2014

STRUCTURE, FUNCTION AND EVOLUTION OF PHOSPHOPROTEIN P0 AND ITS UNIQUE INSERT IN *TETRAHYMENA THERMOPHILA*

Giovanni Pagano
University of Rhode Island, giovanni_pagano@my.uri.edu

STRUCTURE, FUNCTION AND EVOLUTION OF PHOSPHOPROTEIN P0 AND ITS UNIQUE INSERT IN *TETRAHYMENA THERMOPHILA*

BY

GIOVANNI PAGANO

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

BIOLOGICAL AND ENVIRONMENTAL SCIENCES

UNIVERSITY OF RHODE ISLAND

2014

MASTER OF SCIENCE

OF

GIOVANNI PAGANO

APPROVED:

Thesis Committee:

Major Professor Linda A. Hufnagel

Lenore M. Martin

Roberta King

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2014

ABSTRACT

Phosphoprotein P0 is a highly conserved ribosomal protein that forms the central scaffold of the large ribosomal subunit's "stalk complex", which is necessary for recruiting protein elongation factors to the ribosome. Evidence in the literature suggests that P0 may be involved in diseases such as malaria and systemic lupus erythematosus. We are interested in the possibility that the P0 of the "ciliated protozoa" *Tetrahymena thermophila* may be useful as a model system for vaccine research and drug development. In addition, the P0s of *T. thermophila* and other ciliated protozoans contain a 15 to 17 amino acid long insert, unique to the *N*-terminal region. This project sought to further characterize the *T. thermophila* P0 (TtP0) and its unique insert through structural and functional bioinformatics studies.

In order to visualize the three-dimensional structure of TtP0, we created a homology model of the *N*-terminal region of TtP0 and its insert from available P0 structure and sequence data. When the insert was modeled "in-context" in the presence of a previously published crystal structure of the *T. thermophila* ribosomal RNA, we discovered a surprising association between the insert and a highly variable portion of the rRNA, termed expansion segment 7, or ES7. When we investigated if this association could occur in other ciliates, we found very little data for the ES7 sequence in other species, meaning that further analysis on the conservation of this association is not possible at this time. Still, the presence of an association in *T. thermophila* may indicate that the insert has a functional role unique to the ciliates, perhaps in the regulation of P0 function by phosphorylation.

In addition, we also investigated whether the highly conserved nature of P0 meant that it could be useful for phylogenetic and evolutionary studies. By studying P0 sequences from ciliates and other closely related clades, we could determine if P0 provides any information on the early evolution of eukaryotic species. We collected P0 sequences representing all of the eukaryotic supergroups, and used them to create phylogenetic alignments and trees based on the whole molecules, as well as the individual functional domains. Overall, we found that the trees did not resolve very well at the basal branches, but terminal branches had much stronger support. The trees also successfully separated the ciliate P0 sequences into groups matching the previously established taxonomy for the ciliates. Finally, we found evidence that the *N*-terminal domain of P0, called the L10 region, is much more evolutionarily stable than the *C*-terminal 60S region. Thus, the variability of the 60S region appears to contribute to the diversity of ciliate and eukaryotic P0 sequences. Once additional P0 sequences become available for underrepresented clades, they could be used to provide stronger support for the weaker branches of the tree.

Both studies provide a starting framework for further computational-based work on P0, such as homology modeling of P0s from other ciliates or simulations of insert phosphorylation. These studies may also serve as a starting point for *in vitro* or *in vivo* experiments on the protein and its ciliate-specific insert.

ACKNOWLEDGMENTS

The student is grateful to his Major Professor, Dr. Linda A. Hufnagel, for her continued guidance, patience and support in the completion and writing of the thesis. The student would also like to thank his thesis committee members, Dr. Roberta King and Dr. Lenore M. Martin, for their instruction and advice on the tools and methods used in this thesis, as well as for their feedback in the writing and revision of the manuscripts. The student would like to acknowledge the work and help of Justin Schumacher, whose earlier work with Dr. Hufnagel provided the foundation for this project. Finally, the student would like to thank his friends and family for their continued support and encouragement during the thesis research and writing process.

PREFACE

This thesis has been prepared in accordance with the Manuscript Format as laid out by the Graduate School of the University of Rhode Island. The work within represents papers that are being prepared for publication at the time of thesis submission, and the content of the final published papers may vary from the manuscripts presented.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

MANUSCRIPT 1

The Unique *N*-terminal Insert in a Ribosomal Protein, Phosphoprotein P0, of *Tetrahymena thermophila*: Homology Modeling Analysis.

Giovanni Pagano[1], Roberta King[2], Lenore Martin[1], and Linda A. Hufnagel[1]

[1]Department of Cell and Molecular Biology, University of Rhode Island, Kingston, Rhode Island, 02881

[2]Department of Biomedical and Pharmaceutical Sciences, University of Rhode Island, Kingston, Rhode Island, 02881

Manuscript in preparation for publication

ABSTRACT

Phosphoprotein P0 is part of the stalk complex, a component of the eukaryotic large ribosomal subunit necessary for recruiting protein elongation factors. While the protein is highly conserved, our lab has noted a 15-17 amino acid long insert exclusive to the P0 of the ciliated protist *Tetrahymena thermophila,* and other ciliated protists. We hypothesized that this insert may have a function unique to *T. thermophila,* such as regulation of stalk complex function via phosphorylation of the insert. Almost no mention of this insert exists in the literature, and while the *T. thermophila* ribosome has had its structure analyzed by x-ray crystallography, it lacks a structure for the insert and provides limited data on the rest of *Tetrahymena*'s P0 (TtP0). In order to investigate the possible structure and function of the insert in TtP0, we performed several *in silico* analyses. The TtP0 sequence was used with several phosphorylation site prediction tools to detect the likelihood of phosphorylation in the insert. The TtP0 sequence was also combined with existing P0 structure and sequence data to produce a homology model of the *N*-terminal region of TtP0, including the insert. When the insert was modeled in the context of the *T. thermophila* 26S rRNA, the insert associated with a portion of the rRNA that we identified as expansion segment 7 (ES7). This suggests a potential interaction between ES7 and the insert. When the ES7 region of *T. thermophila* and that of three other ciliated protist species were compared, we found little evidence that the insert-ES7 interaction could occur in other ciliates, although more definitive analysis will require the availability of more sequenced genomes from ciliated protists. Overall, this study lays the groundwork for future *in vitro* studies to verify the presence of the insert-ES7 interaction in *T. thermophila*, and

to explore the extent to which similar interactions occur in other species of ciliated protists.

INTRODUCTION

Our laboratory recently reported that a 15-17 amino acid insert is present in the *N*-terminal region of the large subunit ribosomal protein, phosphoprotein P0, of *Tetrahymena thermophila* (TtP0) and other ciliated protists, based on analysis of genomic data (Schumacher et al, 2009, 2010a, b, c ; Schumacher and Hufnagel, MS in preparation). This insert was not present in other prokaryotes or eukaryotes examined. The insert was later also noted by Klinge et al (2011), in their crystallographic study on the *T. thermophila* large ribosomal subunit, but no mention was made of its sequence, structure or possible function. In this paper, we used homology modeling and other analyses to extend previous studies by providing a possible functional organization for the L10 region of TtP0 including the ciliate-specific insert. By developing a model of the three-dimensional shape of P0 through homology modeling, the shape of key P0 functional domains can be understood. Assuming that "form follows function", anatomical data combined with prediction programs can help in developing hypotheses about P0 function that can then be tested experimentally. Here, we report that our homology modeling analysis provides evidence that the insert forms a flexible loop that may have a novel regulatory function via an interaction with the ES7 region of 26S ribosomal RNA. We further report that the ciliate-specific insert of *T. thermophila* contains a potential serine phosphorylation site for Casein II kinases, based on analysis of the predicted protein sequence of TtP0 using phosphorylation site prediction tools (Pagni et al 2007; Blom et al, 1999).

Recently, the structures of the small and large subunits of the *T. thermophila* ribosome were solved by X-ray crystallography (Rabl et al, 2011, PDB Code 2ZXM; Klinge et al, 2011, PDB code 4A1C, 4A1D). The 60S ribosomal subunit structure contained structural coordinates for most of the ribosomal proteins, with the notable exception of the stalk proteins, which include TtP0. The crystallographic data for P0 was not clear enough to resolve its atomic structure, and instead, Klinge et al (2011) reported TtP0 structural coordinates modeled as a polyserine backbone on the L10 protein of *Thermotoga maritima* (PDB code 1zax). The TtP0 backbone model provided the structural context for the more detailed analysis reported here, allowing us to locate (superimpose) our more detailed model within the structural context of the 60S subunit. The Klinge et al model lacks essential structural coordinates for most of TtP0 and the ciliate-specific insert, as well as the amino acid sequence of the entire protein. However, the presence of the L10 backbone data contributed to our decision to create a detailed homology model of the same region in this study.

Phosphoprotein P0 (P0) is a component of the 60S subunit of the eukaryotic ribosome (Figure 1). P0 is able to combine with other phosphoproteins, P1 and P2, to form a "stalk" complex that interacts with extra-ribosomal elongation factors, namely EF-1α and EF2 in eukaryotes (EF-Tu and EF-G in prokaryotes) (Uchiumi et al, 2002). This stalk complex is part of the "GTPase-associated center", which is defined by the GTP-dependent binding of the elongation factors. The protein composition of the ribosomal stalk varies between the three domains of life, but a single P0 molecule (called L10 in eubacteria) is always present in the stalk, acting as a scaffold for other phosphoproteins, usually P1 and P2 (Gordiyenko et al, 2010). The *N*-terminal domain

of the stalk interacts with the ribosomal protein L11P (L12 in eubacteria) (Nomura et al, 2006). The stalk also forms two contacts with the 26S (23S) ribosomal RNA. The loops containing these sites have been termed the "thiostrepton loop" (H42-H44 on Klinge et al Tetrahymena model, position 1070 in *E. coli*) and the "sarcin-ricin loop" (H95 on the Klinge et al Tetrahymena model, position 2660 in *E. coli*), for the antibiotics and ribotoxins that interact with those loops in both prokaryotes and eukaryotes (Uchiumi et al, 2002).

There are several eukaryote-specific inserts in the ribosomal RNA, called expansion segments; two of these, ES7 and ES39, are located in the proximity of the stalk complex (Ben-Shem et al, 2011). Together, these expansion segments account for a large yet variable portion of the eukaryotic-specific RNA that interacts with conserved and eukaryotic-specific proteins (Wilson and Cate, 2012; Ben-Shem et al, 2011; Klinge et al, 2011). In yeast, it has been hypothesized that ES7 may interact with P0 via its tip (Ben-Shem et al, 2011), but no work has been reported that investigates this hypothesis. The Klinge et al model for *T. thermophila* contains a similar expansion segment in proximity to P0, at the tip of a helix termed ES7B (2011).

Three-dimensional models of the stalk complex have been produced from both cryo-EM maps and crystallographic data of large ribosomal subunits from several species. In addition to *T. thermophila* (PDB code 4A1C), crystal structures exist for the archaebacteria *Methanococcus jannaschii* (Kravchenko et al, 2010; PDB code 3JSY) and *Pyrococcus horikoshii* (Naganuma et al, 2010; chain G of 3A1Y), the yeast *Saccharomyces cerevisiae* (Armache et al, 2010; chain s of 3IZS) human and

Drosophila melanogaster (Anger et al, 2013; chain q of 3J3B and chain q of 3J39). The general structure of P0 based on these experiments can be found in Figure 1. The P0 sequence can be divided into three functional domains, two of which (L10 and 60s) have been identified in PFAM (http://pfam.sanger.ac.uk/) as conserved regions (Remacha et al, 1995). The L10 domain [PF00466] is located near the *N*-terminal of P0. This domain is the rRNA binding region of the stalk, and is present in all three domains of life (Liao and Dennis, 1994). The L10 region has been visualized as a five-strand beta sheet with five alpha helices surrounding it. In addition, this region contains the 15-17 aa ciliate-specific insert that is the focus of this paper.

The second domain (middle region or domain II) is found only in archaebacteria and eukaryotes. It lies between the L10 and 60S regions and is unclassified in PFAM. The middle region was first crystallized in the P0 of the archaebacteria *Methanococcus janaschii,* and consists of two alpha helices surrounded by two and three-strand beta sheets. Two flexible linkers connect this domain to the L10 domain (Kravchenko et al, 2010). It has been proposed that this region may form contacts with ribosomal protein L11, the 23S rRNA and EF2, indicating a possible role in GTPase turnover and elongation factor discrimination (Justice et al, 1999; Santos et al, 2004; Naganuma et al, 2010; Kravchenko et al, 2010). A recent cryo-EM structure of the human and *Drosophila* ribosomes with EF2 corroborates this hypothesis, as P0 contains several residues in the middle domain that form contacts with EF2 (Anger et al, 2013).

The 60S domain [PF00428] is found near the *C*-terminal end of P0; it contains the binding sites for the other P proteins, as well as a highly conserved peptide

(consensus sequence SD(D/E)DMGFGLFD) at the *C*-terminal end. This conserved peptide is shared between L10/P0 of all three domains of life, and is thought to be involved in the recruitment of elongation factors to the ribosome (Hiu-Mei Too et al, 2009; Nomura et al, 2012). In addition, a phosphorylation site has been identified in the P0 of yeast, rat and the buds of *Populus* family plants (Ballesta et al 1999, Liu et al 2010). According to these studies, phosphorylation takes place at a serine or threonine residue located a few residues before the *C*-terminal peptide. However, radioisotope labeling and electrophoresis studies of the P proteins of *Tetrahymena pyriformis* (a species distinct from but related to *T. thermophila*) failed to produce evidence of phosphorylation (c.f. Sandermann, Kruger and Kristiansen, 1979). The conserved serine or threonine residue is noted to be absent in *Tetrahymena pyriformis* P proteins, which was suggested to explain the lack of phosphorylation (Ballesta et al, 1999). The 60S region contains alpha helices that protrude from the ribosome (two in eukaryotes, three in archaebacteria) and provide space for P1 and P2 to bind, as heterodimers. The portion of the 60S region beyond the P1/P2 binding sites has yet to be crystalized successfully, likely due to its predicted flexible nature. However, NMR structures of several *C*-terminal peptides from human P proteins are available (Soares et al, 2004).

In addition to its role in protein synthesis, there is evidence suggesting that P0 may have extra-ribosomal functions. Immunocytochemical evidence suggests that P0 can locate to the cell surface in mammals, yeast and single-celled Apicomplexan parasites such as *Plasmodium falciparum* and *Toxoplasma gondii*, organisms that cause malaria and toxoplasmosis respectively (Singh et al, 2002; Sehgal et al, 2003). It was reported that *P. falciparum* parasites were exposed to monoclonal anti-P0

antibodies; their ability to infect mice was blocked, indicating that P0 may play a role in host cell invasion (Rajeshwari et al, 2004). Furthermore, when mice were injected with a highly conserved *C*-terminal domain from the P0 of *P. falciparum,* they were protected from malaria parasite invasion. In addition, when mice were immunized with a plasmid coding for an antigen derived from the P0 of *Leishmania infantum* (a non-apicomplexan parasite), they developed immunity to *Leishmania major.* Based on these findings, P0 has been proposed as a candidate for vaccine research and drug development for the control of protistan parasitic infections (Iborra et al, 2003). Antibodies against P0 and the other P proteins have also been implicated in human diseases, and could be used to detect certain diseases before the appearance of symptoms. For example, elevated levels of antibodies against the L10 region and the *C*-terminal peptide have been found in some systemic lupus erythematosis patients (Heinlen et al, 2010; Uchiumi et al, 1991). A recent study also suggested that elevated levels of the antibodies are also involved in autoimmune hepatitis, which may indicate a common targeting mechanism for both diseases (Calich et al, 2013).

Our lab is investigating the potential of the ciliate *Tetrahymena thermophila*, a eukaryotic microorganism related to the apicomplexans by virtue of their shared membership in the alveolate clade of protists, as a model for vaccine research. Our group has recently obtained immunocytochemical evidence for the location of P0 at the surface of *T. thermophila* (Schumacher et al, 2010a, b, ms in preparation).

MATERIALS AND METHODS

P0 Sequence Identification:

Phosphoprotein P0 protein sequences for eukaryotes, eubacteria and archaebacteria were obtained from NCBI and UniProt through a combination of keyword and BLAST searches. Sequences were selected to represent a wide variety of organisms from the three domains of life, with an emphasis on complete eukaryotic P0 sequences. The TtP0 protein sequence was obtained from NCBI via the Tetrahymena Genome Database website (ciliate.org). The nucleotide sequence for *Goniomonas avonlea* was generously provided by Dr. Eunsoo Kim of the American Museum of Natural History (New York, NY). The nucleotide sequence for *Stentor coeruleus* was provided by Mark Slabodnick of the Marshall lab (University of California, San Francisco) Nucleotide sequences were translated into protein sequences using the ExPASy Translate tool (http://web.expasy.org/translate/) and manually inspected to confirm the correct reading frame. Translated P0 sequences were verified through motif searches and alignments to TtP0, and regions beyond the predicted start and stop sites were removed.

Sequence Alignments:

The selected sequences (104 in total) were aligned using the MCoffee web server (Notredame et al, 2002), which combines the output of several different alignment programs into a single consensus sequence alignment. To further refine our alignments by using P0 structural data, the MCoffee alignment was combined with the coordinates of possible template structures to create an Expresso alignment using default parameters (Notredame et al, 2002). A smaller selection of eukaryotic

sequences (24 total) were aligned with ClustalW, in order to emphasize the presence of the ciliate insert (Larkin et al, 2007).

TtP0 Motif Analysis:

The TtP0 protein sequence was run through the Motif Scan tool at MyHits to detect possible motifs and functional sites on P0 and in the region of the predicted insert (Pagni et al, 2007). The protein sequence was also analyzed using NetPhos 2.0 (Blom, Gammeltoft and Brunak, 1999) and DISPHOS (Iakoucheva et al, 2004) to predict the phosphorylation potential of all serine, threonine or tyrosine residues on the protein. To predict specific kinases that might phosphorylate these same amino acids, NetPhosK with standard parameters (Blom et al, 2004) was also used on the TtP0 amino acid sequence.

Template Identification:

Possible template structural coordinate sets were identified in the Protein Data Bank (http://www.rcsb.org/pdb/). In total, six P0 or L10 structures were considered: chain q of 3J3B (*H. sapiens*), chain q of 3U5I (*S. cerevisiae*), chain A of 1ZAX (*T. maritima*), chain 5 of 4KJ9 (*E. coli*), chain G of 3A1Y (*P. horikoshii*) and chain B of 3JSY (*M. jannaschii*). In addition, the structure of TtP0 (chain G of 4A1C) was chosen as a reference to place the homology model into its approximate location on the 60S ribosomal subunit, providing structural context for later refinements of the insert. To allow for easier manipulation of the TtP0 coordinates from 4A1C, the polyserine backbone was replaced with the actual amino acids of P0 using the "Mutate Amino Acids" option in Discovery Studio. Residues were assigned based on the the Expresso sequence alignment and through visual inspection. These assignments allowed the

coordinates of the TtP0 crystal to be used as a reference for generating our homology model in the context of the full ribosome by superimposing the model template structures over the *T.thermophila* large ribosomal subunit.

Template Superimposition:

To evaluate areas of significant structural homology amongst the templates, the seven P0/L10 chains were superimposed over each other using the "Match" option of UCSF Chimera (Pettersen et al, 2004). A 5.0 Angstrom cutoff for pruning atom pairs was chosen, with the Expresso alignment from the previous step used as the input alignment. The six structures were overlaid over 4A1C with root mean square distance values ranging from 3.138-3.621 Angstroms, using between 26-30 atom pairs. Based on the high conservation of the L10 region in all of the selected structures, along with the variable amount of structural data available for other P0 regions, we decided to focus on modeling the "L10 core" region in our study (Figure 1). This region also contains the insert, which was *de novo* modeled after completion of the backbone modeling.

Backbone Modeling:

After superimposition, we determined the appropriate template for homology modeling, based on the quality of the structural data, as well as the homology of the template sequence to the target sequence. Of the six possible templates, we chose to make models from the Yeast P0 (chain q of 3U5I). This is because *S.cerevisiae* was the nearest relative to *T.thermophila* among the species considered, and because the yeast P proteins are well characterized in the literature.

The backbone of TtP0 was built using the "Build Homology Models" protocol in Discovery Studio, based on the MODELLER algorithm (Eswar et al, 2006). The following parameters were used: Protein Optimize Sidechains: True; Waters: False; Number of Models: 25; Optimization Level: None; Cut Overhangs: True; Disulfide Bridges: False; Cis-Prolines: False; Refine Loops: False.

After building 25 models, we then visually inspected them to find areas of variability in the structures, which were limited to loops between the secondary structure elements. The alpha-helices and beta-strands remained very consistent amongst all 25 models. After inspection, we manually refined the alignment between the *Tetrahymena* and yeast P0 sequences in Discovery Studio. When these adjustments were completed, we built 25 additional homology models using the same protocol as above, using the refined alignment and yeast P0 as a template.

Backbone Refinement:

The best-scoring (lowest energy) model from this second run was then refined using the "side-chain refinement" tool in Discovery Studio, based on the CHI-ROTOR algorithm and CHARMm minimization (Spassov, Yan and Flook, 2007). After refining the side-chains, the entire protein was refined using the Minimization tool of Discovery Studio. The TtP0 backbone model was minimized using multiple runs of the Steepest Descent and Conjugate Gradient protocols using a CHARMm forcefield and "Generalized Born" implicit solvent model at all steps. Minimization was repeated until the Potential Energy reached a plateau.

Following each minimization, a score was calculated for the model using the "Verify Protein Profiles-3D" option in Discovery Studio. The score did not change

much with each minimization, and stayed within the predicted range given by Profiles-3D. After all minimizations were complete, the refined backbone assessed with Discovery Studio's "Protein Health" protocol to verify the validity of the model.

Insert Modeling and Refinement:

The refined and minimized backbone of the TtP0 homology model was used as the template for *de novo* modeling of the ciliate insert. Twenty-five *de novo* models were created, using the "Build Homology Models" function under the same parameters used for the backbone. The lowest-energy model with the insert was refined using the "Smart Minimizer" minimization option, with 1000 steps of Conjugate Gradient modeling and standard parameters. Side chains were refined as above, and a second round of "Smart Minimizer" minimization was used to confirm the model's energy was at a plateau. The "Protein Health" option was used again at this point to assess the lowest-energy model and insert.

To investigate the influence of nearby components in the 26S ribosome complex, the minimized TtP0 homology model was placed into a file containing bases 551-599, 1226-1256 and 1306-1319 of the *T.thermophila* 26S rRNA (PDB code 4A1D), as well as chains F, K and X of 4A1D. The insert was minimized separately, both in and out of context of the nearby rRNA and protein chains using the "Loop Refinement" protocol, based on the LOOPER program (Spassov, Fllok and Yan, 2008) with CHARMm minimization ("CHARMm Polar H" forcefield). One hundred variations of the loop (residues 69-84) were created in the presence (in context) of the rRNA/protein chains and one hundred variations created in the absence (out of context) of the rRNA/protein chains. The quality of the models was verified at all

stages used the "Verify Protein Profiles-3D" and "Protein Health Report" options in Discovery Studio. The twenty lowest-energy conformations of the insert were chosen to demonstrate the variability of the loop conformations. For the "in-context" models, the potential for non-bonded interactions between the insert and the rRNA/protein chains was investigated using Discovery Studio's "Monitor Intermolecular H-Bonds" and "Monitor Distance" tools.

Insert Interactions:

Using the secondary structure data provided in Klinge et al (2011) along with the data gathered from the "in-context" modeling, bases potentially interacting with the insert were observed at the tip of the ES7B region of the *T.thermophila* 26S rRNA. After identifying ES7 as a potential interacting partner, we compared the sequence conservation of ES7 between different species. This would allow us to determine if a similar interaction between the ES7 region and the inserts of other ciliates was predicted. 26S rRNA sequences for eukaryotic species were collected from searches at the Comparative RNA Web Site and Project (Cannone et al, 2002) and the SILVA Ribosomal RNA Database (Quast et al, 2013; Yilmaz et al, 2014), with an emphasis on finding the rRNA sequences of ciliates. Ciliate sequences were inspected for the presence of ES7, and partial or incomplete sequences were discarded. Sequences containing multiple rRNA genes were run through RNAmmer (Lagesen et al, 2007) to trim out non-26S rRNA. In total, rRNA sequences from four species—*Tetrahymena thermophila, Tetrahymena pyriformis*, *Oxytricha trifallax* and *Paramecium tetraurelia*—were chosen and aligned with ClustalW as described under Sequence Alignments.

RESULTS

Alignments of P0:

To determine the location of the ciliate insert compared to conserved regions in P0, several alignments of the predicted protein sequences of *T. thermophila* P0 (TtP0) and its orthologues in other organisms were prepared. These included a MCoffee alignment using 104 P0 and L10 sequences from the three domains of life (Figure 2A) and a ClustalW alignment of 24 eukaryotes (Figure 2B). This included 12 sequences from ciliated protists, representing three of the eleven recognized ciliate classes (Lynn, 2002). The ClustalW alignment of predicted P0 sequences from *T. thermophila* and other ciliates demonstrates the presence of an insert of 15-17 amino acids in *Tetrahymena* sp., and inserts of similar length in other ciliate species. The insert was absent in all the non-ciliates examined. In *T. thermophila*, the insert covered residues 74 to 91.

MCoffee (without PDB structural information) and Expresso (with PDB structural information) alignments were evaluated for use in the homology modeling. Inclusion of 3D structural data improved the quality of the alignment (from 73 to 87 alignment score, out of 99). As expected, the P0 and L10 sequences exhibited strong homology in the center of the protein sequences, and weaker homology at the N-terminus, C-terminus and the site of the insert. Within the insert region, the inserts of the ciliates aligned on one side of a homologous 5 AA peptide, while the kinetoplastid inserts aligned on the opposite side (not shown). All other eukaryotes showed large gaps in this region, except for the homologous peptide.

Motif analysis of TtP0:

To predict the locations of putative functional domains and possible phosphorylation sites, four prediction tools--ExPasy MyHits, NetPhos 2.0, NetphosK and DISPHOS--were used to scan P0. The results are summarized in Figures 3 and 4.

