EVE 161:
Microbial Phylogenomics

Class #6:
The Tree of Life, Example Paper

UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

Where we are going and where we have been

- Previous Class:
  - 5. Phylogeny
- Current Class:
  - 6. Phylogeny example
- Next Class:
  - 7. rRNA Sequencing from uncultured

Eisen Office Hours

- Today 10:30-11:30 Storer 5331
- Monday 10:00-11:00 GBSF 5311

Class Participation Sign Up

- Task: Summarize one figure from one of the papers for class, ~ 5 minutes.
  - There will be 2-3 presenters per class.
  - Each person needs to do a different figure so you need to contact the other person to coordinate.
  - Can meet with Eisen, Ettinger before hand to discuss.
- Sign up for date on Google Sheet goo.gl/bmxCDw - include your name & email address.

Paper Analysis Project

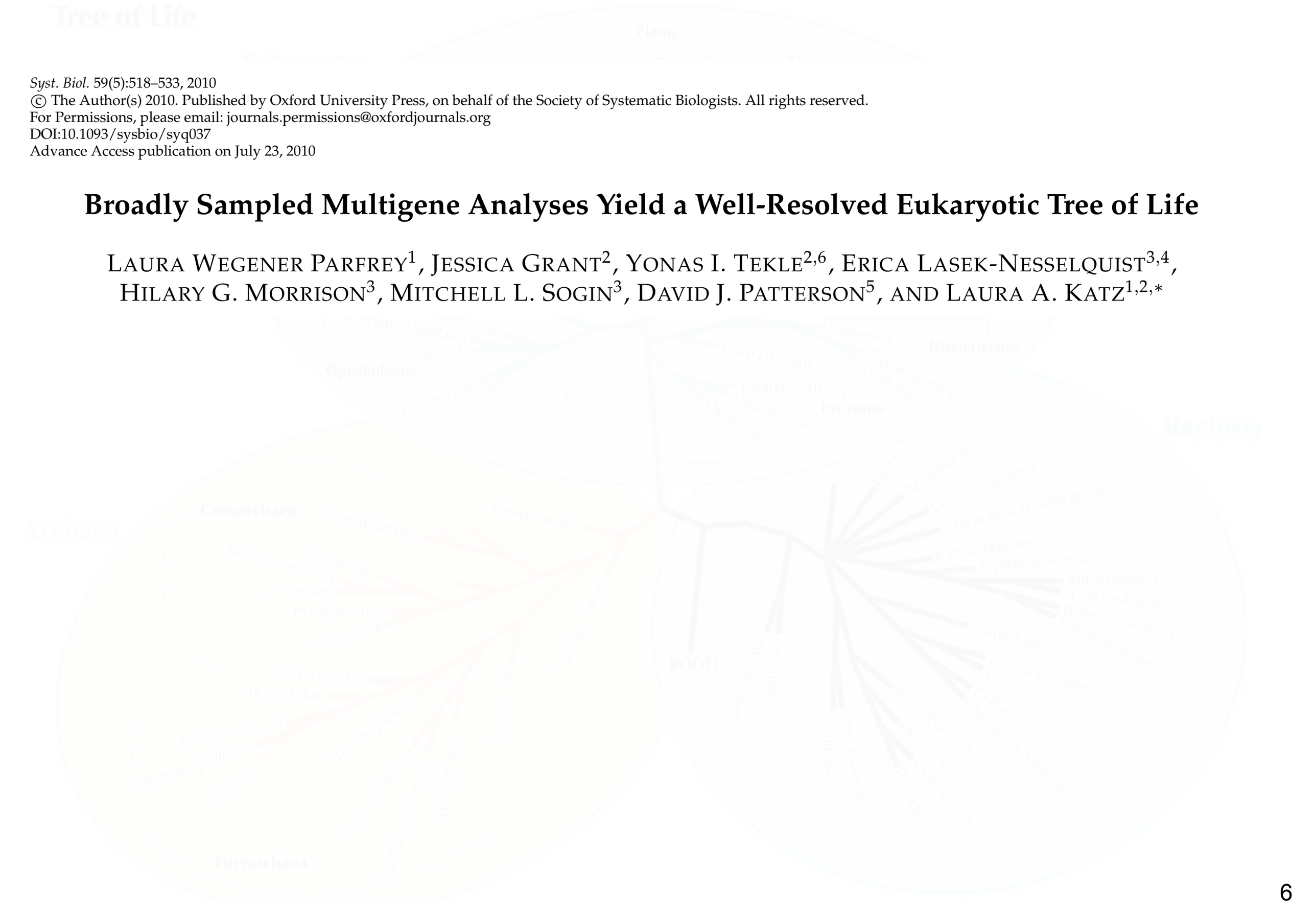
- Select 1-2 papers on one of the topics of the course (approval needed)
- Review the paper and write up a summary of your assessment of the paper (more detail on this later)
- Present a short summary of what you did to the class
- Ask and answer questions about your and other people's reviews in the discussion forum on Canvas

Syst. Biol. 59(5):518–533, 2010

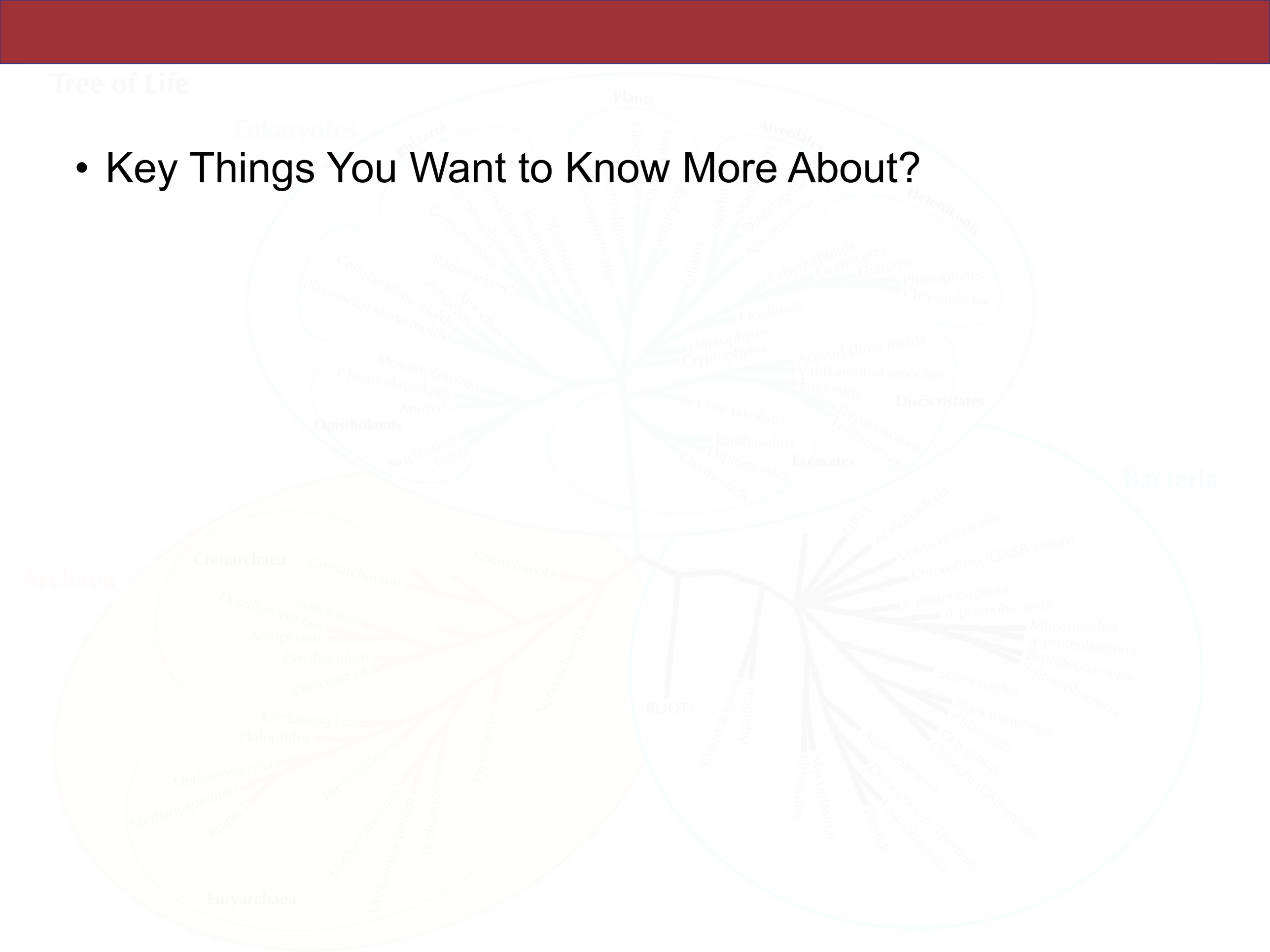
DOI:10.1093/sysbio/syq037
Advance Access publication on July 23, 2010

Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life

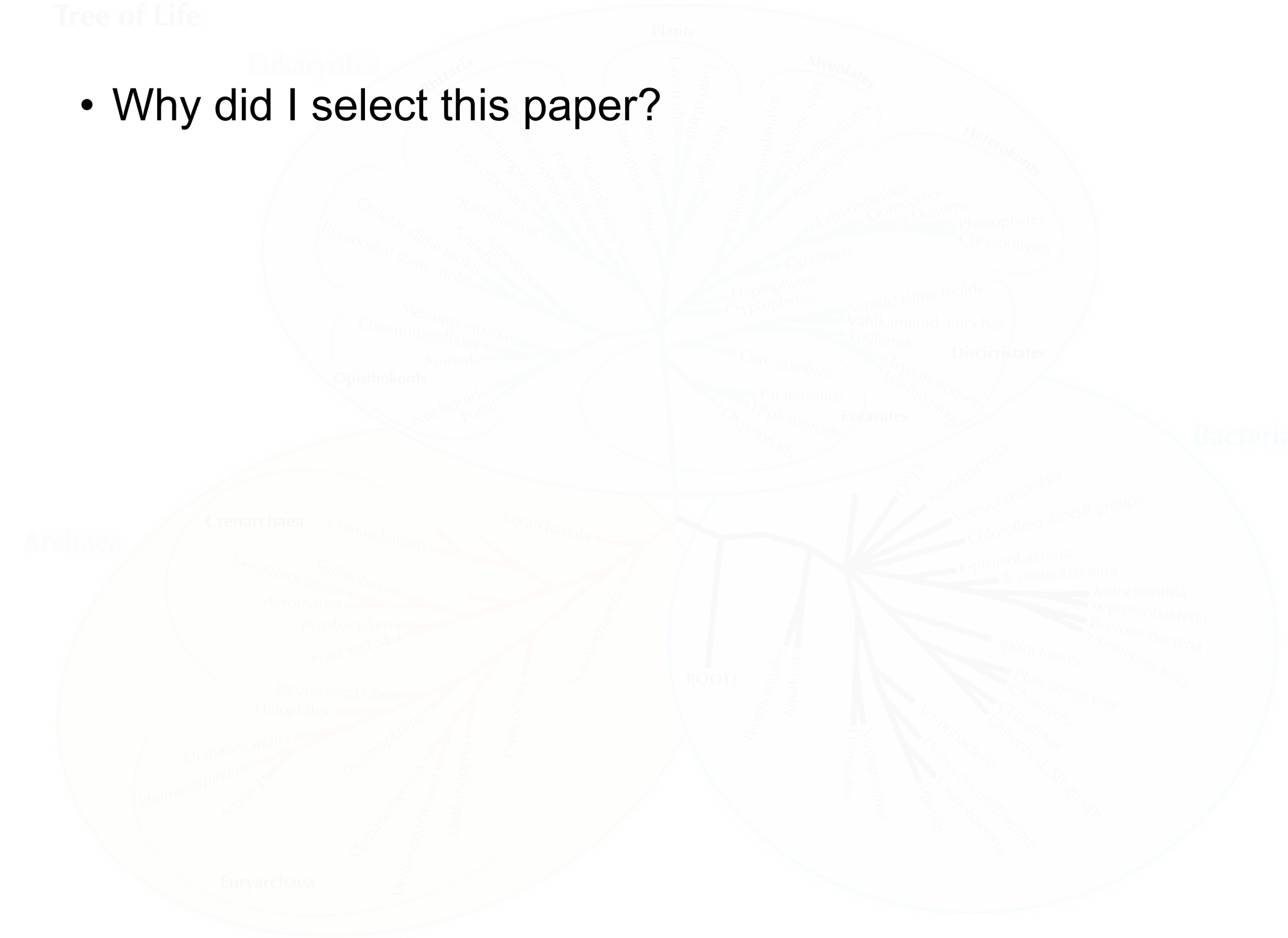
LAURA WEGENER PARFREY[1], JESSICA GRANT[2], YONAS I. TEKLE[2,6], ERICA LASEK-NESSELQUIST[3,4], HILARY G. MORRISON[3], MITCHELL L. SOGIN[3], DAVID J. PATTERSON[5], AND LAURA A. KATZ[1,2,*]



- Key Things You Want to Know More About?



- Why did I select this paper?



Syst. Biol. 59(5):518–533, 2010

DOI:10.1093/sysbio/syq037
Advance Access publication on July 23, 2010

Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life

LAURA WEGENER PARFREY[1], JESSICA GRANT[2], YONAS I. TEKLE[2,6], ERICA LASEK-NESSELQUIST[3,4], HILARY G. MORRISON[3], MITCHELL L. SOGIN[3], DAVID J. PATTERSON[5], AND LAURA A. KATZ[1,2,*]

What Does 1st Authorship Mean?

Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life

LAURA WEGENER PARFREY[1], JESSICA GRANT[2], YONAS I. TEKLE[2,6], ERICA LASEK-NESSELQUIST[3,4], HILARY G. MORRISON[3], MITCHELL L. SOGIN[3], DAVID J. PATTERSON[5], AND LAURA A. KATZ[1,2,*]

[1]Program in Organismic and Evolutionary Biology, University of Massachusetts, 611 North Pleasant Street, Amherst, MA 01003, USA; [2]Department of Biological Sciences, Smith College, 44 College Lane, Northampton, MA 01063, USA; [3]Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA; [4]Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman Street, Providence, RI 02912, USA; [5]Biodiversity Informatics Group, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA; [6]Present address: Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA;

[*]Correspondence to be sent to: Laura A. Katz, 44 College Lane, Northampton, MA 01003, USA; E-mail: lkatz@smith.edu.

Laura Wegener Parfrey and Jessica Grant have contributed equally to this work.

Received 30 September 2009; reviews returned 1 December 2009; accepted 25 May 2010
Associate Editor: Cécile Ané

Abstract

An accurate reconstruction of the eukaryotic tree of life is essential to identify the innovations underlying the diversity of microbial and macroscopic (e.g., plants and animals) eukaryotes. Previous work has divided eukaryotic diversity into a small number of high-level "supergroups," many of which receive strong support in phylogenomic analyses. However, the abundance of data in phylogenomic analyses can lead to highly supported but incorrect relationships due to systematic phylogenetic error. Furthermore, the paucity of major eukaryotic lineages (19 or fewer) included in these genomic studies may exaggerate systematic error and reduce power to evaluate hypotheses. Here, we use a taxon-rich strategy to assess eukaryotic relationships. We show that analyses emphasizing broad taxonomic sampling (up to 451 taxa representing 72 major lineages) combined with a moderate number of genes yield a well-resolved eukaryotic tree of life. The consistency across analyses with varying numbers of taxa (88–451) and levels of missing data (17–69%) supports the accuracy of the resulting topologies. The resulting stable topology emerges without the removal of rapidly evolving genes or taxa, a practice common to phylogenomic analyses. Several major groups are stable and strongly supported in these analyses (e.g., SAR, Rhizaria, Excavata), whereas the proposed supergroup "Chromalveolata" is rejected. Furthermore, ex- tensive instability among photosynthetic lineages suggests the presence of systematic biases including endosymbiotic gene transfer from symbiont (nucleus or plastid) to host. Our analyses demonstrate that stable topologies of ancient evolutionary relationships can be achieved with broad taxonomic sampling and a moderate number of genes. Finally, taxon-rich analyses such as presented here provide a method for testing the accuracy of relationships that receive high bootstrap support (BS) in phylogenomic analyses and enable placement of the multitude of lineages that lack genome scale data. [Excavata; microbial eukaryotes; Rhizaria; supergroups; systematic error; taxon sampling.]

Abstract

An accurate reconstruction of the eukaryotic tree of life is essential to identify the innovations underlying the diversity of microbial and macroscopic (e.g., plants and animals) eukaryotes. Previous work has divided eukaryotic diversity into a small number of high-level "supergroups," many of which receive strong support in phylogenomic analyses. However, the abundance of data in phylogenomic analyses can lead to highly supported but incorrect relationships due to systematic phylogenetic error. Furthermore, the paucity of major eukaryotic lineages (19 or fewer) included in these genomic studies may exaggerate systematic error and reduce power to evaluate hypotheses. Here, we use a taxon-rich strategy to assess eukaryotic relationships. We show that analyses emphasizing broad taxonomic sampling (up to 451 taxa representing 72 major lineages) combined with a moderate number of genes yield a well-resolved eukaryotic tree of life. The consistency across analyses with varying numbers of taxa (88–451) and levels of missing data (17–69%) supports the accuracy of the resulting topologies. The resulting stable topology emerges without the removal of rapidly evolving genes or taxa, a practice common to phylogenomic analyses. Several major groups are stable and strongly supported in these analyses (e.g., SAR, Rhizaria, Excavata), whereas the proposed supergroup "Chromalveolata" is rejected. Furthermore, ex- tensive instability among photosynthetic lineages suggests the presence of systematic biases including endosymbiotic gene transfer from symbiont (nucleus or plastid) to host. Our analyses demonstrate that stable topologies of ancient evolutionary relationships can be achieved with broad taxonomic sampling and a moderate number of genes. Finally, taxon-rich analyses such as presented here provide a method for testing the accuracy of relationships that receive high bootstrap support (BS) in phylogenomic analyses and enable placement of the multitude of lineages that lack genome scale data. [Excavata; microbial eukaryotes; Rhizaria; supergroups; systematic error; taxon sampling.]

Abstract

An accurate reconstruction of the eukaryotic tree of life is essential to identify the innovations underlying the diversity of microbial and macroscopic (e.g., plants and animals) eukaryotes. Previous work has divided eukaryotic diversity into a small number of high-level "supergroups," many of which receive strong support in phylogenomic analyses. However, the abundance of data in phylogenomic analyses can lead to highly supported but incorrect relationships due to systematic phylogenetic error. Furthermore, the paucity of major eukaryotic lineages (19 or fewer) included in these genomic studies may exaggerate systematic error and reduce power to evaluate hypotheses. Here, we use a taxon-rich strategy to assess eukaryotic relationships. We show that analyses emphasizing broad taxonomic sampling (up to 451 taxa representing 72 major lineages) combined with a moderate number of genes yield a well-resolved eukaryotic tree of life. The consistency across analyses with varying numbers of taxa (88–451) and levels of missing data (17–69%) supports the accuracy of the resulting topologies. The resulting stable topology emerges without the removal of rapidly evolving genes or taxa, a practice common to phylogenomic analyses. Several major groups are stable and strongly supported in these analyses (e.g., SAR, Rhizaria, Excavata), whereas the proposed supergroup "Chromalveolata" is rejected. Furthermore, ex- tensive instability among photosynthetic lineages suggests the presence of systematic biases including endosymbiotic gene transfer from symbiont (nucleus or plastid) to host. Our analyses demonstrate that stable topologies of ancient evolutionary relationships can be achieved with broad taxonomic sampling and a moderate number of genes. Finally, taxon-rich analyses such as presented here provide a method for testing the accuracy of relationships that receive high bootstrap support (BS) in phylogenomic analyses and enable placement of the multitude of lineages that lack genome scale data. [Excavata; microbial eukaryotes; Rhizaria; supergroups; systematic error; taxon sampling.]

Abstract

An accurate reconstruction of the eukaryotic tree of life is essential to identify the innovations underlying the diversity of microbial and macroscopic (e.g., plants and animals) eukaryotes. Previous work has divided eukaryotic diversity into a small number of high-level "supergroups," many of which receive strong support in phylogenomic analyses. However, the abundance of data in phylogenomic analyses can lead to highly supported but incorrect relationships due to systematic phylogenetic error. Furthermore, the paucity of major eukaryotic lineages (19 or fewer) included in these genomic studies may exaggerate systematic error and reduce power to evaluate hypotheses. Here, we use a taxon-rich strategy to assess eukaryotic relationships. We show that analyses emphasizing broad taxonomic sampling (up to 451 taxa representing 72 major lineages) combined with a moderate number of genes yield a well-resolved eukaryotic tree of life. The consistency across analyses with varying numbers of taxa (88–451) and levels of missing data (17–69%) supports the accuracy of the resulting topologies. The resulting stable topology emerges without the removal of rapidly evolving genes or taxa, a practice common to phylogenomic analyses. Several major groups are stable and strongly supported in these analyses (e.g., SAR, Rhizaria, Excavata), whereas the proposed supergroup "Chromalveolata" is rejected. Furthermore, ex- tensive instability among photosynthetic lineages suggests the presence of systematic biases including endosymbiotic gene transfer from symbiont (nucleus or plastid) to host. Our analyses demonstrate that stable topologies of ancient evolutionary relationships can be achieved with broad taxonomic sampling and a moderate number of genes. Finally, taxon-rich analyses such as presented here provide a method for testing the accuracy of relationships that receive high bootstrap support (BS) in phylogenomic analyses and enable placement of the multitude of lineages that lack genome scale data. [Excavata; microbial eukaryotes; Rhizaria; supergroups; systematic error; taxon sampling.]

