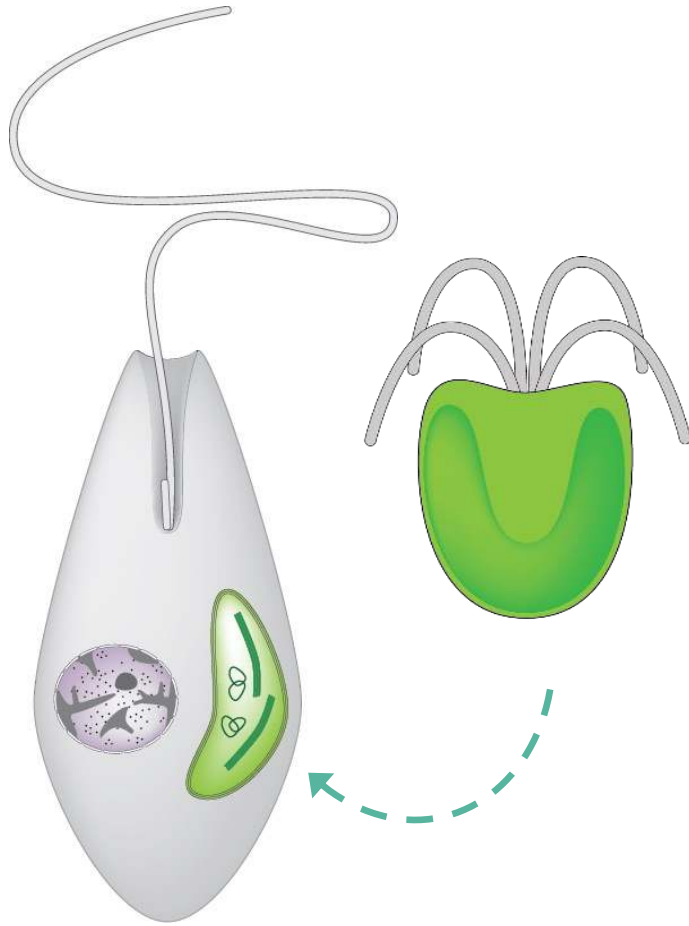
What are the steps in the transition from endosymbiont to organelle?

Which comes first, gene transfer or cellular fixation?

A targeting-ratchet model for the endosymbiotic origin of plastids

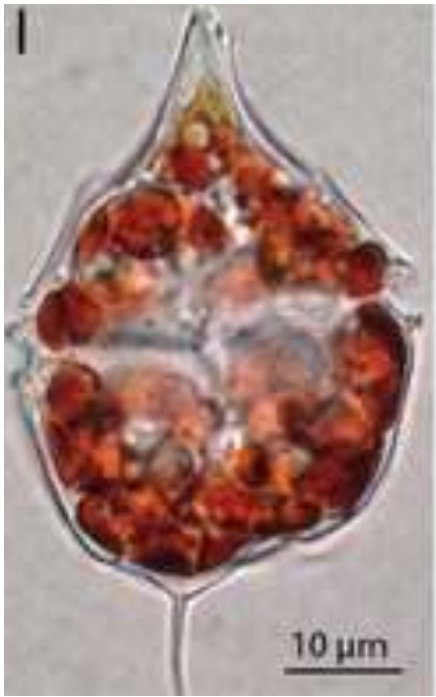
Keeling, 2013



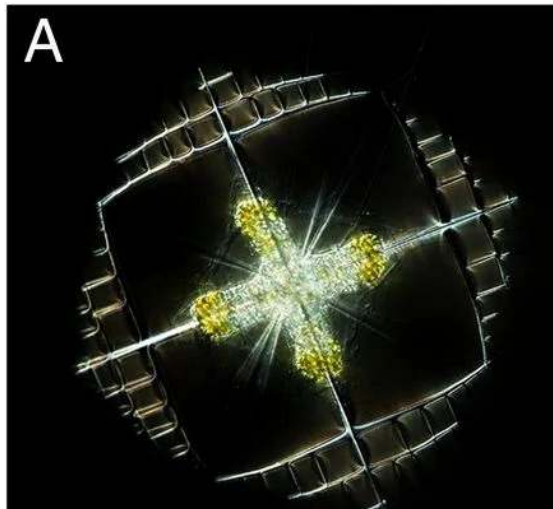


kleptoplasty

transient association between host and endosymbiont, which might resemble the initial steps of the establishing endosymbiosis

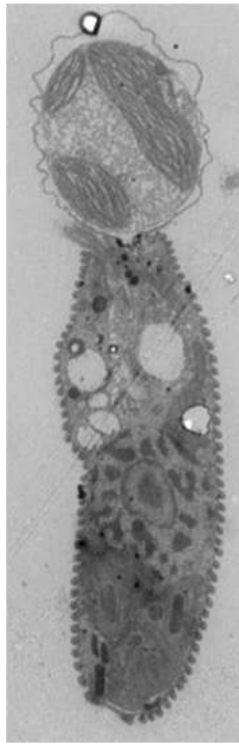


Kleptoplasty is
widespread among
eukaryotes

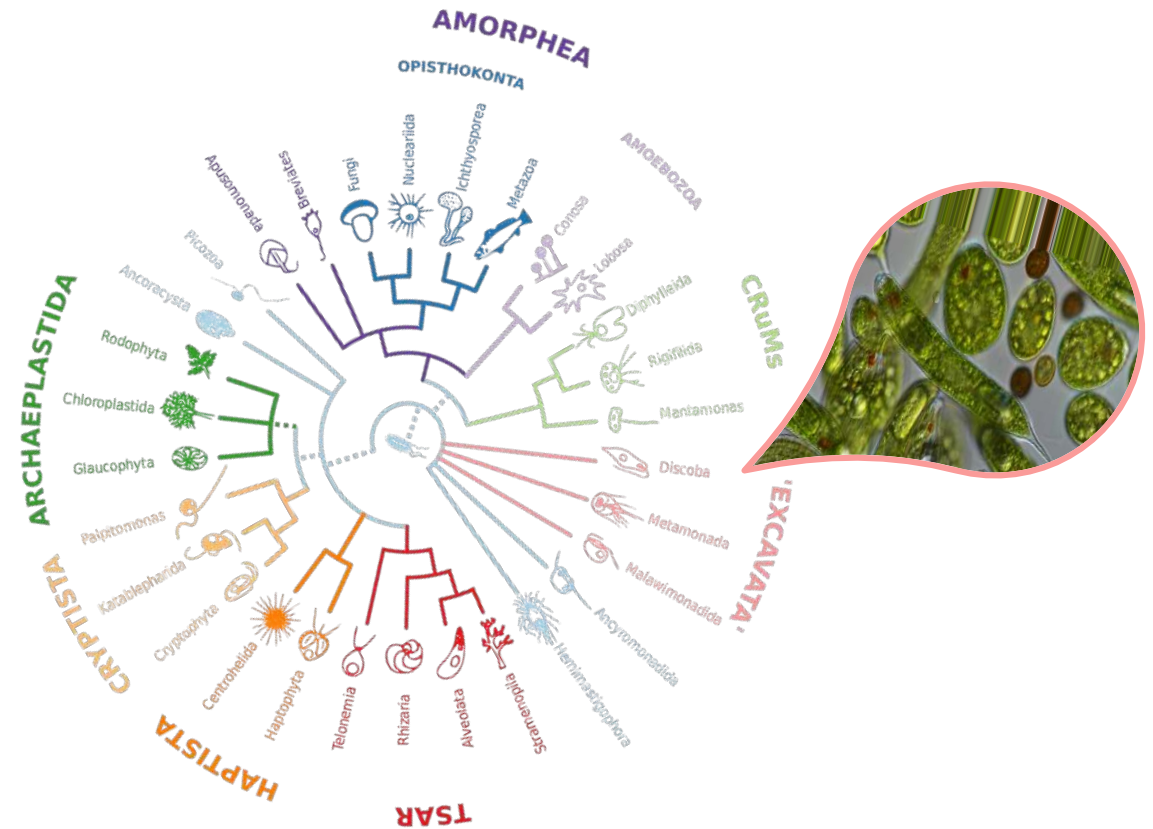


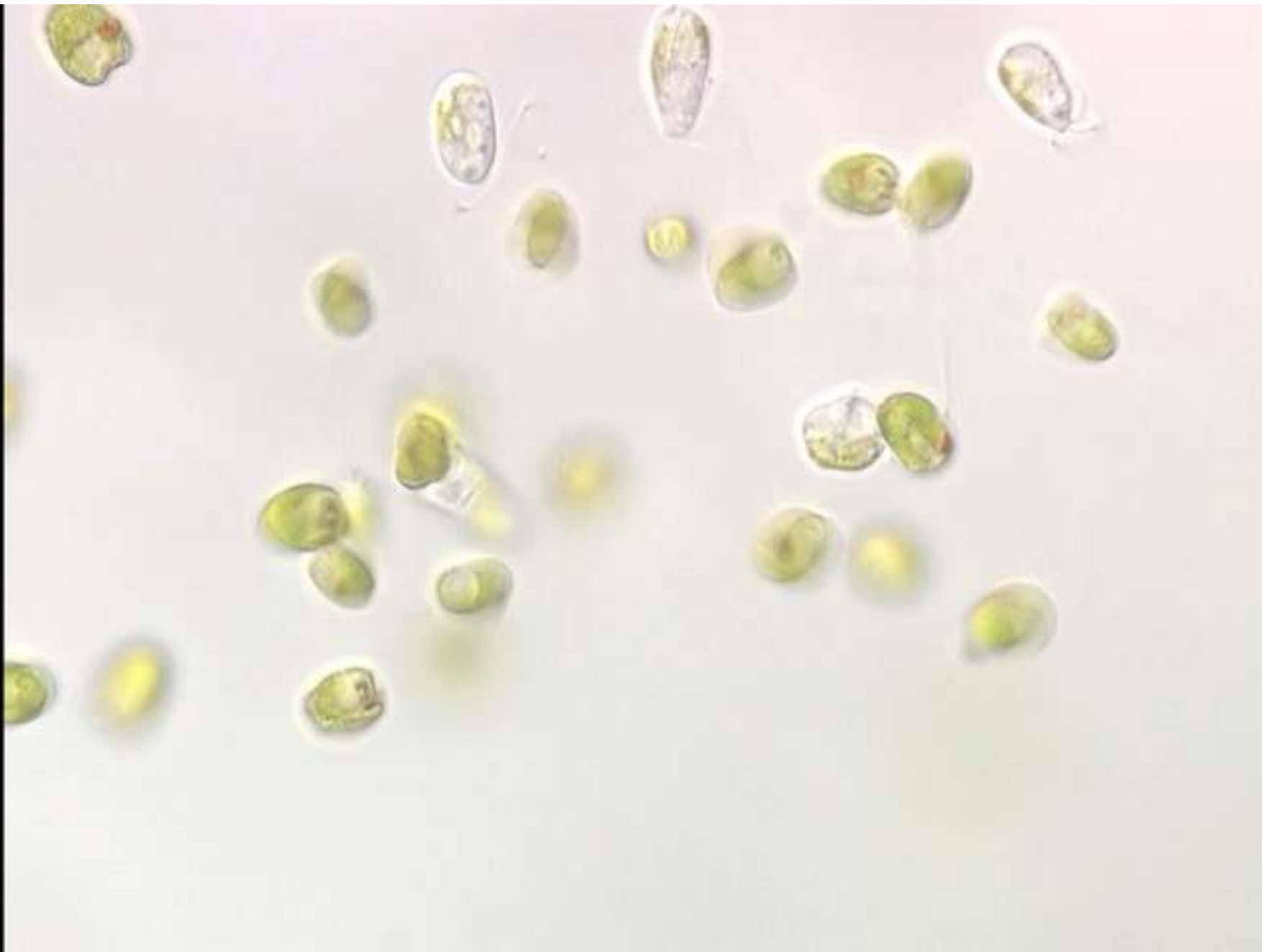
Rapaza viridis – mixotrophic euglenid

feeds on a specific strain of green algae, *Tetraselmis* sp.

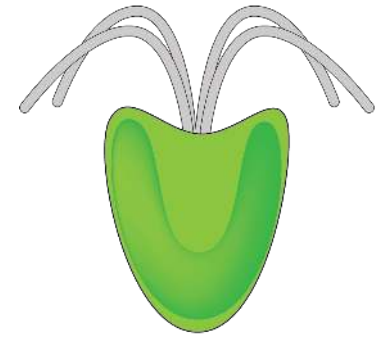
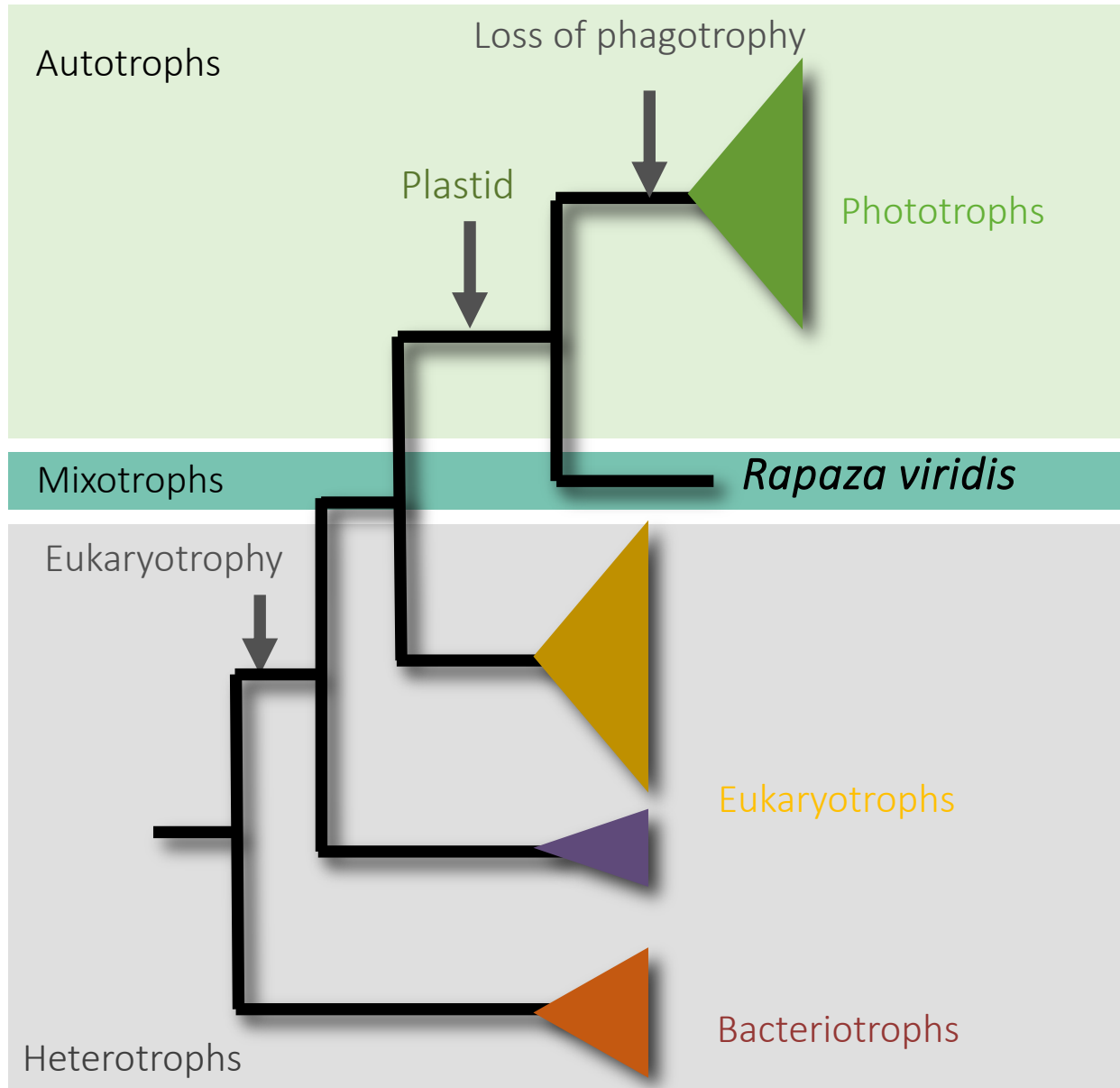


