Abstract

INTRON POSITIONS IN RNA POLYMERASE GENES AND THEIR RELATIONSHIP TO EUKARYOTIC PHYLOGENIES

By MATTHEW C ROBINSON

NOVEMBER, 2010

Director: JOHN STILLER

Co-Director: JINLING HUANG

DEPARTMENT OF BIOLOGY

Over the past two decades, there has been an increasing amount of research devoted to the study of intron evolution and its relationship to eukaryotic phylogeny. Previous studies have shown that a large percentage of intron positions are conserved evolutionarily among three major multicellular eukaryotic groups: animals, plants, and fungi. These studies also have inferred lineage-specific and sometimes massive intron losses, or parallel insertions, based largely on their distributions on molecular sequence-based trees. Interestingly, these studies infer varying numbers of ancestral introns, depending on the algorithms used and phylogenetic associations assumed. The research presented here examines intron evolution in RNA polymerase genes as data for inferring phylogenetic relationships among various eukaryotic lineages. A

phylogenetic tree is inferred based solely on intron position data and these relationships are used to evaluate statistically significant deviations from sequence-based phylogenies. Intron positions were mapped carefully to the various eukaryotic largest and second-largest subunits of RNA polymerases I, II, and III. These sequences were aligned using three different alignment programs (Probcons, T-coffee, and Muscle) and compared using the Altavist web server. Once the proper alignment was established it was analyzed using ProtTest, which tested the alignments against various substitution matrices for the most accurate alignment for use in the phylogenetic analysis. Sequence-based trees were constructed using PHYML as well as RAxML to reduce bias in phylogenetic reconstruction. The intron-based tree was constructed using PAUP v4.0 using intron-positions as binary characteristics. Previous work in our lab has shown that an intron-based tree for RNA polymerase II largest subunit is topographically different from the sequence-based tree, but statistical comparisons were not performed. Such statistical comparisons are rarely made, but are needed to more clearly understand where intron- and sequence-based trees are in clear conflict. This research showed that neither the sequence- or intron-based tress could better explain the data, statistically confirming that both methods produce two different tree topologies. If intron evolution across eukaryotic diversity is to be fully understood, this type of comparison is required to determine where inferences of massive intron gain and loss are in significant conflict with sequence-based phylogenies.

INTRON POSITION IN RNA POLYMERASE GENES AND THEIR RELATIONSHIP TO EUKARYOTIC PHYLOGENIES

A Thesis

Presented To

The Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Masters of Science in Biology

by

Matthew C Robinson

November, 2010

©Copyright 2010
Matthew C Robinson

Intron positions in RNA polymerase genes and their relationship to eukaryotic phylogenies

By

Matthew C. Robinson

Approved by:

Co-director of thesis_____

John Stiller, Ph.D.

Co-director of thesis_____

Jinling Huang, Ph.D.

Committee Member _____

Tim Christensen, Ph.D.

Committee Member _____

Rob Hochberg, Ph.D.

Chair of the Department of Biology

_____

Jeff McKinnon, Ph.D.

Dean of the Graduate School

_____

Paul Gemperline, Ph.D.

ACKNOWLEDGEMENTS

I want to thank Dr. Stiller and Dr. Huang for their support and all they have done while I was working on my thesis. I also want thank the members of my committee, Dr. Tim Christensen and Dr. Robert Hochberg for their guidance during my graduate work.

I want to thank Dr. Terry West, for his help before and while I was a graduate student; as well as Dr. Jason Bond and his lab for letting me run my Bayesian analysis on their computer system. I would finally like to thank my parents for always being there for me and supporting me throughout my life and in my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Many eukaryotic genes are composed of two kinds of sequences, the coding region (exons) and the non-coding region (introns). The focus of most evolutionary research has been on coding regions of these genes; however, over the past two decades research into the importance of non-coding regions has increased dramatically. Introns are divided into three major categories: group I, group II, and spliceosomal introns (Rogers JH 1990). Group I introns, found in rRNA, tRNA, and some protein encoding genes, are self-splicing introns that remove themselves through two transesterification reactions; the first reaction occurs when a free guanosine attacks the 5' splice site of the intron/exon boundary. This allows the exon with a free 3' hydroxyl group to cleave the intron's 3' splice site to completely remove the intron (Cech TR 1990; Saldanha R, Mohr G et al. 1993). Group II introns are found in mitochondria of plants and fungi as well as the chloroplasts of plants. Like group I introns they are also self-splicing, however, their mechanism of splicing differs slightly from the group I variety. Group II introns use two transesterification reactions, the first freeing the 5' end of the intron from the preceding exon. The second reaction involves the creation of a lariat and tail structure created from the free 5' end of the intron binding to the 2' hydryoxyl on an adenine six to seven bases upstream from the 3' end of the excised intron (Saldanha R, Mohr G et al. 1993; Bonen L and Vogel J 2001). Spliceosomal introns are the typical introns present in eukaryotic nuclear protein-encoding genes. The prevailing hypothesis for the origin of spliceosomal introns is that they evolved from what were originally group II

introns. However, unlike group II introns, spliceosomal introns are not self-splicing and require complex machinery called the spliceosome to remove them from the immature mRNA sequence. The spliceosome recognizes the sequence "GT" at the 5' end of the intron as the first cutting site. The free 5' site binds back within the intron to create the same kind of lariat structure found in group II introns, while the spliceosome moves to the "AG" recognition site at the 3' end of the intron and splices the two exons at that point (Rogers JH 1990; Lynch M and Richardson AO 2002).

Two key questions have been investigated with respect to the broad scale evolution of spliceosomal introns. The first key question is whether and how frequently introns were present in ancestral eukaryotes; the second is an effort to understand patterns of gain and loss of introns through the diversification of eukaryotic crown groups (plants, animals, and fungi) (Roy SW and Gilbert W 2005). Large-scale, genome-wide studies encompassing hundreds of eukaryotic genes containing thousands of introns have shown that intron evolution cannot be modeled as a simple process. These studies have shown that eukaryotic species containing high densities of introns are interspersed among species that are intron poor within the same regions of the eukaryotic tree, and sometimes within the same closely related lineage. This has resulted in some ambiguities about processes of intron evolution, with the relative importance placed on inferred intron gain or loss depending on the taxa represented in the study (Roy SW and Gilbert W 2006). The results of these two areas of research have set

the framework for two major theories of intron origins; these are, introns-early versus introns-late.

Walter Gilbert first proposed the introns-early theory that ancient, ancestral organisms contained introns and that these introns were required for assembling the first genes by allowing exons coding for various domains to be shuffled together to create different proteins based on the organization of the domains (Gilbert W, de Souza SJ et al. 1997; Fedorov A, Merican AF et al. 2002; Roy SW and Gilbert W 2005; Roy SW and Gilbert W 2005). Because nearly all extant introns would have been present at the earliest stages of gene evolution, the absence of any given intron in an extant organism indicates loss of that ancestral intron. Thus, intron evolution must be dominated by intron loss, with very little gain. He called this theory “The Exon Theory of Genes”. In it he proposed that the first genes were made up of small segments of DNA (roughly 15 to 20 amino acids) and that new genes were created by the loss of introns between these small segments. He theorized that these early introns were lost via retrotransposition, that is, reverse transcriptase copying spliced mRNA into cDNA and the cDNA recombining back into the genome. Gilbert suggested that, on average, at least two to three of these fusion events occurred, resulting in the increase from early exon lengths of 15 to 20 amino acids to a modern day average of 35 to 40 amino acids (Gilbert W, de Souza SJ et al. 1997).

Three major arguments have been used to support the exon theory of genes. The first is the relationship between modules of proteins and the exons that encode those protein modules. The second form of evidence is the large

number of shared intron positions between plant and animals. The last piece of evidence is the shared intron positions within the genes that have diverged at the progenote (Long M, de Souza SJ et al. 1995). Sverdlov *et al*. showed that numerous intron positions are conserved in orthologous genes in many different eukaryotic species, even between very distantly diverged taxa such as plants and animals. As seeming support for the introns-early theory, Sverdlov hypothesized that early organisms must have harbored many introns and that these introns played a pivotal role in the emergence of the nucleus and cellular organization (Sverdlov AV, Csuros M et al. 2007). Long *et al.* reviewed a study looking at glyceraldehyde-3-phosphate dehydrogenase (GAPDH), which showed identical intron positions between nuclear and chloroplast GAPDH. Other examples with shared intron positions were malate dehydrogenase and aspartate aminotransferase both of which contain shared intron positions between the cytosolic and mitochondrial genes (Long M, de Souza SJ et al. 1995). These observations are consistent with the idea that the ancient prokaryotic ancestors of mitochondria and chloroplasts shared intron positions with early eukaryotes. One of the major problems with the introns-early theory, however, is that extant prokaryotic cells are completely devoid of spliceosomal introns, including the nearest modern day bacterial relatives of mitochondria and chloroplasts. Because complete loss of all ancestral positions from all prokaryotic taxa seems implausible, alternative hypotheses of intron evolution were explored, including how to explain the striking number of common intron positions between animals and green plants.

Because prokaryotic organisms were found to contain no spliceosomal introns, the intron early theory was replaced by the intron-late theory. This theory suggested that introns appeared later in evolution, after the emergence and early diversification of eukaryotic cells (Fedorov A, Merican AF et al. 2002; Roy SW and Gilbert W 2005; Roy SW and Gilbert W 2005). Roy and Gilbert suggested a variation of this theory by postulating that the early explosion of metazoans to multi-cellular life required massive gene shuffling to create all the domains required to carry on the diverse functions associated with developmental complexity (Roy SW and Gilbert W 2005). Fedorov speculated about a mechanism for a later emergence of spliceosomal introns, suggesting they could have arisen from mobile selfish elements with no clear contribution to early genome evolution (Fedorov A, Merican AF et al. 2002). Previous studies on the triosephosphate isomerase gene, which had been raised as support for introns-early, were reexamined and found to also be consistent with the introns-late theory (Logsdon JM, Tyshenko MG et al. 1995). It was also suggested that the relevant increase in intron numbers throughout eukaryotic evolution was correlated with increasing genome complexity (Roy SW and Gilbert W 2005). Evidence presented in support of the introns-late theory is that spliceosomal introns are only present in eukaryotes, suggesting that they did not help to shuffle genetic information in ancestral prokaryotes (Roy SW and Gilbert W 2005). Logsdon and colleagues highlighted a study on xanthine dehydrogenase genes of *Drosophila*, looking at three newly developed intron positions that are thought

to be transposed copies of other introns widely seen in other xanthine dehydrogenase genes (Logsdon JM, Stoltzfus A et al. 1998).

To decide between the two opposing theories of intron evolution, researchers have tried to determine the number of shared intron positions among various eukaryotic taxa, and the relative importance of intron gain versus loss over time. The first major study of intron positions was conducted in 1980s (Shah DM, Hightower RC et al. 1983; Gilbert W, Marchionni M et al. 1986; Marchionni M and Gilbert W 1986; Kersanach R, Brinkmann H et al. 1994); it showed that plants and animals, indeed, share many common intron positions, and that these intron positions could have been inherited from their last common ancestor (Fedorov A, Merican AF et al. 2002). A later study conducted by Rogozin *et al.* indicated that only 1% of introns shared between two species should occur by chance, meaning that it is almost statistically impossible for three or more species to share intron positions except by descent from a common ancestor (Rogozin IB, Wolf YI et al. 2003). Scott Roy found that most fungal genomes contain a range of 0.1 to 5.5 introns per gene, plant genomes ranged between 0.1 and 6.7 introns per gene, animal genomes ranged between 2.6 and 9.3 introns per gene, and even the genomes of the protist group Apicomplexa contained a range of 0.1 to 2.3 introns per gene (Roy SW 2006).

Despite this level of variation within and among taxa many different studies have shown a large percentage of intron positions to be present in the same positions in distantly related organisms. One study of intron positions in plants, animals and fungi by Fedorov and colleagues (2002) showed that plants

and animals share 10% of all intron positions with an additional 7% of positions within six base pairs of each other (an acceptable difference to accommodate intron sliding, the slight shifting of an intron's position within the genes open reading frame). They also showed a 15% match of intron positions between animals and fungi as well as a 13% match between plants and fungi. Interestingly, Fedorov and colleagues observed a percentage of intron positions shared among all three taxa that was higher than expected from Poisson distributions. They suggested that nearly all these shared introns are, in fact, ancestral positions predating the divergence of plants, animals, and fungi (Fedorov A, Merican AF et al. 2002). In 2003 Rogozin *et al.* published their analyses of 684 orthologous genes from animals, plants, fungi, and the apicomplexan protist *Plasmodium*. Their results showed that 24% of intron positions in *Arabidopsis* are shared with humans but that humans only share 12-17% of all intron positions with *Drosophila*, *Caenorhabditis*, and *Anopheles*. Strikingly, they also found that *Plasmodium* shared one third of all its intron positions with at least one member of each of the crown eukaryotic groups (plants, animals, fungi). These discoveries led them to suggest that 25 to 30% of all introns were inherited from the last common ancestor of the three crown eukaryotic taxa (Rogozin IB, Wolf YI et al. 2003). Roy and Gilbert looked at shared intron positions using a maximum-likelihood analysis comparing plants, animals, fungi, with *Plasmodium* as an outgroup, and obtained similar results (Roy SW and Gilbert W 2005). They discovered that almost two-thirds of all animal introns predate the bilaterian ancestor and that two-fifths of plant, animal,

and fungal introns predate the last common ancestor between animals and plants. This suggests that early eukaryotes were more intron rich than previously thought and that intron loss has been a major influence on gene evolution. Interestingly, Roy and Gilbert also observed many shared intron positions in *Plasmodium* suggesting an even larger phylogenetic distribution of ancestral introns (Roy SW and Gilbert W 2005).

By knowing which introns are evolutionarily conversed, and assuming a specific phylogenetic history among taxa, it is possible to estimate rates of intron insertion and deletion. Gilbert and Roy calculated intron insertion rates to be $6x10^{-13}$ to $4x10^{-12}$ per possible intron site per year and intron deletion rates to be $2x10^{-9}$ to $2x10^{-10}$ per year (Roy SW and Gilbert W 2005). These rates suggest that most intron position have existed for a long period of time and, therefore, could retain a strong evolutionary signal much longer than the sequences in which they are found. Intron loss also is generally modeled as an irreversible process meaning once an intron is lost it will not be reinserted into the same location (Roy SW and Gilbert W 2005). With predicted intron gain and loss rates, and conserved intron positions known, general assumptions can be made about intron evolution at various taxonomic levels. For example, there appear to have been extensive losses of introns in many species of bilaterians such as *Drosophila, Caenorhabditis elegans, and C. intestinalis*. In contrast, vertebrate and higher plant introns have remained relatively stable. These differences are thought to be due to selective pressures favoring more compact genomes in some groups, or to other evolutionary factors (Roy SW 2006).

One especially interesting characteristic of intron distributions is the large percentage of shared intron positions between animals and plants, despite their presumed long evolutionary divergence. In fact, in many recent phylogenomic treatments, plants and animals are considered to fall on opposite sides of the root of the eukaryotic tree (Stechmann A and Cavalier-Smith T 2002; Stechmann A and Cavalier-Smith T 2003). These large numbers of shared positions could either be due to parallel gains in either taxa, or the fact that there were many introns present in their last common ancestor that have since been lost in all other taxa (Sverdlov AV, Rogozin IB et al. 2005; Roy SW and Gilbert W 2006; Carmel L, Rogozin IB et al. 2007). Interestingly, in a detailed investigation of patterns of intron gain and loss, Carmel *et al.* found "practically, no parallel gains in closely related lineages, whereas for distant lineages such as animals and plants, parallel gains appear to contribute up to 20% of the shared intron positions" (Carmel L, Rogozin IB et al. 2007). This statement implies one of three conclusions: 1) intron evolution has followed very different patterns at long and short evolutionary distances, 2) there has been an over-estimation of parallel gains and that most shared intron positions were present in the common ancestor of most or all eukaryotes, or 3) the phylogenetic trees on which intron gain and loss are interpreted have overstated the evolutionary distance between plants and animals. Sverdlov's use of Monte Carlo simulations, which predict that intron insertions could happen in only a fraction of the genome, suggest that parallel gains have been very rare and only contributed a small percentage of shared positions at great evolutionary distance (Sverdlov AV, Rogozin IB et al.

2005). This provides theoretical support for over-estimation of parallel gains, thus suggesting that most shared positions were present in the common ancestor of plants and animals.

Current research leaves open many possibilities in the area of intron evolution, there is reasonable evidence that even very early eukaryotes contained introns; however, the relative number of introns is subject to debate, as is whether shared positions in extant taxa date to those early insertions. An important consideration regarding all intron investigations to date is that they have based their findings on patterns of evolution derived from sequence-based phylogenies. If these trees depict incorrect historical relationships, this could be leading to overestimated rates of intron gain and/or loss within or between specific lineages.

Similar to patterns observed in the major studies cited above, previous work in our lab showed a clear topological difference between sequence-based and intron-based trees inferred from the RNA polymerase II (RNAP II) largest subunit (RPB1); however no statistical analysis was conducted to measure the significance of this difference (Harrell 2005). The research reported here deviates from most previous studies by looking at what intron data suggest about the phylogeny of eukaryotes, rather than simply mapping them on sequence-based trees. Specifically broad taxon sampling of intron rich species in highly conserved genes (the two largest subunits of three DNA-dependent RNA polymerases) was used to create a data set of binary characters for phylogenetic analyses. The reasons for using RNA polymerase subunits are several fold.

First, major RNAP subunit genes nearly always exist as single copy orthologs. Second, they are highly conserved throughout evolution making the alignment of protein sequences and inferences of intron positions more reliable. Lastly, RNAP subunits are some of the first and best annotated genes from sequencing projects, and Expressed Sequence Tag (EST) data are generally available for each sequence allowing empirical verification of intron/exon boundaries. Using these six genes helps to remove potential biases created by poorly conserved genes with ambiguous alignments, or artifacts related to differential losses in paralogous gene families.

Sequence-based and intron-based trees were generated from these six RNAP genes to determine if there were substantive differences between the patterns of evolution inferred using the different data sets. The topological differences were further analyzed to determine whether they represented statistically significant variation between the two approaches. This kind of analysis, including careful annotation of all intron positions and rigorous investigation of conflicts between inferred patterns of sequence and intron evolution has not been undertaken previously. Nevertheless, these approaches are required to determine whether significant conflicts exist between implied evolutionary histories of introns and of the exon sequences in which they reside, and what aspects of these histories are most compatible with known mechanisms of molecular evolution.

