MULTI-GENE PHYLOGENETIC ANALYSIS OF THE SUPERGROUP EXCAVATA

By

CHRISTINA CASTLEJOHN

(Under the Direction of Mark A. Farmer)

ABSTRACT

The supergroup Excavata, one of six supergroups of eukaryotes, has been a controversial supergroup within the Eukaryotic Tree of Life. Excavata was originally based largely on morphological data and to date has not been well supported by molecular studies. The goals of this research were to test the monophyly of Excavata and to observe relationships among the nine subgroups of excavates included in this study. Several different types of phylogenetic analyses were performed on a data set consisting of sequences from nine reasonably conserved genes. Analyses of this data set recovered monophyly of Excavata with moderate to strong support. Topology tests rejected all but two topologies: one with a monophyletic Excavata and one with Excavata split into two major clades. Simple gap coding, which was performed on the ribosomal DNA alignments, was found to be more useful for species-level analyses than deeper relationships with the eukaryotes.

INDEX WORDS: Excavata, excavates, monophyly, phylogenetic analysis, gap coding

MULTI-GENE PHYLOGENETIC ANALYSIS OF THE SUPERGROUP EXCAVATA

By

CHRISTINA CASTLEJOHN

B.S., Georgia Institute of Technology, 2002

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2009

MULTI-GENE PHYLOGENETIC ANALYSIS OF THE SUPERGROUP EXCAVATA

By

CHRISTINA CASTLEJOHN

Major Professor: Mark A. Farmer

Committee: James Leebens-Mack
Joseph McHugh

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2009

DEDICATION

To my family, who have supported me in my journey

ACKNOWLEDGEMENTS

I would like to thank Mark Farmer for helping me so much in my pursuit of higher education and my plans for the future. He has been a great teacher and a great friend that has helped me in so many ways throughout this endeavor. In addition, I have greatly benefitted from financial support from the National Science Foundation (NSF) PEET (Partnerships for Enhancing Expertise in Taxonomy) Program (grant no. 0329799), NSF Assembling the Tree of Life Program (grant no. 0830056), and the Department of Cellular Biology at the University of Georgia (UGA). Dr. Gertaud Burger at the University of Montreal, Dr. Jan Andersson at Uppsala University, and Bing Ma deserve my appreciation for providing sequence data to this project.

I'd especially like to thank Joe McHugh and Jim Leebens-Mack for expanding my knowledge of molecular systematics and for helping me with all those pesky details involved in running phylogenetics software. I have also benefitted greatly from the tutelage and friendship of Sarah Jardeleza, who has helped to ease my way through my research. Also, the staffs of the Department of Cellular Biology and the Research Computer Cluster at UGA have been invaluable. Finally, I would like to thank the scientific community at UGA for creating an atmosphere full of lively debate and exchanging of ideas.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Significance of the supergroup Excavata

The supergroup Excavata is one of six supergroups in the current hypothesis of the Eukaryotic Tree of Life: Amoebozoa, Archaeplastida, Chromalveolata, Excavata, Opisthokonta, and Rhizaria (Adl *et al.* 2005) (**Fig. 1.1**). The organisms in the supergroup Excavata represent an ecologically and medically important group whose phylogeny is yet to be resolved. Excavata currently consists of unicellular, flagellated protists, including the species *Trypanosoma brucei*, *T. cruzi*, *Leishmania major, Trichomonas vaginalis,* and *Giardia intestinalis* that infect and cause disease in several million people annually (CDC website 2008). It has been suggested that some organisms assigned to the Excavata may be among the earliest diverging eukaryotes (Gray *et al.* 1999, Gray *et al*. 2004, Lang *et al*. 1997). In particular, *Reclinomonas americana* has been put forward as a candidate due to the fact that its mitochondrion has retained a greater proportion of genes from the α-proteobacterial ancestral symbiont than has the mitochondrion of any other known eukaryote (Gray *et al.* 1999, Gray *et al*. 2004, Lang *et al*. 1997). Studies of this supergroup could reveal the nearest relation to the root of the eukaryotic tree of life.

Formation of the supergroup Excavata

The "excavates" are a poorly resolved supergroup consisting of ten distinct groups (Adl *et al*. 2005; Simpson 2003). These groups, seen in **Fig. 1.2**, have been linked by a suspension feeding groove shared by most members of the groups, similarities in their flagellar apparatus, and other cytoskeletal features (**Table 1.1**, Simpson 2003). Several well-defined clades, including the Euglenozoa, a clade made up of free-living and parasitic unicellular flagellates, lack many of these morphological features but have been linked to the core excavates primarily through molecular studies (Hampl *et al.* 2005, Simpson 2003).

Previous phylogenetic and phylogenomic studies on the excavates

Molecular phylogenetic and phylogenomic analyses have been performed to resolve this supergroup, but monophyly has been recovered rarely and only with poor support values or with in-group taxa removal (Hampl *et al.* 2008, Simpson *et al*. 2006, Simpson *et al*. 2008). Simpson *et al.* (2006), using multigene analyses of six genes, recovered a paraphyletic Excavata comprised of three separate clades even though each other supergroup was recovered as monophyletic (**Fig. 1.3**). The three groups of excavates recovered are as follows: **Group 1)** Diplomonadida, *Carpediemonas*, and Parabasalia; **Group 2)** *Trimastix*, and Oxymonadida; **Group 3)** Euglenozoa, Jakobida, and Heterolobosea. Malawimonads are not consistently recovered with any of the three groups. The analyses performed had problems converging on one topology within each analysis.

In Simpson *et al.* (2006), the authors suggested performing covarion analyses on the data set to try to account for temporal heterogeneity across the alignment, or heterotachy, due to the broad taxonomic sampling. These analyses were performed on their original alignment and were found to have little effect on the resulting trees (Castlejohn *et al.* 2008). Four different topologies were recovered due to problems with convergence, but the same three groups of excavates were recovered in each analysis without the Excavata being recovered as monophyletic (**Fig. 1.4**). After determining that one of the genes used in the original alignment, α-tubulin, showed evidence of horizontal gene transfer, Simpson *et al.* (2008) found that two of these excavate groups were recovered as sister to each other (**Fig. 1.5**). However, the Group 3 excavates were still not closely related to the other groups of excavates.

Hampl *et al.* (2009) performed a phylogenomic study of the excavates in which they used 143 protein-coding genes in taxa throughout the eukaryotic tree of life. With this data set, the majority of the excavates were recovered together with the exception of the malawimonads (**Fig. 1.6**). Then, they systematically removed taxa on long branches to eliminate the effects of long branch attraction, which occurs when taxa on long branches tend to cluster together. They recovered monophyly of the excavates with a bootstrap support of 90, but only through removal of several in-group taxa including all of the Group 1 and Group 2 excavates and several Group 3 excavates (**Fig. 1.7**). Finally, they removed 1,750

long-branching gene sequences rather than long-branching taxa and recovered monophyly of the excavates with a low bootstrap support value of 54 and with low support values at the deeper nodes of the excavate groups (**Fig. 1.8**).

General outline and objectives

Because monophyly has rarely been recovered (and then only with poor support values) for the excavates based on molecular data, it is possible that the morphological characters that were used to infer the monophyly of the excavates are not synapomorphies (shared derived characteristics) of the supergroup but rather pleisiomorphies (ancestral characteristics) of basal eukaryotes. The goals of this research are to determine if the supergroup Excavata is monophyletic and to determine relationships among the ten groups of excavates. We predict that the three groups of excavates recovered in previous analyses will be recovered in these analyses. The relationships between these three groups could possibly be any of the following:

1) The three groups could be recovered as a monophyletic clade.
2) The three groups could be recovered independently of each other.
3) Two of the three groups could be recovered as monophyletic with the third group being independent of this monophyletic clade, possibly due to long branch attraction.

We will test these hypotheses by performing a multi-gene phylogenetic analysis consisting of nine genes, made up of nuclear-encoded protein-coding genes and DNA from ribosomal subunits (**Table 1.2**). Also, morphological data from the coding of gaps in the ribosomal DNA alignments will be included in this analysis.

Table 1.1. Distribution of distinctive excavate features in the ten established groups of Excavata

Taxa: 1, Jakobida; 2, *Malawimonas*; 3, *Trimastix*; 4, *Carpediemonas*; 5, Retortamonadida; 6, Diplomonadida; 7, Heterolobosea; 8, Oxymonadida; 9. Parabasalia; 10, Euglenozoa. +, Presence of feature; ?, arguable homology; ND, no appropriate data. From Simpson (2003).

Feature	1	2	3	4	5	6	7	8	9	10
Feeding groove	+	+	+	+	+	+	+			
I fibre	+	+	+	+	+	+	+	+		
B fibre	+	+	+	+	+			+		
C fibre	+	+	+	+	+			+	?	
Split R1	+	+	+	+	+	+	+			
Singlet root	+	+	+	+	+	?		+		
Flagellar vanes	+	+	+	+	+					
Composite fibre	+	ND	+	+	+					

Table 1.2. Proposed Phylogenetic Markers For This Study

Btub = Beta tubulin; **EF1α** = Elongation Factor 1 alpha; **EF2** = Elongation Factor 2; **cHSP70** = cytosolic Heat Shock Protein 70; **cHSP90** = cytosolic Heat Shock Protein 90; **SSU** = Small Subunit Ribosomal DNA; **LSU** = Large Subunit Ribosomal DNA; **Cal-1** = Calmodulin 1; **NE** = Nuclear-encoded; **PC** = Protein-coding; **Nucl & AA** = Nucleotide and Amino Acid sequences can be used; **LS** = Large Size; **LER** = Low Evolutionary Rate; **SC** = Essentially single copy; **SS** = Secondary structure information can be used

Gene	Actin	Btub	EF1a	EF2	cHSP70	cHSP90	SSU (18S)	LSU (28S)	Cal-1
Function	Component of microfilaments / cytoskeleton	Component of microtubules / cytoskeleton	Binding of AA-tRNA to ribosomes during translation	Polypeptide chain elongation in protein synthesis	Molecular Chaperone that facilitates protein folding	Molecular Chaperone that facilitates protein folding	Protein synthesis	Protein synthesis	Ca++-binding protein
Size	~1100bp	~1300bp	~1300bp	~2600bp	~2000bp	2100-2400bp	1800-2200bp	>4000bp	450bp
Pros	NE, PC, Nucl & AA, Widely used phylogenetic marker	NE, PC, Nucl & AA, Widely used phylogenetic marker	NE, PC, Nucl & AA, Previously used as phylogenetic marker, Easily distinguishable from other closely-related genes	NE, PC, Nucl & AA, Previously used as phylogenetic marker, Easily distinguishable from other closely-related genes	NE, PC, Nucl & AA, Widely used phylogenetic marker	NE, PC, Nucl & AA, Widely used phylogenetic marker	LS, LER, SC, SS, Widely used phylogenetic marker	LS, LER, SC, SS, Different regions can be used for different questions	NE, PC, Nucl & AA, Previously used as phylogenetic marker
Cons	Not single copy in all organisms	Not single copy in all organisms	Complex gene family, Replacement by EFL in some organisms	Complex gene family, Not single copy in all organisms	Complex gene family, Not just cytosolic copies	Complex gene family, Not just cytosolic copies	Long branch attraction problems	Determining which regions to use for this particular question	Not single copy in all organisms, Small in size
References	Baldauf and Palmer 1993, Bhattacharya and Weber 1997, Bhattacharya *et al.* 1998, Bhattacharya *et al.* 2000, Drouin *et al.* 1995, Hennessey *et al.* 1993, Keeling and Doolittle 1996, Simpson *et al.* 2006, Simpson *et al.* 2008		Baldauf 2003, Jardeleza 2007, Keeling and Inagaki 2004, Simpson *et al.* 2006, Simpson *et al.* 2008, Tanabe *et al.* 2004, Yamamoto *et al.* 1997		Ahner *et al.* 2005, Breglia *et al.* 2007, Schutze *et al.* 1999, Simpson *et al.* 2006, Simpson *et al.* 2008, Simpson and Roger 2004, Stechmann & Cavalier-Smith 2003		Busse & Preisfeld 2002, Busse & Preisfeld 2003, Daubin *et al.* 2003, Ma 2005		Hirata *et al.* 2007, Schutze *et al.* 1999, Varga *et al.* 2007

Figure 1.1. Eukaryotic Tree of Life. The eukaryotic tree is separated into six supergroups: Amoebozoa, Opisthokonta, Rhizaria, Archaeplastida, Chromalveolata, and Excavata. From Adl *et al.* (2005).

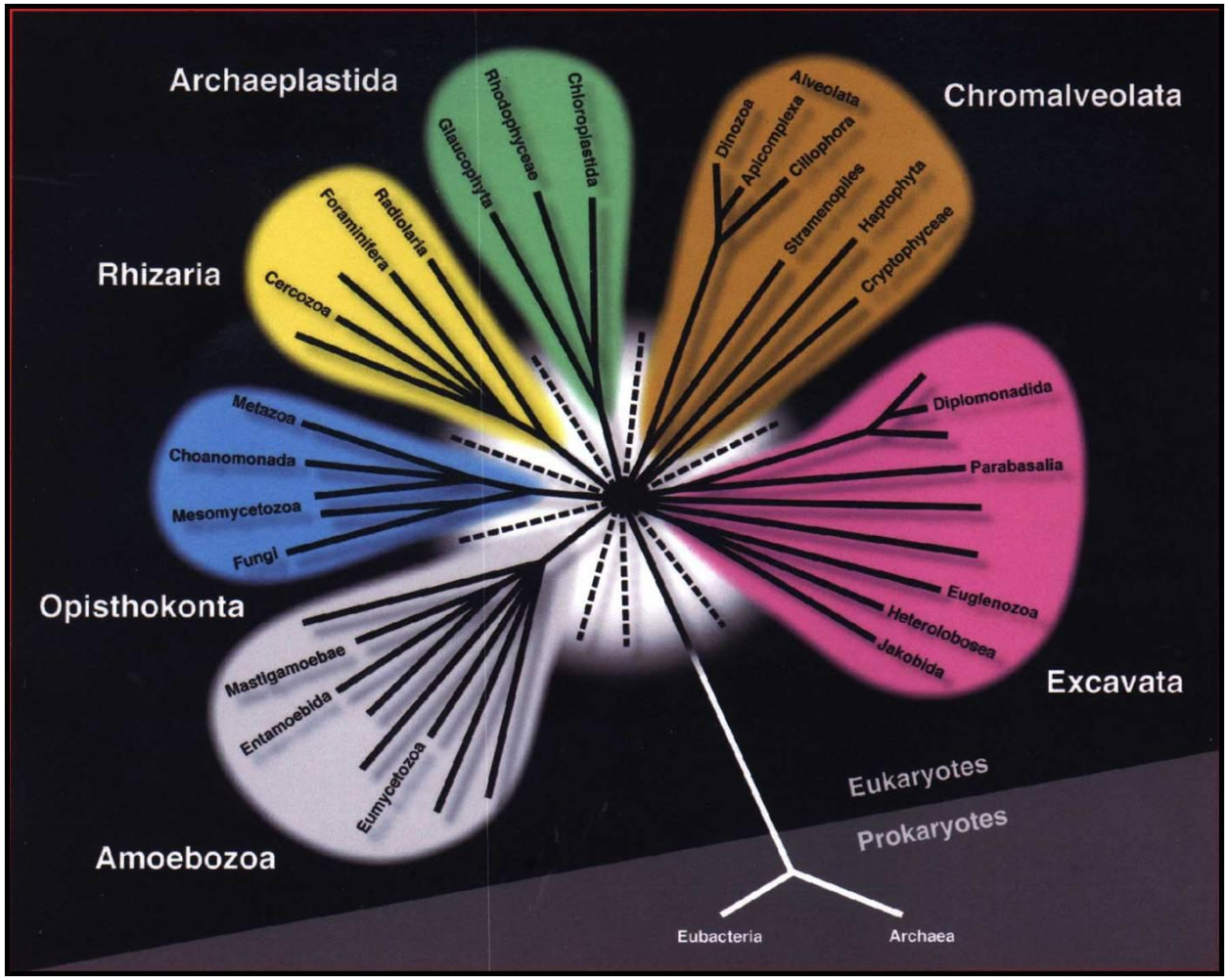

Archaeplastida
Rhizaria
Opisthokonta
Amoebozoa
Chromalveolata
Excavata
Rhodophyceae
Chloroplastida
Glaucophyta
Alveolata
Dinozoa
Apicomplexa
Ciliophora
Stramenopiles
Haptophyta
Cryptophyceae
Radiolaria
Foraminifera
Cercozoa
Metazoa
Choanomonada
Mesomycetozoa
Fungi
Diplomonadida
Parabasalia
Euglenozoa
Heterolobosea
Jakobida
Mastigamoebae
Entamoebida
Eumycetozoa
Eukaryotes
Prokaryotes
Eubacteria
Archaea

Figure 1.2. Light microscopic appearance of the ten groups of Excavata. (a) Jakobid: *Jakoba incarcerata*; (b) *Malawimonas*: *Malawimonas jakobiformis*; (c) *Trimastix*: *Trimastix pyriformis*; (d) *Carpediemonas*: *Carpediemonas membranifera*; (e) retortamonad: *Chilomastix cuspidata*; (f) diplomonad: *Trepomonas agilis*; (g) heteroloboseid: *Percolomonas descissus*; (h) oxymonad: *Monocercomonoides hausmanni*; (i) euglenozoan: *Dimastigella trypaniformis*; (j) parabasalid: *Pseudotrichomonas keilini*. Bar, 10 mm. From Simpson (2003).

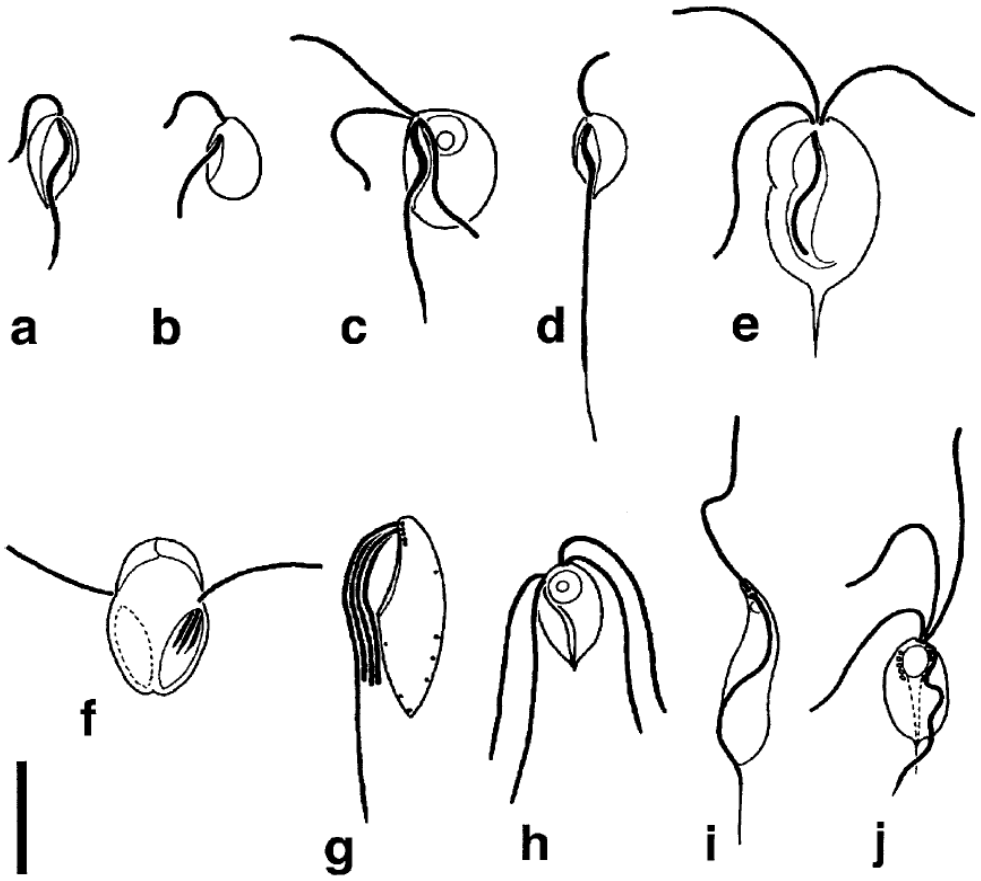

a
b
c
d
e
f
g
h
i
j

Figure 1.3. ML phylogenetic tree of eukaryotes inferred from six slowly evolving nuclear-encoded proteins. Best topology under unlinked model shown (i.e., with gene-specific branch lengths). Numbers on branches represent ML bootstrap support values for the unlinked model (upper numbers) and linked model (lower numbers). Filled circles represent bipartitions receiving .95% support with both methods. Dashes represent values, 50% not critical to the study. Excavates are identified by gray shading. ‘‘1’’ ‘‘2,’’ and ‘‘3’’ indicate well-supported excavate groups. Note that *Malawimonas* is uncertainly placed. ‘‘X’’ and ‘‘Y’’ denote better-supported bipartitions that separate Group 1 from other excavates. From Simpson *et al.* (2006).