Functional domains: MyHits Motif Scan compares an inputted sequence against several profile databases to predict motifs, phosphorylation sites, and other possible protein modifications. Within the TtP0 sequence, two characteristic P0 motifs, the L10 region and the 60S domains, were identified by MyHits, near the *N*-terminal and *C*-terminal ends respectively. The L10 domain predicted by MyHits covers residues 7 to 125, with an E-value of 6.4e-15. The 60S domain covers residues 242-323, with an E-value of 1.9e-05.

To compare the homology of the TtP0 L10 and 60S domains against other eukaryotes, two protein BLASTs were run on each domain, one against all eukaryotes and one against a selected set of species (*Homo sapiens, Saccharomyces cerevisiae, Drosophila melanogaster* and *Triticum aestivum*). In general, the strongest matches for both domains were from other ciliates, fungi, and insects. Hits against the L10 domain had overall stronger e values (ranging from 2e-58 to 6e-10) than hits against the 60S (ranging from 6e-16 to 0.027). When compared to the specific set of organisms, the best matches for L10 came from a P0–like protein called Mrt4 which shares homology with the L10 domain of P0 (2e-11; identity of 29%) and the P0 of *S. cerevisiae* (ScP0) (7e-11; identity of 28%). For the 60S, the best matches came from ScP0 (7e-07, identity: 47%). Due to the repetitive nature of the amino acid sequence

of the 60S, the BLAST algorithm ignored the *C*-terminal end of the 60S region when calculating the best matches.

Phosphorylation sites: MyHits identified five potential Casein Kinase II (CKII) phosphorylation sites [residues 24-27, 78-81, 146-149, 191-194 and 209-212] and two Protein Kinase C (PKC) phosphorylation sites [residues 162-164 and 237-239] from PROSITE profiles. One CKII site was identified in the insert, spanning Ser78 to Asp81 (Figure 3).

Netphos 2.0 is used to evaluate the phosphorylation potential (PP) of the serine, threonine and tyrosine residues in order to predict generic phosphorylation sites. A higher score indicates a stronger confidence that the residue represents a phosphorylation site, with values above 0.5 considered supportive of phosphorylation. In all, sixteen residues scored above the 0.5 cutoff. Two of these residues, Ser78 (PP=0.985) and Tyr80 (PP=0.722) were located in the insert region (Figure 4). One of these (Ser78) was also identified by MyHits. DISPHOS, a tool with a similar function to Netphos 2.0, predicted six possible sites, three threonine residues and three tyrosine residues. Three of these hits were found in the insert at Thr76, Tyr80 and Tyr 83.

Unlike the other tools, NetphosK predicts specific kinases that act on a serine, threonine or tyrosine residue, using a PP score like Netphos 2.0. NetphosK found 25 possible matches to kinases at 20 different residues, with 5 having more than one possible matching kinase. None of these residues were located in the insert. Of all the residues studied, only one, Tyr211, was predicted as a phosphorylation site by all three tools. A summary of the possible phosphorylation sites in TtP0 are included in Table 2.

Other modifications: Three N-myristoylation sites were predicted by MyHits, at residues 147-152, 270-275 and 284-289. One amidation site at residues 181-184 was also predicted. None of these predicted sites are located within the insert region.

Homology modeling of TtP0:

Because of the flexible nature of the *C*-terminal region of the stalk, the modeling studies were limited to the L10-containing *N*-terminal domain of TtP0 (residues 7-125, 203-218), with and without the insert. The quality of the models was verified by generating Protein Health reports, which included a check of the main chain conformations against a Ramachandran plot. In the lowest-energy backbone model, 99 of the 108 non-terminal (not glycine or proline) residues (91.4%) were in allowed regions, eight were in marginal regions, and one was in a disallowed region. After the insert was modeled in and before its conformation was minimized, only eleven of the 121 non-terminal residues (9.1%) were in marginal regions; three of these (ARG84, ALA88, LYS90) were located in the insert. For the out of context model of the insert, thirteen of 121 residues were in marginal regions, including three in the insert (THR75, TYR79 and TYR82) and one (LYS89) was in a disallowed region.The in-context insert of the model included no disallowed residues and many of the same marginal residues as the out-of-context model, with LYS78 and ASP80 instead of TYR79 and TYR82.

Overall, modeling the insert in the presence of the ribosome ("in context") produced a significantly different result than modeling without the ribosome ("out of context"). Out of context, the insert took on a variety of possible orientations and forms, and even formed coils in a few cases (Figure 5, A and B). In context, we

observed more constrained conformations than in the out of context models, along with what appeared to be a close association between the insert and a portion of the rRNA later identified as ES7 (Figure 5, C, D and E). While some variation was still observed, this was restricted to loop models with higher (less negative) energy scores.

Interactions between ribosomal RNA and insert of TtP0:

Once a good fit between TtP0 and the rest of the 60S subunit was achieved for both yeast-derived and *Tetrahymena*-derived ribosomes, the models were investigated further using Discovery Studio. We wanted to determine what intermolecular forces, if any, could be responsible for the close association we observed between the insert and ES7B.

Using the "Monitor" function of Discovery Studio, we identified a number of hydrogen bonds between atoms of the insert and ES7, with measured bond distances between 2 and 3 Angstroms. A diagram and summary of the H bonds predicted by the 10 lowest energy models is shown in Figure 6. Among these models, H bonds were observed between atoms from bases C584 and A 585, and residues ARG83, GLN84 and GLY 86. While the exact atoms in the H bonds varied between different insert conformations, certain atoms, like the HN of GLY86 and O2 and N3 of C584 made H bonds in several of the models. The number of times these atoms were involved in H bonds may indicate their functional importance.

We hypothesized that if the interaction between ES7B and the insert is part of a ciliate-specific regulatory mechanism for the stalk, then the residues or bases involved should be conserved in other ciliate species. To assess the validity of this hypothesis, we compared the available sequences of ES7 and the inserts of several

ciliate species to those of *T. thermophila*. In total, we found five ciliate species whose published 26S rRNA sequences contained ES7. These species are *T thermophila, T. pyriformis, Spathidium amorphiformae, Oxytricha trifallax* and *Paramecium tetraurelia*. Two species closely related to *T. thermophila, T. pyriformis and S. amorphiformae,* both contain 26S rRNA bases homologous to C584 and A585, but a published P0 sequence is not available for either species. Three ciliate species have both a published P0 sequence and a published 26S rRNA sequence with a complete ES7 region: *T. thermophila, O. trifallax* and *P. tetraurelia*. While *T. pyriformis* lacks a published P0 sequence, the ES7 region from its 26S rRNA was included in the comparison because it could provide information about the conservation of ES7 among closely related species. The inserts were compared by visual inspection (Table 1), and the ES7 sequences were compared by a ClustalW sequence alignment (Figure 7).

The Tetrahymena species showed strongly homologous sequences for ES7. The bases we observed interacting with the insert in the homology model (C584 and A585) were found to be conserved in *T. pyriformis*. While no P0 sequence has been published for *T. pyriformis*, the predicted P0 sequences for three other *Tetrahymena* species are available (Table 1). These predicted insert sequences are highly conserved, and the residues that interact with ES7 were observed to be present in all four species. However, the insert and ES7 sequences in *O. trifallax* both differed from those of the two *Tetrahymena* species. Furthermore, the *P. tetraurelia* insert and ES7 sequences showed the largest divergences from those of *T. thermophila*. Also, the aligned *P. tetraurelia* ES7 sequence had a large number of gaps. Since the rest of the *P.*

tetraurelia 26S rRNA sequence, outside the expansion segment, is highly homologous to the same regions in other ciliates, it is clear that ES7 in *Paramecium* is significantly shorter than in the other ciliates.

DISCUSSION

Homology modeling of TtP0:

To explore the structure and function of a previously identified ciliate-specific P0 insert, we created a homology model of the L10 region, including the insert, of TtP0, based on coordinates (chain q of 3U5I) from the P0 of *S. cerevisiae.* This homology model provides the first view of the insert's accessable conformations, as well as evidence of a novel interaction between the insert region of TtP0 and an expansion segment, ES7B, of *T. thermophila* 26S rRNA (discussed in more detail below). In addition, residues 7 to 125, and 203 to 218, in the L10 region of TtP0 were included in the models, providing vital information about the side chains that is missing from the crystal structure of TtP0 reported by Klinge et al (chain G of 4A1C) (2011).

Our homology model consists of one of the three structural domains of TtP0, and contains about one-third of its amino acid residues. Other crystal structures of the stalk complex could be used as templates for constructing additional models of the middle domain and the P1/P2 helices. This was outside the region of interest in the present study, but it will be visited in a future study. However, we decided to focus on the L10 region in the current study, for two reasons. 1), The crystal structure of the *T. thermophila* ribosome lacks coordinates for any residues beyond the L10 region; without these experimentally-derived coordinates as a check, building the other

sections of TtP0 might decrease the homology model's overall quality. 2), Modeling the other regions could distract from the main focus of our study, the structure and function of the ciliate insert.

Phosphorylation of TtP0:

Unlike other eukaryotes, *T. thermophila* lacks a conserved serine that is usually located a few residues before the conserved *C*-terminal peptide (Ballesta et al, 1999); this "missing serine" corresponds to about position 315 in TtP0. Combined with the reported lack of a phosphorylated form of P0 in *T. pyriformis* (Sandermann, Kruger and Kristiansen, 1979), this lack raises the possibility of an alternative phosphorylation site located elsewhere on the molecule, including possibly within the ciliate insert. Although bioinformatics site prediction tools identified several potential phosphorylation sites in the insert region of TtP0 (at Thr76, Ser78, Tyr80 and Tyr83), none of these residues are predicted to interact with ES7 based on the homology modeling. This may mean that if one or more residues within the insert are phosphorylated as part of one or more regulatory mechanisms, it would only interfere with the interaction of TtP0 and ES7 indirectly.

In a previously published study characterizing the *T. thermophila* phosphoproteome networks, several phosphopeptides belonging to TtP0 were detected, with phosphorylation sites at Ser188, Ser191 and Ser290 (Tian et al, 2013). In our study, both Ser188 and Ser191 were identified by NetPhosK as likely phosphorylation sites, for Protein Kinase A and Casein Kinase II respectively, while Ser290 was not identified as a likely phosphorylation site in our analysis. Since all

three of these sites are outside of the L10 region and its insert, the question remains open whether TtP0 phosphorylation occurs within the insert region.

We attempted to determine, through modeling approaches, whether phosphorylation of the insert of TtP0 would have a significant effect on its conformation and interaction with the 26S RNA. However, the nonstandard nature of phosphorylated amino acids makes them difficult to parameterize and incorporate into homology models using the methods outlined in this paper. Another approach will likely be necessary to test whether the L10 insert is phosphorylated, and how phosphorylation could affect the conformation of the loop and its potential interaction with the 26S rRNA. An *in vivo* approach may eventually be necessary to confirm if the insert is in fact phosphorylated, and if so, whether this could serve to regulate the interaction of TtP0 with the 26S rRNA.

The ciliate insert/ES7B interaction:

After refining the model of the ciliate insert in the presence of a portion of the 26S rRNA, we were intrigued to find that the original range of orientations for the insert loop appeared to be restricted by protein-RNA interactions. Hydrogen bonding occurred between residues toward the *C*-terminal end of the TtP0 insert (specifically Gln84, Phe85 and Gly86) and bases C584 and A585 of the rRNA. Both of these bases are located on the tip of a loop of Expansion Segment 7B (ES7B), a eukaryote-specific region of the 26S rRNA in the large ribosomal subunit (Figure 5). A potential interaction between ES7 and P0 of *S. cerevisiae* was suggested by Ben-Shem et al (2011), but no reference was made as to what residues or bases could be involved.

When we searched for additional ES7 and insert sequence data from other ciliate species, in order to determine if a similar interaction might occur in other ciliates, we found P0 and rRNA sequences for two other species, *P. tetraurelia* and *O. trifallax*. When we then performed a ClustalW alignment of the ES7 regions of these species and those of *T. thermophila* and *T. pyriformis*, we observed that the two *Tetrahymena* ES7 sequences were highly homologous overall, and contained similar base sequences in the region that may interact with the ciliate insert. However, we also found that *P. tetraurelia* has a significantly shorter expansion segment than does *Tetrahymena* and *O. trifallax.* Also, the ES7 of *Oxytricha* appeared to have a cytosine residue in a position comparable to the cytosine at the end of the ES7 loop in *Tetrahymena* (C584), but otherwise the overall sequence of the *O. trifallax* ES7 was somewhat more divergent.

Engberg et al (1990) reported that the secondary structures of the 26S rRNA of *T. thermophila* and *T. pyriformis* are virtually identical. Based on their observations, and our own observations on ES7, it seems likely that the P0 insert and ES7 of *T. pyriformis* interact in a way similar to what we would predict for *T. thermophila*, although the sequence of *T. pyriformis* P0 (TpP0) is not yet available. On the other hand, *O. trifallax* and *P. tetraurelia* are both less likely to have an ES7-insert interaction like that of *T. thermophila*, based on observed differences between the P0 and ES7 sequences. While the *O. trifallax* ES7 is of similar length to that of *T. thermophila,* and they share some homology, the inserts have almost no identity. Almost the opposite was observed for *P. tetraurelia*; while its P0 insert contains both a Phe and Gly in similar positions to those of the TtP0 insert, the ES7 of *P. tetraurelia*

is highly divergent from that of *T. thermophila*. *Paramecium* is more closely related to *Tetrahymena* than is *Oxytricha*, so the divergence in their ES7 sequences is surprising. These divergences in ES7 or the insert raise questions about whether or not there has been a divergence in the function of the insert in these other ciliate lineages.

Any conclusions drawn about the conservation in ciliates of the ES7-insert interaction may be limited here, because so few ciliate rRNA and/or P0 sequences are currently available for study. Furthermore, the P0 sequences gathered here represent only three of the eleven known ciliate classes (the Oligohymenophorea, Spirotrichea and Heterotrichea) (Lynn, 2002), so it is still too early to make broad generalizations about the homology relationships between the P0 insert and ES7 regions of all members of the ciliate clade. If the ES7-P0 insert interaction is conserved throughout some or all of the ciliate classes, then it seems likely that the 26S rRNA sequence co-evolved with the ciliate insert. However, before this possibility can be explored further, *in vivo* and *in vitro* experiments on the *T. thermophila* insert and ES7 should be conducted, to verify the existence and importance of the interaction predicted by homology modeling.

REFERENCES

Anger AM, Armache JP, Berninghausen O, Habeck M, Subklewe M, Wilson DN and Beckmann R. (2013), "Structures of the human and *Drosophila* 80S ribosome". Nature 497, 80-85.

Armache JP, Jarasch A, Anger AM, Villa E, Becker T, Bhushan S, Jossinet F, Habeck M, Dindar G, Franckenberg S, Marquez V, Mielke T, Thomm M, Berninghausen O, Beatrix B, Söding J, Westhof E, Wilson DN and Beckmann R. (2010), "Localization of eukaryote-specific ribosomal proteins in a 5.5-Å cryo-EM map of the 80S eukaryotic ribosome". Proc Natl Acad Sci USA. 107(46), 19754–19759.

Ballesta JP, Rodriguez-Gabriel MA, Bou G, Briones E, Zambrano R and Remacha M. (1999), "Phosphorylation of the yeast ribosomal stalk. Functional effects and enzymes involved in the process". FEMS Microbiol Rev. 23(5), 537-550.

Ben-Shem A, Garreau de Loubresse N, Melnikov S, Jenner L, Yusupova G and Yusupov M. (2011), "The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution". Science 334 (6062), 1524-1529.

Blom N, Gammeltoft S, and Brunak S. (1999), "Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites". J Mol Biol 294(5), 1351-1362.

Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S and Brunak S. (2004), "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence ". Proteomics 4(6), 1633-1649. Review.

Calich AL, Viana VS, Cancado E, Tustumi F, Terrabuio DR, Leon EP, Silva CA, Borba EF, and Bonfa E. (2013), "Anti-ribosomal P protein: a novel antibody in autoimmune hepatitis". Liver Int. 33(6), 909-913.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, MÜller KM, Pande N, Shang Z, Yu N and Gutell RR. (2002), "The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs". *BioMed Central Bioinformatics*, 3(2). [Correction: *BioMed Central Bioinformatics*. 3(15).]

Engberg J, Nielsen H, Lenaers G, Murayama O, Fujitani H and Hinashinagakawa T. (1990), "Comparison of primary and secondary 26S rRNA structures in two *Tetrahymena* species: Evidence for a strong evolutionary and structural constraint in expansion segments". J Mol Evol 30(6), 514-521.

Eswar N, Marti-Renom MA, Webb B, Madhusudhan MS, Eramian D, Shen M, Pieper U and Sali A. (2006), "Comparative Protein Structure Modeling With MODELLER". Curr. Protoc. Bioinformatics Supplement 15, 5.6.1-5.6.30.

Gordiyenko Y, Videler H, Zhou M, McKay AR, Fucini P, Biegel E, Müller V and Robinson CV. (2010), "Mass spectrometry defines the stoichiometry of ribosomal stalk complexes across the phylogenetic tree". Mol. Cell Proteomics 9(8), 1774-1783.

Heinlen LD, Ritterhouse LL, McClain MT, Keith MP, Neas BR, Harley JB and James JA. (2010)," Ribosomal P autoantibodies are present before SLE onset and are directed against non-*C*-terminal peptides". J Mol Med (Berl). 88(7), 719-727.

Hiu-Mei Too P, Kit-Wan Ma M, Nga-Szse Mak A, Wong YT, Kit-Ching Tung C, Zhu G, Wing-Ngor Au S, Wong KB and Shaw PC. (2009), "The *C*-terminal fragment of the ribosomal P protein complexed to trichosanthin reveals the interaction between the ribosome-inactivating protein and the ribosome". Nucleic Acids Res. 37(2), 602–610.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z and Dunker AK. (2004), "The importance of Intrinsic disorder for protein phosphorylation". Nucleic Acids Res., 32(3), 1037-1049.

Iborra S, Soto M, Carrión J, Nieto A, Fernández E, Alonso C and Requena JM. (2003), "The *Leishmania infantum* Acidic Ribosomal Protein P0 Administered as a DNA Vaccine Confers Protective Immunity to *Leishmania major* Infection in BALB/c Mice" Infect. Immun. 71(11), 6562-6572.

Justice MC, Ku T, Hsu MJ, Carniol K, Schmatz D and Nielsen J. (1999), "Mutations in Ribosomal Protein L10e Confer Resistance to the Fungal-specific Eukaryotic Elongation Factor 2 Inhibitor Sordarin". J. Biol.Chem. 274, 4869-4875.

Klinge S, Voigts-Hoffmann F, Leibundgut M, Arpagaus S and Ban N. (2011), "Crystal Structure of the Eukaryotic 60S Ribosomal Subunit in Complex with Initiation Factor 6". Science 334, 941-948.

Kravchenko O, Mitroshin I, Nikonov S, Piendl W and Garber M. (2010), "Structure of a two-domain *N*-terminal fragment of ribosomal protein L10 from *Methanococcus jannaschii* reveals a specific piece of the archaeal ribosomal stalk." J.Mol.Biol. 399(2), 214-220

Krieger E, Nabuurs SB and Vriend G. (2003), "Homology Modeling". In "Structural Bioinformatics" (Editors, Bourne, P.E. & Weissig, H.), Wiley & Sons Inc, 509-525.

Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T and Ussery DW. (2007), "RNammer: consistent annotation of rRNA genes in genomic sequences." Nucleic Acids Res. 35(9), 3100-3108.

Liao D and Dennis PP. (1994), "Molecular phylogenies based on ribosomal protein L11, L1, L10, and L12 sequences". J Mol. Evol. 38(4), 405-419.

Liu CC, Lu TC, Li HH, Wang HX, Liu GF, Ma L, Yang CP and Wang BC. (2010), "Phosphoproteomic identification and phylogenetic analysis of ribosomal P-proteins in *Populus* dormant terminal buds". Planta. 231(3), 571-581.

Lynn DH. (2002), *The Ciliate Resource Archive.* http://www.uoguelph.ca/~ciliates/classification/genera.html. Accessed July 21st, 2014.

Naganuma T, Nomura N, Yao M, Mochizuki M, Uchiumi T and Tanaka I. (2010), "Structural Basis for Translation Factor Recruitment to the Eukaryotic/Archaeal Ribosomes". J Biol. Chem. 285: 4747-4756.

Nomura N, Honda T, Baba K, Naganuma T, Tanzawa T, Arisaka F, Noda M, Uchiyama S, Tanaka I, Yao M, Uchiumi T. (2012), "Archaeal ribosomal stalk protein interacts with translation factors in a nucleotide-independent manner via its conserved C terminus. Proc. Natl. Acad. Sci. USA 109(10), 3748-3753.

Nomura T, Nakano K, Maki Y, Naganuma T, Nakashima T, Tanaka I, Kimura M, Hachimori A and Uchiumi T. (2006), "*In vitro* reconstitution of the GTPase-associated centre of the archaebacterial ribosome: the functional features

observed in a hybrid form with *Escherichia coli* 50S subunits". Biochem J. 396(3), 565-571.

Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Hau J, Martin O, Kuznetsov D and Falquet L. (2007), "MyHits: improvements to an interactive resource for analyzing protein sequences". Nucleic Acids Res. 35(Web Server issue), W433-437.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC and Ferrin TE. (2004), "UCSF Chimera--a visualization system for exploratory research and analysis". J Comput Chem. 25(13), 1605-1612.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J and Glöckner FO. (2013), "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". Nucl. Acids Res. 41(D1), D590-D596.

Rabl J, Leibundgut M, Ataide SF, Haag A and Ban N. (2011), "Crystal Structure of the Eukaryotic 40S Ribosomal Subunit in Complex with Initiation Factor 1". Science 331, 730-736.

Rajeshwari K, Patel K, Nambeesan S, Mehta M, Sehgal A, Chakraborty T and Sharma S. (2004), "The P Domain of the P0 protein of *Plasmodium falciparum* Protects against Challenge with Malaria Parasites". Infection and Immunity 72 (9), 5515-5521.

Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Rodriguez Gabriel MA, Guarinos E, and Ballesta JP. (1995), "Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects ". Biochem Cell Biol 73, 959-968.

Sandermann J, Krüger A *and* Kristiansen K. *(*1979*),* "Characterization of acidic proteins in *Tetrahymena pyriformis"*. FEBS Lett. 107, 343–347.

Santos C, Rodriguez-Gabriel MA, Remacha M and Ballesta, JPG. (2004), "Ribosomal P0 Protein Domain Involved in Selectivity of Antifungal Sordarin Derivatives. *Antimicrob. Agents Chemother.* 48(8), 2930-2936.

Schumacher J, Babcock K, Canton S and Hufnagel LA, 2010. Tetrahymena orthologue of the malaria vaccine candidate Phosphoprotein p0: Immunocytochemical localization and a co-expressed membrane protein with unique properties. Proc. Internat. Soc. of Protistol. (ISOP), Canterbury, England.

Schumacher J, Canton S, Babcock K and Hufnagel LA, 2010. An orthologue of the apicomplexan vaccine candidate phosphoprotein P0, a conserved ribosomal protein, is also present at the cell surface in the alveolate protist, *Tetrahymena thermophila*. Proc. Ann. Mtg., Amer. Soc. Cell Biol. (ASCB), Philadelphia.

Schumacher J, Canton S, Babcock K, and Hufnagel LA, 2010. An Orthologue of the Malaria Vaccine Candidate Phosphoprotein p0 in *Tetrahymena thermophila*: Immunocytochemical Localization. Proc. N. Amer. Chapt., Internat. Soc. of Protistol. (ISOP), Lexington, VA

Schumacher J, Corriveau J, Canton S, and Hufnagel LA, 2009. An orthologue of the protozoan vaccine candidate phosphoprotein p0 in *Tetrahymena thermophila*. Proc. N. Amer. Chapt., Internat. Soc. of Protistol. (ISOP), Bristol, RI.

Schumacher J and Hufnagel LA. "Protozoan Vaccine Candidate Homologues in *Tetrahymena thermophila*". Manuscript in preparation.

Sehgal A, Kumar N, Carruthers VB and Sharma S. “Translocation of ribosomal protein P0 onto the *Toxoplasma gondii* tachyzoite surface”. Int J Parasitol. 33(14), 1589-1594.

Singh S, Sehgal A, Waghmare S, Chakraborty T, Goswami A and Sharma S. (2002), “Surface expression of the conserved ribosomal protein P0 on parasite and other cells”. Mol. Biochem. Parasit 119, 121-124

Soares MR, Bisch PM, Campos De Carvalho AC, Valente AP and Almeida FCL. (2004),
“Correlation between conformation and antibody binding: NMR structure of cross-reactive peptides from T. cruzi, human and L. braziliensis”. Febs Lett. 560, 134-140.

Spassov VZ, Flook PK and Yan L. (2008), “LOOPER: a molecular mechanics-based algorithm for protein loop prediction. Protein Eng., Des. Sel. 21, 91–100.

Spassov VZ, Yan L and Flook PK. (2007), “The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: A side-chain prediction algorithm based on side-chain backbone interactions”. Protein Science 16, 494–506.

Uchiumi T, Honma S, Endo Y, Hachimori A. (2002) “Ribosomal proteins at the stalk region modulate functional rRNA structures in the GTPase center”. J Biol Chem. 277(44):4, 1401-1409.

Uchiumi T, Traut RR, Elkon K and Kominami R. (1991), “A human autoantibody specific for a unique conserved region of 28 S ribosomal RNA inhibits

the interaction of elongation factors 1 alpha and 2 with ribosomes". J Biol Chem. 266(4), 2054-62.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W and Glöckner FO (2014), "The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks". Nucl. Acids Res. 42, D643-D648

Figure Legends:

Figure 1: Schematic of the eukaryotic stalk complex

A representation of the general elements of the stalk complex of the 60S ribosomal subunit, including multiple copies of P1 and P2. P0 engages in several protein-protein and protein-RNA interactions to act as the scaffold for the P stalk. P0 consists of the L10, MID, 60S (shown as Helix 1 and Helix 2) and *C*-terminal peptide, which is shared with P1 and P2.