MATERIALS AND METHODS

Determine intron rich species covering a wide range of taxa

Comprehensive and balanced taxa sampling is an important step in looking at the evolution of introns; for this study intron rich species were identified to help reduce problems associated with massive, taxon specific intron loss. Seven animals (*Anopheles gambiae*, *Bos taurus, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Mus musculus,* and *Takifugu rubripes*), five green plants (*Arabidopsis thaliana, Chlamydomonas reinhardtii, Oryza sativa, Ostreococcus lucimarinus, and Populus trichocarpa*), seven fungi (*Aspergillus fumigatus, Cryptococcus neoformans, Phanerochaete chrysosporium, Pichia stipitis, Saccharomyces cerevisiae, Schizosaccharomyces pombe*, and *Ustilago maydis)* and seven protists (*Cryptosporidium parvum, Cyanidioschyzon merolae, Dictyostelium discoideum, Phaeodactylum tricornutum, Plasmodium falciparum, Thalassiosira pseudonana,* and *Trypanosoma brucei*) (table 1). These species represent a broad collective sample of plant, animal, and fungal species with high intron densities relative to similar species in these major taxa, along with a wide sampling of protists to provide adequate outgroups.

Obtaining sequence data

Sequences of the largest and second largest subunits of DNA dependent RNA polymerases I, II and III were obtained from the National Center for Biotechnology Information (NCBI) (maintained by the respective genome project organizations) or from the Joint Genome Institute (JGI) (Table 1). Total genomic

sequences and coding region only (based on ESTs where available) sequences were downloaded for each subunit and species for use in determining intron positions. Protein sequence data were obtained for use in phylogenetic analysis as well as intron position determination.

Multiple sequence alignments

Multiple sequence alignments were performed for the each RNAP subunit to permit comparison of intron positions among sequences and create a data set for phylogenetic analyses. Because no single alignment program can be trusted to predict the correct biological arrangement, three sequence alignment programs were used to help distinguish homologies in regions of low similarity, and to determine whether those regions can be aligned reliably enough to infer shared intron positions accurately. Based on a review of various alignment programs by Edgar and Batzoglou (Edgar RC and Batzoglou S 2006), three alignment programs (Probcons, T-coffee, and Muscle) were chosen to align each of the six subunits.

The first program, Probcons, was developed at Stanford University by Chuong Do in collaboration with Michael Brudno and the Batzoglou research group. Probcons uses a combination of probabilistic modeling and consistency-based alignment techniques to obtain the highest level of accuracy on most standard alignment benchmark samples (Do CB, Mahabhashyam MS et al. 2005).

The T-coffee alignment software was developed by Cedric Notredame at the Comparative Bioinformatics group at The Center for Genomic Regulation in Barcelona. T-coffee uses a progressive alignment to create an initial library. This library then is extended multiple times until the most optimal alignment is achieved (Notredame C, Higgins DG et al. 2000).

The third alignment program utilized was Muscle, developed by Robert Edgar. Muscle uses a three-part algorithm; the first is a draft progressive algorithm that builds its initial alignment. In the next stage the progressive algorithm is iterated to increase alignment accuracy. The final stage is refinement. In this stage the program uses iterations to fine tune the alignment created by stage two (Edgar RC 2004).

Alignments were prepared with each program using an iterative procedure. A first alignment was produced that included all the taxa except for *Trypanosoma brucei* and *Plasmodium falciparum*. Both of these species have large sequence insertions that previous analyses revealed result in poor alignments when all species are aligned at once. This initial alignment then was re-aligned with the *Trypanosoma brucei* sequence, with the restriction that the realignment did not realigned previously determined conserved blocks. This second alignment was re-aligned with the *Plasmodium falciparum* sequence with the same restriction, thereby creating the completed alignment. These final alignments were used in all subsequent phylogenetic and intron position analyses.

The alignment comparison web server AltAVist also was used to help improve the accuracy of alignments. AltAVist is a web-based program that compares results from different sequence alignments and to determine conserved regions recovered in common. It also allows regions to be better aligned by comparing other alignment programs to improve the consensus sequence (Morgenstern B, Goel S et al. 2003). AltAVist was used for trimming the aligned protein sequences for further use in the sequence-based phylogenetic analysis, by highlighting areas that were poorly aligned and should be removed prior to running computational programs.

Determination of intron positions

Two different methods were employed to determine intron positions within the aligned protein sequences. Determination of exact intron positions was imperative to allow an accurate assessment of whether a given intron position was shared among different taxa.

The first method for intron position determination involved searching the genomic database consortia for EST data for each gene used in the study. These consortia obtain EST data by directly sequencing cDNA clones created from mRNA templates, which allows a comparison between the expressed regions of the genes against those of the genomic sequences that contain introns. This provides direct evidence for splice junctions (intron/exon boundaries) within the genomic sequence. The information from EST data

provides experimental validation of the locations of introns within the gene sequence.

The second method was to use a custom Biopython program developed to parse Genbank files for exon boundaries previously established by the original depositor of the sequence file. This program uses the libraries created by Biopython to parse the Genbank file to look for the sequence locations of all exon boundaries. Upon finding the exon boundaries, the program selects the corresponding genomic sequence located between the exon boundaries markers; all the exon sequences for each gene were labeled and numbered. This list of exon sequences then was translated into the coding sequence by using the translate tool located on the ExPASy web server and intron positions were located on already aligned protein sequences.

A two-step process was followed to map introns onto aligned protein sequences using both methods of intron position determination. The first step was to determine all the intron positions for each subunit using the Biopython script. This output gave an initial reading for where each intron was located within the respected protein sequence. The second step in this procedure involved using EST data (when available) to double check the Biopython script determined positions to ensure that the correct position had been annotated and was being used for this study. In the event of any difference between an intron location determined by the Biopython script and EST data, the EST-based position was used because there was experimental evidence from a cDNA library for that sequence, as opposed to the evidence provided by computational *ab*

initio analyses of the sequence for theoretical splice junctions as intron boundaries.

Introns were mapped directly onto all six of the protein subunit alignments, with each intron phase indicated by a different color. The amino acid was colored blue if the intron occurred between two codons for different amino acids (phase 0). Green indicated an intron located between the first and second nucleotide of a codon (phase 1), and red was used for introns located between the second and third nucleotide of a codon (phase 2) (figure 1).

Intron position matrix

Shared and unique intron positions in conserved regions of each subunit were used to create a binary data matrix (1 = presence / 0 = absence of intron at that position). To account for the possibility of intron sliding, two different intron matrices were created to test for the effects on tree topologies of assuming some movement of introns. The first matrix applied a strict rule for assumption of intron homology; that is, only introns in the same location and phase were counted as homologous. The second matrix relaxed this constraint on intron homology. Any intron within six nucleotides (two amino acids) was considered to be homologous. The selection of a six nucleotides permissible window for intron sliding was based on the computational estimates from available literature (Stoltzfus A, Logsdon JM Jr. et al. 1997; Rogozin IB, Lyons-Weiler J et al. 2000; Fedorov A, Merican AF et al. 2002). In addition to accounting for intron sliding, this second matrix also provides some mitigation from ambiguous splice

junctions where solid EST evidence does not exist, which require judgment calls of some intron positions.

In addition to those based on individual intron positions from all species, addition intron matrices were created. For these matrices all intron positions found within defined major groups (animals, plants, fungi, apicomplexans, kinetoplastids, red algae, amoebozoans, and stramenopiles) were condensed, so that if any species from one of these major taxon has an intron at a given position, that intron is coded as present in the group for comparative analyses with other eukaryotic taxa. The reason for condensing intron positions from each group is to remove the substantial bias introduced by independent loss of introns among taxa within each major lineage. This collapsing of introns for each major taxon is based on the assumption that independent gain of introns in a given location is exceedingly rare. By using this matrix, intron positions can be analyzed among major eukaryotic taxa without artificially attracting intron-rich individual species from different lineages to each other in phylogenetic reconstruction. Both the group matrix and the species matrix were used for phylogenetic analyses of relationships among crown groups.

Phylogenetic analysis

Phylogenetic analyses were performed using both the intron position matrix data and sequence data. To select the most appropriate substitution model for the sequence-based phylogenetic analyses, ProtTest was used on each of the aligned subunit sequences. The ProtTest program tests various

phylogenetic substitution models on aligned sequences to determine the most likely one for a given data set (Drummond A and Strimmer K 2001; Guindon S and Gascuel O 2003; Abascal F, Zardoya R et al. 2005).

To help reduce possible biases in the phylogenetic construction two programs were used to construct sequence-based tress, Phylogenetic Inferences using Maximum-Likelihood (PHYML) (Guindon S and Gascuel O 2003; Guindon S, Lethiec F et al. 2005) and MrBayes for Bayesian inference (Huelsenbeck JP, Ronquist F et al. 2001; Ronquist F and Huelsenbeck JP 2003). Settings for PHYML were based on the results from the ProtTest analyses on each subunit; the same substitution model, RetRv, was recovered for all individual subunits as well as the two concatenated sequences. The alpha parameter and proportion of invariable sites were estimated from the data, with the number of substitution rate categories set to four. The same settings also were used for Bayesian inference for one million generations with trees sampled every hundred generations. After one million generations the burn-in was set for one thousand based on empirical observation of likelihood convergence and the majority-rule tree was created.

Intron position trees were created using Dollo parsimony in Phylip v3.6. Dollo parsimony was used, rather than standard maximum (Wagner) parsimony because it permits an intron to be gained only once, and does not allow repeated gains of characters. For both of the crown group intron matrices (sliding, without sliding) kinetoplastids were chosen arbitrarily as the root. For the species-level intron matrices the kinetoplastid *Trypanosoma* was used as the root. As a control for whether Dollo parsimony could be too restrictive for intron gain/loss

reconstruction, trees also were constructed in PAUP v4.0 (Swofford DL 1991) using Wagner parsimony. All trees were made using 1000 bootstrap replicates and included groups compatible with a 50% majority-rule consensus. To prevent established lineages from being broken up by attraction between species that have undergone extensive intron loss, well established major taxa (see above) were constrained to be monophyletic; this allowed each clearly defined lineage to be distinct giving more accurate evolutionary relationships.

Statistical comparison of phylogenetic relationships

To determine whether differences between phylogenetic trees were statistically significant the Kishino-Hasegawa-Templeton (KHT) test and Shimodaira-Hasegawa (SH) test were used to compare trees based on intron data, sequence data, and on the comparison of the sequence data altered to reflect the intron phylogeny. KHT and SH tests were performed in Dollop and proML, respectively, contained in the software package Phylip v3.69 (Felsenstein J 2004). Dollop was used for all the intron specific tree testing to compare differences between the topologies of the various trees created; allowing for intron sliding and inclusion or exclusion of *Chlamydomonas* intron data. The reasoning for exclusion of *Chlamydomonas* was due to poor sequence data and unresolved intron positions for RPA2. ProML was used for sequence-specific tree tests using the Jones-Taylor-Thornton (JTT) model of amino acid change; other parameters differed depending on which data set was used. For the sequence data including *Chlamydomonas* the alpha parameter was 1.45, 4 HMM categories were used with 0.17 as the fraction of invariant sites. For the

sequence data without *Chlamydomonas* the alpha parameter was 1.457, 4 HMM categories were used with 0.167 as the fraction of invariant sites. These parameters were established from the initial phyML runs using these data. All trees were compared with the arbitrary outgroup root of kinetoplastid (intron data) and *Trypanosoma* (sequence data).

The final comparison analyzing differences between the sequence-based and the intron-based phylogenies was done in proML using the parameters established for the larger "with *Chlamydomonas"* data set. To allow comparisons between the sequence-based tree topology and the two different intron-based tree topologies, the sequence-based topology was modified to reflect the intron sliding and no intron sliding tree topologies. To do this the program retree in Phylip v3.69 was used to move plants to become the sister group of the animals, as recovered in intron-based phylogenetic analyses (see results). This newly created tree allowed testing of whether the intron-based topology was significantly worse, in ML analysis of sequence data, than the sequence-based tree recovered from phyML and MrBayes analyses.

RESULTS AND DISCUSSION

Analysis of sequences

To fully evaluate how well the pattern of intron distribution reflects sequence-based phylogenies, accurate sequence data must be obtained and rigorously assessed to ensure the most complete sequences are used, and that

intron positions are inferred as correctly as possible. As described in the materials and methods, the genomic, coding, and protein sequences for each species (table 1) were obtained from NCBI and other genome servers. Inferred protein sequences of each of the largest and second largest subunits of RNA polymerases I, II, and III were aligned using Muscle as a preliminary alignment to assess whether any sequences, as annotated, were missing known essential or highly conserved domains. Initial alignments of the subunits identified sequences that either aligned improperly or were truncated on either or both ends of inferred genes, requiring further manual annotation (table 2). For example, the largest subunit gene of RNA polymerase I (RPA1) from *Populus trichocarpa* was truncated on its 3' end; however, careful examination of neighboring genomic sequence showed that a distinct gene encoding 464 amino acids, with high sequence similarity to the C-terminal region of RPA1, had been annotated just downstream of the annotated RPA1 gene. Another taxon requiring extensive manual annotation of *RPA*1 was the green alga *Chlamydomonas*, which also was truncated at the C-terminus. Review of the region downstream from the annotated RPA1 sequence revealed a region encoding 248 amino acids with strong sequence similarity to the terminal region of RPA1. The third taxon with a mis-annotated *RPA*1 sequence was *Cryptosporidium*; once again there was a problem at the C-terminal end of the inferred protein sequence. Careful analysis of all forward reading frames uncovered a frame shift as a result of an intron that was left in the original annotated protein sequence, resulting in 222 incorrectly

inferred amino acids. This yielded a contiguous RPA1 sequence with a C-terminus with higher similarity to other RPA1 sequences.

For the initial alignment of the second largest RNAP I subunit (RPA2) there were problems with three of the twenty-six species in their deposited, annotated protein sequences. The sequence from *Populus* obtained from NCBI did not align with other RPA2 subunits at the 5' end, instead starting several hundred amino acids downstream. Careful analysis of the genomic region upstream from the annotated start codon did not resolve a better region for amino acid alignment, based either on protein (Blastp) blast or translated protein blast (tBlastn) using other green plant RPA2 sequences as queries. Therefore the *Populus RPA*2 was left unchanged (with N-terminal truncation) in the final alignment. The diatom *Thalassiosira* also was missing some 5' sequence. In this case, however, using RPA2 N-terminal sequences to query upstream genomic sequences revealed an un-annotated intron and a region of high similarity to the 5' regions of other RPA2 genes. These un-annotated introns, when not discovered, present a serious problem for understanding and analyzing intron evolution. Not taking the time to ensure the dataset is complete before testing could lead to incorrect conclusions because of missing data. The *Chlamydomonas RPA*2 sequence was missing a large number of regions that otherwise were conserved in the global alignment. Analysis of the genomic sequence in these regions revealed a large proportion of incomplete genomic data that made it impossible to determine a more accurate protein sequence translation within that region. For this reason, that is, extensive regions of

incomplete data, *Chlamydomonas* was not included in the final alignment procedures for RPA2.

Only one species contained a mis-annotated protein sequence for its largest subunit of RNAP II (RPB1). The diatom, *Phaeodactylum,* had an error in the 5' region resulting from the incorrect choice of a later start codon, which resulted in the loss of conserved proximal domains. This was corrected when conserved RPB1 sequence following a more reasonable methionine start site were found upstream from the original annotated start codon.

The second largest subunit of RNAP II (RPB2) data set contained only two species with problem regions. The first was the 5' region from *Takifugu*; like several examples from other subunits, the 5' region of the protein was missing. Upon manual analysis of the genomic sequence upstream from the annotated theoretical start codon, a six-exon region was discovered that showed strong similarity to other RPB2 sequences in blast searches. The tree species *Populus* also contained a mis-annotated region of *RPB*2. The C-terminal end was missing and subsequently was found downstream in the genomic sequence.

For the largest subunit of RNAP III (RPC1), only two of the twenty-six species studied contained regions of mis-annotation. The first was again from *Takifugu*, which had problems in both the 5' and 3' regions; mis-annotations of the reading frame resulting in a missing exon in each region. According to the initial alignment the apicomplexan *Cryptosporidium* had areas with deleted or missing sequence; however, examination of the genomic sequence revealed no

obviously mis-annotated regions. Therefore, the original sequence was retained for the final alignment.

The last data set, comprising second largest subunits of RNAP III (RPC2) also contained two species with mis-annotated sequences. The mosquito *Anopheles* had a region with a deletion of protein sequence that was annotated as an intron in the NCBI accession, but careful analysis showed it to be part of the coding region based on strong similarity to the missing protein sequence in blast analysis. The *Thalassiosira* inferred RPC2 sequence in the initial alignment started further down the 5' region of the other RPC2 sequences when compared to the rest of the sequences. Examination of the 5' region genomic sequence revealed an exon, which was not included in the protein sequence but was highly similar to 5' regions of other RPC2 genes.

These analyses of each sequence individually, ensuring proper annotation, was a very important step that is not generally taken in automated, large-scale genomic investigations of intron gain and loss. It showed that sequences routinely downloaded from annotated databases are not always correct and careful manual annotation is required to ensure that results obtained using the data are accurate.

Once each subunit was analyzed carefully sequences were re-aligned in the iterative process, using the three alignment programs described in the materials and methods section (Muscle, Probcons, and T-Coffee). As a final check for accurate identification of all sequences, and to ensure that there were

no paralogous or duplicated sequences included in each subunit data set, initial phylogenetic analyses were preformed to verify that each sequence had been classified as the correct subunit. All alignments were trimmed down to only the most highly conserved blocks using the program Mesquite (WP Maddison and DR Maddison 2007). Once trimmed, all three of the largest subunits were combined, re-aligned in Muscle and imported into PhyML for phylogenetic analysis (figure 2). This process was repeated to align all of the second largest subunits globally (figure 3). Both global gene family trees (largest and second largest) showed each subunit family to be monophyletic; that is, no RNAP sequences grouped with subunits from a different polymerase, demonstrating that none had been misidentified and placed into the wrong paralogous gene family.

Sequence-based analysis

Individual subunits

Phylogenies for each of the six subunits were determined individually to recover topologies that could be compared to global phylogenetic trees constructed from the combined signal of all six subunits. Maximum-likelihood trees were created using phyML and Bayesian inference was performed with MrBayes. Regardless of the analytical method used, RPA1 sequences produced tree topologies with almost all established taxonomic groups recovered as monophyletic clades; the only exception was that apicomplexans nested within the green plant group in the Bayesian inference tree (figure 4), whereas

likelihood analysis recovered green plants as monophyletic (figure 5). Likelihood analysis of RPA2 (figure 6) recovered comparable monophyletic clades, but with animals branching off prior to the split of plants and fungi. However, with Bayesian inference (figure 7) the apicomplexans and red alga *Cyanidioschyzon* both nested within the animal group and *Ostreococcus* branched outside of green plants. The overall tree topology for the Bayesian tree showed a different topology from the likelihood analysis in that plants branched off prior to the animal/fungi divergence. The likelihood analysis of RPB1 (figure 8) shows strong monophyletic groupings of major taxa, with plants and animals forming sister clades after the divergence of fungi. Bayesian analysis of RPB1 (figure 9) yielded comparable results, however, contrary to likelihood analysis grouped fungi and animals as sister clades after the divergence of green plants.

