

**Phylogenomic Analyses of 2,786 Genes in 158 Lineages Support a Root of The Eukaryotic
Tree of Life Between Opisthokonts and All Other Lineages**

Mario A. Cerón-Romero ^{a,b,h}, Miguel M. Fonseca ^c, Leonardo de Oliveira Martins^{d,e}, David
Posada^{d,f,g} and Laura A. Katz^{*a,b}

^a Department of Biological Sciences, Smith College, Northampton, Massachusetts, USA.

^b Program in Organismic and Evolutionary Biology, University of Massachusetts Amherst,
Amherst, Massachusetts, USA.

^c CIIMAR - Interdisciplinary Centre of Marine and Environmental Research, University of
Porto, Porto, Portugal.

^d Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo,
Spain.

^e Quadram Institute Bioscience, Norwich, UK

^f CINBIO, Universidade de Vigo, 36310 Vigo, Spain.

^g Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO.

^h Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana-
Champaign, Illinois.

*Author for Correspondence: Laura A. Katz, Department of Biological Sciences, Smith College,
44 College Lane, Northampton, Massachusetts, 01063, USA, lkatz@smith.edu, 413-585-3825

ABSTRACT

Advances in phylogenomics and high-throughput sequencing have allowed the reconstruction of deep phylogenetic relationships in the evolution of eukaryotes. Yet, the root of the eukaryotic tree of life remains elusive. The most popular hypothesis in textbooks and reviews is a root between Unikonta (Opisthokonta + Amoebozoa) and Bikonta (all other eukaryotes), which emerged from analyses of a single-gene fusion. Subsequent, highly cited studies based on concatenation of genes supported this hypothesis with some variations or proposed a root within Excavata. However, concatenation of genes does not consider phylogenetically-informative events like gene duplications and losses. A recent study using Gene Tree Parsimony (GTP) suggested the root lies between Opisthokonta and all other eukaryotes, but only including 59 taxa and 20 genes. Here we use GTP with a duplication-loss model in a gene-rich and taxon-rich dataset (i.e. 2,786 gene families from two sets of 155 and 158 diverse eukaryotic lineages) to assess the root, and we iterate each analysis 100 times to quantify tree space uncertainty. We also contrasted our results and discarded alternative hypotheses from the literature using GTP and the likelihood-based method SpeciesRax. Our estimates suggest a root between Fungi or Opisthokonta and all other eukaryotes; but based on further analysis of genome size, we propose that the root between Opisthokonta and all other eukaryotes is the most likely.

Keywords: root of eukaryotes, phylogenomics, gene tree - species tree reconciliation, gene tree parsimony, maximum likelihood, gene duplication, gene loss.

Significance statement

Finding the root of the eukaryotic tree of life is critical to understanding the timing and mode of evolution of characters across the evolutionary history of eukaryotes. Yet estimating this root is one of the most challenging questions in evolutionary biology because the age (~1.7 billion years), diversity, and complexity of eukaryotes challenge phylogenomic methods. This study evaluates the root using gene trees and species trees reconciliation instead of the more common approach of analyzing concatenated genes. The dataset used in this study includes both more genes and more diverse species than the datasets of previous studies and the analyses here provide support for a root at or within Opisthokonta (i.e. animals, fungi and their microbial relatives). We explicitly tested alternative hypotheses from the literature, and again found support for an Opisthokonta root, providing a framework for the interpretation of the origin and diversification of eukaryotes and their many unusual features.

INTRODUCTION

One of the most controversial topics in the study of the history of life on Earth is the location of the root of the eukaryotic tree of life (EToL), which likely dates to around 1.6-1.8 billion years ago (de Duve 2007; Parfrey et al. 2011). While there has been substantial progress in defining major eukaryotic clades, such as Archaeplastida, Opisthokonta, SAR, and Amoebozoa (Rodríguez-Ezpeleta et al. 2005; Steenkamp et al. 2006; Burki et al. 2007; Hampl et al. 2009; Adl et al. 2012; Jackson and Reyes-Prieto 2014; Cavalier-Smith et al. 2015; Katz and Grant 2015), the location of the root of EToL remains elusive. Initial molecular studies suggested a root in between amitochondriate eukaryotes, such as Microsporidia and Metamonada, based on early evidence that the acquisition of mitochondria was a derived character in the eukaryotic evolution (Cavalier-Smith 1987; Cavalier-Smith 1991; Cavalier-Smith 1993; Baldauf et al. 1996; Hilario and Gogarten 1998). These hypotheses were abandoned after evidence that these amitochondriate eukaryotes lost their mitochondria more recently (Keeling 1998; Roger 1999). Later, another hypothesis was proposed placing the root between the clades Unikonta (Opisthokonta + Amoebozoa) and Bikonta based on the presence of two gene fusions (Stechmann and Cavalier-Smith 2003). The Unikonta-Bikonta hypothesis and its derivatives (see below) remain highly referenced, but rapid changes in eukaryotic taxonomy, phylogenomic methods, and data availability have opened the door for alternative hypotheses that have instigated further research on the root of EToL.

Many recent studies using concatenated genes (i.e. supermatrix) and more inclusive datasets, especially including underrepresented clades of microeukaryotes, overall agree with the Unikonta-Bikonta root but require a series of adjustments on both sides of the tree (Derelle and Lang 2012; Derelle et al. 2015; Brown et al. 2018). Initially, the Unikonta clade contained

Opisthokonta and Amoebozoa, while Bikonta contained the rest of the eukaryotes (Stechmann and Cavalier-Smith 2002; Stechmann and Cavalier-Smith 2003). Later, based on a supermatrix analysis of mitochondrial proteins, a new clade including Unikonta and former bikont lineages (i.e. Apusozoa, Breviata) was defined as Amorphea (Adl et al. 2012; Derelle and Lang 2012), with the root dividing Amorphea and the remaining eukaryotes. A subsequent phylogenomic analysis with an extended dataset of mitochondrial and other bacterial-origin proteins restructured Amorphea as Opimoda, which includes malawimonads and colodictyonids, and classified the rest of the eukaryotes as Diphoda (Derelle et al. 2015). Finally, Ancyromonads were proposed as an early branch on either the Opimoda or the Diphoda side of the tree (Brown et al. 2018). In contrast, hypotheses from studies using alternative approaches to supermatrix deviate substantially from the original Unikonta-Bikonta root (Martin et al. 2003; Rogozin et al. 2009; Wideman et al. 2013).