Rapaza viridis

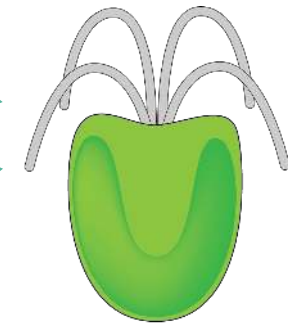




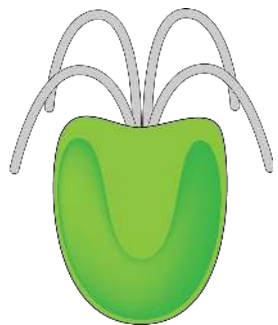
Rapaza viridis phylogeny



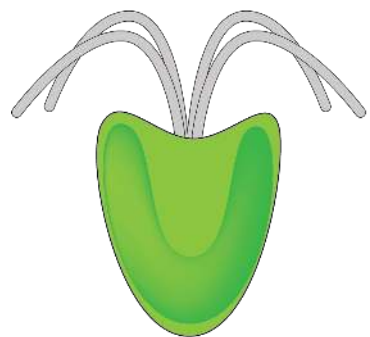
Pyramimonas
Prasinophytes



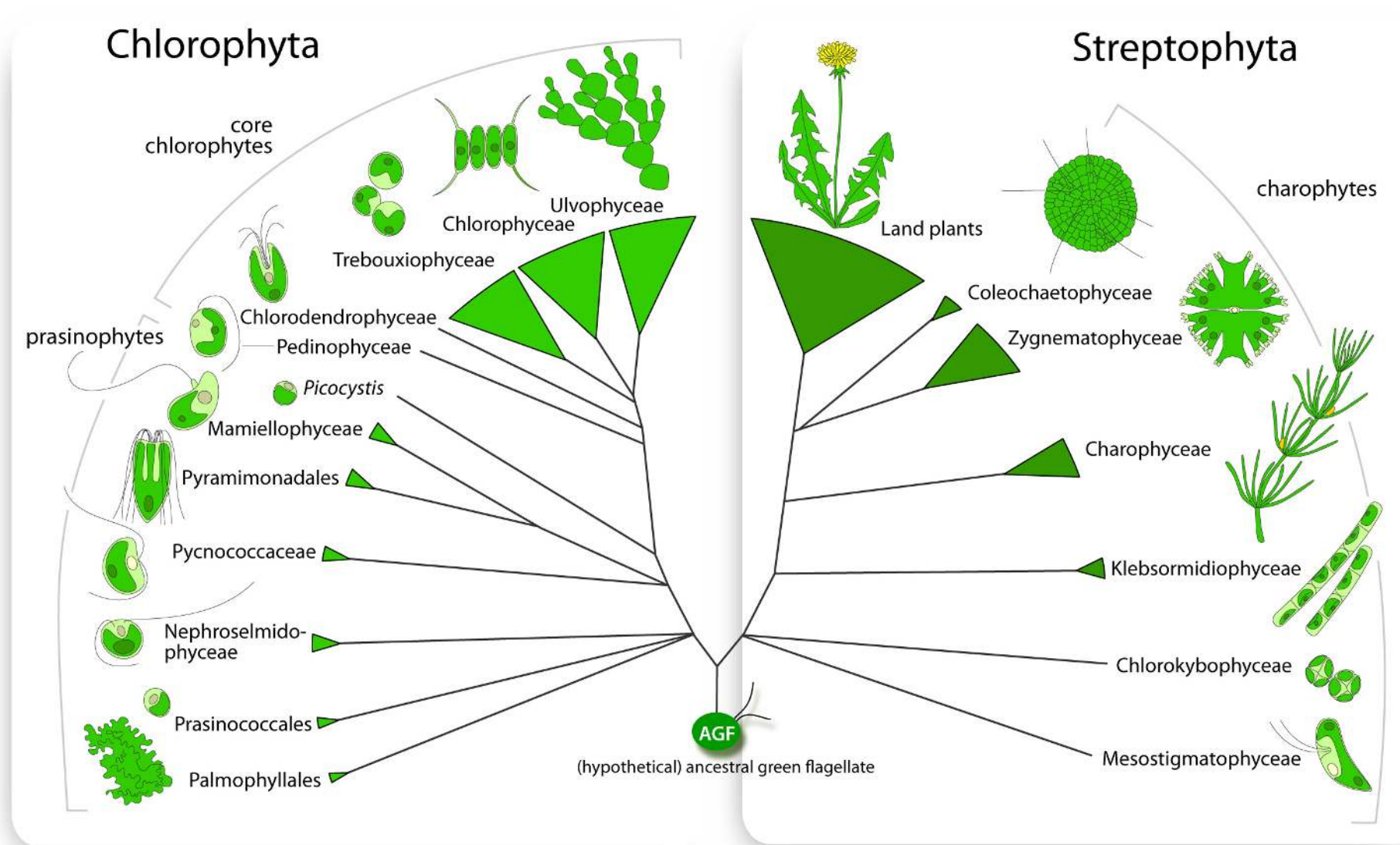
Tetraselmis
'core' chlorophytes

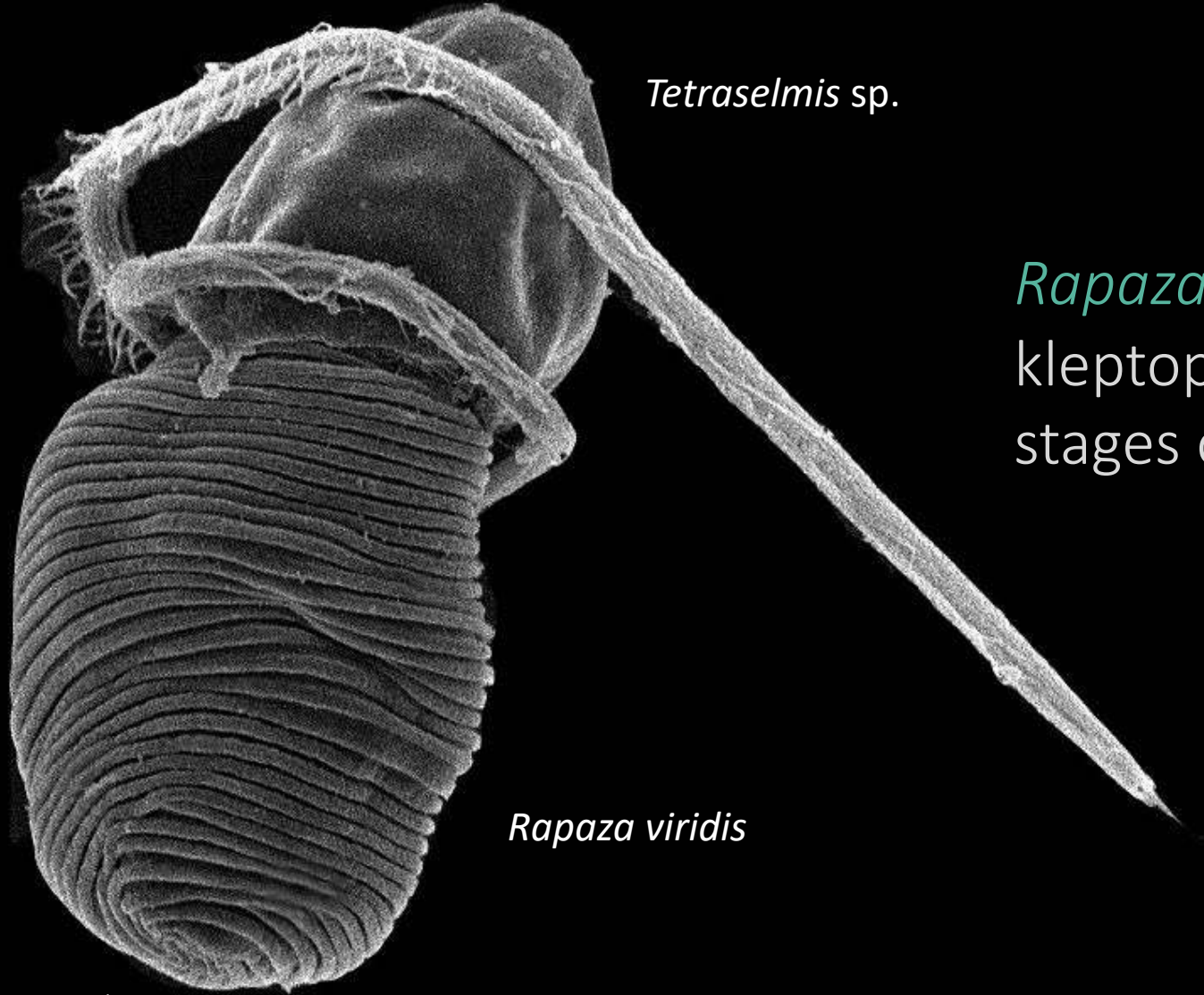


Tetraselmis
'core' chlorophytes



Pyramimonas
Prasinophytes



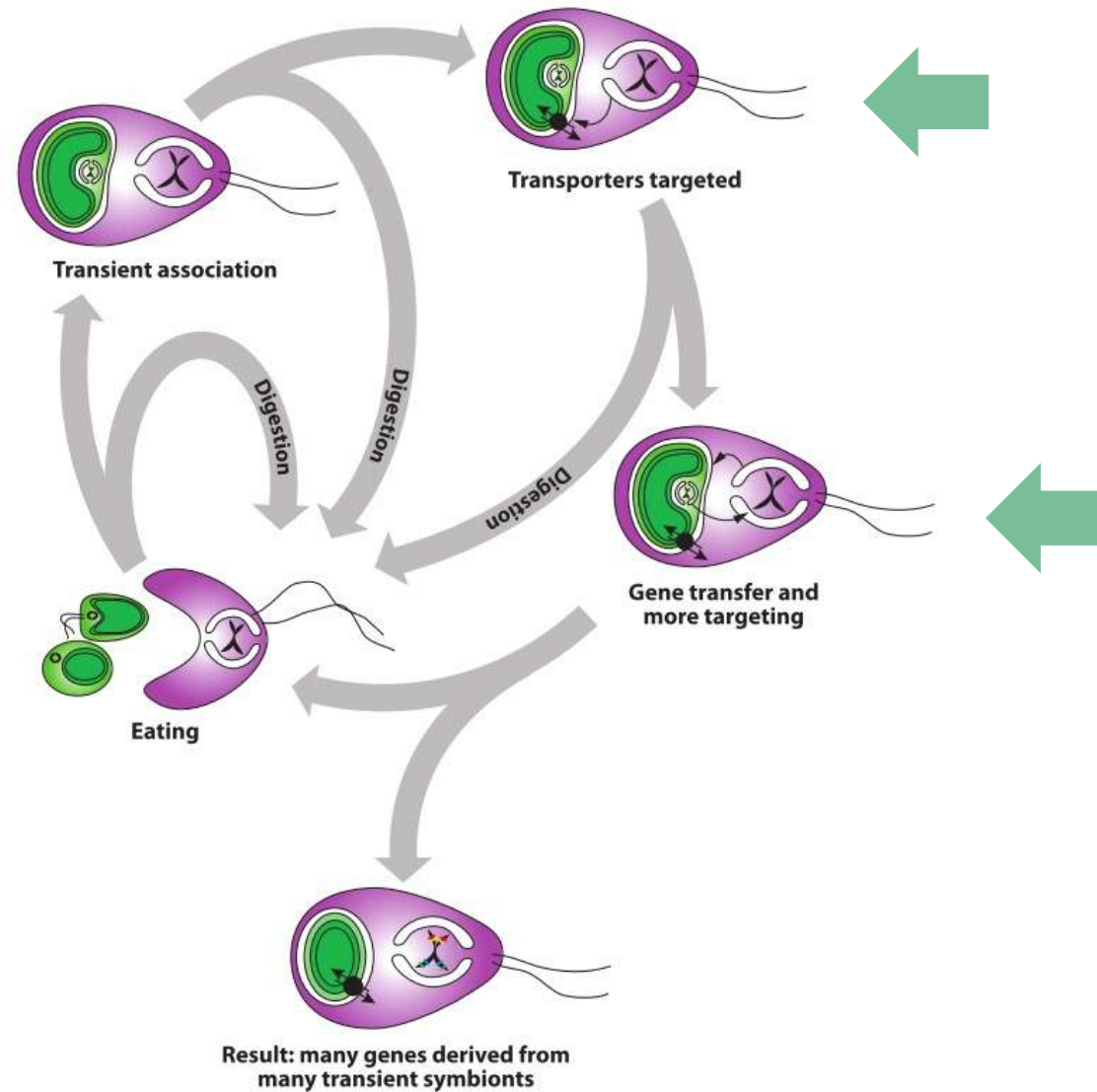


Tetraselmis sp.

Rapaza viridis

Rapaza viridis (Euglenophyceae) and its kleptoplast as a model for studying early stages of endosymbiosis

Which comes first, gene transfer or cellular fixation?



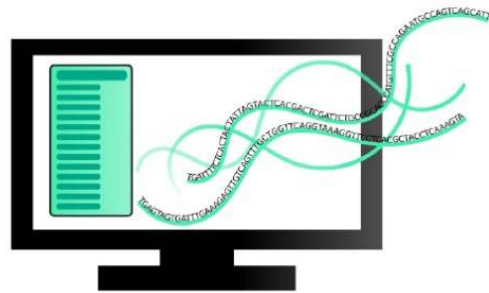


Genes encoding chloroplast proteins are transferred to nuclear genome via **endosymbiotic gene transfer** (EGT).

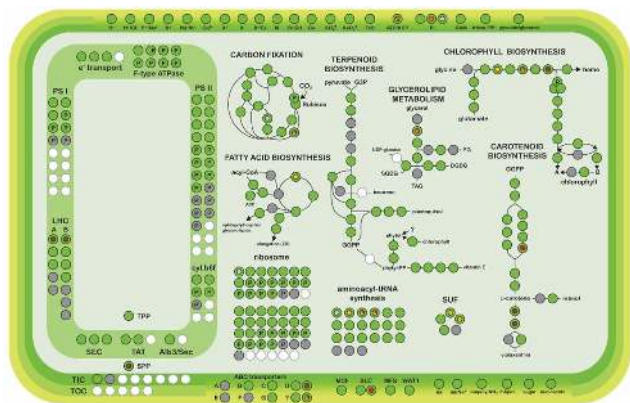
Plastid proteins encoded in nuclear genome are **targeted** to the plastids.

Searching for plastid-targeted proteins in *Rapaza* transcriptome

Sequencing and RNAseq analysis

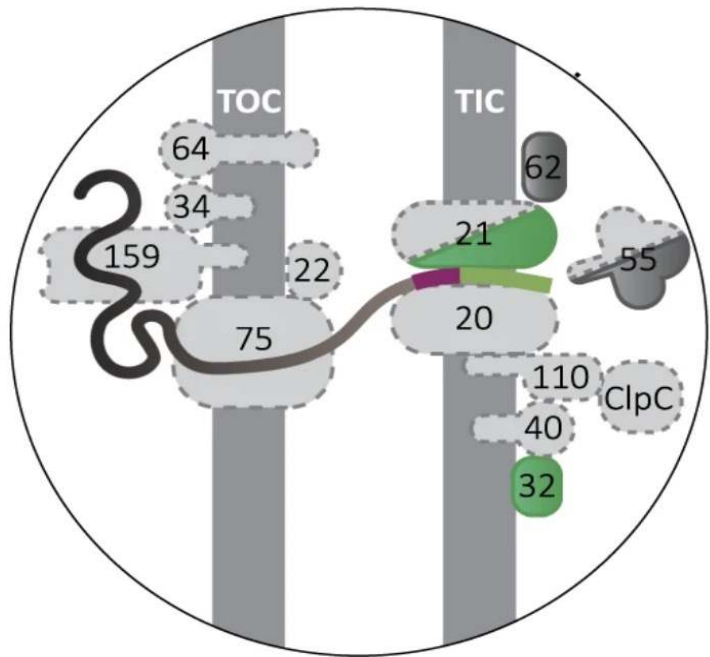


Plastid proteins identification



274 proteins

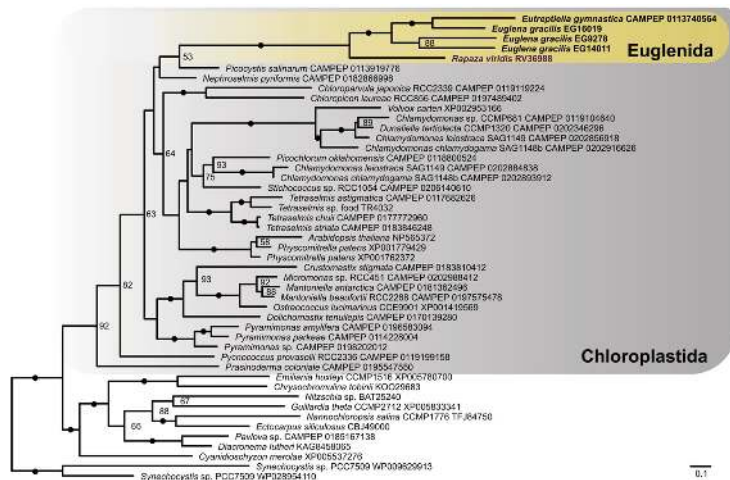
Searching for translocon components



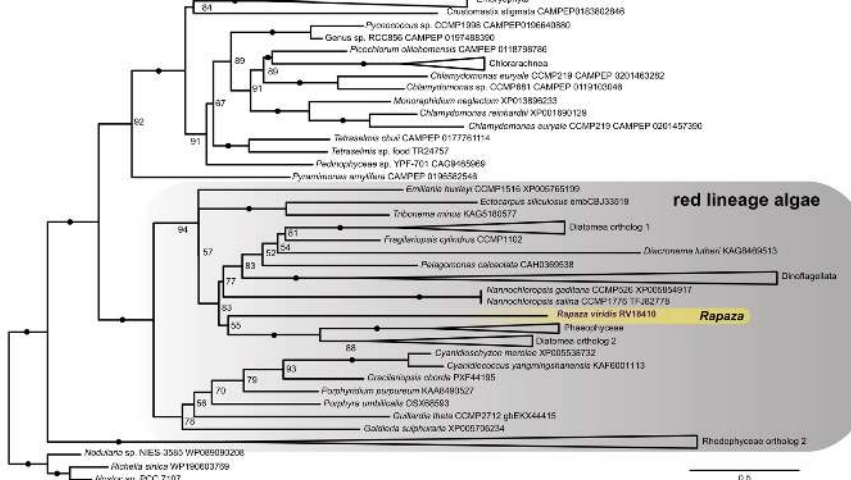
Euglena translocon components
Zahonova et al. 2018

Translocon components of *R. viridis*

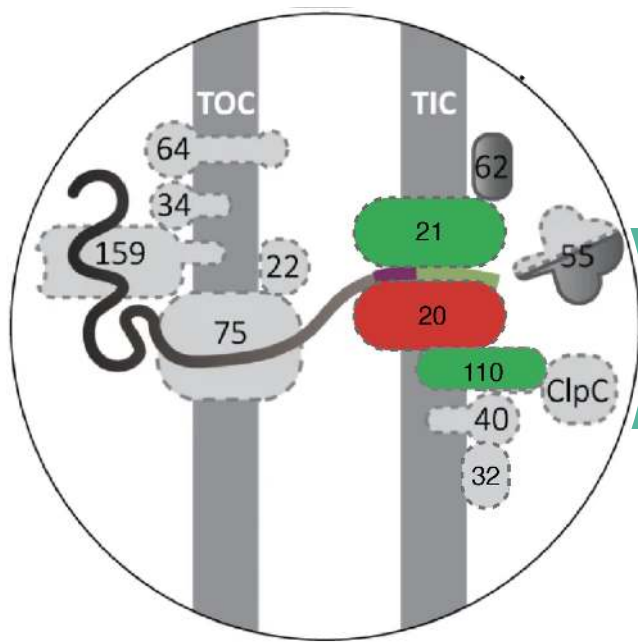
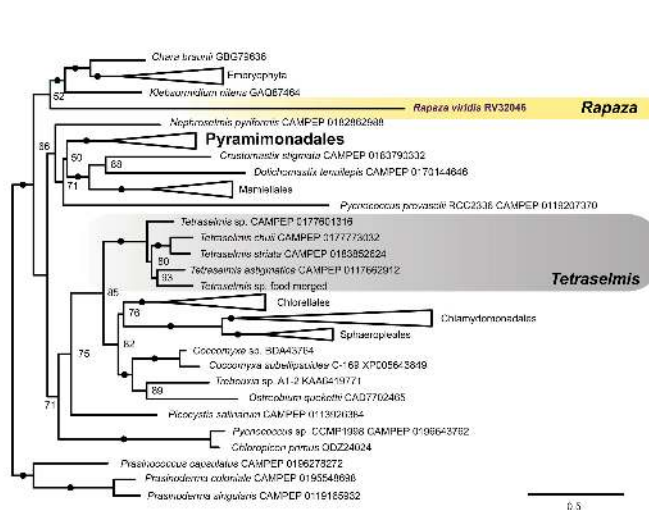
Tic21



Tic20



Tic110

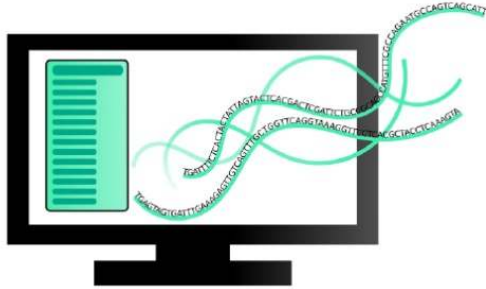




Are there any plastid proteins encoded in nuclear genome of *Rapaza* **targeted to kleptoplasts**?