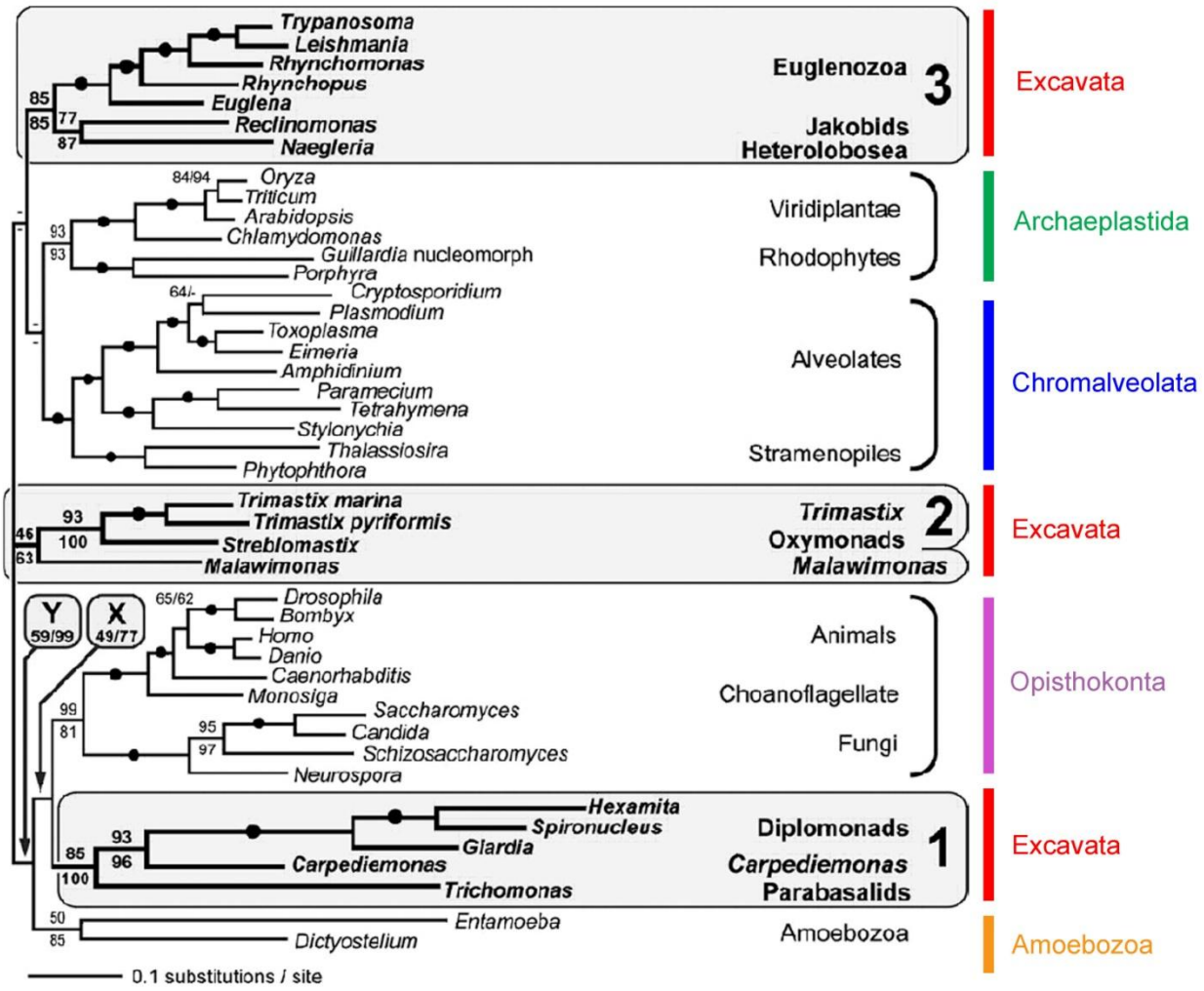

Trypanosoma
Leishmania
Rhynchomonas
Rhynchopus
Euglena
Reclinomonas
Naegleria
85
85
77
87
Euglenozoa
Jakobids
Heterolobosea
3
Excavata
84/94
Oryza
Triticum
Arabidopsis
93
93
Chlamydomonas
Guillardia nucleomorph
Porphyra
Viridiplantae
Rhodophytes
Archaeplastida
64/-
Cryptosporidium
Plasmodium
Toxoplasma
Eimeria
Amphidinium
Paramecium
Tetrahymena
Stylonychia
Thalassiosira
Phytophthora
Alveolates
Stramenopiles
Chromalveolata
Trimastix marina
93
Trimastix pyriformis
46
100
Streblomastix
63
Malawimonas
Trimastix
Oxymonads
Malawimonas
2
Excavata
Y
59/99
X
49/77
65/62
Drosophila
Bombyx
Homo
Danio
Caenorhabditis
Monosiga
99
81
Saccharomyces
95
Candida
97
Schizosaccharomyces
Neurospora
Animals
Choanoflagellate
Fungi
Opisthokonta
Hexamita
Spironucleus
93
Giardia
85
96
Carpediemonas
100
Trichomonas
Diplomonads
Carpediemonas
Parabasalids
1
Excavata
50
Entamoeba
85
Dictyostelium
Amoebozoa
Amoebozoa
0.1 substitutions / site

Figure 1.4. Topologies of Bayesian Analyses of 6-gene Data Set. A) Topology 1 is shown in solid lines. Differences between Topology 1 and Topology 2 are indicated by dotted lines. B) Topology 3 is shown in solid lines. Differences between Topology 3 and Topology 4 are indicated by dotted lines. Bayesian posterior probabilities (PP) are color coded by analysis. Nodes with no values have a PP of 1.00. The large red lines labeled with numbers indicate the excavate groups recovered by Simpson *et al.* (2006). Group 1: Diplomonadida, *Carpediemonas*, and Parabasalia; Group 2: *Trimastix*, and Oxymonadida; Group 3: Euglenozoa, Jakobida, and Heterolobosea. From Castlejohn *et al.* (2008).

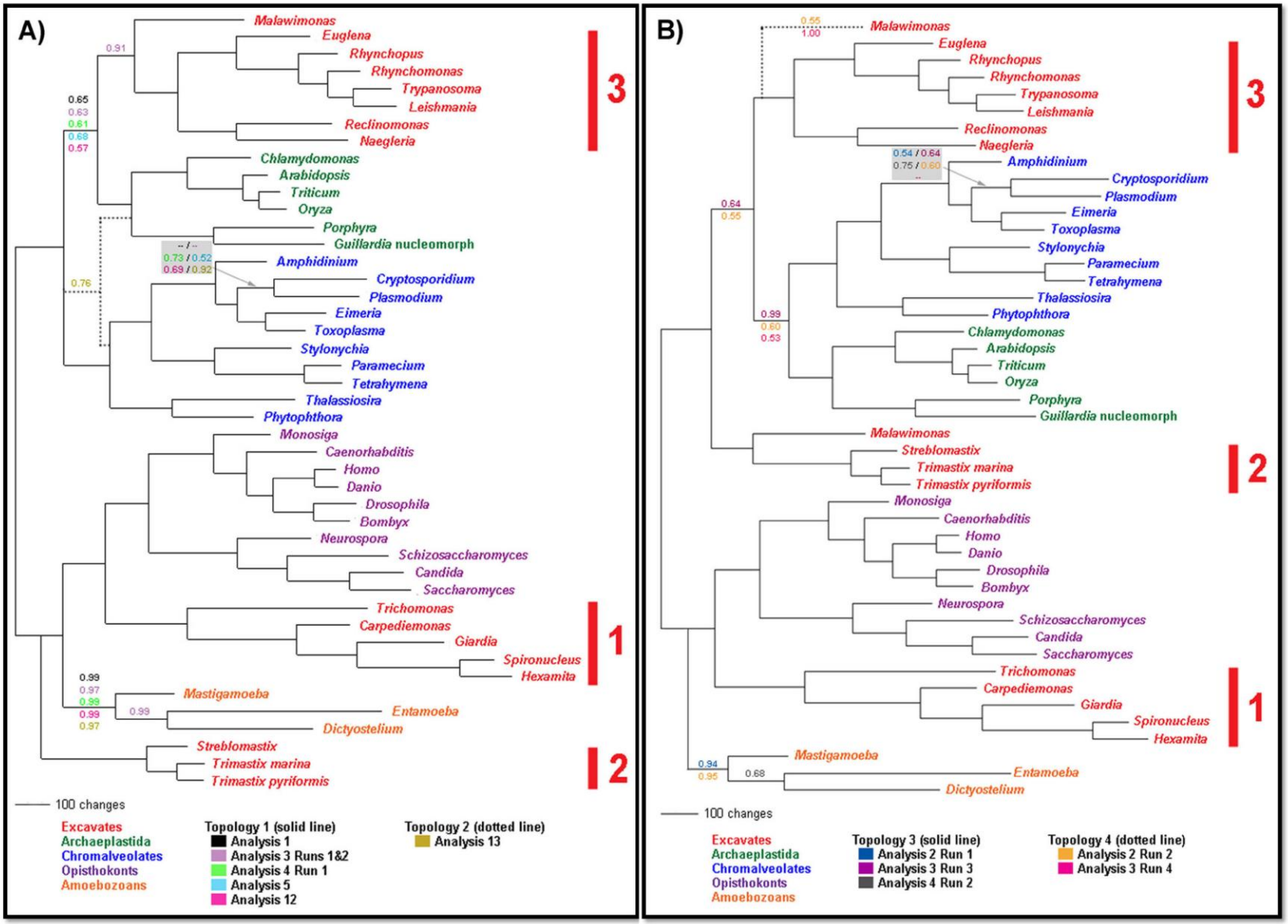

A)
Malawimonas
Euglena
Rhynchopus
Rhynchomonas
Trypanosoma
Leishmania
Reclinomonas
Naegleria
Chlamydomonas
Arabidopsis
Triticum
Oryza
Porphyra
Guillardia nucleomorph
Amphidinium
Cryptosporidium
Plasmodium
Eimeria
Toxoplasma
Stylonychia
Paramecium
Tetrahymena
Thalassiosira
Phytophthora
Monosiga
Caenorhabditis
Homo
Danio
Drosophila
Bombyx
Neurospora
Schizosaccharomyces
Candida
Saccharomyces
Trichomonas
Carpediemonas
Giardia
Spironucleus
Hexamita
Mastigamoeba
Entamoeba
Dictyostelium
Streblomastix
Trimastix marina
Trimastix pyriformis
3
1
2
100 changes
Excavates
Archaeplastida
Chromalveolates
Opisthokonts
Amoebozoans
Topology 1 (solid line)
Analysis 1
Analysis 3 Runs 1&2
Analysis 4 Run 1
Analysis 5
Analysis 12
Topology 2 (dotted line)
Analysis 13
B)
Malawimonas
Euglena
Rhynchopus
Rhynchomonas
Trypanosoma
Leishmania
Reclinomonas
Naegleria
Amphidinium
Cryptosporidium
Plasmodium
Eimeria
Toxoplasma
Stylonychia
Paramecium
Tetrahymena
Thalassiosira
Phytophthora
Chlamydomonas
Arabidopsis
Triticum
Oryza
Porphyra
Guillardia nucleomorph
Malawimonas
Streblomastix
Trimastix marina
Trimastix pyriformis
Monosiga
Caenorhabditis
Homo
Danio
Drosophila
Bombyx
Neurospora
Schizosaccharomyces
Candida
Saccharomyces
Trichomonas
Carpediemonas
Giardia
Spironucleus
Hexamita
Mastigamoeba
Entamoeba
Dictyostelium
3
2
1
100 changes
Excavates
Archaeplastida
Chromalveolates
Opisthokonts
Amoebozoans
Topology 3 (solid line)
Analysis 2 Run 1
Analysis 3 Run 3
Analysis 4 Run 2
Topology 4 (dotted line)
Analysis 2 Run 2
Analysis 3 Run 4

Figure 1.5. Maximum likelihood tree for the six-protein dataset (i.e., with a-tubulin excluded) under the concatenated ('super-gene') model (WAG+Γ). Percentage bootstrap support values are depicted. Single numbers represent values estimated using PHYML (WAG+Γ, 500 replicates). Most values <40% are not shown. For important nodes, double numbers are shown giving the PHYML values followed by values estimated using RAxML (WAG+CAT, 200 replicates). There were no important differences between the ML trees for the concatenated and unlinked models. From Simpson *et al.* (2008).

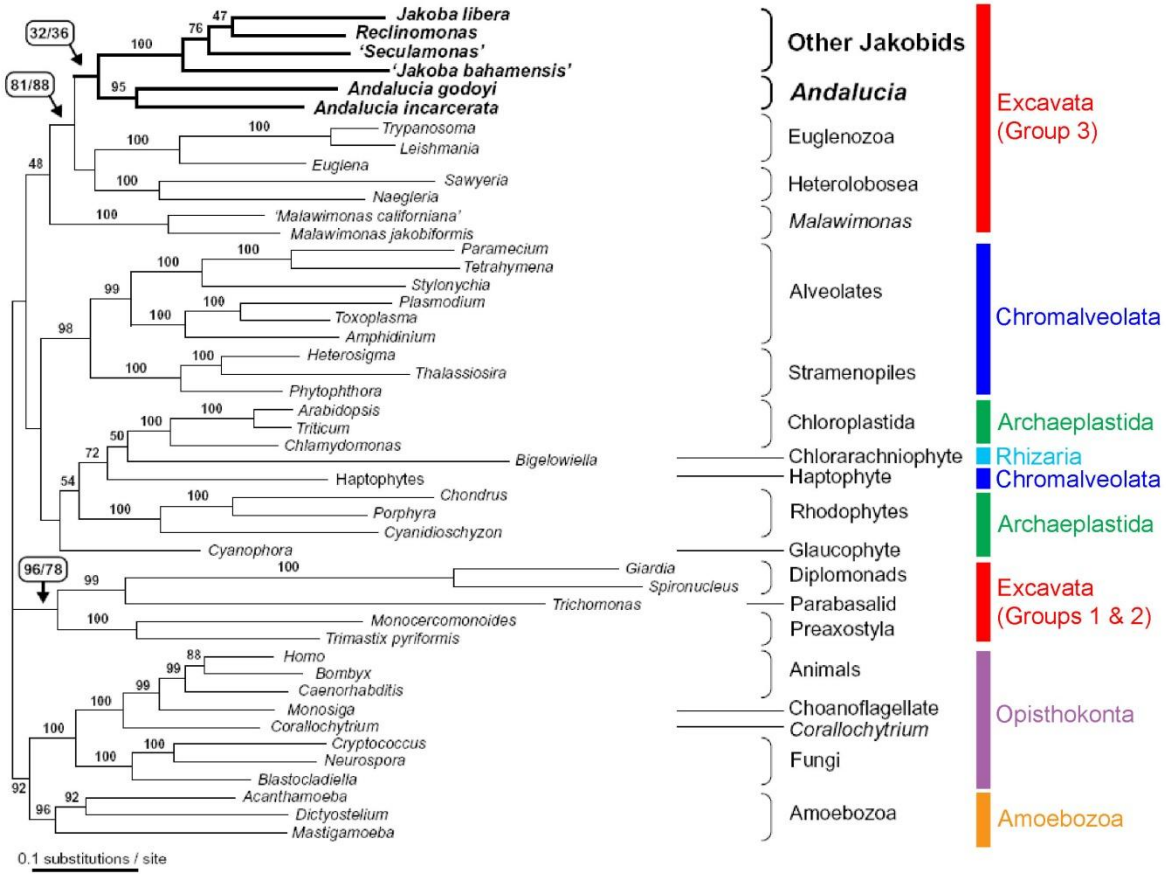

Jakoba libera
Reclinomonas
'Seculamonas'
'Jakoba bahamensis'
Andalucia godoyi
Andalucia incarcerata
Trypanosoma
Leishmania
Euglena
Sawyeria
Naegleria
'Malawimonas californiana'
Malawimonas jakobiformis
Paramecium
Tetrahymena
Stylonychia
Plasmodium
Toxoplasma
Amphidinium
Heterosigma
Thalassiosira
Phytophthora
Arabidopsis
Triticum
Chlamydomonas
Bigelowiella
Haptophytes
Chondrus
Porphyra
Cyanidioschyzon
Cyanophora
Giardia
Spironucleus
Trichomonas
Monocercomonoides
Trimastix pyriformis
Homo
Bombyx
Caenorhabditis
Monosiga
Corallochytrium
Cryptococcus
Neurospora
Blastocladiella
Acanthamoeba
Dictyostelium
Mastigamoeba
32/36
81/88
96/78
0.1 substitutions / site
Other Jakobids
Andalucia
Euglenozoa
Heterolobosea
Malawimonas
Alveolates
Stramenopiles
Chloroplastida
Chlorarachniophyte
Haptophyte
Rhodophytes
Glaucophyte
Diplomonads
Parabasalid
Preaxostyla
Animals
Choanoflagellate
Corallochytrium
Fungi
Amoebozoa
Excavata (Group 3)
Chromalveolata
Archaeplastida
Rhizaria
Chromalveolata
Archaeplastida
Excavata (Groups 1 & 2)
Opisthokonta
Amoebozoa

Figure 1.6. The phylogenetic tree estimated from the main 143 protein-coding gene dataset. This topology received the highest likelihood in the search of unconstrained nodes using the WAG+Γ model; branch-lengths were calculated in RAxML using the WAG+Γ model. The representatives of the 6 supergroups are color-coded. Asterisks indicate the nodes that were not constrained during the exhaustive search. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping/PhyloBayes posterior probability. At nodes that were not constrained during the exhaustive search in the separate analysis (asterisks), the third number indicates the RELL bootstrap value. Branches that received maximum possible support by all methods are indicated by full circles. Dashes indicate bootstrap values <50%, or posterior probabilities <0.5. From Hampl *et al.* (2009).