Introduction

Intro Paragraph 1

Perspectives on the structure of the eukaryotic tree of life have shifted in the past decade as molecular analyses provide hypotheses for relationships among the approximately 75 robust lineages of eukaryotes. These lineages are de ned by ultrastructural identities (Patterson 1999) —patterns of cellular and subcellular organization revealed by electron microscopy—and are strongly supported in molecular analyses (Parfrey et al. 2006; Yoon et al. 2008). Most of these lineages now fall within a small number of higher level clades, the supergroups of eukaryotes (Simpson and Roger 2004; Adl et al. 2005; Keeling et al. 2005). Several of these clades— Opisthokonta, Rhizaria, and Amoebozoa— are increasingly well supported by phylogenomic (Rodr´ıguez-Ezpeleta et al. 2007a; Burki et al. 2008; Hampl et al. 2009) and phylogenetic (Parfrey et al. 2006; Pawlowski and Burki 2009), analyses, whereas support for "Archaeplastida" predominantly comes from some phylogenomic studies (Rodr´ıguez-Ezpeleta et al. 2005; Burki et al. 2007) or analyses of plastid genes (Yoon et al. 2002; Parfrey et al. 2006). In con- trast, support for "Chromalveolata" and Excavata is mixed, often dependent on the selection of taxa in- cluded in analyses (Rodr´ıguez-Ezpeleta et al. 2005; Parfrey et al. 2006; Rodr´ıguez-Ezpeleta et al. 2007a; Burki et al. 2008; Hampl et al. 2009). We use quotation marks throughout to note groups where uncertainties remain. Moreover, it is dif cult to evaluate the overall stability of major clades of eukaryotes because phyloge- nomic analyses have 19 or fewer of the major lineages and hence do not suf ciently sample eukaryotic diver- sity(Rodr´ıguez-Ezpeletaetal.2007b;Burkietal.2008; Hampl et al. 2009), whereas taxon-rich analyses with 4 or fewer genes yield topologies with poor support at deep nodes (Cavalier-Smith 2004; Parfrey et al. 2006; Yoon et al. 2008).

Intro Paragraph 1

Perspectives on the structure of the eukaryotic tree of life have shifted in the past decade as molecular analyses provide hypotheses for relationships among the approximately 75 robust lineages of eukaryotes. These lineages are de ned by ultrastructural identities (Patterson 1999) —patterns of cellular and subcellular organization revealed by electron microscopy—and are strongly supported in molecular analyses (Parfrey et al. 2006; Yoon et al. 2008). Most of these lineages now fall within a small number of higher level clades, the supergroups of eukaryotes (Simpson and Roger 2004; Adl et al. 2005; Keeling et al. 2005). Several of these clades— Opisthokonta, Rhizaria, and Amoebozoa— are increasingly well supported by phylogenomic (Rodr´ıguez-Ezpeleta et al. 2007a; Burki et al. 2008; Hampl et al. 2009) and phylogenetic (Parfrey et al. 2006; Pawlowski and Burki 2009), analyses, whereas support for "Archaeplastida" predominantly comes from some phylogenomic studies (Rodr´ıguez-Ezpeleta et al. 2005; Burki et al. 2007) or analyses of plastid genes (Yoon et al. 2002; Parfrey et al. 2006). In con- trast, support for "Chromalveolata" and Excavata is mixed, often dependent on the selection of taxa in- cluded in analyses (Rodr´ıguez-Ezpeleta et al. 2005; Parfrey et al. 2006; Rodr´ıguez-Ezpeleta et al. 2007a; Burki et al. 2008; Hampl et al. 2009). We use quotation marks throughout to note groups where uncertainties remain. Moreover, it is dif cult to evaluate the overall stability of major clades of eukaryotes because phyloge- nomic analyses have 19 or fewer of the major lineages and hence do not suf ciently sample eukaryotic diver- sity(Rodr´ıguez-Ezpeletaetal.2007b;Burkietal.2008; Hampl et al. 2009), whereas taxon-rich analyses with 4 or fewer genes yield topologies with poor support at deep nodes (Cavalier-Smith 2004; Parfrey et al. 2006; Yoon et al. 2008).

Intro Paragraph 2

Estimating the relationships of the major lineages of eukaryotes is difficult because of both the ancient age of eukaryotes (1.2–1.8 billion years; Knoll et al. 2006) and complex gene histories that include heterogeneous rates of molecular evolution and paralogy (Maddison 1997; Gribaldo and Philippe 2002; Tekle et al. 2009). A further issue obscuring eukaryotic relationships is the chimeric nature of the eukaryotic genome—not all genes are vertically inherited due to lateral gene transfer (LGT) and endosymbiotic gene transfer (EGT)—that can also mislead efforts to reconstruct phylogenetic relationships (Andersson 2005; Rannala and Yang 2008; Tekle et al. 2009). This is especially true among photosynthetic lineages that comprise "Chromalveolata" and "Archaeplastida" where a large portion of the host genome (approximately 8–18%) is

Intro Paragraph 2

Estimating the relationships of the major lineages of eukaryotes is difficult because of both the ancient age of eukaryotes (1.2–1.8 billion years; Knoll et al. 2006) and complex gene histories that include heterogeneous rates of molecular evolution and paralogy (Maddison 1997; Gribaldo and Philippe 2002; Tekle et al. 2009). A further issue obscuring eukaryotic relationships is the chimeric nature of the eukaryotic genome—not all genes are vertically inherited due to lateral gene transfer (LGT) and endosymbiotic gene transfer (EGT)—that can also mislead efforts to reconstruct phylogenetic relationships (Andersson 2005; Rannala and Yang 2008; Tekle et al. 2009). This is especially true among photosynthetic lineages that comprise "Chromalveolata" and "Archaeplastida" where a large portion of the host genome (approximately 8–18%) is

Intro Paragraph 2

Estimating the relationships of the major lineages of eukaryotes is difficult because of both the ancient age of eukaryotes (1.2–1.8 billion years; Knoll et al. 2006) and complex gene histories that include heterogeneous rates of molecular evolution and paralogy (Maddison 1997; Gribaldo and Philippe 2002; Tekle et al. 2009). A further issue obscuring eukaryotic relationships is the chimeric nature of the eukaryotic genome—not all genes are vertically inherited due to lateral gene transfer (LGT) and endosymbiotic gene transfer (EGT)—that can also mislead efforts to reconstruct phylogenetic relationships (Andersson 2005; Rannala and Yang 2008; Tekle et al. 2009). This is especially true among photosynthetic lineages that comprise "Chromalveolata" and "Archaeplastida" where a large portion of the host genome (approximately 8–18%) is

LGT and Endosymbiosis

Intro Paragraph 3

There is a long-standing debate among systematists as to the relative benefits of increasing gene or taxon sampling (Hillis et al. 2003; Cummings and Meyer 2005; Rokas and Carroll 2005). Both approaches improve phylogenetic reconstruction by alleviating either stochastic or systematic phylogenetic error (e.g., Rokas and Carroll 2005; Hedtke et al. 2006). Stochastic error results from too little signal in the data (e.g., single to few gene trees) to estimate relationships and results in poorly resolved trees with low support, especially at deep levels (Swofford et al. 1996; Rokas and Carroll 2005). The problems of stochastic error are amplified for deep relationships, such as relationships among major clades of eukaryotes (Roger and Hug 2006). Many researchers opt to increase the number of genes, exemplified by phylogenomic studies, which alleviates stochastic error and yields well-resolved trees that are highly supported (Rokas and Carroll 2005; Burki et al. 2007; Hampl et al. 2009). However, analyses of many genes are still vulnerable to systematic error and often include very few lineages.

Intro Paragraph 3

There is a long-standing debate among systematists as to the relative benefits of increasing gene or taxon sampling (Hillis et al. 2003; Cummings and Meyer 2005; Rokas and Carroll 2005). Both approaches improve phylogenetic reconstruction by alleviating either stochastic or systematic phylogenetic error (e.g., Rokas and Carroll 2005; Hedtke et al. 2006). Stochastic error results from too little signal in the data (e.g., single to few gene trees) to estimate relationships and results in poorly resolved trees with low support, especially at deep levels (Swofford et al. 1996; Rokas and Carroll 2005). The problems of stochastic error are amplified for deep relationships, such as relationships among major clades of eukaryotes (Roger and Hug 2006). Many researchers opt to increase the number of genes, exemplified by phylogenomic studies, which alleviates stochastic error and yields well-resolved trees that are highly supported (Rokas and Carroll 2005; Burki et al. 2007; Hampl et al.

Intro Paragraph 4

Systematic error results from biases in the data that mislead phylogenetic reconstruction, yielding incorrect sister group relationships that do not reflect historical relationships; the most well known of these is long-branch attraction (Felsenstein 1978). Incongruence can also arise from conflicts between gene trees and species trees resulting from population genetic processes or the chimeric nature of eukaryotic genomes (Maddison 1997; Rannala and Yang 2008). Systematic errors can be detected and eliminated by several methods that are often combined, including using more realistic models of sequence evolution (e.g., Rodríguez-Ezpeleta et al. 2007b), removing rapidly evolving genes and/or taxa that cause errors (Brinkmann et al. 2005), and by increasing taxonomic sampling (Zwickl and Hillis 2002; Hedtke et al. 2006). Increased taxon sampling has been shown to improve phylogenetic accuracy even when the additional taxa contain large amounts of missing data (Philippe et al. 2004; Wiens 2005; Wiens and Moen 2008). In contrast, the abundance of data in phylogenomic studies can yield highly supported, but incorrect relationships caused by these systematic biases (Philippe et al. 2004; Hedtke et al. 2006; Jeffroy et al. 2006; Rokas and Chatzimanolis 2008). Taxon-rich analyses provide a method for testing the accuracy of relationships that receive high BS support in phylogenomic analyses (Zwickl and Hillis 2002; Heath et al. 2008).

Systematic error results from biases in the data that mislead phylogenetic reconstruction, yielding incorrect sister group relationships that do not reflect historical relationships; the most well known of these is long-branch attraction (Felsenstein 1978). Incongruence can also arise from conflicts between gene trees and species trees resulting from population genetic processes or the chimeric nature of eukaryotic genomes (Maddison 1997; Rannala and Yang 2008). Systematic errors can be detected and eliminated by several methods that are often combined, including using more realistic models of sequence evolution (e.g., Rodríguez-Ezpeleta et al. 2007b), removing rapidly evolving genes and/or taxa that cause errors (Brinkmann et al. 2005), and by increasing taxonomic sampling (Zwickl and Hillis 2002; Hedtke et al.

- LBA


Tree of Life
Eukaryotes
Plants
Green algae
Land plants
Red algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Glaucophyte algae
Alveolates
Ciliates
Syndinians
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Discicristates
Core Jakobids
Trypanosomes
Leishmanias
Parabasalids
Diplomonads
Oxymonads
Excavates
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Nucleariids
Fungi
Bacteria
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Gram+
Mycoplasmas
Firmicutes
OP11
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Actinobacteria
Deinococcus/Thermus
Cyanobacteria
Plastids

Introduction

- Questions about Introduction?

Methods

- What data did they use?
- What genes?
- What taxa?
- How did they get the data?

Taxa and Genes

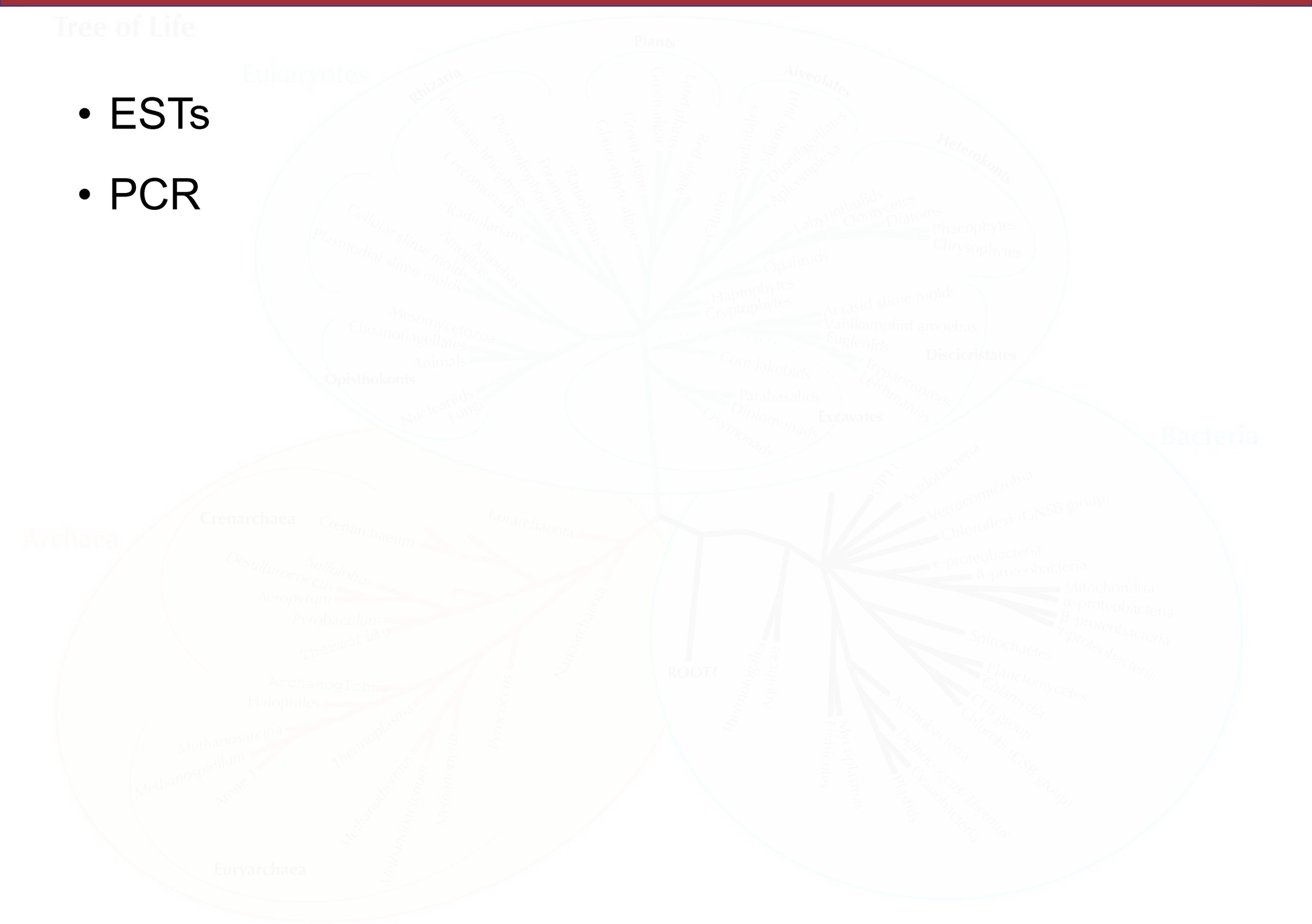
Data set Assembly

Taxa and genes were selected to maximize taxonomic diversity and evenness given the availability of molecular data. This strategy was used to improve phylogenetic accuracy by breaking up long branches with dense sampling across the eukaryotic tree (Hillis 1998). The classifications systems of Patterson (1999) and Adl et al. (2005) were used as guides as we aimed to sample eukaryotic diversity by including representatives of as many lineages defined by ultrastructural identities as possible (Table S2). These lineages have generally proven to be robust as they are well supported in molecular analyses (e.g., Adl et al. 2005; Parfrey et al. 2006; Yoon et al. 2008), including the current study, and they represent monophyletic groups that serve as a proxy for taxonomic diversity. Our data set has representatives from 72 lineages, including 53 of the 71 lineages plus 7 of 200 unplaced genera as defined in Patterson (1999). Additionally, we include 3 unplaced lineages isolated more recently, *Malawimonas jakobiformis* (O'Kelly and Nerad 1999), *Breviata anathema* (Walker et al. 2006),

Taxa and Genes

To maximize gene sampling for diverse taxa, we include markers historically targeted by polymerase chain reaction–based analyses (e.g., SSU-rDNA, *actin*, elongation factor 1α; Table S3) plus commonly sequenced ESTs (e.g., ribosomal proteins, *14-3-3*; Table S3). The comprehensively sampled SSU-rDNA and the historical markers facilitate inclusion of many additional taxa for which only these genes have been characterized (Table S4). The minimum sequence data required for inclusion were nearly full-length SSU-rDNA, which provided the core of information necessary for phylogenetic placement with large amounts of missing data (Wiens and Moen 2008).

- ESTs
- PCR


Tree of Life
Eukaryotes
Plants
Land plants
Green algae
Red algae
Glaucophyte algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
"Radiolarians"
Ciliates
Stramenopiles
Alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Trypanosomes
Leishmanias
Discicristates
Core Jakobids
Parabasalids
Diplomonads
Oxymonads
Excavates
Mesomycetozoa
Choanoflagellates
Animals
Fungi
Nucleariids
Opisthokonts
Bacteria
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Gram+
Mycoplasmas
Thermus/Deinococcus
Cyanobacteria
Plastids
Actinobacteria
CFB group
Chlamydia
Planctomycetes
Spirochaetes
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
δ-proteobacteria
ε-proteobacteria
Chloroflexi (GNSB group)
Verrucomicrobia
Acidobacteria
OP11

Alignments

SSU-rDNA sequences were hand curated for target taxa by removing introns, unalignable regions, non-nuclear rDNAs, and misannotated sequences. This alignment was crucial to overall accuracy because nearly half of the target taxa are represented only by SSU-rDNA, thus several alignment and masking methods were assessed to ensure the robustness of the SSU-rDNA alignment. SSU-rDNA sequences were aligned by HMMER (Eddy 2001), version 2.1.4 with default settings, taking secondary structure into account. HMMER used a set of previously aligned sequences to model the secondary structure of a sequence. The training alignment for building the model, consisting of all available SSU-rDNA eukaryote sequences (as of December 2008) aligned according to their secondary structure, was downloaded from the European Ribosomal Database (Wuyts et al. 2002). An additional SSU-rDNA alignment was constructed in MAFFT 6 implemented in

Alignments

downloaded from the European Ribosomal Database (Wuyts et al. 2002). An additional SSU-rDNA alignment was constructed in MAFFT 6 implemented in SeaView (Galtier et al. 1996) with the E-INS-i algorithm (Katoh and Toh 2008). Both alignments were further edited manually in MacClade v4.08 (Maddison D.R. and Maddison W.P. 2005). To assess the effect of rate heterogeneity on the SSU-rDNA topologies, we partitioned the data matrices into 8 rate classes using the general time-reversible (GTR) model with invariable sites and rate variation among sites following a discrete gamma distribution, as implemented in HyPhy version .99b package (Kosakovsky Pond et al. 2005). We then ran analyses without the fastest and two fastest rate classes, resulting in 1197 and 1019 characters, respectively. However, the reduced data sets resulted in less resolution in the backbone without improving apparent the long-branch attraction. Thus, we used the alignment generated in MAFFT and masked with GBlocks (Talavera and Castresana 2007) and by eye in MacClade, resulting in 867 unambiguously aligned characters.

Protein Database Searches

Assembly of the protein data set relied on a custom-built pipeline and database that combined Perl and Python scripts to identify homologs from diverse eukaryotes. Our goal in developing this pipeline was to ensure that we captured the broadest possible set of sequences given the tremendous heterogeneity among microbial eukaryotes. All available protein and EST

Protein Database Searches

data from our target taxa (Table S4) were downloaded from GenBank in January 2009 and ESTs were analyzed in all 6 translated frames to identify correct sequences for our alignment. A fasta file of 6 sequences representing the six "supergroups" was created for each target gene and used to query our database of target taxa by BLASTp. Results were limited by length, e-value, and identity, and all sequences with greater than 1% divergence within each taxon were retained for assessment of paralogy. The resulting sequences were aligned with ClustalW (Thompson et al. 1994) and the resulting single gene alignments were assessed by eye to remove nonhomologous sequences.

Alignments and Paralogs

The inferred amino acid sequences for each of the protein genes from our data pipeline were combined with the new sequences generated for this study and again aligned in Clustal W (Thompson et al. 1994). The alignment was adjusted by eye in MacClade (Maddison D.R. and Maddison W.P. 2005). As these alignments included all paralogs extracted from the pipeline, individual gene trees were examined to choose appropriate orthologs. For example, in cases where paralogs formed a monophyletic group, the shortest branch sequence was retained. When paralogs fell into multiple locations on the tree, we aimed to maintain orthologous groups that included the greatest taxonomic representation. The individual gene alignments were then concatenated to build a 16 gene, 451-taxon matrix with 6578 unambiguously aligned characters, including SSU-rDNA. All other data sets were constructed by removing taxa and/or genes from this matrix. All data matrices are available at TreeBASE (submission ID S10562).