Searching for plastid-targeted proteins in *Rapaza* transcriptome

Sequencing and RNAseq analysis



Targeting signal prediction

Yes!

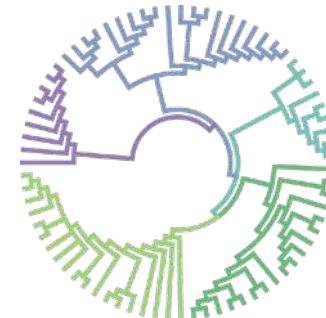


Plastid proteins identification

274 proteins

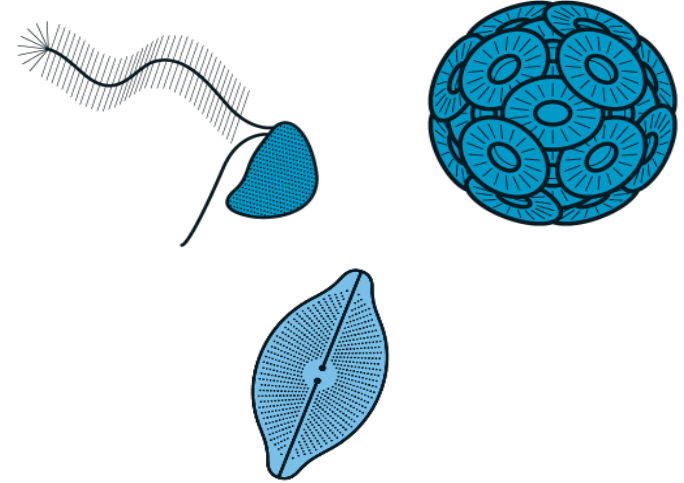
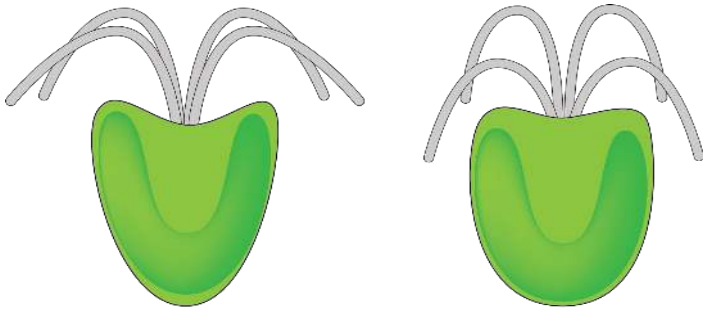


Phylogenetic analyses of candidate proteins



Searching for **plastid-targeted** proteins in *Rapaza* transcriptome

Phylogenetic analyses of candidate proteins



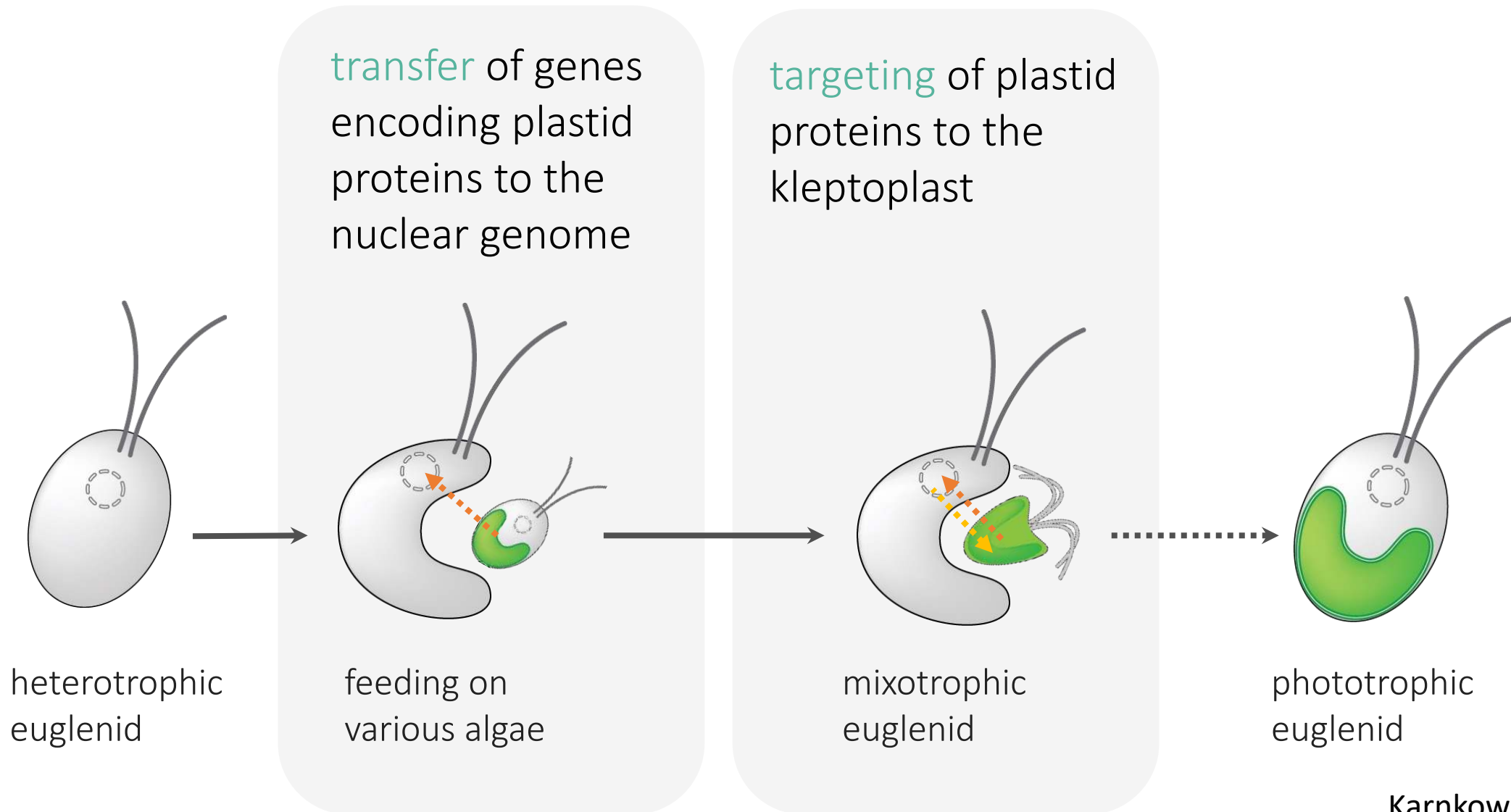
60% originated from green algae

40% from other evolutionary lineages

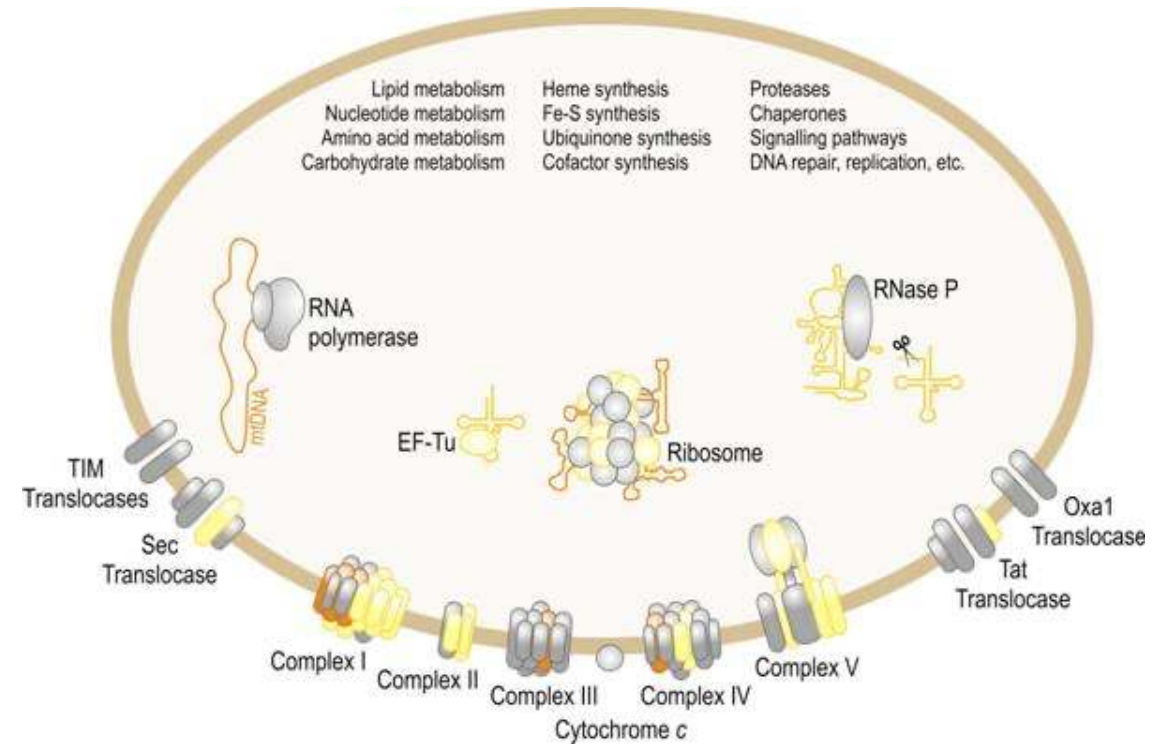
proteins derived from different lineages as the kleptoplast indicate that transfer of genes preceded establishment of the endosymbiont

origin of plastids via kleptoplasty

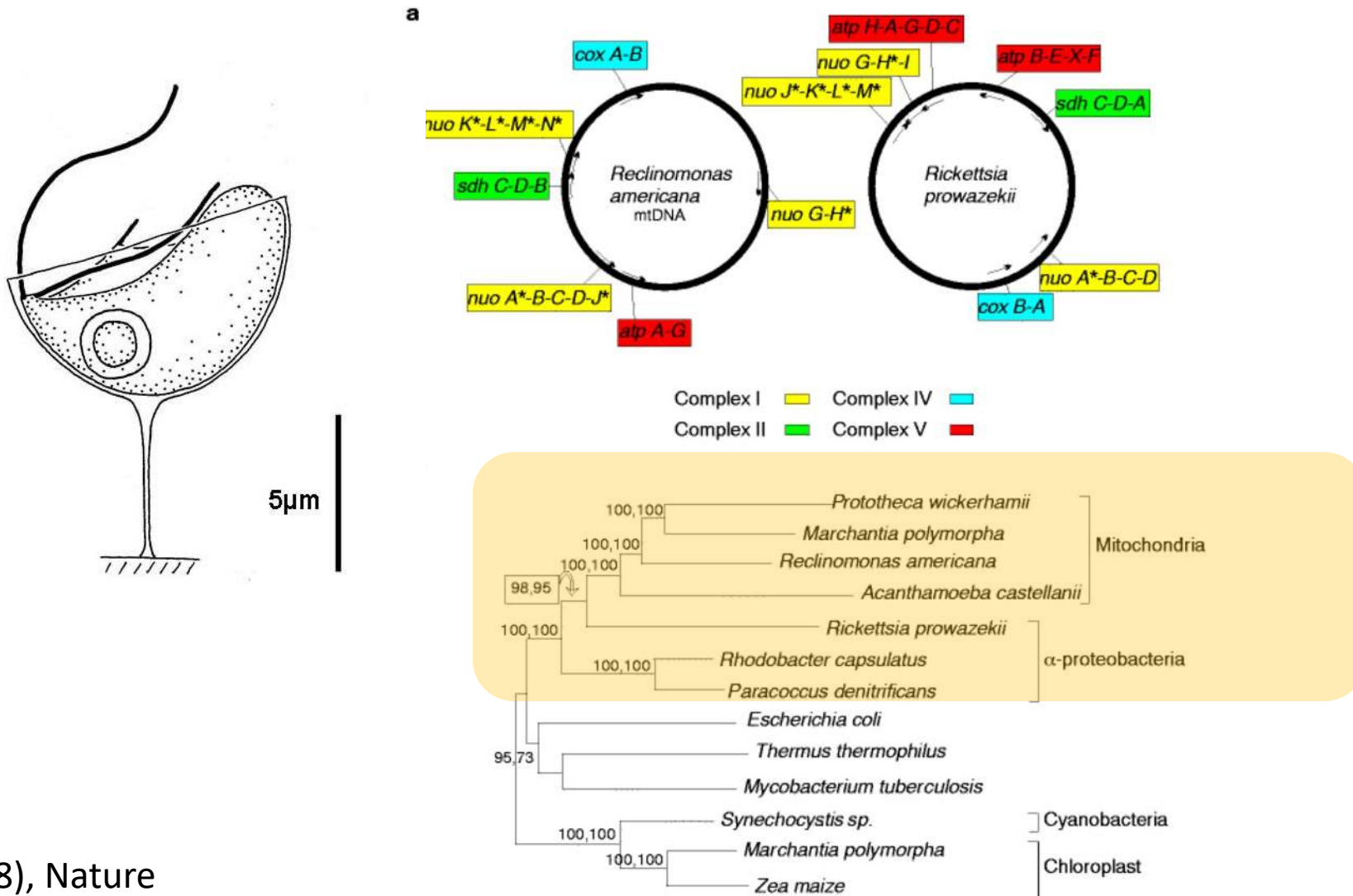
gene transfer comes before cellular fixation