Figure 2: Alignment of eukaryotic P0s demonstrates the presence of a ciliate-specific insert

A: A portion of an unedited MCoffee alignment of 104 P0 and L10 sequences, showing conservation of the main functional regions—the L10 (outlined in green), middle (outlined in orange) and 60S (outlined in blue). Within the L10 region is an area with a large gap, representing the site of the ciliate and kinetoplastid inserts (outlined in red).

B: A portion of an unedited ClustalW alignment of 24 eukaryotic P0 sequences, including 9 ciliate species. *Tetrahymena thermophila* and other ciliates (in red) contain a 15 amino-acid long insert, while species of the Kinetoplastida (green) contain only a partial insert. Other eukaryotes lack an insert completely. Numbers next to species names indicate the residues shown.

Figure 3: Motif Analysis of TtP0.

A summary of a MyHits scan on the TtP0 amino acid sequence. Two well established motif sequences were identified, the *N*-terminal L10 region (dark grey) and the *C*-terminal 60S region (light grey). In addition, several potential Casein kinase II (bold

and italicized) and Protein Kinase C (bold and underlined) phosphorylation sites were identified, with one possible site located in the predicted insert (white, predicted phosphorylation site bold and italicized).

Figure 4: Netphos 2.0 prediction of likely phosphorylation sites on TtP0 suggests that a serine or tyrosine phosphorylation site may exist within the insert.

A NetPhos 2.0 search for generic serine, threonine and tyrosine phosphorylation sites. All potential sites are scored between 0 and 1, with scores of 0.5 or greater representing likely phosphorylation sites. Two potential phosphorylation sites within the insert, at Ser 78 and Tyr 80, are marked with an asterisk.

Figure 5: Homology modeling of TtP0

A view of the whole L10 region (A) and a close-up of the 20 lowest energy models of the insert modeled without the rRNA (B). The model takes on a random appearance, with a wide variety of possible positions.

C, D and E: The 25 lowest-energy in-context models of the TtP0 insert, shown in relation to the ES7B hairpin region of the 26S rRNA of *T. thermophila*. All of them are positioned over the end of the ES7B hairpin, indicating that the insert and the rRNA may interact with each other.

Figure 6: Interactions of the insert and the 26S ribosomal RNA.

Above: A sketch of the secondary structure of ES7 from *T. thermophila*, based on a secondary structure diagram from Klinge et al. A close up view of the region of interaction (in red) is shown with the sequence of the TtP0 insert. Residues and bases capable of H bonding according to the models are shown in blue and green boxes respectively.

Below: A table indicating the H-bonding atoms for the ten lowest-energy loop conformations. Numberings for the bases and residues are relative to the portion of the P0 and ES7 that was included in the modeling, rather than their placement in the actual ribosomes.

Figure 7: Alignment of ciliate ES7B

A portion of an unedited ClustalW alignment of four ciliate LSU rRNA sequences, showing the varying nature of ES7B among the ciliates. The region of ES7 with a potential interaction with the insert is highlighted in gray.

Table 1: The sequences of the ciliate-specific P0 insert among different ciliate species

Portions of the predicted amino acid sequences for the ciliate P0s used for this study, along with their Uniprot accession numbers. There are significant differences between the insert regions of different species, though more closely related ciliates (*T. thermophila* and *I. multifillis,* or *Euplotes* sp. and *O. trifallax*) share some homology within the insert.

Table 2: Predicted phosphorylation sites of TtP0

Summary of predicted phosphorylation of serine, threonine and tyrosine residues of TtP0 from NETPHOS2.0, DISPHOS and NETPHOSK. Residues within the ciliate-specific insert are highlighted in gray. Residues that were predicted to be “likely phosphorylated” (PP>0.5) by the different programs are marked with stars. For NETPHOSK predictions, one or more kinases involved with phosphorylation are indicated.

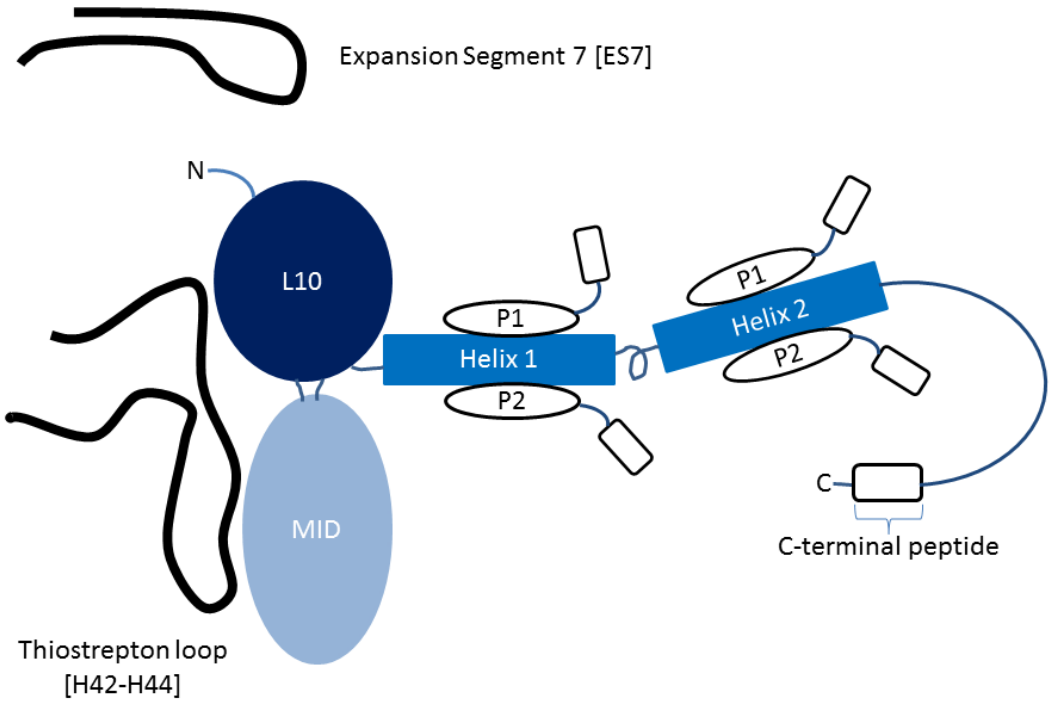
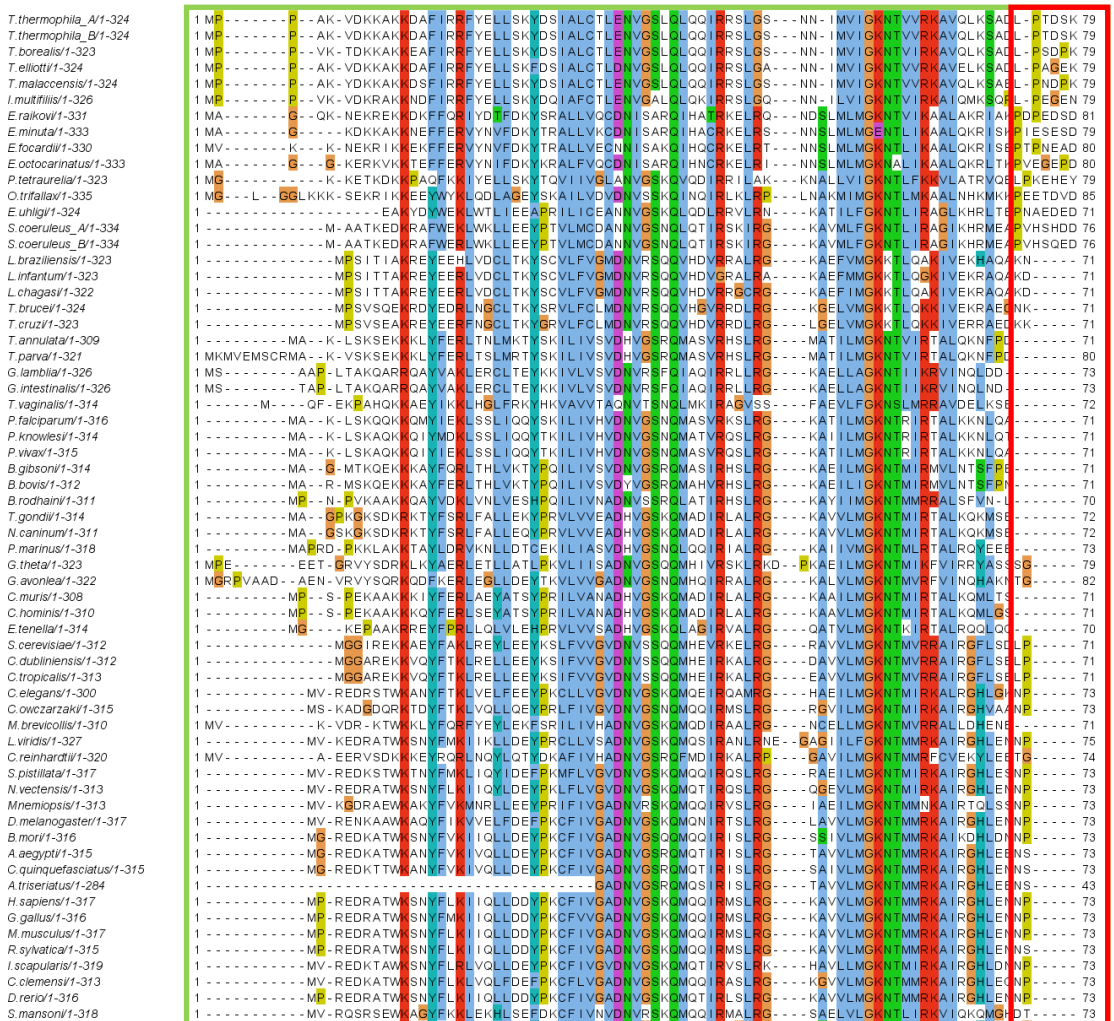
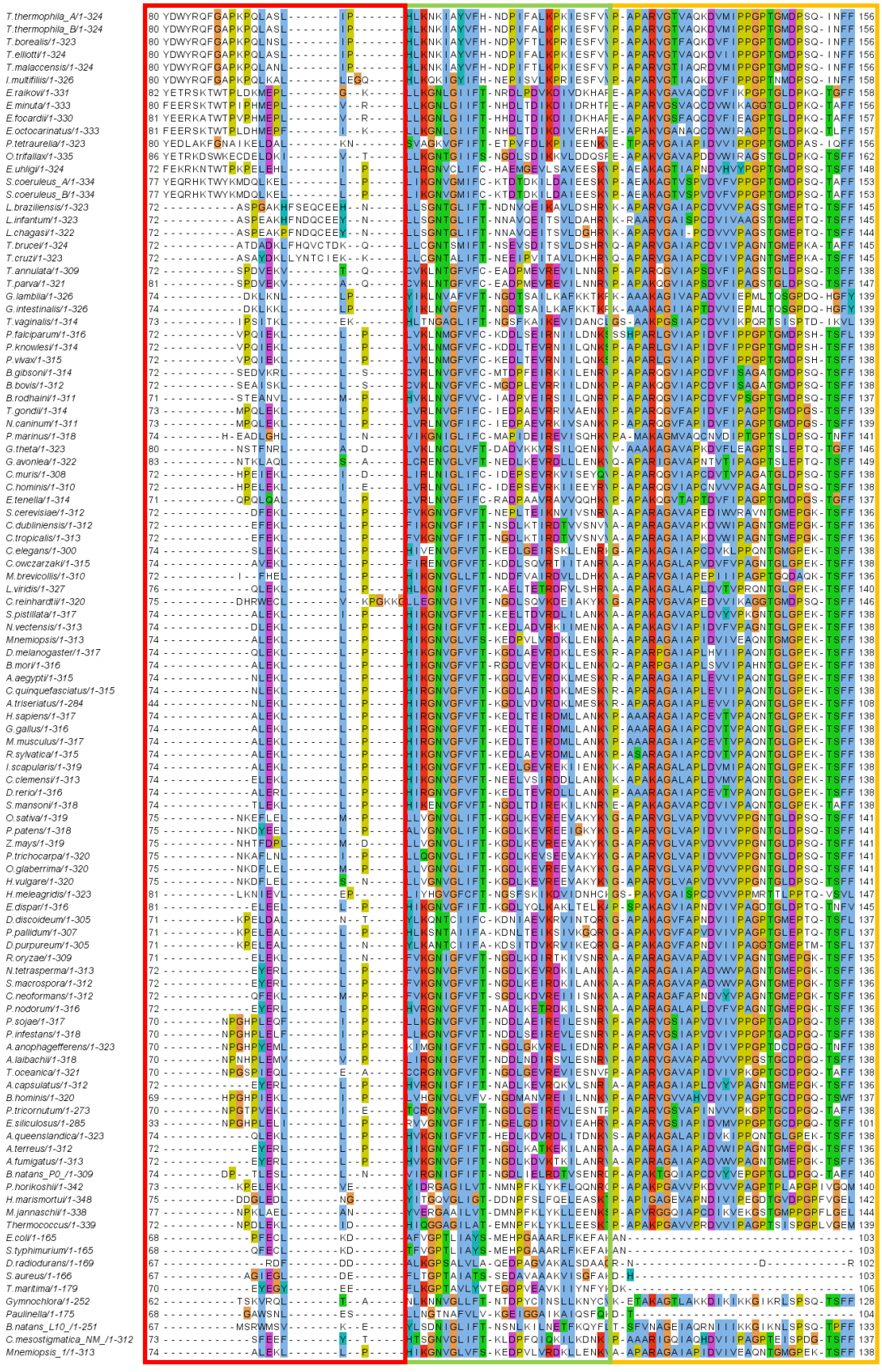
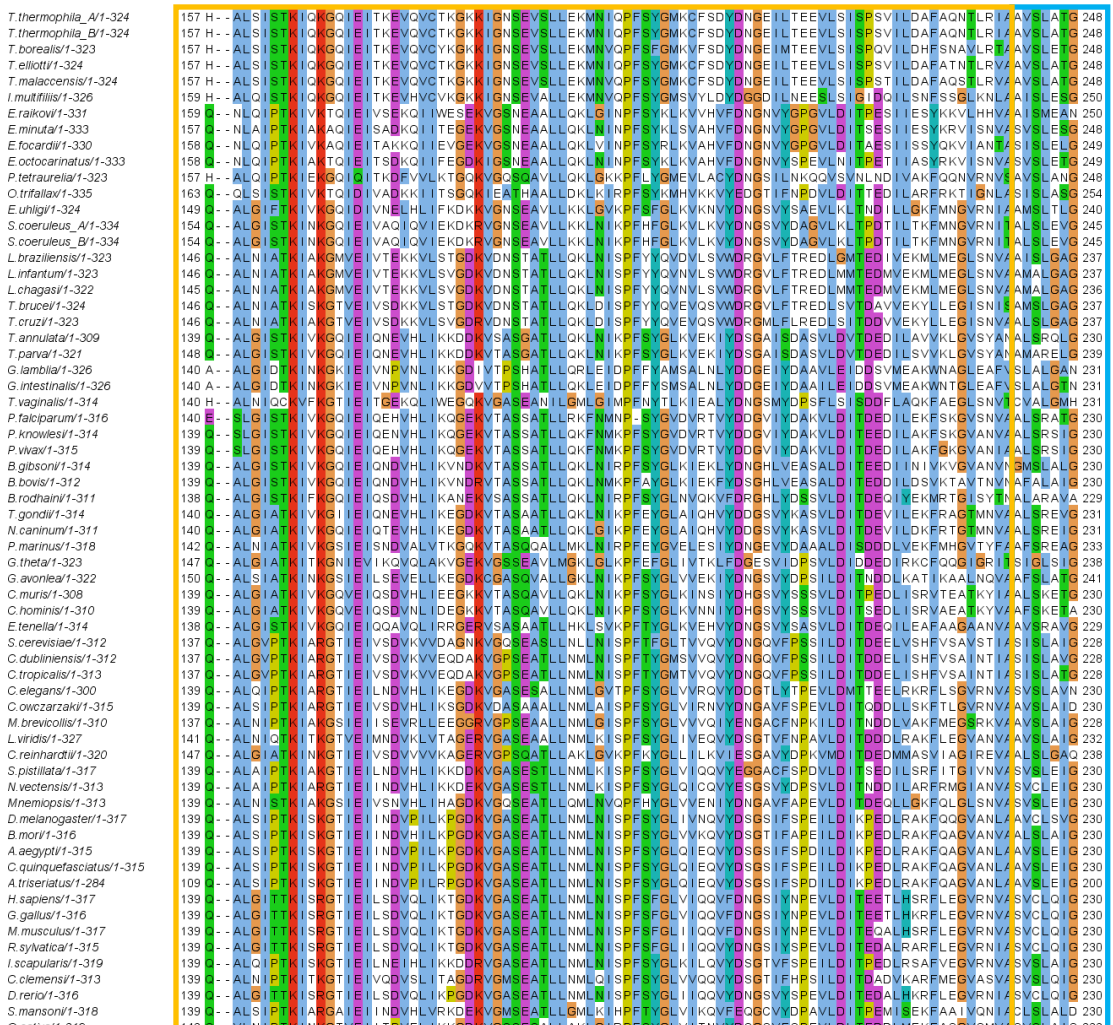

Expansion Segment 7 [ES7]
N
L10
MID
Thiostrepton loop
[H42-H44]
P1
Helix 1
P2
P1
Helix 2
P2
C
C-terminal peptide

Figure 1


T.thermophila_A/1-324
T.thermophila_B/1-324
T.borealis/1-323
T.elliotti/1-324
T.malaccensis/1-324
I.multifiliis/1-326
E.raikovi/1-331
E.minuta/1-333
E.focardii/1-330
E.octocarinatus/1-333
P.tetraurelia/1-323
O.trifallax/1-335
E.uhligi/1-324
S.coeruleus_A/1-334
S.coeruleus_B/1-334
L.braziliensis/1-323
L.infantum/1-323
L.chagasi/1-322
T.brucei/1-324
T.cruzi/1-323
T.annulata/1-309
T.parva/1-321
G.lamblia/1-326
G.intestinalis/1-326
T.vaginalis/1-314
P.falciparum/1-316
P.knowlesi/1-314
P.vivax/1-315
B.gibsoni/1-314
B.bovis/1-312
B.rodhaini/1-311
T.gondii/1-314
N.caninum/1-311
P.marinus/1-318
G.theta/1-323
G.avonlea/1-322
C.muris/1-308
C.hominis/1-310
E.tenella/1-314
S.cerevisiae/1-312
C.dubliniensis/1-312
C.tropicalis/1-313
C.elegans/1-300
C.owczarzaki/1-315
M.brevicollis/1-310
L.viridis/1-327
C.reinhardtii/1-320
S.pistillata/1-317
N.vectensis/1-313
Mnemiopsis/1-313
D.melanogaster/1-317
B.mori/1-316
A.aegypti/1-315
C.quinquefasciatus/1-315
A.triseriatus/1-284
H.sapiens/1-317
G.gallus/1-316
M.musculus/1-317
R.sylvatica/1-315
I.scapularis/1-319
C.clemensi/1-313
D.rerio/1-316
S.mansoni/1-318
O.sativa/1-319
P.patens/1-318
Z.mays/1-319
P.trichocarpa/1-320
O.glaberrima/1-320
H.vulgare/1-320
H.meleagridis/1-323
E.dispar/1-316
D.discoideum/1-305
P.pallidum/1-307
D.purpureum/1-305
R.oryzae/1-309
N.tetrasperma/1-313
S.macrospora/1-312
C.neoformans/1-312
P.nodorum/1-316
P.sojae/1-317
P.infestans/1-318
A.anophagefferens/1-323
A.laibachii/1-318
T.oceanica/1-321
A.capsulatus/1-312
B.hominis/1-320
P.tricornutum/1-273
E.siliculosus/1-285
A.queenslandica/1-323
A.terreus/1-312
A.fumigatus/1-313
B.natans_P0_/1-309
P.horikoshii/1-342
H.marismortui/1-348
M.jannaschii/1-338
Thermococcus/1-339
E.coli/1-165
S.typhimurium/1-165
D.radiodurans/1-169
S.aureus/1-166
T.maritima/1-179
Gymnochlora/1-252
Paulinella/1-175
B.natans_L10_/1-251
C.mesostigmatica_NM_/1-312
Mnemiopsis_1/1-313


T.thermophila_A/1-324
T.thermophila_B/1-324
T.borealis/1-323
T.elliotti/1-324
T.malaccensis/1-324
I.multifiliis/1-326
E.raikovi/1-331
E.minuta/1-333
E.focardii/1-330
E.octocarinatus/1-333
P.tetraurelia/1-323
O.trifallax/1-335
E.uhligi/1-324
S.coeruleus_A/1-334
S.coeruleus_B/1-334
L.braziliensis/1-323
L.infantum/1-323
L.chagasi/1-322
T.brucei/1-324
T.cruzi/1-323
T.annulata/1-309
T.parva/1-321
G.lamblia/1-326
G.intestinalis/1-326
T.vaginalis/1-314
P.falciparum/1-316
P.knowlesi/1-314
P.vivax/1-315
B.gibsoni/1-314
B.bovis/1-312
B.rodhaini/1-311
T.gondii/1-314
N.caninum/1-311
P.marinus/1-318
G.theta/1-323
G.avonlea/1-322
C.muris/1-308
C.hominis/1-310
E.tenella/1-314
S.cerevisiae/1-312
C.dubliniensis/1-312
C.tropicalis/1-313
C.elegans/1-300
C.owczarzaki/1-315
M.brevicollis/1-310
L.viridis/1-327
C.reinhardtii/1-320
S.pistillata/1-317
N.vectensis/1-313
Mnemiopsis/1-313
D.melanogaster/1-317
B.mori/1-316
A.aegypti/1-315
C.quinquefasciatus/1-315
A.triseriatus/1-284
H.sapiens/1-317
G.gallus/1-316
M.musculus/1-317
R.sylvatica/1-315
I.scapularis/1-319
C.clemensi/1-313
D.rerio/1-316
S.mansoni/1-318
O.sativa/1-319
P.patens/1-318
Z.mays/1-319
P.trichocarpa/1-320
O.glaberrima/1-320
H.vulgare/1-320
H.meleagridis/1-323
E.dispar/1-316
D.discoideum/1-305
P.pallidum/1-307
D.purpureum/1-305
R.oryzae/1-309
N.tetrasperma/1-313
S.macrospora/1-312
C.neoformans/1-312
P.nodorum/1-316
P.sojae/1-317
P.infestans/1-318
A.anophagefferens/1-323
A.laibachii/1-318
T.oceanica/1-321
A.capsulatus/1-312
B.hominis/1-320
P.tricornutum/1-273
E.siliculosus/1-285
A.queenslandica/1-323
A.terreus/1-312
A.fumigatus/1-313
B.natans_P0_/1-309
P.horikoshii/1-342
H.marismortui/1-348
M.jannaschii/1-338
Thermococcus/1-339
E.coli/1-165
S.typhimurium/1-165
D.radiodurans/1-169
S.aureus/1-166
T.maritima/1-179
Gymnochlora/1-252
Paulinella/1-175
B.natans_L10_/1-251
C.mesostigmatica_NM_/1-312
Mnemiopsis_1/1-313


T.thermophila_A/1-324
T.thermophila_B/1-324
T.borealis/1-323
T.elliotti/1-324
T.malaccensis/1-324
I.multifiliis/1-326
E.raikovi/1-331
E.minuta/1-333
E.focardii/1-330
E.octocarinatus/1-333
P.tetraurelia/1-323
O.trifallax/1-335
E.uhligi/1-324
S.coeruleus_A/1-334
S.coeruleus_B/1-334
L.braziliensis/1-323
L.infantum/1-323
L.chagasi/1-322
T.brucei/1-324
T.cruzi/1-323
T.annulata/1-309
T.parva/1-321
G.lamblia/1-326
G.intestinalis/1-326
T.vaginalis/1-314
P.falciparum/1-316
P.knowlesi/1-314
P.vivax/1-315
B.gibsoni/1-314
B.bovis/1-312
B.rodhaini/1-311
T.gondii/1-314
N.caninum/1-311
P.marinus/1-318
G.theta/1-323
G.avonlea/1-322
C.muris/1-308
C.hominis/1-310
E.tenella/1-314
S.cerevisiae/1-312
C.dubliniensis/1-312
C.tropicalis/1-313
C.elegans/1-300
C.owczarzaki/1-315
M.brevicollis/1-310
L.viridis/1-327
C.reinhardtii/1-320
S.pistillata/1-317
N.vectensis/1-313
Mnemiopsis/1-313
D.melanogaster/1-317
B.mori/1-316
A.aegypti/1-315
C.quinquefasciatus/1-315
A.triseriatus/1-284
H.sapiens/1-317
G.gallus/1-316
M.musculus/1-317
R.sylvatica/1-315
I.scapularis/1-319
C.clemensi/1-313
D.rerio/1-316
S.mansoni/1-318
O.sativa/1-319
P.patens/1-318
Z.mays/1-319
P.trichocarpa/1-320
O.glaberrima/1-320
H.vulgare/1-320
H.meleagridis/1-323
E.dispar/1-316
D.discoideum/1-305
P.pallidum/1-307
D.purpureum/1-305
R.oryzae/1-309
N.tetrasperma/1-313
S.macrospora/1-312
C.neoformans/1-312
P.nodorum/1-316
P.sojae/1-317
P.infestans/1-318
A.anophagefferens/1-323
A.laibachii/1-318
T.oceanica/1-321
A.capsulatus/1-312
B.hominis/1-320
P.tricornutum/1-273
E.siliculosus/1-285
A.queenslandica/1-323
A.terreus/1-312
A.fumigatus/1-313
B.natans_P0_/1-309
P.horikoshii/1-342
H.marismortui/1-348
M.jannaschii/1-338
Thermococcus/1-339
E.coli/1-165
S.typhimurium/1-165
D.radiodurans/1-169
S.aureus/1-166
T.maritima/1-179
Gymnochlora/1-252
Paulinella/1-175
B.natans_L10_/1-251
C.mesostigmatica_NM_/1-312
Mnemiopsis_1/1-313