Submatrices

Creation of Subdata Matrices

We created an array of data matrices by subsampling our full data matrix of 16 genes (15 protein-coding genes plus SSU-rDNA) and 451 taxa (denoted all:16) in order to assess the impact of taxon sampling, missing data, and gene sampling. First, seven data sets were created to assess the impact of missing data and taxon sampling (summarized in Table 1). The least inclusive of these contained 16 genes and all 88 taxa that had at least 10 of the 16 genes (10:16), which resulted in 17% missing data. Similarly, the 6:16 and 4:16 matrices include all taxa with at least 6 and 4 of the targeted 16 genes, respectively. SSU-rDNA is ubiquitously sampled in our data set and many phylogenetic hypotheses are based on SSU-rDNA genealogies. To address the concern that SSU-rDNA was driving our results, we deleted it from each of the 16 gene data sets resulting in 9:15, 5:15, 3:15, and all:15 matrices.

EGT

genes and taxa.

Photosynthetic lineages have chimeric genomes that are composed of genes originating both from the host eukaryote, the endosymbiotic plastid (through EGT), and, in cases of secondary or greater endosymbiosis, from the symbiont nucleus. If genes of multiple origins were retained in our concatenated data set, the resulting conflicting signal between host, symbiont, and plastid could mislead phylogenetic reconstruction. This chimerism may contribute to the instability observed for photosynthetic lineages without clear sister groups (red algae, green algae, glaucocystophytes, cryptomonads, and haptophytes). Thus, we used 2 methods to detect discordance among loci that could indicate EGT. First, the 16 genes from representatives of each of these photosynthetic lineages were analyzed by top BLASTp hit. We

Phylogeny

Phylogenetic Analyses

Genealogies for this study were constructed almost exclusively in RaxML. The MPI version of RaxML 7.0.4 with rapid bootstrapping was used (Stamatakis et al. 2008). The SSU-rDNA partition was analyzed with GTR+gamma as this was the best fitting model available in RAxML, according to MrModelTest (Nylander 2004). ProtTest (Abascal et al. 2005) was used to select the appropriate model of sequence evolution for the amino acid data using the 9:15 data set. The WAG amino acid replacement matrix was found to be the best-fitting model for the concatenated data, but the rtREV amino acid replacement matrix was the best for some of the individual partitions and both WAG and rtREV were among the top 3 models for all but 1 gene (and with similar likelihood scores). We ran our data under both WAG and rtREV models and found consistent results, indicating that our interpretations are robust to at least

Bootstrapping

this level of model choice. The results presented are from the WAG analyses and the rtREV analyses differed only in level of BS for key nodes (usually ±5 points). In initial analyses, the appropriate number of independent bootstrap replicates was determined for each data set using bootstopping criteria in RAxML 7.0.4 as implemented on Cyberinfrastructure for Phylogenetic Research (CIPRES) portal 2 (Miller et al. 2009). All analyses stopped after 200 or fewer replicates, except all:16, which stopped after 400 replicates. In later analyses, using the MPI version of RAxML, which does not implement a bootstopping criterion, 200 rapid bootstrap replicates followed by a full maximum-likelihood search was used for all analyses except all:16, for which 600 bootstrap replicates were run. Because of the computational cost of the all:16 analysis, this was run as 6 separate analyses: 100 bootstraps followed by a full maximum likelihood search and 5 other runs of 100 bootstraps each. These data were combined in RAxML to complete the analysis. We found no significant difference in comparisons between fast and slow RAxML bootstrap methods (Fig. S1i), which we tested because the fast bootstrapping method in RAxML can produce misleading results particularly for long-branch taxa (Leigh 2008). The results of rapid bootstrapping are shown.

Other Methods

To investigate the stability of our tree topology under different analytic methods, select data sets were analyzed with Bayesian approaches and Parsimony (Fig. S1s–v). Parsimony analysis of 10:16, implemented in Paup* (Swofford 2002), yielded a less resolved version of the RAxML topology (i.e. Excavata as a polytomy) that is generally concordant with the more resolved tree obtained by maximum-likelihood methods. The

Methods

- Questions about Methods?

Results and Discussion

Results and Discussion

Robust Topology of the Eukaryotic Tree of Life

Many major clades were consistently recovered across our analyses (Fig. 1 and Table 1). These stable groups receive moderate to strong support in analyses with limited missing data (Fig. 2) and less support as missing data increases. The Opisthokonta, which includes animals and fungi, and the heterogeneous clade Rhizaria

Fig 1: 451 Taxa and some of the 16 genes

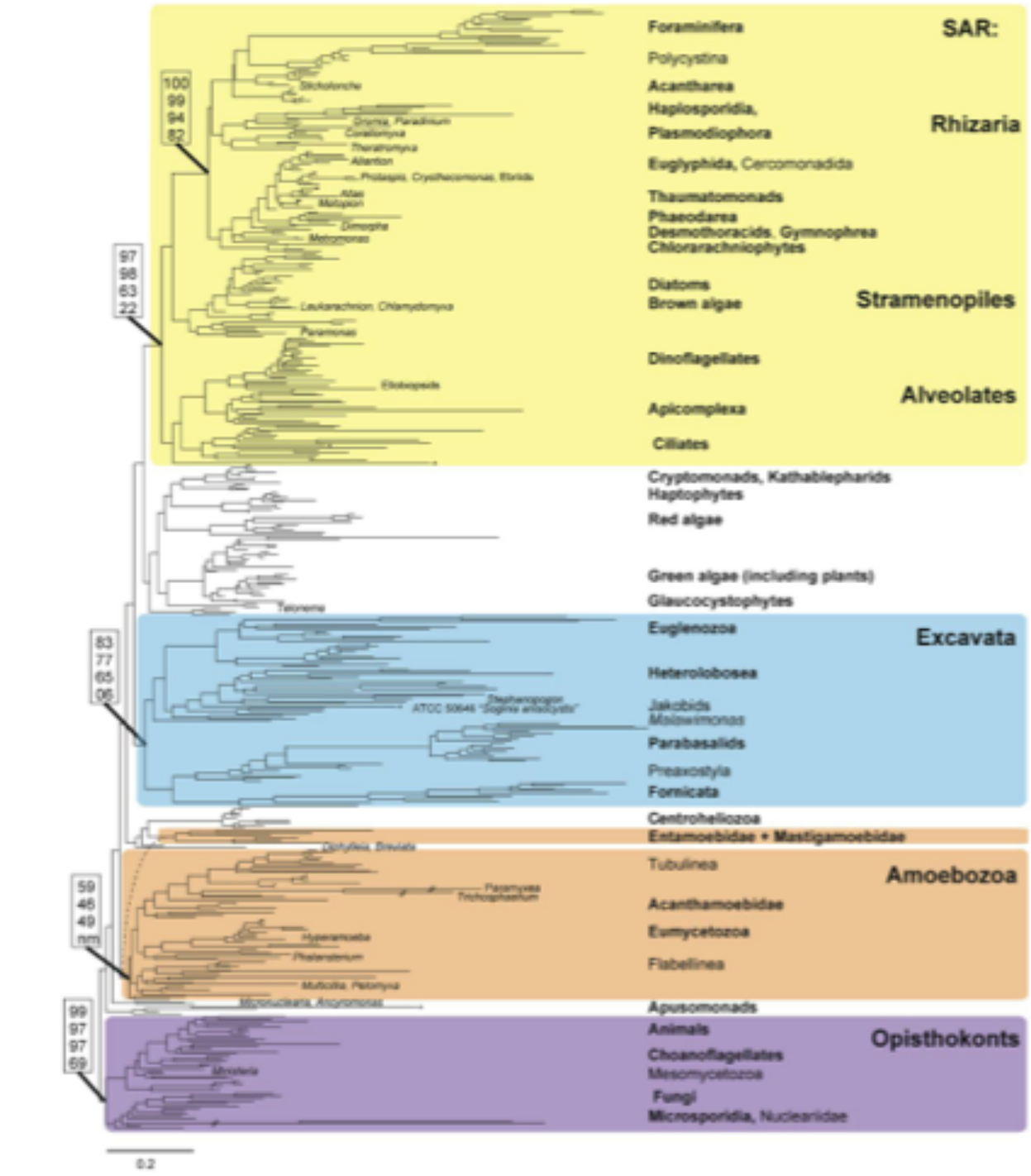
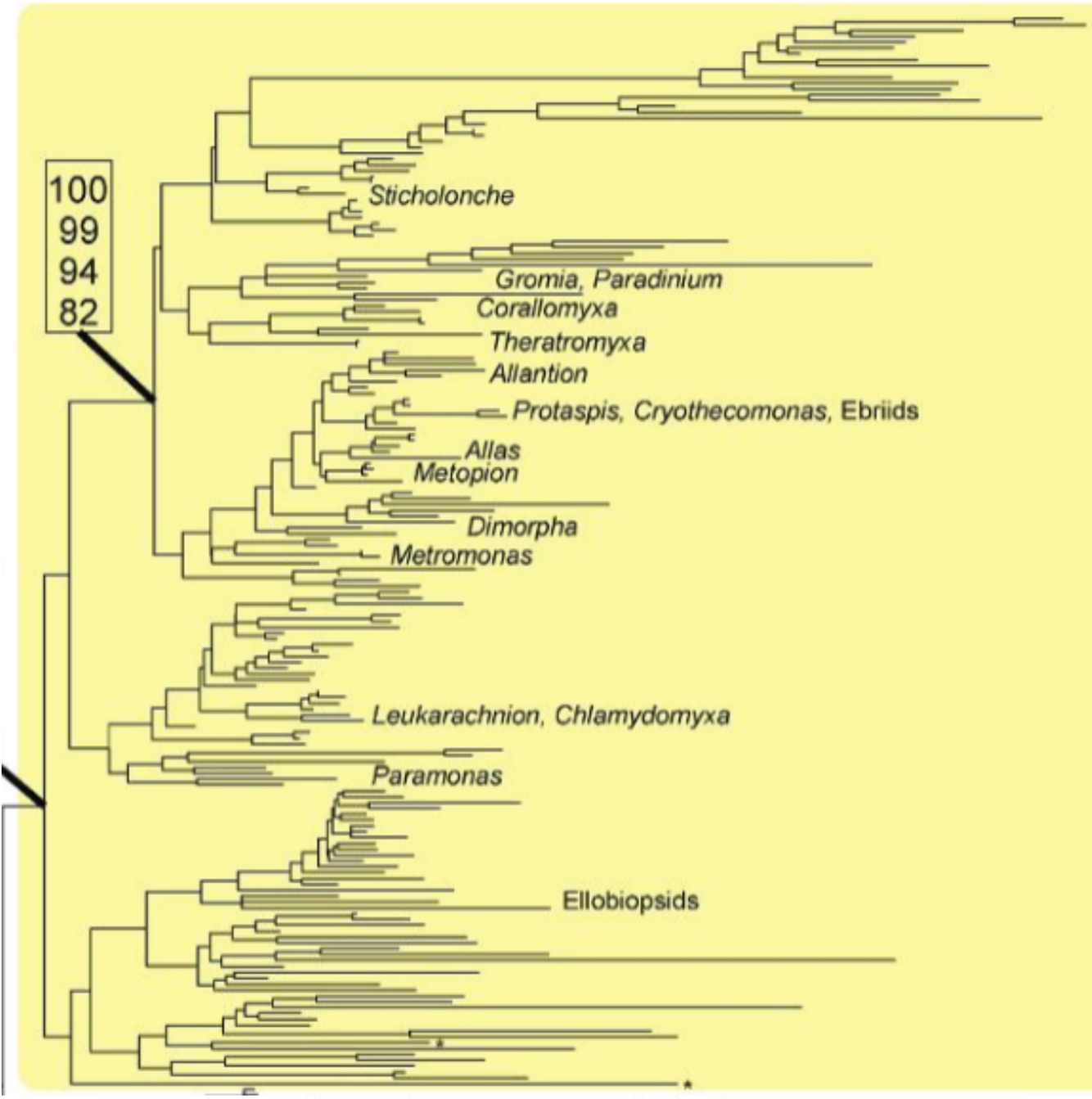


FIGURE 1. Most likely eukaryotic tree of life reconstructed using all 451 taxa and all 16 genes (SSU-rDNA plus 15 protein genes). Major nodes in this topology are robust to analyses of subsets of taxa and genes, which include varying levels of missing data (Table 1). Clades in bold are monophyletic in analyses with 2 or more members except in all:15 in which taxa represented by a single gene were sometimes misplaced. Numbers in boxes represent support at key nodes in analyses with increasing amounts of missing data (10:16, 6:16, 4:16, and all:16 analyses; see Table 1 for more details). Given uncertainties around the root of the eukaryotic tree of life (see text), we have chosen to draw the tree rooted with the well-supported clade Opisthokonta. Dashed line indicates alternate branching pattern seen for Amoebozoa in other analyses. Long branches, indicated by //, have been reduced by half. The 6 lineages labeled by * represent taxa that are misplaced, probably due to LBA, listed from top to bottom with expected clade in parentheses. These are *Protopalina japonica* (Stramenopiles), *Aggregata octopiana* (Apicomplexa), *Mikrocytos mackini* (Haplosporidia), *Centropyxis laevigata* (Tubulinea), *Marteilioides chungmuensis* (unplaced), and *Cochliopodium spiniferum* (Amoebozoa).



Foraminifera

Polycystina

Acantharea

Haplosporidia,

Plasmodiophora

Euglyphida, Cercomonadida

Thaumatomonads

Phaeodarea

Desmothoracids, **Gymnophrea**

Chlorarachniophytes

SAR:

Rhizaria

Diatoms

Brown algae

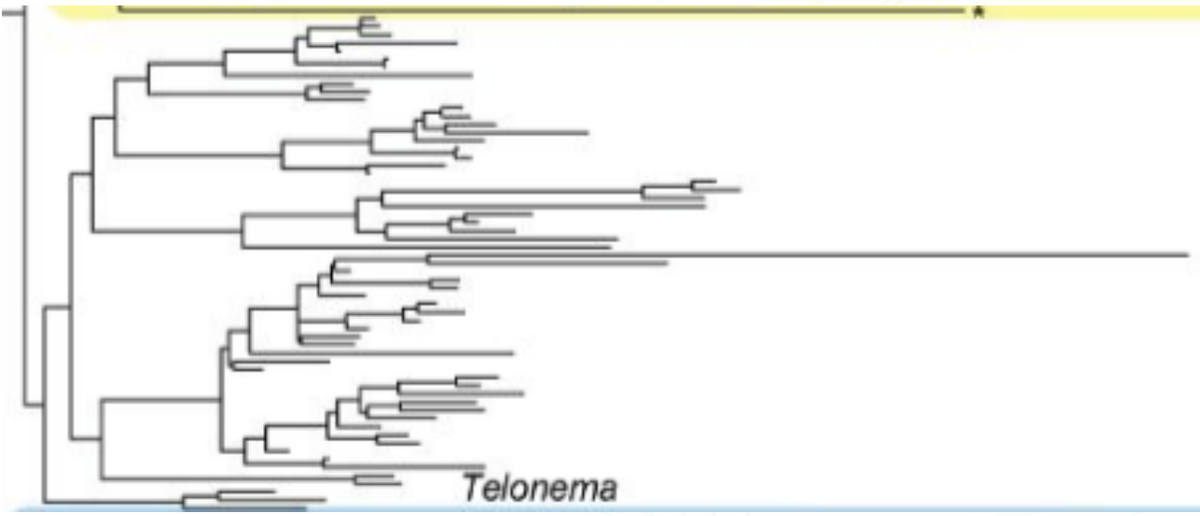
Stramenopiles

Dinoflagellates

Apicomplexa

Ciliates

Alveolates

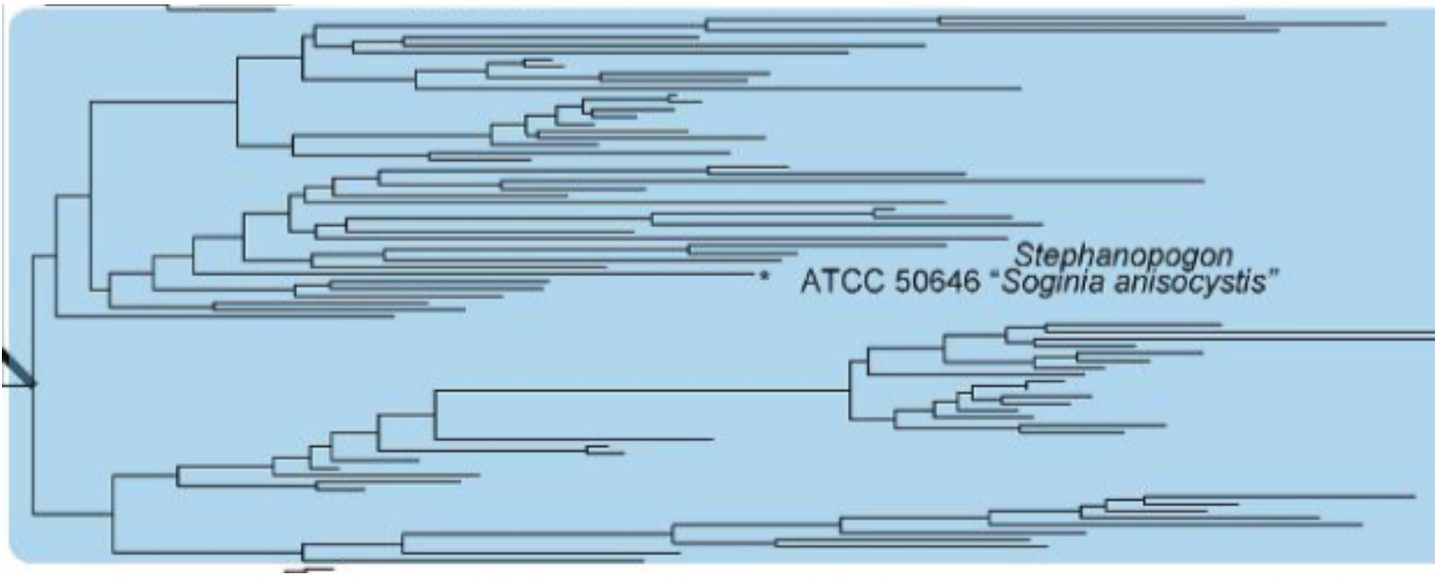


Cryptomonads, Kathablepharids
Haptophytes

Red algae

Green algae (including plants)

Glaucocystophytes



Euglenozoa

Heterolobosea

Jakobids
Malawimonas

Parabasalids

Preaxostyla

Fornicata

Excavata

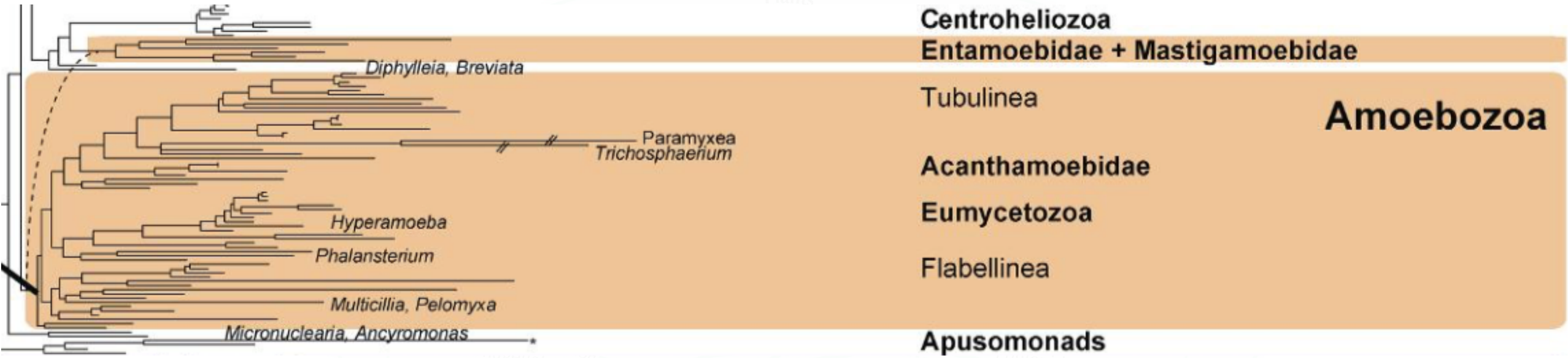

Diphylleia, Breviata
Paramyxea
Trichosphaerium
Hyperamoeba
Phalansterium
Multicillia, Pelomyxa
Micronuclearia, Ancyromonas
Ministeria
Centroheliozoa
Entamoebidae + Mastigamoebidae
Tubulinea
Amoebozoa
Acanthamoebidae
Eumycetozoa
Flabellinea
Apusomonads
Animals
Choanoflagellates
Mesomycetozoa
Fungi
Microsporidia, Nucleariidae
Opisthokonts