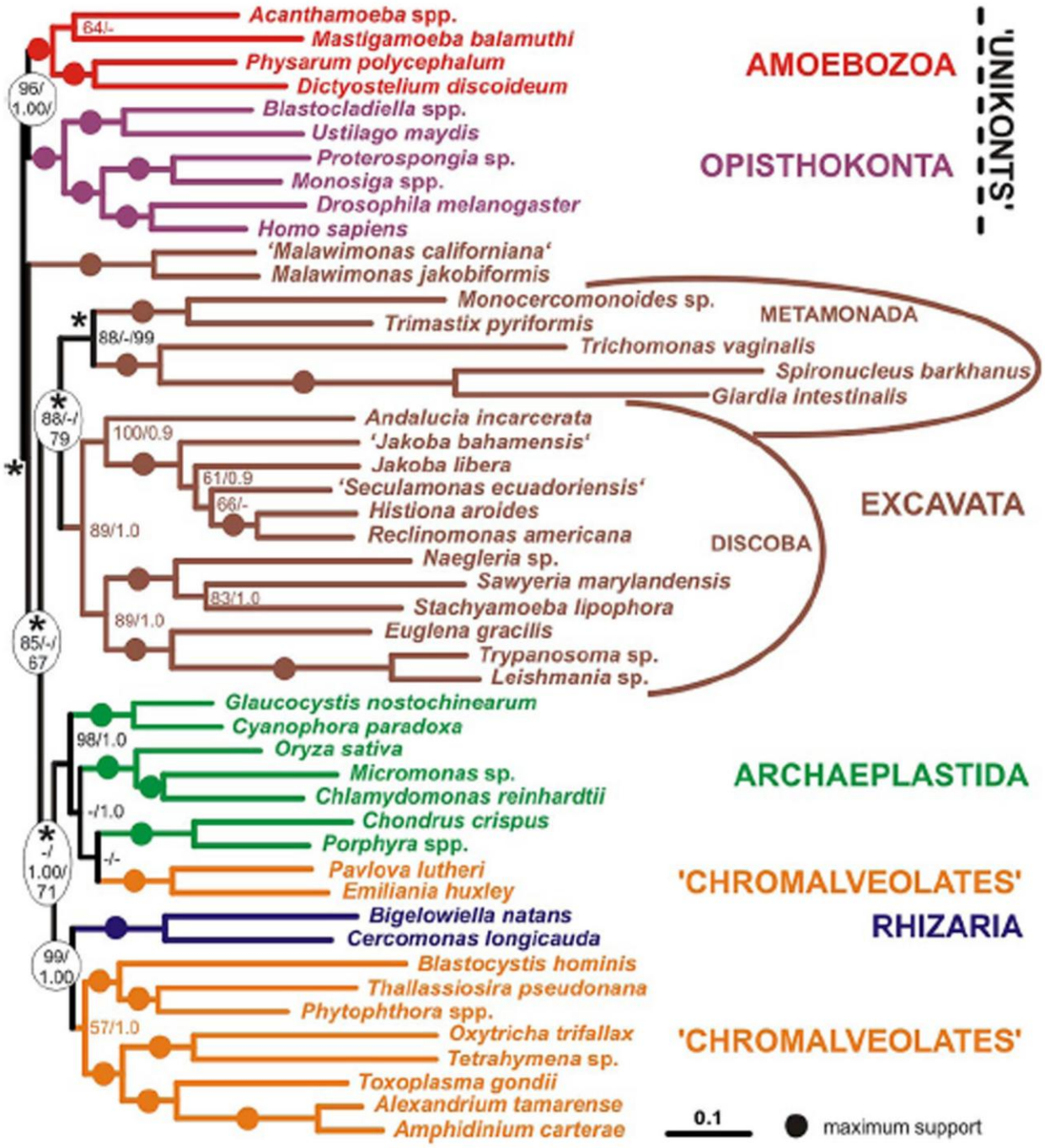

Acanthamoeba spp.
Mastigamoeba balamuthi
Physarum polycephalum
Dictyostelium discoideum
Blastocladiella spp.
Ustilago maydis
Proterospongia sp.
Monosiga spp.
Drosophila melanogaster
Homo sapiens
'Malawimonas californiana'
Malawimonas jakobiformis
Monocercomonoides sp.
Trimastix pyriformis
Trichomonas vaginalis
Spironucleus barkhanus
Giardia intestinalis
Andalucia incarcerata
'Jakoba bahamensis'
Jakoba libera
'Seculamonas ecuadoriensis'
Histiona aroides
Reclinomonas americana
Naegleria sp.
Sawyeria marylandensis
Stachyamoeba lipophora
Euglena gracilis
Trypanosoma sp.
Leishmania sp.
Glaucocystis nostochinearum
Cyanophora paradoxa
Oryza sativa
Micromonas sp.
Chlamydomonas reinhardtii
Chondrus crispus
Porphyra spp.
Pavlova lutheri
Emiliania huxley
Bigelowiella natans
Cercomonas longicauda
Blastocystis hominis
Thallassiosira pseudonana
Phytophthora spp.
Oxytricha trifallax
Tetrahymena sp.
Toxoplasma gondii
Alexandrium tamarense
Amphidinium carterae
64/-
96/
1.00/
*
88/-/99
*
88/-/
79
100/0.9
61/0.9
66/-
89/1.0
89/1.0
*
85/-/
67
83/1.0
98/1.0
-/1.0
*
-/
1.00/
71
-/-
99/
1.00
57/1.0
AMOEBOZOA
OPISTHOKONTA
'UNIKONTS'
METAMONADA
EXCAVATA
DISCOBA
ARCHAEPLASTIDA
'CHROMALVEOLATES'
RHIZARIA
'CHROMALVEOLATES'
0.1
maximum support

Figure 1.7. LB taxa removal. A maximum likelihood tree after removal of 14 long branch taxa based on branch lengths determined by TreeStat. The tree was constructed in RAxML using WAG+Γ model and colored as described in Fig. 1.6. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping. Branches that received maximum support by all methods are indicated by full circles. From Hampl *et al*. (2009).

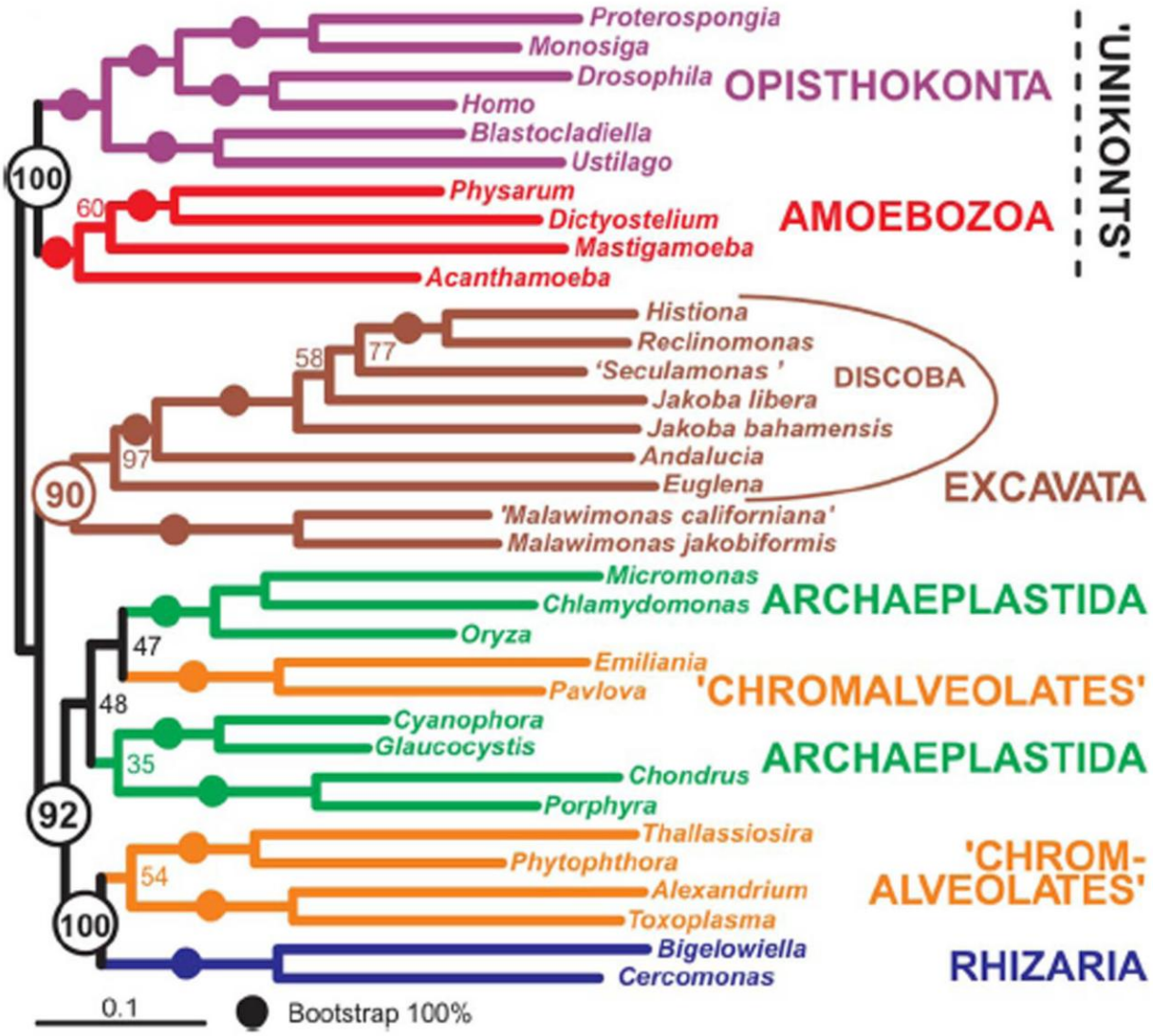
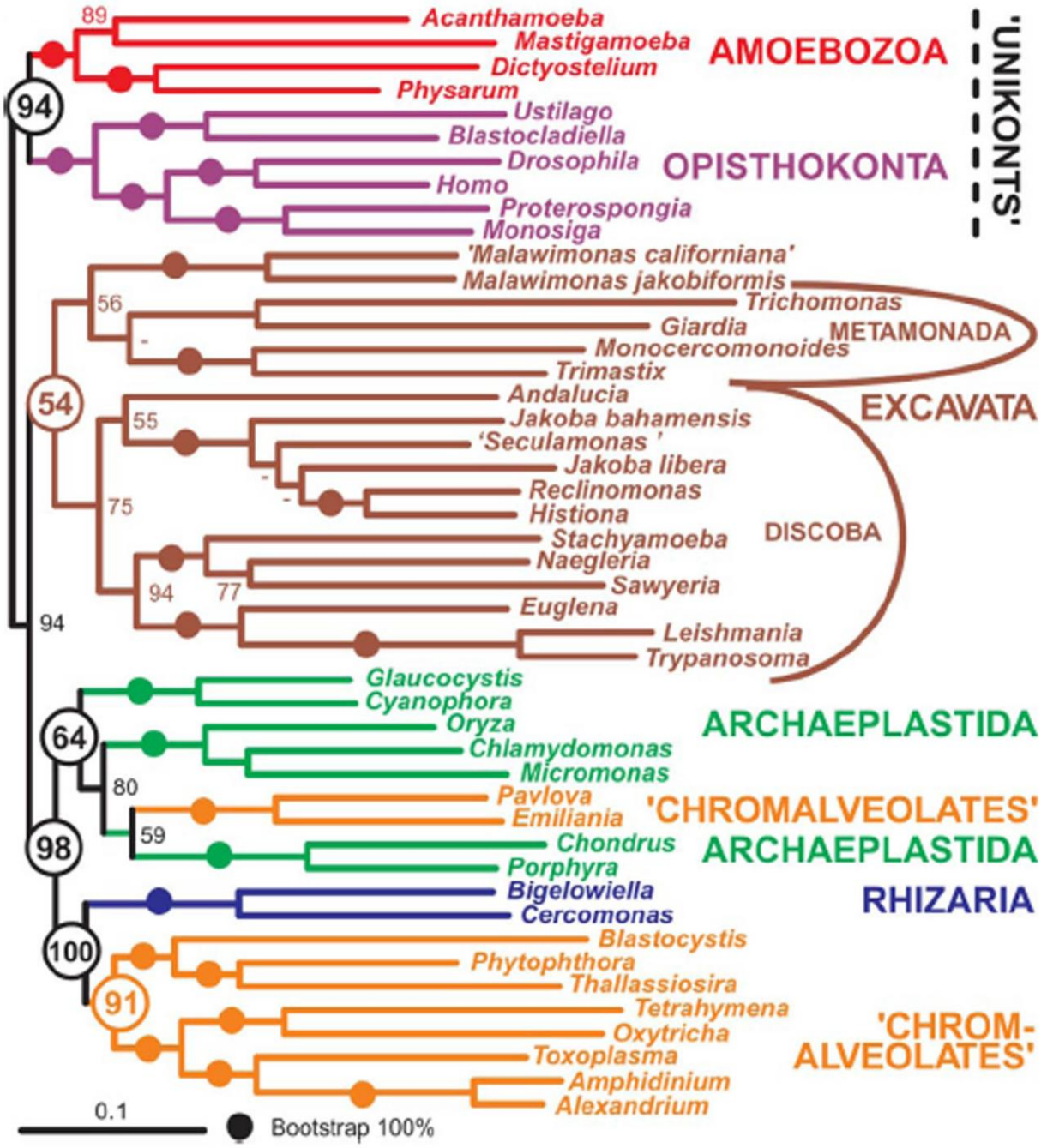

Proterospongia
Monosiga
Drosophila
Homo
Blastocladiella
Ustilago
OPISTHOKONTA
'UNIKONTS'
Physarum
Dictyostelium
Mastigamoeba
Acanthamoeba
AMOEBOZOA
Histiona
Reclinomonas
'Seculamonas '
Jakoba libera
Jakoba bahamensis
Andalucia
Euglena
DISCOBA
EXCAVATA
'Malawimonas californiana'
Malawimonas jakobiformis
Micromonas
Chlamydomonas
Oryza
ARCHAEPLASTIDA
Emiliania
Pavlova
'CHROMALVEOLATES'
Cyanophora
Glaucocystis
Chondrus
Porphyra
ARCHAEPLASTIDA
Thallassiosira
Phytophthora
Alexandrium
Toxoplasma
'CHROM-
ALVEOLATES'
Bigelowiella
Cercomonas
RHIZARIA
100
60
58
77
97
90
47
48
35
92
54
100
0.1
Bootstrap 100%

Figure 1.8. LB gene sequence removal. A maximum likelihood tree after the removal of 1,750 of the longest-branch gene sequences. For every gene tree with the topology constrained as in Fig. 1.6, the distance from each taxon to the root of the tree was calculated using TreeStat. The longest branched sequences were progressively removed from the gene alignments. If representation of a taxon in the concatenated alignment dropped <5% of positions, the taxon was removed from the concatenated alignment. If the number of taxa in a gene alignment dropped below 4, the gene was removed from the concatenation. The tree was constructed in RAxML using the WAG+Γ model and colored as described in Fig. 1.6. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping. Branches that received maximum support by all methods are indicated by full circles, dashes indicate bootstrap values <50%. From Hampl *et al.* (2009).


89
94
56
-
54
55
-
-
75
94
77
94
64
80
59
98
100
91
Acanthamoeba
Mastigamoeba
Dictyostelium
Physarum
AMOEBOZOA
Ustilago
Blastocladiella
Drosophila
Homo
OPISTHOKONTA
Proterospongia
Monosiga
'UNIKONTS'
'Malawimonas californiana'
Malawimonas jakobiformis
Trichomonas
Giardia
Monocercomonoides
Trimastix
METAMONADA
Andalucia
Jakoba bahamensis
'Seculamonas '
Jakoba libera
Reclinomonas
Histiona
Stachyamoeba
Naegleria
Sawyeria
Euglena
Leishmania
Trypanosoma
DISCOBA
EXCAVATA
Glaucocystis
Cyanophora
Oryza
Chlamydomonas
Micromonas
ARCHAEPLASTIDA
Pavlova
Emiliania
'CHROMALVEOLATES'
Chondrus
Porphyra
ARCHAEPLASTIDA
Bigelowiella
Cercomonas
RHIZARIA
Blastocystis
Phytophthora
Thallassiosira
Tetrahymena
Oxytricha
Toxoplasma
Amphidinium
Alexandrium
'CHROM-
ALVEOLATES'
0.1
Bootstrap 100%

CHAPTER 2

MATERIALS AND METHODS

Culturing and Sequence Data

Sequences from nine genes were examined for this analysis: Seven nuclear-encoded protein-coding genes (Actin, Btub, Cal-1, EF1α, EF2, cHSP70, and cHSP90) and two genes from ribosomal DNA (SSU and LSU) (**Table 1.2**). These sequences were obtained from 61 taxa from throughout the Eukaryotic Tree of Life, including 30 taxa from the supergroup Excavata with nine of the ten groups of excavates represented by at least one taxon (**Table 2.1**). Retortamonads were not included in this study due to the limited availability of sequence data, little or no culture availability, and noted problems with sequencing (AG Simpson *pers. comm.* 2008). The lack of a well supported sister group to the excavates required the use of representatives from every other supergroup to be considered as outgroup taxa. Extensive gene name and blast searches were performed for these taxa using several gene databases: GenBank (http://www.ncbi.nlm.nih.gov/), TBestDB (http://tbestdb.bcm.umontreal.ca/searches/welcome.php), GeneDB (http://www.genedb.org/), JGI genome projects (http://genome.jgi-psf.org/), and EuPathDB (http://eupathdb.org/eupathdb/). All sequences were imported into Geneious v.4.5.4 (Biomatters, Ltd.) for viewing. Excavate-specific primers were designed from alignments performed using Muscle v3.7 (Edgar 2004a, Edgar 2004b) on the EBI web server (http://www.ebi.ac.uk/Tools/muscle/index.html) or from alignments using Muscle v3.7 within the Geneious software (**Table 2.2**).

To provide a robust gene sequence data set and to represent a broad range of in-group taxa, several cultures, mostly Euglenozoans, were used in this study (**Table 2.3**). Total RNA was extracted from cultures of these taxa during log phase growth using the RNeasy Plant Mini Kit (Qiagen, Cat#

74903), respectively. cDNA libraries were made from the total extracted RNA for each organism using Superscript II Reverse Transcriptase (Invitrogen, Cat# 18064-014). All DNA was diluted to concentrations ranging from 10 ng/μL to 100 ng/μL. Polymerase Chain Reactions (PCRs) were performed on the gDNA and cDNA with the following reaction mix: 12.4 μL of molecular-grade water, 2.5 μL of 10x Bioline Buffer, 1 μL of DMSO, 5 μM $MgCl_2$, 4 mM dNTPs, 0.6 μM of each primer (forward and reverse), 1 μL of diluted DNA and 0.1 μL of Biolase DNA Polymerase (Bioline, Cat# BIO-21043). The annealing temperatures for the PCR amplifications ranged from 37-56°C. Most amplification programs consisted of 30 reiterating cycles. PCR amplification of EF1α had a slightly different program consisting of 35 cycles with gradually increasing extension times for each cycle. PCR products were sequenced either by direct sequencing or by cloning via TOPO-TA cloning (Invitrogen, Cat# K4500-40) or pGEM T-Easy Vector cloning (Promega, Cat# A1380). Sequence data was edited in Geneious to remove extraneous sequence and to assemble forward and reverse reads of multiple clones. Cal-1 sequences obtained in this study were procured with the assistance of Franklyn Aguebor at the University of Georgia. The final distribution of taxa and gene sequence data used in this study are shown in **Table 2.1**. Gene sequences obtained in this study will be deposited in GenBank.