Due to its tractability, the supermatrix approach for species tree reconstruction has been very popular in studies attempting to find the root of the EToL (Baptiste et al. 2002; Derelle and Lang 2012; He et al. 2014; Derelle et al. 2015; Brown et al. 2018). Although this approach offers a good resolution when there is not much discordance among the evolutionary histories of the concatenated markers, there are aspects that need consideration when deploying it on highly diverse datasets. The supermatrix approach requires a critical step of distinguishing orthologous sequences from paralogous sequences, a difficult task when the evolutionary scale is >1 billion years of eukaryotic evolution (Vallender 2009; Tekaia 2016; Glover et al. 2019). The supermatrix approach also requires the challenging step of choosing the correct set of markers, which often ends up restricting the analysis to reduced datasets with confusing phylogenetic signals. For instance, Derelle and Lang (2012) and He et al. (2014) proposed a different root of EToL despite using datasets with similar characteristics: about 40 genes of mitochondrial origin that allow the

use of a bacterial outgroup. The different results from these two studies were attributed to the use of different criteria to identify the mitochondrial genes (Williams 2014), different approaches for ortholog calling, different models of protein evolution (Derelle et al. 2015), and the effect of missing data and poorly sampled lineages on the models of protein evolution (Al Jewari and Baldauf 2022).

Phylogenomic methods referred to as 'tree-aware methods' offer an alternative that fixes the main problems of the supermatrix, including the need to carefully choose an appropriate set of markers and to correctly identify orthologs. These methods produce a species tree that best represents a set of gene trees based on an optimization criterion (Mallo and Posada 2016). Some tree-aware methods use an optimization criterion based on biological events: duplications, losses, incomplete lineage sorting (ILS), and gene transfers (Wehe et al. 2008; Chaudhary et al. 2010; Mirarab and Warnow 2015; Bayzid and Warnow 2018). Other methods use phylogenetic distances, their posterior probabilities, or quartet similarity scores (Chaudhary et al. 2015; De Oliveira Martins et al. 2016; Molloy and Warnow 2020; Zhang et al. 2020). A popular approach in tree-aware methods is gene tree parsimony (GTP), which infers the species tree that requires the lowest number of events according to their optimization criterion (Wehe et al. 2008; Chaudhary et al. 2010). More recently, additional methods have improved different aspects such as the computational speed (Mirarab and Warnow 2015; Molloy and Warnow 2020; Zhang et al. 2020) or the inclusion of models of evolution through a parametric framework (Boussau et al. 2013; De Oliveira Martins et al. 2016; Morel et al. 2022).

The suitability of the tree-aware methods to study the root of the EToL relies on their optimization criteria, inputs, outputs, and computing requirements. For instance, methods with optimization criteria based on ILS (e.g. Mirarab and Warnow 2015; Vachaspati and Warnow

2015) are not ideal because ILS has a negligible effect on highly divergent datasets (Maddison and Knowles 2006). Also, many methods are not applicable because they require rooted gene trees and/or generate unrooted species trees (e.g. Chaudhary et al. 2015; Bayzid and Warnow 2018; Molloy and Warnow 2020; Zhang et al. 2020; Willson et al. 2022). In contrast, GTP methods, such as those included in the tool iGTP (Wehe et al. 2008; Chaudhary et al. 2010), and the likelihood-based software SpeciesRax (Morel et al. 2022) produce a rooted species tree from unrooted gene trees using optimization criteria that include duplications and losses, which are critical in eukaryotic evolution (Wolfe and Shields 1997; Otto and Whitton 2000; Dehal and Boore 2005). Other parametric tools offer similar characteristics but they required significantly more computational resources (e.g. Boussau et al. 2013; De Oliveira Martins et al. 2016; Morel et al. 2020)

Despite the growing advances and interest in tree-aware methods to reconstruct species trees in the last couple of decades, we are only aware of one study implementing this type of approach to estimate a rooted ETOL: based on only 20 gene trees, an analysis using iGTP estimated a root between Opisthokonta and the rest of eukaryotes (Katz et al. 2012), which is consistent with gene-fusion analyses (Stechmann and Cavalier-Smith 2002). Here, we leverage a much larger dataset of 2,786 gene families including up to 158 species distributed across the whole ETOL, gathered with our phylogenomic pipeline PhyloToL (Cerón-Romero et al. 2019). We pay particular attention to filtering out contamination and possible lateral gene transfers, both common in microeukaryotes, and we apply a robust processing of multiple sequence alignments before gene tree reconstruction. Then, we deploy iGTP with a duplications-losses criterion to find the root of ETOL and compared the resulting root against previously published rootings using both iGTP and SpeciesRax.

RESULTS

Building the phylogenomic datasets

We used the database of PhyloToL, which contains more than 13,000 gene families and 1,007 taxa (i.e. including Archaea, Bacteria and Eukaryotes; Grant and Katz 2014; Cerón-Romero et al. 2019) to select the gene families for this study. Initially, we filtered gene families that were present in at least 25 taxa of at least 4 major eukaryotic clades (i.e. Opisthokonta, Amoebozoa, Archaeplastida, Excavata, and SAR; Table 1; Dataset S1). Then, we built alignments and phylogenetic trees to select the 2,786 gene families that were only found in eukaryotes or in which eukaryotes are monophyletic (see Materials and methods; Dataset S2). By this point, we expected to have the most conserved gene families and have already filtered out interdomain LGT, including LGT from chloroplast and mitochondria.

In order to balance phylogenetic diversity and computation speed, we built four datasets that each included the 2,786 gene families and 153 to 158 eukaryotic species from between 140 to 158 genera (Table 1; Datasets S1 and S2). The four datasets varied based on taxon selection criteria: for the 'SEL+' (i.e. selected) dataset, we selected representative species of all major eukaryotic clades based on our assessment of data quality and taxonomic breadth; and for the 'RAN+' (i.e. random) dataset, we randomly chose even numbers of species among the major eukaryotic clades. We also generated two additional databases by excluding the fast-evolving Microsporidia (i.e. SEL- and RAN-) as the inclusion of these lineages can generate phylogenetic artifacts, such as long-branch attraction (Embley and Hirt 1998; Hirt et al. 1999; Van de Peer et al. 2000).

Inference on the location of the EToL root

Though we set out to deploy two summary methods to infer the root of the eukaryotic tree of life, due to the complexity of the data, we were constrained to focus on only one method for the analyses presented here. Our original intent was to use both a Bayesian supertree approach with the software *guenomu* (De Oliveira Martins et al. 2016) and a gene tree parsimony approach with the software package iGTP (Chaudhary et al. 2010). Unfortunately, *guenomu* failed to converge in an estimate of species trees after being run for multiple weeks on a cluster with more than 400 cores, likely due to the complexity of the algorithm and underlying uncertainty in the gene trees, so we continued only with iGTP.

Using iGTP with a duplication-loss criterion, we estimated the most parsimonious rooted tree of eukaryotes for each of our four datasets, all of which indicated Fungi as the earliest branching group (Figure 1, Dataset S3). Other less parsimonious but frequent alternatives indicate glaucophytes or the taxon *Fabomonas tropica* as the earliest branching group or taxon. Across all replicates of the analysis, the second most frequent earliest branching group was Opisthokonta (i.e. the remaining opisthokonts when the earliest branching group was Fungi). These results leave open the possibility of a root between Opisthokonta and the other eukaryotes, which we discuss below.