Fig. 2: 88 Taxa each w/ 10 or more of the 16 genes

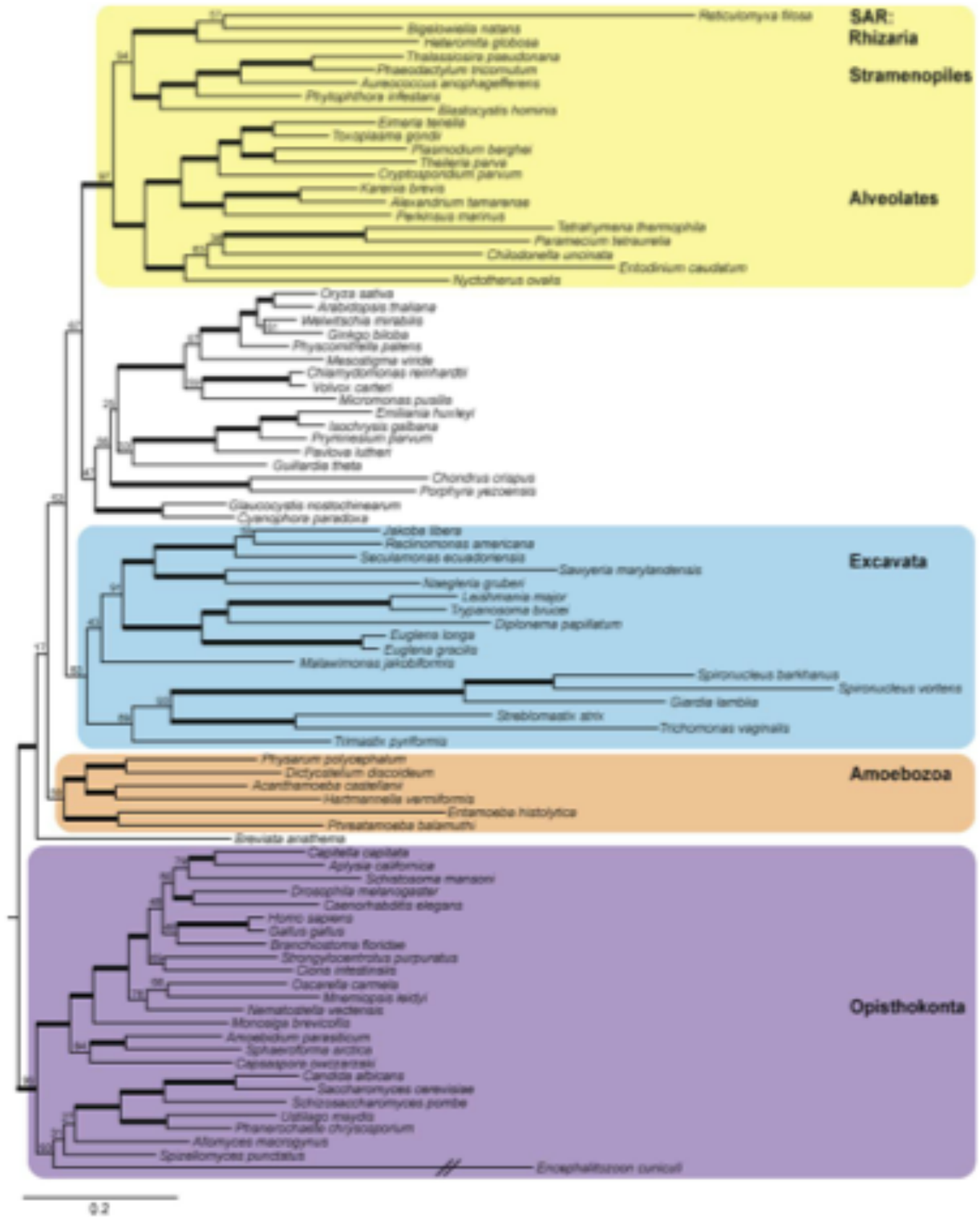
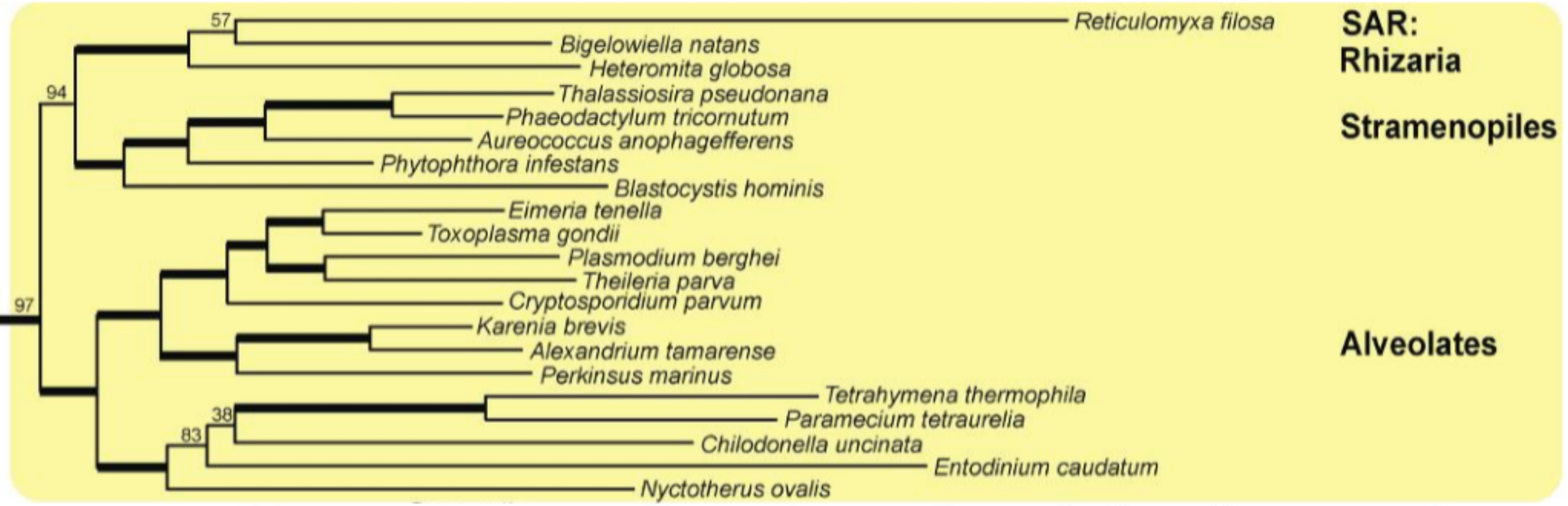
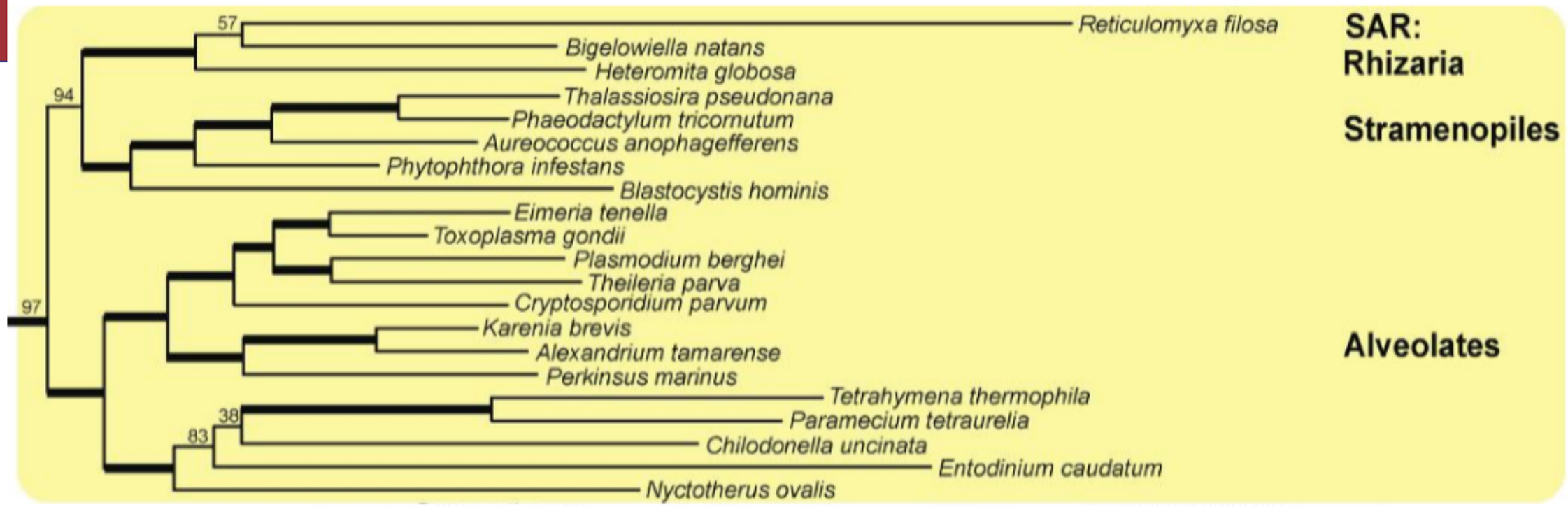
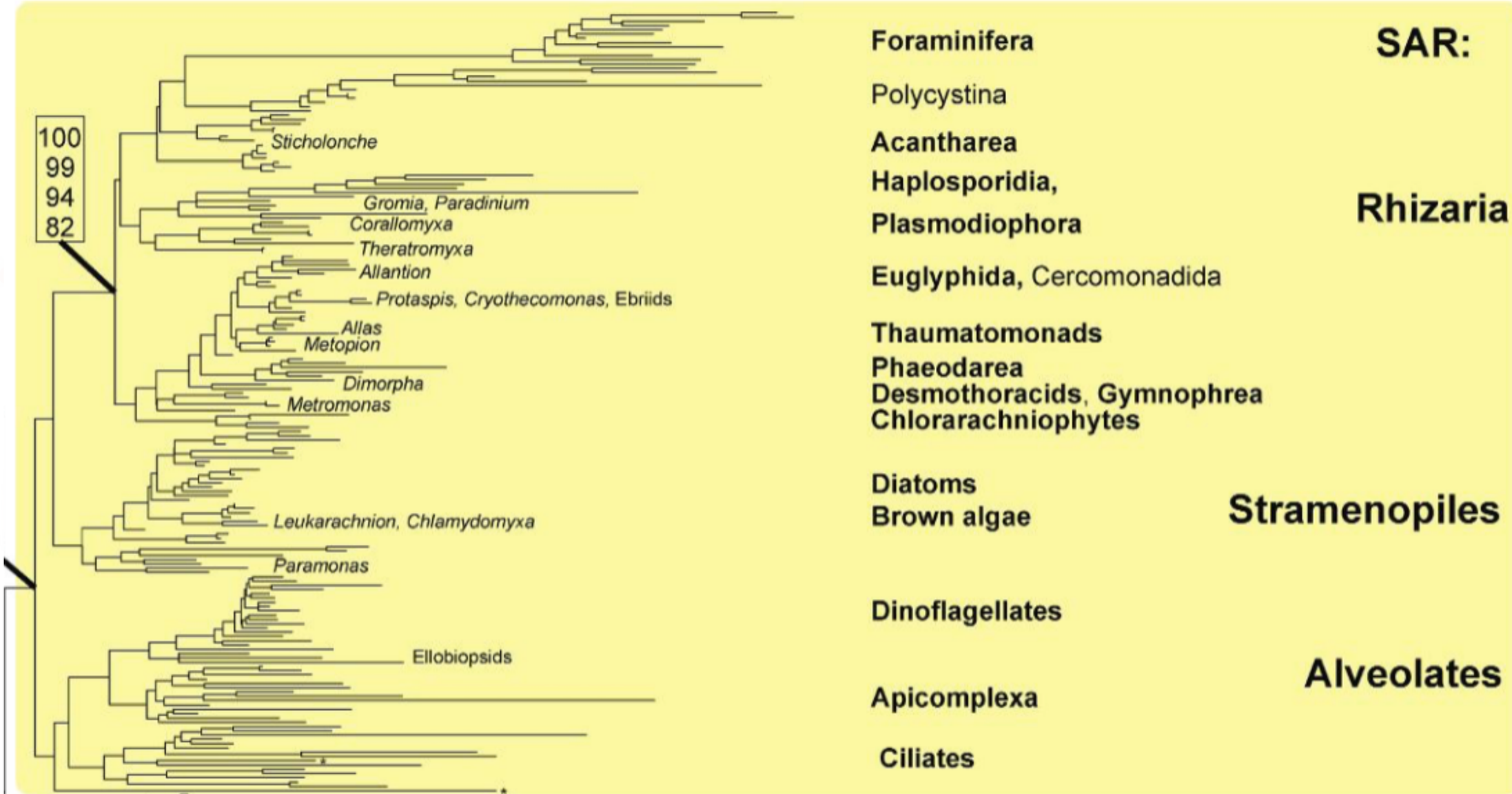
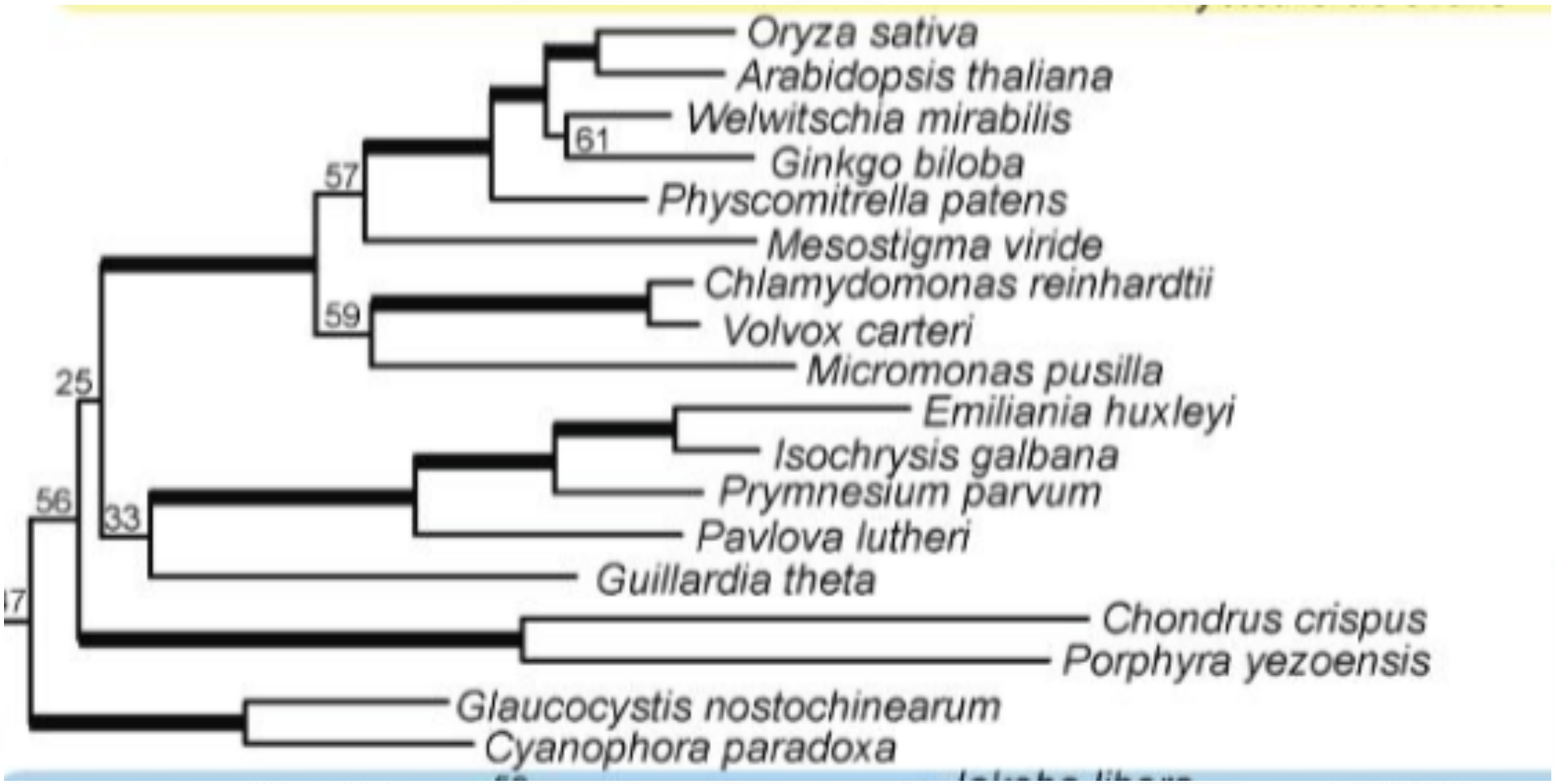


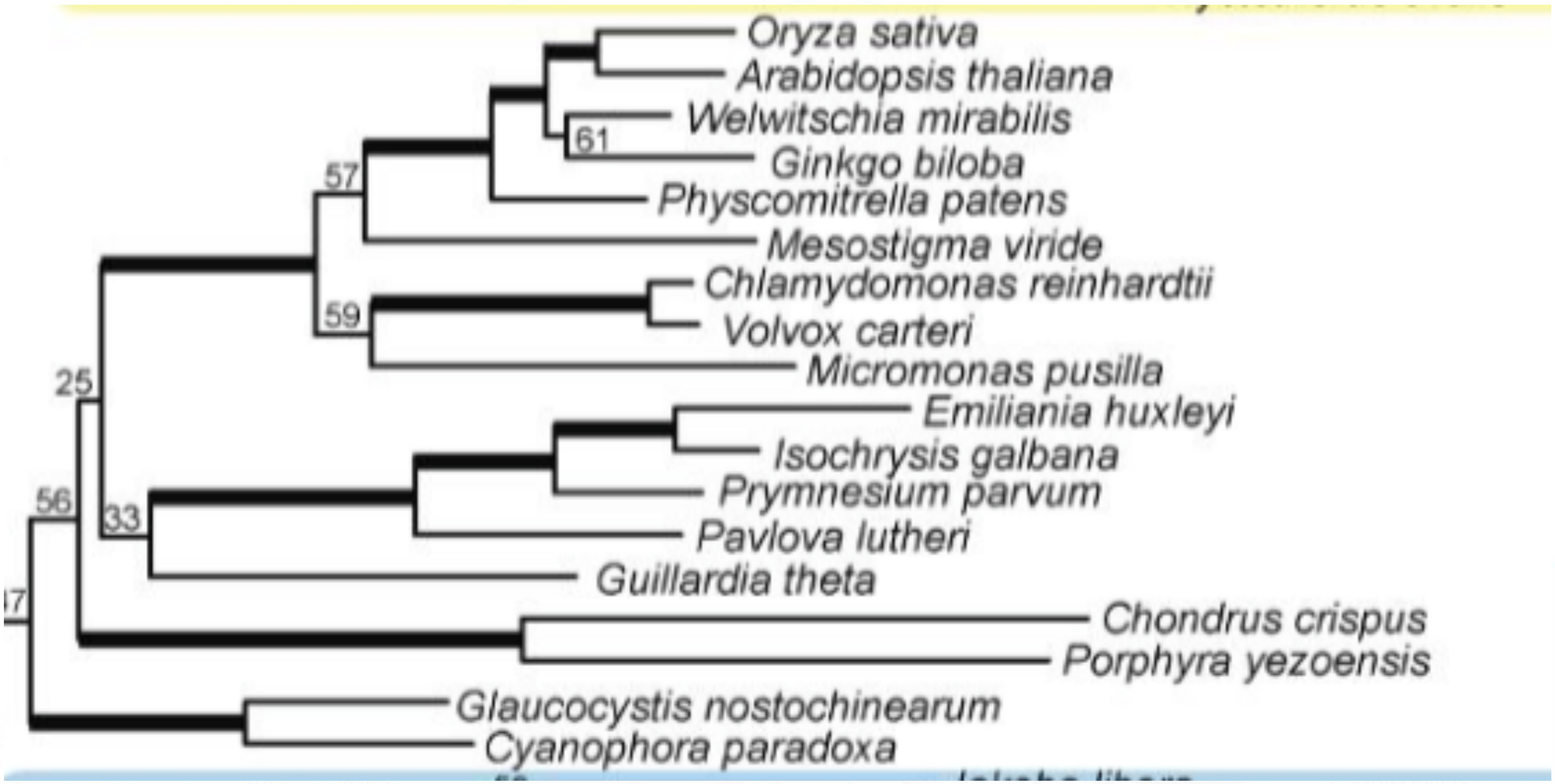
FIGURE 2. Most likely eukaryotic tree of life reconstructed with 10:16, which includes 88 taxa (each with 10 or more of the genes analyzed in this study) and 16 genes (SSU-rDNA plus 15 protein genes). Thickened lines receive >95% bootstrap support. Other notes as in Methods section and Figure 1.