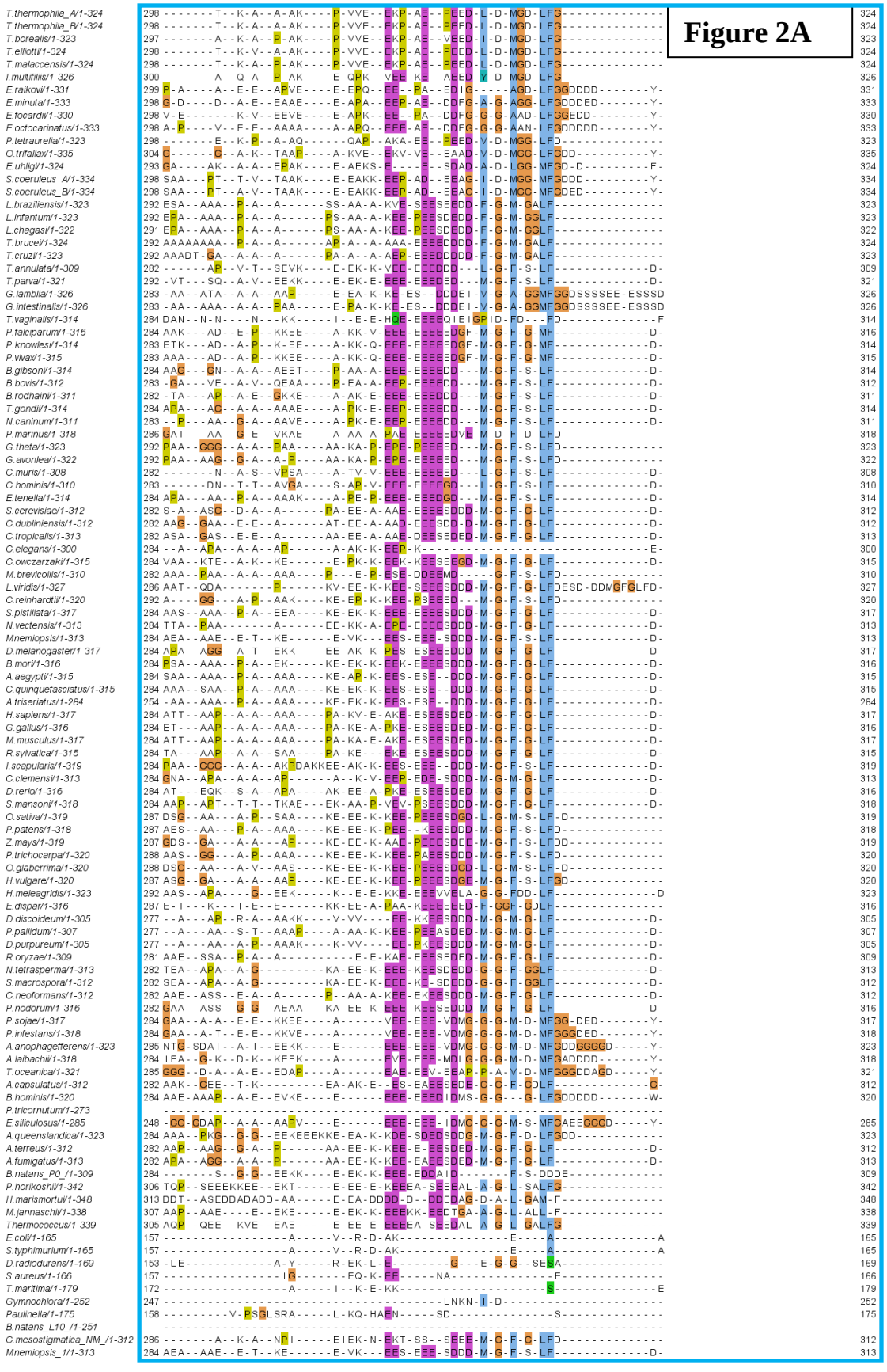


Figure 2A

```
T.thermophila/54-98    -IMVIGKNTVVRKAVQLKSADLP-TDSKYDWYRQFGAPKPQLASLIP
I.multifiliis/54-98    -ILVIGKNTVIRKAIQMKSQPLP-EGENYDWYRQFGAPKPQLKALLE
P.tetraurelia/52-98    ALLVIGKNTLFKKVLATRVQELPKEHEYYEDLAKFGNAIKELDALKN
E.focardii/53-99       SLMLMGKNTLIKAALQKRISEPTPNEADYEERKATWTPVPHMEPLVR
E.minuta/52-98         SLMLMGENTLIKAALQKRISKPIESESDFEERSKTWTPIPHMEPLVR
E.octocarinatus/53-99  SLMLMGKNALIKAALQKRLTKPVEGEPDFEERSKTWTPLDHMEPFIK
E.raikovi/54-100       SLMLMGKNTVIKAALAKRIAKPDPEDSDYETRSKTWTPLDKMEPLGK
O.trifallax/58-104     AKMIMGKNTLMKAALNHKMKKPEETDVDYETRKDSWKECDELDKIVT
E.uhligi/44-90         ATILFGKNTLIRAGLKHRLTEPNAEDEDFEKRKNTWTPKPELEHLIP
L.braziliensis/49-87   AEFVMGKKTLQAKIVEKHAQAKN--------ASPGAKHFSEQCEEHN
T.brucei/49-87         GELVMGKKTLQKKIVEKRAEGNK--------ATDADKLFHQVCTDKQ
T.annulata/51-80       ATILMGKNTVIRTALQKNFPD-----------------SPDVEKVTQ
P.falciparum/51-80     ATILMGKNTRIRTALKKNLQA-----------------VPQIEKLLP
T.gondii/52-81         AVVLMGKNTMIRTALKQKMSE-----------------MPQLEKLLP
C.muris/51-80          AAILMGKNTMIRTALKQMLTS-----------------HPEIEKLID
P.marinus/53-83        AIIVMGKNTMLRTALRQYEEEH----------------EADLGHLLN
T.vaginalis/52-81      AEVLFGKNSLMRRAVDELKSE-----------------IPSITKLEK
D.discoideum/50-79     GAVLMGKKTMIRKVIRDLADS-----------------KPELDALNT
S.cerevisiae/49-78     AVVLMGKNTMVRRAIRGFLSD-----------------LPDFEKLLP
M.musculus/51-80       AVVLMGKNTMMRKAIRGHLEN-----------------NPALEKLLP
D.melanogaster/51-80   AVVLMGKNTMMRKAIRGHLEN-----------------NPQLEKLLP
C.elegans/51-80        AEILMGKNTMIRKALRGHLGK-----------------NPSLEKLLP
Z.mays/52-83           SVVLMGKNTLIRRCIKVYAEKTG---------------NHTFDPLMD
H.sapiens/51-80        AVVLMGKNTMMRKAIRGHLEN-----------------NPALEKLLP
```

Figure 2B

MPPAKVDKKAKKDAFIRRFYELL***SKYD***SIALCTLENVGSLQLQQIRRSLGSNN
IMVIGKNTVVRKAVQLKSADLPTD***SKYD***WYRQFGAPKPQLASLIPHLKNKIAY
VFHNDPIFALKPKIESFVVPAPARVGTVAQKDVMIPPGP***TGMD***PSQINFFHAL
SI**STK**IQKGQIEITKEVQVCTKGKKIGNSEV***SLLE***KMNIQPFSYGMKCF***SDYD***
NGEILTEEVLSISPSVILDAFAQN**TLR**IAAVSLATGYVTAPSVPHFIQNAFKD
LAAIGMETGYKFKEIENAGQAVAVSAPAAKTETKAAAKPVVEEKPAEPEEDLD
MGDLFG

Figure 3

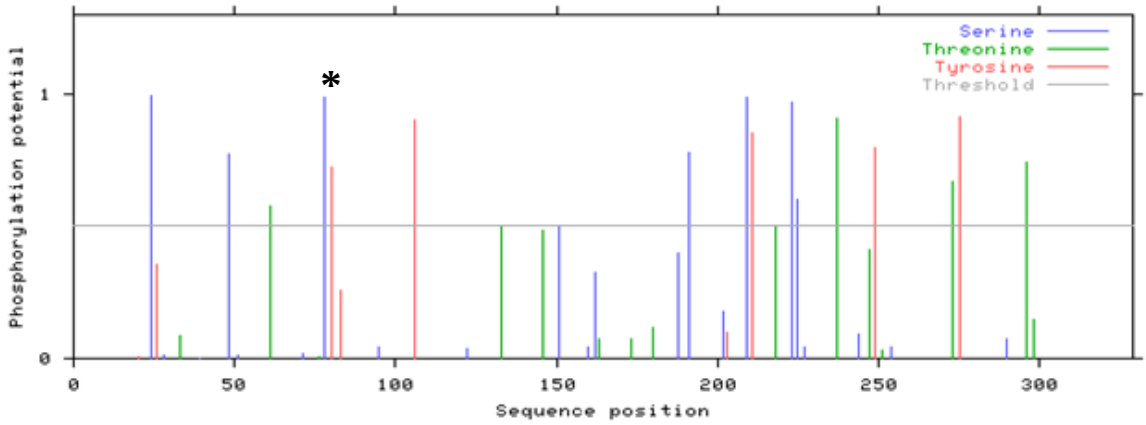
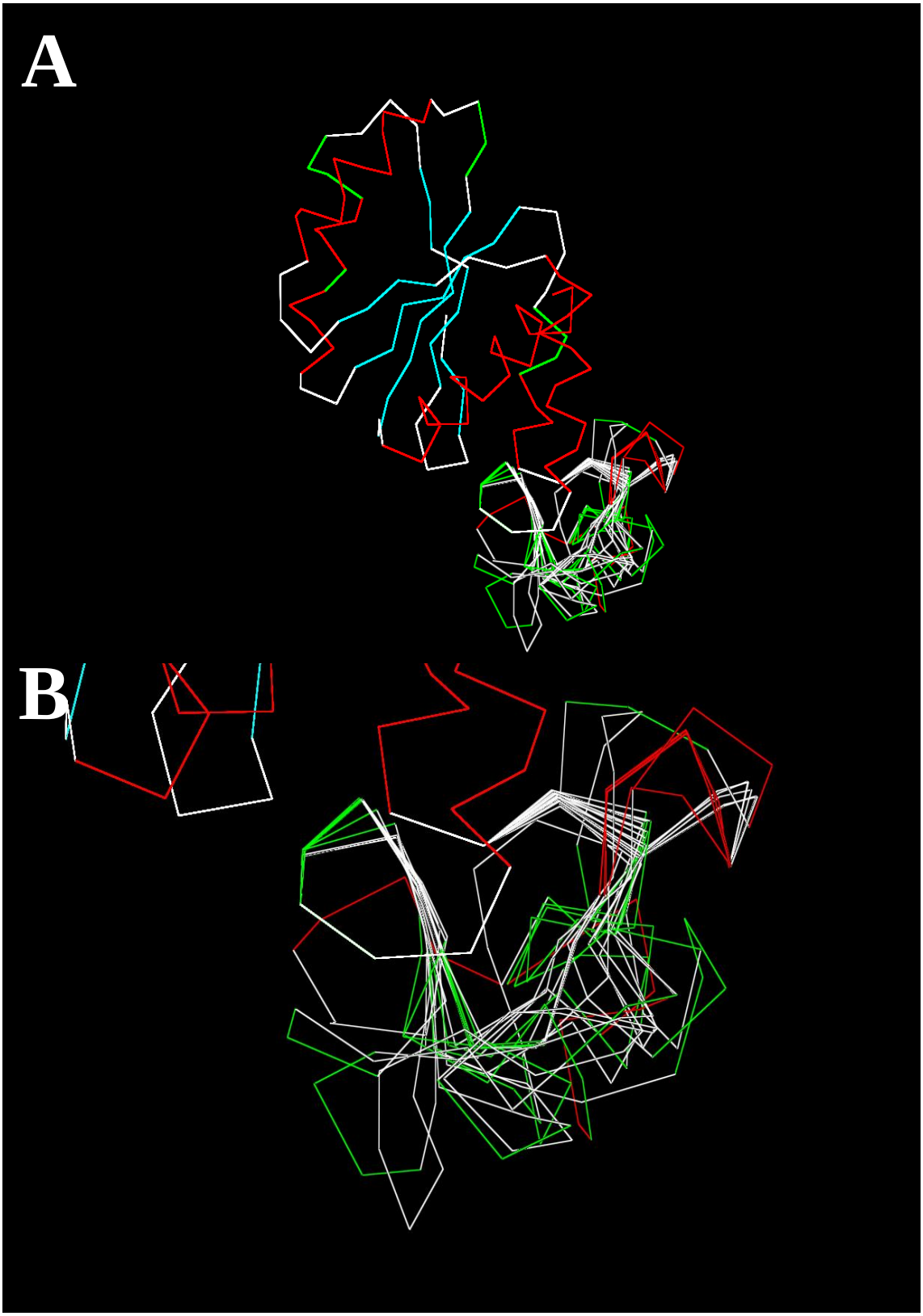

Serine
Threonine
Tyrosine
Threshold
Phosphorylation potential
1
0
*
0
50
100
150
200
250
300
Sequence position

Figure 4


A
B

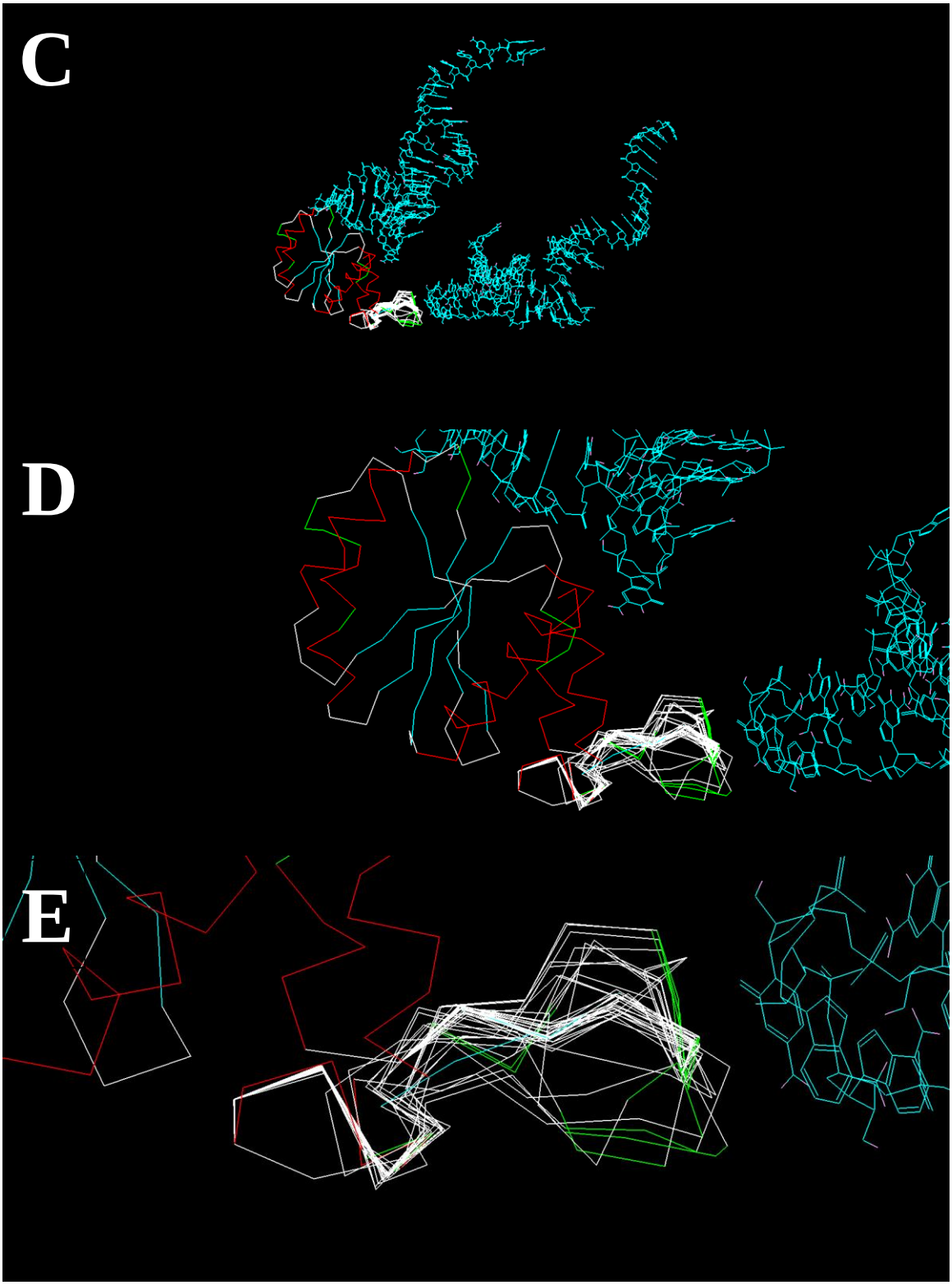


Figure 5

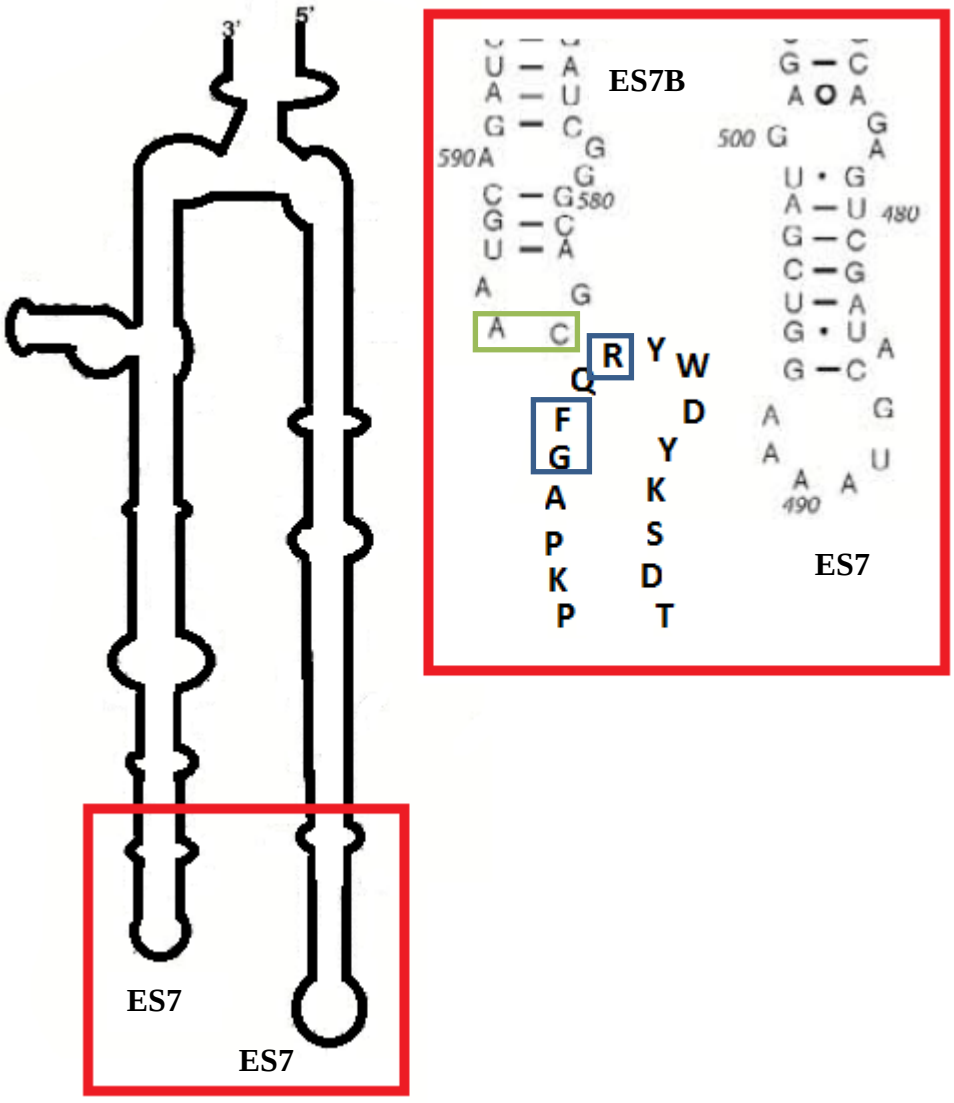


Figure 6

Model	H Bonding Atoms	Model	H Bonding Atoms
1	T:GLN79:HE21 - 1:C34:O2	6	T:GLN79:HE21 - 1:C34:O2
	T:GLY81:HN - 1:C34:N3		T:GLY81:HN - 1:C34:N3
2	T:GLN79:HE22 - 1:C34:OP2	7	T:GLN79:HE21 - 1:C34:OP2
	T:PHE80:HN - 1:C34:N3		T:GLY81:HN - 1:C34:N3
	T:GLY81:HN - 1:C34:O2		
	T:GLY81:HN - 1:C34:N3	8	T:GLY81:HN - 1:C34:O2
			T:GLY81:HN - 1:C34:N3
3	T:GLN79:HE21 - 1:C34:O2		
	T:GLY81:HN - 1:C34:O2	9	T:GLN79:HE21 - 1:C34:O2'
	T:GLY81:HN - 1:C34:N3		T:GLY81:HN - 1:C34:O2
			T:GLY81:HN - 1:C34:N3
4	T:ARG78:HH12 - 1:A35:O4'		
	T:ARG78:HH21 - 1:A35:O5'	10	T:GLY81:HN - 1:C34:O2
	T:GLN79:HE22 - 1:C34:OP2		T:GLY81:HN - 1:C34:N3
	T:GLY81:HN - 1:C34:O2		
	T:GLY81:HN - 1:C34:N3		
5	T:GLN79:HE21 - 1:C34:O2		
	T:GLY81:HN - 1:C34:N3		

Figure 6 (continued)


O.trifallax/740-1178 GTCAAAAGACTTGAAATCGTTGAGAAGGAAGGGGTAGAAATTTATTCTTC
P.tetraurelia/694-947 GTTAAAAGACTTGAAATCGTTGAGGAGAAAAGCG----------
T.thermophila/390-823 GCTAAAAGACTTGAAACCGTTGAGAAGGAAGCTGTAGAAGAGCAATAAAC
T.pyriformis/379-811 GTTAAAAGACTTGAAACCGTTGAGAAGGAAGCTGTAGAAGAGCAATAAAC
O.trifallax/740-1178 GGTGCATGCAGAGTTTAGTCGCCTAACCATTGCGGGCTAAGGATACGTAA
P.tetraurelia/694-947 ----------GTAG
T.thermophila/390-823 TGGACGGCGCATAAGGGGGAAGTACTAATCACTGCAGAGTCGATACGTAA
T.pyriformis/379-811 TGGACGGCGCATAAGGGGGAAGTGTTACTCACTGCGGAGTCGATACG--A
O.trifallax/740-1178 AAGGTCCTGGTTTGTACCTGGGGAAGTGACTGGGTGAGTGTGCATCGTGA
P.tetraurelia/694-947 AAGA----------GAAATGA----------
T.thermophila/390-823 AAGGTCG------ATGAGTAAGGAAATGGTACAGAACTTGCTACACCGGT
T.pyriformis/379-811 AAGGTCG------ATGAGTAAGGAAAGGACACAGAACTT-CTACGCCGGT
O.trifallax/740-1178 GCCAAGATGGGGTCGGACAGCCACAAAGGCTCTGTACAACCGGTTTCCTT
P.tetraurelia/694-947 ----------TTTCATT
T.thermophila/390-823 CAGAAGACAAAATGGGTTCAGATTGAAGG--------AGTCACCTGAGAT
T.pyriformis/379-811 CAGAAGACAAAATGAGTTCAGATTGAAGG--------AGTCACCTGAGAT
O.trifallax/740-1178 CGGGAGGCAGTGTGCGGATGTCGGTGGAGGTTCGGCCTGA---GGAAGCT
P.tetraurelia/694-947 TAGAAG----TATGTAGTTAT--GTAGGTGT----------CT
T.thermophila/390-823 CGGGCAGCAATGCAGATCAAAAGGAAAACTTCAAACTGGACTGAGGGGCC
T.pyriformis/379-811 CGGGGGTCAAACCAGATCAAAAGGGAAACTTCAGACTGGACTGAGGGGCC
O.trifallax/740-1178 TC--GGCGATCTTGGCAAAATGGTTTTTACCCACCCGTCTTGAAACACGG
P.tetraurelia/694-947 TC--TGCGGT--------AATGGTACTT----------CATAG
T.thermophila/390-823 TAAGGGCGATTTTGTCAAAATGGCTTCTACTGACCCGTCTTGAAACACGG
T.pyriformis/379-811 TAAGGGCGATTTTGTCAAAATGGCTTCTACTGACCCGTCTTGAAACACGG
O.trifallax/740-1178 ACCAAGGAGTCTAACATGTATGCGAGTGTGCTAGTGGAAAAACTAACACG
P.tetraurelia/694-947 GCCTAGCTGTA-GACACAAGTGCGAGTTTTAGGGTGGAAAAACCCGACGG
T.thermophila/390-823 ACCAAGGAGTCTATCAATTAAGCGAGTGATAGGGTGGAAAAACCCGTCCG
T.pyriformis/379-811 ACCAAGGAGTCTATCAATTAAGCGAGTGATAGGGTGGAGAAACCCGTCCG
O.trifallax/740-1178 CGTAATGAAGGTGATT----GATGCCAAGC-GCAAGC-AGCAGCATCGAC
P.tetraurelia/694-947 CGCAACGAAAGTGAGTATAAGGTGCGAATCCGTAAGA-GGCAGCATCGGC
T.thermophila/390-823 CGAAACGAAAGTGAGTACAAGGTGCCAAGCCGCAAGGTAGCAGCATCACC
T.pyriformis/379-811 CGAAACGAAAGTGAGTACAAGGTGCCAAGCCGCAAGGTAGCAGCATCACC
O.trifallax/740-1178 CGACCATGATCCTCTGGTGAAAGGTTTGAGTACGAGCATAAATGTTAGGA
P.tetraurelia/694-947 CAACCTTGATTTTCGAATGAAAGGATTGAGCAAGAGCATTTTTGGTAGGA
T.thermophila/390-823 CG-CCTTGAGTCTCCGC-GAAGGGTTCGAGGAAGAGCTTAATTGTTAGGA
T.pyriformis/379-811 CGACCTAGATTCTCCGAAGAAGGGTTCGAGGAAGAGCTTAATTGTTAGGA

Figure 7

Organism with insert	Sequence	Source:
Tetrahymena thermophila (MAC)	LPTDSKYDWYRQFGAPKP	Broad Institute TCD (predicted amino acid sequences)
Tetrahymena malaccensis (MAC)	LPNDPKYDWYRQFGAPKP	
Tetrahymena elliotti (MAC)	LPAGEKYDWYRQFGAPKP	
Tetrahymena borealis (MAC)	LPSDPKYDWYRQFGAPKP	
Ichthyophthirius multifiliis	LPEGENYDWYRQFGAPKP	Uniprot G0QS62
Paramecium tetraurelia	PKEHEYYEDLAKFGNAIK	Uniprot A0CFB3
Euplotes minuta	PIESESDFEERSKTWTPIP	Uniprot Q52H32
Euplotes focardii	PTPNEADYEERKATWTPVP	Uniprot Q52H52
Euplotes raikovi	PDPEDSDYETRSKTWTPLD	Uniprot Q52H31
Euplotes octocarinatus	PVEGEPDFEERSKTWTPLD	Uniprot F6M1F1
Oxytricha trifallax	PEETDVDYETRKDSWKECD	Uniprot J9IG69
Eufolliculina uhligi	PNAEDEDFEKRKNTWTPKP	Uniprot Q9U7P1

Table 1

Residue	Netphos 2.0	DISPHOS	NetphosK	Predicted function
S-24	*		*	cdc2 0.53
S-28			*	PKA 0.64
S-39			*	DNAPK 0.51
S-48	*		*	PKC 0.54
S-48	*		*	PKA 0.81
T-61	*		*	PKC 0.57
T-76		*		
S-78	*			
Y-80	*	*		
Y-83		*		
S-95			*	PKC 0.74
Y-106	*			
T-133			*	PKG 0.55
S-151			*	DNAPK 0.60
S-151			*	ATM 0.68
T-163			*	PKC 0.90
T-180			*	PKC 0.84
S-188			*	PKA 0.75
S-191	*		*	CKII 0.55
S-202			*	PKC 0.73
S-202			*	cdc2 0.51
S-209	*		*	CKII 0.54
S-209	*		*	cdc2 0.50
Y-211	*	*	*	INSR 0.52
T-218			*	CKII 0.58
S-223	*			
S-225	*		*	cdc2 0.52
S-225	*		*	GSK3 0.51
T-237	*			
T-247			*	cdc2 0.52
Y-249	*			
T-273	*		*	PKC 0.85
Y-275	*			
T-296	*	*		
T-298		*	*	PKC 0.83

Table 2

MANUSCRIPT 2

Phylogenetic Analysis of the Eukaryotic 60s Ribosomal Phosphoprotein P0 of the Ciliophora: Large Scale Tree-Building and Conserved Domain Analysis.