Comparison to published EToL hypotheses

We also used iGTP to evaluate various hypotheses from the literature including a root between Opisthokonta and others (Stechmann and Cavalier-Smith 2002; Katz et al. 2012), between Discoba (Excavata) and others (He et al. 2014), and the Unikonta – Bikonta root (Stechmann and Cavalier-Smith 2003). Additionally, we included an alternative root with

Ancyromonadida + Metamonada as sister to all other eukaryotes (personal communication Tom Williams, Celine Petitjean), which emerged from studies of probabilistic gene tree-species tree reconciliation analyses with amalgamated likelihood estimation (ALE; Szöllősi et al. 2013; Szöllősi et al. 2013 b). For the Unikonta-Bikonta root, we chose the one with the lowest reconciliation cost in a preliminary analysis (see Material and methods; Dataset S4) after comparing all the derived hypotheses under the same umbrella (Stechmann and Cavalier-Smith 2003; Derelle and Lang 2012; Derelle et al. 2015; Brown et al. 2018). Here, iGTP estimates the reconciliation cost of a species tree given constrained phylogenetic relations among major eukaryotic clades to reflect the different hypotheses of the root of ETOL (Figure S1). In addition to these four hypotheses, we also calculated and compared the reconciliation cost of a species tree reflecting our initial estimate, placing the root between Fungi and the other eukaryotes. The results show that while Opisthokonta–others is the most parsimonious root followed by Fungi–others for SEL+ and RAN+, the opposite is true for SEL- and RAN- (Figure 2, Table S1).

We assessed the difference in reconciliation costs between Fungi–others and every other hypothesis in all four datasets. We determined that reconciliation cost values were not normally distributed based on Shapiro-Wilk tests ($n = 100$, $p > 0.05$; Table S1). Then, we performed Wilcoxon rank sum tests to determine if there was a statistically significant difference between the median reconciliation costs of the Fungi–others hypothesis against every other hypothesis. Our results show that for datasets SEL+ and RAN+ there are no significant statistical differences between the median reconciliation costs of Opisthokonta–others and Fungi–others ($n = 100$, $p > 0.05$; Tables S1 and S2). For all four taxon sets, the median estimated reconciliation costs for species trees inferred to match the remaining published hypotheses were all statistically significantly higher (i.e. less parsimonious) than our rooted species tree (Fungi–others, t-student, $p < 0.05$; Figure 2 and Tables S1 and S2).

We confirmed our results using SpeciesRax (Morel et al. 2022), a method that uses maximum likelihood to calculate the probability of observing a set of gene trees given a rooted species tree and a model based on a reconciliation criterion. We compare the reconciliation likelihood of every root hypothesis per dataset using as input the most parsimonious species trees with their underlying gene trees from the iGTP analysis. Also, given our data curation, which sought to remove the effect of LGTs, we decided to use the duplication-loss model instead of the default duplication-transfer-loss model. Our results with SpeciesRax were consistent with those of iGTP (Figure 2) having Fungi–others and Opisthokonta–others as the most likely roots in every dataset (Figure S2).

Assessing the effect of missing data in Fungi over root calculations

We tested if missing data in Fungi, due to reduced genomes (Figure S3), artifactually contribute to the most parsimonious root between Fungi/Opisthokonta and the rest of eukaryotes. We ran iGTP in two subsets from the SEL+ dataset: 1) 336 genes that contain at least 10 metazoan and 10 fungi species and 2) 246 genes that contain at least 10 metazoans and no fungi. For the first subset, the most parsimonious root was between the taxon *Fabomonas tropica* (an “orphan” taxon with substantial levels of missing data; Figures S3 and S4) and the others, followed by the root Fungi–others (Figure S5). In both topologies, the next earliest divergent group was *other* Opisthokonta. For the small set of genes present only in 10 or more metazoan and no fungi, a root Opisthokonta–others still appeared as one of the most parsimonious roots (Figure S6).

DISCUSSION

This study, which represents the most taxon-inclusive analysis yet to address the root of the eukaryotic tree of life, analyzed 2,786 gene trees for four taxon sets of up to 158 diverse eukaryotes, with each analysis iterated 100 times by changing both gene tree order and root. As in Katz et al. (2012), we used gene tree parsimony as implemented in the software iGTP to estimate the root of EToL that minimizes gene duplications and losses. The use of a tree-aware method to find the root of EToL offers an alternative to the use of supermatrix methods (Mallo and Posada 2016), most notably to take advantage of the wealth of sequencing data and to avoid the ortholog calling step that can be challenging for such a highly diverse taxon dataset (Vallender 2009; Tekaia 2016; Glover et al. 2019). Moreover, given the importance of gene duplications and losses for the evolution of eukaryotic genomes (Wolfe and Shields 1997; Otto and Whitton 2000; Dehal and Boore 2005), their inclusion in the estimation of the most likely root of EToL also represents an advance over previous studies. Tree-aware methods can also account for LGT and ILS, but such events should be less relevant for our datasets because of our pre-processing of the data (see methods) and level of divergence (Maddison and Knowles 2006), respectively.

Across our analyses, we found support for either Fungi or Opisthokonta as sister to all other eukaryotes, and also that previously published hypotheses were significantly less parsimonious (by iGTP) and less likely (by SpeciesRax; Figures 1, 2, and S2; Table S2). Martin et al. (2003) argued for a Fungi + others root based on the fact that fungi have osmotrophic feeding and most other eukaryotes are phagotrophic (with the exceptions including autotrophic lineages; Martin et al. 2003). Early evidence for this hypothesis comes from pre-Ediacaran fossils that look similar to fungi (Butterfield 2005; Butterfield 2009; Loron et al. 2019), which can

be twice as old as fossils used for the current estimates of the origin of fungi (450 Ma; Redecker et al. 2000). However, given the overwhelming evidence from molecular data that Opisthokonta is monophyletic (Baldauf and Palmer 1993; Burki et al. 2007; Katz and Grant 2015; Brown et al. 2018), the “fungi first” hypothesis seems unlikely. In contrast, the Opisthokonta root is consistent with a previous study using iGTP with a significantly smaller dataset (Katz et al. 2012) and initial analyses of a gene fusions (Stechmann and Cavalier-Smith 2002). Under this scenario, the placement of Fungi at the root can be explained as a phylogenetic artifact, which according to our results, is more likely associated with the nature of the data than with the method.