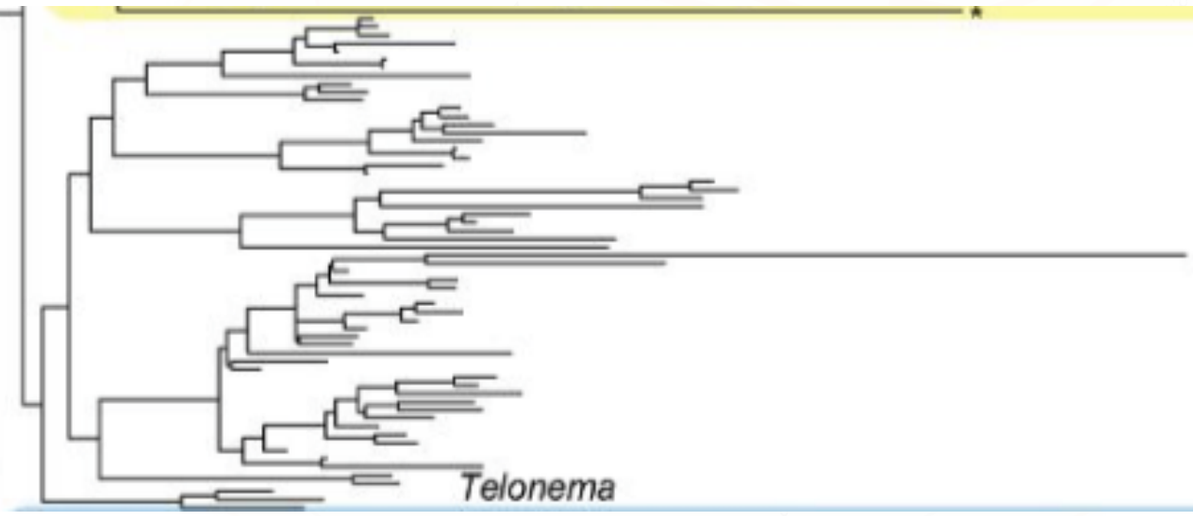

57
94
97
38
83
Reticulomyxa filosa
Bigelowiella natans
Heteromita globosa
Thalassiosira pseudonana
Phaeodactylum tricornutum
Aureococcus anophagefferens
Phytophthora infestans
Blastocystis hominis
Eimeria tenella
Toxoplasma gondii
Plasmodium berghei
Theileria parva
Cryptosporidium parvum
Karenia brevis
Alexandrium tamarense
Perkinsus marinus
Tetrahymena thermophila
Paramecium tetraurelia
Chilodonella uncinata
Entodinium caudatum
Nyctotherus ovalis
SAR:
Rhizaria
Stramenopiles
Alveolates


57
94
97
38
83
Reticulomyxa filosa
Bigelowiella natans
Heteromita globosa
Thalassiosira pseudonana
Phaeodactylum tricornutum
Aureococcus anophagefferens
Phytophthora infestans
Blastocystis hominis
Eimeria tenella
Toxoplasma gondii
Plasmodium berghei
Theileria parva
Cryptosporidium parvum
Karenia brevis
Alexandrium tamarense
Perkinsus marinus
Tetrahymena thermophila
Paramecium tetraurelia
Chilodonella uncinata
Entodinium caudatum
Nyctotherus ovalis
SAR:
Rhizaria
Stramenopiles
Alveolates


100
99
94
82
Sticholonche
Gromia, Paradinium
Corallomyxa
Theratromyxa
Allantion
Protaspis, Cryothecomonas, Ebriids
Allas
Metopion
Dimorpha
Metromonas
Leukarachnion, Chlamydomyxa
Paramonas
Ellobiopsids
Foraminifera
Polycystina
Acantharea
Haplosporidia,
Plasmodiophora
Euglyphida, Cercomonadida
Thaumatomonads
Phaeodarea
Desmothoracids, Gymnophrea
Chlorarachniophytes
Diatoms
Brown algae
Dinoflagellates
Apicomplexa
Ciliates
SAR:
Rhizaria
Stramenopiles
Alveolates


Oryza sativa
Arabidopsis thaliana
Welwitschia mirabilis
61
Ginkgo biloba
57
Physcomitrella patens
Mesostigma viride
Chlamydomonas reinhardtii
59
Volvox carteri
Micromonas pusilla
25
Emiliania huxleyi
Isochrysis galbana
Prymnesium parvum
56
33
Pavlova lutheri
Guillardia theta
Chondrus crispus
Porphyra yezoensis
Glaucocystis nostochinearum
Cyanophora paradoxa



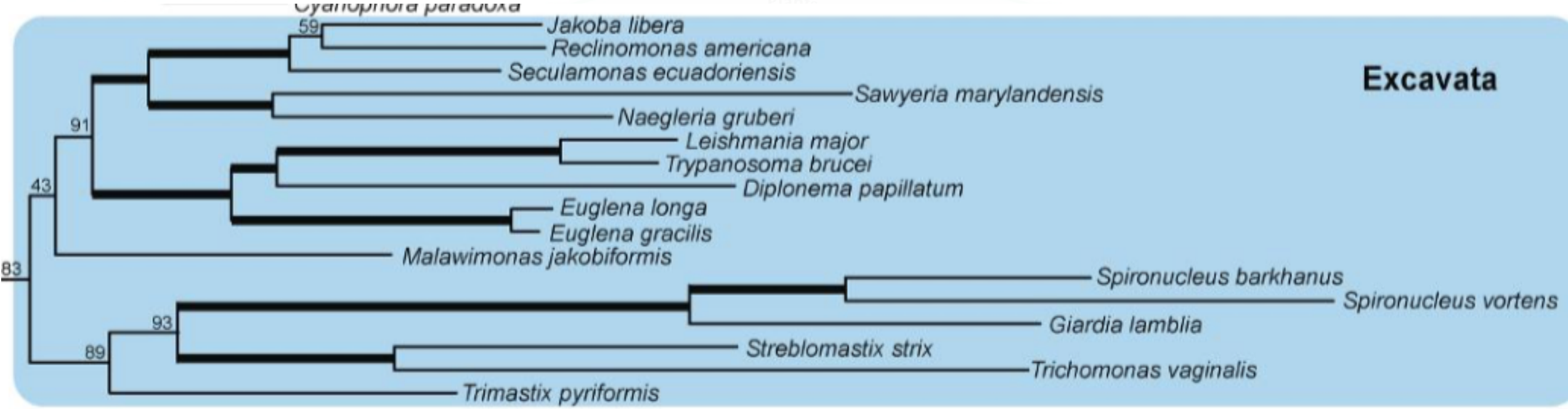


Cryptomonads, Kathablepharids
Haptophytes

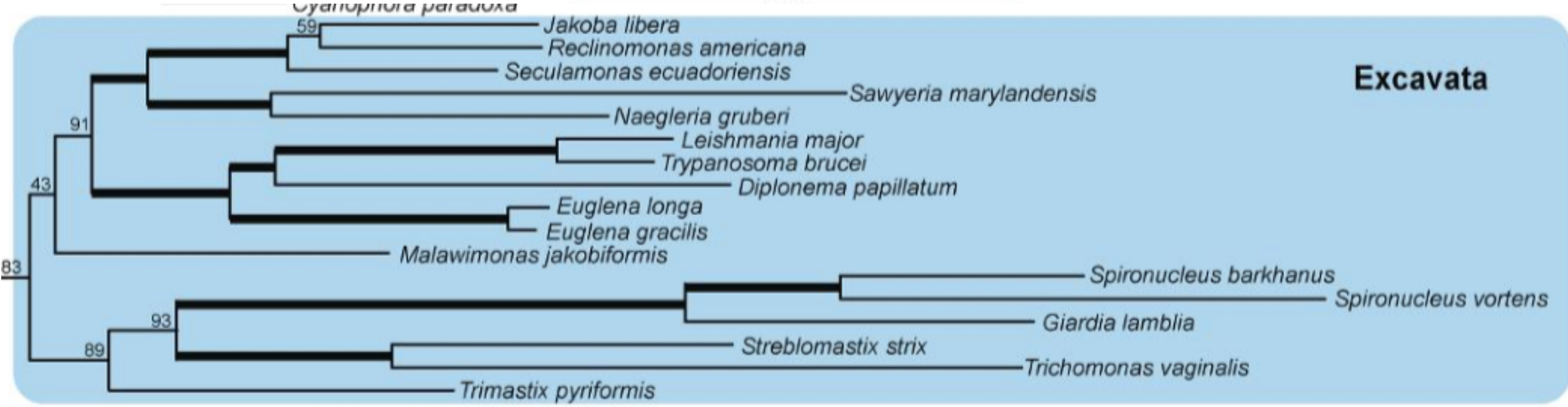
Red algae

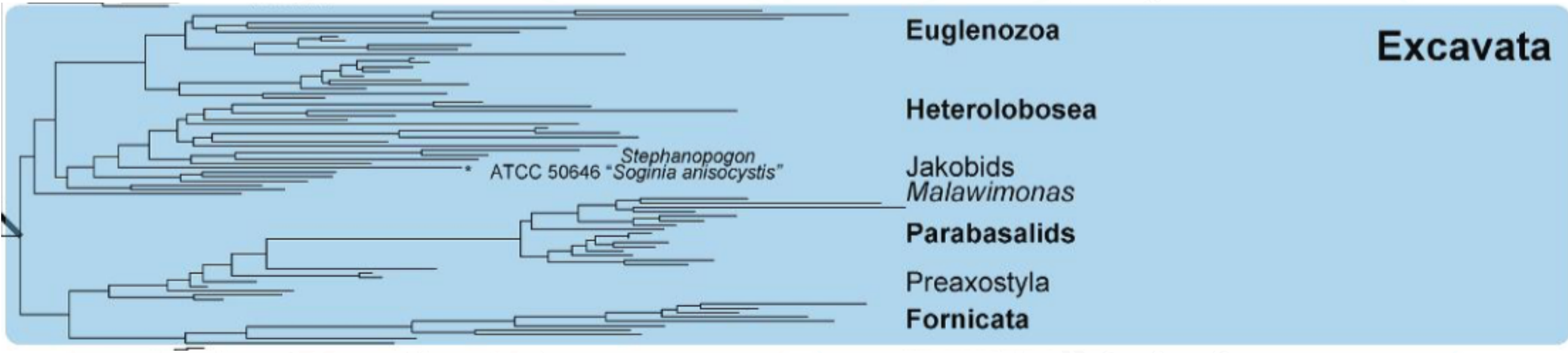
Green algae (including plants)

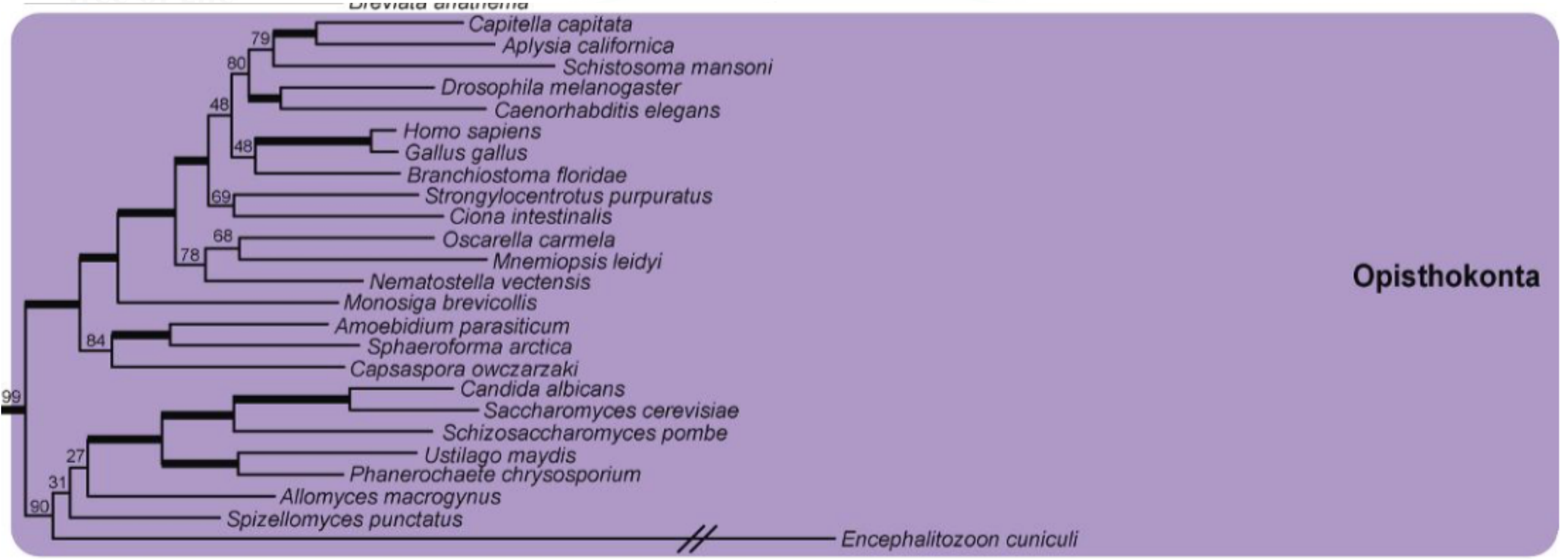
Glaucocystophytes

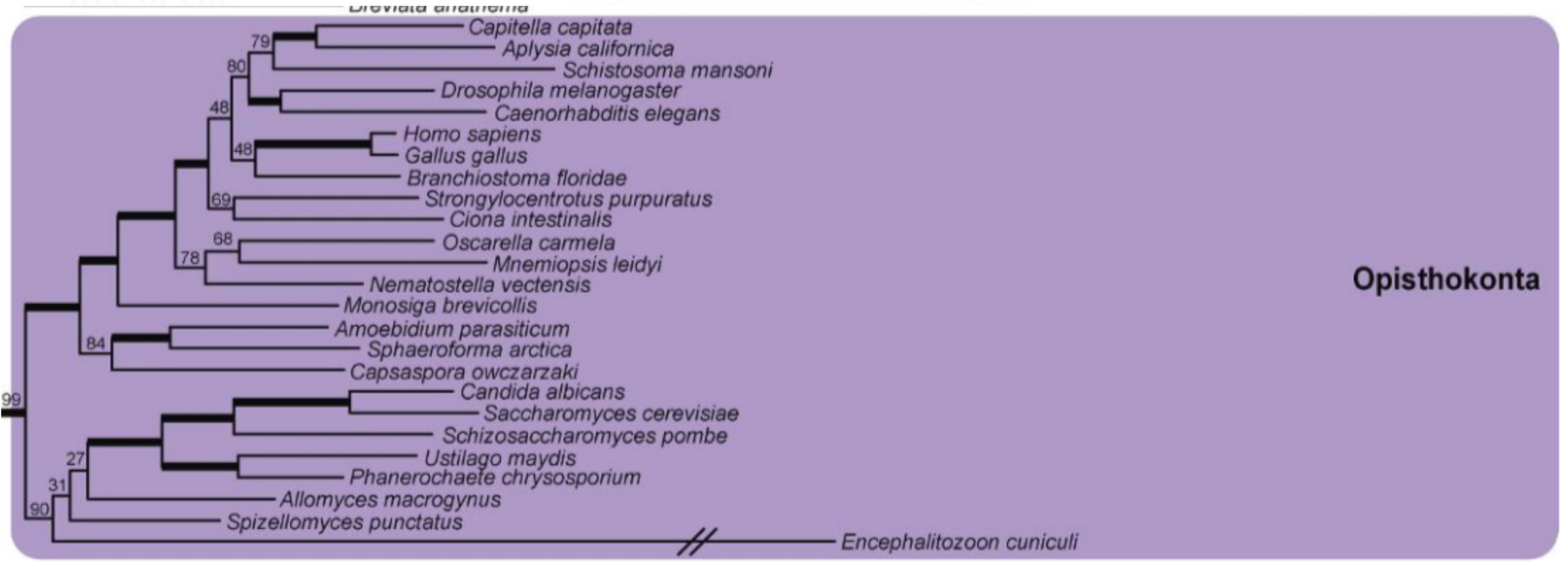

Cyanophora paradoxa
Jakoba libera
Reclinomonas americana
Seculamonas ecuadoriensis
Sawyeria marylandensis
Naegleria gruberi
Leishmania major
Trypanosoma brucei
Diplonema papillatum
Euglena longa
Euglena gracilis
Malawimonas jakobiformis
Spironucleus barkhanus
Spironucleus vortens
Giardia lamblia
Streblomastix strix
Trichomonas vaginalis
Trimastix pyriformis
59
91
43
83
93
89
Excavata

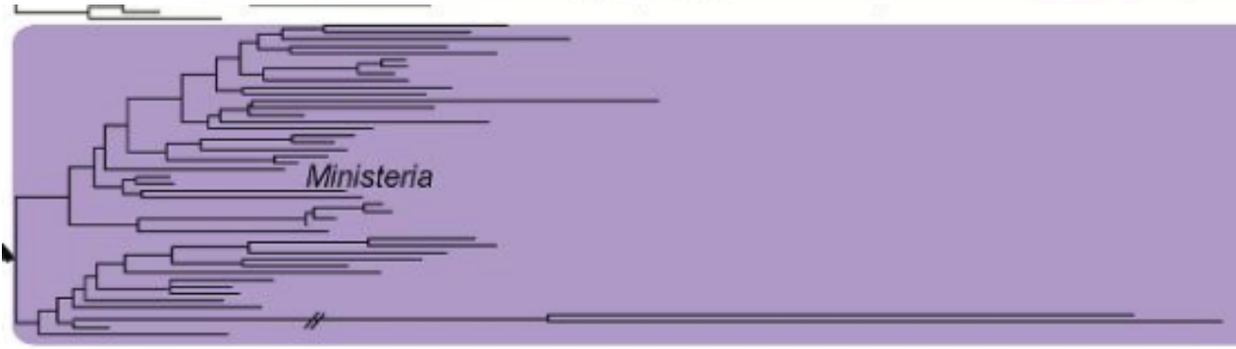
Tree of Life
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Nanoarchaeota
Euryarchaea
ROOT?
Thermotogales
Aquifex
Cytophagales/Flavobacteria
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Actinobacteria
Deinococcus/Thermus
Cyanobacteria
Plastids
Mycoplasmas
Firmicutes
Bacteria


Cyanophora paradoxa
59
Jakoba libera
Reclinomonas americana
Seculamonas ecuadoriensis
Sawyeria marylandensis
Naegleria gruberi
91
Leishmania major
Trypanosoma brucei
Diplonema papillatum
43
Euglena longa
Euglena gracilis
Malawimonas jakobiformis
83
Spironucleus barkhanus
Spironucleus vortens
93
Giardia lamblia
Streblomastix strix
89
Trichomonas vaginalis
Trimastix pyriformis
Excavata


Euglenozoa
Heterolobosea
Stephanopogon
ATCC 50646 "Soginia anisocystis"
Jakobids
Malawimonas
Parabasalids
Preaxostyla
Fornicata
Excavata







Animals

Choanoflagellates
Mesomycetozoa

Fungi
Microsporidia, Nucleariidae

Opisthokonts

(reviewed in Roger and Simpson 2009; Tekle et al. 2009). In our analyses, we find at best moderate support for "Unikonta" (Table 1), but concatenated analyses such as these cannot resolve the root.

In exploring the tradeoffs between increasing taxo-

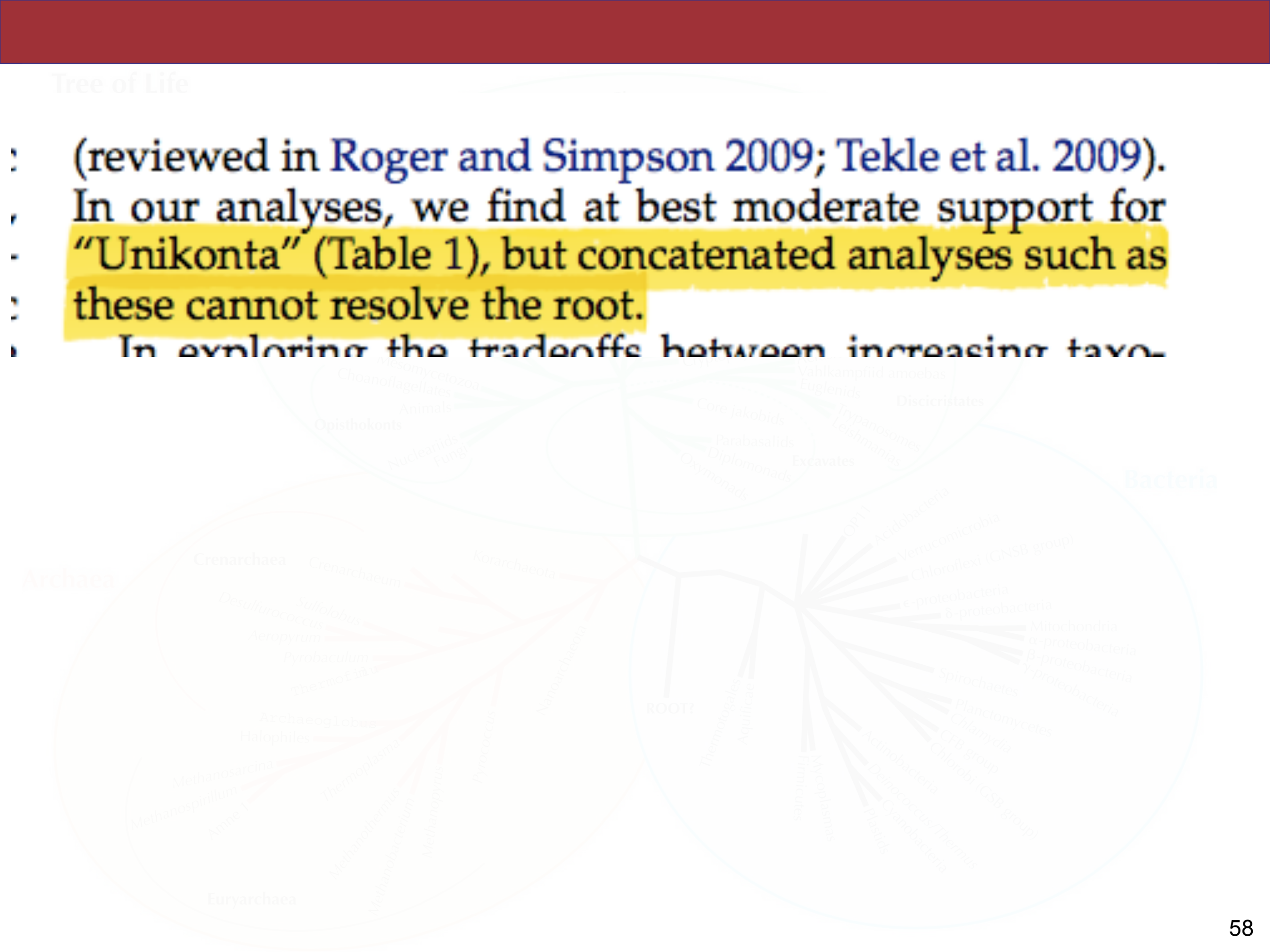


TABLE 1. Support for major clades of eukaryotes in analyses containing varying levels of taxon inclusion and missing data

	10:16	6:16	4: 16	all:16	9:15	5:15	3:15	all:15
Supported clades								
Opisthokonta	**99**[a]	**97**[a]	**97**[a]	69	**100**[a]	**99**[a]	**85**	19
Rhizaria	**100**[a]	**99**[a]	**94**[a]	**82**	**100**[a]	**100**[a]	47	29
SAR	**97**[a]	**98**[a]	63	22	**100**[a]	**100**[a]	32	19
Rhizaria + stramenopiles	**94**[a]	**94**[a]	57	26	**92**[a]	**96**[a]	29	18
Excavata	**83**	**77**	65	6	**84**	**76**	44	19
Amoebozoa	59	46	49	nm	68	56	44	5
"Unikonta"	63	39	21	nm	54	50	15	3
Weak/unsupported hypotheses								
"Archaeplastida"	nm	nm	nm	nm	nm	nm	nm	nm
"Chromalveloata"	nm	nm	nm	nm	nm	nm	nm	nm
Cryptomonads + haptophytes	33	50	nm	29	38	56	22	25
Haptophytes + SAR	nm	nm	15	nm	nm	nm	nm	nm
Alveolates + stramenopiles	nm	nm	nm	nm	nm	nm	nm	nm
Red algae + green algae	nm	nm	nm	nm	nm	nm	nm	nm
Red, Green, Glauco, Hapto, Crypt	47	32	nm	9	39	27	16	8
Data set statistics								
Number of taxa	88	111	160	451	88	111	160	240
Number of lineages	26	30	45	72	26	30	45	54
% Missing data (characters)	17	25	38	69	19	28	43	59

Note: Supported clades are stable across analyses, albeit with decreasing support as the percentage of missing data increases. Bootstrap support values from RAxML analyses. Support values greater than 75 are indicated by bold text and greater than 85 are indicated with a. nm = nonmonophyletic. Column headings describe the data sets. For example, "10:16" includes all taxa that have at least 10 of the 16 genes, with a total of 88 taxa representing 26 lineages and containing 17% missing data. The "all:15" includes the protein-coding genes from all taxa and contains 59% missing data. See Table S2 for lineagesand Figure S1a–h for individual trees.