mitochondria

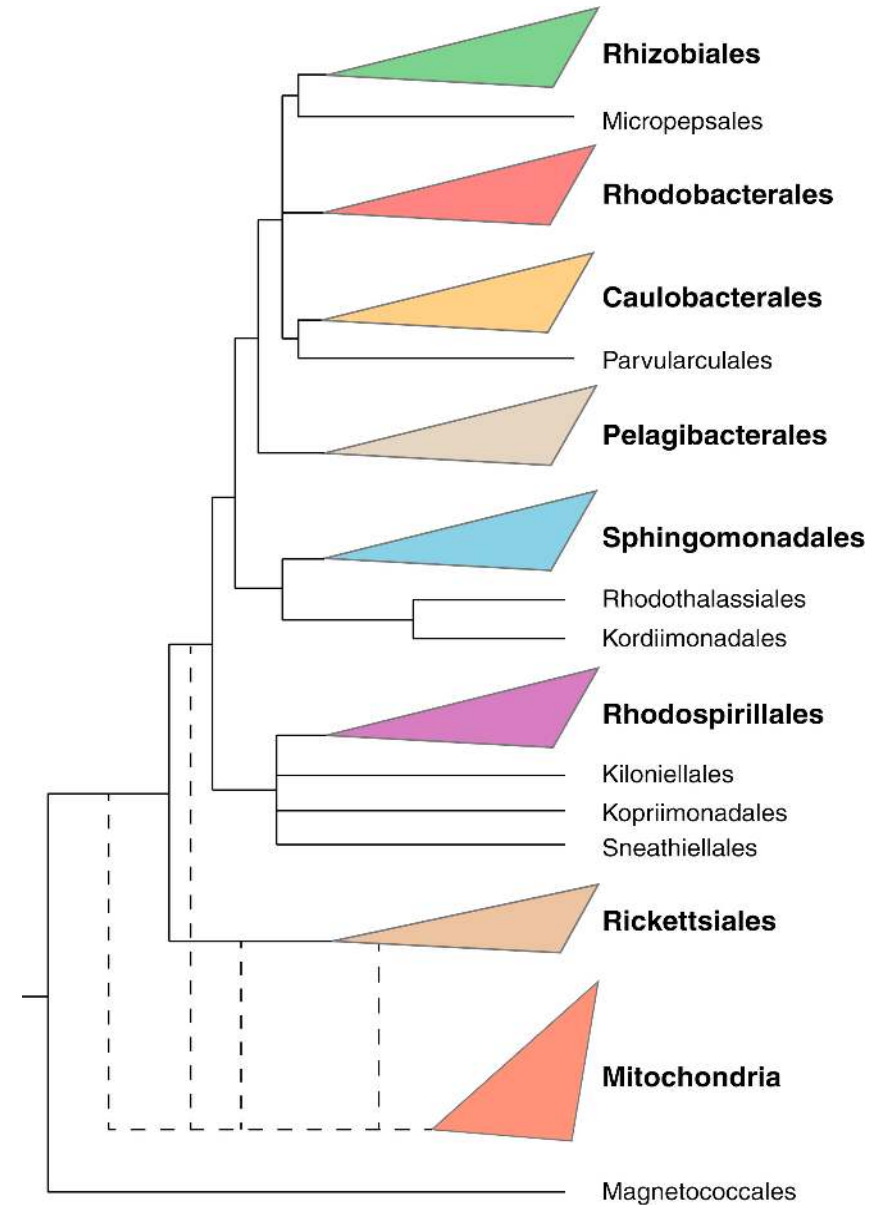


The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria

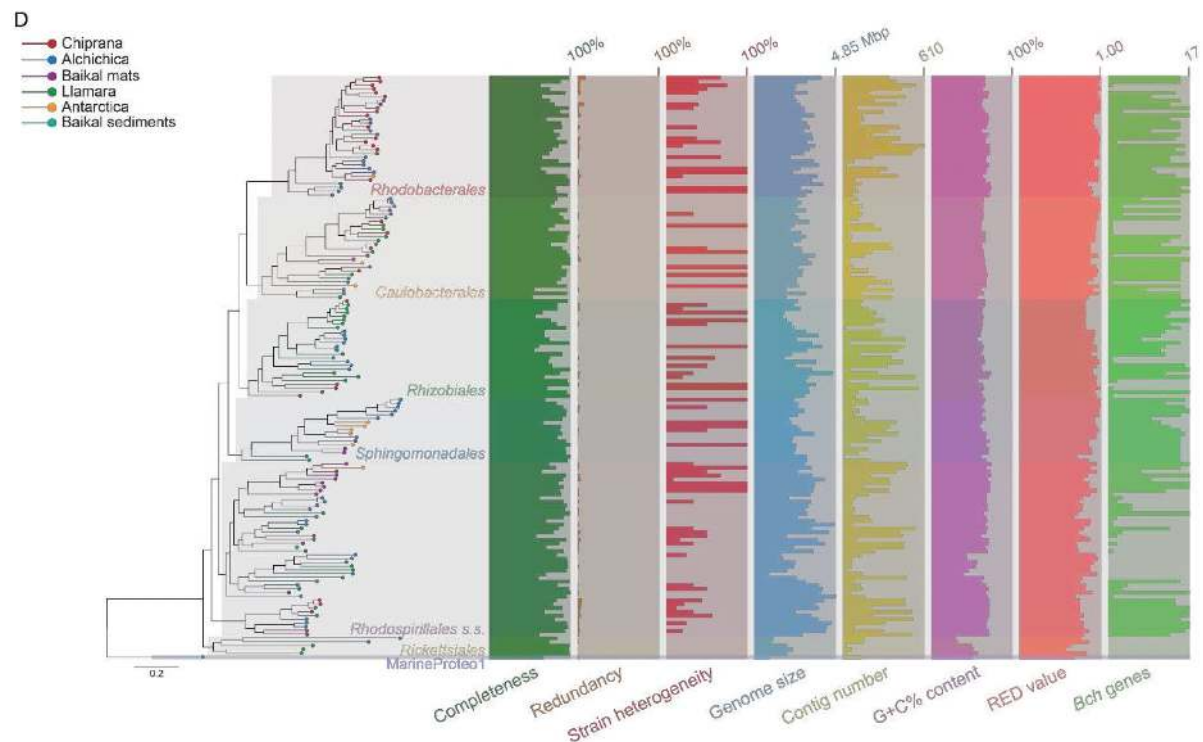
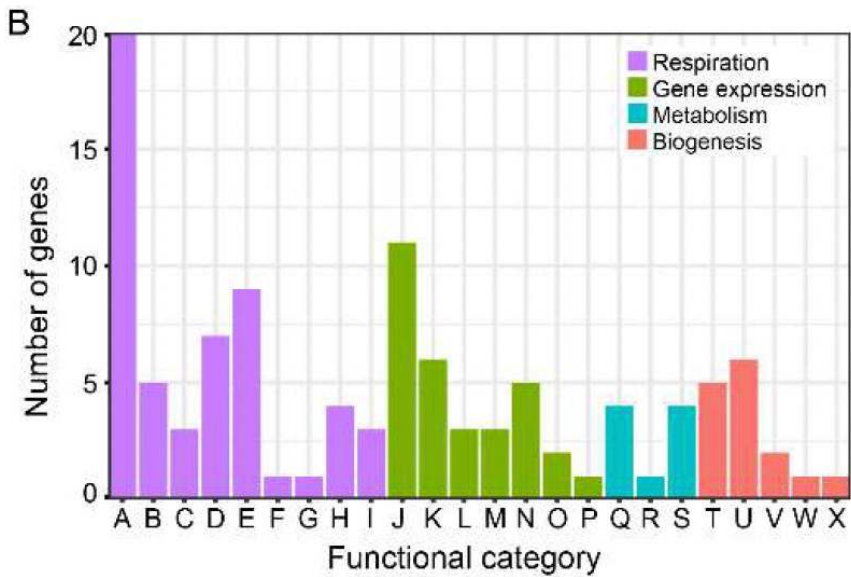
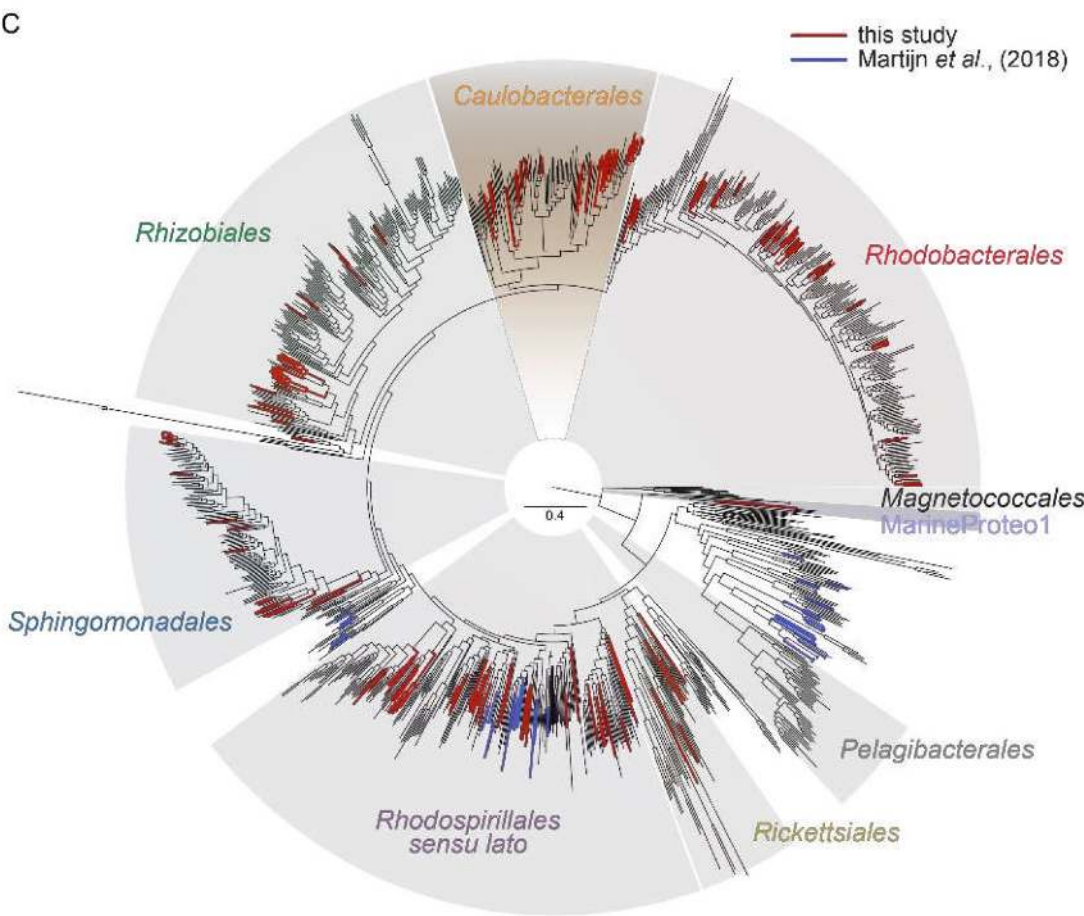


Alphaproteobacterial phylogenomics and origin of mitochondria

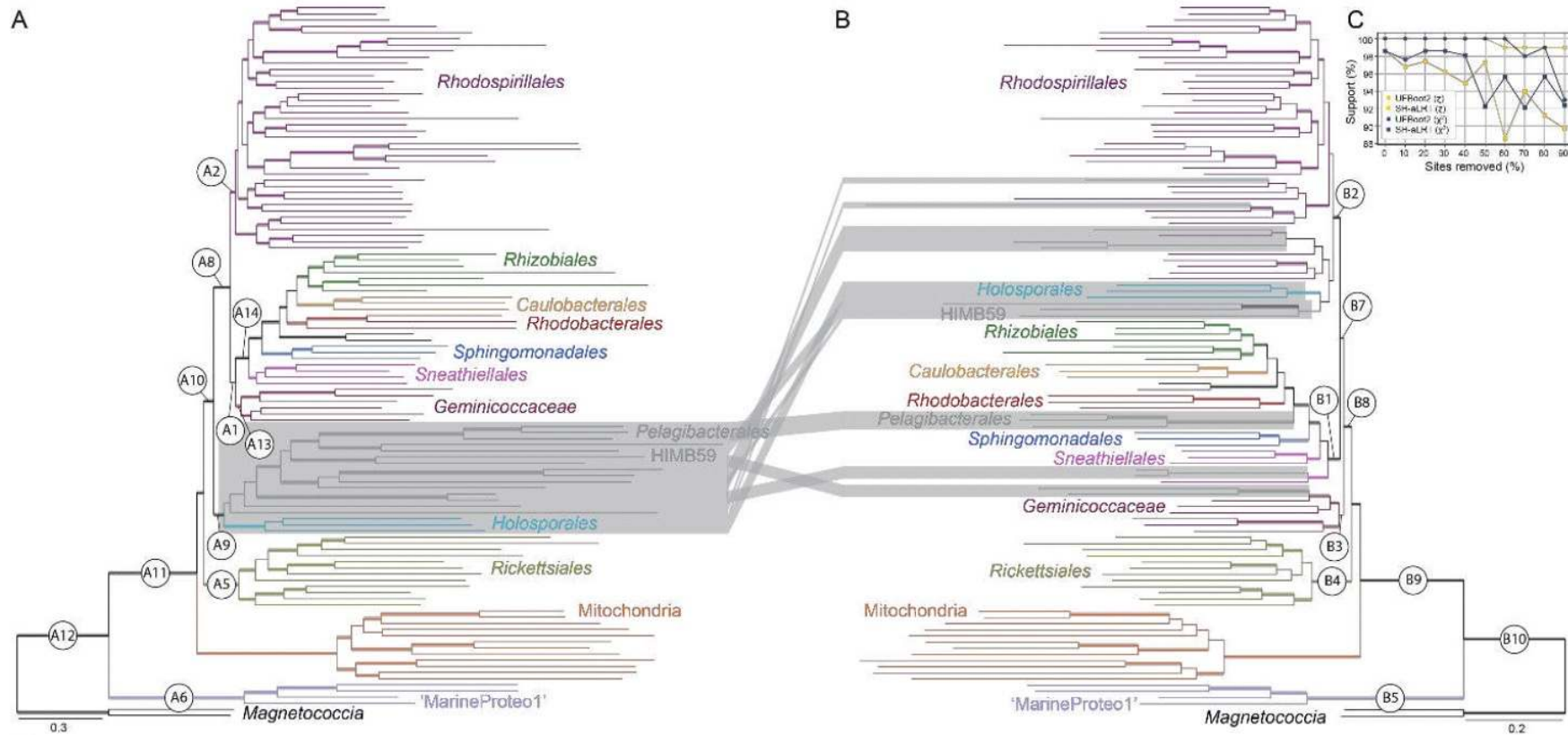
- The tree of alphaproteobacteria is not resolved
- Number of genes of alphaproteobacterial origin in mitochondrial genomes is limited (not much phylogenetic signal)



Environmental MAGs allow to expand the taxa sampling



Expanded dataset favour mitochondria as sister to known Alphaproteobacteria



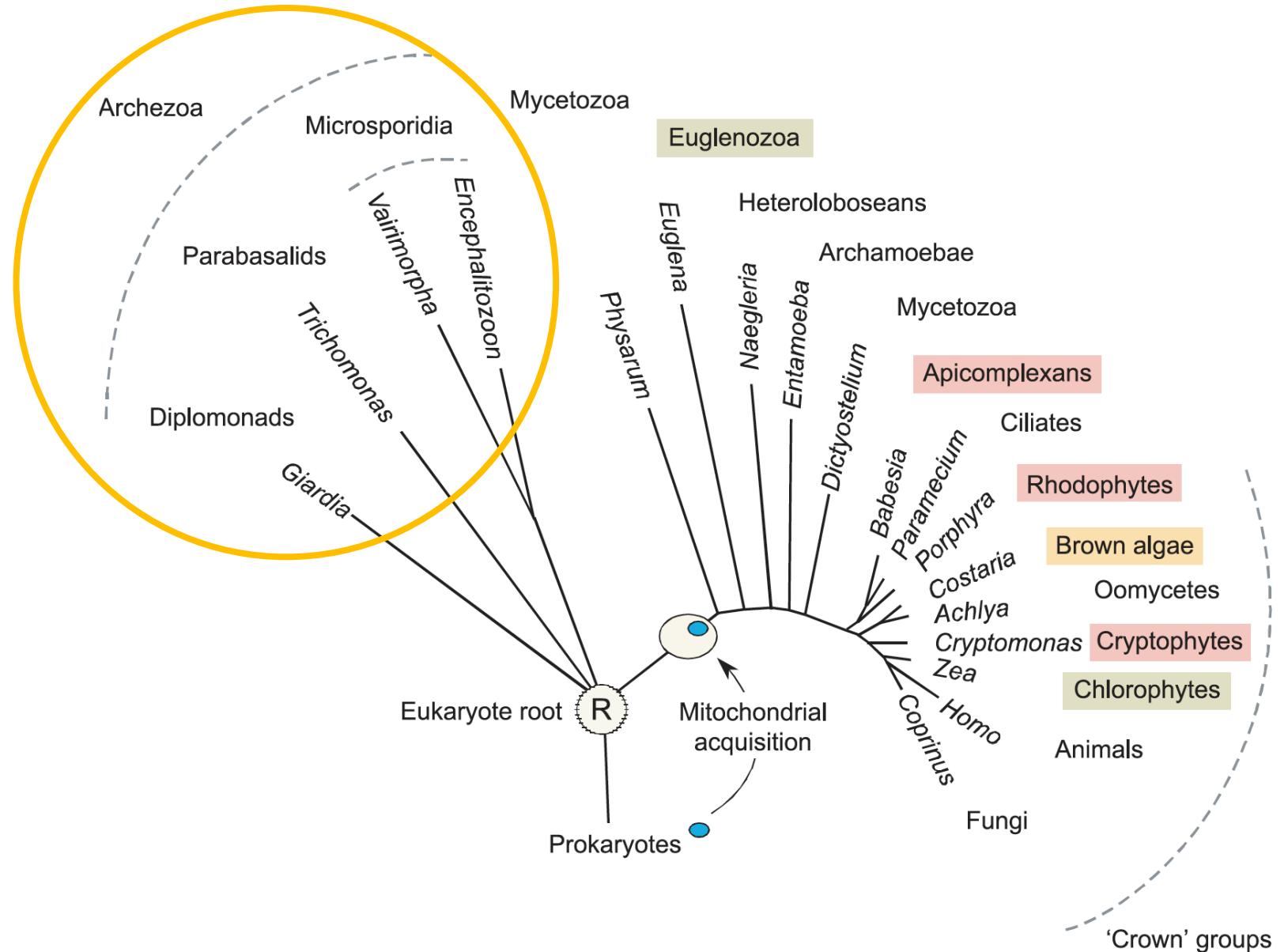
(A) Phylogenetic tree for the *Alphaproteobacteria* and mitochondria derived from a site-heterogeneous analysis of an untreated dataset.

(B) Phylogenetic tree for the *Alphaproteobacteria* and mitochondria derived from a site-heterogeneous analysis of a dataset from which 50% of the most compositionally heterogeneous sites had been removed.

Which came first nucleus or mitochondrion?