Multiple Sequence Alignments (MSAs) and Determination of Model of Sequence Evolution (MoSE)

MSAs were performed on the protein-coding gene sequences using Muscle via the EBI web server. SSU and LSU sequences were indirectly aligned based on their secondary structure using SINA (Silva 98 Incremental Aligner) on the Silva website (http://www.arb-silva.de/, Pruesse *et al.* 2007). All of the MSAs were then edited by eye for alignment errors in Geneious. Edited alignments were run through GBlocks v0.91b (Castresana 2000, Talavera and Castresana 2007) with the least stringent settings in an effort to find conserved regions in the alignments for use in phylogenetic analyses (Full alignments). The Codon model was used for the protein-coding gene sequence alignments, and the DNA model was used for the SSU and LSU sequence alignments. In an effort to test for possible differences in phylogenetic

signal due to codon usage, alternative sequence alignments were made for the protein-coding gene sequence alignments by removing the third codon position to approximate amino acid data (Third Codon Removed alignments). The best MoSE was determined for all of the alignments using PAUP*4.0b10 (Swofford 2003) in combination with ModelTest v3.7 (Posada and Crandall 1998) (**Table 2.4**).

Single Gene Analyses

Maximum Likelihood (ML) analyses with 1000 rapid bootstrap (BS) replicates were performed on the full alignments and the third codon removed alignments of each gene individually using RAxML v7.0.4 (Stamatakis *et al.* 2005, Stamatakis *et al.* 2008) via the Cipres Portal (http://www.phylo.org/sub_sections/portal/) to determine if there was any evidence of horizontal gene transfer (HGT). All analyses were performed under the GTR MoSE because it is the only nucleotide model that RAxML supports, but other parameters (I and Γ) from the ModelTest output were complied with during these analyses (**Table 2.4**). As no evidence of HGT was apparent, sequence alignments of all nine genes were concatenated for use in subsequent analyses. In order to determine how the concatenated alignment should be partitioned, the parameters of the individual gene alignments (Full and Third Codon Removed alignments) were estimated using PhyML v2.4.4 (Guindon and Gascuel 2003) via the LIRMM online server (http://atgc.lirmm.fr/phyml/, Guindon *et al.* 2005) and PhyML v3.0 via the Montpelier online server (http://www.atgc-montpellier.fr/phyml/). The data for all partitioned analyses were partitioned with each gene in a separate partition based on a comparison of the estimated parameters of each gene.

Phylogenetic Analyses of Concatenated Data Set 1 (Full Alignments + SSU and LSU)

ML analyses with 1000 BS replicates were performed on Concatenated Data Set 1 under the specified best MoSE using RAxML via the Cipres Portal. Both partitioned and unpartitioned ML analyses were performed. All trees were rooted with the Stramenopiles, *Phytophthora* and *Thalassiosira pseudonana*. Due to the unusual topology of these trees in which well-established supergroups (e.g.,

Opisthokonta) were not recovered as monophyletic, further analyses were performed only on the concatenated data set using the Third Codon Removed alignments + SSU and LSU alignments (Concatenated Data Set 2).

Phylogenetic Analyses of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU)

All analyses (ML, Bayesian, and weighted and unweighted Maximum Parsimony (MP)) were performed on Concatenated Data Set 2 under the specified best MoSE, and all topologies were rooted with the Stramenopiles, *Phytophthora* and *T. pseudonana*. RAxML via the Cipres Portal was used to perform ML analyses with 1000 BS replicates on the partitioned and unpartitioned Concatenated Data Set 2. MrBayes v3.1.2 was used to perform Bayesian analyses on the partitioned and unpartitioned Concatenated Data Set 2 each with 4 runs of approximately 4 million generations (Altekar *et al*. 2004, Huelsenbeck *et al.* 2001, Ronquist and Huelsenbeck 2003). The covarion model (Tuffley and Steel 1998) was also applied to the partitioned and unpartitioned data set using MrBayes to see if heterotachy was playing a major role in the analyses with 4 runs of approximately 8 million generations and 4 runs of approximately 10 million generations, respectively. Single gene jackknifing, in which the data set is analyzed after removal of a single gene from the data set, was performed with removal of each gene individually using RAxML via the Cipres Portal.

Additionally, weighted and unweighted MP analyses were performed on the data set to find the most parsimonious tree (1 run) and to generate BS support values (1000 BS replicates) using a New Technology search with Sectorial Search, Ratchet, Drift, and Tree Fusing selected in TNT v1.1 via the LIRMM online server (http://phylogeny.lirmm.fr/phylo_cgi/index.cgi) (Dereeper *et al.* 2008, Goloboff *et al.* 2000). For weighted MP analyses, transitions were given a cost of 1 while transversions were given a cost of 2, giving a transitions:transversions ratio of 2:1. The Consistency Index (CI) and the Retention Index (RI) for these analyses were determined by running specific scripts on the output trees in TNT v1.1. These MP analyses were repeated on Concatenated Data Set 2 after the removal of the heterolobosean,

Naegleria gruberi, to determine if this taxon was responsible for the amoebozoans being recovered with the excavates.

Topology Tests

Twenty possible alternative tree topologies were analyzed using the Approximately Unbiased (AU) topology test to see if any of the topologies was significantly more likely than the other topologies (Shimodaira 2002). These topologies were based on the resultant trees from previously mentioned analyses (Further discussion in Chapter 3). The different topologies were created using Winclada (BETA) v1.00.00 (Nixon 1999) starting with the MP tree with the best score. The trees were then imported into TreeView v1.6.6 (Page 1996) and immediately exported in Newick format. The site likelihoods for these topologies were calculated using RAxML v7.0.4 with the best MoSE with four Γ rate categories. AU tests were performed using CONSEL v0.1i (Shimodaira and Hasegawa 2001).

Gap Coding

The alignments for SSU and LSU from SINA that were edited by eye in Geneious (without being run through Gblocks) were used to obtain morphological data from the molecular data set. The alignments were run through SeqState v1.40 (Müller 2005) utilizing the simple gap-coding program. Parsimony uninformative characters were removed from the resultant matrices using Winclada. These matrices were analyzed for characters significant to the excavates.

Table 2.1. Distribution of Taxa and Sequences

Accession numbers from GenBank, TBestDB, GeneDB, JGI Genome Projects, EuPathDB, and Dr. Jan Andersson;
Yellow - Coded as missing data; Organisms with only a generic name are from more than one species within the genus

Taxon Name	Supergroup	Subgroup	Actin	Btub	EF1α	EF2	cHSP70	cHSP90	SSU	LSU	Cal-1
Carpediemonas membranifera	Excavata	*Carpediemonas*	--	AY117422	--	--	AY131204	DQ295219	AY117416	--	--
Giardia intestinalis	Excavata	Diplomonads	XM_001704601	XM_001707320	XM_001704477	XM_001704768	XM_001707918	AB092407, AB092408	AF473852	EuPathDB: GL50803_r0021	XM_001705768
Hexamita inflata	Excavata	Diplomonads	--	AY277792	U37081	--	--	AY462239	HXM16RR	--	--
Spironucleus barkhanus	Excavata	Diplomonads	J. Andersson Contig560, Contig537, Contig64	U29441	U29442	DQ295234	AY131206	DQ295222	DQ273887.1	--	DQ812518
Diplonema papillatum	Excavata	Euglenozoans / Diplonemids	Lab	Lab	EFL	DPL00001934	DPL00000060, DPL00002701, DPL00000040	AY122623, DPL00000122, DPL00001912	AF119811	B. Ma	DPL00003786
Rhynchopus	Excavata	Euglenozoans / Diplonemids	RSL00000357	RSL00000832, RSL00000393	EFL	--	AY288513	AY122622	AY425013	RSL00000098	--
Entosiphon sulcatum	Excavata	Euglenozoans / Euglenids	--	AF095840	FJ807254	--	--	DQ683347	AF220826	--	--
Euglena gracilis	Excavata	Euglenozoans / Euglenids	AF057161	AF182554	ELL00002603	AF213663, ELL00002564	AY288512, ELL00002598	AY288511	AF283308	X53361	ELL00000058
Petalomonas cantuscygni	Excavata	Euglenozoans / Euglenids	--	Lab	EFL	Lab	--	DQ683346	AF386635	B. Ma	--
Peranema tricophorum	Excavata	Euglenozoans / Euglenids	--	Lab	FJ807243	--	PTL00000010, PTL00000007	DQ683345	AF386636	AY130826	Lab
Bodo saltans	Excavata	Euglenozoans / Kinetoplastids	--	DQ450531	FJ807252	--	AY651257	AY122632	AY490233	EF681898	Lab
Leishmania major	Excavata	Euglenozoans / Kinetoplastids	XM_883541	XM_001685771	XM_001682206	XM_001686536	XM_001684511	XM_001685707	X53915	CP000079	EuPathDB: LmjF09.0920
Rhynchomonas nasuta	Excavata	Euglenozoans / Kinetoplastids	--	DQ295213	FJ807250	--	AY288514	AY122625	AY425023	DQ086724	--
Trypanosoma brucei	Excavata	Euglenozoans / Kinetoplastids	XM_822112	XM_001218932	XM_817372	XM_817610	XM_824105	XM_818214	M12676	NC_008409	XM_824266
Trypanosoma cruzi	Excavata	Euglenozoans / Kinetoplastids	XM_802169	XM_811597	XM_814346	XM_803948	XM_812645	XM_799387	AF359495	L22334	XM_802997
Naegleria gruberi	Excavata	Heteroloboseans	AF101729	Z13961	DQ295229	DQ295230	AY288516	AY122634	M18732	AB298288	U04381
Stachyamoeba lipophora	Excavata	Heteroloboseans	SLL00000016, SLL00000516	--	SLL00000002, SLL00001883	SLL00000094, SLL00001585	SLL00001720, SLL00001420, SLL00002214	SLL00000816, SLL00002232	SLL00000112	SLL00001582	SLL00001104
Andalucia incarcerata	Excavata	Jakobids	EU334881	EU334883	EU334884	EU334885	--	EU334886	EU334887	--	--
Histiona aroides	Excavata	Jakobids	HAL00001020	HAL00000551	HAL00000058	HAL00001149, HAL00000477	HAL00001640, HAL00001245	HAL00001244	HAL00000355	HAL00000043	HAL00000031
Jakoba bahamensis	Excavata	Jakobids	JBL00000106	JBL00000610	JBL00000183	JBL00000178	--	JBL00000507, JBL00000446	JBL00000661, JBL00000818, JBL00000193	JBL00002032, JBL00000437, JBL00000837, JBL00000901	JBL00000072
Jakoba libera	Excavata	Jakobids	JLL00000765	AF267184, JLL00000772	JLL00000774	JLL00000743	--	JLL00000621, JLL00000644	AY117418	JLL00002533	JLL00000155
Reclinomonas americana	Excavata	Jakobids	RAL00001684, RAL00001650	AF267188, RAL00001606	DQ295232	DQ295231, RAL00001355	DQ295225, RAL00000687	DQ295221, RAL00001632	AY117417	RAL00001253, RAL00000007	RAL00001437
Seculamonas ecuadoriensis	Excavata	Jakobids	SEL00000733	SEL00000718, SEL00000734	SEL00000743	SEL00000697	SEL00001720, SEL00000016, SEL00000035, SEL00001837	SEL00000626	DQ190541	SEL00001092	SEL00000502

Taxon Name	Supergroup	Subgroup	Actin	Btub	EF1α	EF2	cHSP70	cHSP90	SSU	LSU	Cal-1
Malawimonas californiana	Excavata	*Malawimonas*	MCL00000711, MCL00000769	MCL00000748	MCL00000750	MCL00000554, MCL00001634, MCL00000326	MCL00002186	MCL00001811, MCL00002203	MCL00000562	MCL00000650, MCL00001442	MCL00000552, MCL00001064
Malawimonas jakobiformis	Excavata	*Malawimonas*	EF455790, MJL00000669	AF267185, MJL00000639, MJL00000461	DQ295227, MJL00000647	DQ295228	DQ295224	DQ295220	AY117420	MJL00004078, MJL00001688, MJL00001637, MJL00001047, MJL00002975	MJL00000073
Monocercomonoides sp.	Excavata	Oxymonads	--	EF474124	EF474125	AY831447	AY831449	AY831450	AY831435	--	--
Streblomastix strix	Excavata	Oxymonads	SSL00000053	DQ363673	AY188862, SSL00000161, SSL00000767	SSL00000615, SSL00000358	SSL00000508, SSL00000413	AY188866, SSL00000241, SSL00000037	AY188886	SSL00000240, SSL00000251, SSL00000740	SSL00000871
Trichomonas vaginalis	Excavata	Parabasalids	XM_001301716	XM_001318453	XM_001325448	XM_001321756	XM_001317754	XM_001317510	AY338476	AF202181	XM_001323102
Trimastix pyriformis	Excavata	*Trimastix*	--	Lab	DQ295235	DQ295236	EU327685	EU327684	AF244903	--	--
Trimastix marina	Excavata	*Trimastix*	--	DQ295218	--	--	DQ295226	DQ295223	AF244905	--	--
Porphyra	Archaeplastida	Rhodophytes	AB039831	AY221630	AB048204	AY010231	DQ356007.1	--	AB235853	EF033597	EF368215
Arabidopsis thalania	Archaeplastida	Viridiplantae	NM_112046	AY054693	NM_125432	AC009894	AY059885	NM_124985	X16077	X52320	Z12024
Chlamydomonas reinhardtii	Archaeplastida	Viridiplantae	XM_001699016	XM_001694020	XM_001702295	XM_001703163	XM_001701274	XM_001695212	M32703	EU410621	M20729
Oryza sativa	Archaeplastida	Viridiplantae	AB047313	NM_001055214	AF030517	NM_001050746	AP003231	AB111810	AF069218	M11585	AF042840
Triticum aestivum	Archaeplastida	Viridiplantae	AY663392	U76745	M90077	AF475129	AF005993	DQ665783	AY049040	AY049041	U48242
Dictyostelium discoideum	Amoebozoa	Amoebozoans	XM_638773	L14000	XM_640747	XM_631629	X75263	XM_642390	AM168039	X00601	M64089
Entamoeba histolytica	Amoebozoa	Amoebozoans	M16396	XM_652078	XM_646777	XM_645917	XM_645366	XM_648040	AB426549	X65163	XM_646616
Cryptosporidium parvum (IOWA II)	Chromalveolata	Alveolates / Apicomplexans	XM_001388245	XM_627803	XM_001388307	XM_627193	XM_625373	XM_626924	X64341	AF040725	XM_001388155
Eimeria tenella	Chromalveolata	Alveolates / Apicomplexans	GeneDB: EIMER_contig_00020916	U19609	Gene_DB: EIMER_contig_00004439	GeneDB: EIMER_contig_00020987	GeneDB: EIMER_contig_00030871	AF042329	AF026388	AF026388	Z71757
Plasmodium falciparum	Chromalveolata	Alveolates / Apicomplexans	EF472536	X16075	XM_001350245	XM_001348624	XM_001349300	XM_001348962	AF145334	U21939	M59770
Toxoplasma gondii	Chromalveolata	Alveolates / Apicomplexans	TGU10429	M20025	AM055942	GeneDB: 20.m03912	U85649	AY292370	M97703	AF076901	Y08373
Paramecium tetraurelia	Chromalveolata	Alveolates / Ciliates	CR855973	X67237	AF172083	XM_001433618	CR933371	XM_001447758	X03772	AF149979	M34540
Stylonychia	Chromalveolata	Alveolates / Ciliates	DQ108617	AF510208	X57926	AF213664	AF227962	--	AF164124	AF508773	M76407
Tetrahymena thermophila	Chromalveolata	Alveolates / Ciliates	XM_001016672	XM_001023006	XM_001032213	AF534908	AY028633	XM_001009780	X56165	X54512	X52242
Heterocapsa triquetra	Chromalveolata	Alveolates / Dinoflagellates	EF640328, HTL00001474, HTL00000001	AF482413	EFL	HTL00000969, HTL00001989, HTL00000080	AY729868, HTL00001491, HTL00000576	AY729855, HTL00001511, HTL00001295	AF022198	AF260401	EU153195
Isochrysis galbana	Chromalveolata	Haptophytes	AY729843, ISL00000902	AY729818, ISL00001380, ISL00000444	EFL	ISL00001456, ISL00004505, ISL00007110	ISL00000711	AY729856, ISL00001487	AJ246266	DQ202390, ISL00000054, ISL00003004	ISL00000660
Pavlova lutheri	Chromalveolata	Haptophytes	PLL00000105	AY729820, PLL00000031, PLL00001519	EFL	PLL00000385	X59555	AY729858, PLL00003075	AF102369	PLL00001069, PLL00000833, PLL00001988	PLL00000234, PLL00001545
Thalassiosira pseudonana	Chromalveolata	Stramenopiles / Diatoms	JGI: Thaps3/chr_22: 805000-806500	JGI: Thaps3/chr_9:671000-673234	JGI: Thaps3/chr3:1974630-1976121	JGI: Thaps3/chr_6:1462832-1465396	JGI: Thaps3/chr_6:1187800-1189500	JGI: Thaps3/chr_6:1191772-1194168	AF374481	JGI: Thaps3/chr_17:653530-655390	CP001160

Taxon Name	Supergroup	Subgroup	Actin	Btub	EF1α	EF2	cHSP70	cHSP90	SSU	LSU	Cal-1
Phytophthora	Chromalveolata	Stramenopiles / Oomycetes	AY244551	EU079613	AJ249839	CF891679	AY456093	EU079629	AY742761	X75631	M83535
Monosiga brevicollis	Opisthokonta	Choanoflagellates	AY026072	XM_001743918	AY026073	AY026074	XM_001743188	AY226081	AF100940	AY026374	CH991568
Candida albicans	Opisthokonta	Fungi	X16377	M19398	XM_706806	Y09664	Z30210	X81025	M60302	X70659	M61128
Neurospora crassa	Opisthokonta	Fungi	XM_956040	BX897679	D45837	AF258620	U10443	XM_956205	X04971	FJ360521	L02964
Saccharomyces cerevisiae	Opisthokonta	Fungi	V01288	V01296	X00779	M59369	J05637	NC_001148	Z75578	AY048154	M14760
Schizosaccharomyces pombe	Opisthokonta	Fungi	NM_001021513	AF042827	NM_001022750	D83975	AB012387	NM_001019786	X58056	Z19578	NM_001018772
Bombyx mori	Opisthokonta	Metozoa	NM_001126255	NM_001043500	NM_001044045	DQ443396	NM_001043931	NM_001043411	DQ347470	AY038991	DQ311341
Caenorhabditis elegans	Opisthokonta	Metozoa	NM_076440	NM_077184	U51994	M86959	M18540	Z75530	EU196001	X03680	NM_070985
Danio rerio	Opisthokonta	Metozoa	BC045846	NM_198809	NM_200009	AY391422	BX120005	NM_131328	BX537263	AF398343	NM_213351
Drosophila melanogaster	Opisthokonta	Metozoa	NM_078497	M20419	M11744	X15805	L01501	X03810	M21017	M21017	NM_078986
Homo sapiens	Opisthokonta	Metozoa	NM_001614	NM_001069	NM_001958	NM_001961	AF352832	NM_003299	X03205	NR_003287	NM_006888
Cercomonas	Rhizaria	Cercomonads	EF455793	AF119173	--	--	--	--	AF411271	DQ386165	--
Paracercomonas marina	Rhizaria	Cercomonads	CLL00000009	--	CLL00000114	CLL00000113	--	--	CLL00000486, CLL00000290	DQ386164, CLL00000239	CLL00000105