We propose that reductions in genome size and subsequent gene loss within Fungi contribute to a spurious placement of Fungi at the root. Multiple studies have shown that gene loss is a pervasive factor of evolution in both Fungi (Braun et al. 2000; Nagy et al. 2014; Stajich 2017) and other Opisthokonta (Albalat and Cañestro 2016; Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020). Moreover, the significantly smaller fungi genomes, as compared to the metazoan genomes (Figure S3), suggest that gene loss is much more intense in Fungi. Although iGTP and SpeciesRax count on gene loss events for their score estimates, we believe that the striking differences in genome size between Fungi and other Opisthokonta affects such calculations. Several studies indicate that interdomain LGTs are frequent in fungi (Rosewich and Kistler 2000; Wenzl et al. 2005; Lawrence et al. 2011), and such genes might also contribute to pulling Fungi to the root of the eukaryotic tree of life. However, given our data curation and procedures to remove the effect of LGTs (see methods), we do not expect LGT to play any major role in splitting Fungi and the other Opisthokonta in our estimates of the root.

Missing data is another well-known factor that affects phylogenetic methods and previous studies have shown its negative effect on gene tree parsimony approaches for species tree

inference (Burleigh et al. 2011; Davis et al. 2019). If iGTP cannot distinguish between gene loss and missing data, we would expect clades with significant levels of missing data to be placed at the root, a similar scenario to the one that we propose for Fungi with their relatively small genomes. For instance, a root Glaucophyta–others (i.e. Glaucophyta–(Opisthokonta + others)) appears as one of the four most parsimonious (though always less parsimonious than Fungi–others) across taxon sets in our results (Figure 1). Besides this being a root with no support from the literature, it also implies that Archaeplastida is not monophyletic. Although analyses with molecular data have shown mixed results about the monophyly of Archaeplastida (Katz and Grant 2015; Cenci et al. 2018; Leebens-Mack et al. 2019; Price et al. 2019; Strasser et al. 2019), this result may be due to missing data; despite our efforts to choose genes with well-represented species, the glaucophytes are the minor clade with the fewest data across gene trees (Figure S4) likely due to incomplete sequencing of transcriptomes. Given that the rest of the eukaryotic clades are better represented in our datasets, except for some “orphan” taxa such as *Fabomonas tropica* or *Glaucocystis nostochinearum*, missing data is not expected to influence the major results of this study. In fact, Opisthokonta and Fungi, the clades that are consistently placed at the root of ETOL by our analyses, are mostly represented by whole genomes in our taxon datasets (more than 80% and 90%, respectively; Dataset S1).

Since a large proportion of our Opisthokonta sample is made of Fungi (i.e. 37%), it could also be argued that high rates of gene loss in Fungi promote an artifactual placement of the root between Opisthokonta and the other eukaryotes. We tested for this effect in two analyses: 1) 336 genes present in 10 or more species of both fungi and metazoan, and 2) 246 genes present in more than 10 metazoans but absent from fungi. The significant reduction of the datasets and power in both tests in comparison to the analyses of all 2,786 genes gives opportunities for spurious results, namely, “orphan” single taxa being placed at the root. Nevertheless, the

retention of a Fungi/Opisthokonta root and the absence of all other previously published roots (e.g. Unikonta-Bikonta) among the most parsimonious results in both analyses (Figures S4 and S5) suggests that high rates of gene loss in fungi do not determine the major findings of the fuller analyses.

Recent studies on the root of ETOL have focused on a conflict between the Unikonta-Bikonta root (the Derelle et al. (2015) variant, Opimoda- Diphoda) and the Discoba–others root (Derelle and Lang 2012; He et al. 2014; Derelle et al. 2015; Al Jewari and Baldauf 2022), which were formulated using similar data and methods despite being completely different roots (Derelle and Lang 2012; He et al. 2014; Derelle et al. 2015). Those studies were based on supermatrix approaches on around 40 mitochondrial genes, and a recent study shows that sensitivity to poorly sampled lineages is a major factor in explaining the discrepancies between them (Al Jewari and Baldauf 2022). Given that the studies that have been mostly consistent with the Unikonta-Bikonta root used similar approaches and involved new orphan taxa (Derelle and Lang 2012; Derelle et al. 2015; Brown et al. 2018), it is not surprising that they ended up proposing changes to both sides of the tree by placing the orphan taxa very close to the root. The lack of consistency of hypotheses generated using supermatrix approaches is reflected in our results, where the Opisthokonta–others root performed is better supported, with the Opimoda- Diphoda root not even passing our preliminary analysis comparing among the different variants of the Unikonta-Bikonta root (Figure S1, Dataset S4).

In conclusion, our estimates of the root of the eukaryotic tree of life based on 2,786 genes support a root in or between Opisthokonta and the rest of the eukaryotes (i.e. Opisthokonta–others). We show that these results are consistent across datasets, and none of the most referenced published hypotheses are more parsimonious or more likely. There are

caveats to be considered, but they do not seem to affect the major findings of the study. For instance, our results consistently point to a root between either Fungi or Opisthokonta and the rest of the eukaryotes. Based on comparisons of genome size and the overwhelming evidence on the monophyly of Opisthokonta, we argue that Opisthokonta–others is the correct root and that Fungi–others is an artifact given the data (eg. small genome sizes among fungi), but further studies are needed to resolve such issues. A possible step to tackle this would be to replicate our analysis but using DupTree, a GTP tool that only considers duplications (no losses). Also, although we sought to remove genes that included interdomain LGTs (see Material and methods), it is possible that we missed a few cases and interdomain LGT could have had a minor contribution to pulling fungi to the root. Another alternative is to try SpeciesRax or GeneRax with the duplication-loss-transfer model. Though, in our experience, such analyses might not finish in a reasonable time with datasets as large and complex as ours. Finally, we acknowledge that missing data likely affect our estimates, as we argue that the spurious position of a root between glaucophytes and other eukaryotes is due to a lack of genome data. Future studies with more complete genome data are required to validate the estimates presented here.

MATERIALS AND METHODS

Taxa selection

We started with the database of our phylogenomic pipeline PhyloToL (Grant and Katz 2014; Cerón-Romero et al. 2019), which contains 1,007 taxa including Bacteria, Archaea, and Eukaryotes. From this database, we generated four subsets of 158, 155, 155, and 153 eukaryotes under three different criteria: 1) selecting taxa based on the quality of the data and maximizing the diversity based on their taxonomy (SEL+), 2) selecting taxa randomly among the major eukaryotic clades Opisthokonta, Amoebozoa, Archaeplastida, Excavata, SAR and some orphan lineages (RAN+), and 3) The same taxa as in datasets SEL+ and RAN+ but without microsporidians (SEL- and RAN-; Table 1 and Dataset S1). For the last two datasets, we excluded microsporidians in order to avoid long-branch attraction due to microsporidians fast-evolutionary rates.