[a]Support values greater than 85.

Just Rhizaria

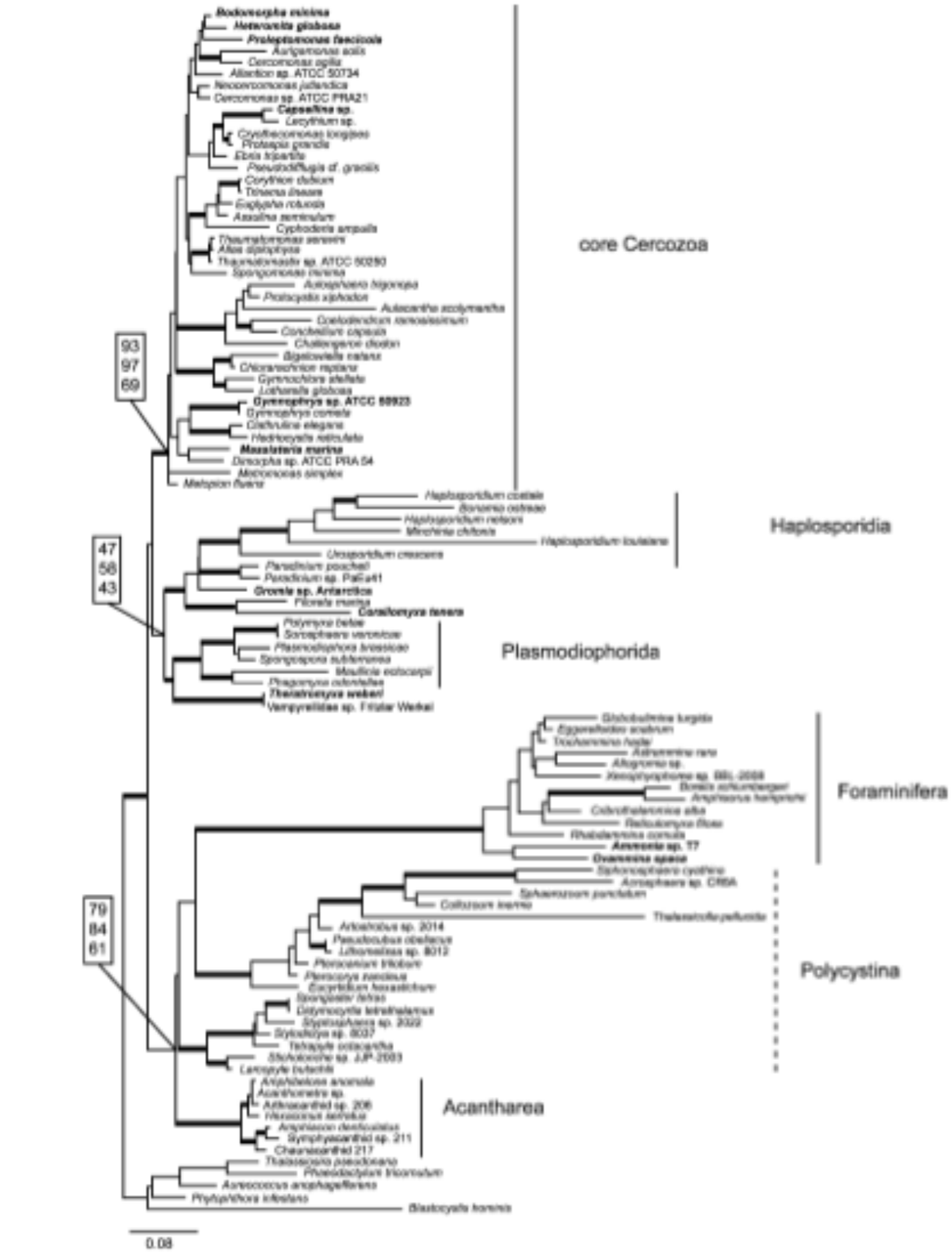


FIGURE 3. Maximum likelihood tree of Rhizaria reconstructed with 103 Rhizaria taxa and 16 genes. The SSU-rDNA partition was analyzed with GTR+gamma and proteins with rtREV. Thickened lines receive >80% bootstrap support in all analyses. Node support in boxes from Rhizaria:4-gene, Rhizaria:16-gene, all:16 analyses. Taxa with new data are in bold. Dashed lines indicate nonmonophyly.

Just Excavata

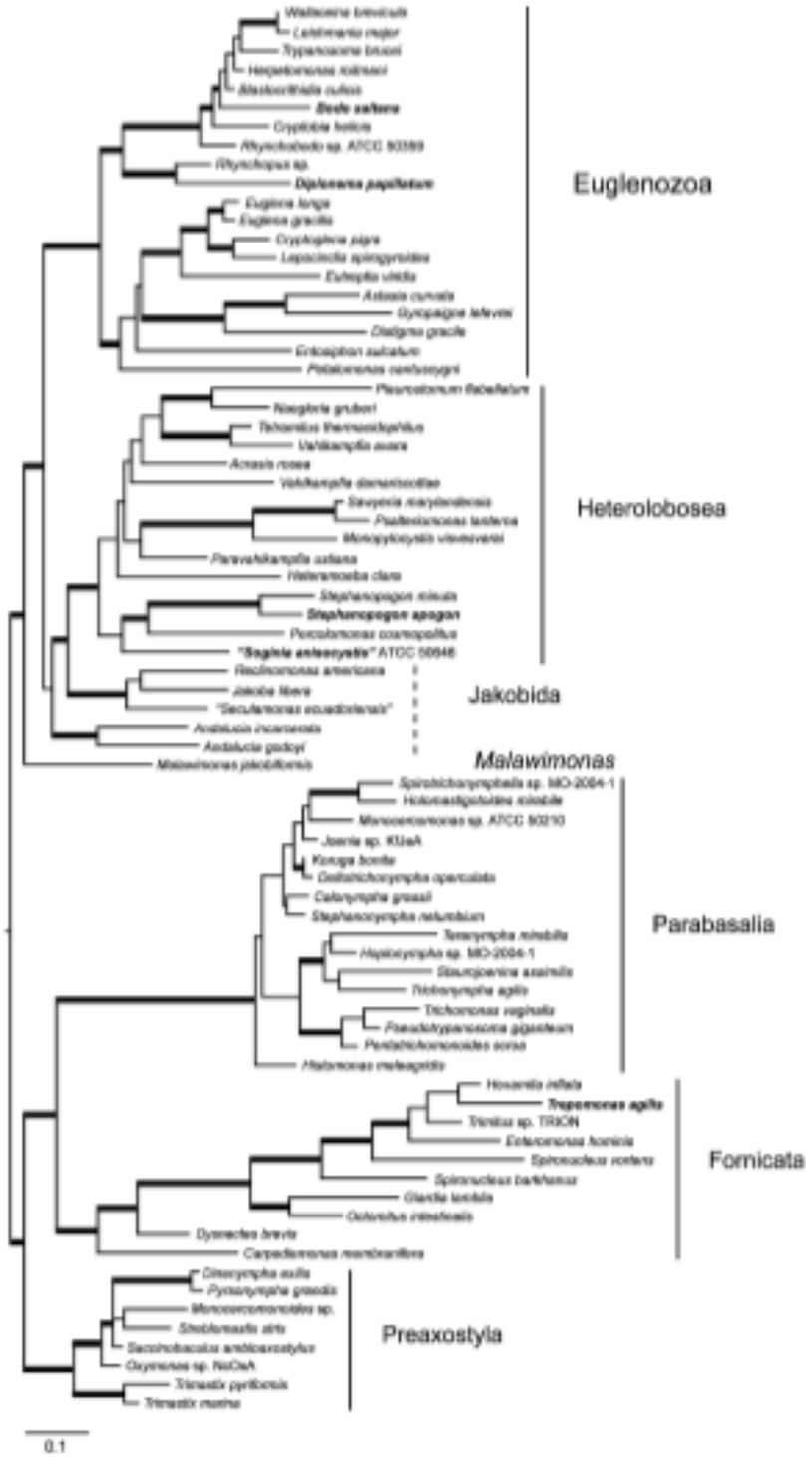


FIGURE 4. Maximum-likelihood tree of Excavata with 75 taxa and 16 genes. The SSU-rDNA partition was analyzed with GTR+gamma and proteins with rtREV. See Figure 3 for other notes.

Consensus

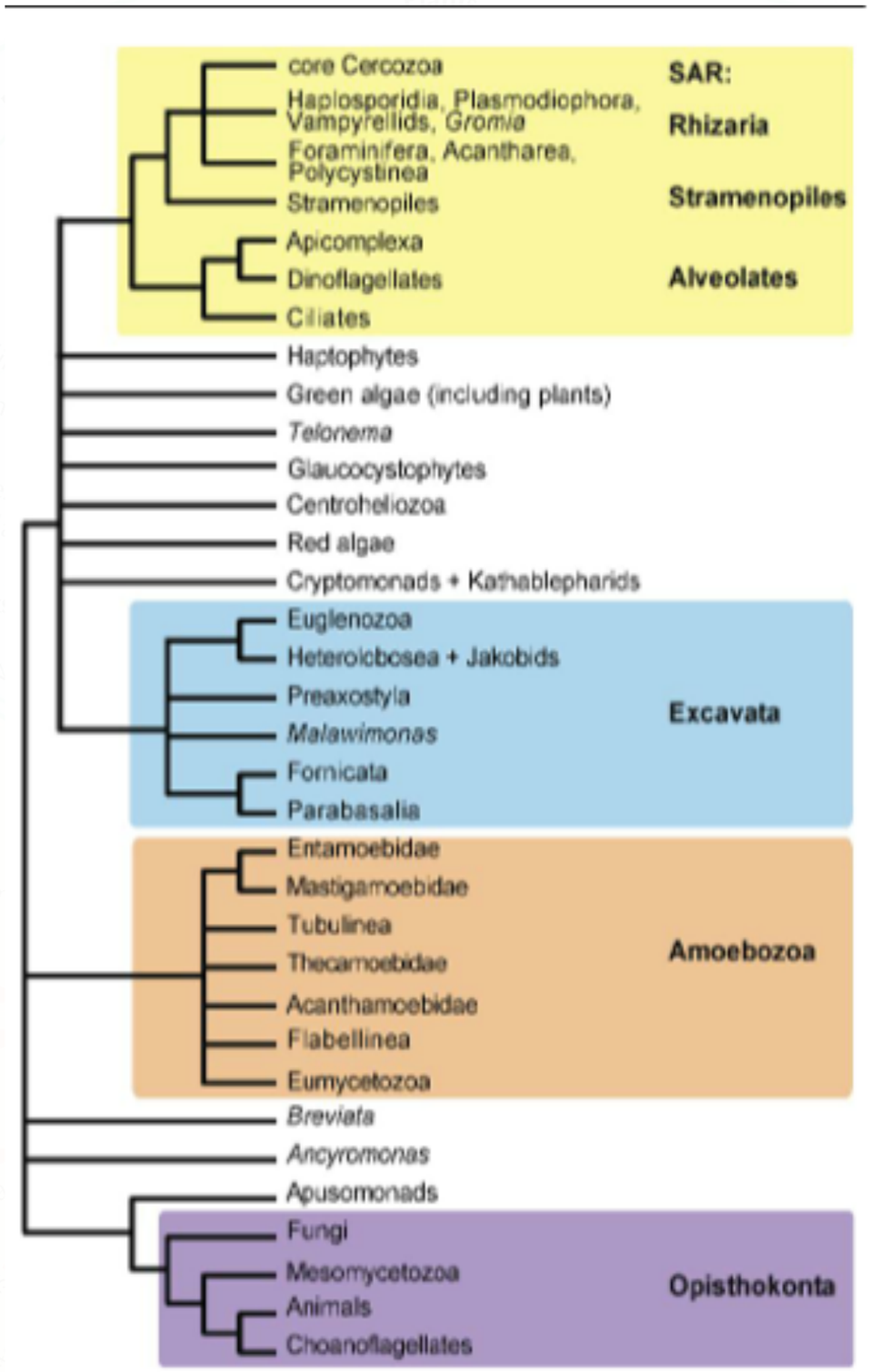
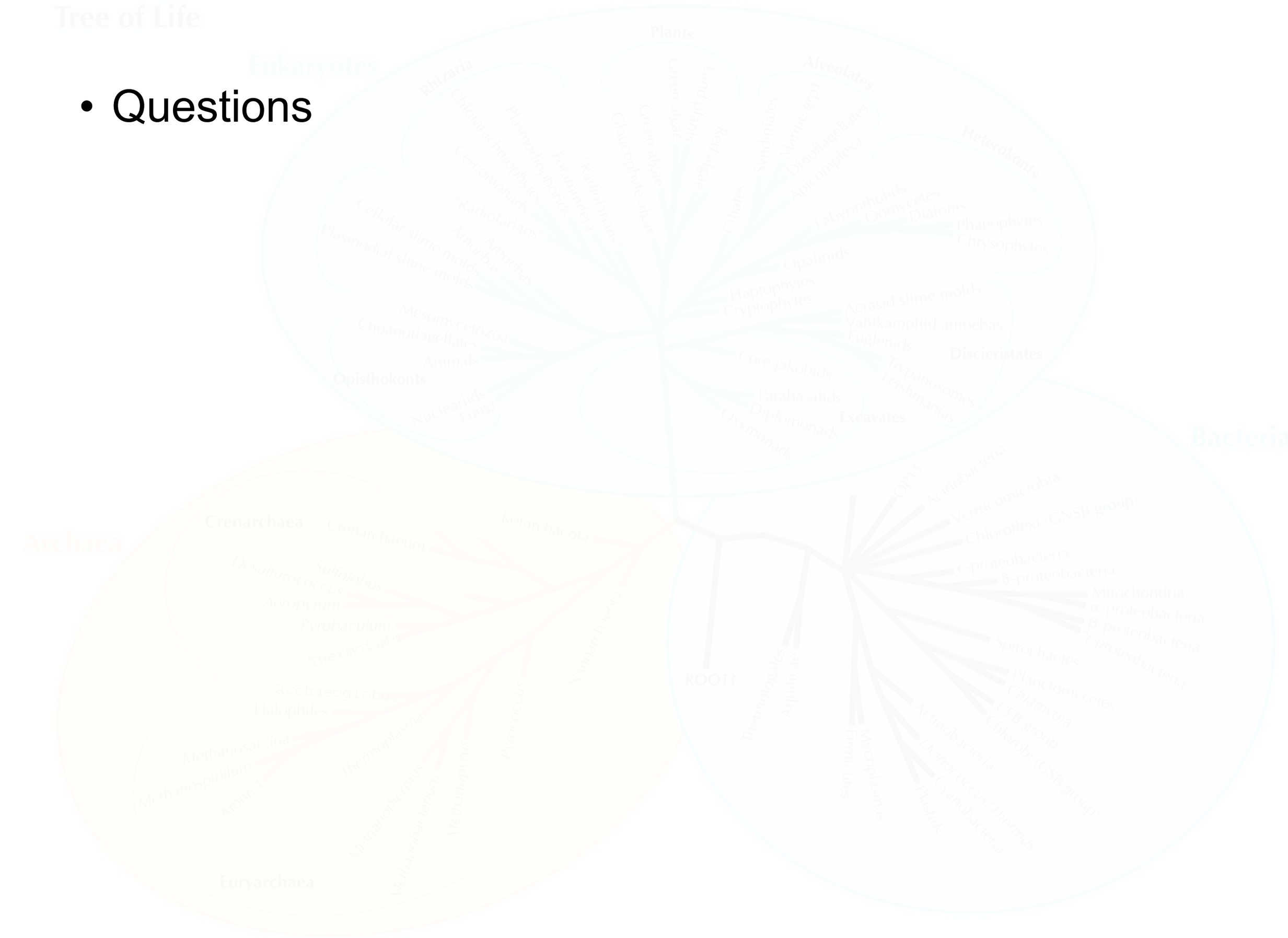


FIGURE 5. Summary of major findings—the evolutionary relationships among major lineages of eukaryotes. Clades have been collapsed into those that we view to be strongly supported. The many polytomies represent uncertainties that remain.