Table 2.2. Primer Sequences

Gene	Primer Name	Direction	Sequence 5' - 3'	Reference
Actin	ACT_88XF_Orig	Forward	TGGGACGACATGGARAARATHTGG	From Simpson *et al.* 2008
	SEA_ACTBXR	Reverse	TAAGCAYTTBYKRTGSACRAT	Modified from Simpson *et al.* 2008
Btub	BtubA	Forward	GCGGYCARTGYGGNAACCA	Modified from Simpson *et al.* 2008
	BtubB	Reverse	CCGTGAAYTCCATYTCRTCCAT	Modified from Simpson *et al.* 2008
	SEA_BtubB	Reverse	CCCAGTRAAYTCCATYTCRTCCAT	Modified from Simpson *et al.* 2008
EF2	SEA_EF2F1B	Forward	TGTGATCGCCCAYGTNGAYCAYGGNAA	Modified from Simpson *et al.* 2008
	EF2R1C_A	Reverse	TCCARTGGTSRAAMACRCAYTGYGGGAA	Modified from Simpson *et al.* 2008
Cal-1	Cal1_K186	Forward	GGATGGCHGANSAMYTGWCSVA	Modified from Karabinos and Bhattacharya 2000
	Cal1_1R	Reverse	GGCATCATCATYTTSACRAAYTC	This Study
	Cal1_2R	Reverse	GSACRAAYTCCTCRTAGTTGATYTG	This Study

Table 2.3. Culturing Information

ATCC = American Type Culture Collection; CCAP = Culture Collection of Algae and Protozoa

Organism	Culture ID	Medium	Temp (°C)
Diplonema papillatum	ATCC 50162	**Artificial Seawater** (from Ultramarine Sea Salts, Waterlife, Ltd.) w/ Chloramphenicol (100µg/ml) and 1% Heat-Inactivated Horse Serum	25
Petalomonas cantuscygni	CCAP 1259/1	**ESNW** (Enriched Soil extract Natural Seawater medium) w/ rice (modified from Harrison *et al.* 1980)	20
Peranema tricophorum	CCAP 1260/1B	**Hay Infusion** (Boiled 2.5g hay in 1L distilled water for 30 minutes, filtered, and autoclaved) w/ sterile egg yolk	25
Bodo saltans	CCAP 1907/2	**CCAP S/W**	15
Trimastix pyriformis	ATCC 50562	**ATCC 802**	25

Table 2.4. Best MoSE for Alignments
I = Invariant Sites; Γ = Gamma Distribution
*Based on Akaike Information Criterion results from ModelTest v3.7

Gene	Alignment	Best MoSE*
Actin	Full	GTR + I + Γ
	Third Codon Removed	TIM + I + Γ
Btub	Full	GTR + I + Γ
	Third Codon Removed	TrN + I + Γ
Cal-1	Full	GTR + I + Γ
	Third Codon Removed	GTR + Γ
EF1α	Full	GTR + I + Γ
	Third Codon Removed	GTR + I + Γ
EF2	Full	GTR + I + Γ
	Third Codon Removed	GTR + I + Γ
cHSP70	Full	GTR + I + Γ
	Third Codon Removed	GTR + I + Γ
cHSP90	Full	GTR + I + Γ
	Third Codon Removed	GTR + I + Γ
SSU	SINA	GTR + I + Γ
LSU	SINA	GTR + I + Γ
Concatenated data set 1	All Full alignments + SSU and LSU	GTR + I + Γ
Concatenated data set 2	All Third Codon Removed alignments + SSU and LSU	GTR + I + Γ

CHAPTER 3

RESULTS

Single Gene Analyses

Alpha-tubulin, a gene used in previous phylogenetic studies of the supergroup Excavata, seems to have been subject to at least one horizontal gene transfer (HGT) event (Simpson *et al.* 2006, Simpson *et al.* 2008). In order to prevent HGT from affecting the results in this study, each gene, with and without the third codon removed, was analyzed individually to check for evidence of HGT. There was no evidence in any of the individual ML gene trees that indicated an HGT event. Although the topologies did not recover monophyly for most groups, all of the well-supported nodes in the single gene trees were for widely accepted relationships, such as monophyly of the plants. Because there seemed to be no strong evidence of HGT in any of the genes, all of the genes analyzed were used in subsequent analyses.

Each of the genes was then analyzed to determine the best way to partition the concatenated data set. The parameters of each gene were estimated and analyzed to determine if any of the genes should be partitioned together. A number of parameters (gamma shape parameters, proportions of invariant sites, nucleotide frequencies, and instantaneous rate matrices) of each gene were compared to the parameters of the other genes. After reviewing the parameters, no clear distinction between the gene parameters could be used to partition the concatenated data set. Therefore, all partitioned analyses had each gene in a separate partition.

Phylogenetic analyses of Concatenated Data Set 1 (Full Alignments + SSU and LSU)

The topology of the Maximum Likelihood (ML) analysis of Concatenated Data Set 1, both partitioned and unpartitioned, was not generally concordant with accepted relationships within the

Eukaryotic Tree of Life (**Fig. 3.1**). Particularly, some of the opisthokonts (the fungi) were recovered as sister to an unexpected clade of chromalveolates, oxymonads, amoebozoans, and a heterolobosean. In addition, the haptophytes within the chromalveolates are found among the Archaeplastida between the viridiplantae and the rhodophyte, *Porphyra*. However, some expected relationships were recovered. The rhizarians were recovered as sister to the stramenopiles among the chromalveolates. Seven of the nine groups of excavates in this study were recovered, and the Group 1 excavates were recovered as a monophyletic clade. Neither Group 2 nor Group 3 excavates were recovered as monophyletic clades, and monophyly of Excavata was not recovered in these analyses. Due to the unexpected topology of the ML analyses of Concatenated Data Set 1 in which well-established supergroups (e.g., Opisthokonta) were not recovered as monophyletic, all further analyses were performed on Concatenated Data Set 2.

Phylogenetic analyses of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU)

The topologies from the ML and Bayesian analyses of Concatenated Data Set 2 were largely congruent (**Figs. 3.2a, 3.2b**). All of these analyses recovered the monophyly of Excavata with moderate to strong support. The parabasalid, *Trichomonas vaginalis*, was recovered as sister to the oxymonad, *Streblomastix strix*, which is not consistent with previous analyses (Hampl *et al.* 2009, Simpson *et al.* 2006, Simpson *et al.* 2008). Due to this placement, neither Group 1 nor Group 2 excavates were recovered as previously described, and the oxymonads were not recovered as monophyletic. However, all taxa within Groups 1 and 2 were recovered as a monophyletic clade (Group 1/2). Group 3 excavates were monophyletic and recovered as sister to the malawimonads.

The other supergroups in the analyses were mostly recovered as monophyletic clades with the exception of the haptophytes being recovered within the archaeplastidans as seen in **Fig 3.1**. Also, the location of the rhizarians changes between sister to the chromalveolates and sister to the stramenopiles within the chromalveolate clade. The partitioned analyses did recover the amoebozoans as sister to the opisthokonts (with low support) rather than as sister to the excavates, which is consistent with the unikont theory (Cavalier-Smith 2003). The covarion model seemed to have no effect on the topology of the

unpartitioned Bayesian analysis. It had a slight effect on the topology of the partitioned Bayesian analysis with recovery of the amoebozoans as sister to the excavates and the rhizarians as sister to the chromalveolates. Therefore, heterotachy did not seem to have been a significant factor in this particular data set.

Both weighted and unweighted Maximum Parsimony (MP) analyses were also performed on this data set (**Figs. 3.3a, 3.3b**). The most parsimonious tree had some unexpected relationships in its topology (**Fig 3.3a**). The excavates were not recovered as a monophyletic clade, but Group 1/2 was recovered. Its location as sister to *Porphyra* was very unusual based on results from previous studies. Also, a clade was recovered with the amoebozoans as sister to the heteroloboseans. However, seven of the nine excavate groups were recovered as monophyletic. When the analysis was repeated with 1000 bootstrap (BS) replicates, the topology agreed with some previously seen relationships (**Fig 3.3b**). All of the outgroups, with the exception of the amoebozoans, were basically recovered as seen in **Figs 3.2a and 3.2b**. The amoebozoans were again recovered in a clade with the heteroloboseans. The excavates were not recovered as a monophyletic clade, but rather as a polytomy with the amoebozoans. Group 1/2 and seven out of nine groups of excavates were recovered as monophyletic.

Due to the fact that the heterolobosean, *Naegleria gruberi*, was seen in unexpected clades with the amoebozoans in **Fig 1.1** and **Fig 3.3**, it was hypothesized that this particular taxon was pulling the amoebozoans closer to the excavates and preventing monophyly of the excavates from being recovered. To test this hypothesis, *N. gruberi* was removed from the data set and the MP analyses were repeated (**Fig 3.4**). This hypothesis was refuted due to the recovery of the amoebozoans with the excavates in both the most parsimonious tree and the tree from 1000 BS replicates. The CI and RI values for all of the MP analyses were within normal limits based on the large amount of taxa included in the analyses (Sanderson and Donoghue 1989).

Single Gene Jackknifing

Single gene jackknifing was performed on the Concatenated Data Set 2 to determine if any one gene was having a marked effect on the ML topology. ML analyses were performed on the data set with each gene removed one at a time. Most of the analyses recovered the topology of ML tree of Concatenated Data Set 2 (**Fig 3.2a**). All but three of these analyses recovered monophyly of the excavates: without cHSP70, without LSU, and without SSU. Removal of cHSP70 from the data set resulted in monophyly of Group 3 with the malawimonads and Group 1/2, but these groups were separated by all of the outgroups. Removal of LSU from the data set resulted in a similar topology, but with less separation between the two groups. Removal of SSU from the data set resulted in an almost monophyletic excavate clade with the exception of *N. gruberi* being recovered with the amoebozoans. The changes seen in these topologies were not well supported.

Topology Tests

Twenty different topologies were tested using the Approximately Unbiased (AU) topology test to determine if any topologies could be statistically rejected based on Concatenated Data Set 2 (**Fig 3.5**). Because the amoebozoans were consistently recovered within the excavates in the MP trees (**Figs 3.3** and **3.4**), major groups of excavates were switched with the amoebozoans to create alternate topologies. Tree 1 was the ML tree from **Fig 3.2a**. Trees 2-9 were based on this topology with the major excavate groups switched with the amoebozoans. Tree 10 was exactly like Tree 1 with the exception of the location of the parabasalid. *Trichomonas vaginalis* was grafted to the base of the *Carpediemonas*-Diplomonadida clade to recreate the Group 1 excavates. Trees 11-18 were based on the topology of Tree 10 with the same major excavate groups switched with the amoebozoans as seen in Trees 2-9. Tree 19 was the most parsimonious tree (**Fig 3.3a**). Tree 20 was **Fig 3.4a (**the most parsimonious tree recovered with *N. gruberi* removed from the data set) with *N. gruberi* grafted to *Stachyamoeba lipophora*, the other heterolobosean. The AU test rejected ($p < 0.050$) all but two topologies: Tree 1 (the ML tree (**Fig 3.2a**)) and Tree 4 (the ML tree with Group 1/2 switched with the amoebozoans). Therefore, although monophyly of the

excavates was well supported, a topology in which excavates were not monophyletic could not be statistically rejected.

Gap Coding

The SSU and LSU alignments from SINA were aligned to 50,000 and 150,000 positions, respectively. The sequences were aligned to alignments based on secondary structure, indirectly aligning them based on secondary structure. Due to the nature of these alignments, there were many gaps throughout these alignments that could be based solely on structural characteristics. Simple gap coding was performed on these alignments to provide morphological data based on these secondary structure alignments. Gap coding of the SSU alignment resulted in 3183 coded gaps with 829 that were parsimony informative. Gap coding of the LSU alignment resulted in 3837 coded gaps with 1179 that were parsimony informative. The gap coding matrices for each alignment were examined by eye for characters that supported previously recovered relationships. Several gaps were found that supported smaller groups, such as the malawimonads, the oxymonads, and *Trimastix*, but no gaps were found that supported the excavates as a whole.

Figure 3.1. ML Tree of Concatenated Data Set 1 (Full Alignments + SSU and LSU). The tree was inferred using RAxML with the GTR+I+Γ model of substitution. Organisms are color-coded by supergroup with all Excavata colored red. Support values are displayed at the nodes with the RAxML bootstrap (BS) support values for the unpartitioned analysis followed by BS support values for the partitioned analysis. Black circles on branches indicate full support for the node. Taxa with asterisks had long branches (>1.0 based on scale) in the partitioned analysis.

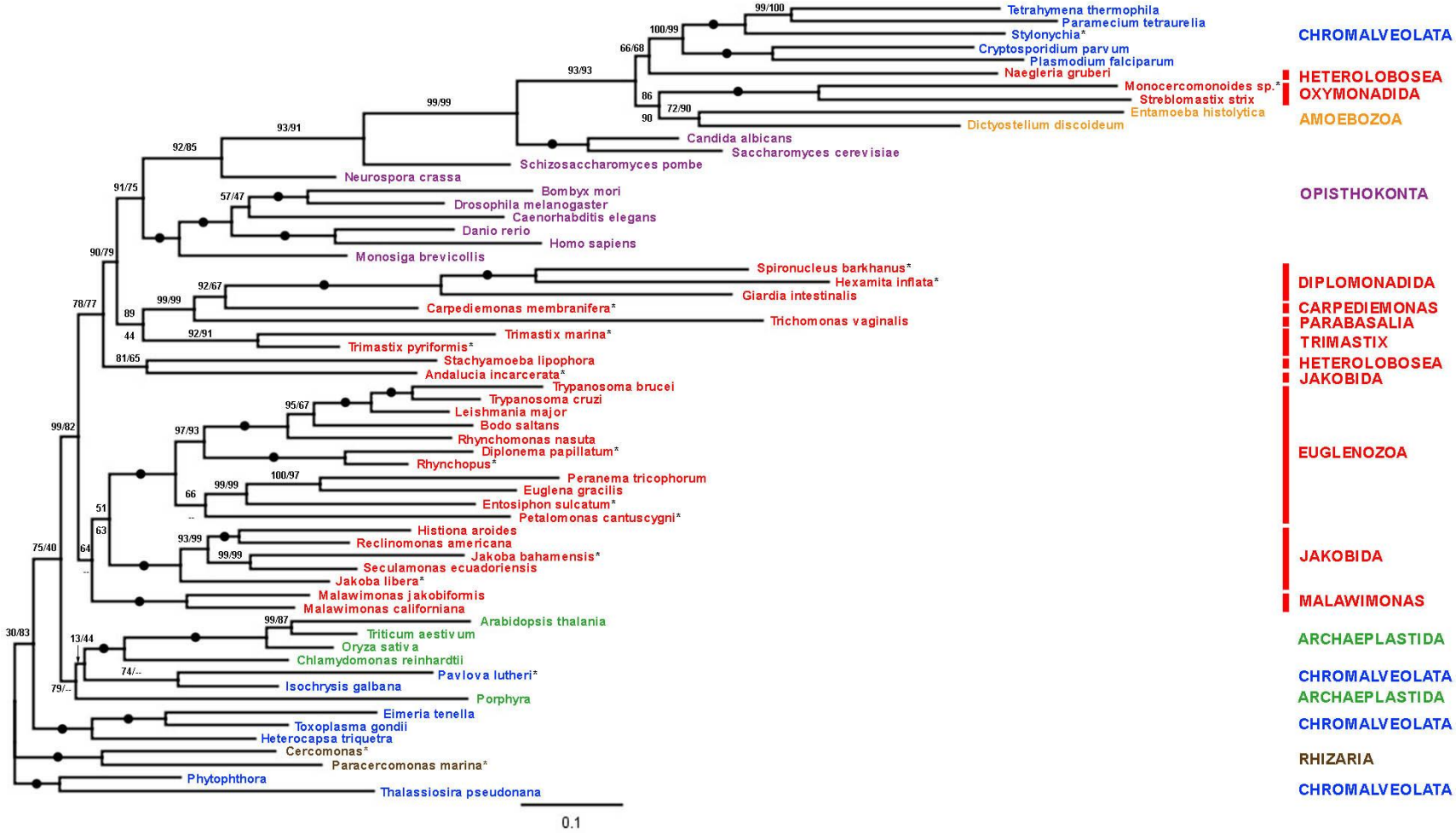

Tetrahymena thermophila
Paramecium tetraurelia
Stylonychia*
Cryptosporidium parvum
Plasmodium falciparum
Naegleria gruberi
Monocercomonoides sp.*
Streblomastix strix
Entamoeba histolytica
Dictyostelium discoideum
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Bombyx mori
Drosophila melanogaster
Caenorhabditis elegans
Danio rerio
Homo sapiens
Monosiga brevicollis
Spironucleus barkhanus*
Hexamita inflata*
Giardia intestinalis
Carpediemonas membranifera*
Trichomonas vaginalis
Trimastix marina*
Trimastix pyriformis*
Stachyamoeba lipophora
Andalucia incarcerata*
Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Diplonema papillatum*
Rhynchopus*
Peranema tricophorum
Euglena gracilis
Entosiphon sulcatum*
Petalomonas cantuscygni*
Histiona aroides
Reclinomonas americana
Jakoba bahamensis*
Seculamonas ecuadoriensis
Jakoba libera*
Malawimonas jakobiformis
Malawimonas californiana
Arabidopsis thalania
Triticum aestivum
Oryza sativa
Chlamydomonas reinhardtii
Pavlova lutheri*
Isochrysis galbana
Porphyra
Eimeria tenella
Toxoplasma gondii
Heterocapsa triquetra
Cercomonas*
Paracercomonas marina*
Phytophthora
Thalassiosira pseudonana
0.1
CHROMALVEOLATA
HETEROLOBOSEA
OXYMONADIDA
AMOEBOZOA
OPISTHOKONTA
DIPLOMONADIDA
CARPEDIEMONAS
PARABASALIA
TRIMASTIX
HETEROLOBOSEA
JAKOBIDA
EUGLENOZOA
JAKOBIDA
MALAWIMONAS
ARCHAEPLASTIDA
CHROMALVEOLATA
ARCHAEPLASTIDA
CHROMALVEOLATA
RHIZARIA
CHROMALVEOLATA

Figure 3.2a. ML tree of Unpartitioned Analysis of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU). The tree was inferred using RAxML with the GTR+I+Γ model of substitution. Support values are displayed at the nodes with the RAxML BS support values, the Bayesian posterior probability (PP), and the Bayesian PP for the covarion analysis of the unpartitioned data set. Black circles on branches indicate full support for the node. Organisms are color-coded as in Fig. 3.1. The yellow box shows the support values for monophyly of Excavata.