Gene family selection

PhyloToL contains more than 13,000 gene families that were chosen for their representation in diverse eukaryotes. For this study, we focused on gene families that contain at least 25 taxa representing at least four of the five major eukaryotic clades. Additionally, at least two of the major clades had to contain at least two 'minor clades' (e.g. we consider Glaucophyta and Rhodophyta as minor clades in the major clade Archaeplastida). In a pilot analysis, we produced an alignment and a phylogenetic tree for each gene family using the default settings of a previous version of PhyloToL (GUIDANCE V1.3.1 sequence cutoff = 0.3 and column cutoff = 0.4; RAxML quick tree with model PROTGAMMALG and no bootstraps; Stamatakis 2006; Penn et al. 2010; Grant and Katz 2014). Then, we kept the gene families that were exclusive of

eukaryotes or the ones in which eukaryotes were monophyletic. From a total of 3,002 gene families that met our criteria, 2786 passed the initial steps of PhyloToL when including only the data from the dataset SEL+. These 2,786 gene families were used for further analyses with all datasets (Dataset S2).

Multiple Sequence Alignments (MSA) and gene tree inference

MSAs for the four datasets were produced with PhyloToL (GUIDANCE V2.02 sequence cutoff = 0.3, column cutoff = 0.4, number of iterations = 5; Sela et al. 2015; Cerón-Romero et al. 2019). The default parameters of PhyloToL include up to five iterations of GUIDANCE V2.02 with 10 bootstraps and MAFFT V7 (Kato and Standley 2013) with algorithm E-INS-i for less than 200 sequences or “auto” option if more than 200 sequences, and maxiterate = 1000. Instead, here we run up to five iterations of GUIDANCE with 20 bootstraps and the simple MAFFT algorithm FFT-NS-2. Then, we performed a GUIDANCE run with 100 bootstraps and the default MAFFT parameters for PhyloToL.

Gene trees were inferred with RAxML v.8.2.4 (Stamatakis 2014) with 10 ML searches for best-ML tree (option “-# 10”), using the rapid hill-climbing algorithm (option “-f d”) and no bootstrap replicates. The protein evolution model used was evaluated during the gene tree inference (option “-m PROTCATAUTO”) by testing all models available in RAxML (e.g. JTT, LG, WAG, etc) with optimization of substitution rates and of site-specific evolutionary rates which were categorized into four distinct rate categories for computational efficiency.

Inference of rooted species trees

To infer a rooted ETOL, we used two summary methods/tools for species tree inference: the Bayesian-based *guenomu* and the gene tree parsimony tool iGTP. While iGTP considers

that the disagreement between gene trees and the species tree is due to either duplications, duplications-losses, or deep coalescence; *guenomu* considers the effect of these and other evolutionary processes in a multivariate manner. With *guenomu* we did not see convergence in two independent replicates in a reasonable time, which may reflect a lack of convergence of MrBayes or underlying uncertainty in the gene trees; therefore, we chose to continue further analyses with iGTP only, which relies on point estimates of the gene phylogenies.

We ran iGTP for the four datasets with the analysis option that accounts for gene duplications and losses. In our application of iGTP, we decided to iterate each analysis 100 times to explore the tree space. Given the complexity of the datasets and the heuristic nature of some key steps of the iGTP algorithm (e.g. gene tree rooting and starting species tree generation), we faced two systematic challenges in a preliminary analysis with iGTP as the inferred species tree was affected by: 1) the order of the leaves in the input unrooted gene tree Newick strings (i.e. the input trees were treated as rooted even though we specified that they were not) and 2) the input gene order in the 100 replicates. Therefore, we randomly shuffled the order of the leaves in the unrooted gene trees (keeping the same topology) and randomly shuffled the order of the input gene trees in each of the 100 replicates per dataset.

Comparing different EToL root hypotheses

For the datasets SEL+, RAN+, SEL- and RAN-, we compared 5 different hypotheses of the root of EToL. These hypotheses are: 1) the most parsimonious root according to the iGTP analysis (i.e. Fungi–others), 2) between Opisthokonta and the rest of eukaryotes, 3) between Discoba (Excavata) and rest of the eukaryotes, 4) between Unikonta and Bikonta, and 5) between Metamonada (Excavata) + Ancyromonadida and the rest of eukaryotes. For the Unikonta-Bikonta root, different alternative topologies according to the multiple changes in the

definition of both clades were evaluated using the dataset SEL-, but only the one with the lowest reconciliation cost was used for further comparisons (Figure S1, Dataset S4). In order to compare the hypotheses, we constrained species trees (fixing the relationships among major clades and allowing the relationships within minor clades to be inferred by iGTP) according to every hypothesis and calculated the reconciliation cost per hypothesis in each dataset.

In order to test if our results were robust to different methods, we also compared the root hypotheses using the likelihood-based tool SpeciesRax. Initially, we set up to run GeneRax (--strategy EVAL --si-strategy SKIP) instead of SpeciesRax to optimize the gene trees given the alignments and calculate a joint likelihood. However, after a week of running time in a computer with 16 cores and 32 GB of RAM only 10% of the gene families had been optimized for the first dataset. Therefore, we opted for SpeciesRax (--strategy SKIP --si-strategy EVAL). SpeciesRax takes the best iGTP constrained species trees per hypothesis and their underlying gene trees to calculate the reconciliation likelihood. Since we removed LGT and contamination from our dataset using a series of filters, we applied the model UndatedDL instead of UndatedDTL, which takes into consideration duplications and losses and ignores potential gene transfers.

Computational resources

The production of alignments following the strategy described above for each of the four datasets required 10 weeks of running time (around 40 weeks in total) in 75 threads and around 120 GB of RAM. The gene tree inference for each dataset required around 4 weeks (around 16 weeks in total) in 24 threads and 24 GB of RAM. Each iGTP analysis (with 100 replicates) requires 1 week of running time in 100 threads and ~100 GB of RAM. Given that there were six iGTP analyses per dataset, the running time for all datasets was around 24 weeks.

Quantification and statistical analysis

We iterated each iGTP analysis 100 times to quantify tree space uncertainty. The reconciliation costs of each root hypothesis were non-normally distributed (Shapiro–Wilk, $n = 100$, $p > 0.05$; Table S1). Therefore, we compared the median reconciliation cost of Fungi–others (i.e. other eukaryotes) against the one of every other hypothesis in all four datasets using Wilcoxon rank-sum tests. The results of these tests were summarized in Tables S1 and S2 and displayed in figure 2.

ACKNOWLEDGEMENTS

We thank members of the Katz lab for comments on earlier versions of the manuscript. This work was supported by a National Institute of Health award (R15HG010409) and three awards from the National Science Foundation (OCE-1924570, DEB-1651908 and DEB-1541511) to L.A.K.

AUTHOR CONTRIBUTIONS

LA.K and M.A.C.R conceived of the study and broad approach, and designed the experiments in collaboration with M.M.F and D.P; M.A.C.R, M.M.F and L.O.M performed the analyses; M.A.C.R and L.A. K wrote the manuscript with input from all authors.