Giovanni Pagano and Linda A. Hufnagel

Department of Cell and Molecular Biology , University of Rhode Island, Kingston, Rhode Island, 02881

Manuscript in preparation for publication

ABSTRACT

The large subunit ribosomal protein, phosphoprotein P0, is a necessary component for protein elongation factor recruitment. Orthologues of P0 are present in both prokaryotic and eukaryotic species, and the protein is thought to be one of the most highly conserved ribosome proteins. In this study, we investigated if P0 could serve as a good target for phylogenetic studies by itself, and if analysis of the phylogeny of P0 would reveal events during early eukaryotic evolution, as well as the evolution of the Ciliophora. P0 and L10 protein sequences from organisms representing the major eukaryotic supergroups were aligned and used to build phylogenetic trees based on the entire protein, as well as the individual functional protein domains of P0. We found that P0 could provide support for higher-level taxa, but failed to provide strong support for the earliest roots of the trees. The ciliates could be resolved into previously defined Classes, but the monophyly of the Alveolata Group was not supported in all of the trees. Domain trees of P0 seemed to indicate that the *C*-terminal 60S region may contribute significantly to P0 diversity, while the *N*-terminal L10 region appeared to be more conserved in eukaryotes. We also discuss how the phenomenon of long-branch attraction may have factored into our results, as well as how it could be avoided in future phylogenetic studies on P0.

INTRODUCTION

Ribosomal phosphoprotein P0 (P0) is a component of the 60S subunit of the eukaryotic ribosome. P0 is able to form a "stalk" complex, with the phosphoproteins P1 and P2, that interacts with extra-ribosomal elongation factors (EF-1α and EF2; EF-Tu and EF-G in prokaryotes) as part of the "GTPase-associated center" (Uchiumi et al,

2002). P0 is present in organisms from all three domains of life; the P0 analog in eubacteria is known as L10, while the archaebacterial equivalent is also called P0. While the exact composition of the ribosomal stalk varies between the three domains of life, the stalk always contains a single copy of L10/P0, acting as a scaffold for other phosphoproteins, usually P1 and P2 (L7/L12 in eukaryotes) (Gordiyenko et al, 2010). The stalk interacts with the ribosomal protein L12 (L11P in eukaryotes) via its *N*-terminal (Nomura et al, 2006). The stalk also forms two contacts with the 23S/26S ribosomal RNA, at positions 1070 and 2660 (*E. coli* numbering) (Uchiumi et al, 2002).

The P0 sequence can be divided into three functional domains, two of which have been identified in PFAM (http://pfam.sanger.ac.uk/) as conserved domains (Remacha et al, 1995). The first domain is the L10 region [PF00466], located near the *N*-terminal of P0. This domain binds to H43-H44 of the 26S rRNA, tethering the stalk to the large ribosomal subunit. The L10 region is present in all three domains of life. The second region is the 60S region [PF00428], found near the *C*-terminal end of P0. This region contains the alpha helices that provide the binding site for P1/P2 dimers, as well as a highly conserved peptide (consensus sequence SD(D/E)DMGFGLFD) at the very end of the protein. While eubacteria lack the 60S domain, the conserved peptide is shared between all three domains of life, and is thought to be involved in the recruitment of elongation factors to the ribosome (Too et al, 2009; Nomura et al, 2012). The third region (middle region, or MID) lies between the L10 and 60S regions, is not present in eubacteria, and is unclassified in PFAM. There is little known about its function, although it has been hypothesized that the region is involved

in binding to EF2 (Santos et al, 2004; Justice et al, 1999). Recently published cryo-EM structures of human and *Drosophila* ribosomes with EF2 attached corroborate this hypothesis, as P0 contains several residues in the middle region that form contacts with EF2 (Anger et al, 2013).

P0 is also thought to be one of the 29 most highly conserved eukaryotic ribosomal proteins that form the core of the universal eukaryotic ancestor (Harris et al, 2003). Generally, phylogenetic studies on eukaryotes have been based on the sequence of small subunit ribosomal RNA (c.f. Cavalier-Smith, 1987; Doolittle, 1987; Woese, 1987; Zillig, 1987) or more recently, concatenated alignments of highly conserved genes (c.f. Parfrey et al, 2009; Katz et al, 2012). One of these concatenated gene studies recently focused on ribosomal proteins, but only small subunit proteins were utilized (Leigh and Chang, 2012). Because of its highly conserved nature, P0 may provide a valuable addition to these phylogenetic studies. In an early phylogenetic analysis, Liao and Dennis (1994) showed that L10/P0 sequences could be used to distinguish between eubacteria, archaebacteria and eukaryotes. More recently, Pucciarelli et al (2005) concluded, from a study on the P0 sequences of a limited number of organisms, that P0 could be useful for investigating "the phylogenetic origin of early eukaryotes". Today, many more sequenced eukaryotic genomes are available; therefore, a much more comprehensive and detailed analysis of the evolution of L10/P0 is possible, with a greatly improved opportunity for discovering new information about early eukaryotic lineages.

Tetrahymena thermophila is a unicellular eukaryotic microorganism that belongs to a Phylum of protists known as the Ciliophora (ciliated protists, aka ciliates).

Along with the apicomplexan parasites (e.g. *Plasmodium*, *Toxoplasma*, *Eimeria*) , the Ciliophora belong to a protistan clade known as the Alveolata, which in turn has been proposed to be part of the "SAR" (Stramenopiles, Alveolata, Rhizaria) Supergroup of eukaryotes (Adl et al, 2012). The ciliates are unique in that they contain two different kinds of nuclei with two differing genomes, a vegetative, transcriptionally active macronucleus (MAC) and a genetic, transcriptionally silent micronucleus (MIC) (c.f.Karrer, 2000). So far, only macronuclear genes have been utilized in phylogenetic studies, because gene predictions have only been carried out on macronuclear genome sequences, and because gene expression is almost exclusively limited to the macronucleus.

The amino acid sequence of the P0 ortholog of *T. thermophila* (TtP0) was originally obtained through preliminary genome sequence analysis and verified through PCR-based methods by Pucciarelli et al (2005). Further characterization by gene sequence analysis and immunocytochemistry was more recently carried out in our laboratory (Canton et al, 2009; Schumacher et al, 2009; Schumacher et al, 2010a, b, c; ms in preparation). Through Clustal W-based sequence alignments of TtP0 with P0 sequences from ciliates and other organisms, it was revealed that an additional 15-17 amino acid-long insert is present in the L10 region of *T. thermophila* and other ciliates. However, this insert was not found in any other prokaryotes or eukaryotes. Alignments that included a larger sample of eukaryotes showed that a smaller, apparently unrelated insert is present in the same location in members of the Kinetoplastida, an Order of excavate protists (Schumacher et al, 2009).The ciliate-specific insert was also noted more recently in *T. thermophila* by Klinge et al (2011).

Through homology modeling experiments, evidence was provided that the insert of *T. thermophila* may interact with expansion segment 7 (ES7) of the 26S ribosomal RNA of *T. thermophila* (Pagano et al, ms in preparation). This evidence for a functional role of the insert suggests that it may be useful for phylogenetic studies on the early diversification and systematics of the Ciliophora.

In the present study, we wanted to obtain more information about the early evolution of eukaryotic P0, as well as about the evolution of the L10 insert in the ciliate lineage. We created sequence alignments and phylogenetic trees using L10 and P0 protein sequences from a wide variety of eukaryotes and prokaryotes.
Trees were created from complete and modified P0/L10 sequences, as well as from each of the three functional domains, L10, 60S and MID. These trees were then compared to the taxonomic classifications proposed by Adl et al (2012), and to phylogenetic trees based on other methods, such as concatenated sequences of conserved genes and small ribosomal proteins. We provide evidence that P0 may be useful for assigning ciliates to different clades, and that the later branches of P0's evolution are consistent with other phylogenetic studies. The early stages of P0's evolution in eukaryotes are still ambiguous after this study.

MATERIALS AND METHODS

P0 homologue identification:

The TtP0 protein sequence was obtained from NCBI via the Tetrahymena Genome Database website (ciliate.org). Using this sequence, predicted and verified P0/L10 protein sequences for eukaryotes, archaebacteria and eubacteria were collected from NCBI and UniProt (species selected are given in Table 3) through BLAST

searches. The nucleotide sequence for *Goniomonas avonlea* was generously provided by Dr. Eunsoo Kim of the American Museum of Natural History (New York, NY). The two similar, but not identical, putative nucleotide sequences for P0 of *Stentor coeruleus* that were derived from the same sequenced genome were provided by Mark Slabodnick of the University of California, San Francisco.

Nucleotide sequences were translated into predicted amino acid sequences using the ExPasy translate tool (http://web.expasy.org/translate/), then manually inspected to determine the correct reading frame. To confirm the translation, a PFAM motif scan was performed on each of the sequences, to verify the presence of the L10 and 60S regions. ClustalW alignments of the translated sequences to TtP0 were also performed to determine if there were any extra amino acids beyond the start and stop codons that needed to be removed (Larkin et al, 2007).

P0 and L10 sequence alignments:

The TtP0 amino acid sequence was aligned against P0/L10 sequences from 90 (eukaryotes only) or 100 (eukaryotes, archaebacteria and eubacteria) organisms using MCoffee, run under default parameters (Notredame, Higgins and Heringa, 2000). Due to a problem in MCoffee where the first input sequence (usually TtP0) was assigned a lower homology score than it should normally have, a duplicate TtP0 sequence was included in the alignment. The TtP0 duplicates always appeared at the same location in the trees, thus providing one type of control during tree building. Based on these alignments, poorly-aligned terminal regions were removed from all 101 sequences, and the remaining amino acids were realigned in MCoffee. The amino acids corresponding to positions 1-5 and 274-324 of TtP0 were removed. After the *N*- and

C- terminals were trimmed, we also removed the inserts from the ciliate and kinetoplastid P0s, and realigned the 101 sequences. For both of these edited alignments, the P0 of *T. thermophila* was arbitrarily chosen as the reference point for trimming the terminals and inserts.

In addition, the 91 eukaryotic P0 sequences were divided into three portions, based on the PFAM annotations for TtP0. These regions correspond to residues 1-124 (L10), 125-248 (MID) and 249-324 (60S) of TtP0. These three portions were realigned in MCoffee using default parameters.

Phylogenetic tree building:

The alignments described above were used to create phylogenetic trees, using both Maximum Likelihood (ML) and Fitch-Margoliash (FM, a method based on distance matrices) algorithms. The RAxML web server at CIPRES was used to construct 1000 bootstrapped ML trees under a Protein CAT model and JTT matrix, followed by a majority-rule consensus tree (Miller, Pfeiffer and Schwartz, 2010; Stamatakis, 2014). The alignments were also used to make 1000 Fitch-Margoliash trees (from distance matrices) and a majority-rule consensus tree, using the PROTDIST, FITCH and CONSENSE programs available in the PHYLIP software package (Felsenstein, 2005). The consensus trees were displayed and rooted using the Interactive Tree of Life website (Letunic and Bork, 2006). The archaebacterium *Pyrococcus horikoshii* was chosen as the root for all of the consensus trees, based on its evolutionary distance from the eukaryotes and the presence of a 60S domain.

RESULTS

P0/L10 sequence diversity:

The P0 and L10 sequences of 101 different organisms—92 eukaryotes, 4 archaebacteria and 5 eubacteria—were used as the basis for phylogenetic alignment. Four organelle-derived L10 sequences, three from nucleomorph genomes (*Bigelowiella natans, Gymnochlora stellata* and *Chroomonas mesostigmatica*) and one from a chromatophore genome (*Paulinella chromatophora*), were also included. A full list of the species used and the classes to which they belong is given in Table 3. All five of the major eukaryotic supergroups identified by Adl et al (2012) (SAR, Archaeplastida, Excavata, Amoebozoa and Opisthokonta) were represented in the alignments, but emphasis was placed on species from the Alveolata, a subgroup within the SAR. These include 13 ciliate species representing three of the eleven different Classes (Lynn, 2002)—the Oligohymenophorea (*Tetrahymena, Paramecium* and *Ichthyophthirius*), the Heterotrichea (*Eufolliculina* and *Stentor*) and the Spirotrichea (*Oxytricha* and *Euplotes*). Also representing the Alveolata were 13 species from the Phylum Apicomplexa (three *Babesia, three Plasmodium,* two *Theileria, Toxoplasma*, *Neospora,* two *Cryptosporidium and Eimeria*). Also to be noted, the Kinetoplastida, an Order belonging to the supergroup Excavata, were represented by five species, three *Leishmania* and two *Trypanosoma*.

Trees derived from complete and trimmed L10 and P0 sequences:

After a MCoffee alignment of the complete P0 and L10 sequences was performed, we observed that the *N*- and *C*-terminals contained a significant amount of gaps and were poorly-aligned, compared to the rest of the protein. To gauge the effect of these poorly aligned terminals on the resulting trees, we removed them from the

sequences and realigned the remaining sequence data to produce a "trimmed terminals" alignment. Finally, the region containing the ciliate and kinetoplastid inserts was removed, along with the terminals, to determine what effect the presence or absence of the insert had on the quality of the trees. This produced a third "trimmed terminals and insert" alignment. All three alignments were used to build 1000 ML and FM trees (Figures 8-13).

Maximum Likelihood Trees:

Complete P0 tree: Within this tree, the apicomplexans formed a monophyletic group with reasonable support (between 58 and 100 percent) for its terminal nodes (Figure 8). Rather than forming a single group however, the ciliates instead fragmented into three separate groups, based on class. The Oligohymenophorea separated out close to the archaebacteria, near the base of the tree. The Spirotrichea associated closely with the kinetoplastids and the Heterotrichea grouped with the stramenopile supergroup and *Bigelowiella natans,* the lone rhizarial representative in this study.

The Excavata also split up across the tree. As noted above, the kinetoplastids were found on the same branch as the spirotrich ciliates, whereas *Giardia lamblia and Giardia intestinalis* were grouped with the slime molds of the supergroup Amoebozoa. Finally, the remaining excavates (*Trichomonas vaginalis* and *Hordeum meleagridis*) were located on the same branch as the eubacterial L10 sequences, which were situated on an exceptionally long branch.

The rest of the eukaryotic P0s generated monophyletic branches. The opisthokonts (including the fungi) and the Archaeplastida formed monophyletic

branches far from the archaebacterial root of the tree. As with the other groups, support values for the more terminal branches were reasonably strong , with values ranging from 61% to 100%, whereas many of the basal branches exhibited less than 50% support (i.e. no support value shown), indicating less certain placement on the tree.

Trimmed Terminal P0 tree: Several differences were observed between these trees and the trees derived from whole P0 sequences. Notably, the alveolates were monophyletic, with the ciliates contained within a clade that included the apicomplexans (Figure 9). The alveolate clade consisted of two subgroups—the Oligohymenophorea and Spirotrichea in one group, and the Heterotrichea and Apicomplexans in the other. The excavate kinetoplastids, too, moved to a different location than the previous trees. Rather than grouping with the spirotrichs, they formed a branch with *B. natans* and two cryptophytes, *G. avonlea* and *Guillardia theta*. The remaining excavates (*G. lamblia*, *T.vaginalis* and *H. melesgridis*) formed two neighboring branches, located closer to the archaebacterial root. The eubacterial sequences were found on a longer branch, close to the L10-like nucleomorph sequences near the base of the tree. The stramenopiles moved also, farther away from the heterotrichs, towards the opisthokonts; they still formed a single clade as in the previous tree. Finally, the Archaeplastida and Opisthokonta remained in the same location as they did on the whole P0 trees, and support values for these clades were consistent between the trees.

Trimmed Terminals and Insert tree: Overall, the basal branches of this tree appeared shorter than in the other trees, which is likely an effect of removing most of

the poorly-aligned amino acids from the input sequence alignment (Figure 10). The bootstrap support values changed slightly compared to the other trees, though some branches retained their consistently strong values. The ciliate clade once again fragmented into three groups—the spirotrichs (in close association with the kinetoplastids), the heterotrichs (in association with the apicomplexans and *P. marinus*), and finally the oligohymeophorea. The remaining excavates also moved farther apart; *T. vaginalis* and *H. meleagridis* remained close to the archaebacteria, while the two *Giardia* species formed a branch with the nuclear-derived P0 from *B. natans.* Meanwhile, the eubacteria formed a very long branch, the longest among the three trees, near the archaebacteria and *Entamoeba dispar,* one of the Amoebozoa. The positions of other clades on the tree were consistent with their locations in the Trimmed Terminals tree.

Fitch-Margoliash Trees:

When interpreting the results of the FM trees, it is important to note that the branch lengths are based on the support values, where longer branches represent higher bootstrap values. Therefore, FM trees cannot be used to predict the amount of changes that may have occurred between different P0s, or to determine if long-branch attraction has occurred in the tree. Much like the ML trees, the FM trees (Figures 11, 12 and 13) have moderate to strong terminal branch support values, and much weaker basal branch support.

Whole P0 Tree: The spirotrichs and kinetoplastids are closely yet weakly related on this tree, and the oligohymenophorea are grouped on a nearby branch (Figure 11). A similar spirotrich-kinetoplastid association is present in the ML tree,

although it is also weakly supported. The heterotrichs form their own branch farther away from the other ciliates, unlike the ML tree where they associate with the stramenopiles. Instead, the stramenopiles and *B. natans* associate with moderate bootstrap support (53%). The apicomplexans form a monophyletic group, with terminal branch support between 50 and 100% within the clade. Also, eubacteria form a branch closer to the archaebacterial root as would be expected, compared to the long-branch seen in the ML tree. The other supergroups form clades in similar locations to the ML tree.

Trimmed Terminals tree: Unlike the ML tree, the alveolates are not monophyletic, forming three separate branches in this tree (Figure 12). Oligohymenophoreans form their own branch earlier in the tree, followed by the spirotrichs and finally the heterotrichs. The spirotrichs are very weakly associated with the kinetoplastids and the *C. mesostigmatica* nucleomorph sequence. Heterotrichs and apicomplexans associate closely in the FM tree, albeit with somewhat weak bootstrap support (28%). Support values within the apicomplexan clade have improved from the Whole P0 tree, with a range from 65% to 100%. The P0 of *B. natans* is now associated with *Goniomonas* rather than the stramenopiles, which form their own clade. Other clades are present in similar positions compared to the previous trees.

Trimmed Terminals and Insert tree: Once again, the heterotrichs and apicomplexans closely associate in this tree with a support value of 27% (Figure 13). This is much weaker support than in the ML version of the tree, which has a 52% support value for the heterotrich-apicomplexan branch. The oligohymenophorea and spirotrichs associate with a 33% support value; both associate very weakly with the

kinetoplastids (9% support). This relationship is slightly different than in the ML tree (Figure 10), where the kinetoplastids are more closely associated to the spirotrichs than the oligohymenophorea.

Maximum Likelihood trees derived from individual P0 domains:

To examine the phylogeny of different functional regions of P0, and to uncover the effect that each functional domain of P0 may have had on the protein's overall phylogeny, we divided the eukaryotic P0s into three parts, and created trees for each of the domains, based on 1000 iterations (Figures 14, 15 and 16). Only ML trees were prepared, because the DM trees did not appear to be as useful in our earlier work with three domain trees. Eubacterial L10 sequences were excluded from these trees because the Eubacteria only contain the L10 region, and would not contribute any meaningful data to the trees of the MID and 60S domains. As with the three-domain trees, bootstrap support values above 50% were observed more often for terminal branches than for more basal branches, while branch lengths were much longer than those seen in the three-domain trees.

L10 Domain: The ciliate groups resolved into two uneven parts (See Figure 14). The heterotrichs and spirotrichs were located closer to the other alveolates than the oligohymenophorea, which formed a group with *E. dispar* (Amoebozoa). The kinetoplastids were located on a long branch at the top of the tree, near *B. natans* and in proximity to other excavates and the stramenopiles. One other notable change was the interruption of the opisthokonts by a long branch containing the Dictyostelia species and the other Amoebozoa representatives except for *E. dispar*.

MID Domain tree: The spirotrich and heterotrich ciliates were closely associated, while the Oligohymenophorea were more distantly situated, forming a branch with the kinetoplastids (Figure 15). The apicomplexans were distant from all three classes of ciliates, forming a branch near the root of the tree and showing slightly more fragmentation than in other trees. As for the other excavates, the *Giardia species* formed a very long branch near the apicomplexans, spirotrichs and heterotrichs. *T. vaginalis* and *H. meleagridis* were on the same branch as *G. avonlea* and *G. theta*. Stramenopiles, Archaeplastida and opisthokonts were all monophyletic, as in the other trees.

60S Domain: Unlike in the trees from other regions, *O. trifallax* split off from the spirotrichs to form a group with the oligohymenophorea, kinetoplastids and Amoebozoa (Figure 16). On a nearby branch, the other spirotrichs, the heterotrichs and stramenopiles were grouped together. Many of the branches in this larger group are longer than other branches in the tree. The kinetoplastids were associated more closely with the ciliates than the apicomplexans, as in the whole P0 tree. Also, the apicomplexans were located closer to the fungi, and *P. marinus,* formed a long branch near the Viridiplantae. The excavate clade was quite fragmented in this tree, forming long branches in three separate regions of the tree.

DISCUSSION

Topology of the Maximum likelihood three-domain trees:

Overall, the terminal branches had strong support values, suggesting that, for the ciliates, P0 may be useful for distinguishing species from each other and for identifying the class to which each species belongs. Also, while unicellular eukaryotes

were inconsistently positioned, the Viridiplantae (green plants) and Opisthokonta (both single-celled fungi and multicellular animals) consistently grouped near the top of the tree, far away from the archaebacterial root. This observation is made stronger by the large number of opisthokonts sampled. Even though P0 is a highly conserved protein, it still appears to provide limited information about the early evolution of the eukaryotes, as there was poor support for the basal nodes, less than 50% in most cases. The poor support for these nodes makes it difficult to identify in which eukaryotic clades the P0 is more closely related to the ancestral prokaryotic L10.

Phylogeny of the ciliates (ML): We only examined P0s from three of the eleven classes of ciliates proposed by Lynn (2002); thus, any conclusions drawn for the phylogeny of the ciliates would be preliminary. However, the three ciliate classes studied (Oligohymenophorea, Spirotrichea and Heterotrichea) consistently formed separate clades, supporting the class distinctions established by Lynn and Small (1997), with strong bootstrap support in all trees. However, of the three trees based on the entire P0 sequences, only the Trimmed Terminals tree (Figure 9) showed some support for the monophyletic association of the apicomplexans and ciliates, in keeping with other evidence that supports a clade called the Alveolata (Adl et al, 2012). However, the bootstrap value was below 50% for the node linking the ciliates to the apicomplexans, and some ciliates appeared to associate more strongly with the apicomplexans, while others did not. Perhaps removing the ciliate-specific insert from the ciliate sequences weakened support for the monophyletic nature of the ciliate clade. Adl et al (2012) proposed a SAR Supergroup consisting of the Stramenopiles, Alveolates and Rhizaria. With the exception of the Whole P0 tree (Figure 8), the

stramenopiles and rhizaria did not associate very well with the alveolates. Thus, ML trees based on entire P0 sequences do not appear to provide strong and consistent support for the SAR supergroup. The position of P0 of the rhizarian, *B. natans*, appears to be quite unstable in these three-domain trees. Therefore, additional rhizarial sequences may be necessary to stabilize the branch to which *B. natans* belongs.