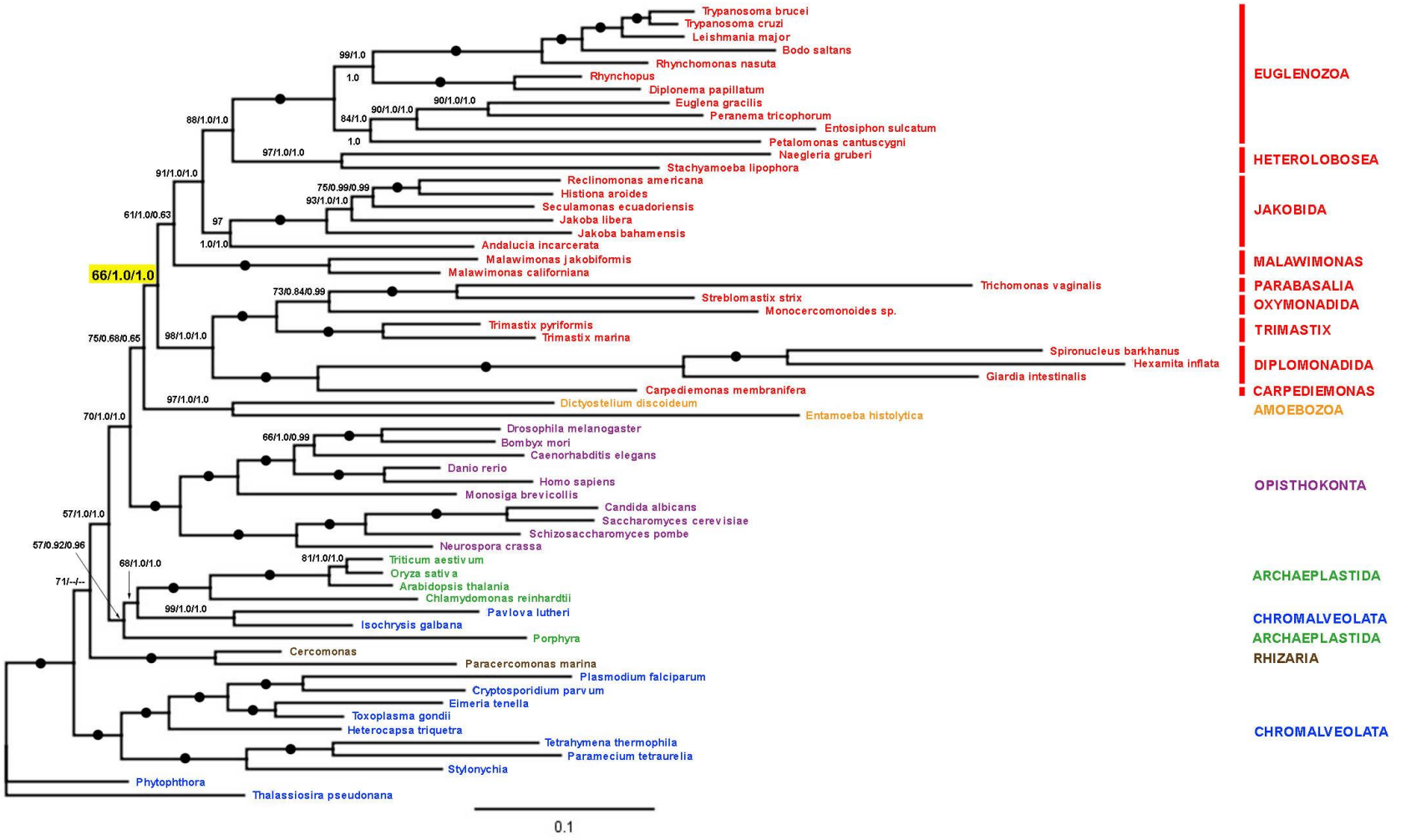

Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Rhynchopus
Diplonema papillatum
Euglena gracilis
Peranema tricophorum
Entosiphon sulcatum
Petalomonas cantuscygni
Naegleria gruberi
Stachyamoeba lipophora
Reclinomonas americana
Histiona aroides
Seculamonas ecuadoriensis
Jakoba libera
Jakoba bahamensis
Andalucia incarcerata
Malawimonas jakobiformis
Malawimonas californiana
Trichomonas vaginalis
Streblomastix strix
Monocercomonoides sp.
Trimastix pyriformis
Trimastix marina
Spironucleus barkhanus
Hexamita inflata
Giardia intestinalis
Carpediemonas membranifera
Dictyostelium discoideum
Entamoeba histolytica
Drosophila melanogaster
Bombyx mori
Caenorhabditis elegans
Danio rerio
Homo sapiens
Monosiga brevicollis
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Triticum aestivum
Oryza sativa
Arabidopsis thalania
Chlamydomonas reinhardtii
Pavlova lutheri
Isochrysis galbana
Porphyra
Cercomonas
Paracercomonas marina
Plasmodium falciparum
Cryptosporidium parvum
Eimeria tenella
Toxoplasma gondii
Heterocapsa triquetra
Tetrahymena thermophila
Paramecium tetraurelia
Stylonychia
Phytophthora
Thalassiosira pseudonana
99/1.0
1.0
90/1.0/1.0
84/1.0
88/1.0/1.0
97/1.0/1.0
91/1.0/1.0
75/0.99/0.99
93/1.0/1.0
61/1.0/0.63
97
1.0/1.0
66/1.0/1.0
73/0.84/0.99
75/0.68/0.65
98/1.0/1.0
97/1.0/1.0
70/1.0/1.0
66/1.0/0.99
57/1.0/1.0
57/0.92/0.96
71/--/--
68/1.0/1.0
81/1.0/1.0
99/1.0/1.0
0.1
EUGLENOZOA
HETEROLOBOSEA
JAKOBIDA
MALAWIMONAS
PARABASALIA
OXYMONADIDA
TRIMASTIX
DIPLOMONADIDA
CARPEDIEMONAS
AMOEBOZOA
OPISTHOKONTA
ARCHAEPLASTIDA
CHROMALVEOLATA
ARCHAEPLASTIDA
RHIZARIA
CHROMALVEOLATA

Figure 3.2b. ML tree of Partitioned Analysis of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU). The tree was inferred using RAxML with the GTR+I+Γ model of substitution. Support values are displayed at the nodes with the RAxML BS support values, the Bayesian posterior probability (PP), and the Bayesian PP for the covarion analysis of the partitioned data set. Black circles on branches indicate full support for the node. Organisms are color-coded as in Fig. 3.1. The yellow box shows the support values for monophyly of Excavata.

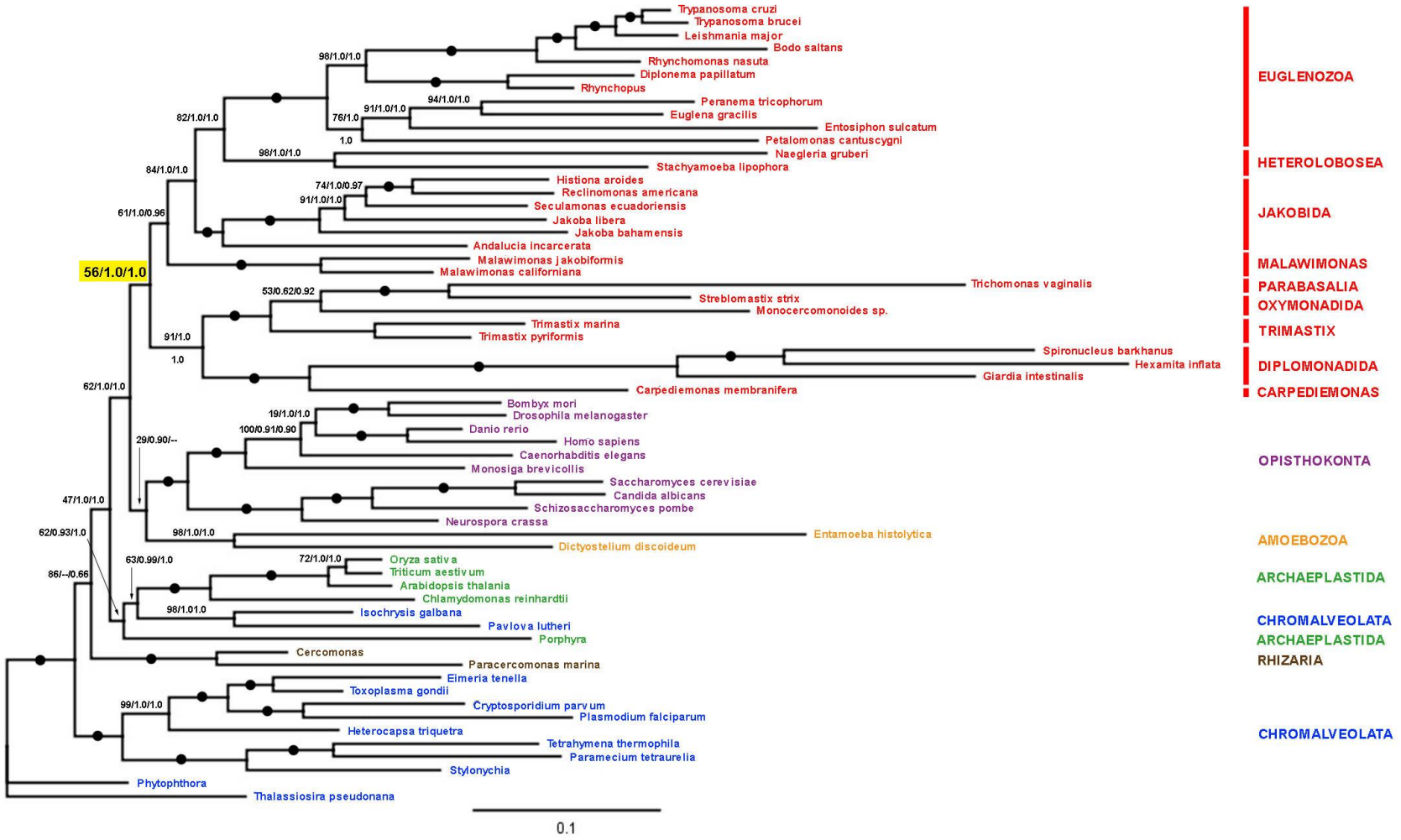

Trypanosoma cruzi
Trypanosoma brucei
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Diplonema papillatum
Rhynchopus
Peranema tricophorum
Euglena gracilis
Entosiphon sulcatum
Petalomonas cantuscygni
Naegleria gruberi
Stachyamoeba lipophora
Histiona aroides
Reclinomonas americana
Seculamonas ecuadoriensis
Jakoba libera
Jakoba bahamensis
Andalucia incarcerata
Malawimonas jakobiformis
Malawimonas californiana
Trichomonas vaginalis
Streblomastix strix
Monocercomonoides sp.
Trimastix marina
Trimastix pyriformis
Spironucleus barkhanus
Hexamita inflata
Giardia intestinalis
Carpediemonas membranifera
Bombyx mori
Drosophila melanogaster
Danio rerio
Homo sapiens
Caenorhabditis elegans
Monosiga brevicollis
Saccharomyces cerevisiae
Candida albicans
Schizosaccharomyces pombe
Neurospora crassa
Entamoeba histolytica
Dictyostelium discoideum
Oryza sativa
Triticum aestivum
Arabidopsis thalania
Chlamydomonas reinhardtii
Isochrysis galbana
Pavlova lutheri
Porphyra
Cercomonas
Paracercomonas marina
Eimeria tenella
Toxoplasma gondii
Cryptosporidium parvum
Plasmodium falciparum
Heterocapsa triquetra
Tetrahymena thermophila
Paramecium tetraurelia
Stylonychia
Phytophthora
Thalassiosira pseudonana
98/1.0/1.0
82/1.0/1.0
76/1.0
91/1.0/1.0
94/1.0/1.0
1.0
98/1.0/1.0
84/1.0/1.0
74/1.0/0.97
91/1.0/1.0
61/1.0/0.96
56/1.0/1.0
53/0.62/0.92
91/1.0
1.0
62/1.0/1.0
19/1.0/1.0
100/0.91/0.90
29/0.90/--
47/1.0/1.0
62/0.93/1.0
98/1.0/1.0
72/1.0/1.0
63/0.99/1.0
86/--/0.66
98/1.01.0
99/1.0/1.0
0.1
EUGLENOZOA
HETEROLOBOSEA
JAKOBIDA
MALAWIMONAS
PARABASALIA
OXYMONADIDA
TRIMASTIX
DIPLOMONADIDA
CARPEDIEMONAS
OPISTHOKONTA
AMOEBOZOA
ARCHAEPLASTIDA
CHROMALVEOLATA
ARCHAEPLASTIDA
RHIZARIA
CHROMALVEOLATA

Figure 3.3a. Most Parsimonious Tree of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU). The tree was inferred using TNT from both weighted and unweighted analyses. Weighted analyses had a 2:1 transitions to transversions ratio. Organisms are color-coded as in Fig. 3.1. Tree score = 40678; CI = 0.272; RI = 0.421.

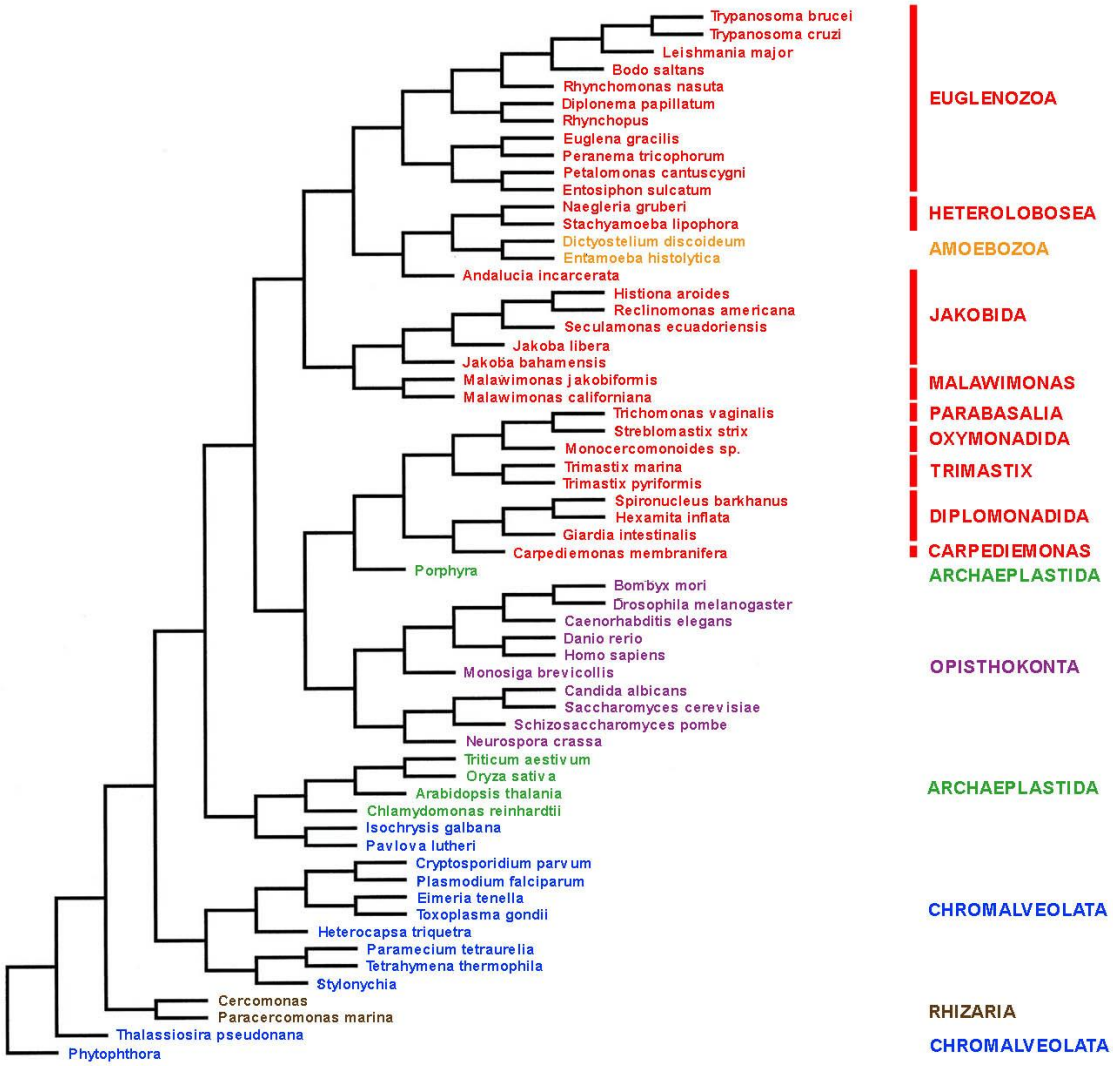

Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Diplonema papillatum
Rhynchopus
Euglena gracilis
Peranema tricophorum
Petalomonas cantuscygni
Entosiphon sulcatum
Naegleria gruberi
Stachyamoeba lipophora
Dictyostelium discoideum
Entamoeba histolytica
Andalucia incarcerata
Histiona aroides
Reclinomonas americana
Seculamonas ecuadoriensis
Jakoba libera
Jakoba bahamensis
Malawimonas jakobiformis
Malawimonas californiana
Trichomonas vaginalis
Streblomastix strix
Monocercomonoides sp.
Trimastix marina
Trimastix pyriformis
Spironucleus barkhanus
Hexamita inflata
Giardia intestinalis
Carpediemonas membranifera
Porphyra
Bombyx mori
Drosophila melanogaster
Caenorhabditis elegans
Danio rerio
Homo sapiens
Monosiga brevicollis
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Triticum aestivum
Oryza sativa
Arabidopsis thalania
Chlamydomonas reinhardtii
Isochrysis galbana
Pavlova lutheri
Cryptosporidium parvum
Plasmodium falciparum
Eimeria tenella
Toxoplasma gondii
Heterocapsa triquetra
Paramecium tetraurelia
Tetrahymena thermophila
Stylonychia
Cercomonas
Paracercomonas marina
Thalassiosira pseudonana
Phytophthora
EUGLENOZOA
HETEROLOBOSEA
AMOEBOZOA
JAKOBIDA
MALAWIMONAS
PARABASALIA
OXYMONADIDA
TRIMASTIX
DIPLOMONADIDA
CARPEDIEMONAS
ARCHAEPLASTIDA
OPISTHOKONTA
ARCHAEPLASTIDA
CHROMALVEOLATA
RHIZARIA
CHROMALVEOLATA

Figure 3.3b. MP Tree of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU) with 1000 BS Replicates. The tree was inferred using TNT from both weighted and unweighted analyses. Weighted analyses had a 2:1 transitions to transversions ratio. Support values are displayed at the nodes with the BS values for the unweighted analysis followed by the BS values for the weighted analysis. Black circles on branches indicate full support for the node. Organisms are color-coded as in Fig. 3.1. Tree score = 40679; CI = 0.270; RI = 0.416.