DATA ARCHIVING

Raw and analyzed data are deposited in the DRYAD database

(<https://datadryad.org/stash/share/9jrGM0UhndWT0tYmoawBnwfkpB-cuNfnQwL9fPuxBiU>)

DECLARATION OF INTERESTS

None

REFERENCES

- Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, et al. 2012. The revised classification of eukaryotes. *J Eukaryot Microbiol* 59:429–493.
- Al Jewari C, Baldauf SL. 2022. Conflict over the eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. *Systematic Biology*:syac029.
- Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet* 17:379–391.
- Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* 90:11558–11562.
- Baldauf SL, Palmer JD, Doolittle WF. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences* 93:7749–7754.
- Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller M, et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proceedings of the National Academy of Sciences* 99:1414–1419.
- Bayzid MS, Warnow T. 2018. Gene tree parsimony for incomplete gene trees: addressing true biological loss. *Algorithms for Molecular Biology* 13:1.
- Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23:323–330.
- Braun EL, Halpern AL, Nelson MA, Natvig DO. 2000. Large-Scale Comparison of Fungal Sequence Information: Mechanisms of Innovation in Neurospora crassa and Gene Loss in Saccharomyces cerevisiae. *Genome Res.* 10:416–430.
- Brown MW, Heiss AA, Kamikawa R, Inagaki Y, Yabuki A, Tice AK, Shiratori T, Ishida K-I, Hashimoto T, Simpson AGB, et al. 2018. Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution* 10:427–433.
- Burki F, Shalchian-Tabrizi K, Minge M, Skjæveland Å, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics Reshuffles the Eukaryotic Supergroups. *PLOS ONE* 2:e790.
- Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. 2011. Genome-Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896 Gene Trees. *Systematic Biology* 60:117–125.
- Butterfield NJ. 2005. Probable Proterozoic fungi. *Paleobiology* 31:165–182.

- Butterfield NJ. 2009. Modes of pre-Ediacaran multicellularity. *Precambrian Research* 173:201–211.
- Cavalier-Smith T. 1987. Eukaryotes with no mitochondria. *Nature* 326:332–333.
- Cavalier-Smith T. 1991. Archamoebae: the ancestral eukaryotes? *Biosystems* 25:25–38.
- Cavalier-Smith T. 1993. Kingdom protozoa and its 18 phyla. *Microbiol Rev* 57:953–994.
- Cavalier-Smith T, Fiore-Donno AM, Chao E, Kudryavtsev A, Berney C, Snell EA, Lewis R. 2015. Multigene phylogeny resolves deep branching of Amoebozoa. *Molecular Phylogenetics and Evolution* 83:293–304.
- Cenci U, Sibbald SJ, Curtis BA, Kamikawa R, Eme L, Moog D, Henrissat B, Maréchal E, Chabi M, Djemiel C, et al. 2018. Nuclear genome sequence of the plastid-lacking cryptomonad *Goniomonas avonlea* provides insights into the evolution of secondary plastids. *BMC Biology* 16:137.
- Cerón-Romero MA, Maurer-Alcalá XX, Grattepanche J-D, Yan Y, Fonseca MM, Katz LA. 2019. PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes. *Molecular Biology and Evolution* 36:1831–1842.
- Chaudhary R, Bansal MS, Wehe A, Fernández-Baca D, Eulenstein O. 2010. iGTP: A software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 11:574.
- Chaudhary R, Fernández-Baca D, Burleigh JG. 2015. MulRF: a software package for phylogenetic analysis using multi-copy gene trees. *Bioinformatics* 31:432–433.
- Davis WJ, Amses KR, Benny GL, Carter-House D, Chang Y, Grigoriev I, Smith ME, Spatafora JW, Stajich JE, James TY. 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota, Fungi). *Mol Phylogenet Evol* 133:152–163.
- De Oliveira Martins L, Mallo D, Posada D. 2016. A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction. *Syst Biol* 65:397–416.
- Dehal P, Boore JL. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology* 3:e314.
- Derelle R, Lang BF. 2012. Rooting the Eukaryotic Tree with Mitochondrial and Bacterial Proteins. *Molecular Biology and Evolution* 29:1277–1289.
- Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, Lang BF, Eliáš M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *PNAS* 112:E693–E699.
- de Duve C. 2007. The origin of eukaryotes: a reappraisal. *Nat Rev Genet* 8:395–403.
- Embley TM, Hirt RP. 1998. Early branching eukaryotes? *Curr Opin Genet Dev* 8:624–629.

- Fernández R, Gabaldón T. 2020. Gene gain and loss across the metazoan tree of life. *Nat Ecol Evol* 4:524–533.
- Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, Martin M-J, Muffato M, Patricio M, Pereira C, et al. 2019. Advances and Applications in the Quest for Orthologs. *Mol Biol Evol* 36:2157–2164.
- Grant JR, Katz LA. 2014. Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. *PLoS Curr* 6:ecurrents.tol.c24b6054aebf3602748ac042ccc8f2e9.
- Guijarro-Clarke C, Holland PWH, Paps J. 2020. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat Ecol Evol* 4:519–523.
- Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups.” *PNAS* 106:3859–3864.
- He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL. 2014. An alternative root for the eukaryote tree of life. *Curr Biol* 24:465–470.
- Hilario E, Gogarten JP. 1998. The Prokaryote-to-Eukaryote Transition Reflected in the Evolution of the V/F/A-ATPase Catalytic and Proteolipid Subunits. *J Mol Evol* 46:703–715.
- Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, Embley TM. 1999. Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *PNAS* 96:580–585.
- Jackson CJ, Reyes-Prieto A. 2014. The Mitochondrial Genomes of the Glaucophytes *Gloeochaete wittrockiana* and *Cyanoptylche gloeocystis*: Multilocus Phylogenetics Suggests a Monophyletic Archaeplastida. *Genome Biology and Evolution* 6:2774–2785.
- Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30:772–780.
- Katz LA, Grant JR. 2015. Taxon-Rich Phylogenomic Analyses Resolve the Eukaryotic Tree of Life and Reveal the Power of Subsampling by Sites. *Systematic Biology* 64:406–415.
- Katz LA, Grant JR, Parfrey LW, Burleigh JG. 2012. Turning the Crown Upside Down: Gene Tree Parsimony Roots the Eukaryotic Tree of Life. *Systematic Biology* 61:653–660.
- Keeling PJ. 1998. A kingdom’s progress: Archezoa and the origin of eukaryotes. *Bioessays* 20:87–95.
- Lawrence DP, Kroken S, Pryor BM, Arnold AE. 2011. Interkingdom gene transfer of a hybrid NPS/PKS from bacteria to filamentous Ascomycota. *PLoS One* 6:e28231.

- Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679–685.
- Loron CC, François C, Rainbird RH, Turner EC, Borensztajn S, Javaux EJ. 2019. Early fungi from the Proterozoic era in Arctic Canada. *Nature* 570:232–235.
- Maddison WP, Knowles LL. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. *Systematic Biology* 55:21–30.
- Mallo D, Posada D. 2016. Multilocus inference of species trees and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371:20150335.
- Martin W, Rotte C, Hoffmeister M, Theissen U, Gelius-Dietrich G, Ahr S, Henze K. 2003. Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. *IUBMB Life* 55:193–204.
- Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31:i44–i52.
- Molloy EK, Warnow T. 2020. FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics* 36:i57–i65.
- Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution* 37:2763–2774.
- Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A. 2022. SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Molecular Biology and Evolution* 39:msab365.
- Nagy LG, Ohm RA, Kovács GM, Floudas D, Riley R, Gácsér A, Sipiczki M, Davis JM, Doty SL, de Hoog GS, et al. 2014. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat Commun* 5:4471.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* 34:401–437.
- Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *PNAS* 108:13624–13629.
- Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res* 38:W23–28.
- Price DC, Goodenough UW, Roth R, Lee J-H, Kariyawasam T, Mutwil M, Ferrari C, Facchinelli F, Ball SG, Cenci U, et al. 2019. Analysis of an improved *Cyanophora paradoxa* genome assembly. *DNA Research* 26:287–299.
- Redecker D, Kodner R, Graham LE. 2000. Glomalean Fungi from the Ordovician. *Science* 289:1920–1921.

- Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes. *Current Biology* 15:1325–1330.
- Roger AJ. 1999. Reconstructing Early Events in Eukaryotic Evolution. *Am Nat* 154:S146–S163.
- Rogozin IB, Basu MK, Csűrös M, Koonin EV. 2009. Analysis of Rare Genomic Changes Does Not Support the Unikont–Bikont Phylogeny and Suggests Cyanobacterial Symbiosis as the Point of Primary Radiation of Eukaryotes. *Genome Biol Evol* 1:99–113.
- Rosewich UL, Kistler HC. 2000. Role of Horizontal Gene Transfer in the Evolution of Fungi. *Annu Rev Phytopathol* 38:325–363.
- Sela I, Ashkenazy H, Katoh K, Pupko T. 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* 43:W7–W14.
- Stajich JE. 2017. Fungal Genomes and Insights into the Evolution of the Kingdom. *Microbiol Spectr* 5.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
- Stechmann A, Cavalier-Smith T. 2002. Rooting the Eukaryote Tree by Using a Derived Gene Fusion. *Science* 297:89–91.
- Stechmann A, Cavalier-Smith T. 2003. The root of the eukaryote tree pinpointed. *Current Biology* 13:R665–R666.
- Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. *Mol Biol Evol* 23:93–106.
- Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. 2019. New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. *Molecular Biology and Evolution* 36:757–765.
- Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. *Syst Biol* 62:901–912.
- Szöllősi GJ, Tannier E, Lartillot N, Daubin V. 2013. Lateral gene transfer from the dead. *Syst Biol* 62:386–397.
- Tekaia F. 2016. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights* 9:GEI.S37925.

- Vachaspati P, Warnow T. 2015. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics* 16:S3.
- Vallender EJ. 2009. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. *Methods* 49:50–55.
- Van de Peer Y, Ben Ali A, Meyer A. 2000. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene* 246:1–8.
- Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24:1540–1541.
- Wenzl P, Wong L, Kwang-won K, Jefferson RA. 2005. A functional screen identifies lateral transfer of beta-glucuronidase (gus) from bacteria to fungi. *Mol Biol Evol* 22:308–316.
- Wideman JG, Gawryluk RMR, Gray MW, Dacks JB. 2013. The ancient and widespread nature of the ER-mitochondria encounter structure. *Mol Biol Evol* 30:2044–2049.
- Williams TA. 2014. Evolution: Rooting the Eukaryotic Tree of Life. *Current Biology* 24:R151–R152.
- Willson J, Roddur MS, Liu B, Zaharias P, Warnow T. 2022. DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Systematic Biology* 71:610–629.
- Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.
- Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Molecular Biology and Evolution* 37:3292–3307.

Table 1. Summary of taxon selection for each study. Genera in **bold** are only in the taxonomy informed selected datasets (i.e. SEL+ and SEL-), and underlined genera are only in the randomly selected within clades datasets (i.e. RAN+ and RAN-). The genera with an asterisk (*) are microsporidians, which we excluded from datasets SEL- and RAN- because they often fall on very long branches (Embley and Hirt 1998; Hirt et al. 1999; Van de Peer et al. 2000). The numbers outside the parentheses are the number of species and the number inside the parentheses are those represented by whole genome data (More details are in Dataset S1).