Results and Discussions

- Questions



Conclusions

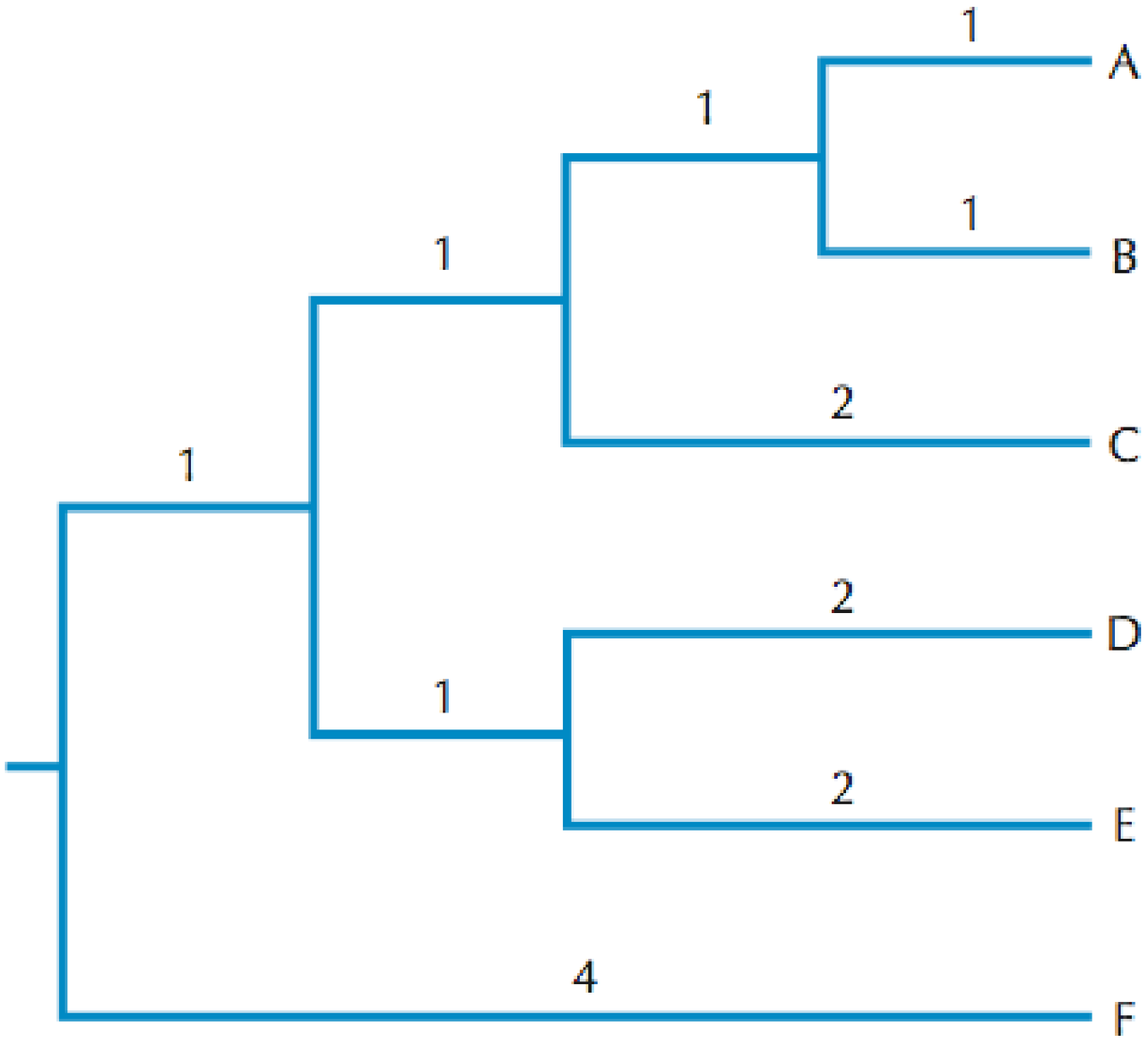
CONCLUSIONS

The robust tree of life emerging from this study demonstrates the benefits of improved taxon sampling for reconstructing deep phylogeny as our analyses produce stable topologies that include a broad representation of eukaryotes. The current study, combined with insights from other studies referenced herein, has refined the eukaryotic tree of life from over 70 major lineages (Patterson 1999) to ~16 major groups (Fig. 5, http://eutree.lifedesks.org/). Most significantly, we attribute the stability of major clades (e.g., Excavata, Amoebozoa, Opisthokonta, and SAR) to broader taxonomic sampling combined with analyses of sufficient characters (16 genes or 6578 characters). In our view, inclusion of more taxa coupled with carefully chosen genes is necessary to further resolve the 16 or so major lineages of microbial eukaryotes for which sister group relationships remain uncertain.

UPGMA

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic mean (UPGMA) algorithm

The True Tree



Distance Matrix For True Tree

TABLE 27.6. Distance matrix

OTUs	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

Collapse Diagonal

TABLE 27.7. Diagonal matrix—Step 1

OTUs	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

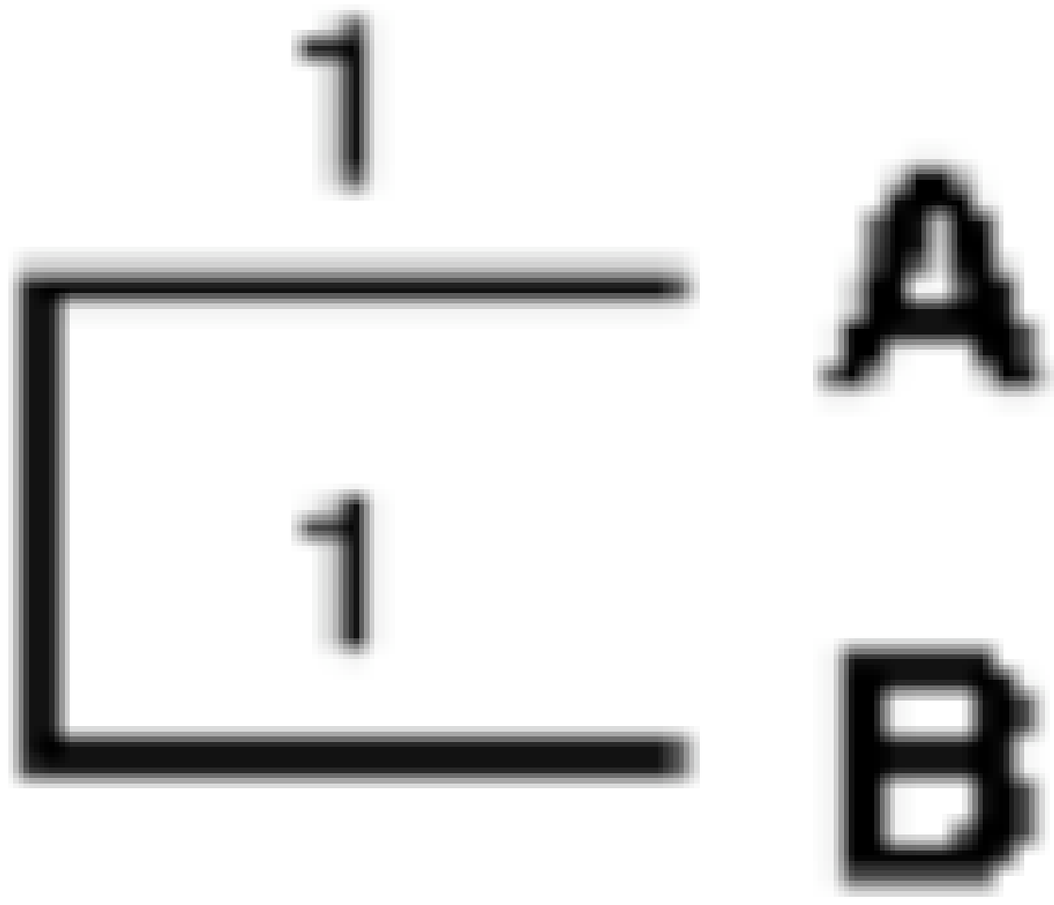
Identify Lowest D

OTUs	A	B	C	D	E
B	**2**				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Join Those Two Taxa

- Create branch with length = D
- 2
- A----------B

Make D from Node Equal



Create New Distance Matrix

- Merge Two OTUs joined in previous step (AB)

- $D_{x,AB} = 0.5 * (D_{x,A} + D_{x,B})$

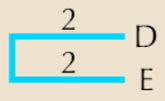
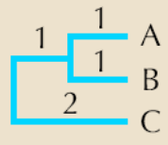
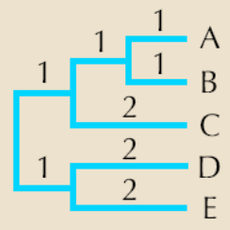
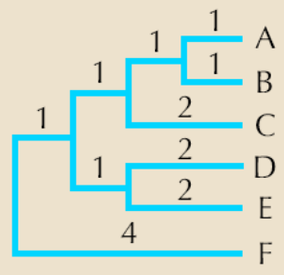
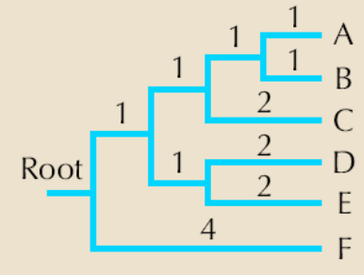
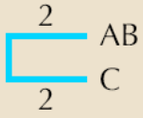
New Matrix

TABLE 27.8. Diagonal matrix—Step 2

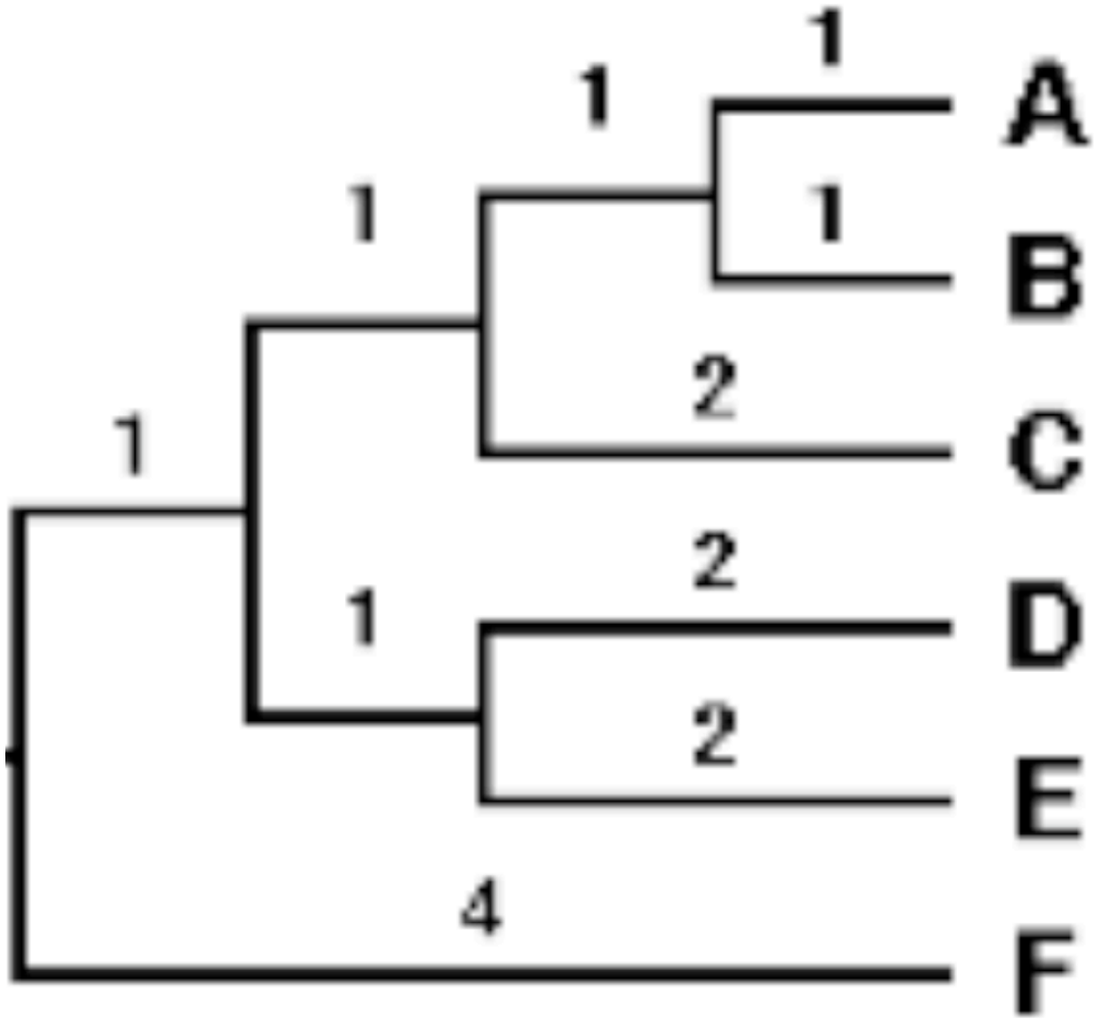
OTUs	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

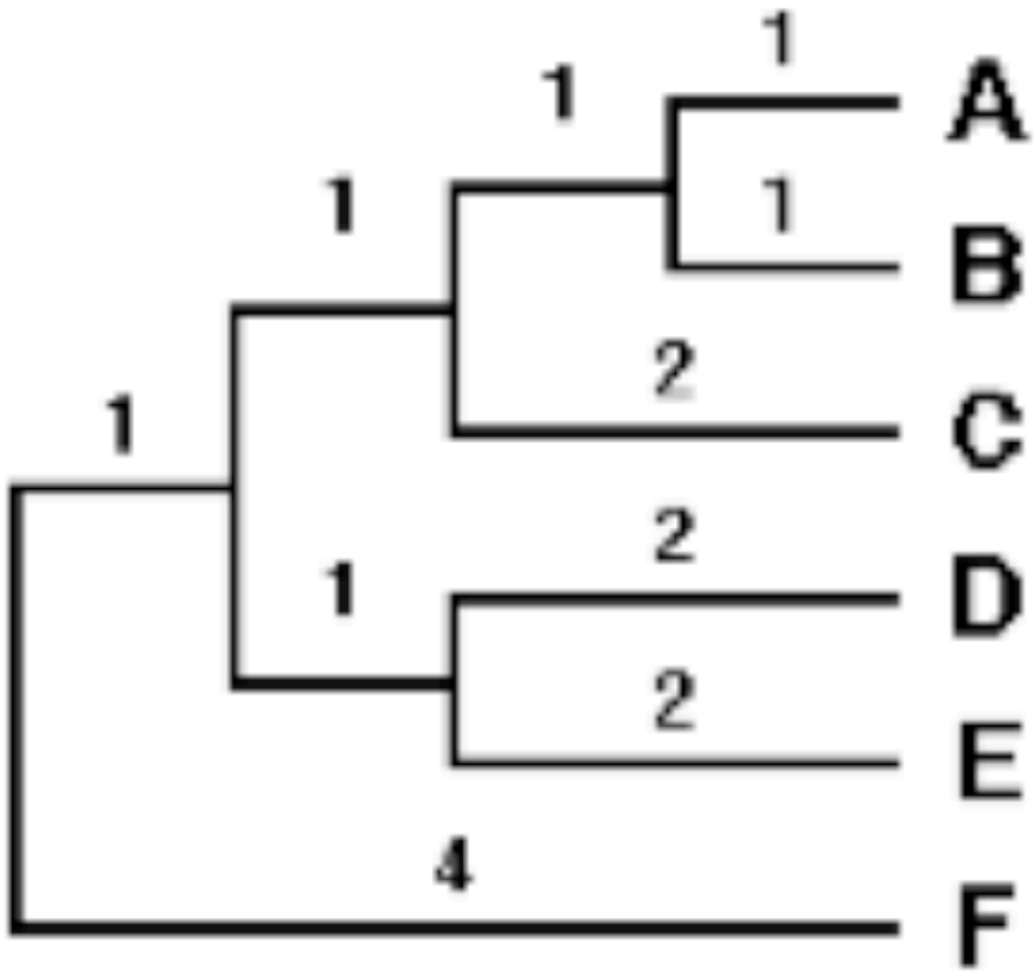
UPGMA

TABLE 27.9. Example of UPGMA tree construction

Step	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Distance matrix	OTUs | A B C D E B | 2 C | 4 4 D | 6 6 6 E | 6 6 6 4 F | 8 8 8 8 8	OTUs | AB C D E C | 4 D | 6 6 E | 6 6 4 F | 8 8 8 8	OTUs | AB C DE C | 4 DE | 6 6 DE | 8 8 8	OTUs | ABC DE DE | 6 F | 8 8	OTUs | ABCDE F | 8	No new matrix
Identify	A↔B = 2 smallest D	AB↔C = 4 D↔E= 4	AB↔DE = 6 C↔DE = 6	ABC↔DE	ABCDE↔F	
Taxa joined	A and B	D and E	AB and C	ABC and DE	ABCDE and F	
Subtree						
Comments on tree drawing	The distance between A and B is 2 units. A subtree is drawn with the branch point halfway between the two. Thus, each branch is 1 unit in length.	Branching done as in Step 1. Because the distance from AB to C is also 4, that pair could have been selected as well.	First a subtree is drawn with AB and C:  The the AB subtree is attached to the AB branch at a point equal to the length of the A and B branches.	The tree is first done as in Step 3 with the ABC and DE subtrees replacing the branches.	The tree is now complete but unrooted.	The tree can then be rooted using midpoint rooting which tries to balance all the tips to reach the same end point. Note this is the tree that we started with to build the distance matrix.

Compare to True Tree

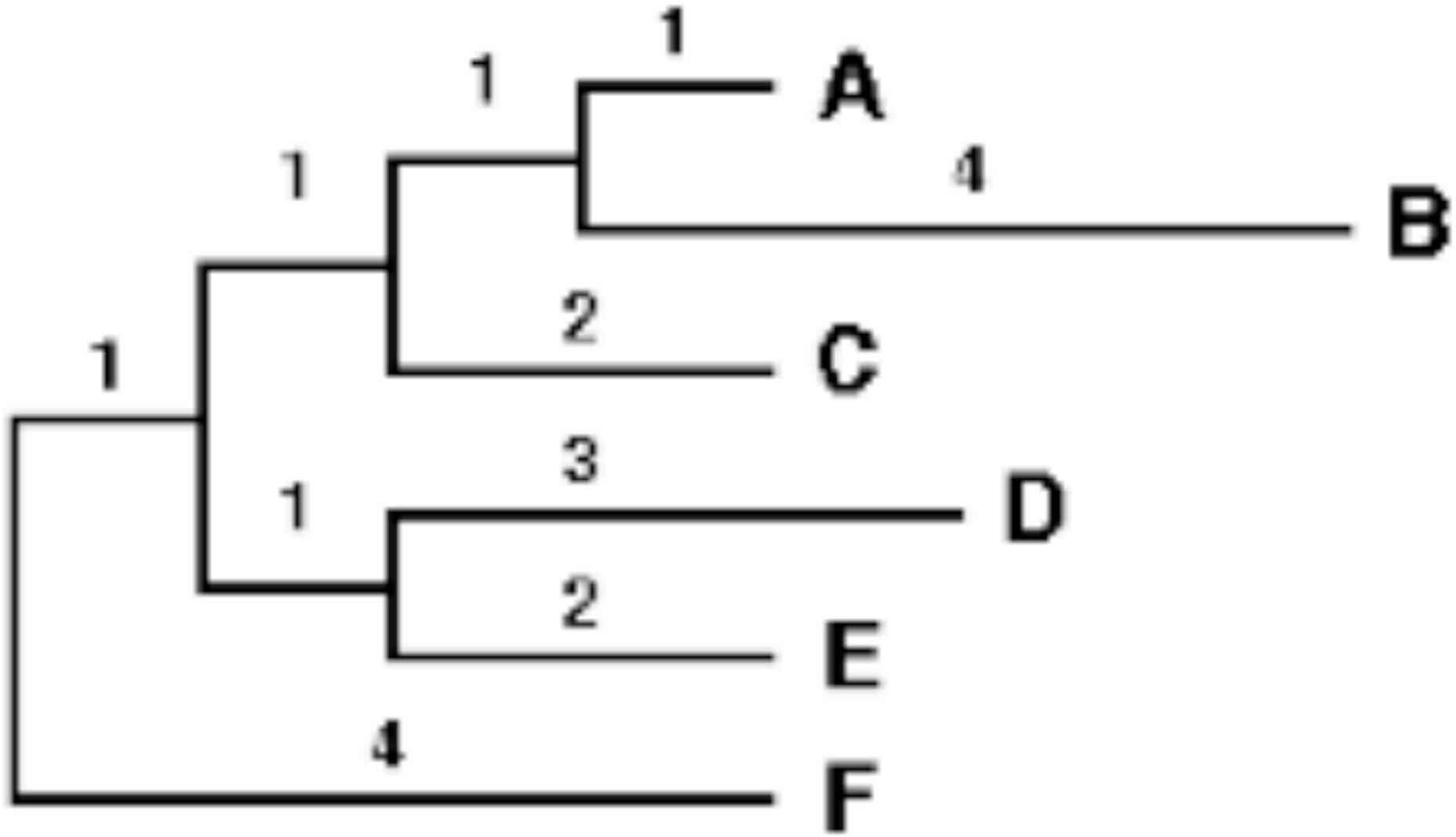




But …

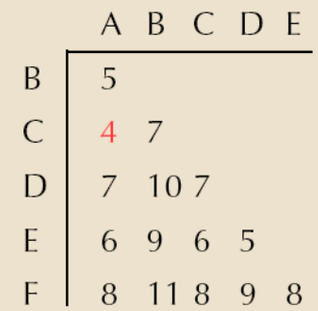
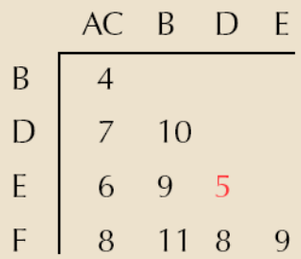
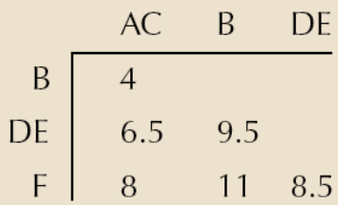
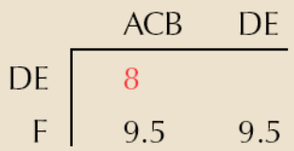
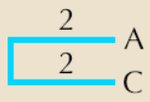
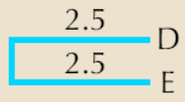
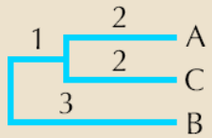
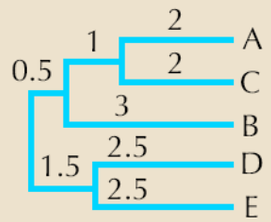
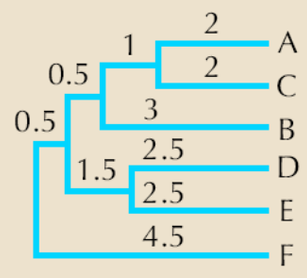
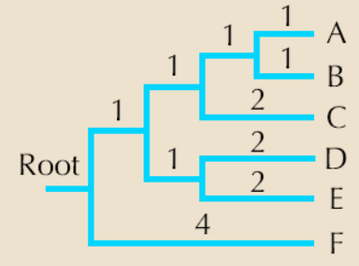
- What is evolutionary rates not equal