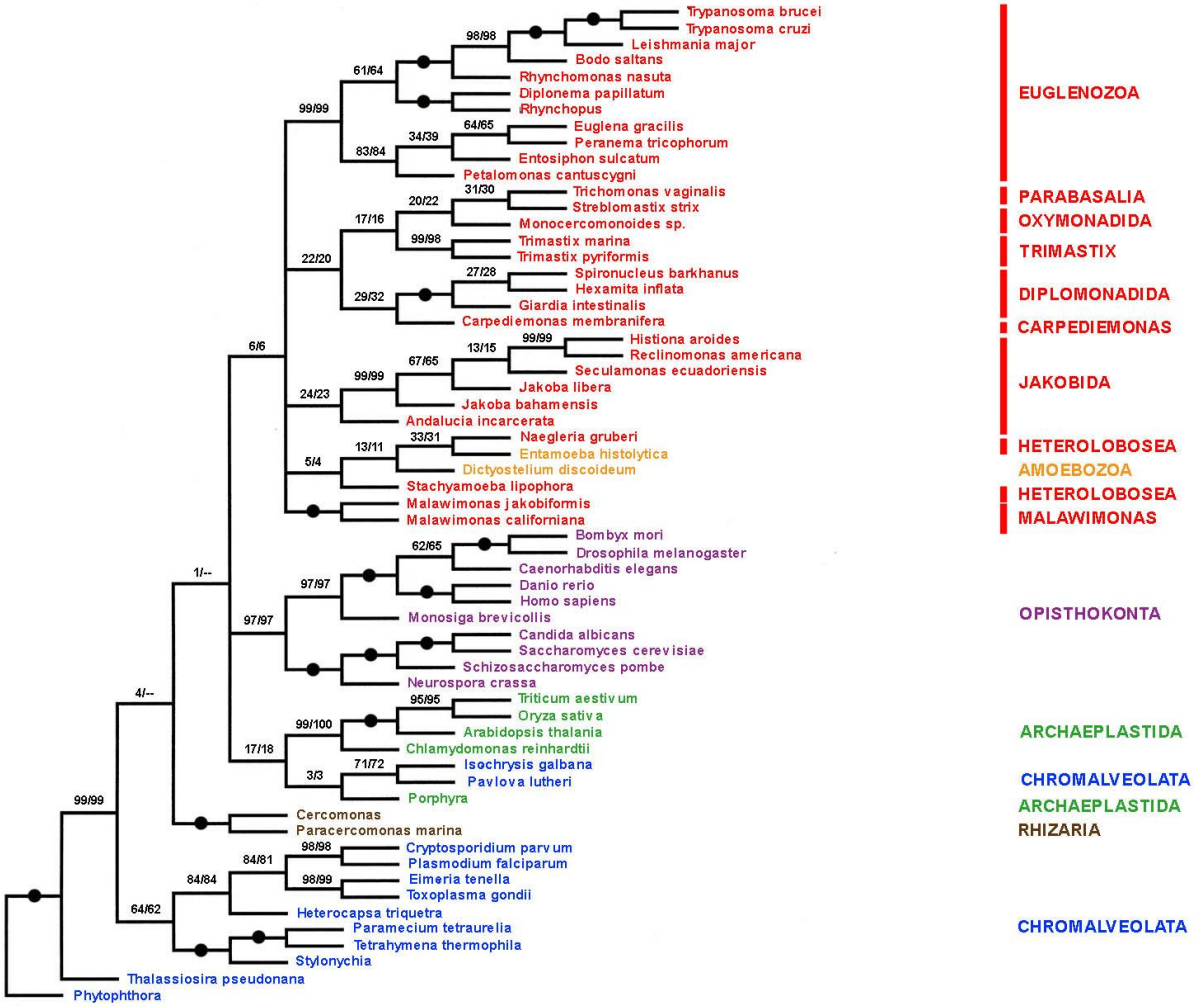

Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Diplonema papillatum
Rhynchopus
Euglena gracilis
Peranema tricophorum
Entosiphon sulcatum
Petalomonas cantuscygni
Trichomonas vaginalis
Streblomastix strix
Monocercomonoides sp.
Trimastix marina
Trimastix pyriformis
Spironucleus barkhanus
Hexamita inflata
Giardia intestinalis
Carpediemonas membranifera
Histiona aroides
Reclinomonas americana
Seculamonas ecuadoriensis
Jakoba libera
Jakoba bahamensis
Andalucia incarcerata
Naegleria gruberi
Entamoeba histolytica
Dictyostelium discoideum
Stachyamoeba lipophora
Malawimonas jakobiformis
Malawimonas californiana
Bombyx mori
Drosophila melanogaster
Caenorhabditis elegans
Danio rerio
Homo sapiens
Monosiga brevicollis
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Triticum aestivum
Oryza sativa
Arabidopsis thalania
Chlamydomonas reinhardtii
Isochrysis galbana
Pavlova lutheri
Porphyra
Cercomonas
Paracercomonas marina
Cryptosporidium parvum
Plasmodium falciparum
Eimeria tenella
Toxoplasma gondii
Heterocapsa triquetra
Paramecium tetraurelia
Tetrahymena thermophila
Stylonychia
Thalassiosira pseudonana
Phytophthora
98/98
61/64
99/99
64/65
34/39
83/84
31/30
20/22
17/16
99/98
22/20
27/28
29/32
6/6
99/99
67/65
13/15
99/99
24/23
33/31
13/11
5/4
62/65
97/97
1/--
97/97
95/95
99/100
17/18
71/72
3/3
4/--
99/99
98/98
84/81
98/99
84/84
64/62
EUGLENOZOA
PARABASALIA
OXYMONADIDA
TRIMASTIX
DIPLOMONADIDA
CARPEDIEMONAS
JAKOBIDA
HETEROLOBOSEA
AMOEBOZOA
HETEROLOBOSEA
MALAWIMONAS
OPISTHOKONTA
ARCHAEPLASTIDA
CHROMALVEOLATA
ARCHAEPLASTIDA
RHIZARIA
CHROMALVEOLATA

Figure 3.4a. Most Parsimonious Tree of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU) after Removal of *N. gruberi*. The tree was inferred using TNT from both weighted and unweighted analyses. Weighted analyses had a 2:1 transitions to transversions ratio. Organisms are color-coded as in Fig. 3.1. Tree score = 39419; CI = 0.279; RI = 0.427.

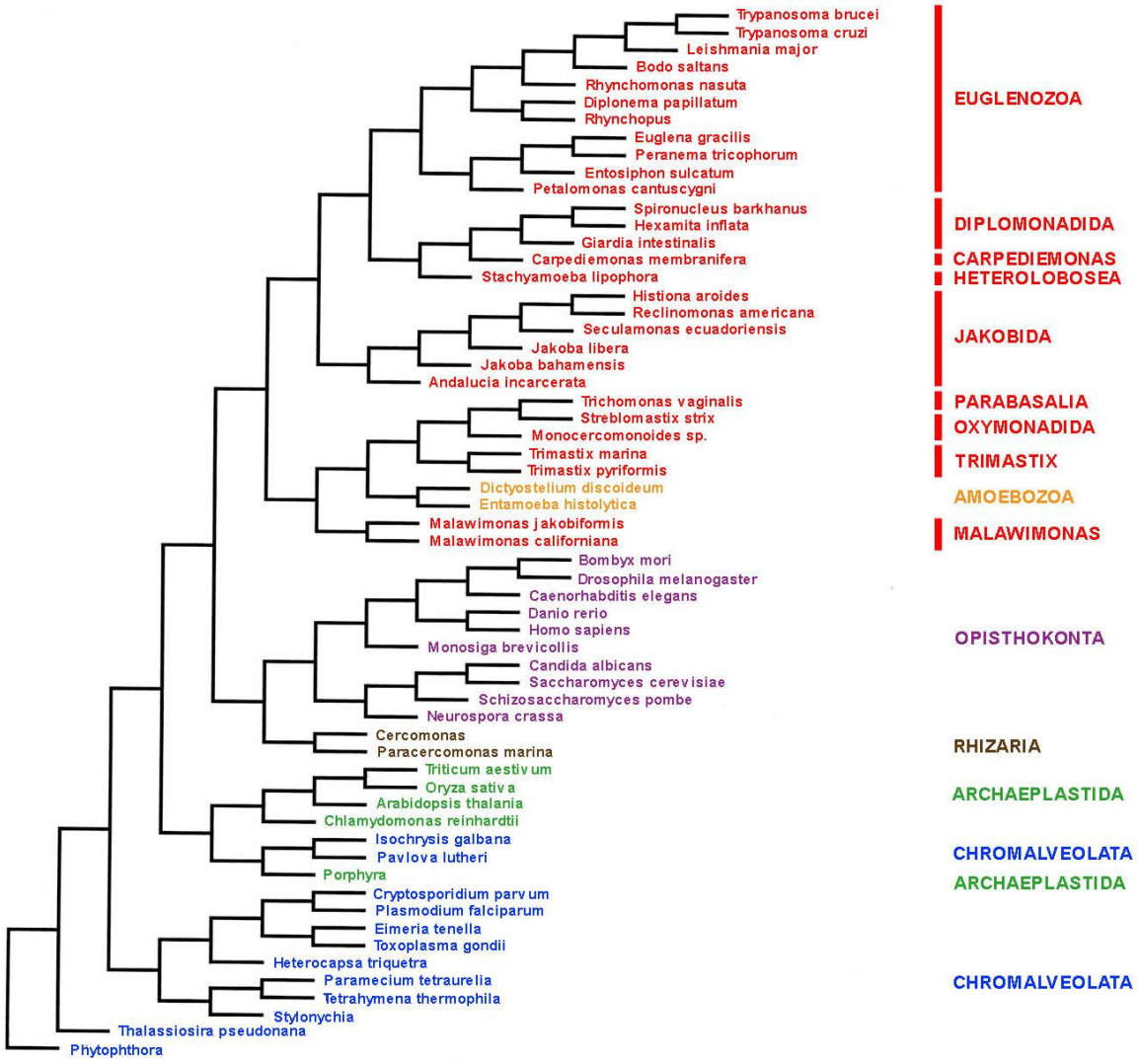

Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Diplonema papillatum
Rhynchopus
Euglena gracilis
Peranema tricophorum
Entosiphon sulcatum
Petalomonas cantuscygni
Spironucleus barkhanus
Hexamita inflata
Giardia intestinalis
Carpediemonas membranifera
Stachyamoeba lipophora
Histiona aroides
Reclinomonas americana
Seculamonas ecuadoriensis
Jakoba libera
Jakoba bahamensis
Andalucia incarcerata
Trichomonas vaginalis
Streblomastix strix
Monocercomonoides sp.
Trimastix marina
Trimastix pyriformis
Dictyostelium discoideum
Entamoeba histolytica
Malawimonas jakobiformis
Malawimonas californiana
Bombyx mori
Drosophila melanogaster
Caenorhabditis elegans
Danio rerio
Homo sapiens
Monosiga brevicollis
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Cercomonas
Paracercomonas marina
Triticum aestivum
Oryza sativa
Arabidopsis thalania
Chlamydomonas reinhardtii
Isochrysis galbana
Pavlova lutheri
Porphyra
Cryptosporidium parvum
Plasmodium falciparum
Eimeria tenella
Toxoplasma gondii
Heterocapsa triquetra
Paramecium tetraurelia
Tetrahymena thermophila
Stylonychia
Thalassiosira pseudonana
Phytophthora
EUGLENOZOA
DIPLOMONADIDA
CARPEDIEMONAS
HETEROLOBOSEA
JAKOBIDA
PARABASALIA
OXYMONADIDA
TRIMASTIX
AMOEBOZOA
MALAWIMONAS
OPISTHOKONTA
RHIZARIA
ARCHAEPLASTIDA
CHROMALVEOLATA
ARCHAEPLASTIDA
CHROMALVEOLATA

Figure 3.4b. MP Tree of Concatenated Data Set 2 (Third Codon Removed Alignments + SSU and LSU) with 1000 BS Replicates after Removal of *N. gruberi*. The tree was inferred using TNT from both weighted and unweighted analyses. Weighted analyses had a 2:1 transitions to transversions ratio. Support values are displayed at the nodes with the BS values for the unweighted analysis followed by the BS values for the weighted analysis. Black circles on branches indicate full support for the node. Organisms are color-coded as in Fig. 3.1. Tree score = 39419; CI = 0.276; RI = 0.420.

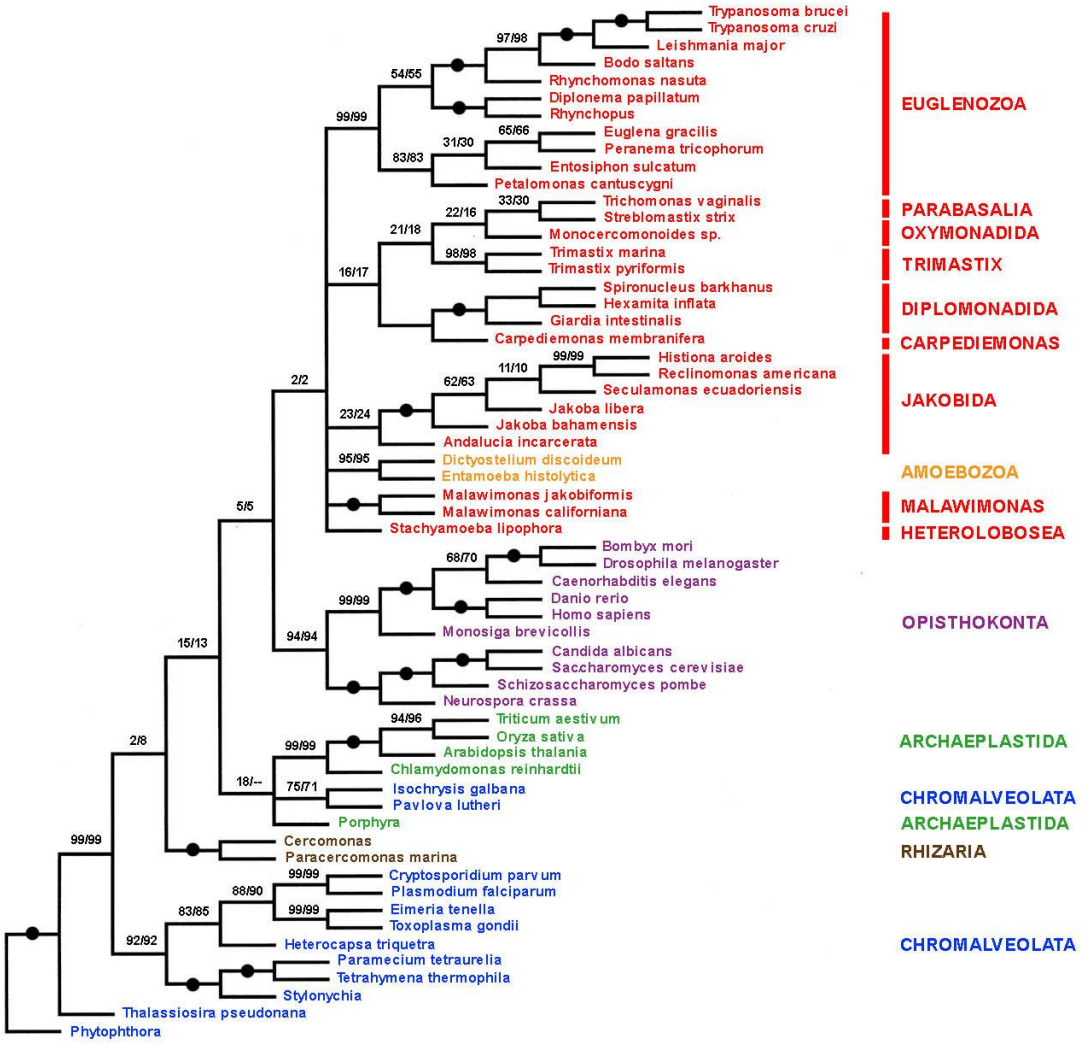
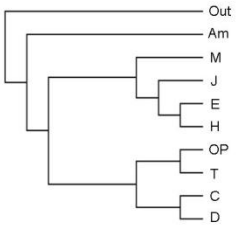
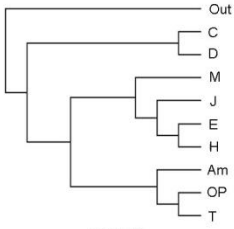

Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Bodo saltans
Rhynchomonas nasuta
Diplonema papillatum
Rhynchopus
Euglena gracilis
Peranema tricophorum
Entosiphon sulcatum
Petalomonas cantuscygni
Trichomonas vaginalis
Streblomastix strix
Monocercomonoides sp.
Trimastix marina
Trimastix pyriformis
Spironucleus barkhanus
Hexamita inflata
Giardia intestinalis
Carpediemonas membranifera
Histiona aroides
Reclinomonas americana
Seculamonas ecuadoriensis
Jakoba libera
Jakoba bahamensis
Andalucia incarcerata
Dictyostelium discoideum
Entamoeba histolytica
Malawimonas jakobiformis
Malawimonas californiana
Stachyamoeba lipophora
Bombyx mori
Drosophila melanogaster
Caenorhabditis elegans
Danio rerio
Homo sapiens
Monosiga brevicollis
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Neurospora crassa
Triticum aestivum
Oryza sativa
Arabidopsis thalania
Chlamydomonas reinhardtii
Isochrysis galbana
Pavlova lutheri
Porphyra
Cercomonas
Paracercomonas marina
Cryptosporidium parvum
Plasmodium falciparum
Eimeria tenella
Toxoplasma gondii
Heterocapsa triquetra
Paramecium tetraurelia
Tetrahymena thermophila
Stylonychia
Thalassiosira pseudonana
Phytophthora
97/98
54/55
99/99
65/66
31/30
83/83
33/30
22/16
21/18
98/98
16/17
99/99
11/10
62/63
2/2
23/24
95/95
5/5
68/70
99/99
94/94
15/13
94/96
99/99
2/8
18/--
75/71
99/99
88/90
99/99
83/85
92/92
EUGLENOZOA
PARABASALIA
OXYMONADIDA
TRIMASTIX
DIPLOMONADIDA
CARPEDIEMONAS
JAKOBIDA
AMOEBOZOA
MALAWIMONAS
HETEROLOBOSEA
OPISTHOKONTA
ARCHAEPLASTIDA
CHROMALVEOLATA
ARCHAEPLASTIDA
RHIZARIA
CHROMALVEOLATA

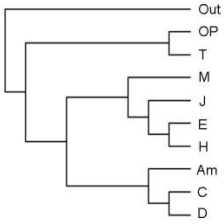
Figure 3.5. Alternate Topologies for Approximately Unbiased (AU) Topology Test. Out = outgroups from root of tree to Opisthokonta in the ML Tree in Fig. 3.2a; Am = Amoebozoa; C = *Carpediemonas*; D = Diplomonadida; E = Euglenozoa; H = Heterolobosea; J = Jakobida; M = *Malawimonas*; O = Oxymonadida; P = Parabasalia; T = *Trimastix*; OP = Oxymonadida and Parabasalia clade as seen in Fig 3.2a. Trees 2-18 are based on the original ML tree of Concatenated Data Set 2 (Fig 3.2a) with modifications based on MP trees (Fig 3.3 and 3.4). Tree 20 was a modified version of Fig 3.4a with *N. gruberi* grafted as sister to *S. lipophora*. AU test p-values for each topology are shown below the corresponding tree. Tree with p-values < 0.050 are rejected as statistically unlikely. Asterisks indicate trees that cannot be statistically rejected.