Archezoa hypothesis

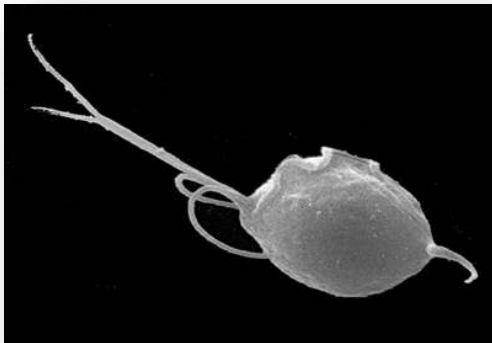
Thomas Cavalier-Smith, 1989



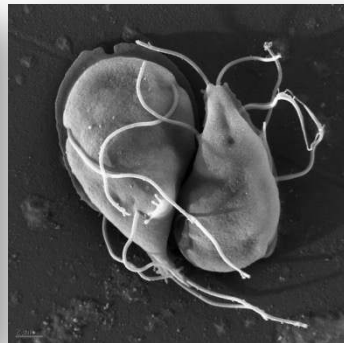
Embley and Martin (2006) Nature

Archezoa

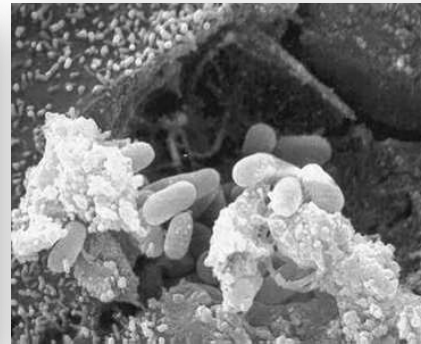
- early branching eukaryotes
- lack of introns
- no sexual reproduction
- lack of peroxisomes
- lack of Golgi apparatus
- lack of mitochondria



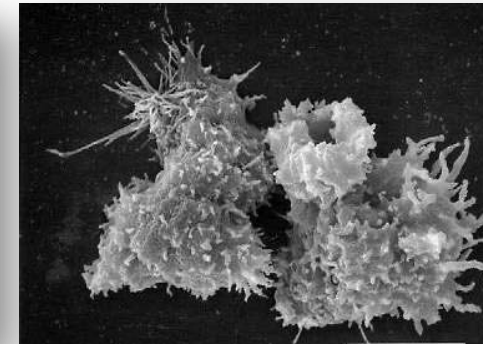
Trichomonas



Giardia



Microsporidia

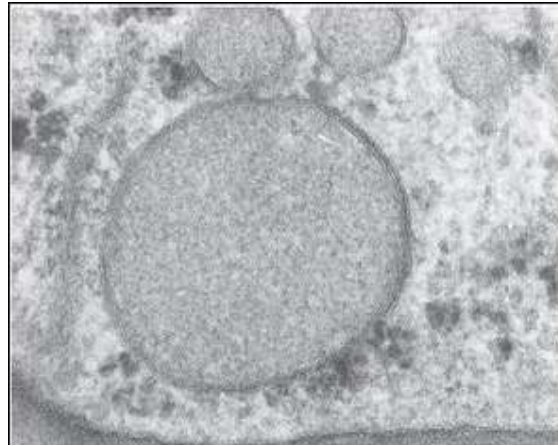


Archamoebae

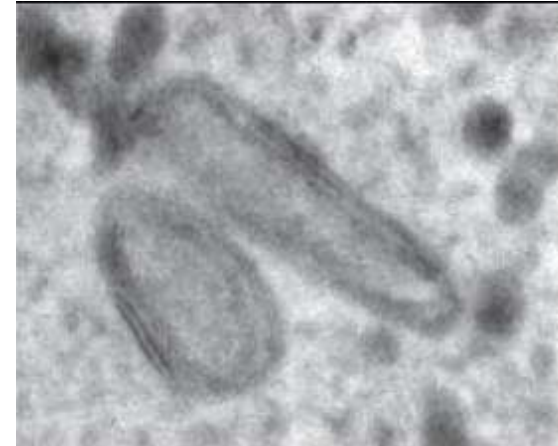
Archaezoa possess mitochondria related organelles (MROs)



mitochondrion



hydrogenosome

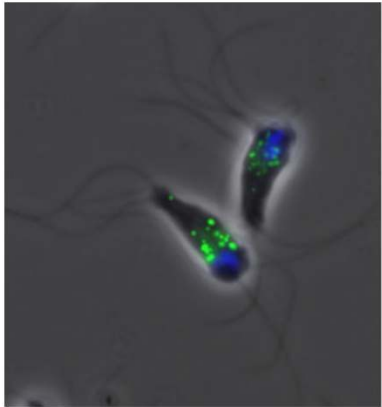


mitosomes

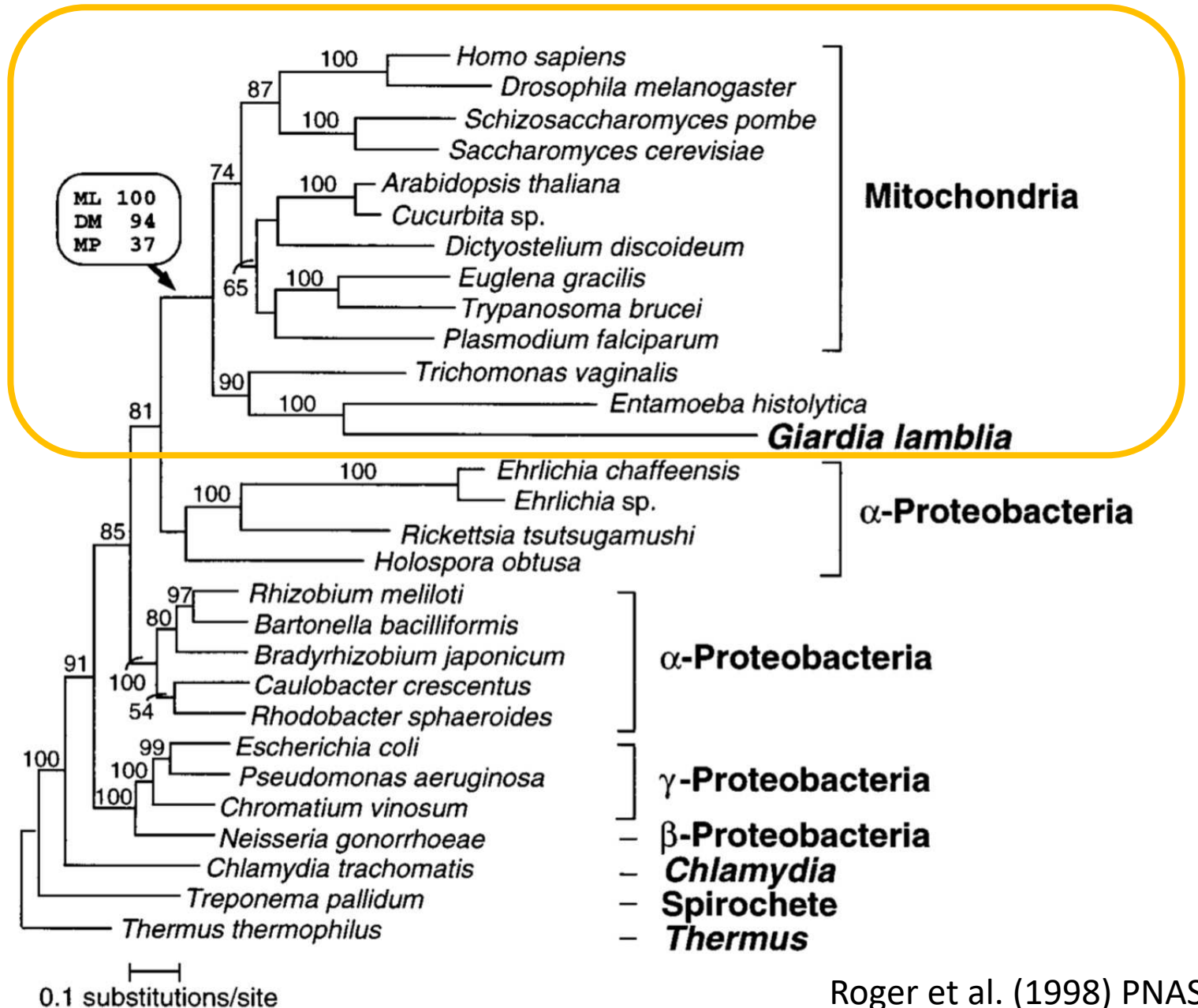
Common origin of mitochondria and MROs

Cpn60 chaperonin

- mitochondrial chaperonins
- ATP transporters
- mitochondrial membrane transport proteins
- Fe-S cluster assembly proteins



cpn60 localization
Jerlström-Hultqvist et al. 2013



Roger et al. (1998) PNAS

MROs are widespread at the tree of eukaryotes



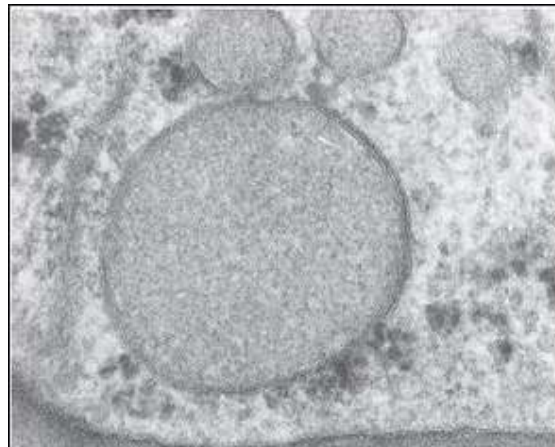
Archezoa hypothesis rejected

- All “archezoa” possess:
 - mitochondrial genes in nuclear genomes
 - degenerate derivatives of mitochondria
 - they do not group together on the modern tree of life

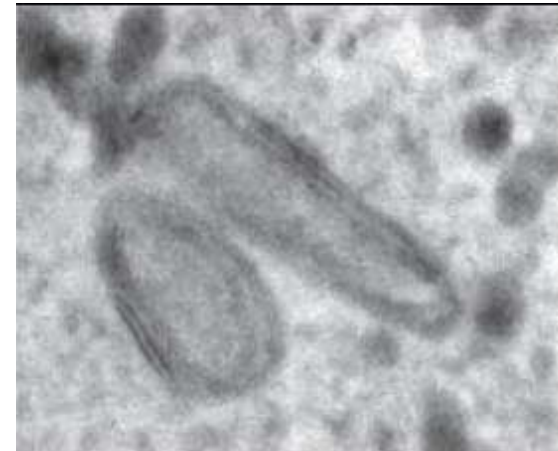
Common ancestor of all eukaryotes possessed mitochondria



mitochondrion

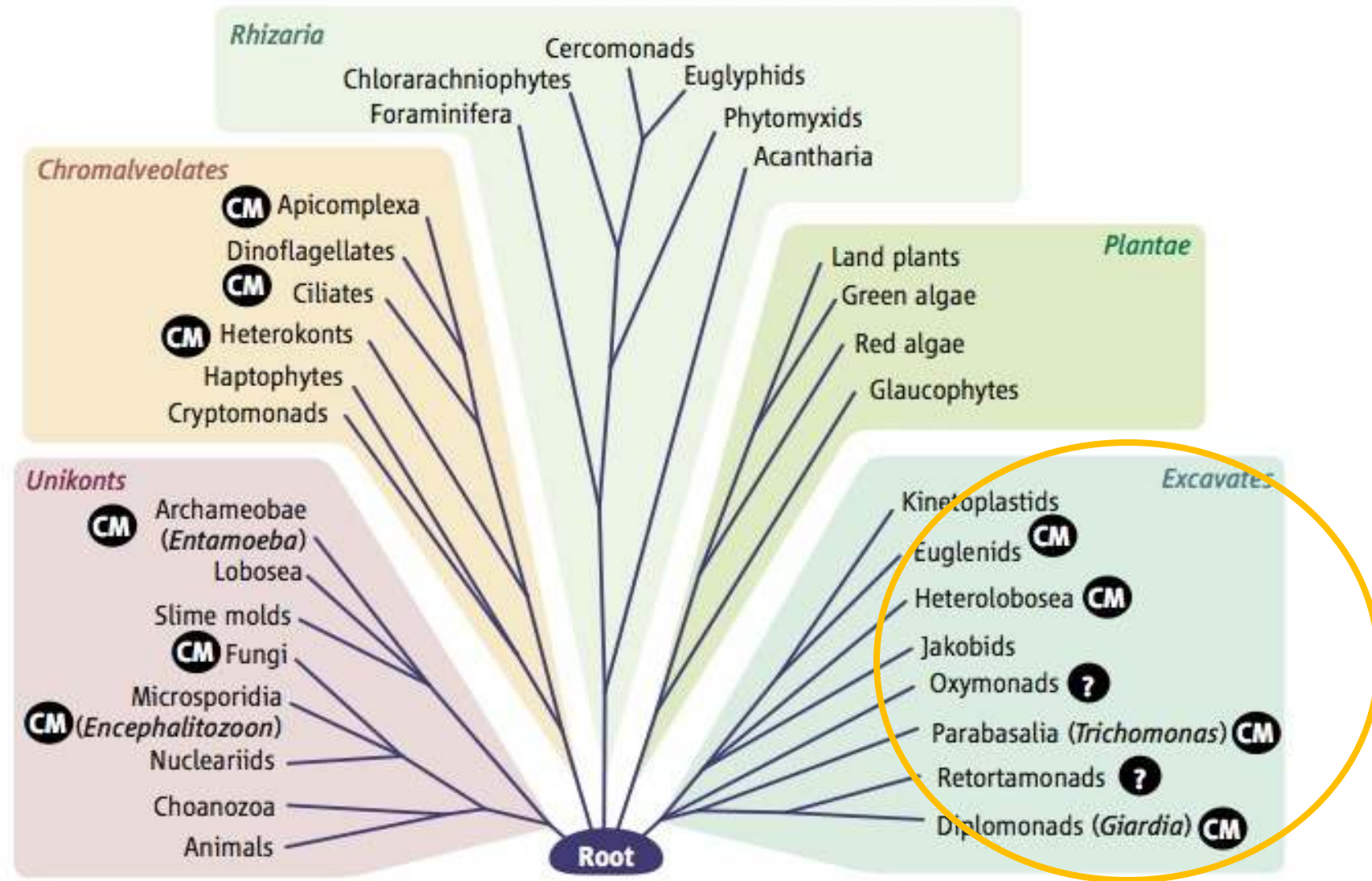


hydrogenosome



mitosomes

Does amitochondriate eukaryote exist?

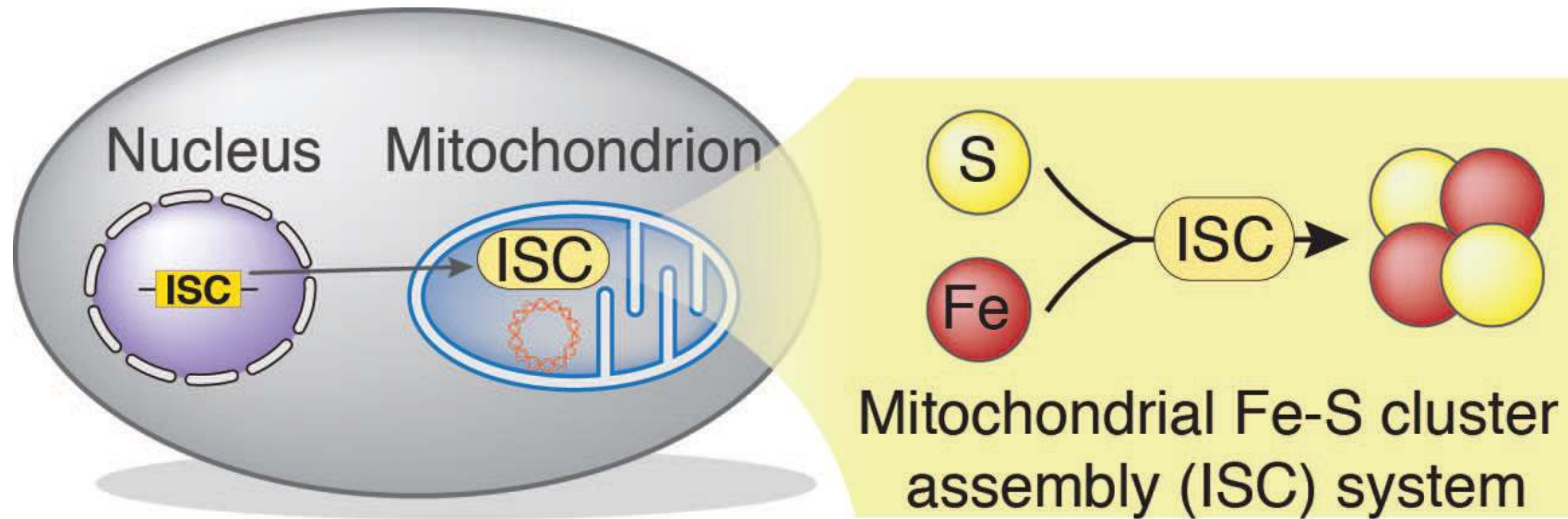


CM

cryptic mitochondria

Keeling, 2007

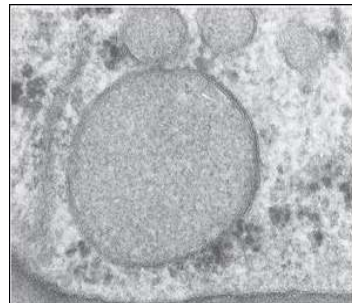
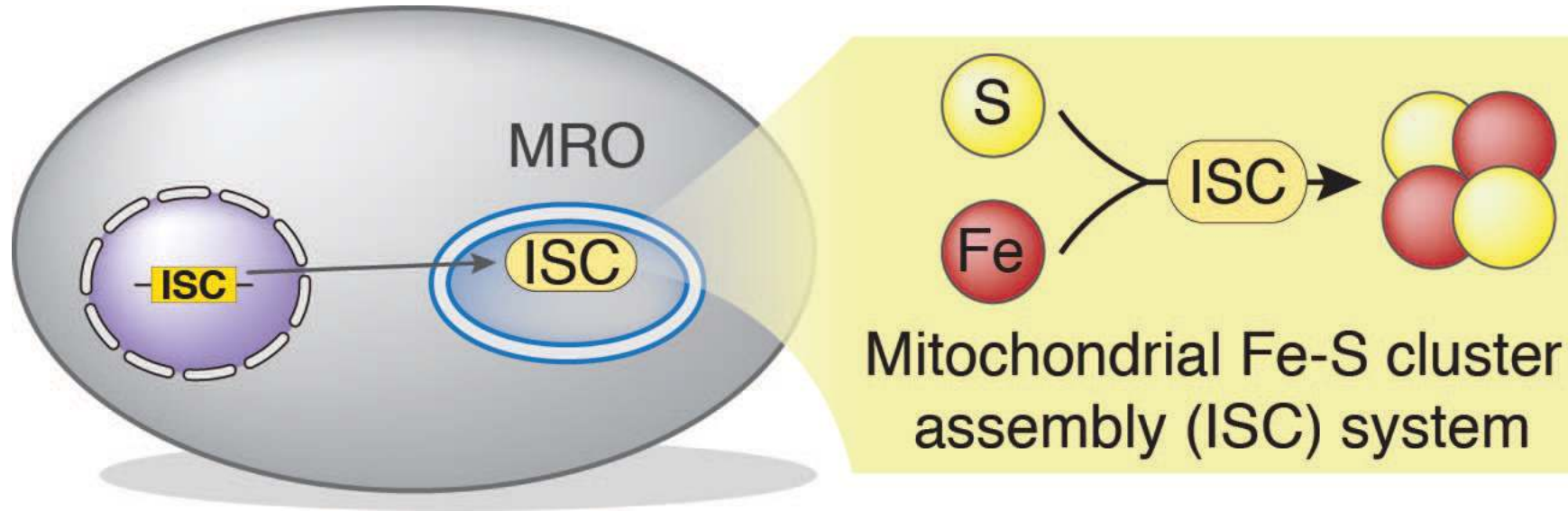
A eukaryotic cell with typical mitochondria