Unequal rates

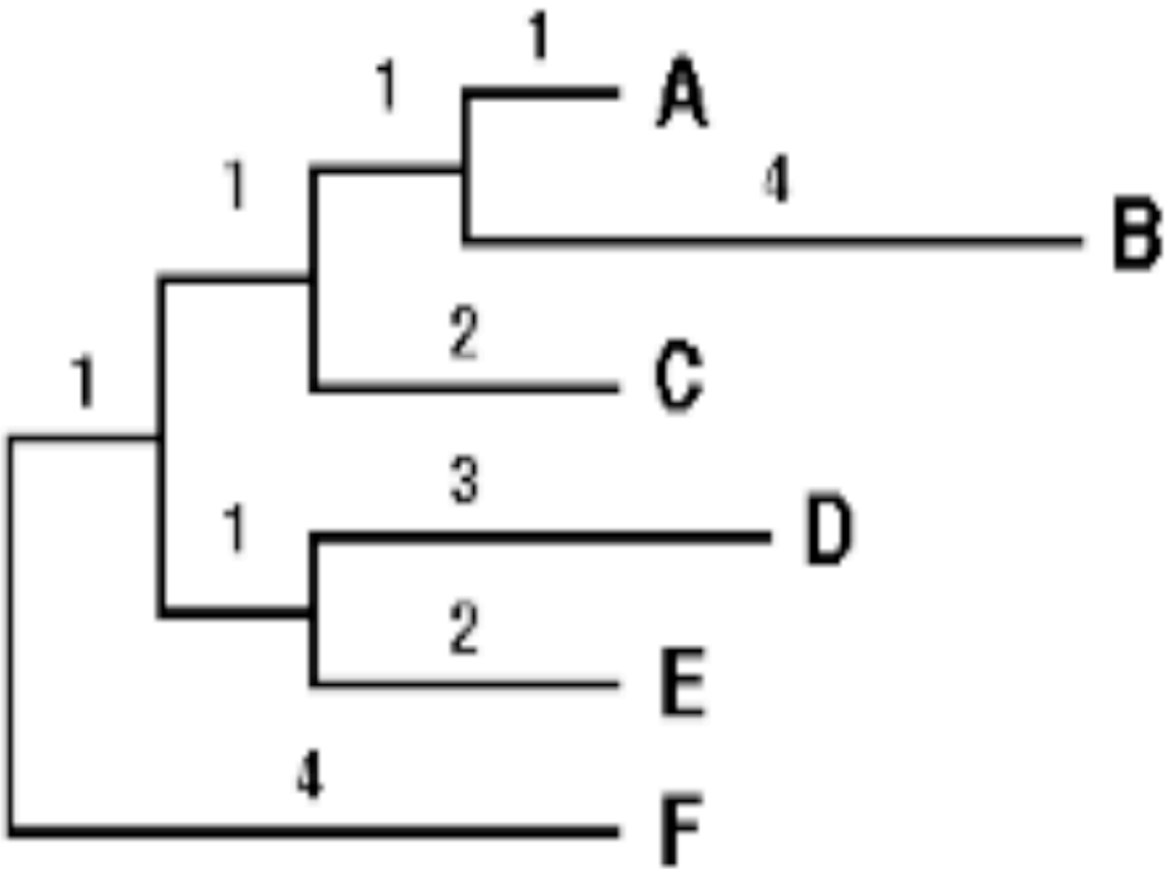


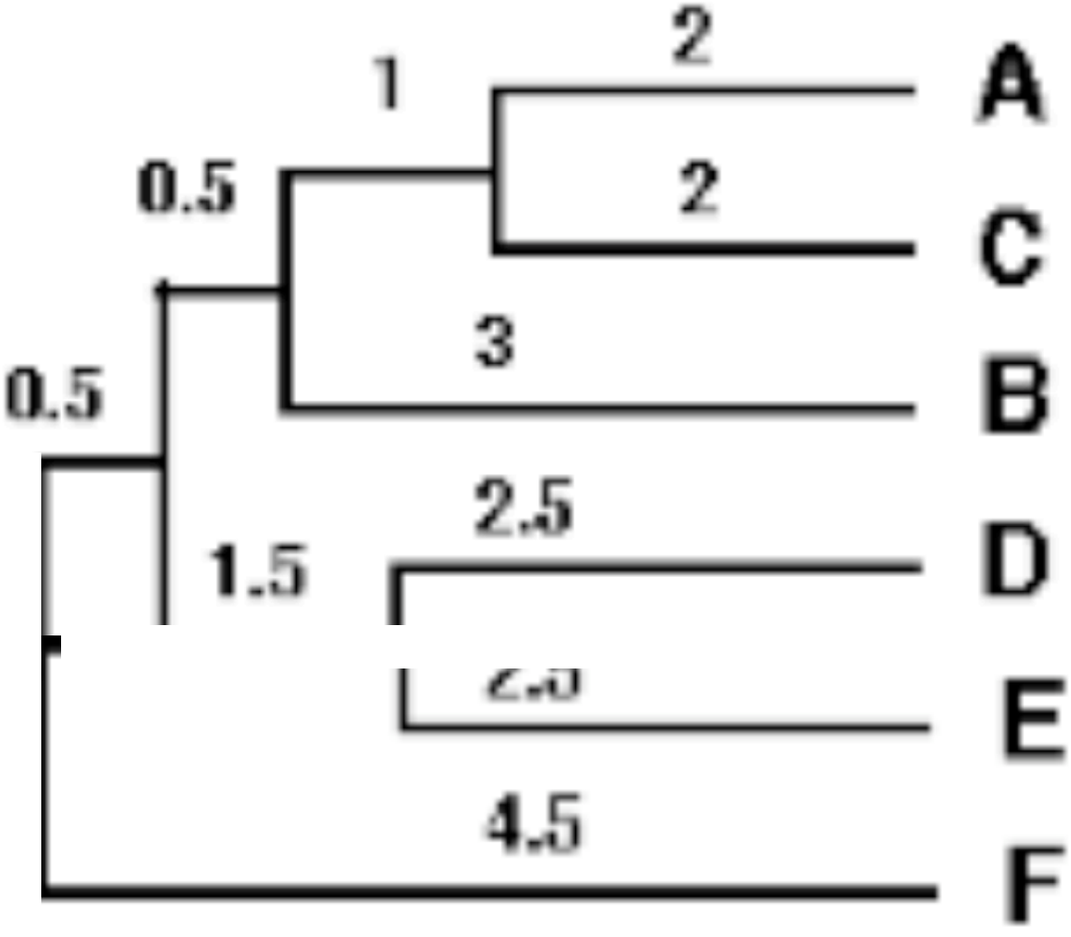
UPGMA with Unequal rates

TABLE 27.10. UPGMA tree construction errors

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Distance matrix						No new matrix
Identify smallest D	A↔C = 4	D↔E = 5	AC↔B = 4	ACB↔DF = 8	ABCDE↔F = 9	
Taxa joined	A and C	D and E	A and C with B	ABC with DE	ABCDE with F	
Subtree						
Comments					Note how this is not the same as the starting tree.	

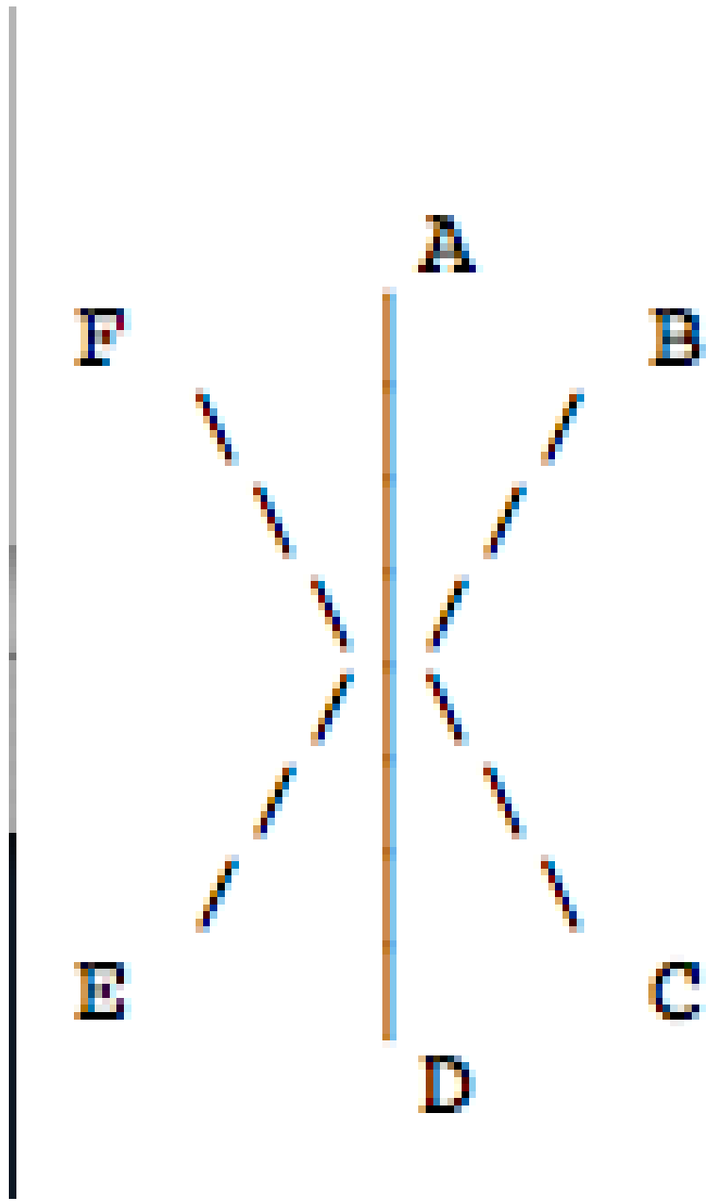
Compare to True Tree





Neighbor Joining

Start with Star Tree



Calculate S

- For each OTU, calculate a measure (S) as follows.
- S is the sum of the distances (D) between that OTU and every other OTU, divided by N-2 where N is the total number of OTUs.
- This is a measure of the distance an OTU is from all other OTUs

Calculate Pairwise Distances

- Calculate the distance D_{ij} between each OTU pair (e.g., I and J).

Identify Closest Pair

- Identify the pair of OTUs with the minimum value of $D_{ij} - S_i - S_j$

Joining Taxa

- As in UPGMA, join these two taxa at a node in a subtree.

Calculate Branches

- Calculate branch lengths. Unlike UPGMA, neighbor joining does not force the branch lengths from node X to I (D_{xi}) and to J (D_{xj}) to be equal, i.e., does not force the rate of change in those branches to be equal. Instead, these distances are calculated according to the following formulas
- $D_{xi} = 1/2\ D_{ij} + 1/2\ (S_i - S_j)$
- $D_{xj} = 1/2\ D_{ij} + 1/2\ (S_j - S_i)$

Calculate New Matrix

- Calculate a new distance matrix with I and J merged and replaced by the node (X) that joins them. Calculate the distances from this node to the other tips (K) by:

- $D_{xk} = (D_{ik} + D_{jk} - D_{ij})/2$

Distance Matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

S

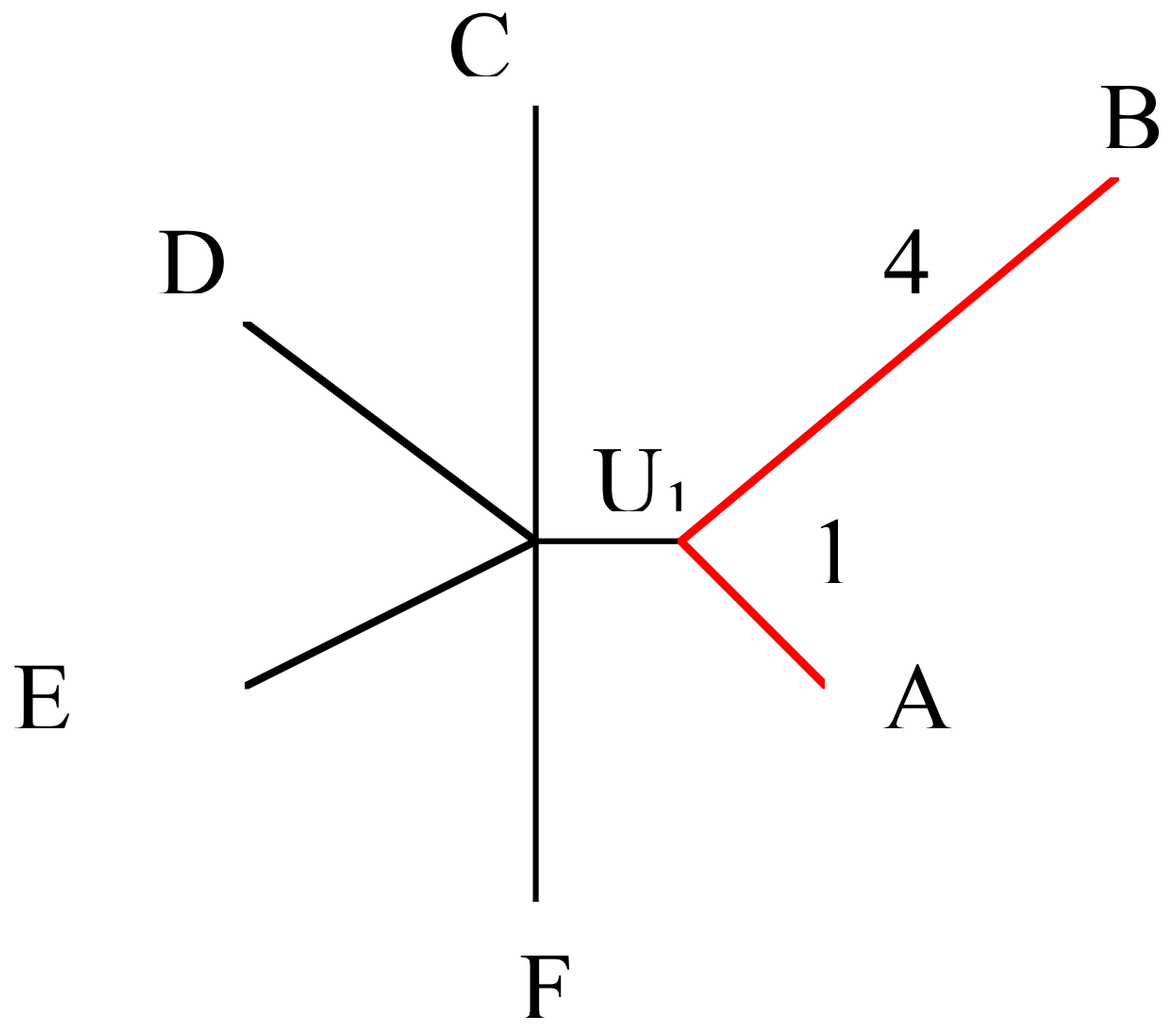
- $S_A = (5+4+7+6+8) / 4 =$ 7.5
- $S_B = (5+7+10+9+11) / 4 =$ 10.5
- $S_C = (4+7+7+6+8) / 4 =$ 8
- $S_D = (7+10+7+5+9) / 4 =$ 9.5
- $S_E = (6+9+6+5+8) / 4 =$ 8.5
- $S_F = (8+11+8+9+8) / 4 =$ 11

M

- $M_{ij} = D_{ij} - S_i - S_j$
- Smallest are
- $M_{AB} = 5 - 7.5 - 10.5 = -13$
- $M_{DE} = 5 - 9.5 - 8.5 = -13$
- Choose one of these (AB here)

New Tree

- Create a node (U) that joins pair with lowest M_{ij} such that
- $S_{IU} = D_{IJ}/2 + (S_I - S_J) / 2$
- Merge U with other taxa
- Join I and J according to S above and make all other taxa in form of a star


C
B
D
4
U_1
1
E
A
F

New Matrix

- $D_{XU} = D_{IX} + D_{JX} - D_{IJ}$ where I and J are those selected from above.

TABLE 27.11. Neighbor-joining example

Cycle 1

Distance matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Step 1

S calculations

S_x = (sum all D_x)/(N – 2), where N is the # of OTUs in the set.

$S_A = (5+4+7+6+8)/4 = 7.5$
$S_B = (5+7+10+9+11)/4 = 10.5$
$S_C = (4+7+7+6+8)/4 = 8$
$S_D = (7+10+7+5+9)/4 = 9.5$
$S_E = (6+9+6+5+8)/4 = 8.5$
$S_F = (8+11+8+9+8)/4 = 11$

Step 2

Calculate pair with smallest (M), where $M_{ij} = D_{ij} - S_i - S_j$.

Smallest are
$M_{AB} = 5 - 7.5 - 10.5 = -13$
$M_{DE} = 5 - 9.5 - 8.5 = -13$
Choose one of these (AB here).

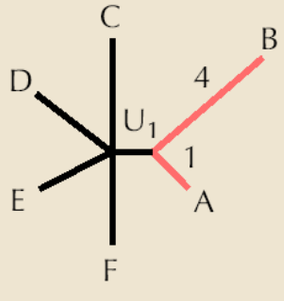
Step 3

Create a node (U) that joins pair with lowest M_{ij} such that $S_{IU} = D_{ij}/2 + (S_i - S_j)/2$.

U_1 joins A and B:
$S_{AU_1} = D_{AB}/2 + (S_A - S_B)/2 = 1$
$S_{BU_1} = D_{AB}/2 + (S_B - S_A)/2 = 4$

Step 4

Join i and j according to S above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.



Step 5

Calculate new distance matrix of all other taxa to U with $D_{xU} = D_{ix} + D_{jx} - D_{ij}$, where i and j are those selected from above.

Cycle 2

	U_1	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

Step 1

$S_{U_1} = (3+6+5+7)/3 = 7$
$S_C = (3+7+6=8)/3 = 8$
$S_D = (6+7+5+9)/3 = 9$
$S_E = (5+6+5+8)/3 = 8$
$S_F = (7+8+9+8)/3 = 10.6$

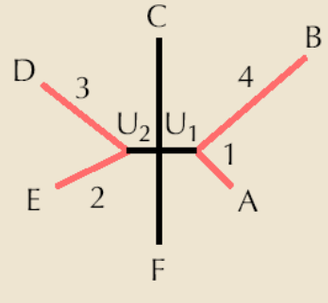
Step 2

Smallest is
$M_{CU_1} = 3 - 7 - 8 = -12$
$M_{DE} = 5 - 9 - 8 = -12$
Choose one of these (DE here).

Step 3

U_2 joins D and E:
$S_{DU_2} = D_{DE}/2 + (S_D - S_E)2 = 3$
$S_{EU_2} = D_{DE}/2 + (S_E - S_D)/2 = 2$

Step 4



Step 5

Cycle 3

	U_1	C	U_2
C	3		
U_2	3	4	
F	7	8	6

Step 1

$S_{U_1} = (3+3+7)/2 = 6.5$
$S_C = (3+4+8)/2 = 7.5$
$S_{U_2} = (3+4+6)/2 = 6.5$
$S_F = (7+8+6)/2 = 10.5$

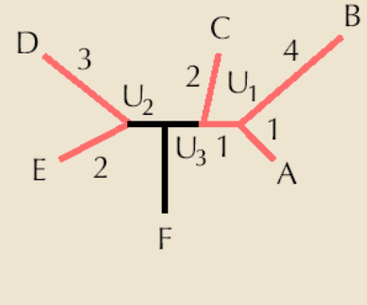
Step 2

Smallest is
$M_{CU_1} = 3 - 6.5 - 7.5 = -11$

Step 3

U_3 joins C and U_1:
$S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$
$S_{U_1U_3} = D_{CU_1}/2 + (S_{U1} - S_C)/2 = 1$

Step 4



Step 5

Cycle 4

	U_2	U_3
U_3	2	
F	6	6

Step 1

$S_{U_2} = (2+6)/1 = 8$
$S_{U_3} = (2+6)/1 = 8$
$S_F = (6+6)/1 = 12$

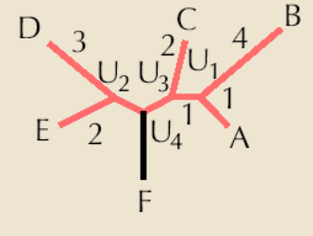
Step 2

Smallest is
$M_{U_2F} = 6 - 8 - 12 = -14$
$M_{U_3F} = 6 - 8 - 12 = -14$
$M_{U_2U_3} = 2 - 8 - 8 = -14$
Choose one of these ($M_{U_2U_3}$ here).

Step 3

U_4 joins U_2 and U_3:
$S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U2} - S_{U3})/2 = 1$
$S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U3} - S_{U2})/2 = 1$

Step 4



Step 5

Cycle 5

	U_4
F	5

Step 1

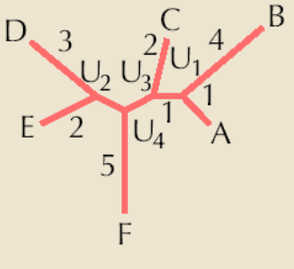
Because N – 2 = 0, we cannot do this calculation.

Step 2

Step 3

For last pair connect with branch = D.
Here $D_{U_4F} = 5$.

Step 4



Step 5

Comments

Note this is the same tree we started with (drawn in unrooted form here).

Compare to True Tree