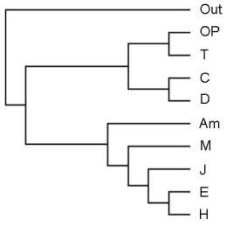
Tree 1 (ML Tree Fig. 3.2a)
p-value = 0.984*



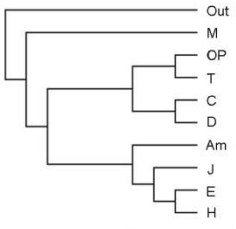
Tree 2
p-value = 0.002



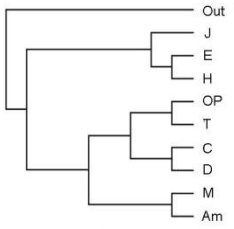
Tree 3
p-value < 0.001



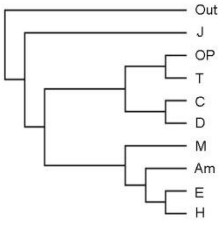
Tree 4
p-value = 0.071*



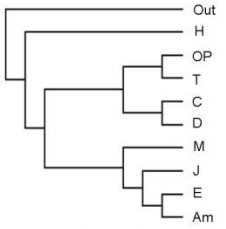
Tree 5
p-value = 0.010



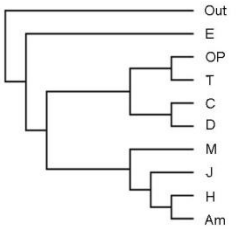
Tree 6
p-value = 0.010



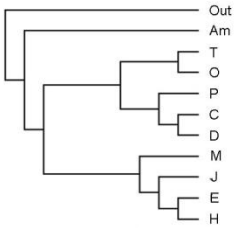
Tree 7
p-value < 0.001



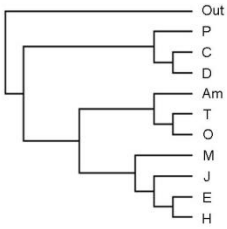
Tree 8
p-value < 0.001



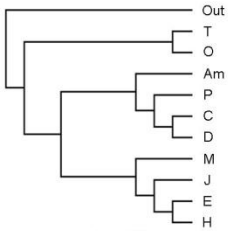
Tree 9
p-value < 0.001



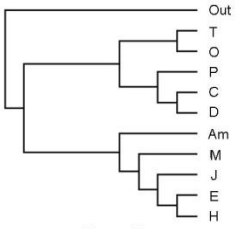
Tree 10
p-value < 0.001



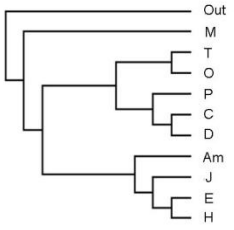
Tree 11
p-value < 0.001



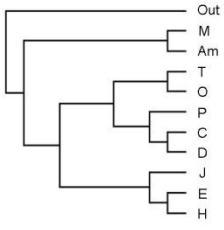
Tree 12
p-value < 0.001



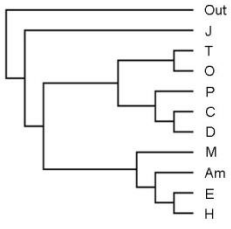
Tree 13
p-value < 0.001



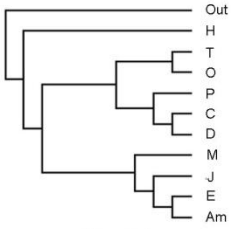
Tree 14
p-value < 0.001



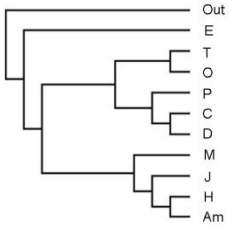
Tree 15
p-value < 0.001



Tree 16
p-value < 0.001



Tree 17
p-value < 0.001



Tree 18
p-value < 0.001

Tree 19 (MP Tree Fig 3.3a)
p-value < 0.001

Tree 20
(MP Tree Fig 3.4a modified)
p-value = 0.003

CHAPTER 4

DISCUSSION AND CONCLUSIONS

Phylogeny of Excavata

The validity of the supergroup Excavata has been highly debated over the last few years due to the lack of support for monophyly of this group based solely on molecular data. This study has recovered monophyly of Excavata with moderate to strong support values using several different methods. Because several of the analyses did not recover monophyly of Excavata, AU tests were performed on alternate topologies based on the conflicting trees. Two topologies could not be statistically rejected using this method. A monophyletic Excavata was seen in the topology with the strongest support, although paraphyly of the excavates could not be ruled out. Based on these analyses, the supergroup Excavata seems to be supported as a whole.

One major relationship recovered within the excavates in these analyses was unexpected. In previous analyses, the parabasalids were recovered basal to a clade of diplomonads and *Carpediemonas* (Group 1 excavates) (Hampl *et al.* 2009, Simpson *et al.* 2006, Simpson *et al.* 2008). This relationship was seen in the analysis of Concatenated Data Set 1, but in analyses of Concatenated Data Set 2, the parabasalid, *T. vaginalis*, was consistently recovered as sister to the oxymonad, *S. strix*, preventing a monophyletic Oxymonadida. Although this result does not agree with the findings of previous studies, the metamonads (Group 1/2) were still recovered as a monophyletic clade. The addition of other taxa from these groups to future analyses could result in monophyly of the oxymonads. Also, this result makes more intuitive sense based on the morphological characteristics of these organisms. Oxymonads, parabasalids, and euglenozoans do not have the ventral feeding groove seen in the other excavate groups. With the relationship of the oxymonads and the parabasalids seen in this analysis, loss of the ventral feeding

groove would only need to have occurred twice: once for the euglenozoans and once for the oxymonad/parabasalid clade. This is a more parsimonious explanation of the loss of the ventral feeding groove than could be explained by the previously recovered relationships with separation between all three of these groups.

Phylogeny of Outgroups

Other unexpected relationships were seen throughout the trees. Although monophyly of the amoebozoans and the opisthokonts were consistently recovered, these two supergroups were rarely recovered as sister to each other. This result is not consistent with the unikont theory, which has been supported by previous analyses (Cavalier-Smith 2003, Hampl *et al.* 2009, Simpson *et al.* 2006, Simpson *et al.* 2008). Also, the haptophytes were consistently recovered within the Archaeplastida between the Viridiplantae and the Rhodophyta. Although this relationship does not agree with the chromalveolate hypothesis, it was also previously reported in analyses with these groups (Hampl *et al.* 2009). Another unexpected relationship recovered was the rhizarians within the chromalveolates as sister to the stramenopiles, although this relationship has been seen in some recent studies (Tekle *et al.* 2009). Although there appear to be some unexpected relationships among the outgroups, all of these relationships have been reported in previous analyses.

Efficacy of Selected Genes

The genes used in this study were selected based on their ubiquitous presence in eukaryotes and their relatively conserved natures across the Eukaryotic Tree of Life. Based on the single gene trees for each of the genes, these genes should not be used in single gene trees for analysis of the deeper relationships within the Eukaryotic Tree of Life. However, these genes should be useful in further multi-gene phylogenetic studies. Based on the single gene jackknifing, three genes in particular are useful for resolving these deep level relationships: cHSP70, SSU, and LSU. The removal of each of these genes

from the data set led to strange topologies, in which well-established clades exhibited uncharacteristic relationships.

Third Codon Removal

Highly conserved data is most useful for resolving the deeper relationships within the Eukaryotic Tree of Life. For previous studies on Excavata, amino acid data sets have typically been used (Hampl *et al.* 2009, Simpson *et al.* 2006, Simpson *et al.* 2008). However, SSU and LSU, which have been widely used in phylogenetic analyses, are not protein-coding genes. In order to efficiently use these genes in combination with the protein-coding genes, the data set needed to be nucleotide-based rather than amino acid-based. The initial data set was a concatenation of the full alignments of these genes, but the resultant topologies from the analyses of this data set had very unexpected relationships. The strange relationships could have been due to homoplasious signal in the third codon position. The majority of substitutions in the first and second codon position are nonsynonymous (i.e., substitutions that change the amino acid). However, most substitutions for the third codon position are synonymous (i.e., substitutions that do not change the amino acid). These sites, in which synonymous substitutions are more likely, have higher evolutionary rates than those that are limited by selection at the amino acid level and can saturate with multiple substitutions (Graur and Li 2000). Codon usage preferences for different organisms are also seen in the third codon position. These characteristics of the third codon position could possibly lead to strongly misleading signal in the data, which would explain the strange relationships recovered in the analysis of the full alignments. In general, the first and second codon positions should be more conserved due to the higher possibility of nonsynonymous changes, although synonymous changes can occur at these positions for Leucine, Serine, and Arginine, each of which are encoded by six different codons (Inagaki *et al.* 2004). The analyses of Concatenated Data Set 2, which included only these conserved codon positions for the protein-coding genes, seemed to be appropriate for the analysis of the deeper nodes of the Eukaryotic Tree of Life.

Gap Coding

SSU and LSU were aligned indirectly based on their secondary structures. Gap coding was performed on these alignments to determine if morphological data from these gaps could be useful in supporting any of the relationships within the excavates. The analysis of the gap coding matrices found no gaps that supported only one supergroup, including Excavata, which indicates that gap coding would not be useful for future analyses of deep level relationships. In addition, because no deeper level relationships were supported by individual coded gaps, this data set was not concatenated with the nucleotide data sets to avoid the use of data inappropriate for the current study. However, several genus- and species-level relationships were supported by specific gaps. Gap coding could potentially be useful for future analyses of aimed at these levels.

Conclusions

The initial hypotheses for this study predicted that there would be one of three final outcomes: 1) the supergroup Excavata recovered as a monophyletic clade, 2) two of the three previously recovered groups of excavates recovered as a monophyletic clade, or 3) each of the three previously recovered groups of excavates recovered as separate monophyletic clades. The first hypothesis, recovery of the supergroup Excavata as monophyletic, is reasonably well supported by the majority of the analyses performed in this study, although the second hypothesis could not be completely ruled out. This result supports the original grouping of the excavates based largely on morphological characteristics. Within Excavata, most relationships among the excavates were as expected from previous studies with the exception of the recovery of the parabasalids as most closely related to the oxymonads rather than as basal to the Group 1 excavates. This relationship provides a more parsimonious explanation for the loss of the ventral feeding groove in the euglenozoans, parabasalids, and oxymonads than previous findings. Future studies could be improved by the use of more genes that are relatively conserved throughout the Eukaryotic Tree of Life and by the addition of more ingroup taxa, especially parabasalids, oxymonads, and retortamonads.

REFERENCES

Adl S, Simpson AG, and Farmer MA *et al.* (2005) "The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists." *J. Eukaryot. Microbiol*. **52(5):** 399–451.

Ahner A, Whyte FM, and Brodsky JL. (2005) "Distinct but overlapping functions of Hsp70, Hsp90, and an Hsp70 nucleotide exchange factor during protein biogenesis in yeast." *Archives of Biochemistry and Biophysics* **435:**32–41.

Altekar G, Dwarkadas S, Huelsenbeck JP, and Ronquist F. (2004) "Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference." *Bioinformatics* **20:**407-415.

Baldauf SL and Palmer JD. (1993) "Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins." *Proceedings of the National Academy of Sciences, USA* **90:**11558–11562.

Baldauf SL. (2003) "The deep roots of eukaryotes." *Science* **300(5626):**1703-6.

Bhattacharya D and Weber K. (1997) "The actin gene of the glaucocystophyte *Cyanophora paradoxa*: analysis of the coding region and introns, and an actin phylogeny of eucaryotes." *Current Genetics* **31:**439–446.

Bhattacharya D, Weber K, An SS, and Berning-Koch W. (1998) "Actin phylogeny identifies *Mesostigma viride* as a flagellate ancestor of the land plants." *J. Mol. Evol.* **47:** 544–550.

Bhattacharya D, Aubry J, Twait EC, Jurk S. (2000) "Actin gene duplication and the evolution of morphological complexity in land plants.**"** *J Phycol* **36(5):**813-820.

Breglia SA, Slamovits CH, and Leander BS. (2007) "Phylogeny of Phagotrophic Euglenids (Euglenozoa) as Inferred from Hsp90 Gene Sequences." *Journal of Eukaryotic Microbiology* **54(1):**86-92.

Busse I and Preisfeld A. (2002) "Phylogenetic position of Rhynchopus sp. and Diplonema ambulator as indicated by analyses of euglenozoan small subunit ribosomal DNA." *Gene* **284**(1-2): 83-91.

Busse I and Preisfeld A. (2003) "Application of spectral analysis to examine phylogenetic signal among euglenid SSU rDNA data sets (Euglenozoa)." *Organisms Diversity & Evolution* **3**(1): 1-12.

Castlejohn C, Leebens-Mack J, and Farmer MA. (2008) "Use of the covarion model of evolution to resolve relationships within the Excavates." Poster presented at Protist2008 conference.

Castresana J. (2000) “Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.” *Molecular Biology and Evolution* **17:**540-552.

Cavalier-Smith T. (2003) “Protist phylogeny and the high-level classification of Protozoa.” *Europ. J. Protistol.* **39:**338–348.

CDC website (Division of Parasitic Diseases). http://www.cdc.gov/ncidod/dpd.htm (Retrieved 9/3/08).

Daubin V, Lerat E, Perriere G. (2003). “The source of laterally transferred genes in bacterial genomes.” *Genome Biol* **4**(9)**:** R57.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, and Gascuel O. (2008) “Phylogeny.fr: robust phylogenetic analysis for the non-specialist.” *Nucleic Acids Res*. **36**(Web Server issue):W465-9. Epub.

Drouin G, Moniz de Sá M, and Zuker M. (1996) “The *Giardia lamblia* actin gene and the phylogeny of eukaryotes.” *J. Mol. Evol.* **41(6):** 841–849.

Edgar RC. (2004a) “MUSCLE: a multiple sequence alignment method with reduced time and space complexity.” *BMC Bioinformatics* **5:**113.

Edgar RC. (2004b) ”MUSCLE: multiple sequence alignment with high accuracy and high throughput.” *Nucleic Acids Research* **32(5):**1792-1797.

Goloboff P, Farris S, and Nixon K. (2000) “TNT (Tree analysis using New Technology) ver. 1.1.” Published by the authors, Tucumán, Argentina.

Graur D and Li W. (2000) “Rates and patterns of nucleotide substitution.” Fundamentals of molecular evolution. Sinauer, Sunderland, Massachusetts:99-164.

Gray MW, Burger G, and Lang BF. (1999) “Mitochondrial evolution.” *Science* **283:**1476–1481.

Gray MW, Lang BF, and Burger G. (2004) “Mitochondria of protists.” *Annu. Rev. Genet.* **38:**477–524.

Hampl V, Horner DS, Dyal P, Kulda J, Flegr J, Foster P, and Embley TM. (2005) “Inference of the phylogenetic position of oxymonads based on nine genes: support for Metamonada and Excavata.” *Mol. Biol. Evol.* **22:**2508–2518.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, and Roger AJ. (2009) “Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”.” *PNAS* **106(10):**3859-3864.

Harrison PJ, Waters RE, *et al.* (1980). "A Broad-Spectrum Artificial Seawater Medium for Coastal and Open Ocean Phytoplankton." Journal of Phycology **16**(1): 28-35.

Hennessey ES, Drummond DR, and Sparrow JC. (1993) "Molecular genetics of actin function." **291**: 657-671.

Hirata K, Kusaba M, and Chuma I *et al.* (2007) "Speciation in *Pyricularia* inferred from multilocus phylogenetic analysis." *Mycological Res* **111(7):**799-808.

Huelsenbeck JP, Ronquist F, Nielsen R, and Bollback JP. (2001) "Bayesian inference of phylogeny and its impact on evolutionary biology." *Science* **294:**2310-2314.

Inagaki Y, Simpson AGB, Dacks JB, and Roger AJ. (2004) "Phylogenetic Artifacts Can Be Caused by Leucine, Serine, and Arginine Codon Usage Heterogeneity: Dinoflagellate Plastid Origins as a Case Study." *Systematic Biology* **53(4):**582-593.

Jardeleza, Sarah. (2007) "Examining the Euglenophyte Mucilaginous Clade with EF1α." Poster, University of Georgia.

Karabinos A and Bhattacharya D (2000). "Molecular evolution of calmodulin and calmodulin-like genes in the cephalochordate Branchiostoma." *J Mol Evol* **51:**141–148.

Keeling PJ and Doolittle WF. (1996) "A non-canonical genetic code in an early diverging eukaryotic lineage**."** *EMBO J* **15(9):**2285-2290.

Keeling P and Inagaki Y. (2004) "A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1α." *PNAS* **101:**15380-15385.

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, and Gray MW. (1997) "An ancestral mitochondrial DNA resembling a eubacterial genome in miniature." *Nature* **387:**493–497.

Ma B. (2005) "Phylogeny of deep-level relationships within Euglenozoa based on combined small subunit and large subunit ribosomal DNA sequences." Thesis for Masters degree, University of Georgia.

Müller KF. (2005) "SeqState - primer design and sequence statistics for phylogenetic DNA data sets." *Applied Bioinformatics* **4**:65-69.

Nixon KC. (1999) "Winclada (BETA) ver. 0.9.9." PUBLISHED BY THE AUTHOR, ITHACA, NY**.**

Page RDM. (1996) "TREEVIEW: An application to display phylogenetic trees on personal computers." *Computer Applications in the Biosciences* **12**:357-358.

Posada D and Crandall KA. (1998) "Modeltest: testing the model of DNA substitution." *Bioinformatics* **14 (9):**817-818.

Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, and Glöckner FO**.** (2007) "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." *Nuc. Acids Res.* **35(21):**7188-7196.

Ronquist F and Huelsenbeck JP. (2003) "MRBAYES 3: Bayesian phylogenetic inference under mixed models." *Bioinformatics* **19:**1572-1574.

Sanderson MJ and Donoghue MJ. (1989)"Patterns of variation in levels of homoplasy." Evolution **43(8):**1781-1795.

Schutze J, Krasko A, Custodio MR, Efremova SM, Muller IM, and Muller WEG. (1999) "Evolutionary Relationships of Metazoa within the Eukaryotes Based on Molecular Data from Porifera." *Proceedings: Biological Sciences* **266(1414):**63-73.

Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol. ,* **51**, 492-508 (2002).

Shimodaira H and Hasegawa M. (2001) "CONSEL: for assessing the confidence of phylogenetic tree selection." *Bioinformatics* **17:**1246-1247.

Simpson AGB. (2003) "Cytoskeletal organisation phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota)." *Int. J. Syst. Evol. Microbiol.* **53:**1759–1777.

Simpson AGB, Inagaki Y, and Roger AJ. (2006) "Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes." *Mol. Biol. Evol.* **23**:615–625.

Simpson AGB, Perley TA, and Lara E. (2008) "Lateral transfer of the gene for a widely used marker, [alpha]-tubulin, indicated by a multi-protein study of the phylogenetic position of Andalucia (Excavata)." *Molecular Phylogenetics and Evolution* **47(1):**366-377.

Simpson AGB and Roger AJ. (2004) "Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa." *Molecular Phylogenetics and Evolution* **30(1):**201-212.

Stamatakis A, Hoover P, and Rougemont J. (2008) "A Fast Bootstrapping Algorithm for the RAxML Web-Servers." *Systematic Biology* **57(5):**758-771.

Stamatakis A, Ott M, and Ludwig T. (2005) "RAxML-OMP: An Efficient Program for Phylogenetic Inference on SMPs." *In Proceedings of 8th International Conference on Parallel Computing Technologies (PaCT2005), Lecture Notes in Computer Science*, **3506:**288-302, Springer Verlag.

Stechmann A and Cavalier-Smith T. (2003) "Phylogenetic analysis of eukaryotes using heat-shock protein Hsp90." *Journal of Molecular Evolution* **57(4):**408-419.

Swofford DL. (2003) "PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4." Sinauer Associates, Sunderland, Massachusetts.

Talavera G and Castresana J. (2007) "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments." *Systematic Biology* **56:**564-577.

Tanabe Y, Saikawa M, Watanabe MM, and Sugiyama J. (2004) "Molecular phylogeny of Zygomycota based on EF-1α and RPB1 sequences: limitations and utility of alternative markers to rDNA." *Molecular Phylogenetics and Evolution* **30**:438-449.

Tekle YI, Wegener Parfrey L, and Katz LA. (2009) "Molecular Data Are Transforming Hypotheses on the Origin and Diversification of Eukaryotes." *BioScience* **59**:471–481.

Tuffley C and Steel M. (1998) "Modeling the covarion hypothesis of nucleotide substitution." *Mathematical Biosciences* **147:**63-91.

Varga J, Frisvad JC, and Samson RA. (2007) "Polyphasic taxonomy of Aspergillis section Candidi based on molecular, morphological, and physiological data."*Studies in Mycology* **59:**75-88.

Yamamoto A, Hashimoto T, Asaga E, Hasegawa M, and Goto N. (1997) "Phylogenetic position of the mitochondrion-lacking protozoan *Trichomonas tenax*, based on amino acid sequences of elongation factors 1α and 2*." J Mol Evol* **144:**98–105.