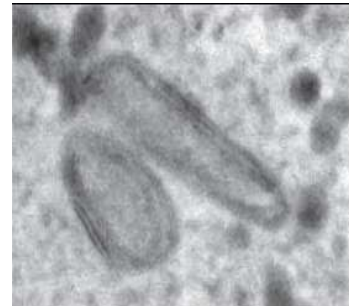
Oxidative phosphorylation, and other essential processes such as Fe-S cluster formation take place in mitochondria

Fe-S clusters are cofactors of essential enzymes

An anaerobic eukaryote with mitochondrion related organelles (MROs)



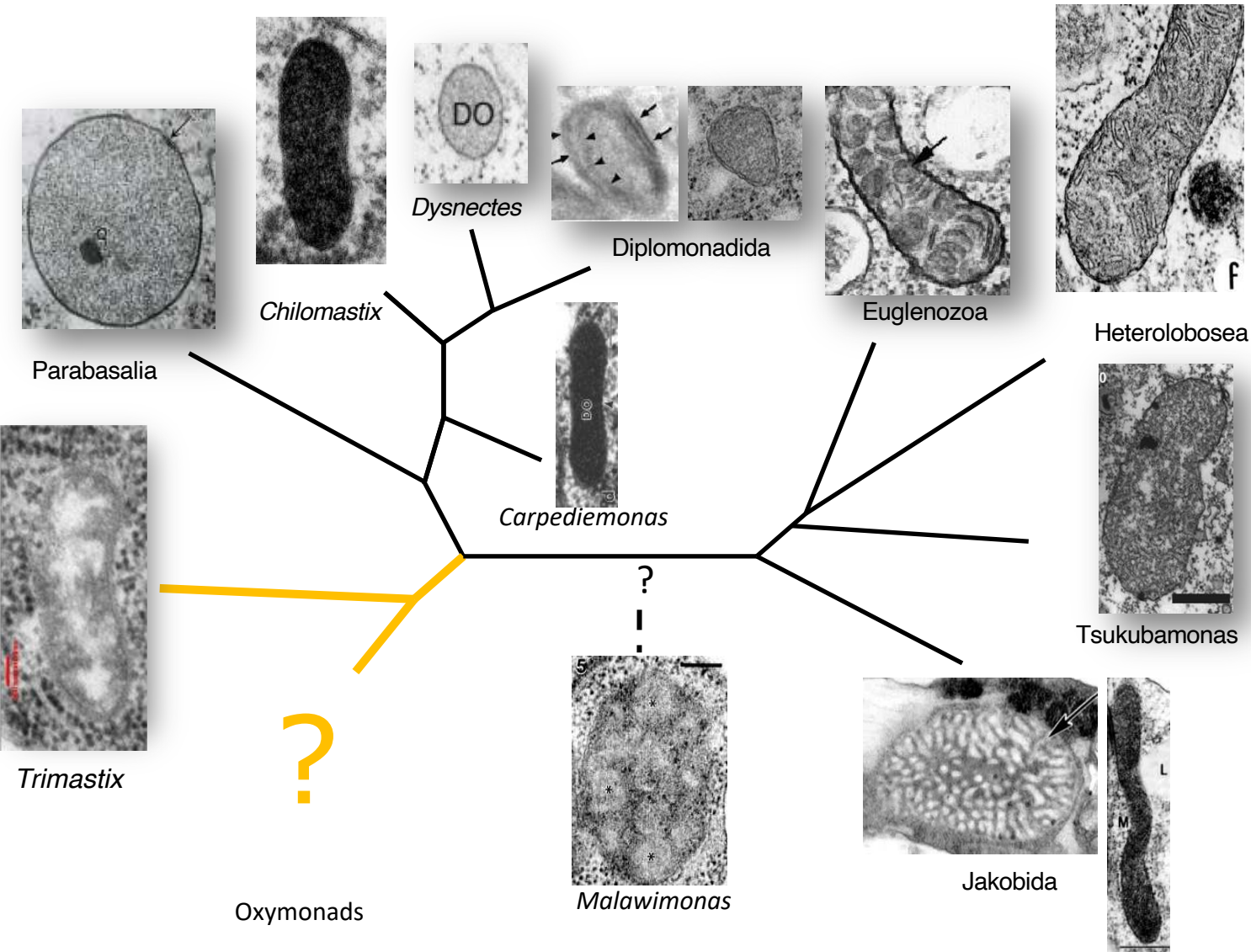
hydrogenosome

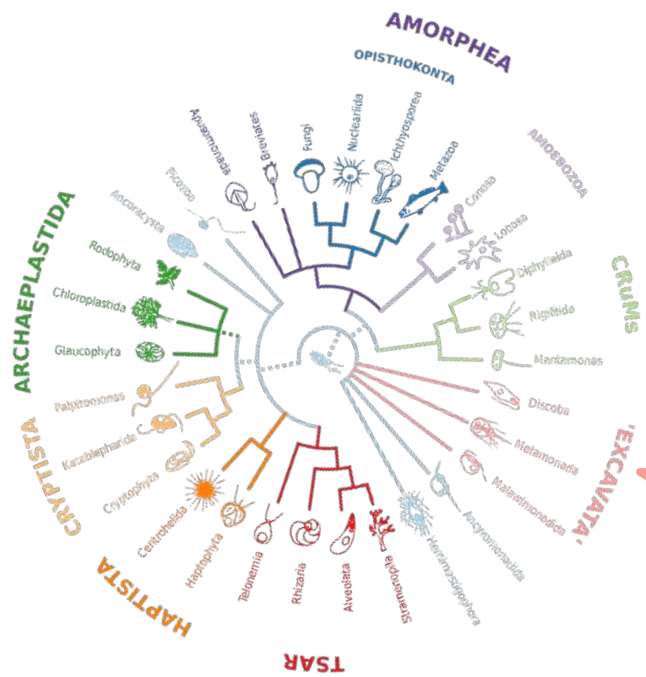


mitosome

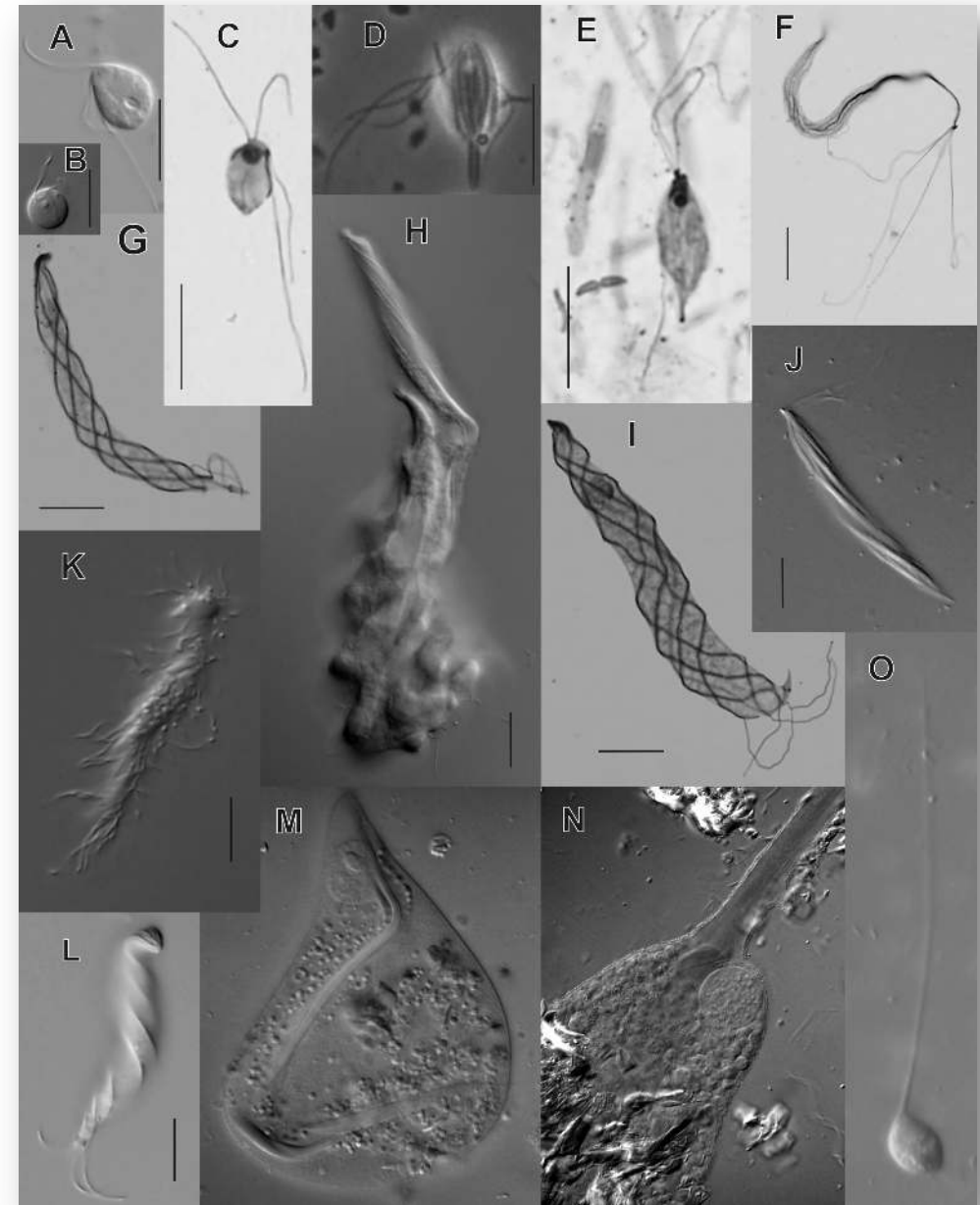
Loss of oxidative phosphorylation in anaerobic eukaryotes

mitochondria in 'Excavata'



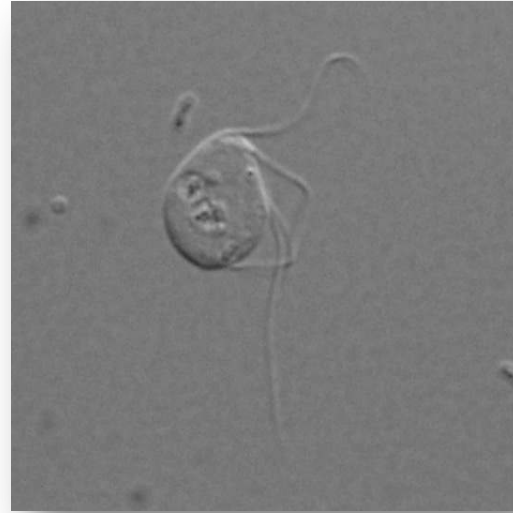
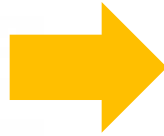


Metamonada
 Parabasalia (*Trichomonas*)
 Fornicata (*Giardia*)
 Preaxostyla
 Trimastix
 oxymonads



- found in the intestinal tracts of animals
- sexual reproduction debatable
- no peroxisomes
- no Golgi apparatus
- no mitochondria