Bayesian analysis of RPB2 (figure 10) separated *Chlamydomonas* from the rest of the green plant clade and showed stramenopiles (*Thalassiosira* and *Phaeodactylum*) branching from the green plant lineage. The fungus *Ustilago* was also grouped outside of the rest of the fungal species branching off very early in the tree. Animals and fungi formed monophyletic groups, with animals as a sister clade to plants after the divergence of fungi. In likelihood analysis of RPB2 (figure 11) the overall grouping of the species was very similar to Bayesian inference (figure 10); however, in the likelihood tree plants and animals formed sister clades to each other with fungi branching off prior to the animal/plant divergence. When analyzed by maximum likelihood (figure 12) RPC1 sequences produced a tree with all major taxa as monophyletic group. It also featured

animals and fungi as a sister clades branching after the divergence of green plants. The Bayesian tree (figure 13) produced a tree with animals, plants, and fungi all monophyletic; however, it could not resolve the branching point for *Dictyostelium* and the *Phaeodactylum/Thalassiosira* clade. The likelihood analysis of RPC2 (figure 14) was similar to what was observed in RPC1 with all major groups monophyletic and green plants diverging before the animal/fungi split. Bayesian analysis of RPC2 (figure 15) agreed with the likelihood analysis grouping animals and fungi as sister clades with plants branching prior to the divergence of animals and fungi.

Concatenated subunits

To recover an overall tree topology from all of the RNA polymerase subunits, alignments were concatenated together to create one large data set. This complete data set was used because tree topologies created from one gene often are subject to various biases; that is, the evolution, or at least phylogeny of that gene may not reflect the evolution of the species as a whole, as shown by variations in topologies obtained from each individual subunit (figures 16-19). Because the *Chlamydomonas* RPA2 sequence was missing large regions, two different data sets were created. The first set included *Chlamydomonas* sequences (figures 16 and 17) and the second data set did not (figures 18 and 19). Both of these data sets were analyzed in PhyML and MrBayes to determine whether they produced comparable phylogenies. With the data set that included *Chlamydomonas,* the tree topology agreed with currently “accepted” assumptions that animals and fungi are sister groups, with plants branching further away. This

was recovered in both phylogenetic analyses (ML and Bayesian) and had strong statistical support in both cases (figures 16 and 17 respectively). The tree topology using second data set (without *Chlamydomonas*) was the same as the topology with *Chlamydomonas*, and all major taxonomic groupings were monophyletic (figures 18 and 19). Animals and fungi again grouped together as sister taxa with plants branching before the divergence of this "opisthokont" clade. Since the two data sets produced the same topology regardless of the inclusion of *Chlamydomonas*, the data set of all 6 subunits excluding *Chlamydomonas* was used later for statistical comparison to the intron-based tree. Exclusion of *Chlamydomonas* sequences was based on the notion that including incomplete taxa in phylogenetic analysis is often associated with difficulties in the assembly of the phylogeny resulting in problems in tree resolution (Wiens JJ 2003; Philippe H, Snell EA et al. 2004; Wiens JJ 2006)

Intron-based analysis

To analyze intron gains and losses, intron positions from each gene were mapped directly on the aligned protein sequences. Intron numbers varied greatly among species in broader comparisons, with vertebrate animals and green plants containing the highest densities, whereas most protist genes were relative deprived of introns (table 3). From this mapping two distinct positional matrices were created; one with each mapped position as a distinct binary data point (no intron sliding) and one with intron positions within six nucleotides (two amino acids) counted as the same position (allowing for intron sliding). The number of intron positions scored for each subunit varied from one another, both within a

given matrix method and also between different methods (table 4). The RPB1 data set experienced the least change in intron numbers between the sliding and no-sliding approaches, with differences in only six total positions out of 76 scored. All other subunits had much higher degrees of variation between intron numbers inferred using the two methods.

To reduce the effects of intron gain and loss on phylogenies intron position matrices were collapsed within respective major taxonomic groupings: animals (*Anopheles, Bos Caenorhabditis, Danio, Drosophila, Mus,* and *Takifugu*), green plants (*Arabidopsis, Chlamydomonas, Oryza, Ostreococcus,* and *Populus*), fungi (*Aspergillus, Cryptococcus, Phanerochaete, Pichia, Saccharomyces, Schizosaccharomyces,* and *Ustilago*), apicomplexans (*Cryptosporidium* and *Plasmodium*), kinetoplastids (*Trypanosoma*), red algae (*Cyanidioschyzon*), amoebozoans (*Dictyostelium*), and stramenopiles (*Phaeodactylum* and *Thalassiosira*). Because of incomplete sequence data from *Chlamydomonas* three different group matrices were created based on whether intron sliding was allowed or not. The first matrix included *Chlamydomonas* in the plant data set and all six subunits of RNA polymerase. The second intron matrix did not include *Chlamydomonas* in plants but still contained all six subunits. The third matrix included *Chlamydomonas* in the plant grouping but only contained 5 subunits (RPA1, B1, B2, C1, C2).

Each of the six different matrices was analyzed using Dollo parsimony in Phylip v3.69 and rooted with the kinetoplastid group. The reason Dollo parsimony was used is that it assumes that an intron will only be gained once in

any given position, but can be lost multiple times from that position. It has been shown in a number of studies (Roy SW and Gilbert W 2006; Carmel L, Rogozin IB et al. 2007; Sverdlov AV, Csuros M et al. 2007) that intron gain is a rare event relative to intron loss over broad scale evolution; therefore Dollo parsimony appears to be the most reasonable computational model based on current assumptions about biological processes. All six equally parsimonious output trees from each matrix were converted into one consensus tree by the majority rule for further analysis. The trees created from the three data sets described above, with intron sliding not allowed were the same (figure 20). The overall topology of the tree follows the currently "accepted" phylogenetic relationships (Hasegawa M, Iida Y et al. 1985; Baldauf SL and Palmer JD 1993); except that animals and plants form sister clades with fungi more distantly related. This flipping of plants and fungi as the nearest relative to animals has been observed in intron based phylogenies in a previous study of RPB1 alone in the laboratory (Harrell 2005). The three matrices permitting intron sliding all produced the same tree topology (figure 21); however, there were some differences in the topology of the no sliding trees. Although these trees also show the same switch between fungi and plants as the sister group to animals, there was movement of stramenopiles from an earlier branching node in previous trees (figure 20) to just before fungi. The intron sliding tree (figure 21) groups the stramenopiles closer to more intron rich taxa (animals, plants, fungi); presumably because the stramenopiles contain a higher density of introns than most protists and by

allowing sliding these introns are more often interpreted as in shared positions with other intron rich taxa.

Because intron sliding data and the no intron sliding data resulted in the same major phylogenetic flipping of plants and fungi as the sister group to animals, but differed in the placement of the stramenopiles group, statistical analysis was performed to determine the significance of this difference. Kishino-Hasegawa-Templeton testing (table 5 and 6) was performed on both sets of data (no intron sliding and intron sliding respectively), results from this testing showed that the sliding tree is significantly worse then the no sliding tree when given the no sliding intron data lacking *Chlamydomonas.* However, the no intron sliding tree is not significantly worse than the sliding tree when given the sliding intron data lacking *Chlamydomonas*. Therefore, since only the no intron sliding tree was a possible alternative model for sliding data, these trees were further tested against topological variations between intron and sequence-based trees.

Comparison of sequence- and intron-based phylogeny

For the comparison between the sequence- and intron-based phylogeny each tree was tested against the best tree recovered from the alternative corresponding data set. Specifically, the intron-based tree was tested statistically against the best sequence-based tree using sequence data, and the sequence-based tree was tested against the best intron-based trees recovered from intron matrices (sliding and no sliding). Because the intron-based phylogenies were created using constrained major taxa rather than all species in the study, a new

tree topology was created by modifying the sequence-based phylogeny to reflect the intron-based phylogenies (both no sliding and sliding trees) using the retree program in Phylip (J Felsenstein 2004) (figures 22-24).

Once all of the trees were assembled the first test used intron data without sliding with dollop in the Phylip package. The results from the KHT test showed that the grouped sequence-based tree was significantly worse than the original no sliding intron tree topology given the no intron sliding data (table 7). To determine if the tree topology was significantly better than the intron sliding topology, the grouped trees were also tested using the intron sliding data in dollop. The results from this KHT testing were similar to the no intron sliding test in that the grouped sequence tree was significantly worse when compared to the intron tree that allowed intron sliding (table 8). The final test was to analyze the different topologies against the sequence data, the results of the SH testing confirmed the same results as the two intron analysis; the alternative models (intron-based trees) were both significantly worse then the original model (sequence-based tree) (table 9).

CONCLUSIONS

The aim of this study was to look at the phylogenies derived from sequence- and intron-based data and statistically compare the two data types to see if there were significant differences between the two data types. This type of analysis can help shed light on the relative importance of intron gain versus loss, and how intron evolution relates to eukaryotic phylogenies. To determine the most accurate sequence-based phylogeny, all of the RNA polymerase subunits were carefully checked for proper annotation and aligned using three different multiple sequence alignment programs. Once aligned, both likelihood and Bayesian analysis were preformed on each subunit as well as concatenated sequences to look at sequence-based phylogeny. These resulted showed animals and fungi grouping together as a sister clade with plants diverging before the animal/fungi split. For the intron data, each intron was coded into a position matrix. To account for intron sliding 2 different intron matrices were created; the first did not allow sliding; therefore, only introns in the exact position were considered to be homologous. For the second position matrix, introns within 6 nucleotides were considered to be homologous. These two matrices were analyzed using dollo parsimony; this showed very different tree topologies from the sequence-based analysis, with animals and plants grouping as sister clades and fungi diverging prior to animals and plants.

Statistical comparisons of these different topologies showed that each tree model was the best tree to its original data compared to the alternative tree topologies. In the case of this study, sequence-based methods recovered a

phylogeny with plants diverging before animals and fungi, while the intron-based methods recovered a phylogeny where plants and animals group closer together, with fungi diverging prior. Both of these trees represent the best tree given the data. This statistically significant difference is strong support for the argument that the evolution of the introns has not followed the pattern of evolution inferred from molecular sequences. This is important to clarify because raises questions about what kinds data should be used for recreating species phylogeny, and what data produce the most accurate phylogeny.

If one considers the sequence phylogeny to be the most accurate, then the evolution of intron positions becomes very complex. Under the prevailing theory that introns appeared early in eukaryotic evolution there are two possible routes to modern intron distributions. The first is that the ancestral eukaryote contained a remarkably high number of introns. This scenario accounts for the large percentage of shared intron positions between deeply diverged taxa in sequence-based phylogenies (plants and animals), but strongly emphasizes the importance of intron loss. The second possible scenario is one favoring intron gain, where ancestral eukaryotes contained a small number of introns and introns were gained, throughout evolution, often in parallel, in the various higher eukaryotic lineages. In this scenario intron gain is very common, with some introns preferring “hot spots”; it is these locations that show up as shared positions between divergent species such as plants and animals. Either scenario involves assumptions of complex patterns of intron evolution.

To help tease apart these two different scenarios of intron evolution increasing the species contained in the study and also the gene number would provide a larger dataset for more comprehensive analyses. One species of high interest would be the marine crustacean *Daphnia*, which recently had its genome sequenced completely by the *Daphnia* Genomics Consortium. Recent sequence analyses in *Daphnia* have shown it to contain a large number of introns, and some of these introns have inserted in parallel in paralogous loci or allelic variants. This case provides evidence that introns can, in fact, insert in parallel in the same spot during evolution (Li W 2009; Omilian AR 2008). Therefore, this would be an ideal species to add to this study to determine whether any of these newly arisen intron positions are shared with other eukaryotic species, especially intron rich taxa such as plants. If some of these new introns indeed share positions with plant introns, this would be strong support for the notion of introns inserting into “hot spots” and, therefore place a larger importance on parallel intron gain during eukaryotic evolution.

In this study, taking the intron-based phylogeny as more accurate than sequence-based phylogeny reduces the complexity of intron gain and loss in evolution. In this phylogeny the plants and animals are more closely related because of their high percentage of shared intron positions. This results in a conflict with sequence homology assumptions that suggest animals and fungi are most closely related. Clearly both methods of phylogenetic reconstruction present difficulties in producing the most parsimonious hypotheses of gene evolution. If both methods produce trees with significant differences between

them, then further research into intron evolution is needed to elucidate how introns are gained and lost. By more fully understanding intron gain and loss rates, questions about intron evolution can be addressed in a more complete manner, possibly shedding light on larger patterns and processes of eukaryotic evolution. While the results of this study do not indicate that intron positions provide a more accurate evolutionary history than molecular sequences in phylogenetic analysis, they do highlight the problem that both methods produce vastly different tree topologies, each significantly rejecting the other.

Table 1. List of species

Species	**Group (Subgroup)**	**Database**
Caenorhabditis elegans	Animals (Roundworms)	WormBase
Drosophila melanogaster	Animals (Insects)	Flybase
Mus musculus	Animals (Mammals)	Mouse Sequencing Consortium
Takifugu rubripes	Animals (Fishes)	DOE Join Genome Institute T. (Fugu) rubripes v4.0
Bos Taurus	Animals (Mammals)	Cattle Genome Sequencing International Consortium
Danio rerio	Animals (Fishes)	Welcome Trust Sanger Institute
Anopheles gambiae	Animals (Insects)	The International Consortium for the Sequencing of Anopheles Genome
Arabidopsis thaliana	Plants (Land plants)	Arabidopsis Information Resource
Chlamydomonas reinhardtii	Plants (Green algae)	DOE Join Genome Institute Chlamy v3.0
Oryza sativa	Plants (Land plants)	Rice Genome Annotation
Ostreococcus lucimarinus	Plants (Green algae)	DOE Join Genome Institute Ostreococcus v2.0
Populus trichocarpa	Plants (Land plants)	DOE Join Genome Institute Populus trichocarpa v1.1
Aspergillus fumigatus	Fungi (Ascomycetes)	TIGR
Saccharomyces cerevisiae	Fungi (Ascomycetes)	Genome Sequencing Center at Washington University
Pichia stipitis	Fungi (Ascomycetes)	DOE Join Genome Institute
Ustilago maydis	Fungi (Basidiomycetes)	Broad Institute

Cryptococcus neoformans	Fungi (Basidiomycetes)	TIGR
Schizosaccharomyces pombe	Fungi (Ascomycetes)	Schizosaccharomyces pombe Gene Database
Phanerochaete chrysosporium	Fungi (Basidiomycetes)	DOE Join Genome Institute Phanerochaete chrysosporium v2.0
Cryptosporidium parvum	Protist (Apicomplexans)	University of Minnesota
Cyanidioschyzon merolae	Protist (Red algae)	National Institute of Genetics, Japan
Dictyostelium discoideum	Protist (Amoebozoa)	The Dictyostelium discoideum Sequencing Consortium
Plasmodium falciparum	Protist (Apicomplexans)	Broad Institute
Trypanosoma brucei	Protist (Kinetoplasts)	Trypanosoma brucei Consortium
Phaeodactylum tricornutum	Protist (Stramenopiles)	Diatom Consortium
Thalassiosira pseudonana	Protist (Stramenopiles)	DOE Joint Genome Institute

Table 2. Species requiring manual annotation

Subunit	Species	Problem	Number of Exons Changed	Number of Introns Changed
RPA1	*Populus*	Truncated 3’ region	+3	+3
	Chlamydomonas	Truncated 3’ region	+1	+1
	Cryptosporidium	Poor alignment in the 3’ region	+1	+1
RPA2	*Thalassiosira*	Missing 5’ region	+1	+1
RPB1	*Phaeodactylum*	Incorrect starting sequence	+1	+1
RPB2	*Takifugu*	Missing 5’ and 3’ data	+6	+6
	Populus	Mis-annotated 3’ end	+5	+5
RPC1	*Takifugu*	Missing 5’ region	+5	+4
			+1	
RPC2	*Anopheles*	Mis-annotated intron sequence	+1	-1
	Thalassiosira	Incorrect starting sequence	+1	+1

Table 3. Intron numbers by species

Species	Subunit						Total
	RPA1	RPA2	RPB1	RPB2	RPC1	RPC2	
Anopheles	0	3	1	1	8	3	16
Arabidopsis	20	26	10	23	26	37	142
Aspergillus	1	2	3	1	1	2	10
Bos	33	14	25	24	30	27	153
Caenorhabditis	8	9	8	9	9	6	49
Chlamydomonas	14	N/A	28	20	27	28	117
Cryptococcus	6	9	11	6	15	5	52
Cryptosporidium	0	0	0	0	0	0	0
Cyanidioschyzon	0	0	0	0	0	0	0
Danio	33	14	24	24	30	27	152
Dictyostelium	2	2	1	3	1	2	11
Drosophila	10	2	2	3	5	1	23
Mus	33	14	25	24	30	27	153
Oryza	20	26	10	23	26	37	142
Ostreococcus	1	3	1	2	1	0	8
Phaeodactylum	4	5	1	0	5	2	17
Phanerochaete	10	9	9	6	13	13	60
Pichia	0	0	0	0	0	0	0
Plasmodium	1	1	0	2	3	0	7
Populus	21	23	10	23	25	36	138
Saccharomyces	0	0	0	0	0	0	0

Schizosaccharomyces	0	0	6	1	1	0	8
Takifugu	35	14	25	20	31	27	152
Thalassiosira	9	8	4	4	4	5	34
Trypanosoma	0	0	0	0	0	0	0
Ustilago	0	0	1	0	0	0	1

Table 4. Intron numbers by analysis method

Subunit	No Sliding	Sliding
RPA1	119	102
RPA2	83	70
RPB1	76	70
RPB2	99	72
RPC1	105	91
RPC2	125	99
Total	607	504

Table 5. Kishino-Hasegawa-Templeton testing between sliding and no sliding intron data given no sliding data

Tree	Steps	Diff Steps*	S.D.	Significantly worse
Intron tree (no sliding)	26.0			Best Tree
Intron tree (sliding)	33.0	0.7	0.3003	Yes

*Variance of step differences between trees, taken across characters

Table 6. Kishino-Hasegawa-Templeton testing between sliding and no sliding intron data given sliding data

Tree	Steps	Diff Steps*	S.D.	Significantly worse
Intron tree (no sliding)	62.0	0.5	0.4364	No
Intron tree (sliding)	57.0			Best Tree

*Variance of step differences between trees, taken across characters

Table 7. Kishino-Hasegawa-Templeton testing of the no sliding intron data

Tree	Steps	Diff Steps*	S.D.	Significantly worse
Grouped sequence tree	98.0	7.1	1.2051	Yes
Intron tree (no sliding)	27.0			Best Tree

*Variance of step differences between trees, taken across characters

Table 8. Kishino-Hasegawa-Templeton testing of the sliding intron data

Tree	Steps	Diff Steps*	S.D.	Significantly worse
Grouped sequence tree	157.0	9.9	1.5344	Yes
Intron tree (sliding)	58.0			Best Tree

*Variance of step differences between trees, taken across characters

Table 9. Shimodaira-Hasegawa testing of the sequence data

Tree	logL	Diff logL	p-value	Significantly worse
Sequence-based tree	-58987.6			Best Tree
Modified sequence-based tree*	-59089.7	-102.0	0.000	Yes
Modified sequence-based tree#	-59083.5	-95.9	0.000	Yes