Topology of the Fitch-Margoliash three-domain trees:

For each of the three-domain data sets, 1000 bootstrapped trees were prepared by the Fitch-Margoliash tree building method, and used to generate three consensus trees. Both the FM and ML trees showed weak basal branch support and stronger terminal branch support values. The Whole P0 FM and ML trees place the alveolate Classes on separate branches of the tree, and show a close but weak association between the kinetoplastids and heterotrichs. The Trimmed Terminals ML and FM trees (Figures 9 and 12 respectively) show the biggest differences in the placement of the alveolates; the FM tree contains three distinct branches for the alveolates, versus a single branch with two alveolate groups. However, both still show an association between the heterotrichs and apicomplexans. The trees made from P0s without the terminals or ciliate-specific insert differ in how closely the kinetoplastids, spirotichs and oligohymenophorea are associated. The FM tree places the two ciliate groups closer together, while the ML tree group the kinetoplastids and spirotrichs together. Once again, the heterotrichs and kinetoplastids form a branch together in both types of trees.

FM versus ML trees: After comparing the trees derived from the two tree-building methods, they seem to agree on the general placement of the Alveolata and

Kinetoplastida on branches closer to the archaebacterial root of the trees. The exact topology of the branches varies somewhat between the ML and FM trees, and both methods still have problems with weak basal support values. One key difference between the ML and FM trees is the addition of branch length data in the ML trees, which makes it possible to quantify the number of changes between the P0s of different organisms. This addition seems to make the ML trees more useful for phylogenetic analyses than the FM trees. The FM trees, however, can still be valuable as an aid for analyzing trees made from other methods to see if their topology is consistent. A third tree-making method (such as Bayesian inference) that provides branch length estimations could be combined with the ML and FM trees to further strengthen the conclusions of this study.

Phylogeny of the ciliates (FM): The FM (as well as the ML) trees provide some evidence that the heterotrichs are more closely associated with the apicomplexans than the spirotrichs or the oligohymenophorea. This relationship holds even when the insert in removed from the heterotrichs, though the branch is not strongly supported in either tree. Lynn and Small (1997) noted a "bifurcation" in the Phylum Ciliophora that divides it into two Subphyla, the Postciliodesmatophora (includes Heterotrichs) and the Intramacronucleata (includes Spirotrichs and Oligohymenophorea). The Lynn and Small Subphylum split is reflected in the separation of the ciliate Classes in the FM trees. Additional P0 sequences from other Classes of the Subphyla Postciliodesmatophora and Intramacronucleata, such as the Karyorelictea and Litostomatea, will be needed in order to further explore and characterize the evolution of phosphoprotein P0 within the Ciliophora, and to see if P0

evolution is consistent with current views concerning the systematics and evolution of the Ciliophora based on other phylogenetic studies (c.f. Gao and Katz, 2014).

Phylogeny of the SAR Supergroup (FM): On the whole, the FM trees also fail to provide strong and consistent evidence for the SAR supergroup. In all three FM trees, the P0 of *B. natans* and the clade containing the Stramenopiles form terminal branches further from the archaebacteria than the alveolates, but closer to the prokaryotes than the opisthokonts and green plants. Only the Whole P0 tree (Figure 11) has them closely associated, with a moderate bootstrap value of 52%. The other two FM trees instead place the *B. natans* P0 on its own branch or with the *Goniomonas*, indicating uncertainty in its placement, similar to the ML trees, albeit to a less drastic degree. In addition, both the stramenopiles and *B. natans* are not as closely associated with the alveolates as other clades, specifically the Amoebozoa and Excavata. This lack of monophyly for the SAR Supergroup is complicated and possibly explained by the weak basal branches, as well as the lack of Rhizarial P0s in the trees. As with the ML trees, the inclusion of additional P0 sequences for the SAR could help to resolve the question concerning the monophyly of the clade.

Phylogeny of the P0 functional domains (ML only):

Overall, the topologies of the single-domain trees appeared to be quite different from those of the three-domain trees. One of the major differences is that many of the branch lengths were significantly longer in the single domain trees. This may be due to the smaller lengths of the individual domains. Since branch lengths reflect the average number of substitutions per amino acid position, having fewer possible residues to measure increases the contribution of each residue substitution to

the branch length. The phenomenon of long-branch attraction (Bergstein, 2005), however, can result in some false positioning of species or clades, but comparison with three-domain trees may help to resolve potential problems of this type.

L10 region: With regard to the L10 region tree, the oligohymenophorea form a clade that is more distinct, whereas the spirotrichs and heterotrichs exhibit a closer association. Thus, the L10 region may have diverged more extensively in the Oligohymenophorea. Surprisingly, the L10 region of *P. tetraurelia* seems to be quite distinct from that of the *Tetrahymenidae* (bootstrap value of 86% for this split). Additional P0 sequences from other species of the Peniculids may be useful in providing support for this divergence. The effect of the ciliate-specific L10 insert may help to exaggerate the branch lengths for the various ciliate groups. Further analyses in which the insert is removed and the edited L10 regions are used to build new trees may help to clarify the effect that the insert has on the tree structure. Also, there is still weak bootstrap support for a close relationship between the apicomplexans and the various classes of ciliates in this tree.

It was also noted that one group of Opisthokonta (the Supergroup that includes the multicellular animals) appeared to move to a location closer to the archaebacteria rather than further away (Figure 14). This is likely to be an artifact of the tree-building, due to the inability of the L10 region to provide significant information about the early ancestors of the modern supergroups. The placement of the *Dictyostelia* on a long branch within this group is suspect, and might be due to long-branch attraction. One possibly significant result is the clear separation of the kinetoplastids from the ciliates, given how closely the kinetoplastids cling to the

ciliates in other trees, which may be a false positive, as with the clearly erroneous association of the eubacteria with members of the Excavata, in the complete P0 trees (Figure 8).

MID region: In this tree, the ciliate, apicomplexan and kinetoplastid clades are distinct, but they fragment and disperse to different sections of the tree. The Oligohymenophorea and Heterotrichea are closer together, and the Spirotrichea form a group with the kinetoplastids in a different section of the tree. The apicomplexans lie at the root of the tree (see Figure 15), in contrast to the rest of the trees. This difference in the apparent earliest group between the L10 and MID tree is likely due to poor basal branch support in the tree, as well as ambiguity about which group should be placed first. Overall, the rest of the groupings are similar to results from the other trees, but the fragmented nature of the tree appears to reflect and may be derived from the sequence diversity of the MID region.

60S region: This is the only tree where one of the ciliate groups, the heterotrichs, becomes split up, with *O. trifallax* and the four *Euplotes* species coming to lie in quite separate locations on the tree. The branch between *O.trifallax* and the oligohymenophorea has a 53% support value, just above the cutoff. The P0 60S region of *O. trifallax* may not have undergone as many changes compared to the other two regions, making *O. trifallax* more closely related, by a small margin, to the oligohymneophoreans. In the 60S tree the ciliates group nearest to the stramenopiles; thus the 60S region tree provides the best support for the SAR clade, out of the three regional trees.

Many of the terminal (Genus or species level) branches of the 60S tree (Figure 16) are long, especially those of the ciliates. This indicates that more substitutions or changes have occurred in this region. This large amount of change may be due to the presence of repetitive sequences of amino acids (like alanine and glutamic acid) in the 60S region. Such repetitive sequences would make replication errors more likely. The 60S region of repetitive sequence has been termed the 'hinge", because it is thought to be a flexible portion of the protein necessary for interacting with the elongation factors (Gonzalo and Reboud, 2003). It is worth noting that in our sequence alignments, the 60S regions of different P0s aligned poorly, which was why the hinge sequences were removed in the Trimmed Terminal trees. The variability of the 60S region may be a large contributor to the diversification of ciliate P0s and of eukaryotic P0s in general.

The 60S region may also hold clues to how P0 evolved from L10. In their phylogenetic study of the stalk proteins, Shimmin et al (1989) suggested that this region shares homology with the stalk protein P1/L12 (described earlier), and proposed a model of P0 evolution where P0 arose from the fusion of ancestral L10 and L12 genes. A comparison of the 60S region with ribosomal proteins like P1 and P2 could be the basis of a future study, since the P1 and P2 gene/protein sequences of *T. thermophila* and many other eukaryotes have not been identified or characterized yet.

Long-branch attraction in L10 and P0:

In all trees, we observed some branches where organisms known to be evolutionarily distant were grouped together, such as the eubacteria and a couple of the excavates (see Figure 8). Long branches between prokaryotes and eukaryotes are expected, given their long history of divergence from each other. However, this

divergence does not account for the unusual placement of these branches, which are caused by a phenomenon known as long–branch attraction. Long-branch attraction occurs when two divergent sequences have undergone enough changes that they appear more homologous than they actually are, causing tree-building programs to falsely group them together (Bergstein, 2005). Bergstein reviewed four methods for tree building (maximum likelihood, maximum parsimony, distance matrix and Bayesian inference), and found that ML trees were less vulnerable to long-branch attraction. It was also noted that protein sequences were less likely than gene sequences to form false branches, due to the larger number of possible amino acids versus nucleotides.

However, even though ML methods and protein sequences were used in the present study, long-branch attraction still appeared to cause false branches to appear in all of the three-domain ML trees (Figures 8, 9 and 10). There are two likely factors contributing to their appearance; the poorly supported nature of the basal branches, and the presence of poorly-aligned terminals in the whole P0 alignment. Removing the terminals and leaving the strongly-aligned portions of P0 seemed to address some of the noise, but removing the insert seemed to reintroduce some problems, such as the fracturing of the ciliate clade. Removing poorly-aligned regions did not strengthen the support of basal branches, so another method is necessary to improve the resolution of basal branches. The simplest method might be to add more sequences to the alignment, especially in the case of fragmented clades like the Excavata and Alveolata. The sequences used in this study represent most of the excavates and ciliates whose P0s have been sequenced, so this work will need to be revisited in the

future, as sequencing projects continue. Other ways to improve the quality of the basal branches may also need to be investigated, as it is still unclear from these findings whether P0 could be utilized to trace the earliest stages of the eukaryotic tree of life.

Conclusion:

Using maximum likelihood and distance matrix methods, several phylogenetic trees were created from alignments of whole L10 and P0 sequences, as well as the individual functional domains of P0. Both methods produced trees with poorly supported basal branches and stronger terminal branches, reflecting uncertainty in the early evolution of P0. Despite the unbalanced support of the branches, the results suggest a relationship between the P0s of ciliates and kinetoplastids, although the support was not strong. Surprisingly, there was also generally poor support for a relationship between the P0s of ciliates and apicomplexans, both members of a well-established clade, the Alveolata. The postulated SAR Supergroup was also not well-represented by P0's phylogeny in the current study, although this may be partially due to representation of the Rhizaria in the tree by a single species. However, support was strong for the Genera and Classes of ciliates that had been previously established through other studies, and thus P0 may be useful as a basis for classification of organisms at higher taxonomic levels. Of the two known functional regions of eukaryotic P0, the *C*-terminal 60S region may be the most significant contributor to the evolutionary diversification of P0, while the *N*-terminal L10 region seems to be the most conserved. As new genomes are sequenced and more P0 sequences become available, it should be possible to revisit the phylogeny of L10 and P0, and draw stronger conclusions about the evolutionary transition from prokaryotic L10 to eukaryotic P0.

REFERENCES

Adl SM, Simpson AG, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, Le Gall L, Lynn DH, McManus H, Mitchell EA, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Schoch CL, Smirnov A and Spiegel FW. (2012), "The revised classification of eukaryotes". J Eukaryot Microbiol. 59(5), 429-493.

Anger AM, Armache JP, Berninghausen O, Habeck M, Subklewe M, Wilson DN and Beckmann R. (2013), "Structures of the human and Drosophila 80S ribosome". Nature 497, 80-85.

Bergsten J. (2005), "A review of long branch attraction". Cladistics 21, 163-193.

Canton S, Chu H, Corriveau J, Schumacher J and Hufnagel LA. (2009), "Orthologues in Tetrahymena thermophila of Proteins related to Human Infectious and Genetic Diseases". Proc. FASEB summer conference on Ciliate Molecular Biology, Saxton's River, VT.

Cavalier-Smith T. (1987), "The origin of eukaryote and archaebacterial cells". *Annals of the New York Academy of Sciences* 503, 17-54.

Doolittle WF. (1987), "The evolutionary significance of the archaebacteria". *Annals of the New York Academy of Sciences* 503, 72-77.

Felsenstein J. (2005), "PHYLIP (Phylogeny Inference Package) version 3.6." Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gao F and Katz LA. (2014), “Phylogenomic analyses support the bifurcation of ciliates into two major clades that differ in properties of nuclear division”. Mol. Phylogenet. Evol. 70, 240-243.

Gonzalo P and Reboud JP. (2003), “The puzzling lateral flexible stalk of the ribosome”. Biol. Cell 95, 179-193.

Gordiyenko Y, Videler H, Zhou M, McKay AR, Fucini P, Biegel E, Müller V and Robinson CV. (2010), “Mass spectrometry defines the stoichiometry of ribosomal stalk complexes across the phylogenetic tree”. Mol. Cell Proteomics 9(8), 1774-1783.

Harris JK, Kelley ST, Spiegelman GB, and Pace NR. (2003), “The genetic core of the universal ancestor”. Genome Res. 13(3),407-412.

Hiu-Mei Too P, Kit-Wan Ma M, Nga-Szse Mak A, Wong YT, Kit-Ching Tung C, Zhu G, Wing-Ngor Au S, Wong KB and Shaw PC. (2009), “The *C*-terminal fragment of the ribosomal P protein complexed to trichosanthin reveals the interaction between the ribosome-inactivating protein and the ribosome”. Nucleic Acids Res. 37(2), 602–610.

Justice MC, Ku T, Hsu MJ, Carniol K, Schmatz D and Nielsen J. (1999), “Mutations in Ribosomal Protein L10e Confer Resistance to the Fungal-specific Eukaryotic Elongation Factor 2 Inhibitor Sordarin”. J. Biol.Chem. 274, 4869-4875.

Karrer KM. (2000), “Tetrahymena genetics: Two nuclei are better than one”, in Tetrahymena thermophila, Asai DJ and Forney JD, eds. Methods in Cell Biology 62, 128-186.

Katz LA, Grant JR, Parfrey LW and Burleigh JG. (2012), “Turning the Crown Upside Down: Gene Tree Parsimony Roots the Eukaryotic Tree of Life”. Syst. Biol. 61(4), 653-660.

Klinge S, Voigts-Hoffmann F, Leibundgut M, Arpagaus S and Ban N. (2011), “Crystal Structure of the Eukaryotic 60S Ribosomal Subunit in Complex with Initiation Factor 6”. Science 334, 941-948.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG. (2007), "ClustalW and ClustalX version 2". Bioinformatics 23(21), 2947-2948.

Leigh T and Chang WJ. (2012), “Phylogenetic analyses on the evolution of eukaryotes using concatenated ribosomal protein sequences”. Proceedings of Protist 2012, Oslo.

Letunic I and Bork P. (2006), "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation". Bioinformatics 23(1), 127-128.

Letunic I and Bork P (2011), "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy". Nucleic Acids Res. 39 (suppl 2), W475-W478.

Liao D and Dennis PP. (1994), “Molecular phylogenies based on ribosomal protein L11, L1, L10, and L12 sequences”. J Mol. Evol. 38(4), 405-419.

Lynn DH. (2002), *The Ciliate Resource Archive.* http://www.uoguelph.ca/~ciliates/classification/genera.html. Accessed July 21st, 2014.

Lynn DH and Small EB. (1997), "A Revised Classification of the Phylum Ciliophora Doflein, 1901". Rev. Soc. Mex. Hist. Nat. 47, 65-78.

Miller MA, Pfeiffer W and Schwartz T. (2010), "Creating the CIPRES Science Gateway for inference of large phylogenetic trees." Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, 1-8.

Nomura T, Nakano K, Maki Y, Naganuma T, Nakashima T, Tanaka I, Kimura M, Hachimori A and Uchiumi T. (2006), "In vitro reconstitution of the GTPase-associated centre of the archaebacterial ribosome: the functional features observed in a hybrid form with Escherichia coli 50S subunits". Biochem J. 396(3), 565-571.

Notredame C, Higgins DG and Heringa J. (2000), "T-Coffee: A novel method for multiple sequence alignments". J. Mol Biol 302, 205-217.

Pagano G, King R, Martin LM and Hufnagel LA. "The Unique *N*-terminal Insert in a Ribosomal Protein, Phosphoprotein P0, of Tetrahymena thermophila: Homology Modeling Analysis". Manuscript in preparation.

Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Hau J, Martin O, Kuznetsov D and Falquet L. (2007), "MyHits: improvements to an interactive resource for analyzing protein sequences". Nucleic Acids Res. 35(Web Server issue), W433-W4377.

Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ and Katz LA. (2010), "Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life". Syst. Biol. 59(5), 518-533.

Pucciarelli S, Marziale F, Di Giuseppe G, Barchetta S and Miceli C. (2005), "Ribosomal cold-adaptation: characterization of the genes encoding the acidic

ribosomal P0 and P2 proteins from the Antarctic ciliate Euplotes focardii". Gene 360(2), 103-110.

Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Rodriguez Gabriel MA, Guarinos E, and Ballesta JP. (1995), "Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects ". Biochem Cell Biol 73, 959-968.

Stamatakis A. (2014), "RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies". Bioinformatics, open access.

Santos C, Rodriguez-Gabriel MA, Remacha M and Ballesta, JPG. (2004), "Ribosomal P0 Protein Domain Involved in Selectivity of Antifungal Sordarin Derivatives. Antimicrob. Agents Chemother. 48(8), 2930-2936.

Schumacher J, Babcock K, Canton S and Hufnagel LA. (2010), "Tetrahymena orthologue of the malaria vaccine candidate Phosphoprotein p0: Immunocytochemical localization and a co-expressed membrane protein with unique properties". Proc. Internat. Soc. of Protistol. (ISOP), Canterbury, England.

Schumacher J, Canton S, Babcock K and Hufnagel LA. (2010), "An orthologue of the apicomplexan vaccine candidate phosphoprotein P0, a conserved ribosomal protein, is also present at the cell surface in the alveolate protist, Tetrahymena thermophila". Proc. Ann. Mtg., Amer. Soc. Cell Biol. (ASCB), Philadelphia.

Schumacher J, Canton S, Babcock K, and Hufnagel LA. (2010), "An Orthologue of the Malaria Vaccine Candidate Phosphoprotein p0 in Tetrahymena

thermophila: Immunocytochemical Localization". Proc. N. Amer. Chapt., Internat. Soc. of Protistol. (ISOP), Lexington, VA

Schumacher J, Corriveau J, Canton S, and Hufnagel LA. (2009), "An orthologue of the protozoan vaccine candidate phosphoprotein p0 in Tetrahymena thermophila". Proc. N. Amer. Chapt., Internat. Soc. of Protistol. (ISOP), Bristol, RI.

Schumacher J and Hufnagel LA. "Protozoan Vaccine Candidate Homologues in Tetrahymena thermophila". Manuscript in preparation.

Shimmin LC, Ramirez C, Matheson AT and Dennis PP. (1989), "Sequence Alignment and Evolutionary Comparison of the L10 Equivalent and L12 Equivalent Ribosomal Proteins from Archaebacteria, Eubacteria and Eucaryotes". J Mol Evol 29, 448-462.

Uchiumi T, Honma S, Endo Y, and Hachimori A. (2002), “Ribosomal proteins at the stalk region modulate functional rRNA structures in the GTPase center”. J Biol Chem. 277(44):4, 1401-1409.

Woese CR. (1987), “Bacterial evolution”. *Microbiological Reviews* 51(2), 98-122.

Zillig W. (1987), “Eukaryotic traits in archaebacteria: Could the eukaryotic cytoplasm have arisen from archaebacterial origin?” *Annals of the New York Academy of Sciences* 503, 78-82.

Figure Legends:

In all trees, brackets identifying the clades of interest in this study have been provided. C (blue bracket) indicates members of the Ciliophora, A (green bracket) indicates members of the Apicomplexa, and K (red bracket) indicates members of the Kinetoplastida.

Fig. 1: The maximum likelihood consensus tree of whole L10 and P0 sequences. Inferred from the amino acid sequences of P0/L10 from 101 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 2: The Fitch-Margoliash consensus tree of whole L10 and P0 sequences. Inferred from the amino acid sequences of P0/L10 from 101 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 3: The maximum likelihood consensus tree of L10 and P0 sequences without their poorly-aligned *N*- and *C*- terminals. Inferred from the amino acid sequences of P0/L10 from 101 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 4: The Fitch-Margoliash consensus tree of L10 and P0 sequences without their poorly-aligned *N*- and *C*- terminals. Inferred from the amino acid sequences of

P0/L10 from 101 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 5: The maximum likelihood consensus tree of L10 and P0 sequences, without poorly-aligned terminals or the ciliate-specific inserts. Inferred from the amino acid sequences of P0/L10 from 101 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 6: The Fitch-Margoliash consensus tree of L10 and P0 sequences, without poorly-aligned terminals or the ciliate-specific inserts. Inferred from the amino acid sequences of P0/L10 from 101 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 7: The maximum likelihood consensus tree of the L10 region of eukaryotic P0s. Inferred from the amino acid sequences of P0 from 91 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 8: The maximum likelihood consensus tree of the MID region of eukaryotic P0s. Inferred from the amino acid sequences of P0 from 91 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent

two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Fig. 9: The maximum likelihood consensus tree of the 60S region of eukaryotic P0s. Inferred from the amino acid sequences of P0 from 91 species, with *P. horikoshii* as an outgroup. The two *S. coeruleus* sequences, labeled A and B, represent two distinct P0 hits from the same genome, while the two *T. thermophila* sequences are identical due to a quirk in MCoffee.

Table 3: A list of the eukaryotic, archaebacterial and eubacterial species represented in the phylogenetic alignments and trees, along with their Classes. Species are organized according to their Supergroups, with members of the Ciliophora and Apicomplexa placed first.