Monocercomonoides - microaerophilic, commensal of animals

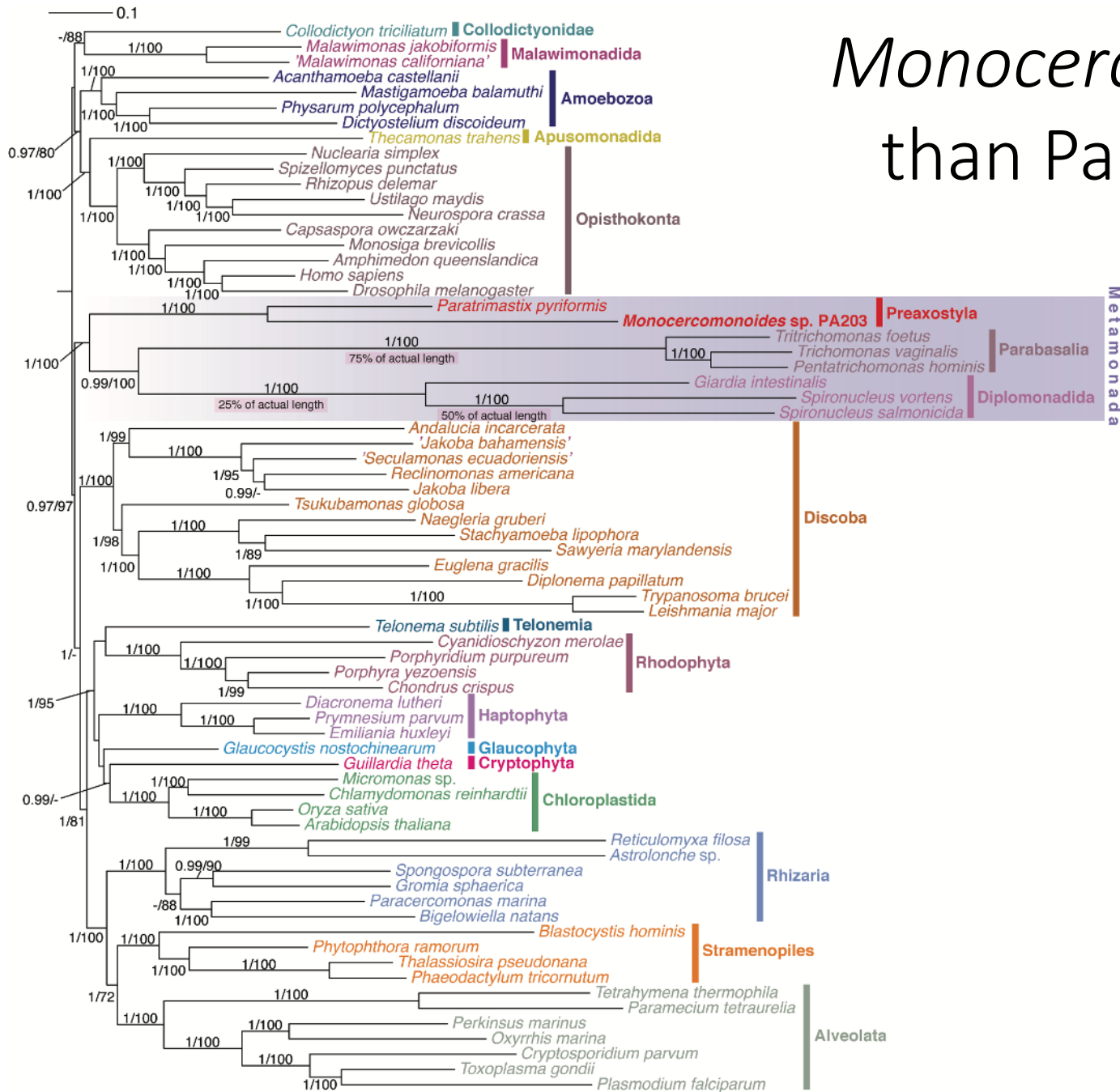


lack of mitochondria in *Monocercomonoides* under the microscope



genome analysis of *Monocercomonoides*

Monocercomonoides is less divergent than Parabasalids and Diplomonads

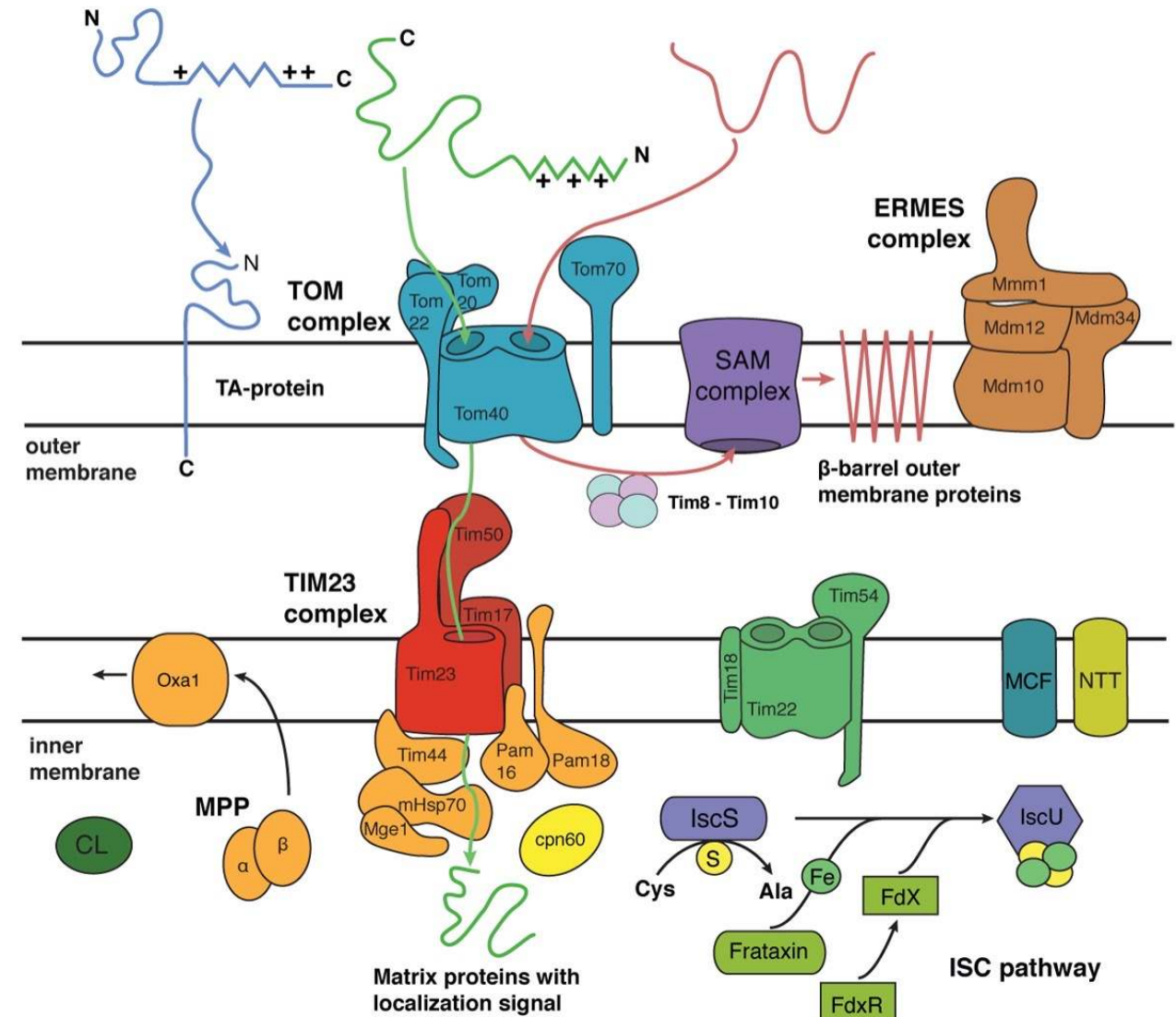


Searching for mitochondrial proteins

Mitochondrial outer membrane targeted proteins (TA)

Proteins with mitochondrial localization signal

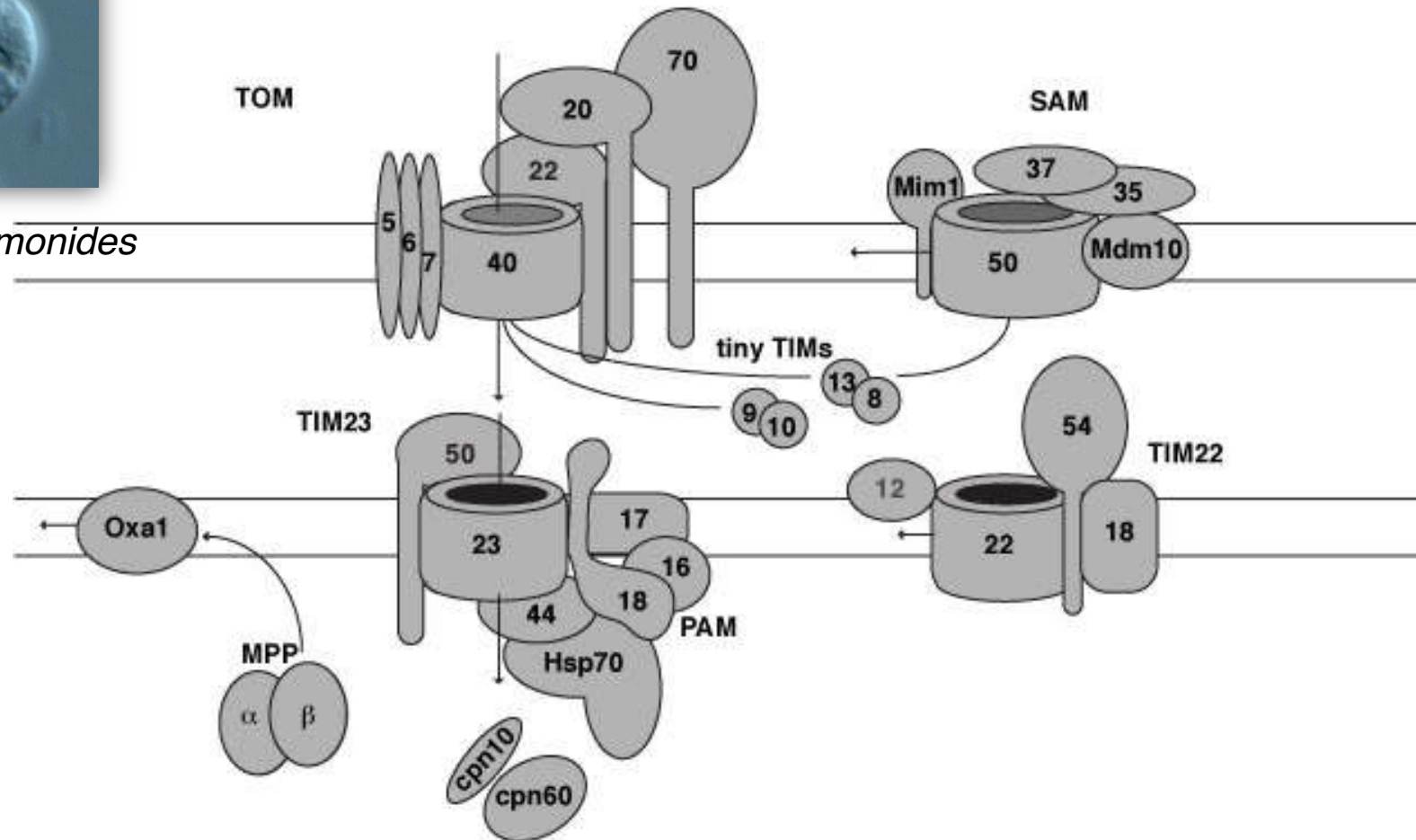
β -barrel mitochondrial outer membrane proteins



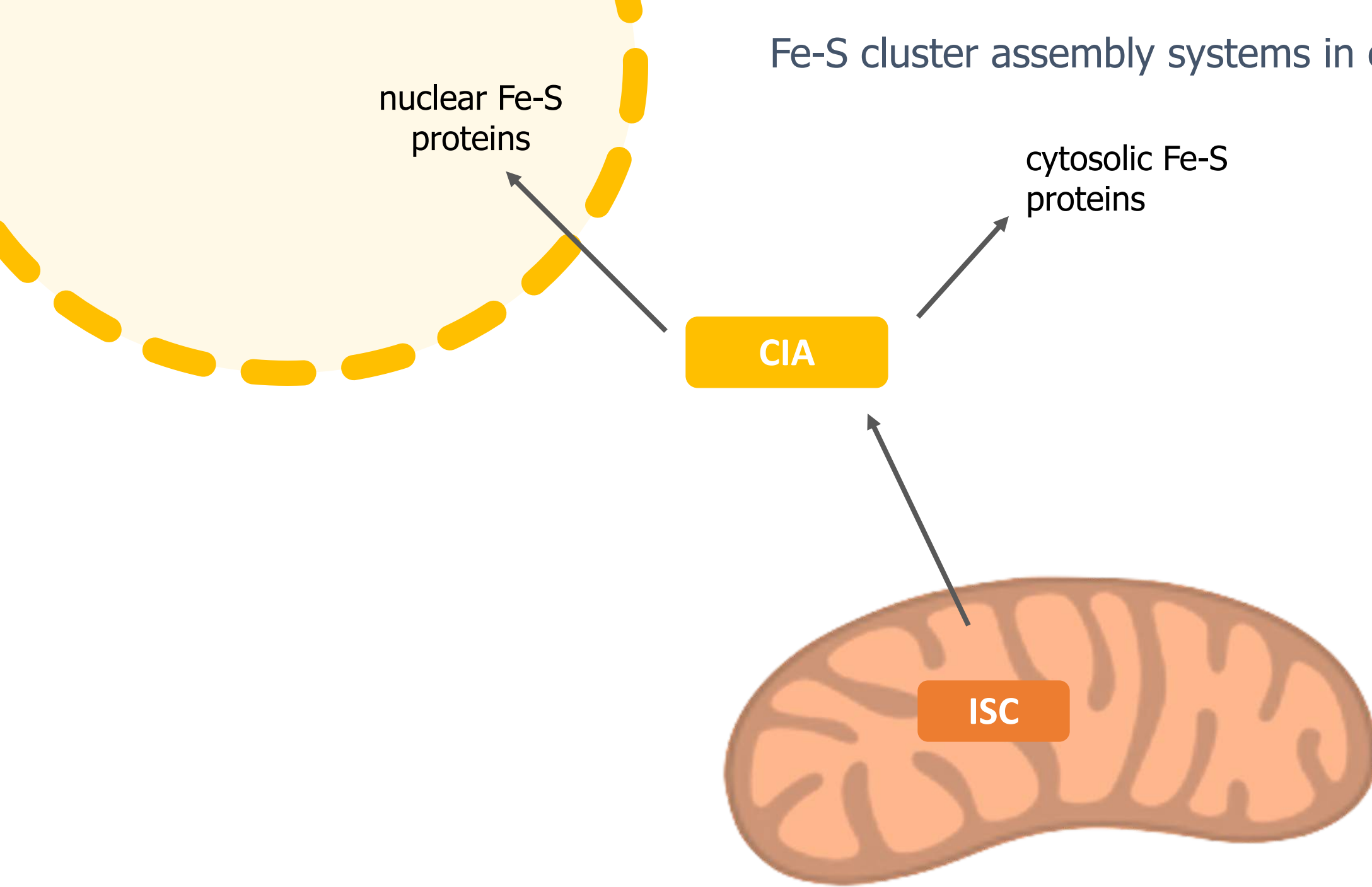
mitochondrial membrane transport proteins



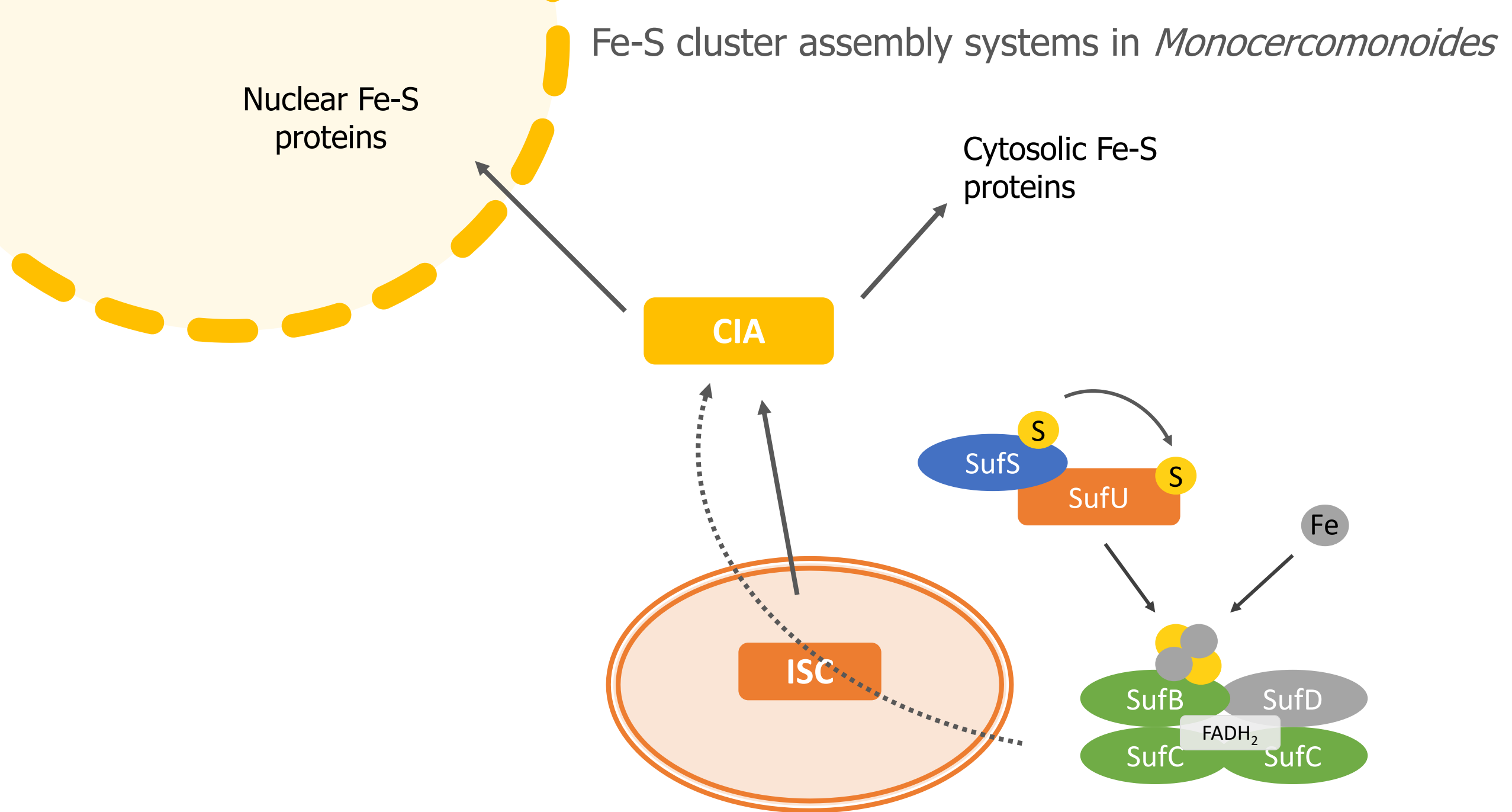
Monocercomonoides



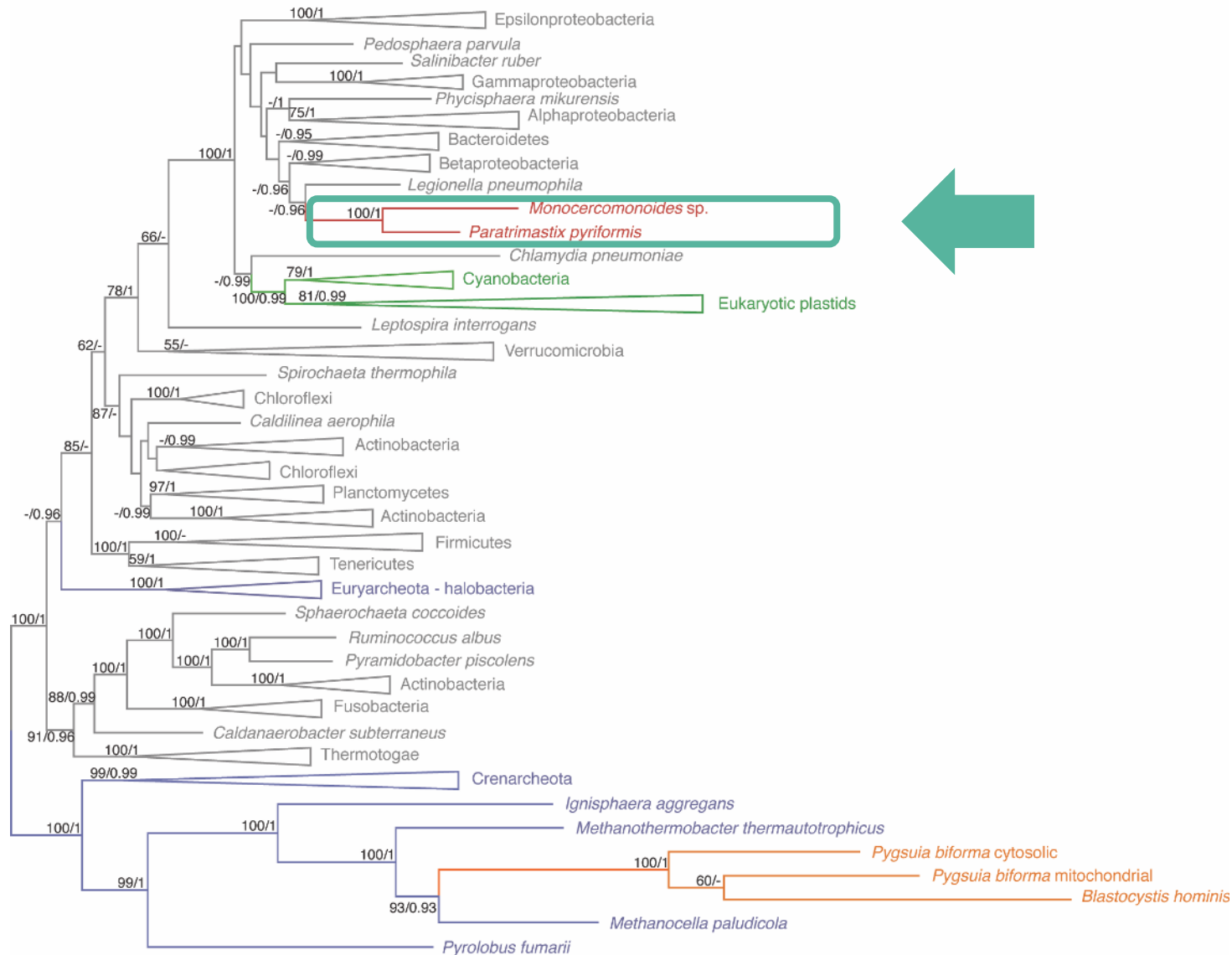
Fe-S cluster assembly systems in eukaryotes