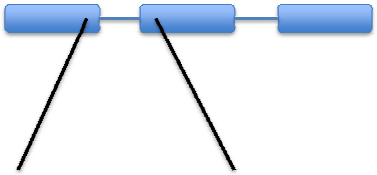
* = Reflects the intron tree without sliding

= Reflects the intron tree with sliding

Figure 1. Intron position color code

Phase 0 Introns:

Located between two codons for different amino acids (colored blue)



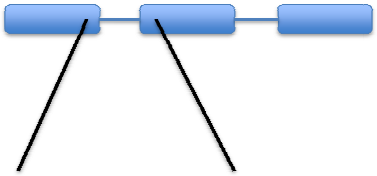
CTA|GTTAGCTA*****AGCGCTAG|ATC

His- Thr

HT

Phase 1 Introns:

Located between the first and second nucleotide of a codon (color green)



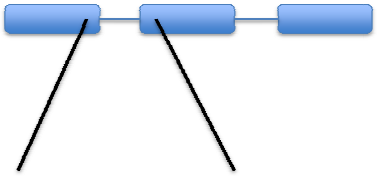
C|GTTAGCTA*****AGCGCTAG|TAATC

H is-Thr

HT

Phase 2 Introns:

Located between the second and third nucleotide of a codon (colored red)

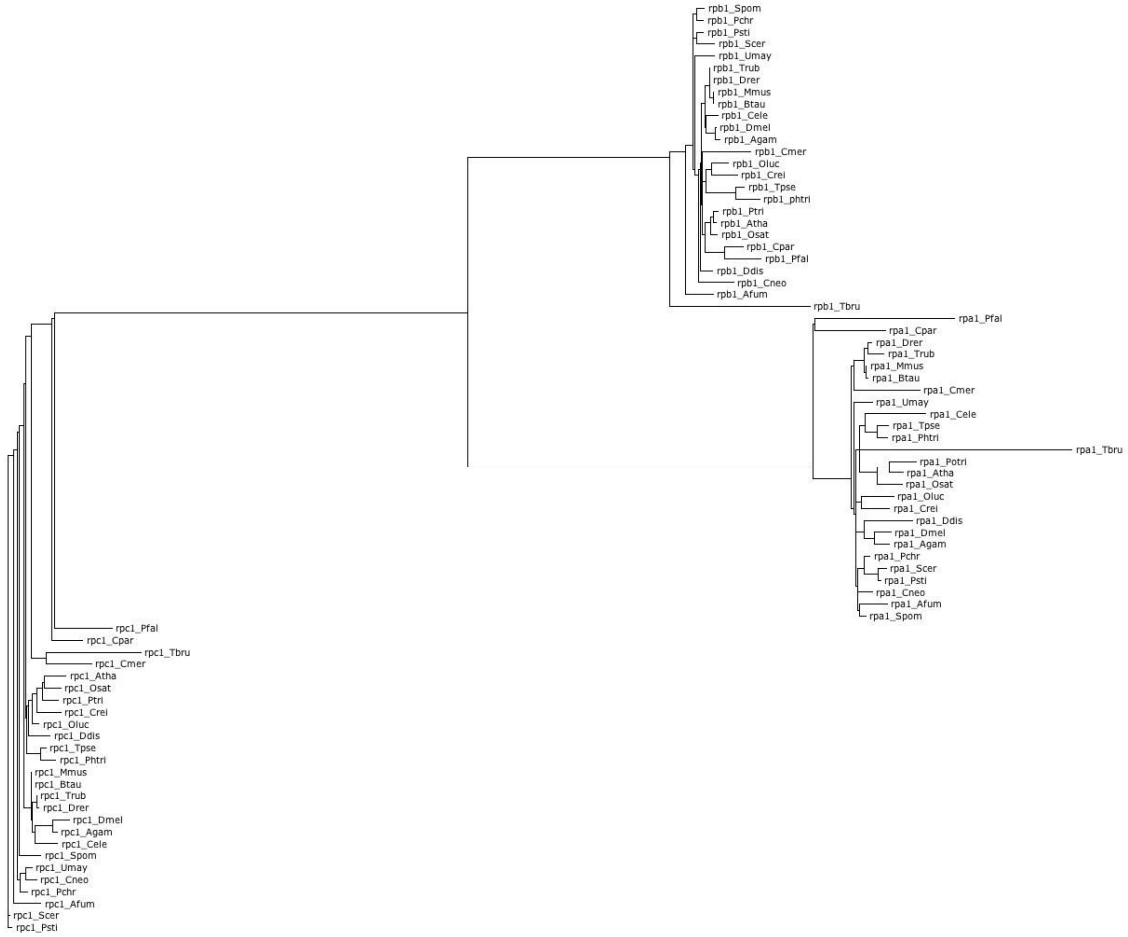


CT|GTTAGCTA*****AGCGCTAG|AATC

Hi s-Thr

HT

Figure 2. Tree of largest subunits



Phylogenetic tree recovered by maximum likelihood on protein sequence from the largest subunits of RNA polymerase I, II, and III for the 26 species. The tree was unrooted.

Subunit notation (format: subunit_species)

rpa1 = RNA polymerase I largest subunit

rpa2 = RNA polymerase I second largest subunit

rpb1 = RNA polymerase II largest subunit

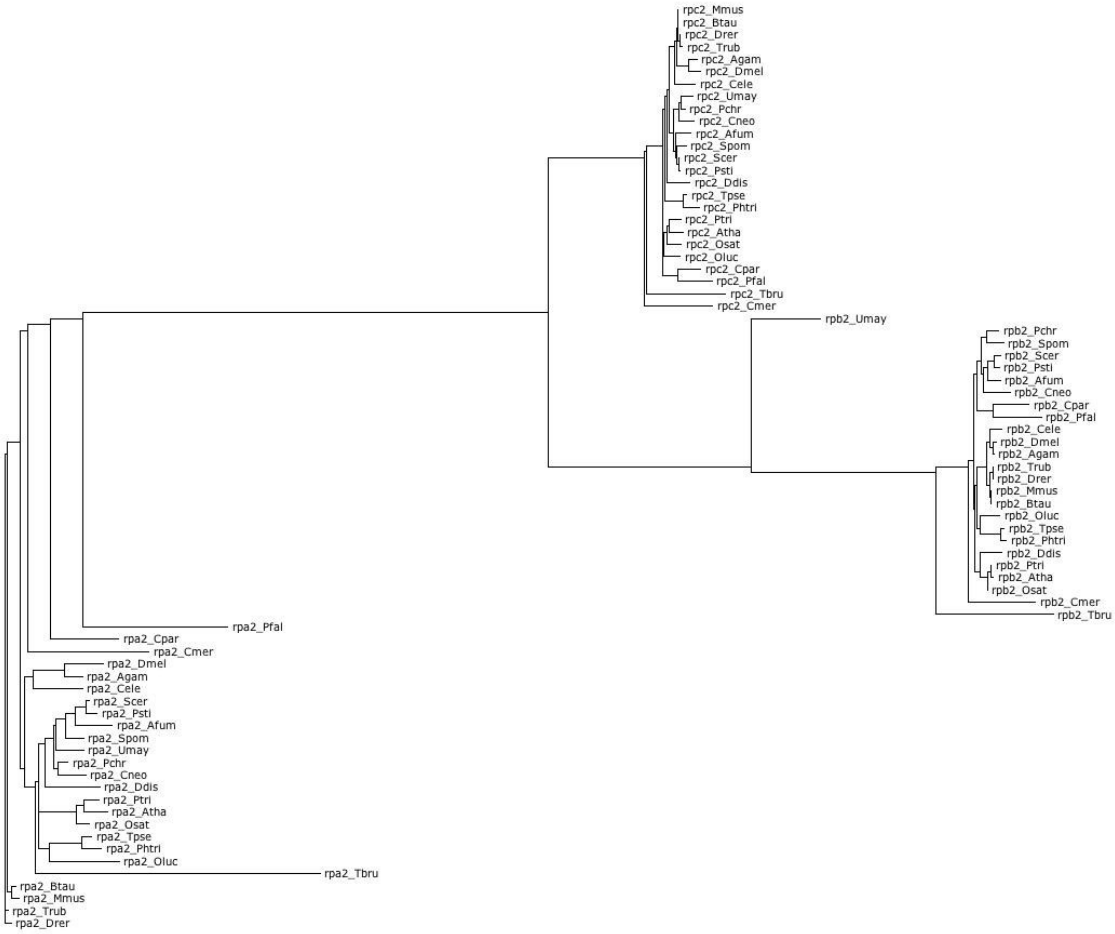
rpb2 = RNA polymerase II second largest subunit

rpc1 = RNA polymerase III largest subunit

rpc2 = RNA polymerase III second largest subunit

Afum = *Aspergillus fumigatus*, Agam = *Anopheles gambiae,* Atha = *Arabidopsis thaliana,* Btau = *Bos Taurus,* Cele = *Caenorhabditis elegans*, Cmer = *Cyanidioschyzon merolae,* Cneo = *Cryptococcs neoformans,* Cpar = *Cryptosporidium parvum,* Crei = *Chlamydomonas reinhardtii,* Ddis = *Dictyostelium discoideum,* Dmel = *Drosophila melanogaster,* Drer = *Danio rerio,* Mmus = *Mus musculus,* Oluc = *Ostreococcus lucimarinus,* Osat = *Oryza sativa,* Pchr = *Phanerochaete chrysosporium,* Pfal = *Plasmodium falciparum,* Phtri = *Phaeodactylum tricornutum,* Potri = *Populus trichocarpa,* Psti = *Pichia stipitis,* Scer = *Saccharomyces cerevisiae,* Spom = *Schizosaccharomyces pombe,* Tbru = *Trypanosoma brucei,* Tpse = *Thalassiosira pseudonana,* Trub = *Takifugu rubripes,* and Umay = *Ustilago maydis*

Figure 3. Tree of second largest subunits



Phylogenetic tree recovered by maximum likelihood on protein sequence from the second largest subunits of RNA polymerase I, II, and III for the 26 species. The tree was unrooted.

Subunit notation (format: subunit_species)

rpa1 = RNA polymerase I largest subunit

rpa2 = RNA polymerase I second largest subunit

rpb1 = RNA polymerase II largest subunit

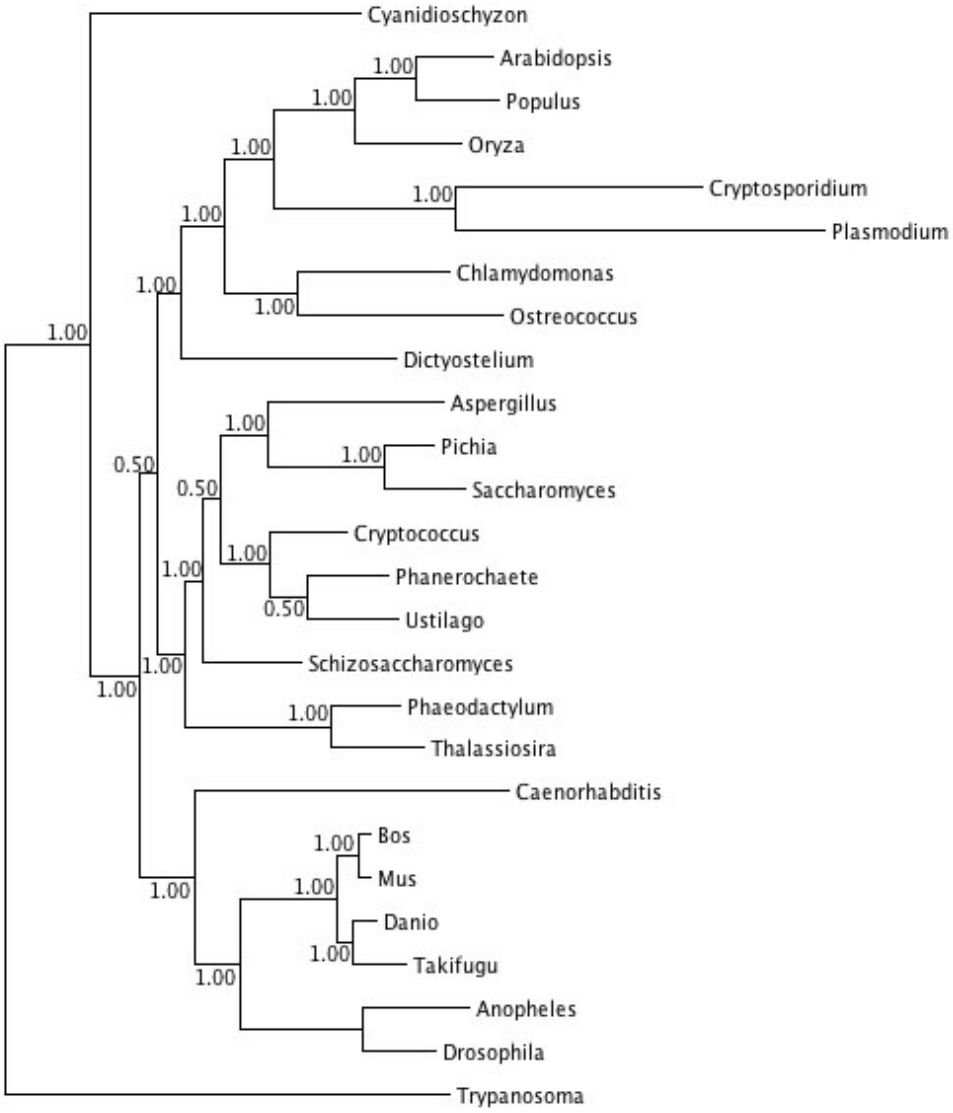
rpb2 = RNA polymerase II second largest subunit

rpc1 = RNA polymerase III largest subunit

rpc2 = RNA polymerase III second largest subunit

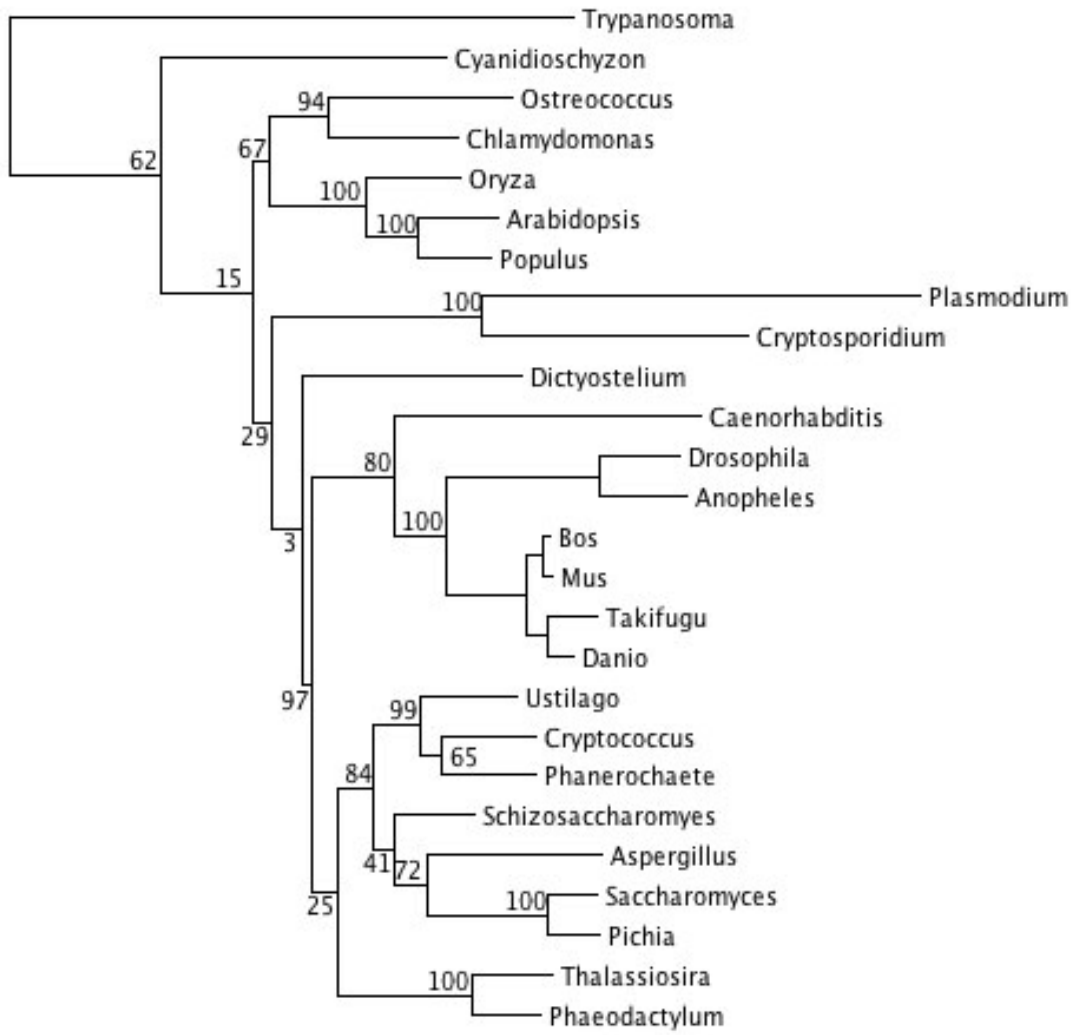
Afum = *Aspergillus fumigatus*, Agam = *Anopheles gambiae,* Atha = *Arabidopsis thaliana,* Btau = *Bos Taurus,* Cele = *Caenorhabditis elegans*, Cmer = *Cyanidioschyzon merolae,* Cneo = *Cryptococcs neoformans,* Cpar = *Cryptosporidium parvum,* Crei = *Chlamydomonas reinhardtii,* Ddis = *Dictyostelium discoideum,* Dmel = *Drosophila melanogaster,* Drer = *Danio rerio,* Mmus = *Mus musculus,* Oluc = *Ostreococcus lucimarinus,* Osat = *Oryza sativa,* Pchr = *Phanerochaete chrysosporium,* Pfal = *Plasmodium falciparum,* Phtri = *Phaeodactylum tricornutum,* Potri = *Populus trichocarpa,* Psti = *Pichia stipitis,* Scer = *Saccharomyces cerevisiae,* Spom = *Schizosaccharomyces pombe,* Tbru = *Trypanosoma brucei,* Tpse = *Thalassiosira pseudonana,* Trub = *Takifugu rubripes,* and Umay = *Ustilago maydis*