Figure 8a—Whole P0 Tree, ML

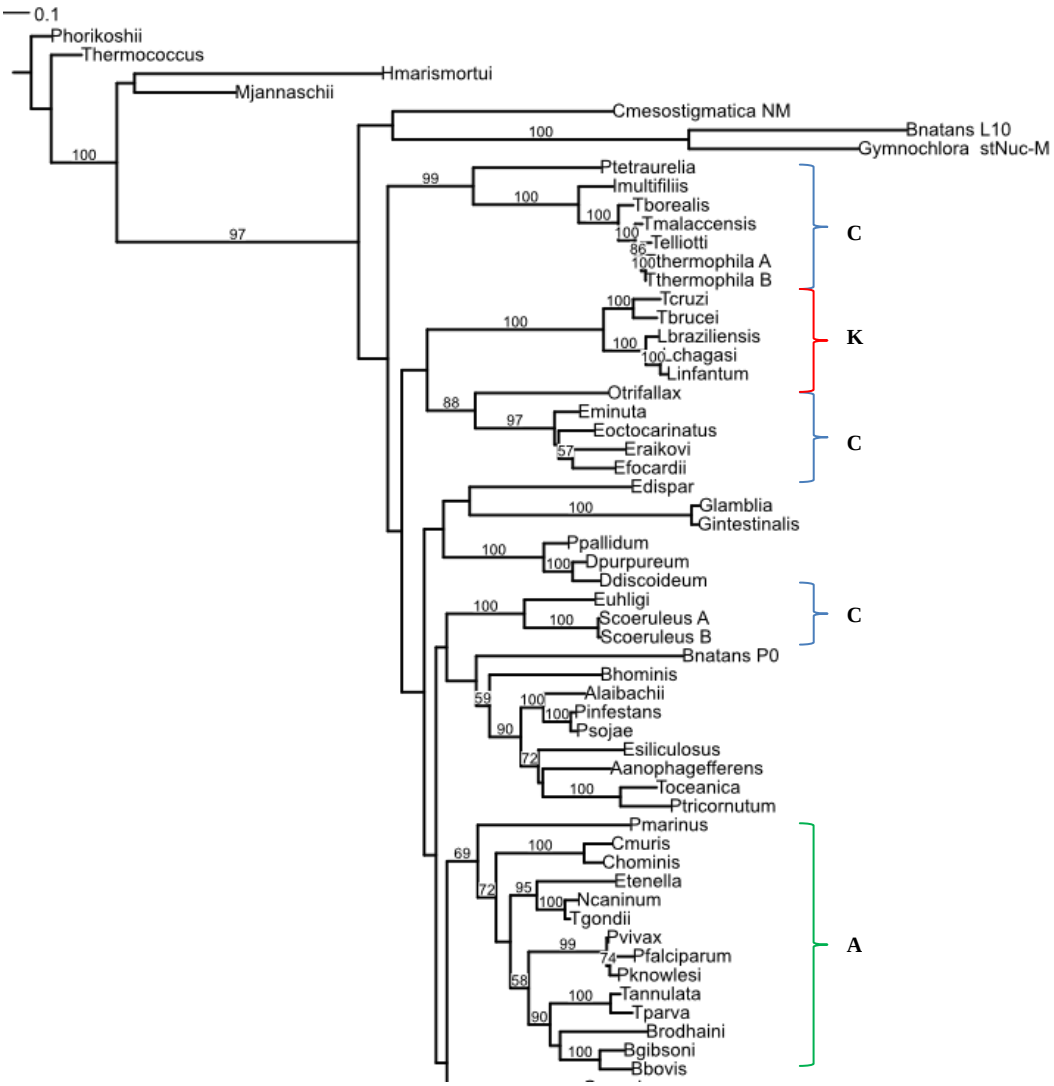


Figure 8b—Whole P0 Tree, ML

Figure 9a—Trimmed Terminals Tree, ML

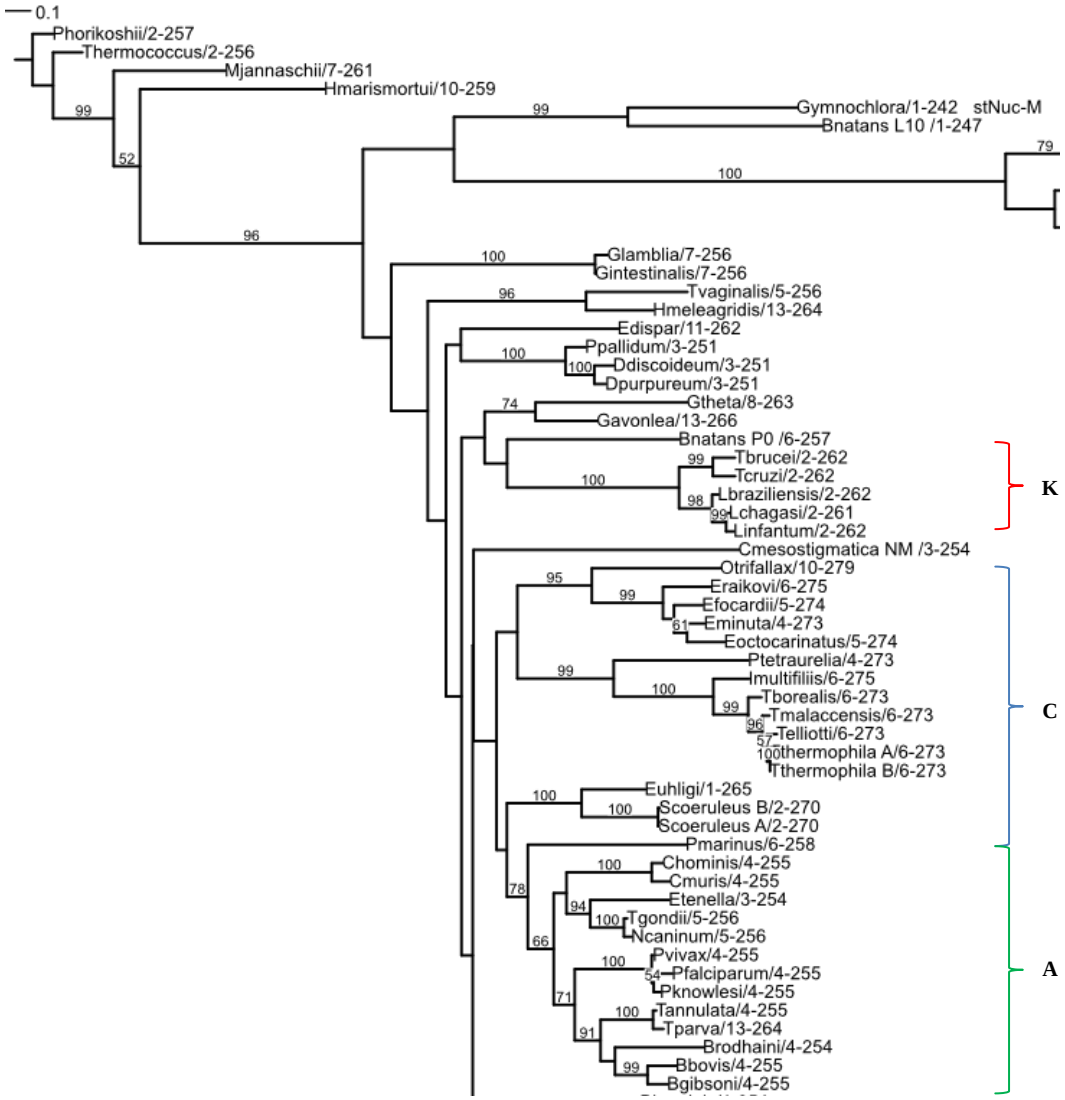


Figure 9b—Trimmed Terminals Tree, ML

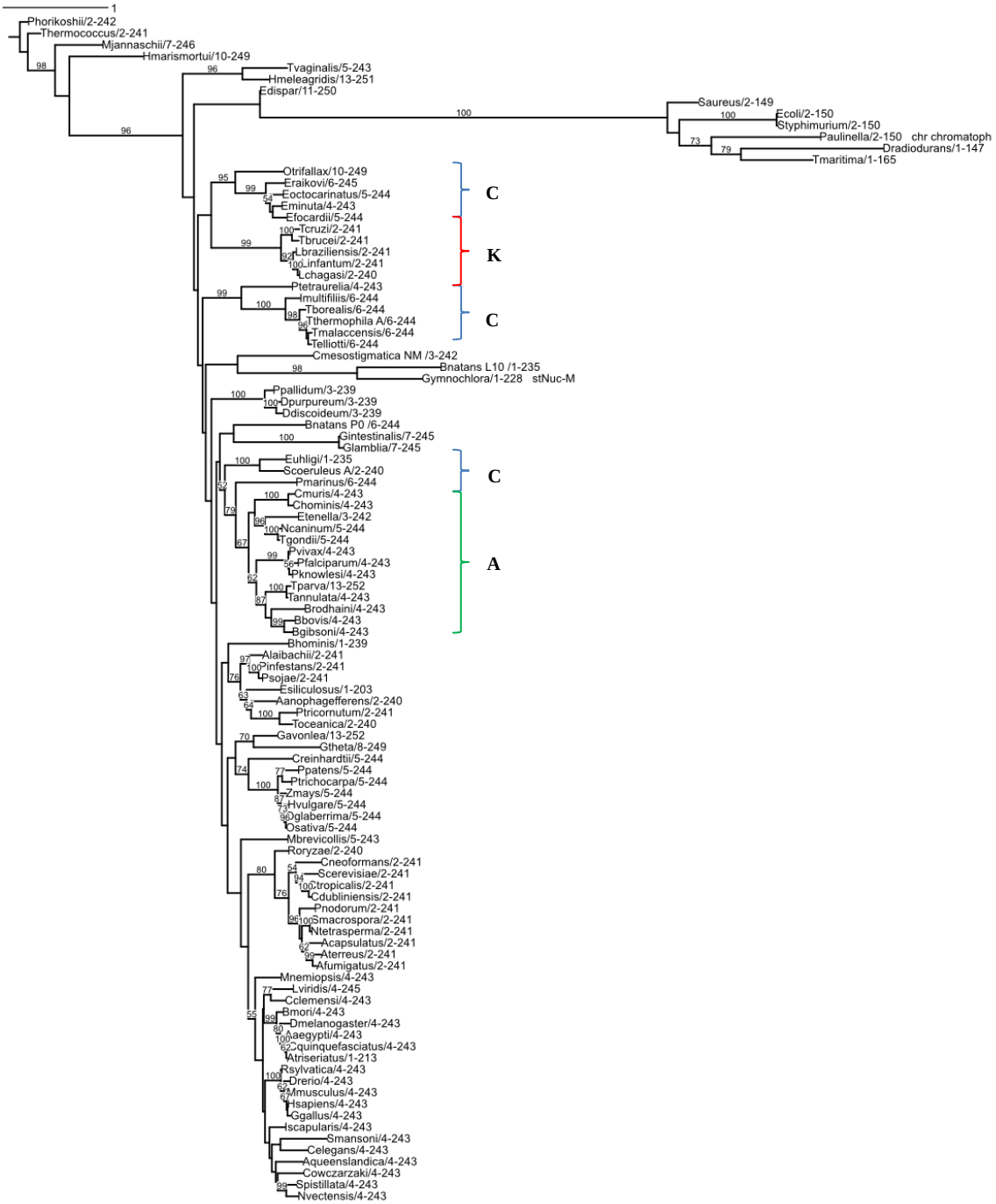


Figure 10a—Trimmed Terminals and Insert Tree, ML

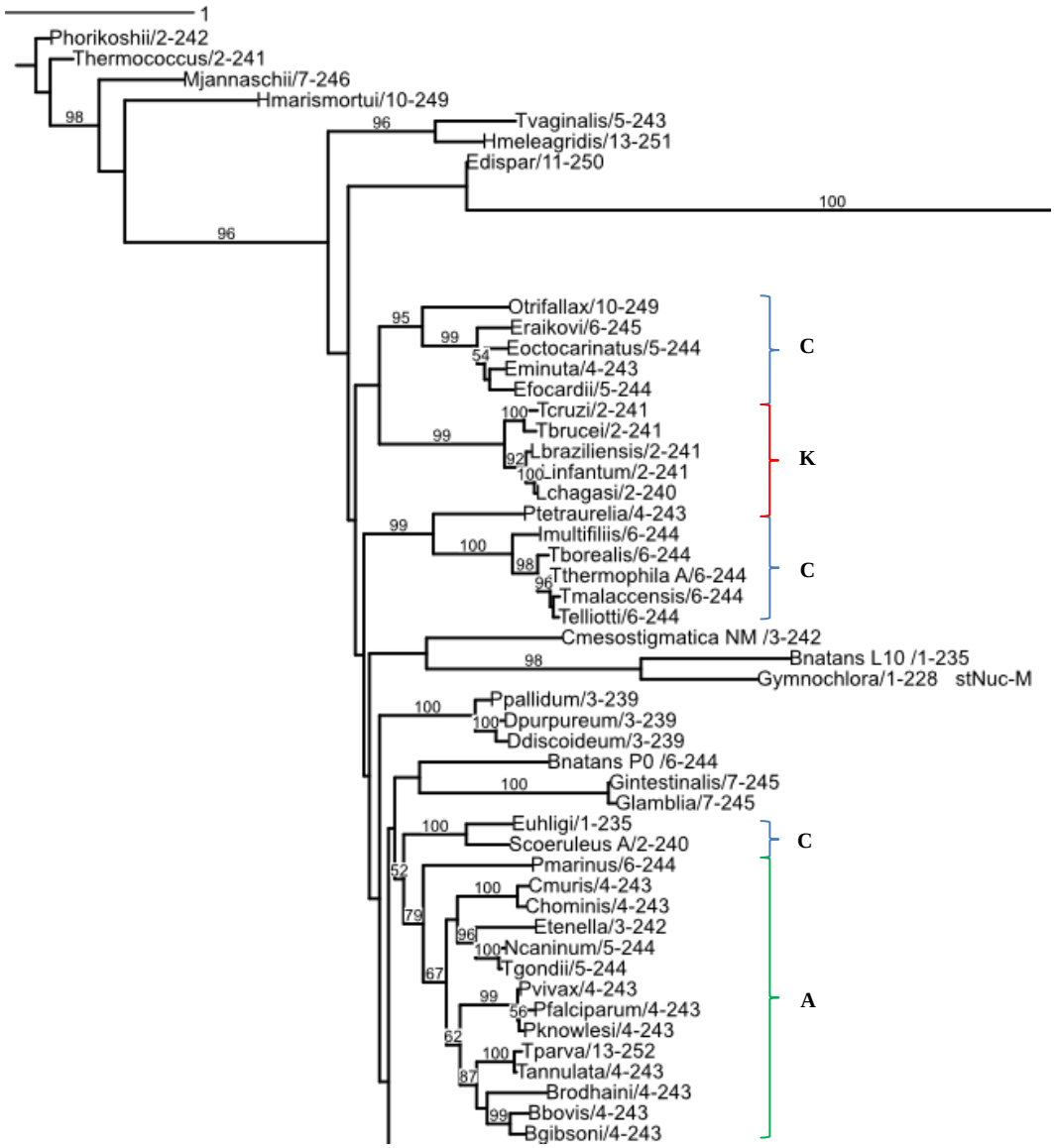


Figure 10b—Trimmed Terminals and Insert Tree, ML

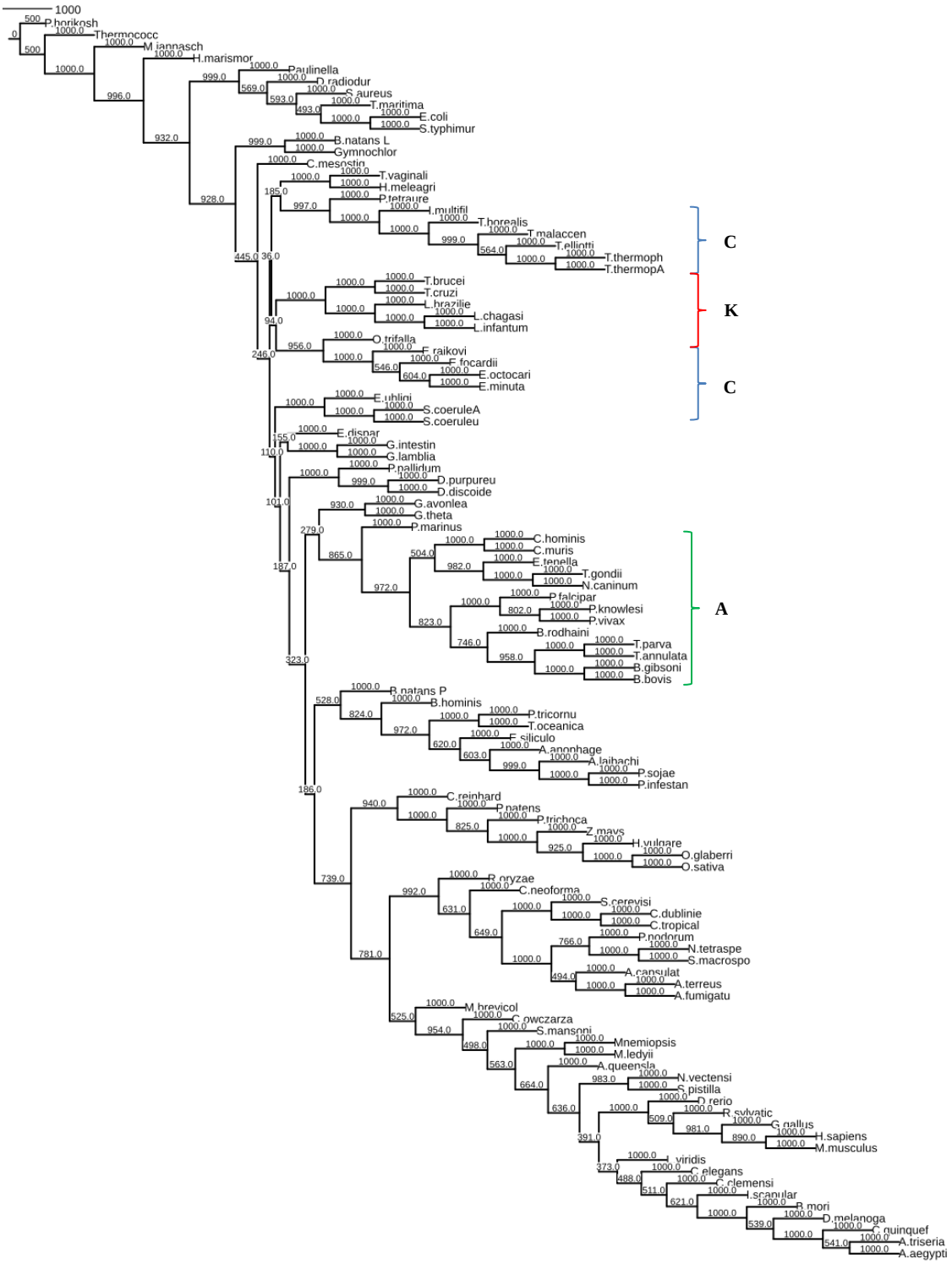


Figure 11a—Whole P0 Tree, FM

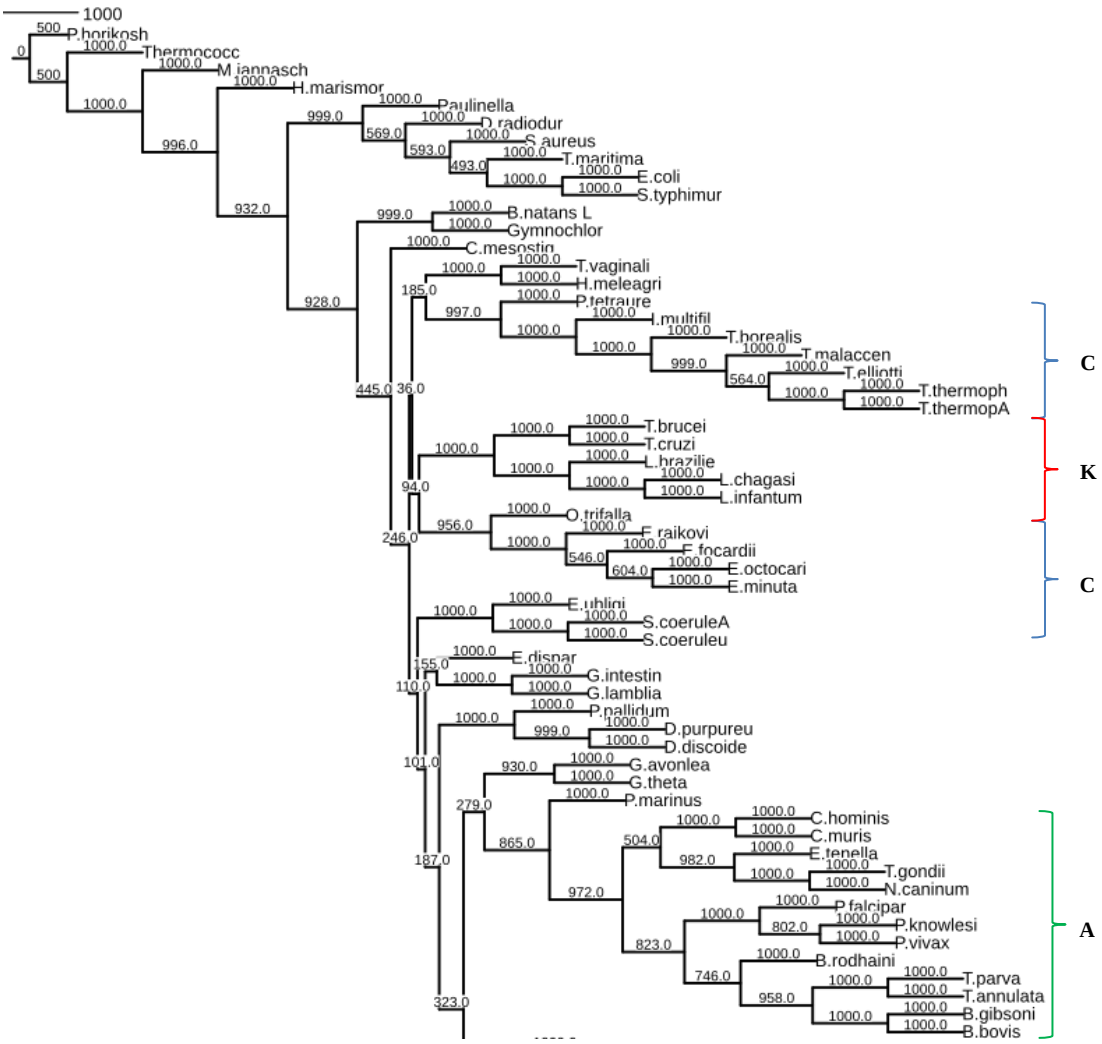


Figure 11b—Whole P0 Tree, FM

Figure 12a—Trimmed Terminals Tree, FM

Figure 12b—Trimmed Terminals Tree, FM

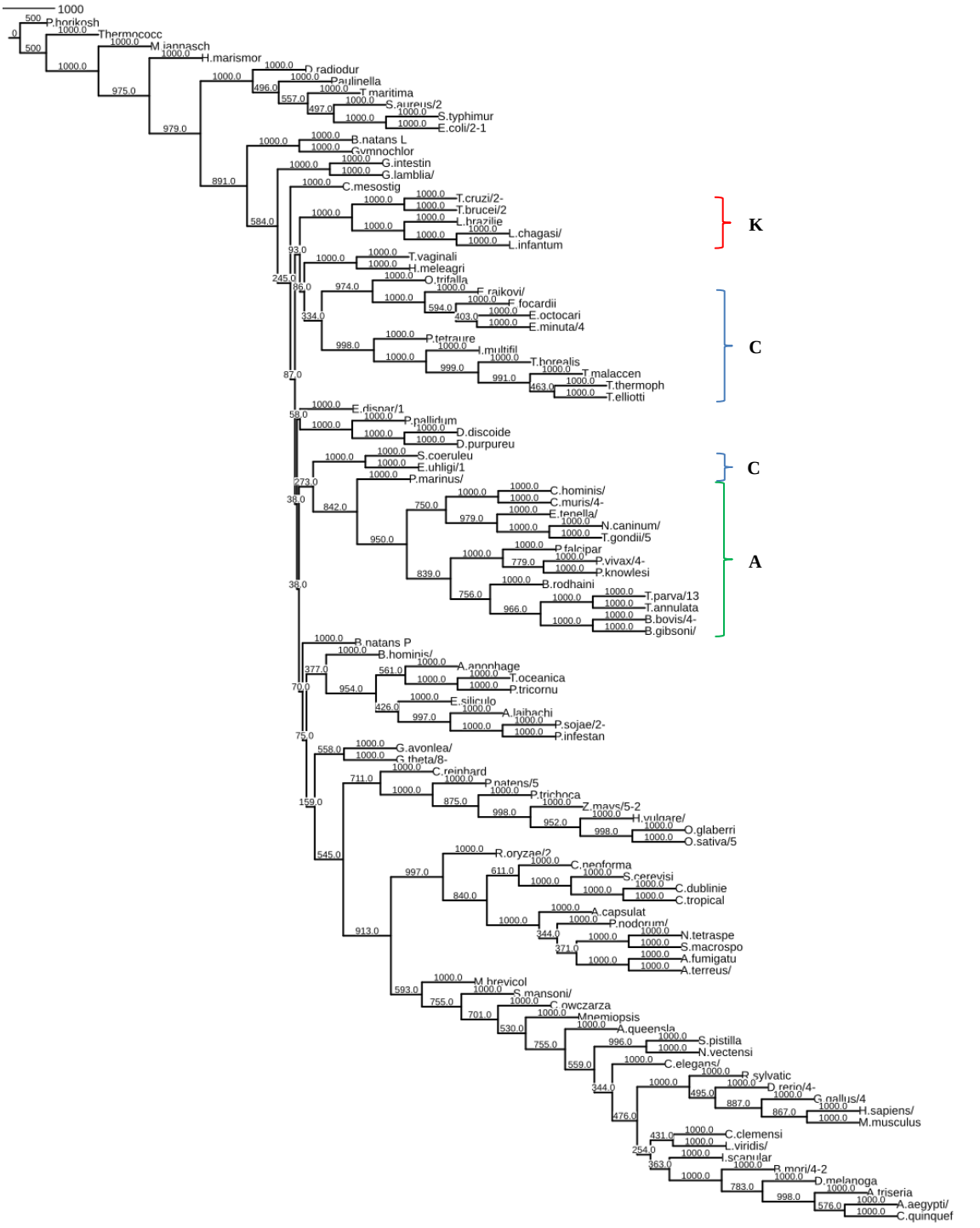


Figure 13a—Trimmed Terminals and Insert Tree, FM

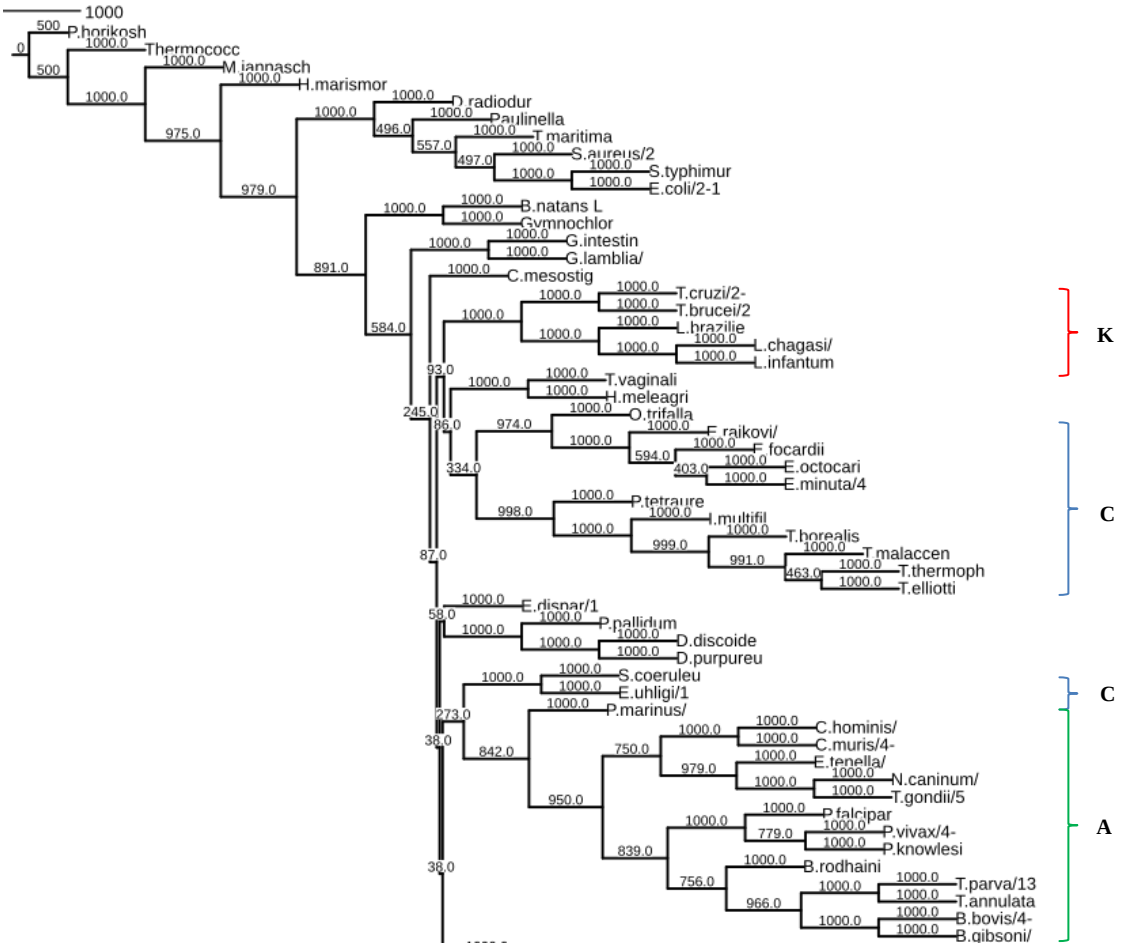


Figure 13b—Trimmed Terminals and Insert Tree, FM

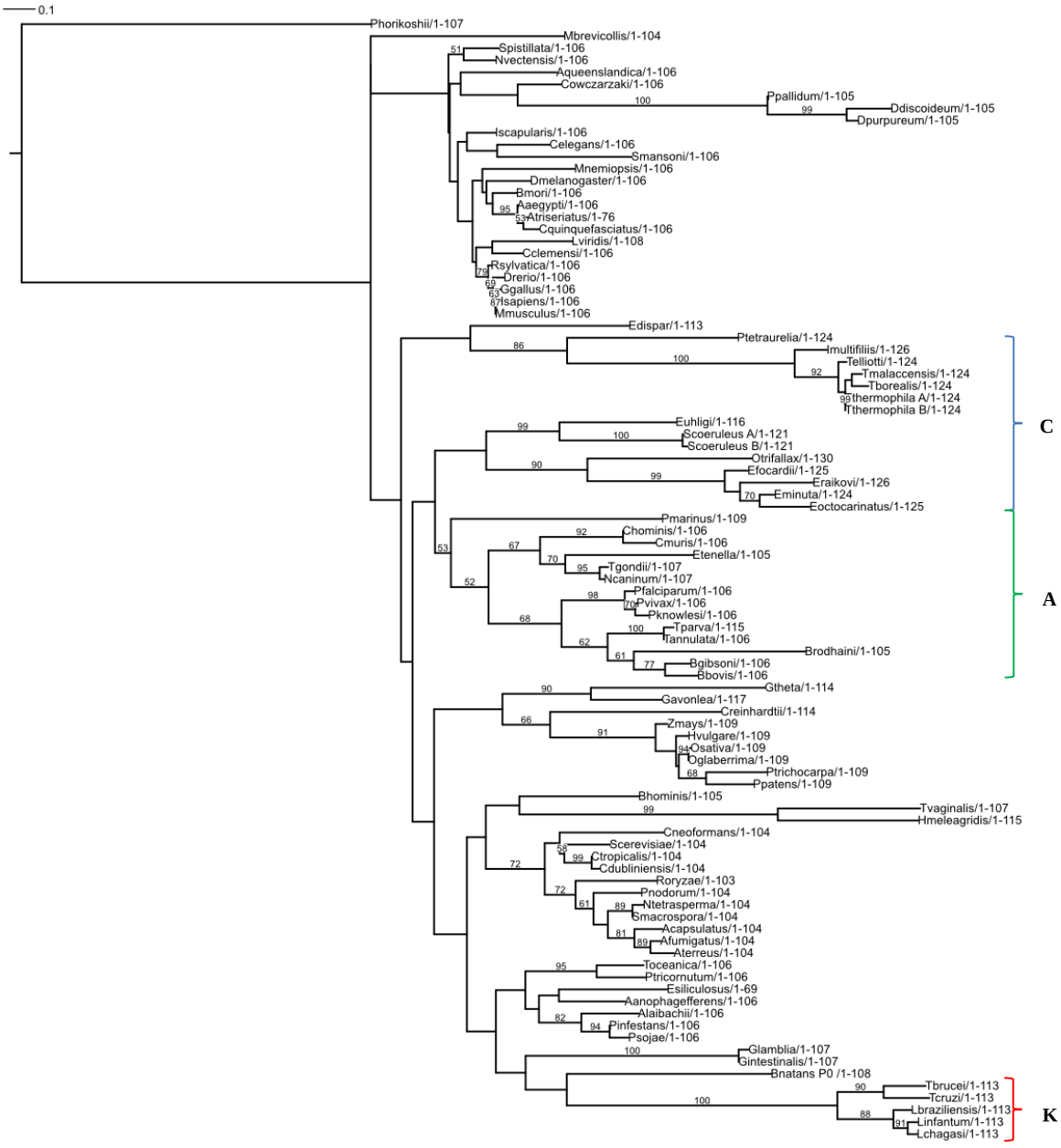


Figure 14a—L10 Region Tree, ML

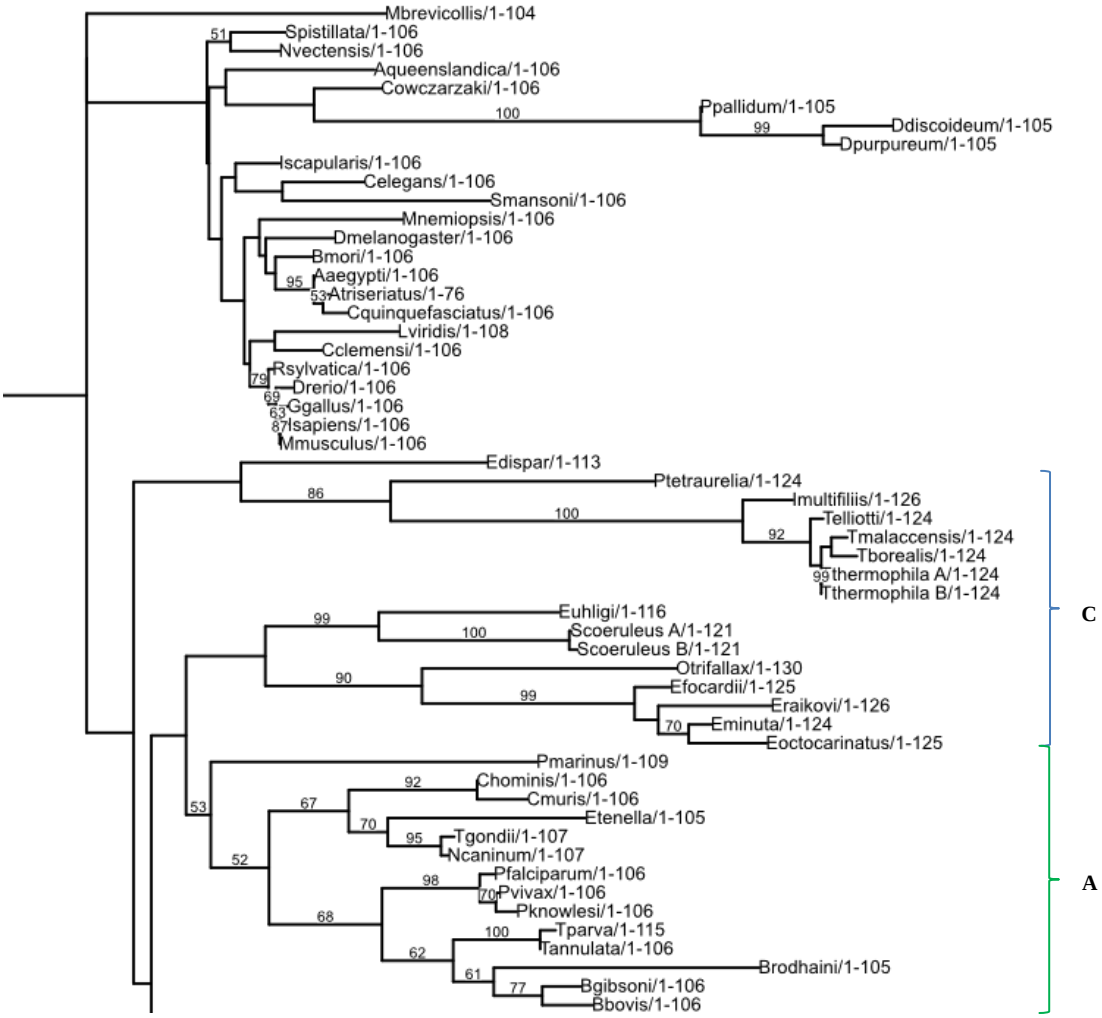


Figure 14b—L10 Region Tree, ML

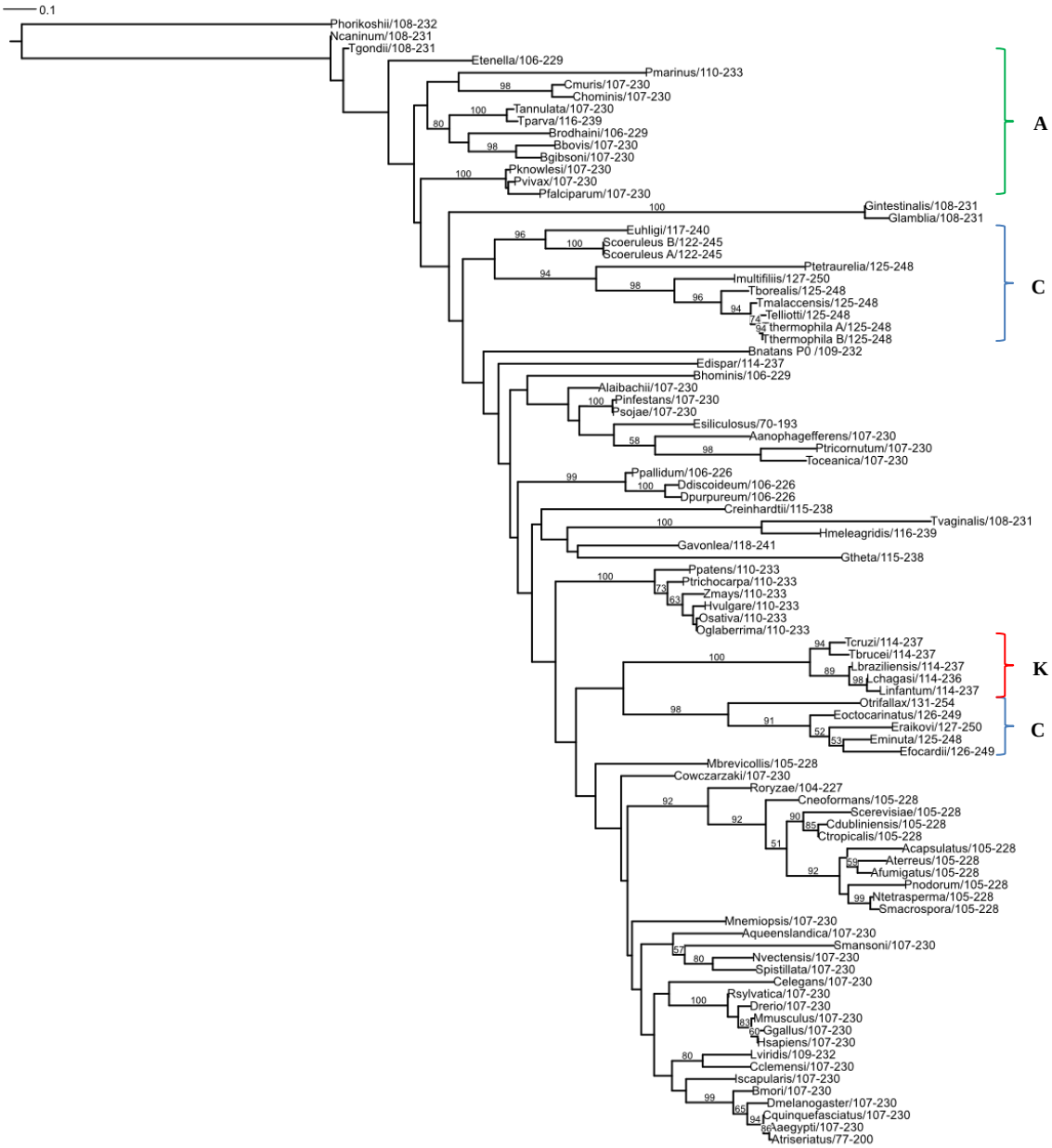


Figure 15a—MID Region Tree, ML

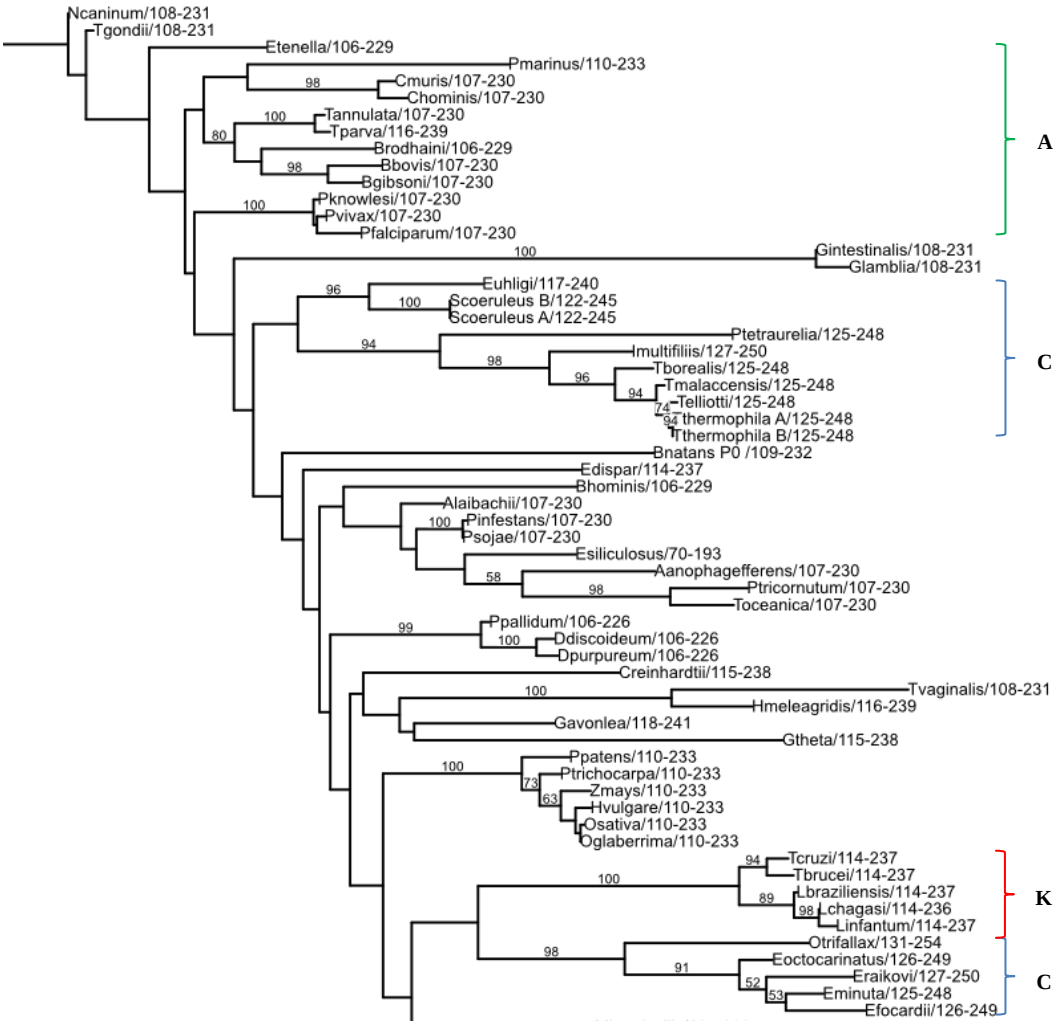


Figure 15b—MID Region Tree, ML

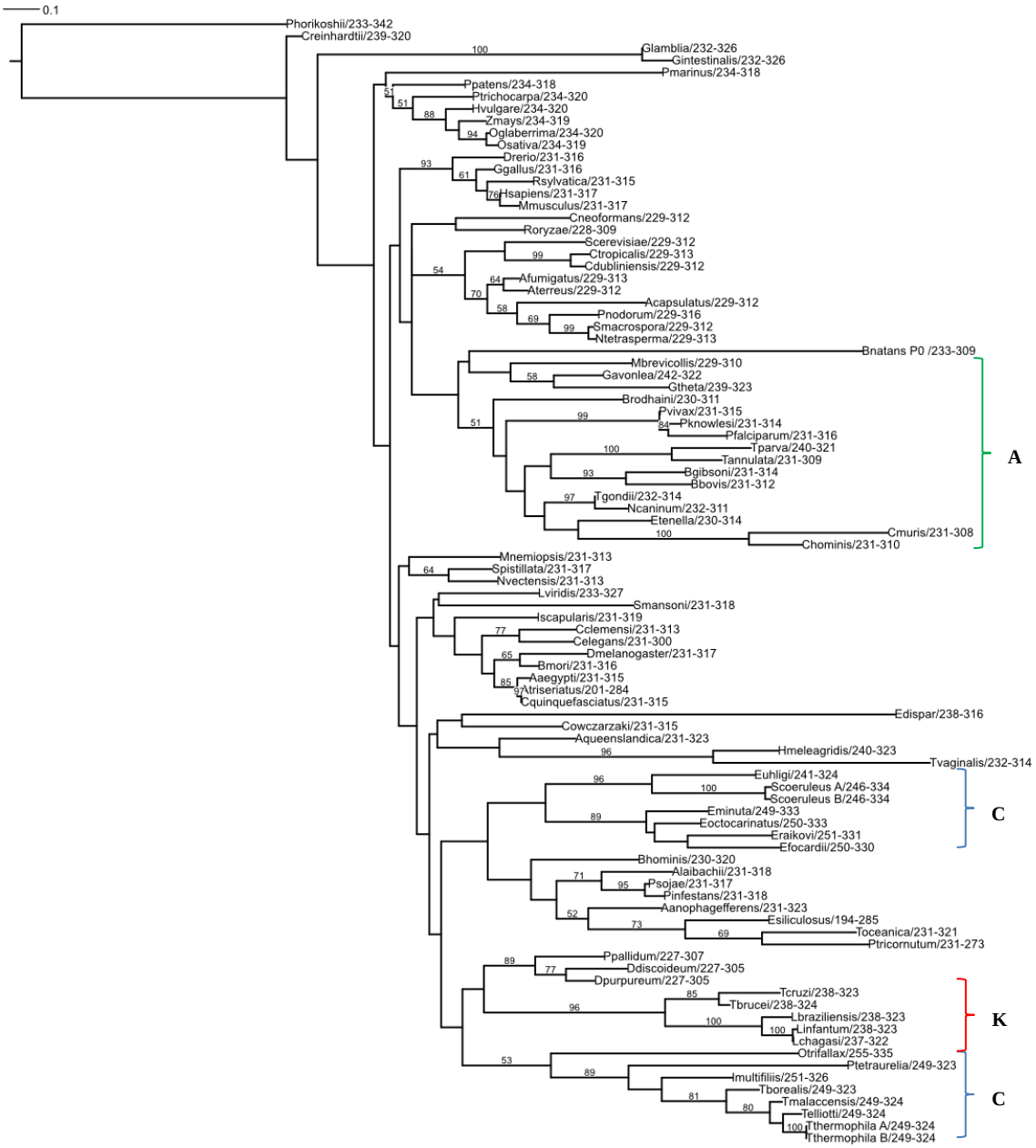


Figure 16a—60S Region Tree, ML

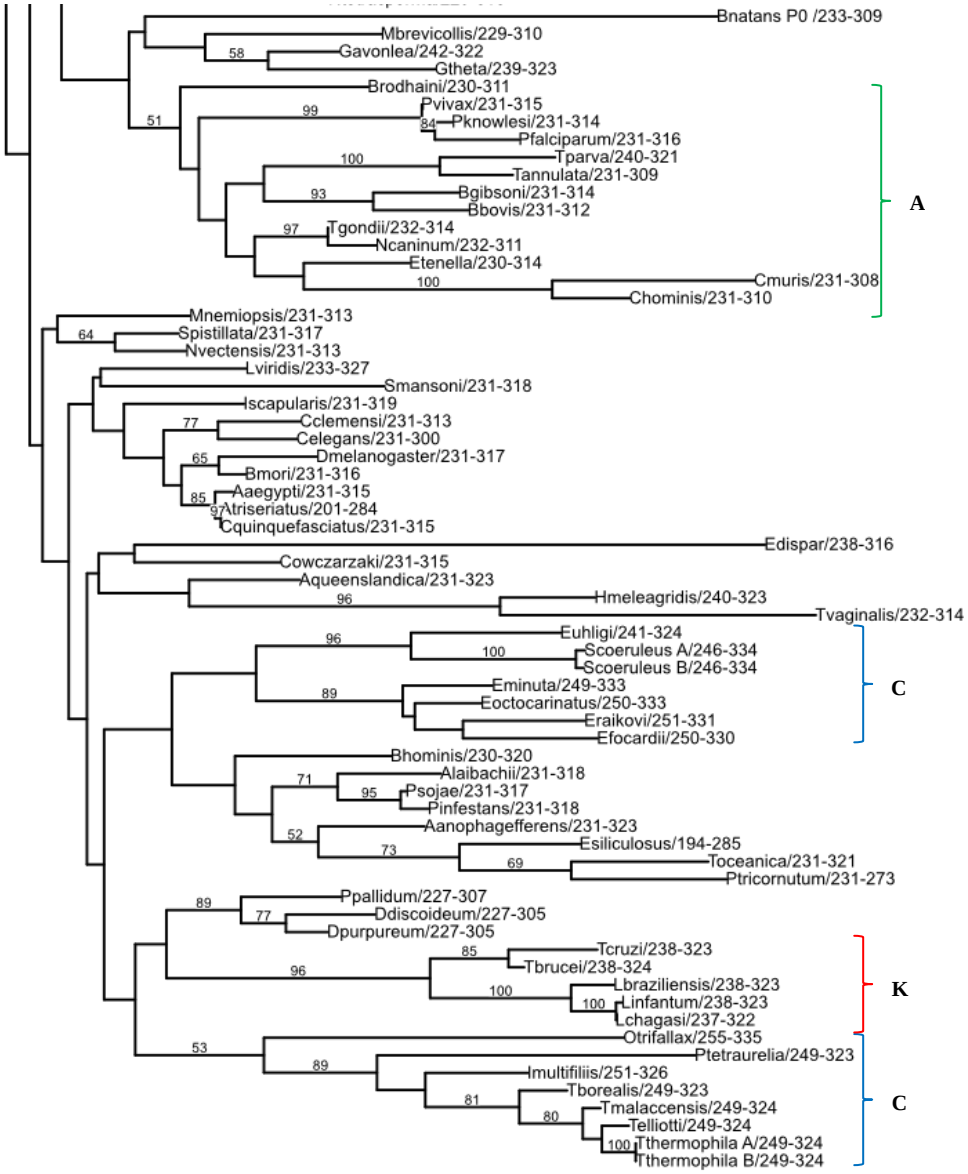


Figure 16b—60S Region Tree, ML

Species	Class	Species	Class	Species	Class
Eukaryotes		*SAR-Rhizaria*		*C. dubliniensis*	Saccharomycetes
SAR-Alveolata		*B. natans*	Chlorarachniophyte	*C. tropicalis*	Saccharomycetes
E. uhligi	Heterotrichea	*Excavata*		*C. elegans*	Secernentea
S.coeruleus	Heterotrichea	*G. intestinalis*	Diplomonads	*S. pistillata*	Stylophora
I. multifiliis	Oligohymenophorea	*G. lamblia*	Diplomonads	*M. leidyi*	Tentaculata
P. tetraurelia	Oligohymenophorea	*L. braziliensis*	Kinetoplastida	*S. mansoni*	Trematoda
T. borealis	Oligohymenophorea	*L. chagasi*	Kinetoplastida	*C. neoformans*	Tremellomycetes
T. elliotti	Oligohymenophorea	*T. brucei*	Kinetoplastida	*R. oryzae*	Zygomycetes
T. malaccensis	Oligohymenophorea	*T. cruzi*	Kinetoplastida	*Archaeplastida*	
T. thermophila	Oligohymenophorea	*H. meleagridis*	Parabasalia	*P. patens*	Bryopsida
E. focardii	Spirotrichea	*T. vaginalis*	Parabasalia	*P. trichocarpa*	Eudicots