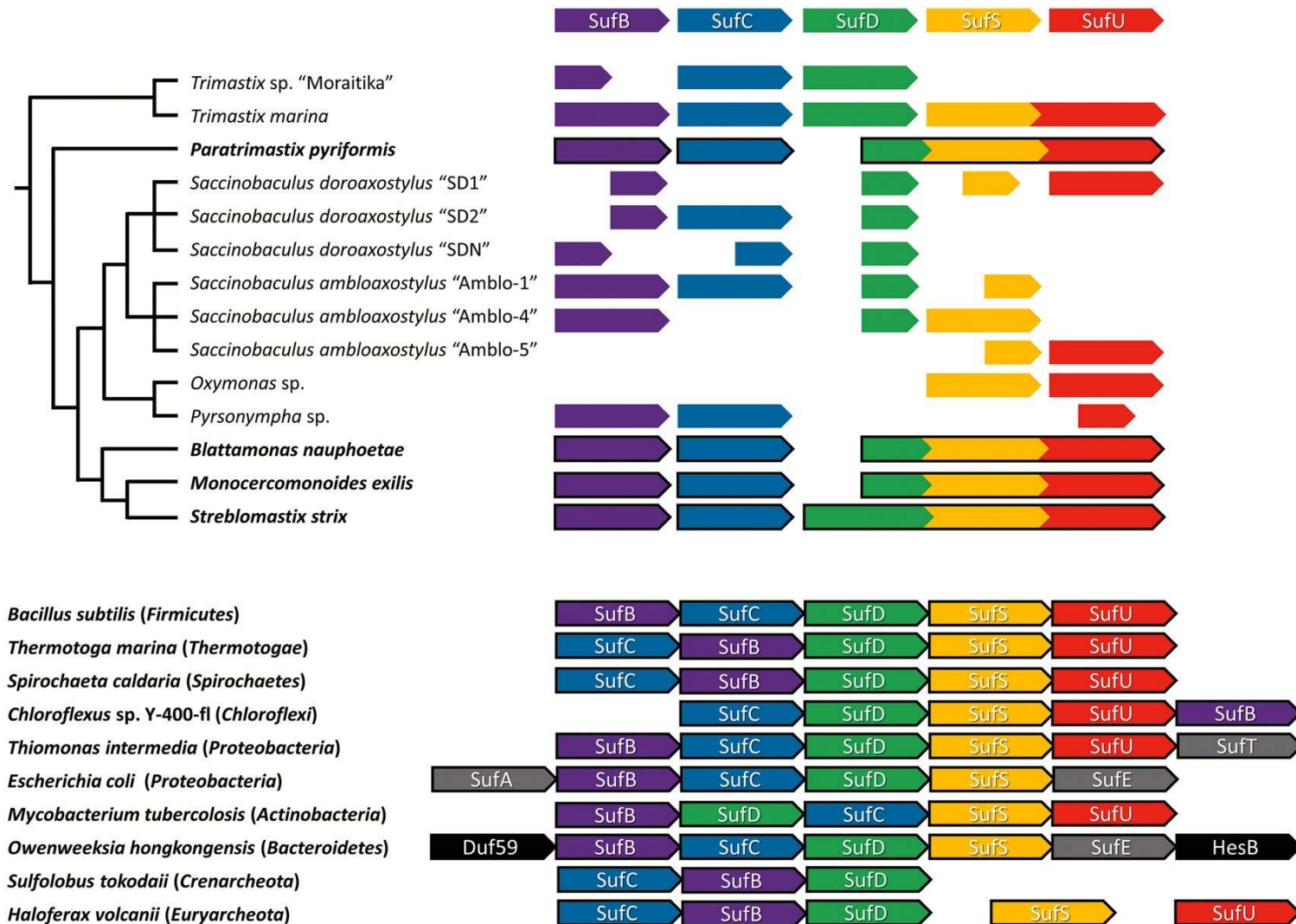
Fe-S cluster assembly systems in *Monocercomonoides*



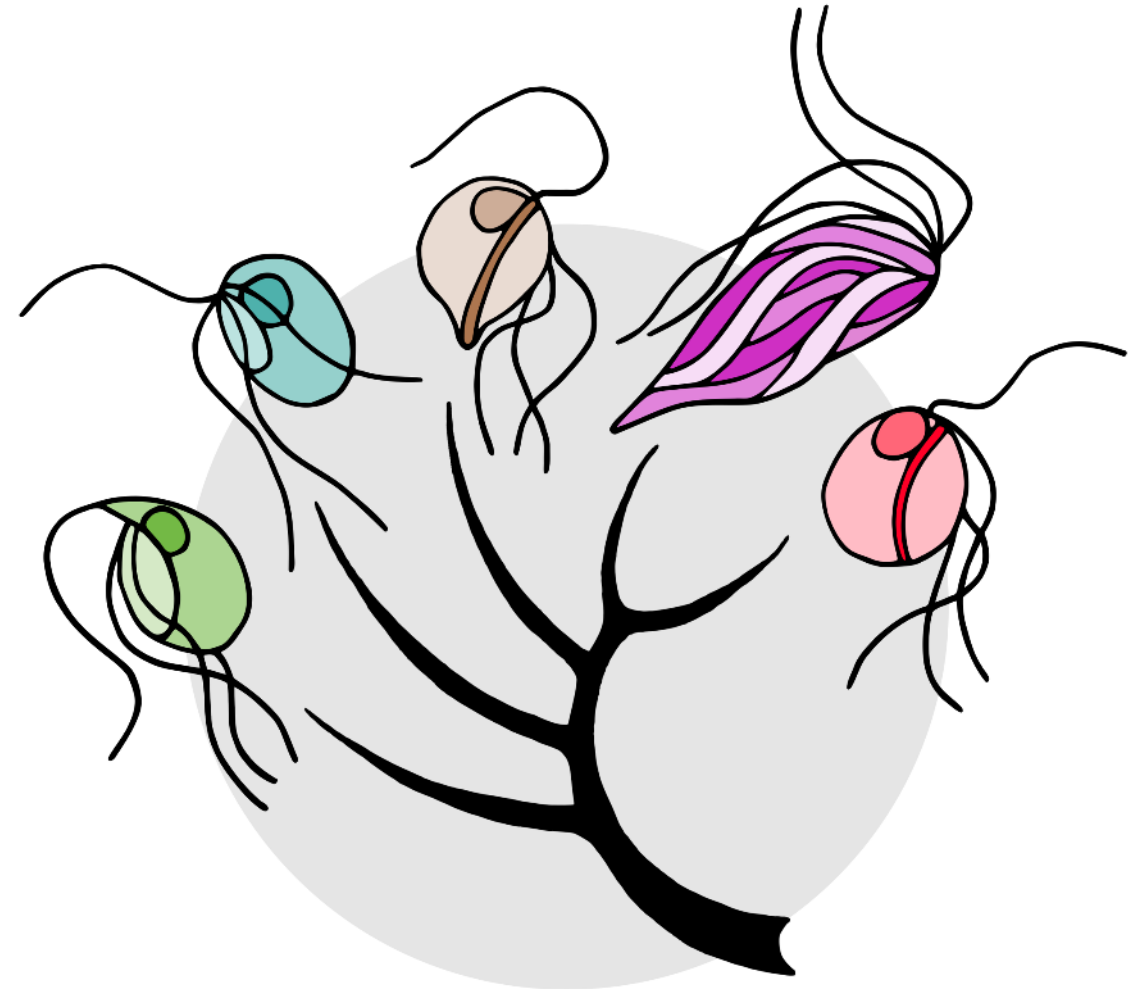
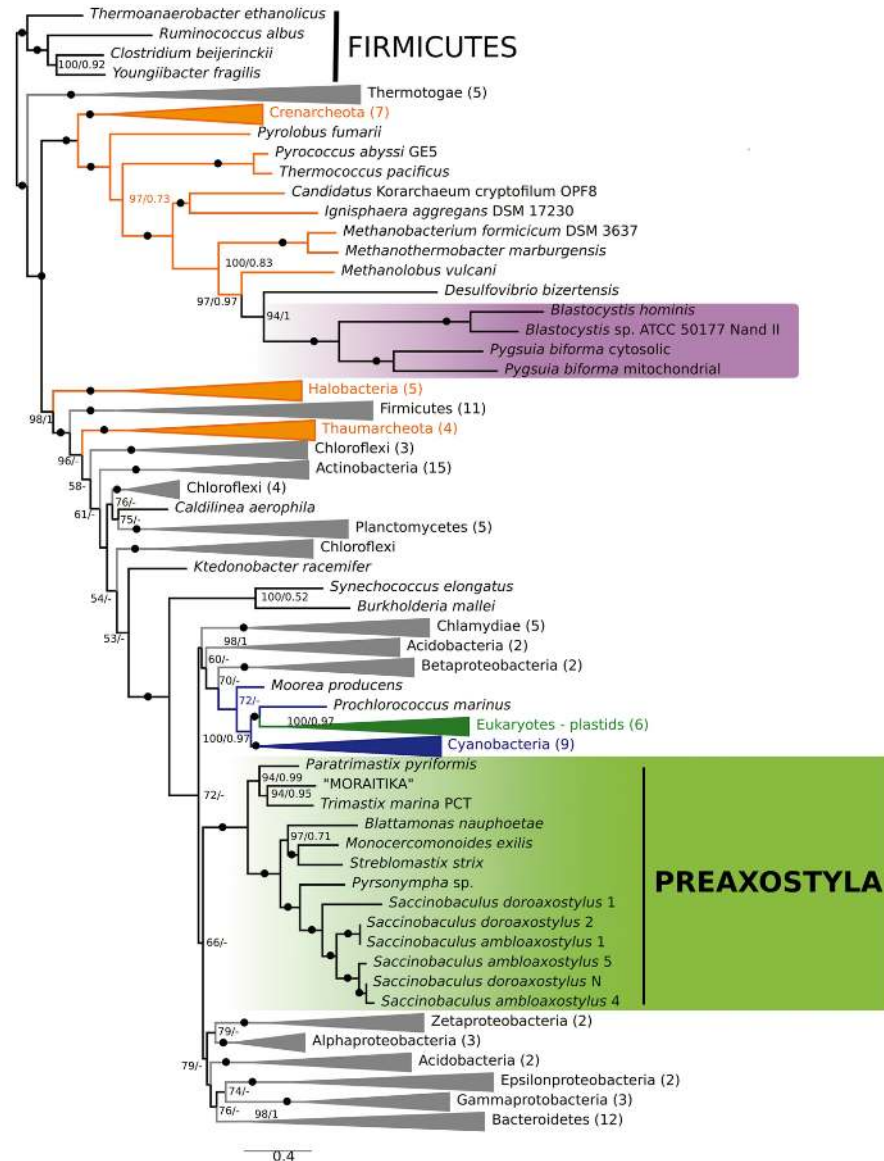
Lateral gene transfer (LGT) of SUF system



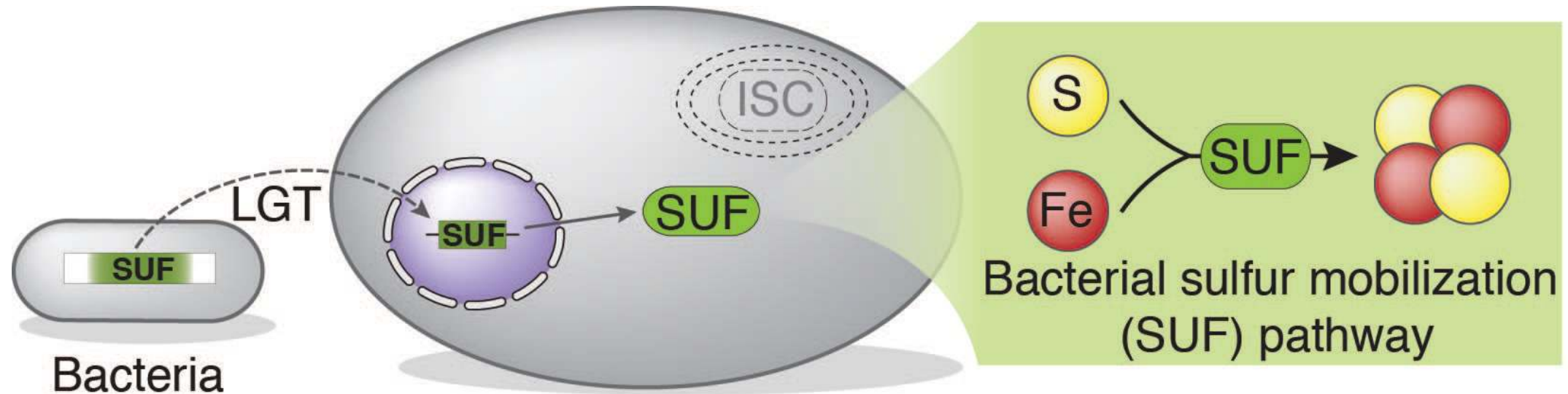
SUF proteins in Preaxostyla and representatives of bacteria



SUF system is widespread in Preaxostyla



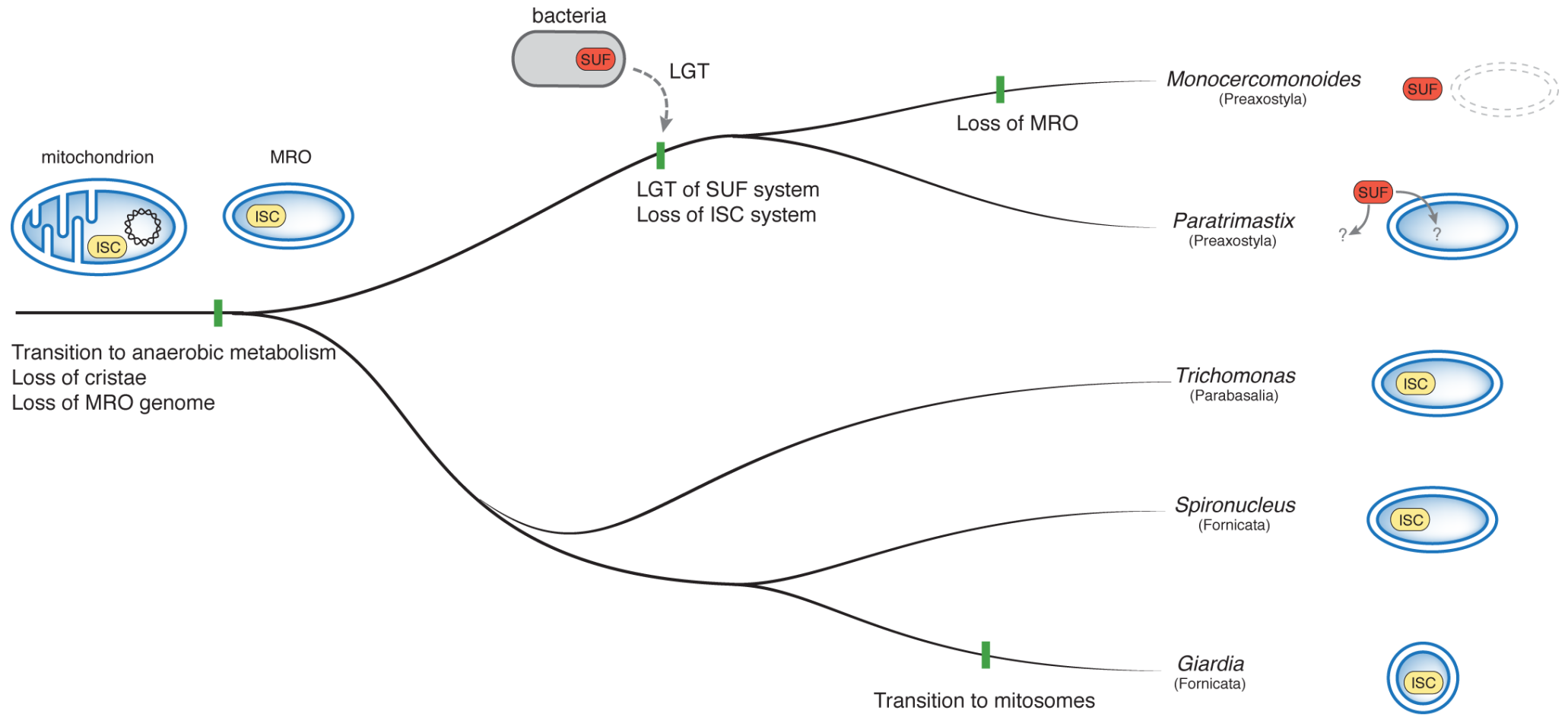
Monocercomonoides an **amitochondriate** eukaryote



Lateral gene transfer (LGT) of bacterial genes encoding sulfur mobilization pathway (SUF) and loss of reduced mitochondrion

Loss of a mitochondrial organelle

LGT of SUF system resulted in relocation of the pathway to the cytosol



Endosymbiosis can be undone!