Figure 4. RPA1 MrBayes



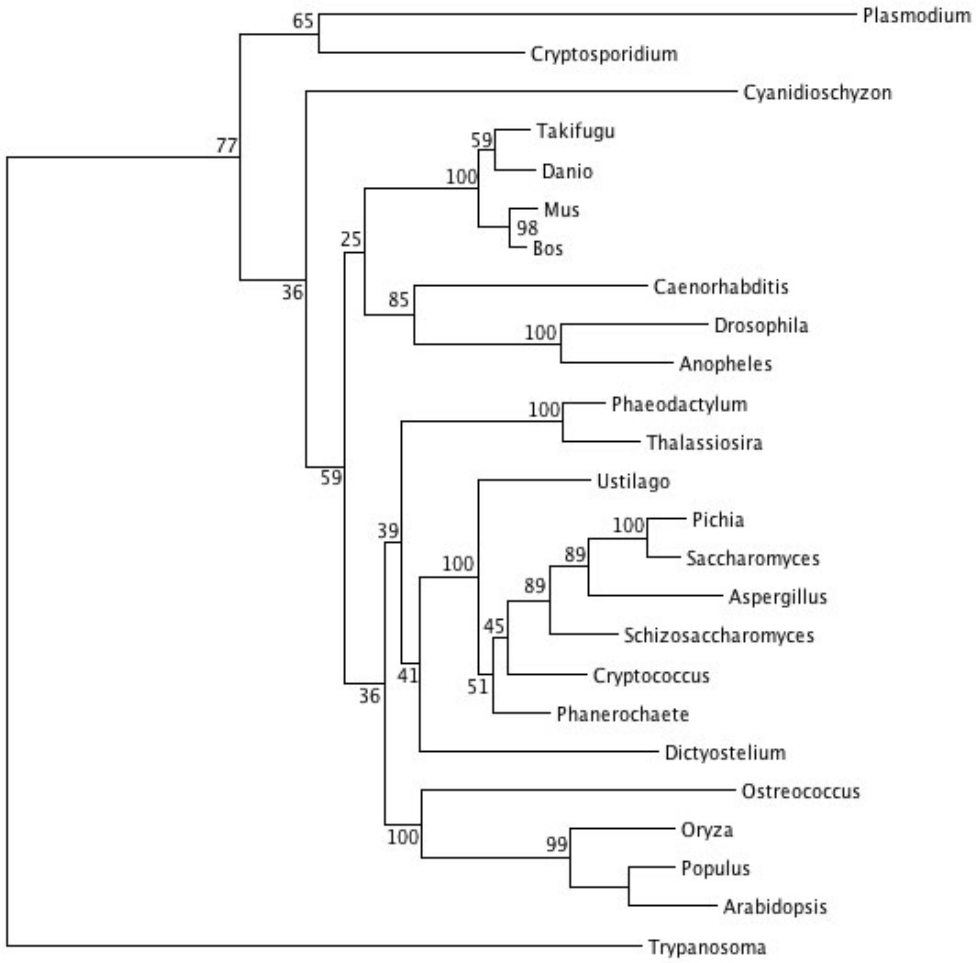
Phylogenetic tree recovered by Bayesian inference on protein sequence of RPA1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 5. RPA1 PhyML



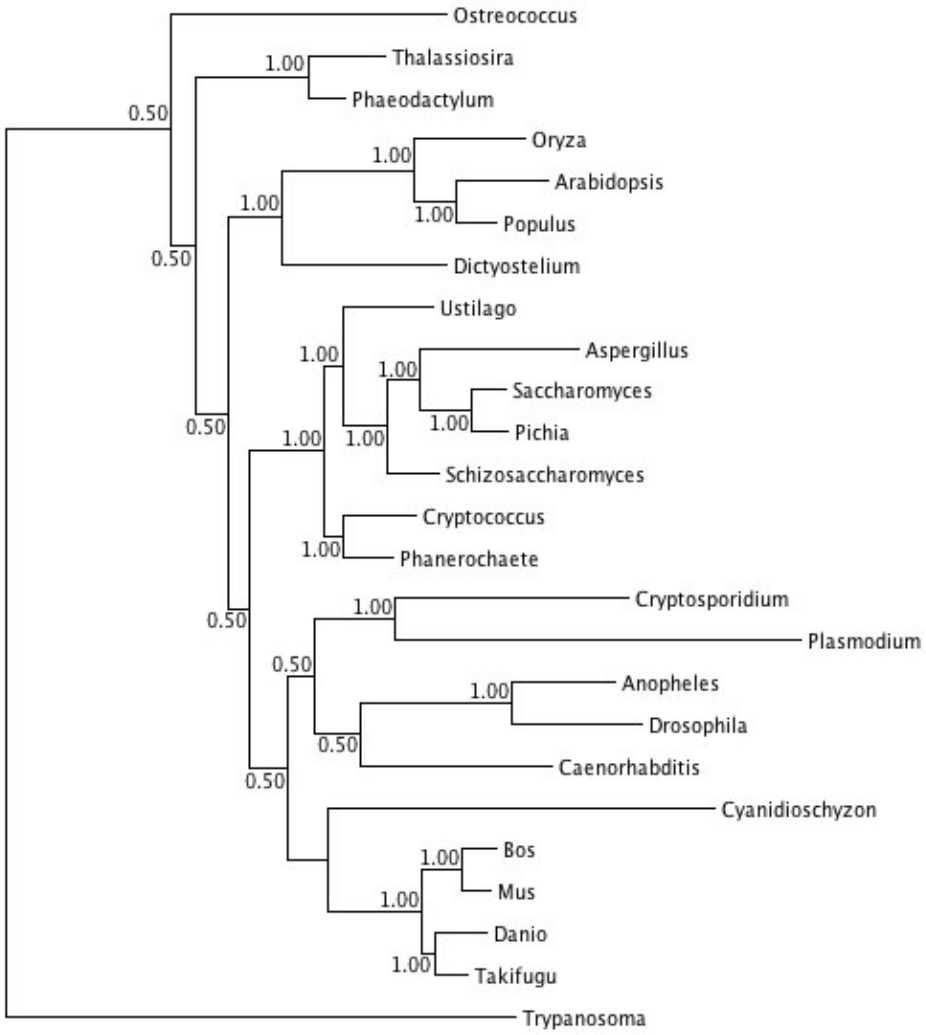
Phylogenetic tree recovered by maximum likelihood on protein sequence of RPA1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 6. RPA2 PhyML



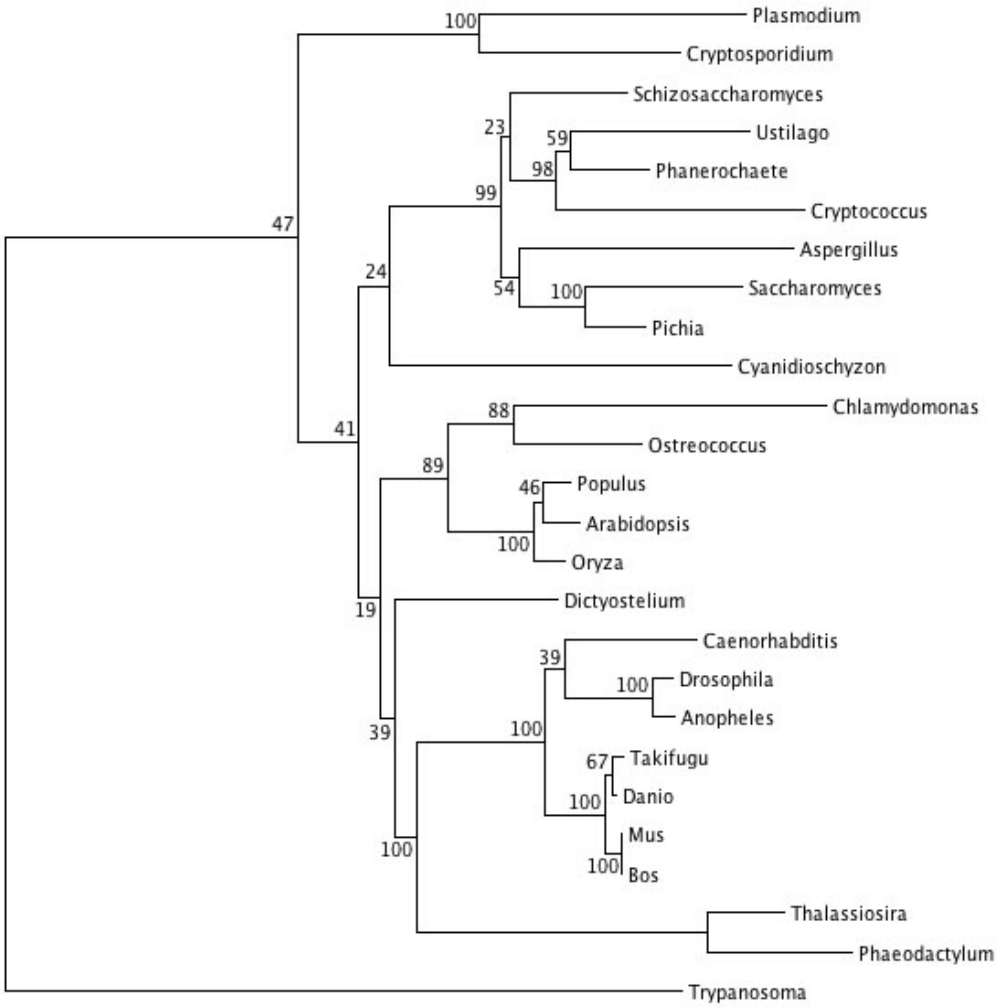
Phylogenetic tree recovered by maximum likelihood on protein sequence of RPA1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 7. RPA2 MrBayes



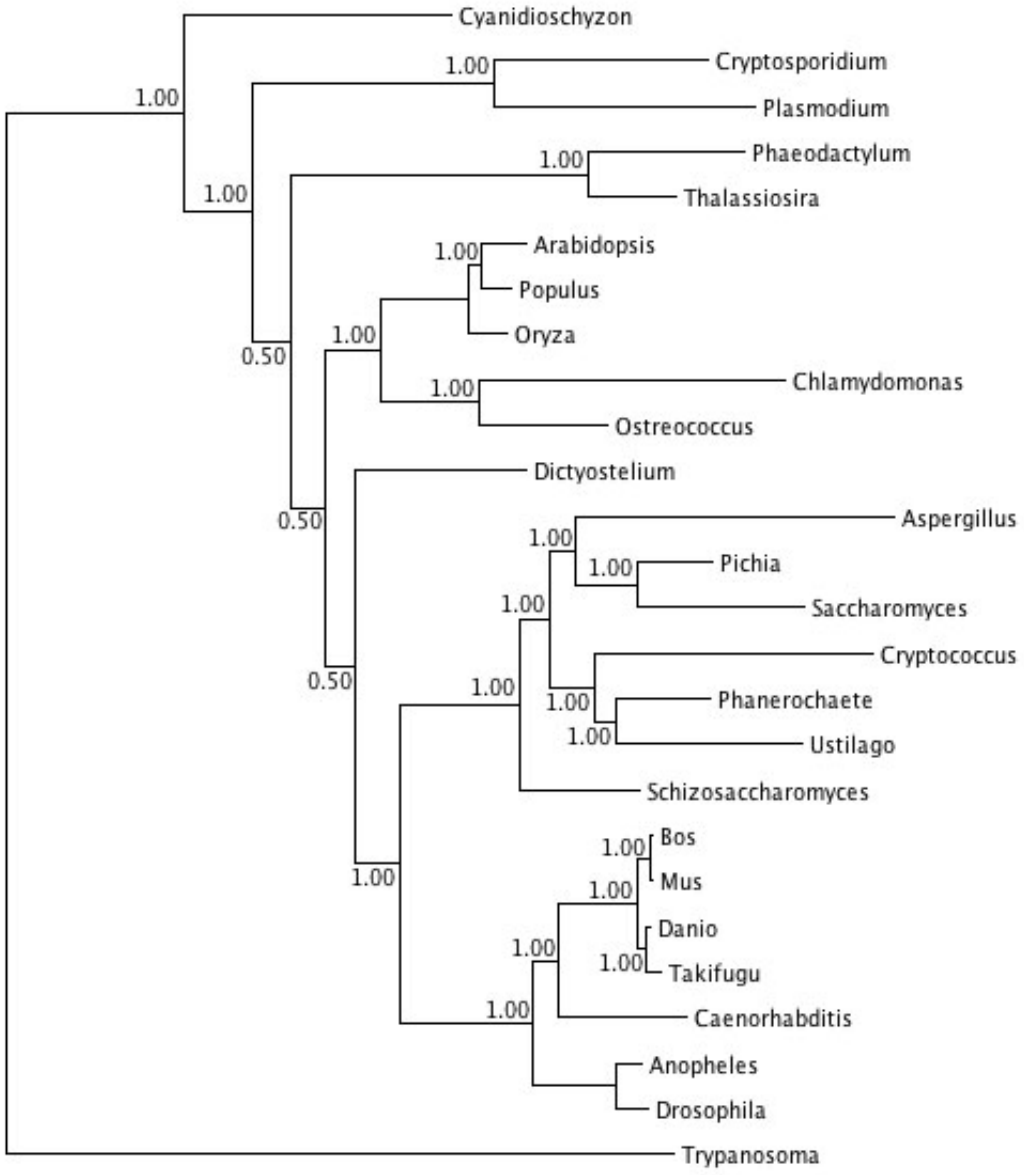
Phylogenetic tree recovered by Bayesian inference on protein sequence of RPA2 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 8. RPB1 PhyML



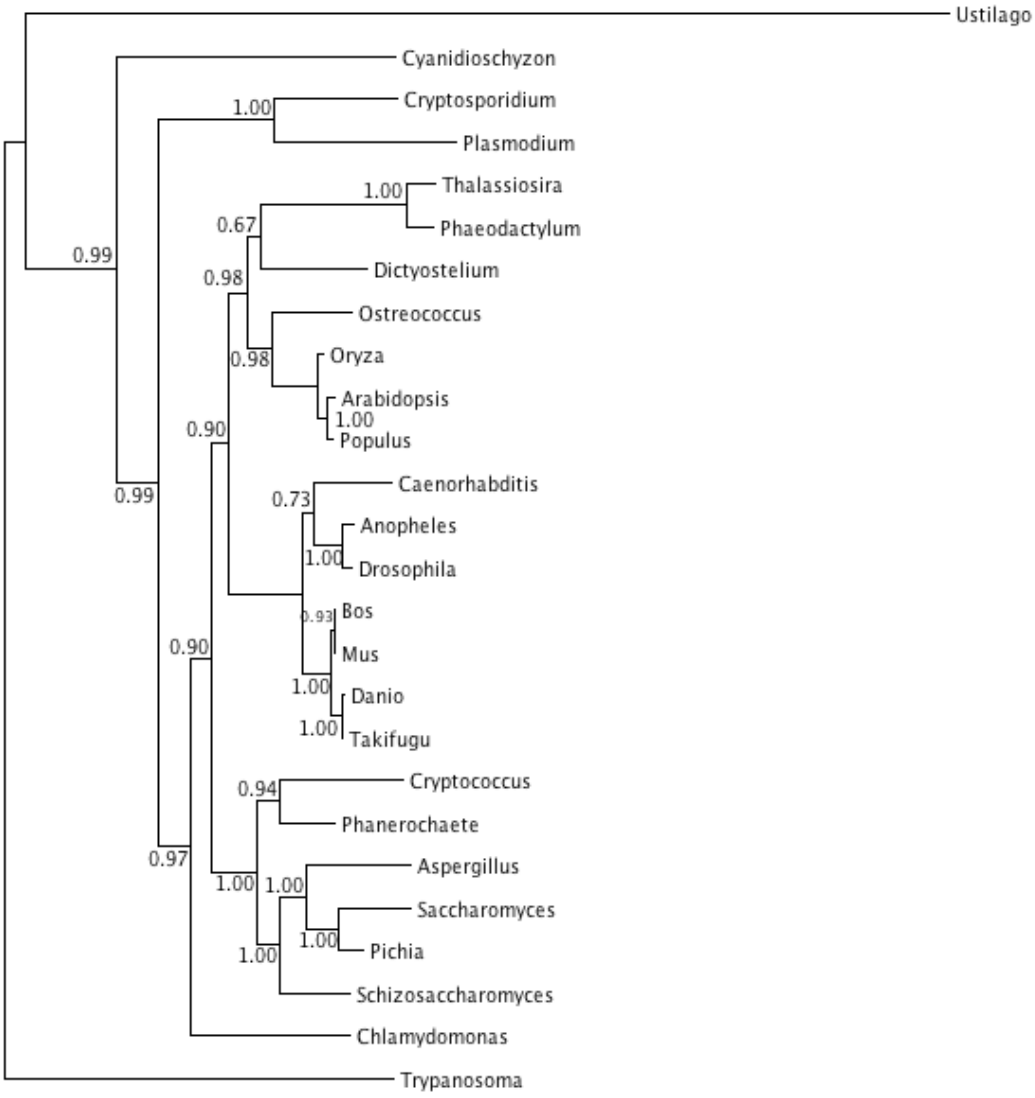
Phylogenetic tree recovered by maximum likelihood on protein sequence of RPB1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 9. RPB1 MrBayes



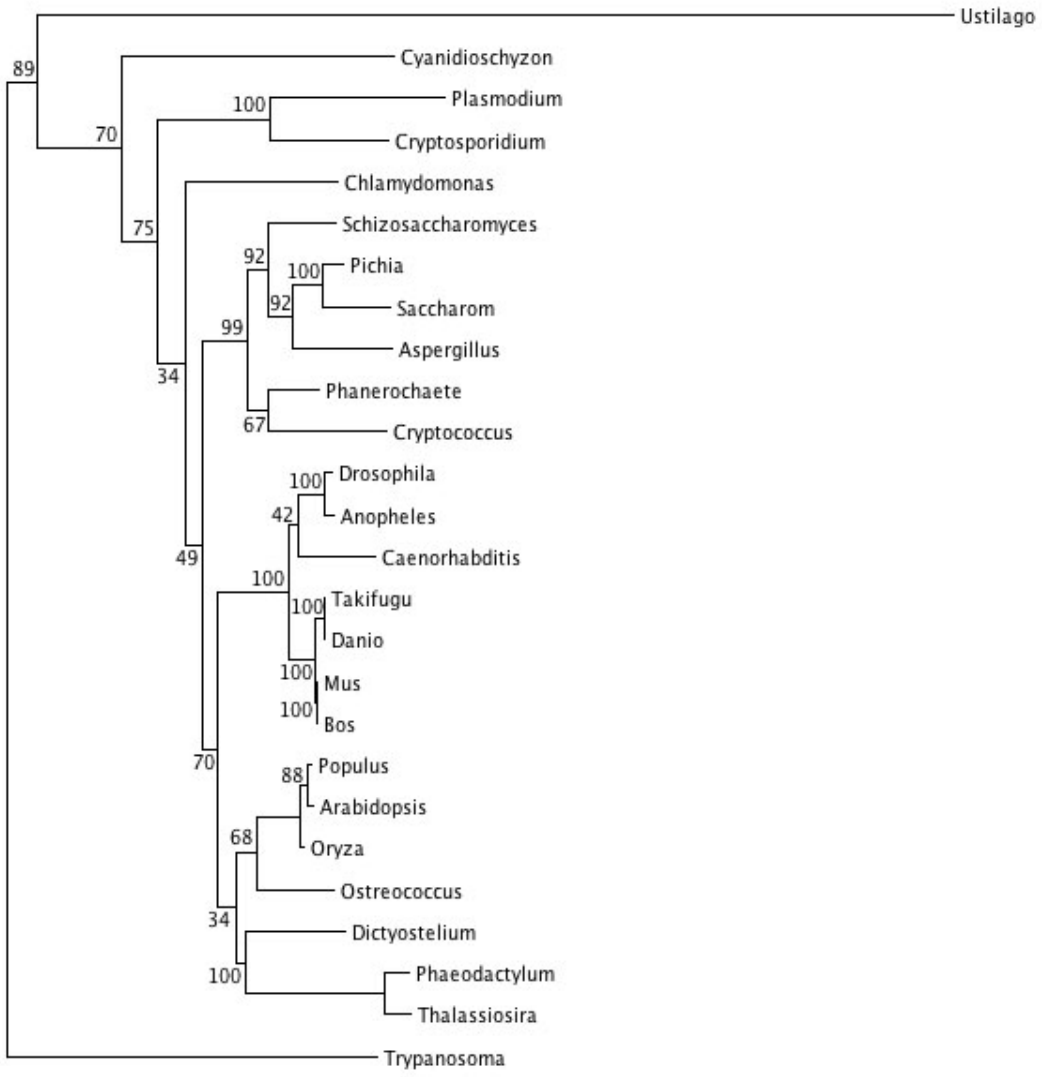
Phylogenetic tree recovered by Bayesian inference on protein sequence of RPB1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 10. RPB2 MrBayes