E. minuta	Spirotrichea	*Opisthokonta*		*H. vulgare*	Monocots
E. octocarinatus	Spirotrichea	*D. rerio*	Actinopterygii	*O. glaberrima*	Monocots
E. raikovi	Spirotrichea	*R. sylvatica*	Amphibia	*O. sativa*	Monocots
O. trifallax	Spirotrichea	*L. viridis*	Anopla	*Z. mays*	Monocots
B. bovis	Aconoidasida	*N. vectensis*	Anthozoa	*Amoebozoa*	
B. gibsoni	Aconoidasida	*N. tetrasperma*	Ascomycetes	*E. dispar*	Archamoebae
B. rodhaini	Aconoidasida	*S. macrospora*	Ascomycetes	*D. discoideum*	Dictyostelia
P. falciparum	Aconoidasida	*G. gallus*	Aves	*P. pallidum*	Dictyostelia
P. knowlesi	Aconoidasida	*C. reinhardtii*	Chlorophyceae	*D. purpureum*	Dictyostelia
P. vivax	Aconoidasida	*M. brevicollis*	Choanoflagellata	*Other eukaryotes*	
T. annulata	Aconoidasida	*A. queenslandica*	Demospongiae	*G. avonlea*	Cryptophyceae
T. parva	Aconoidasida	*P. nodorum*	Dothideomycetes	*G. theta*	Cryptophyceae
T. gondii	Conoidasida	*A. capsulatus*	Eurotiomycetes	*Organelles [NM-nucleomorph]*	
N. caninum	Conoidasida	*A. fumigatus*	Eurotiomycetes	*B. natans (NM)*	Chlorarachniophyte
C. muris	Conoidasida	*A. terreus*	Eurotiomycetes	*G. stellata (NM)*	Chlorarachniophyte
C. hominis	Conoidasida	*C.owczarzaki*	Filasterea	*C. mesostigmatica (NM)*	Cryptophyceae
E. tenella	Conoidasida	*D. melanogaster*	Insecta	*P.chromatophora*	Imbricatea
SAR-Stramenopiles		*B. mori*	Insecta	**Archaebacteria**	
P. tricornutum	Bacillariophyceae	*A. aegypti*	Insecta	*H. marismortui*	Halobacteria
B. hominis	Blastocystae	*C. quinquefasciatus*	Insecta	*M. jannaschii*	Methanococci
T. oceanica	Coscinodiscophyceae	*A. triseriatus*	Insecta	*P. horikoshii*	Thermococci
A. laibachii	Oomycota	*I. scapularis*	Insecta	*Thermococcus*	Thermococci
P. infestans	Oomycota	*H. sapiens*	Mammalia	**Eubacteria**	
P. sojae	Oomycetes	*M. musculus*	Mammalia	*S. aureus*	Bacilli
P. marinus	Perkinsea	*C. clemensi*	Maxillopoda	*D. radiodurans*	Deinococci
A. anophagefferens	Pelagophyceae	*R. oryzae*	Mucormycotina	*E. coli*	Gammaproteobacteria
E. siliculosis	Phaeophyceae	*S. cerevisiae*	Saccharomycetes	*S. typhimurium*	Gammaproteobacteria
				T. maritima	Thermotogae

Table 3