Major Clade	Genera	Taxa (genomes)	
		SEL+	RAN+
Amoebozoa	<i>Acanthamoeba</i> , <i>Acytostelium</i> , <i>Clydonella</i> , <i>Dictyostelium</i> , <i>Endostelium</i> , <i>Entamoeba</i> , <i>Filamoeba</i> , <i>Flamella</i> , <i>Gocevia</i> , <i>Hartmanella</i> , <i>Mastigamoeba</i> , <i>Mayorella</i> , <i>Neoparamoeba</i> , <i>Ovalopodium</i> , <i>Paramoeba</i> , <i>Parvamoeba</i> , <i>Pessonella</i> , <i>Physarum</i> , <i>Polysphondylium</i> , <i>Stenamoeba</i> , <i>Stereomyxa</i> , <i>Thecamoeba</i> , <i>Unda</i> , <i>Vannella</i> , <i>Vermistella</i> , <i>Vexillifera</i>	22(3)	23(4)
Fungi	<i>Aspergillus</i> , <i>Batrachochytrium</i> , <i>Candida</i> , <i>Cryptococcus</i> , <i>Dacryopinax</i> , <i>Encephalitozoon</i> *, <i>Enterocytozoon</i> *, <i>Laccaria</i> , <i>Malassezia</i> , <i>Melampsora</i> , <i>Nematocida</i> *, <i>Neurospora</i> , <i>Nosema</i> *, <i>Phanerochaete</i> , <i>Piromyces</i> , <i>Puccinia</i> , <i>Rhizophagus</i> , <i>Saccharomyces</i> , <i>Schizosaccharomyces</i>	13(11)	13(10)
Other Opisthokonta	<i>Amphimedon</i> , <i>Anopheles</i> , <i>Apis</i> , <i>Aplysia</i> , <i>Branchiostoma</i> , <i>Caenorhabditis</i> , <i>Capitella</i> , <i>Capsaspora</i> , <i>Carteriospongia</i> , <i>Ciona</i> , <i>Culex</i> , <i>Drosophila</i> , <i>Equus</i> , <i>Fonticula</i> , <i>Gallus</i> , <i>Helobdella</i> , <i>Homo</i> , <i>Hydra</i> , <i>Hydractinia</i> , <i>Leucetta</i> , <i>Lubomirskia</i> , <i>Macaca</i> , <i>Mnemiopsis</i> , <i>Monosiga</i> , <i>Nematostella</i> , <i>Oikopleura</i> , <i>Ornithorhynchus</i> , <i>Oscarella</i> , <i>Pan</i> , <i>Pleurobrachia</i> , <i>Rattus</i> , <i>Saccoglossus</i> , <i>Salpingoeca</i> , <i>Schistosoma</i> , <i>Sphaeroforma</i> , <i>Trichinella</i> , <i>Trichoplax</i>	21(12)	21(14)
Archaeplastida	<i>Amborella</i> , <i>Arabidopsis</i> , <i>Bathycoccus</i> , <i>Chlorella</i> , <i>Chondrus</i> , <i>Coleochaete</i> , <i>Compsopogon</i> , <i>Crustomastix</i> , <i>Cyanidioschyzon</i> , <i>Cyanophora</i> , <i>Cyanoptyche</i> , <i>Erythrolobus</i> , <i>Galdieria</i> , <i>Glaucocystis</i> , <i>Mantoniella</i> , <i>Mesostigma</i> , <i>Micromonas</i> , <i>Nephroselmis</i> , <i>Ostreococcus</i> , <i>Physcomitrella</i> , <i>Picochlorum</i> , <i>Picocystis</i> , <i>Porphyra</i> , <i>Porphyridium</i> , <i>Pycnococcus</i> , <i>Rhodella</i> , <i>Rhodorus</i> , <i>Ricinus</i> , <i>Volvox</i>	20(7)	18(4)
SAR	<i>Alexandrium</i> , <i>Ammonia</i> , <i>Amphidinium</i> , <i>Amphiprora</i> , <i>Amphora</i> , <i>Astrosyne</i> , <i>Aureococcus</i> , <i>Bigelowiella</i> , <i>Blastocystis</i> , <i>Bolidomonas</i> , <i>Brandtodinium</i> , <i>Brevimastigomonas</i> , <i>Bulimina</i> , <i>Cafeteria</i> , <i>Chattonella</i> , <i>Chlorarachnion</i> , <i>Chrysoreinhardia</i> , <i>Corallomyxa</i> , <i>Corethron</i> , <i>Cryptosporidium</i> , <i>Ectocarpus</i> , <i>Eimeria</i> , <i>Euglypha</i> , <i>Euplotes</i> , <i>Extubocellulus</i> , <i>Florenciella</i> , <i>Fragilariopsis</i> , <i>Fucus</i> , <i>Gonyaulax</i> , <i>Gregarina</i> , <i>Gymnodinium</i> , <i>Gymnophrys</i> , <i>Karlodinium</i> , <i>Lankesteria</i> , <i>Leptophrys</i> , <i>Lingulodinium</i> , <i>Lotharella</i> , <i>Nannochloropsis</i> , <i>Nitzschia</i> , <i>Ochromonas</i> , <i>Oxytricha</i> , <i>Paracercomonas</i> , <i>Pelagodinium</i> , <i>Perkinsus</i> , <i>Phaeodactylum</i> , <i>Phaeomonas</i> , <i>Phyllostaurus</i> , <i>Phytophthora</i> , <i>Plasmodium</i> , <i>Pyrodinium</i>	40(17)	39(7)

	<i>Pythium</i> , Reticulomyxa , <i>Rhizochromulina</i> , Saprolegnia , <i>Sarcinochrysis</i> , <i>Scrippsiella</i> , Sorites , Spumella , <i>Stylonychia</i> , <i>Synchroma</i> , <i>Tetrahymena</i> , Thalassionema , Thalassiosira , Thraustochytrium , Toxoplasma , Vitrella		
Excavata	Euglena , Eutreptiella , <i>Giardia</i> , <i>Histiona</i> , <i>Histomonas</i> , <i>Jakoba</i> , <i>Leishmania</i> , Malawimonas , Monocercomonoides , Naegleria , <i>Neobodo</i> , Percolomonas , Reclinomonas , <i>Sawyeria</i> , <i>Seculamonas</i> , <i>Spironucleus</i> , <i>Stachyamoeba</i> , <i>Strigomonas</i> , Trichomonas , <i>Trimastix</i> , Tritrichomonas , <i>Trypanosoma</i>	22(7)	21(12)
Other eukaryotes	<i>Acanthocystis</i> , <i>Calcidiscus</i> , Choanocystis , Chrysochromulina , <i>Chrysoculter</i> , Collodictyon , <i>Cryptomonas</i> , Diphylleia , <i>Emiliana</i> , <i>Fabomonas</i> , <i>Goniomonas</i> , <i>Hanusia</i> , <i>Hemiselmis</i> , <i>Isochrysis</i> , Palpitomonas , <i>Pavlova</i> , <i>Phaeocystis</i> , Pleurochrysis , <i>Prymnesium</i> , <i>Raphidiophrys</i> , <i>Rhodomonas</i> , <i>Rigifila</i> , Roombia , Subulatomonas , <i>Telonema</i> , <i>Thecamonas</i> , <i>Tsukubamonas</i>	20(1)	20(1)

Figure 1. A root between Fungi and all other eukaryotes is the most parsimonious hypothesis based on 100 iterations of iGTP using all four taxon sets. Here we report the four most parsimonious topologies in 100 iterations of the analysis and note the number of times the first hypothesis appeared before any alternative in square parenthesis (i.e. a fungal root was present in the six iterations of iGTP with the lowest reconciliation scores in the SEL+ analyses). SEL+: Taxonomically informed taxa selection including microsporidians, SEL-: taxonomically informed taxa selection excluding the long-branch microsporidians, RAN+: random taxa selection including microsporidians, RAN-: random taxa selection excluding microsporidians. The caret (^) implies a non-monophyletic clade. For example, in datasets SEL+ and RAN+, the microsporidians do not fall in the same clade as the rest of the opisthokonts. We show the relative reconciliation costs compared to the optimum (lowest value) for each dataset. After Fungi–others, other parsimonious roots involve clades underrepresented in our dataset such as Glaucophyta or Apusozoa (see also Figure S4).

Figure 2. Constraining the species tree to match varying hypotheses of the root of ETOL supports a root at or within Opisthokonta and is inconsistent with other hypotheses. We show the relative reconciliation costs compared to the optimum (lowest value) for each dataset. The five hypotheses here are: A) Fungi–others (our estimate from the previous analysis; see results and Fig. 1), B) Opisthokonta–others (Stechmann and Cavalier-Smith 2002; Katz et al. 2012), C) (Ancyromonadida + Metamonada)–others, D) Discoba–others (He et al. 2014), and E) Unikonta-Bikonta (Stechmann and Cavalier-Smith 2002; Derelle et al. 2015). The empty circles on the cartoon phylogenies indicate where in the tree the constraint was applied, and other notations are as in Figure 1. Overall, there are significant differences (asterisks) between Fungi–others and any other hypotheses in all datasets, except Opisthokonta–others in datasets SEL+ and

RAN+ (Tables S1 and S2; significance level of 0.05). This result is consistent that Opisthokonta—others as the root and Fungi—others as a potential artifact.