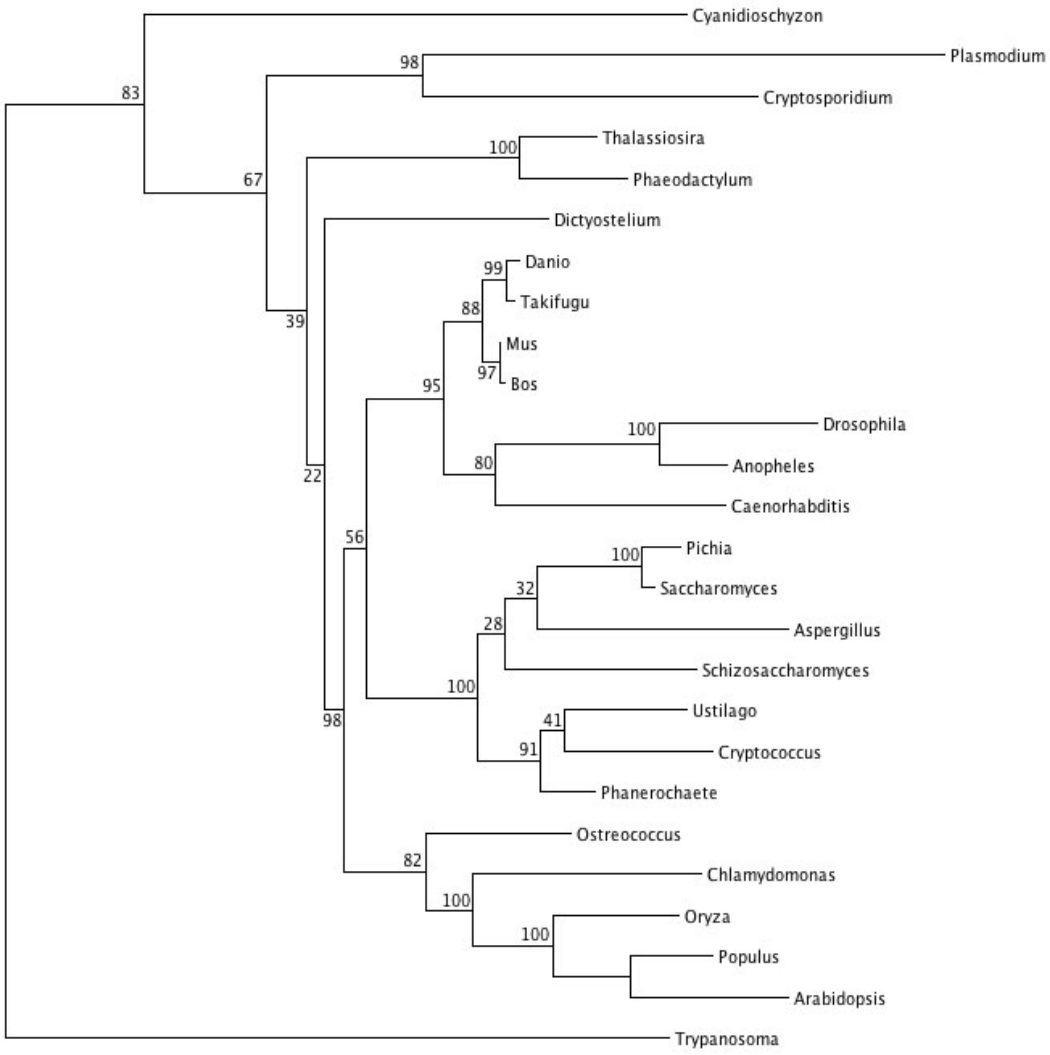
Phylogenetic tree recovered by Bayesian inference on protein sequence of RPB2 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 11. RPB2 PhyML



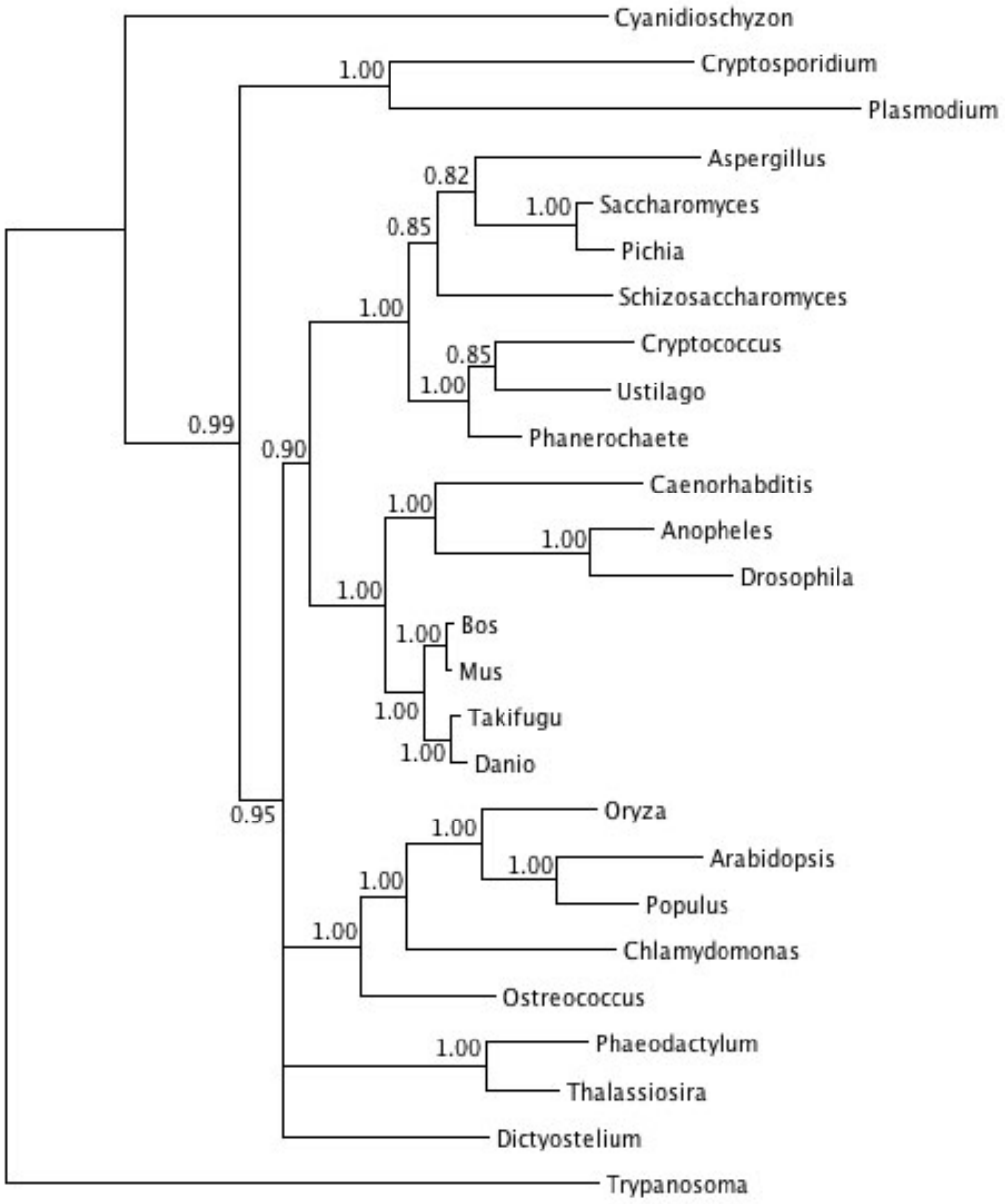
Phylogenetic tree recovered by maximum likelihood on protein sequence of RPB2 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 12. RPC1 PhyML



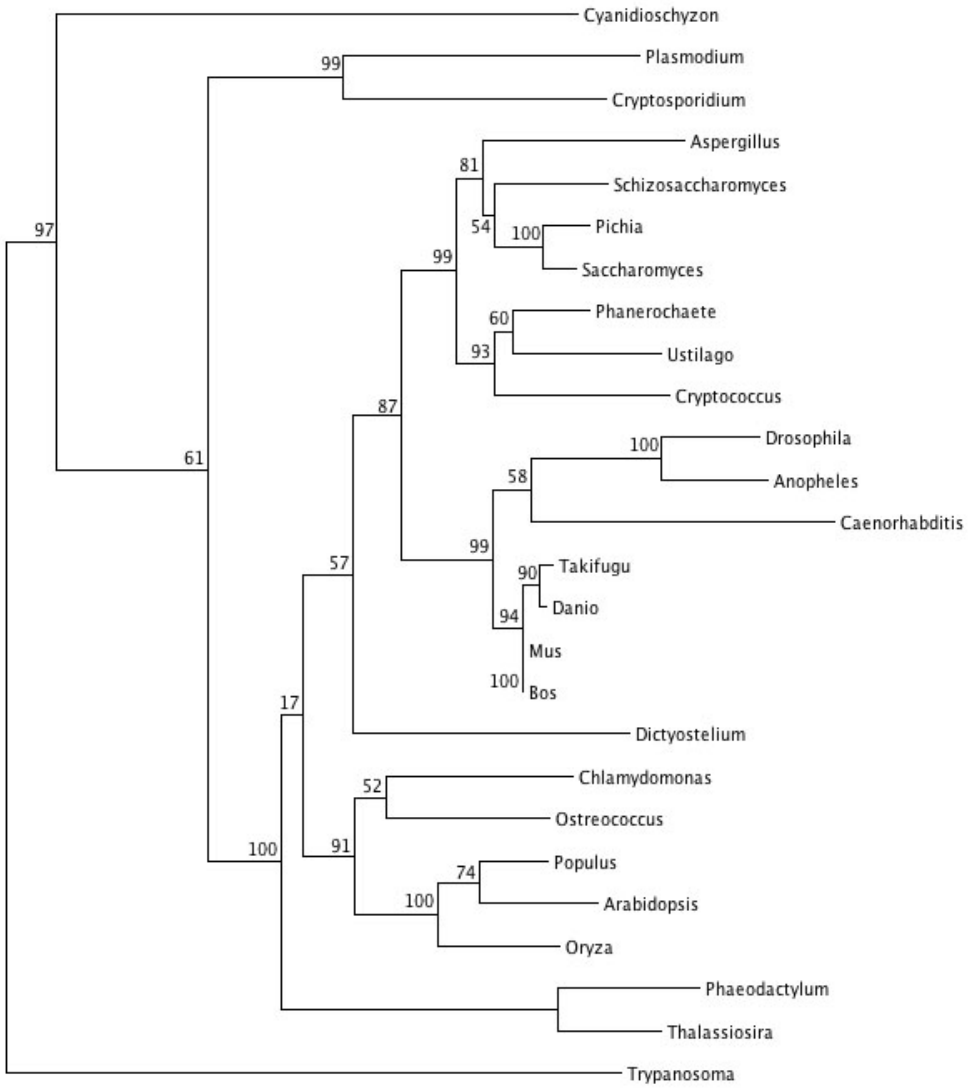
Phylogenetic tree recovered by maximum likelihood on protein sequence of RPC1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 13. RPC1 MrBayes



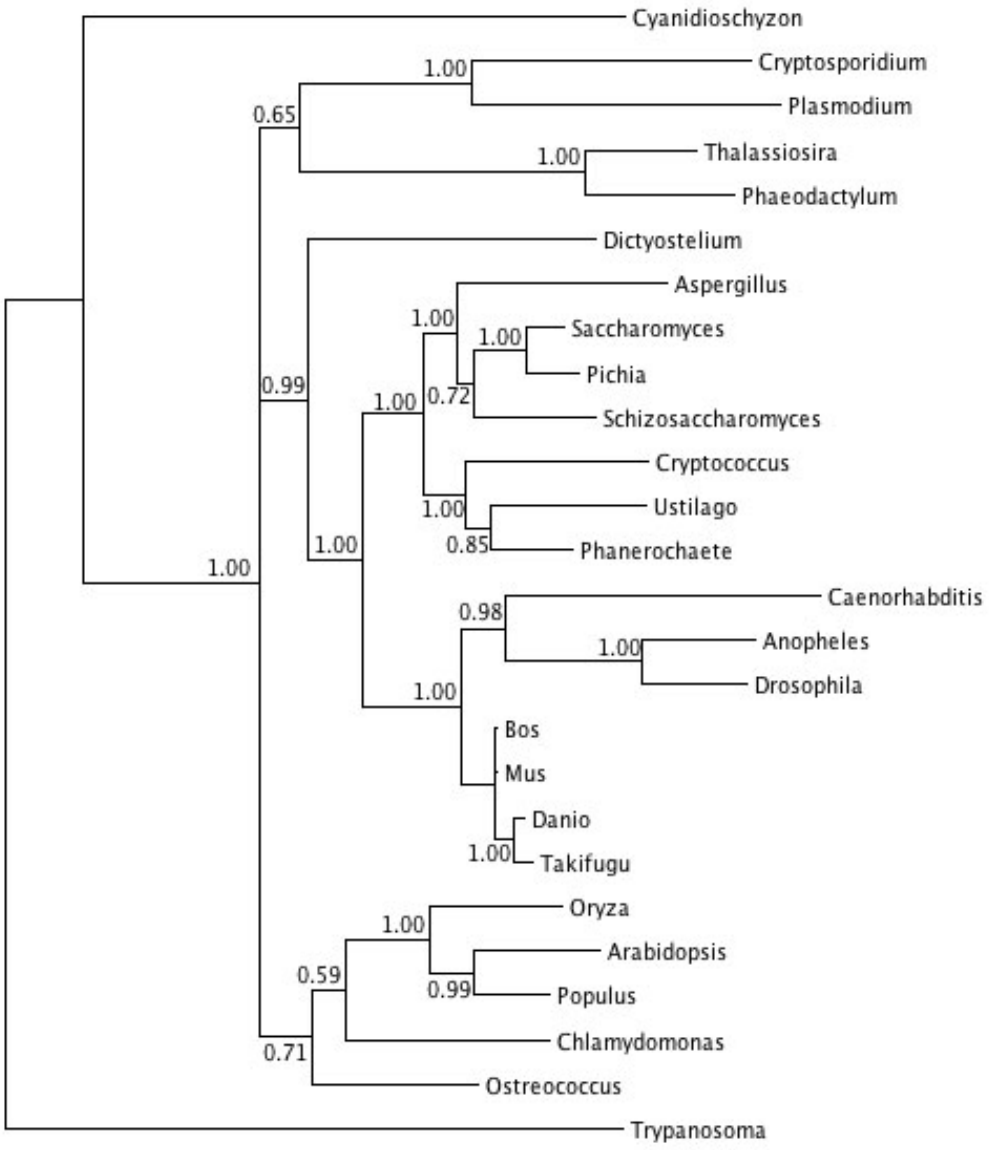
Phylogenetic tree recovered by Bayesian inference on protein sequence of RPC1 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 14. RPC2 PhyML



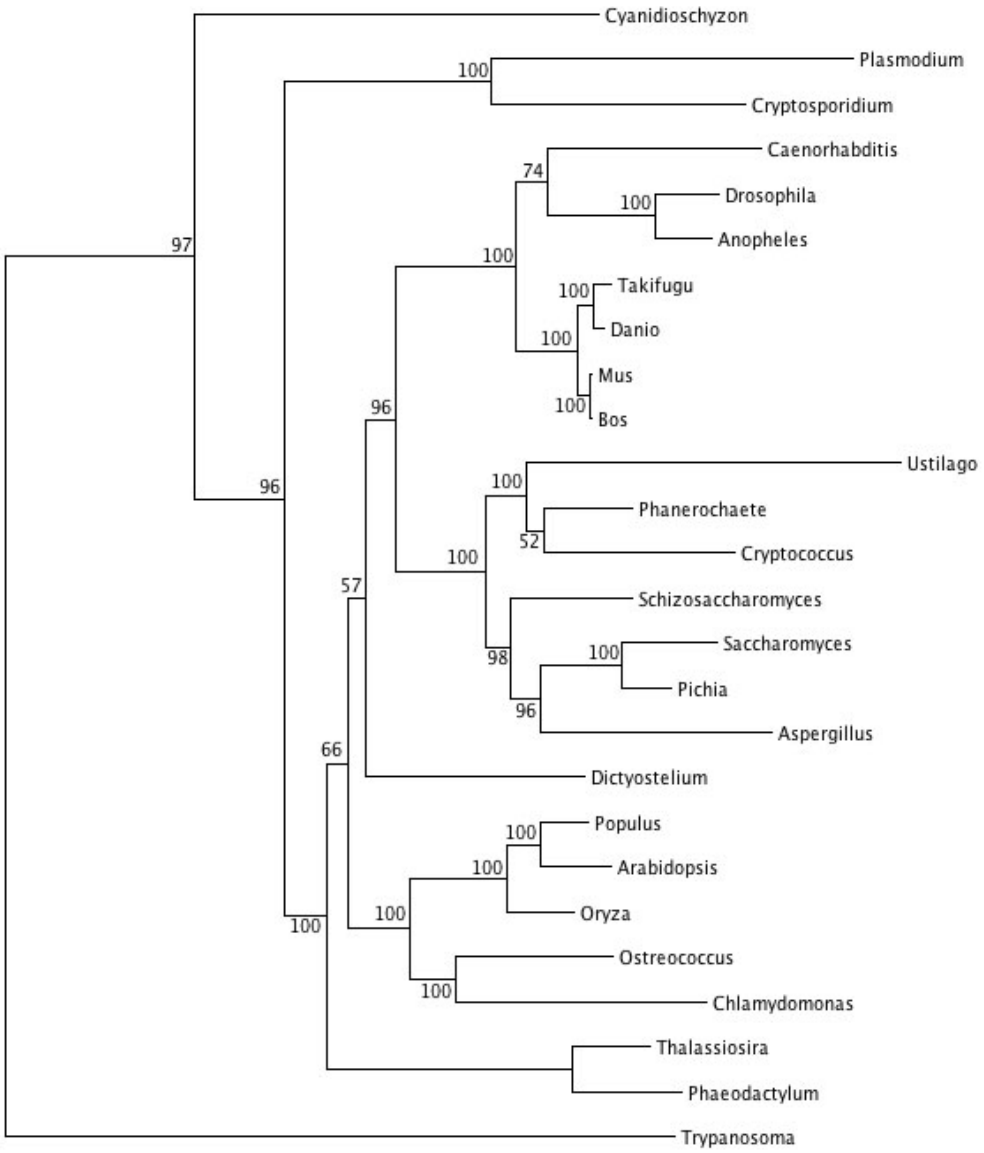
Phylogenetic tree recovered by maximum likelihood on protein sequence of RPC2 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 15. RPC2 Mrbayes



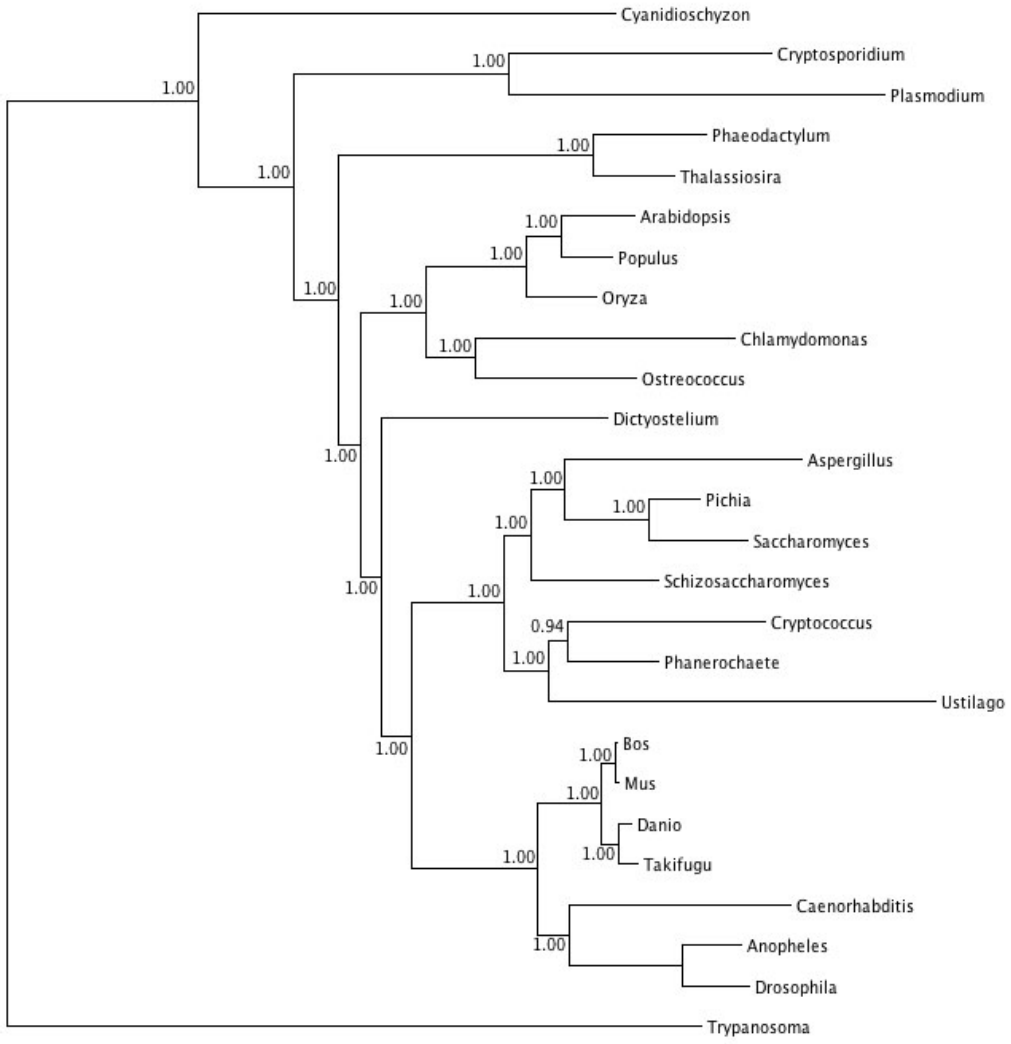
Phylogenetic tree recovered by Bayesian inference on protein sequence of RPC2 for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 16. Concatenated tree including *Chlamydomonas* (PhyML)



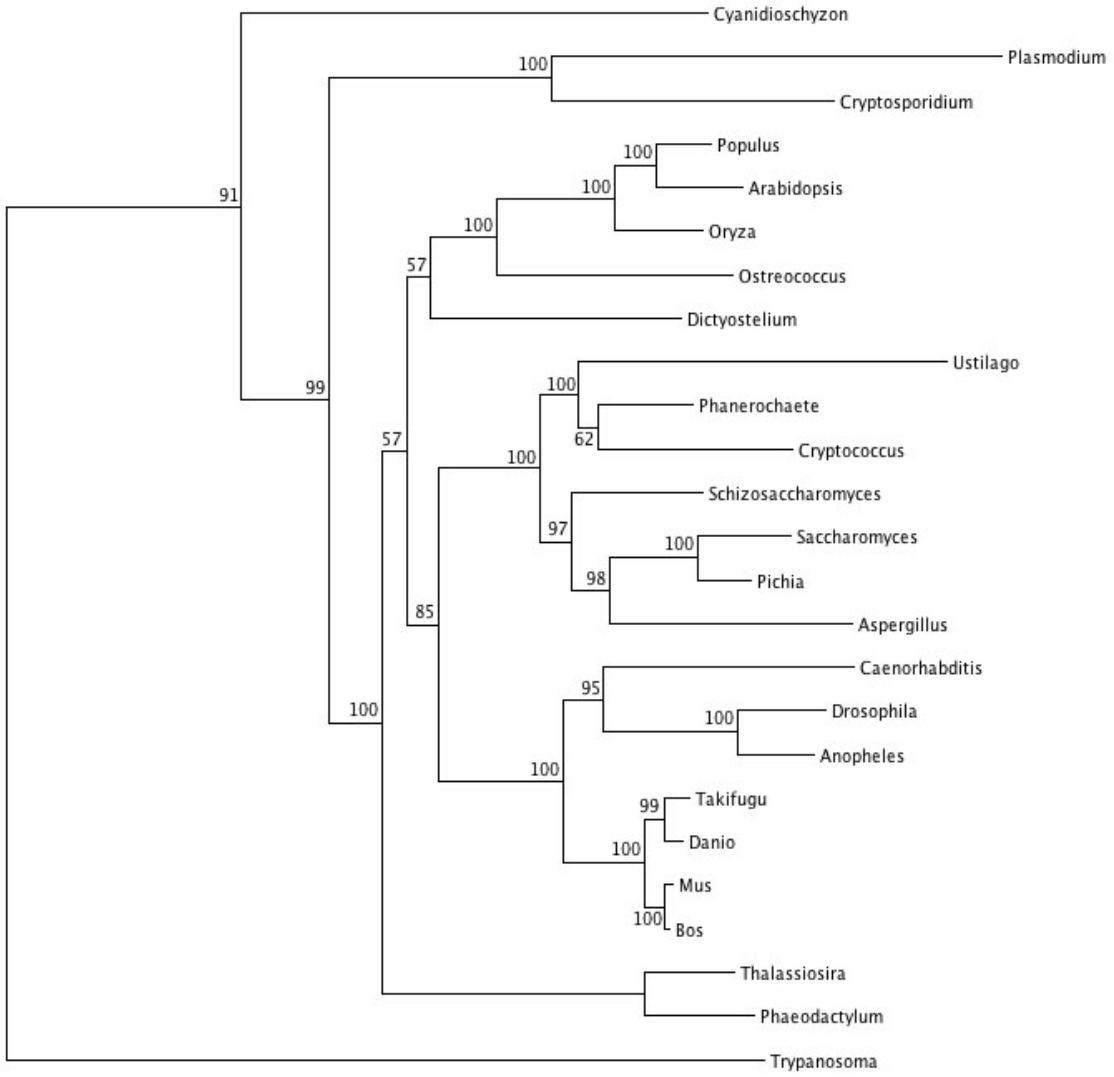
Phylogenetic tree recovered by maximum likelihood on all concatenated subunits for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 17. Concatenated tree including *Chlamydomonas* (MrBayes)



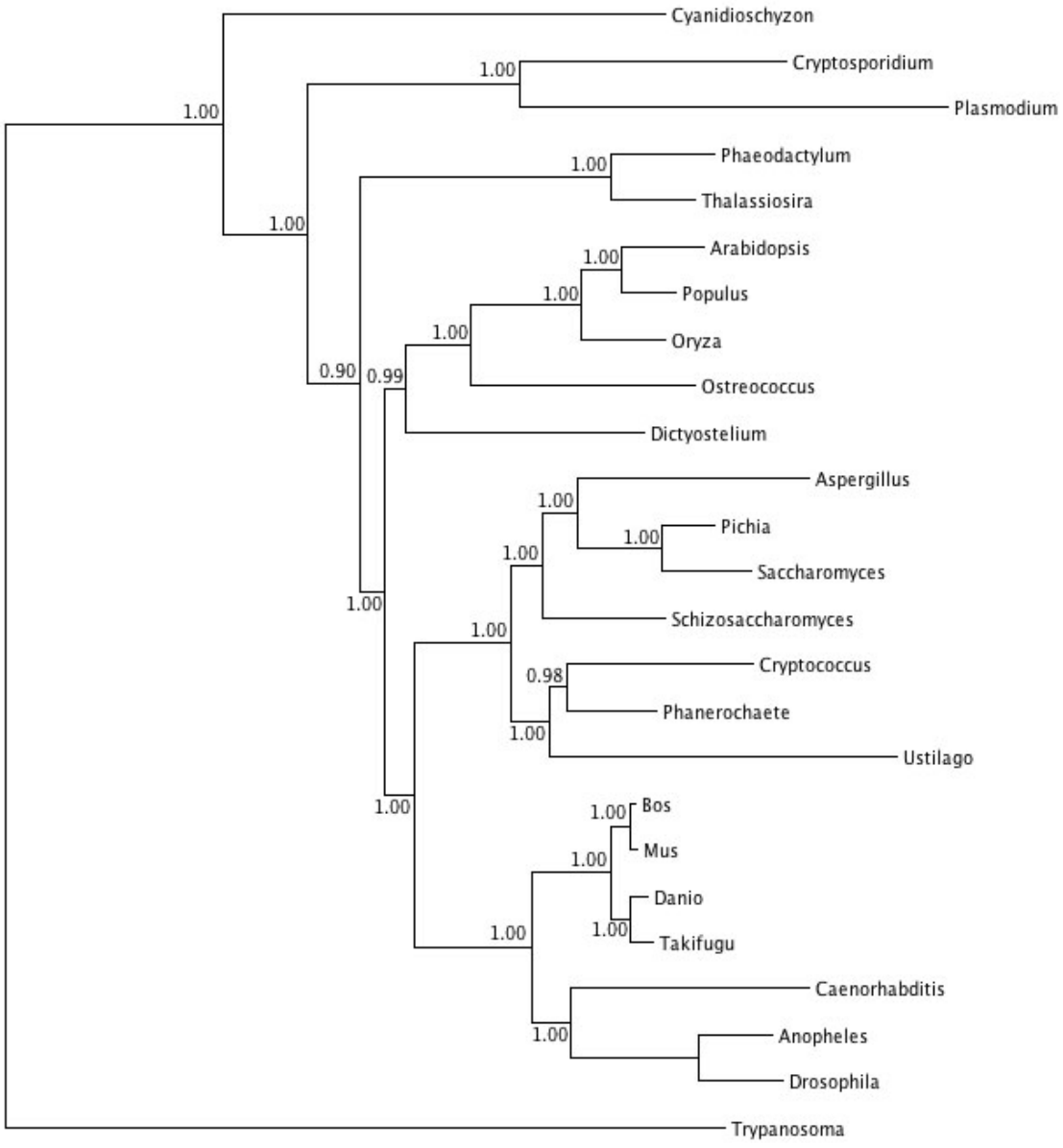
Phylogenetic tree recovered by Bayesian inference on all concatenated subunits for the 26 species. The tree was rooted with *Trypanosoma* species.

Figure 18. Concatenated tree excluding *Chlamydomonas* (PhyML)



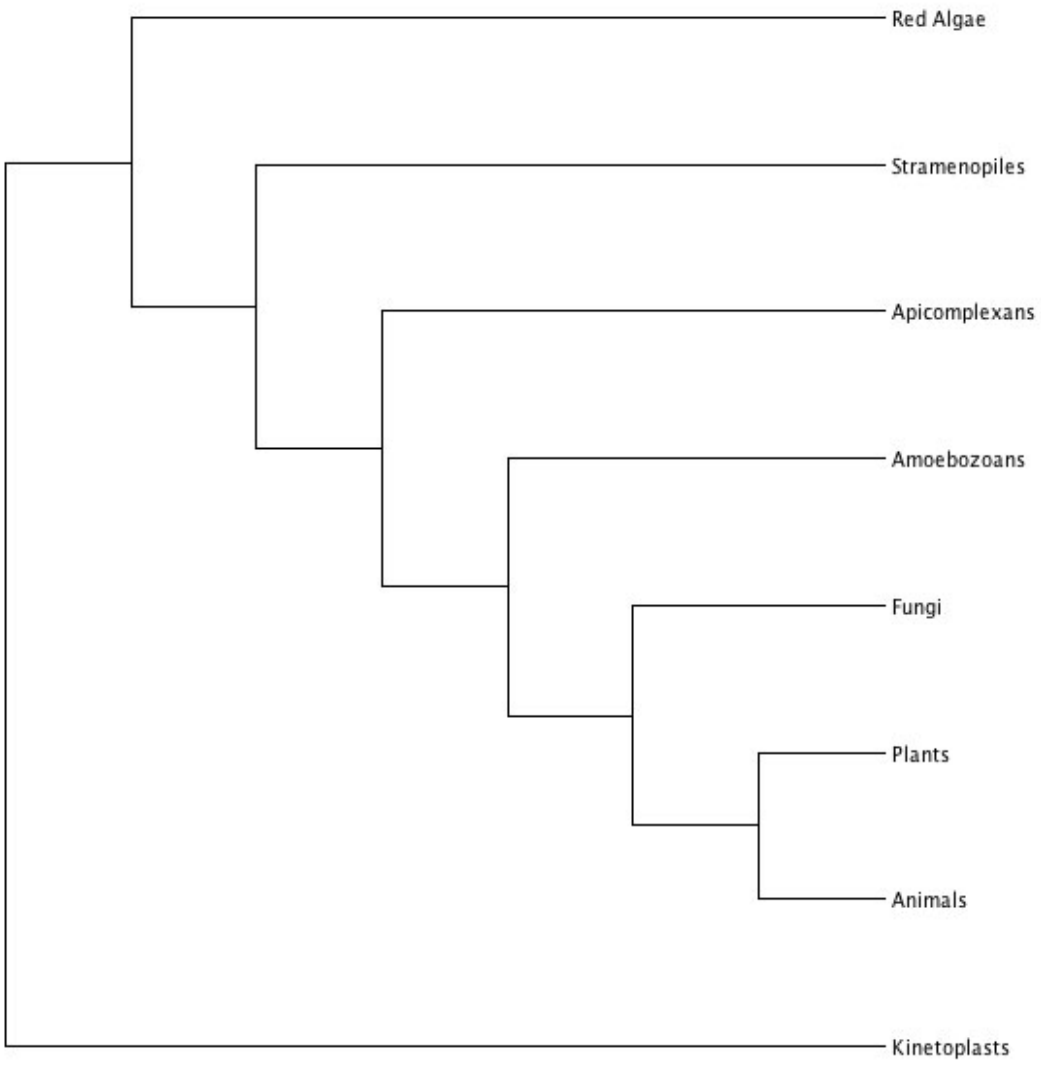
Phylogenetic tree recovered by maximum likelihood on concatenated subunits (RPA1, B1, B2, C1, C2) for the 25 species (*Chalmydomonas* was excluded due to poor RPA2 sequence annotation). The tree was rooted with *Trypanosoma* species.

Figure 19. Concatenated tree excluding *Chlamydomonas* (MrBayes)



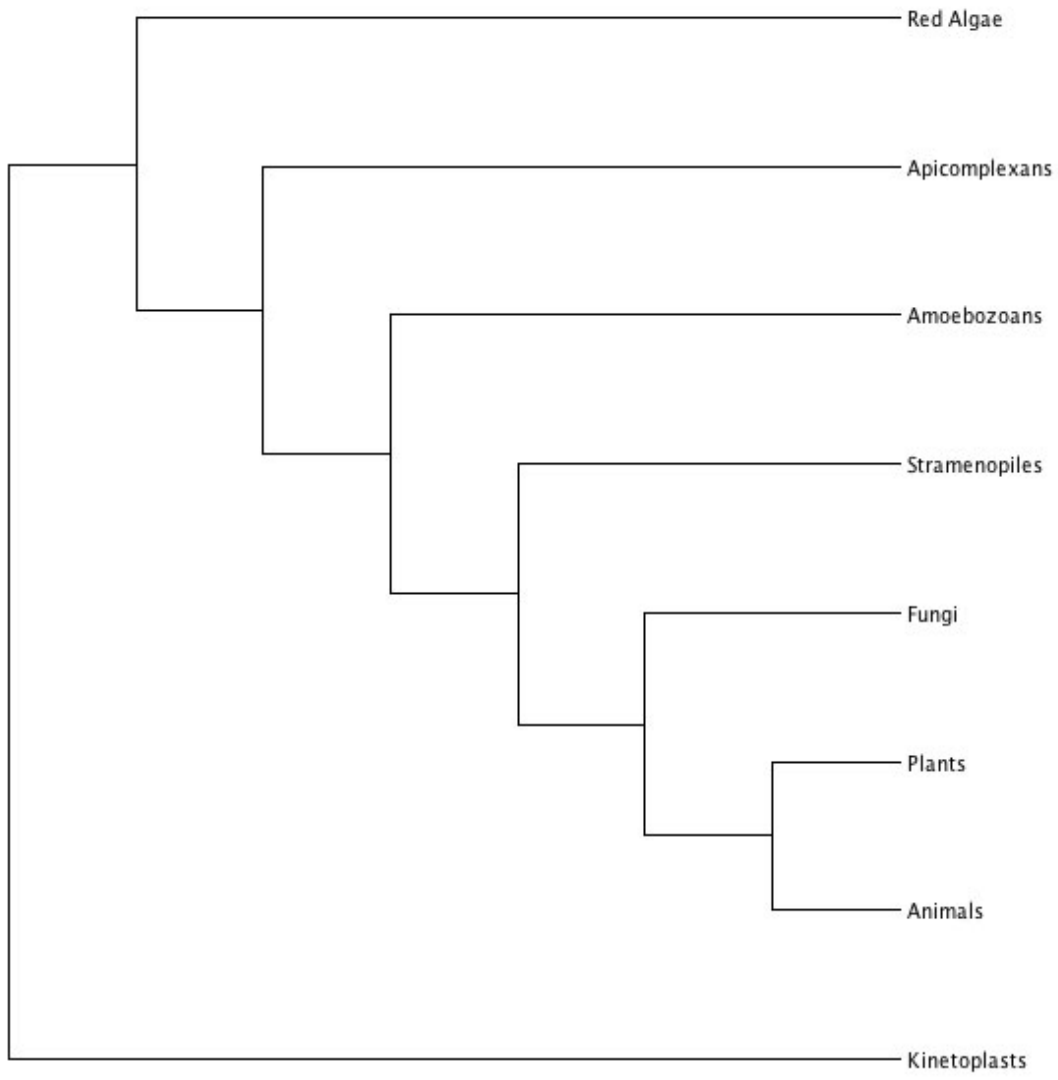
Phylogenetic tree recovered by Bayesian inference on concatenated subunits (RPA1, B1, B2, C1, C2) for the 25 species (*Chalmydomonas* was excluded due to poor RPA2 sequence annotation). The tree was rooted with *Trypanosoma* species.

Figure 20. No intron sliding tree in Dollo



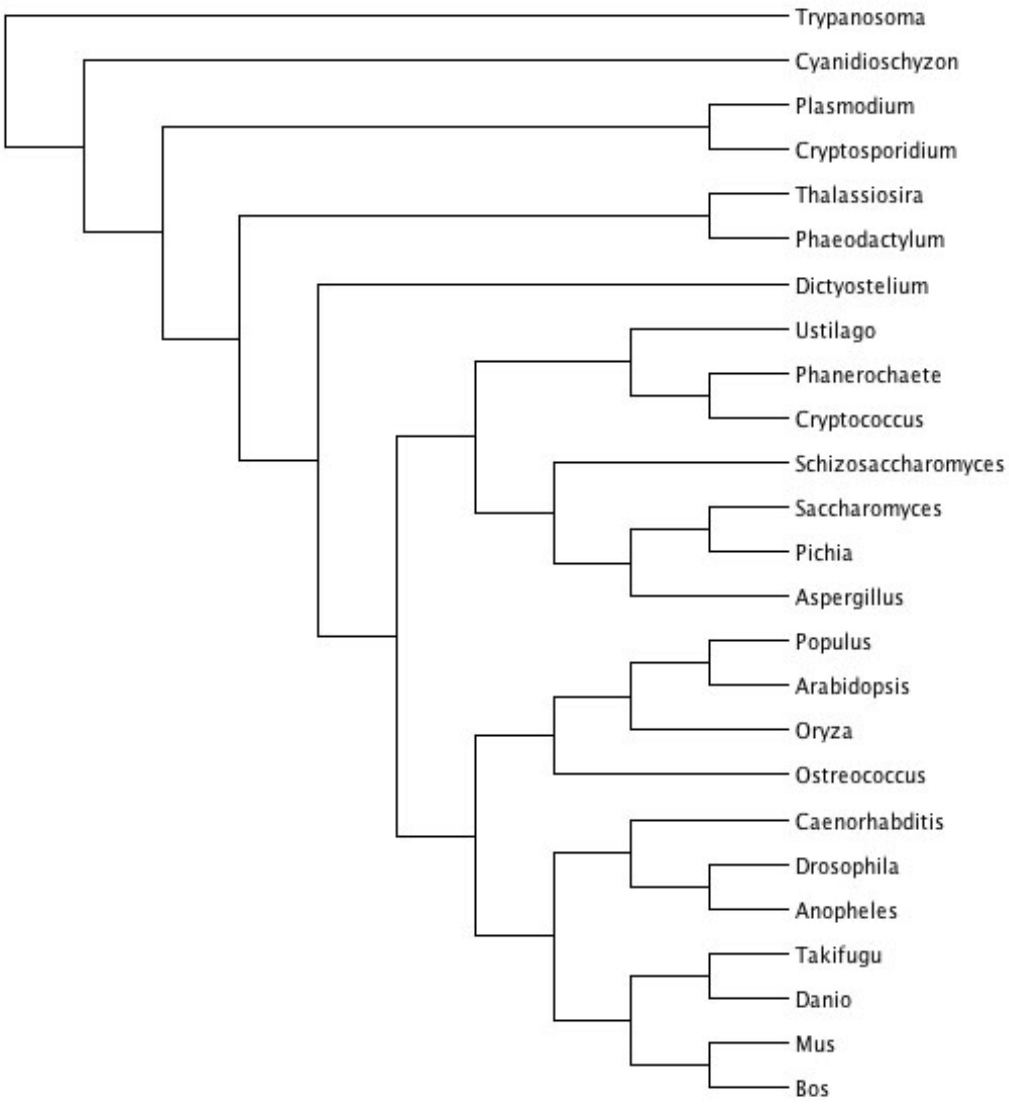
Phylogenetic tree recovered by dollo parsimony using grouped intron position matrix that did not allow intron sliding.

Figure 21. Intron sliding tree in Dollo



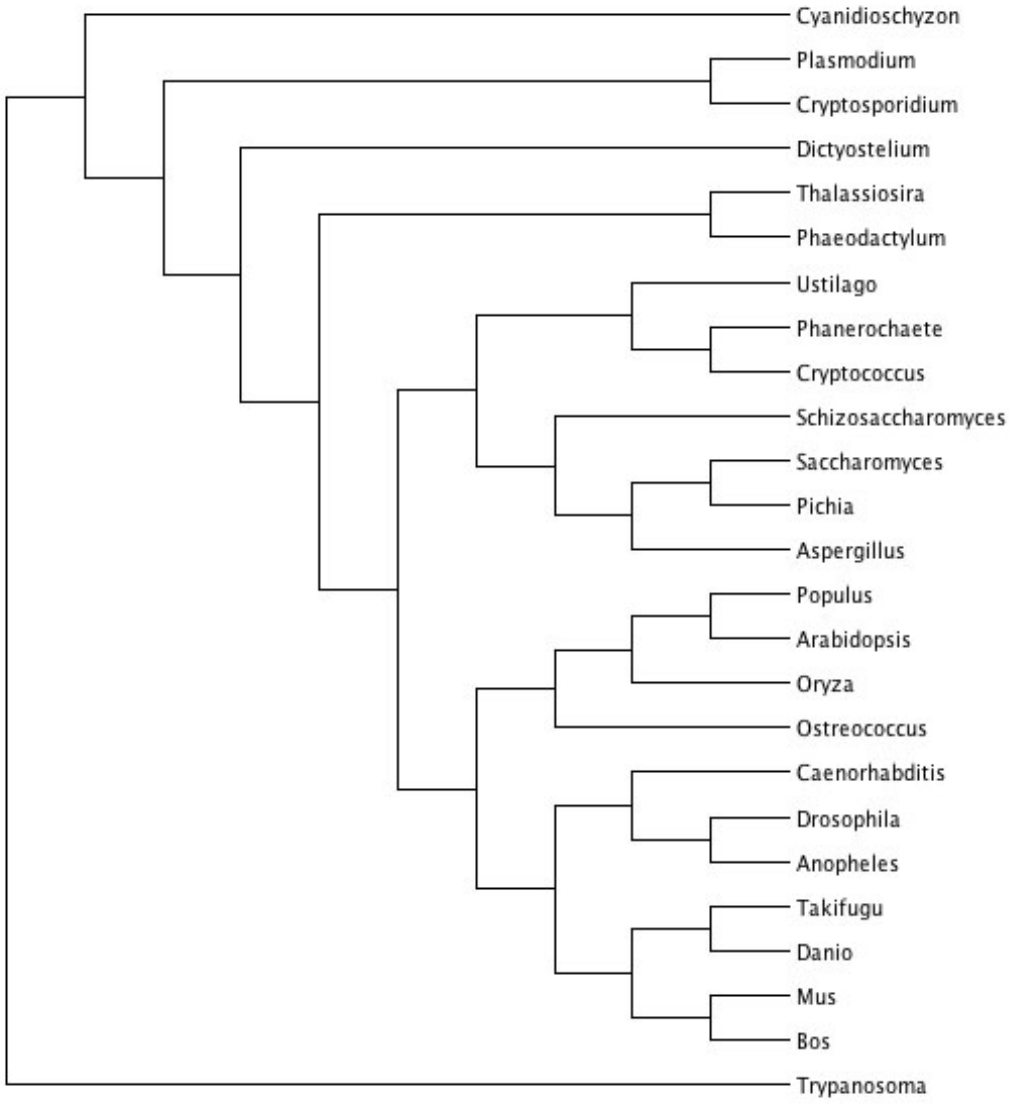
Phylogenetic tree recovered by dollo parsimony using grouped intron position

Figure 22. Sequence-based tree to reflect the no sliding intron tree



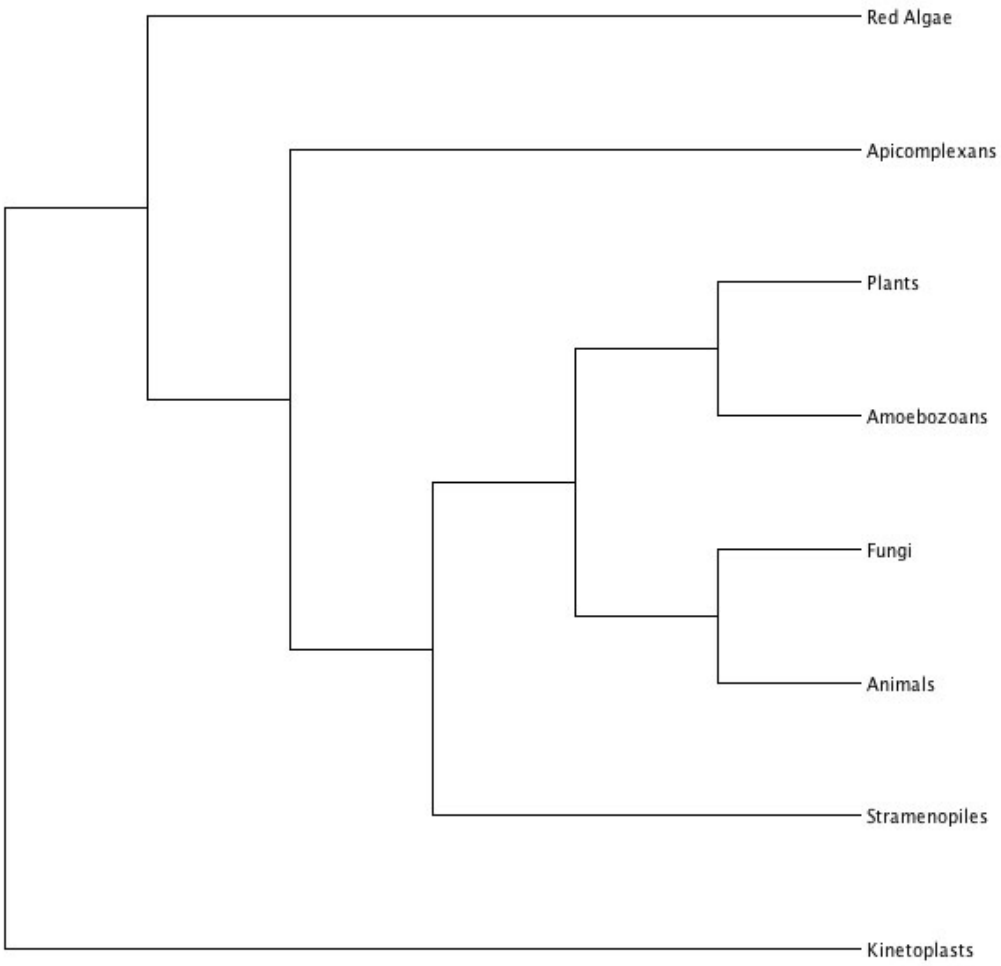
Sequence-based phylogenetic tree modified with retree (Phylip) to reflect the no sliding intron-based phylogenetic trees. The tree was rooted with *Trypanosoma*.

Figure 23. Sequence-based tree to reflect the sliding intron tree



Sequence-based phylogenetic tree modified with retree (Phylip) to reflect the sliding intron-based phylogenetic trees. The tree was rooted with *Trypanosoma*.

Figure 24. Sequence-based tree group tree



Sequence-based phylogenetic tree modified with retree (Phylip) with major groups (red algae, apicomplexans, plants, animals, fungi, amoebozoans, stramenopiles, and kinetoplasts) collapsed for comparison with the intron-based phylogenetic trees. The tree was rooted with the kinetoplasts group.

REFERENCES

Abascal F, Zardoya R, et al. (2005). "ProtTest: selection of best-fit models of protein evolution." Bioinformatics 21(9): 2104-2105.

Baldauf SL and Palmer JD (1993). "Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins." Proc Natl Acad Sci U S A 90(24): 11558-11562.

Bonen L and Vogel J (2001). "The ins and outs of group II introns." Trends Genet 17(6): 322-331.

Carmel L, Rogozin IB, et al. (2007). "Patterns of intron gain and conservation in eukaryotic genes." BMC Evol Biol 7: 192.

Cech TR (1990). "Self-splicing of group I introns." Annu. Rev. Biochem. 59: 543-568.

Do CB, Mahabhashyam MS, et al. (2005). "ProbCons: Probabilistic consistency-based multiple sequence alignment." Genome Res 15(2): 330-340.

Drummond A and Strimmer K (2001). "PAL: an object-oriented programming library for molecular evolution and phylogenetics." Bioinformatics 17(7): 662-663.

Edgar RC (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res 32(5): 1792-1797.

Edgar RC and Batzoglou S (2006). "Multiple sequence alignment." Curr Opin Struct Biol 16(3): 368-373.

Fedorov A, Merican AF, et al. (2002). "Large-scale comparison of intron positions among animal, plant, and fungal genes." Proc Natl Acad Sci U S A 99(25): 16128-16133.

Felsenstein J (2004). PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle, Distributed by the author.

Gilbert W, de Souza SJ, et al. (1997). "Origin of genes." Proc Natl Acad Sci U S A 94(15): 7698-7703.

Gilbert W, Marchionni M, et al. (1986). "On the antiquity of introns." Cell 46(2): 151-153.

Guindon S and Gascuel O (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol 52(5): 696-704.

Guindon S, Lethiec F, et al. (2005). "PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference." Nucleic Acids Res 33(Web Server issue): W557-559.

Harrell, L. E. (2005). Is Glaucocystophyta the protistan ancestor of green plants? Biology. Greenville, East Carolina University. Masters of Biology: 87.

Hasegawa M, Iida Y, et al. (1985). "Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences." J Mol Evol 22(1): 32-38.

Huelsenbeck JP, Ronquist F, et al. (2001). "Bayesian inference of phylogeny and its impact on evolutionary biology." Science 294(5550): 2310-2314.

Kersanach R, Brinkmann H, et al. (1994). "Five identical intron positions in ancient duplicated genes of eubacterial origin." Nature 367(6461): 387-389.

Li W, Tucker AE, et al. (2009). "Extensive, recent intron gains in *Daphnia* populations." Science 326(5957): 1260-1262.

Logsdon JM, J., Stoltzfus A, et al. (1998). "Molecular evolution: recent cases of spliceosomal intron gain?" Curr Biol 8(16): R560-563.

Logsdon JM, J., Tyshenko MG, et al. (1995). "Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory." Proc Natl Acad Sci U S A 92(18): 8507-8511.

Long M, de Souza SJ, et al. (1995). "Evolution of the intron-exon structure of eukaryotic genes." Curr Opin Genet Dev 5(6): 774-778.

Lynch M and Richardson AO (2002). "The evolution of spliceosomal introns." Curr Opin Genet Dev 12(6): 701-710.

Marchionni M and Gilbert W (1986). "The triosephosphate isomerase gene from maize: introns antedate the plant-animal divergence." Cell 46(1): 133-141.

Morgenstern B, Goel S, et al. (2003). "AltAVisT: comparing alternative multiple sequence alignments." Bioinformatics 19(3): 425-426.

Notredame C, Higgins DG, et al. (2000). "T-Coffee: a novel method for fast and accurate multiple sequence alignment." J Mol Biol 302(1): 205-217.

Omilian AR, Scofield DG, et al. (1998). "Intron presence-absence polymorphisms in *Daphnia*." Mol Biol Evol 25(10): 2129-2139.

Philippe H, Snell EA, et al. (2004). "Phylogenomics of eukaryotes: impact of missing data on large alignments." Mol Biol Evol 21(9): 1740-1752.

Rogers JH (1990). "The role of introns in evolution." FEBS Lett 268(2): 339-343.

Rogozin IB, Lyons-Weiler J, et al. (2000). "Intron sliding in conserved gene families." Trends Genet 16(10): 430-432.

Rogozin IB, Wolf YI, et al. (2003). "Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution." Curr Biol 13(17): 1512-1517.

Ronquist F and Huelsenbeck JP (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics 19(12): 1572-1574.

Roy SW (2006). "Intron-rich ancestors." Trends Genet 22(9): 468-471.

Roy SW and Gilbert W (2005). "Complex early genes." Proc Natl Acad Sci U S A 102(6): 1986-1991.

Roy SW and Gilbert W (2005). "Rates of intron loss and gain: implications for early eukaryotic evolution." Proc Natl Acad Sci U S A 102(16): 5773-5778.

Roy SW and Gilbert W (2005). "Resolution of a deep animal divergence by the pattern of intron conservation." Proc Natl Acad Sci U S A 102(12): 4403-4408.

Roy SW and Gilbert W (2006). "The evolution of spliceosomal introns: patterns, puzzles and progress." Nat Rev Genet 7(3): 211-221.

Saldanha R, Mohr G, et al. (1993). "Group I and group II introns." Faseb J 7(1): 15-24.

Shah DM, Hightower RC, et al. (1983). "Genes encoding actin in higher plants: intron positions are highly conserved but the coding sequences are not." J Mol Appl Genet 2(1): 111-126.

Stechmann A and Cavalier-Smith T (2002). "Rooting the eukaryote tree by using a derived gene fusion." Science 297(5578): 89-91.

Stechmann A and Cavalier-Smith T (2003). "The root of the eukaryote tree pinpointed." Curr Biol 13(17): R665-666.

Stoltzfus A, Logsdon JM Jr., et al. (1997). "Intron "sliding" and the diversity of intron positions." Proc Natl Acad Sci U S A 94(20): 10739-10744.

Sverdlov AV, Csuros M, et al. (2007). "A glimpse of a putative pre-intron phase of eukaryotic evolution." Trends Genet 23(3): 105-108.

Sverdlov AV, Rogozin IB, et al. (2005). "Conservation versus parallel gains in intron evolution." Nucleic Acids Res 33(6): 1741-1748.

Swofford DL (1991). PAUP: Phylogenetic Analysis Using Parsimony, version 3.1. Champaign, Illinois, Computer program distributed by the Illinois Natural History Survey.

Wiens JJ (2003). "Missing data, incomplete taxa, and phylogenetic accuracy." Syst Biol 52(4): 528-538.

Wiens JJ (2006). "Missing data and the design of phylogenetic analyses." J Biomed Inform 39(1): 34-42.

WP Maddison and DR Maddison. (2007). "Mesquite: a modular system for evolutionary analysis. Version 2.0 http://mesquiteproject.org."