Rethinking large scale phylogenomics with PhyloToL 6, a flexible toolkit to enable phylogeny-informed data curation and analysis

Auden Cote-L'Heureux[1]*^, Marie Leleu[1]*, Godwin Ani[1,2], Rebecca Gawron[1], Laura A. Katz[1,2]

*Authors contributed equally

[1] Department of Biological Sciences, Smith College, Northampton, Massachusetts, United States of America
[2] Program in Organismic Biology and Evolution, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America

^Current Address: Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America

Key words: Phylogenomics, Bioinformatic tools, Microeukaryotes,

*Laura A. Katz

Email: lkatz@smith.edu

Author Contributions: ACL designed research, performed research, analyzed data, and wrote the paper; ML designed research, performed research, analyzed data, and wrote the paper; GA performed research, analyzed data, and wrote the paper; RG analyzed data; LAK designed research, performed research, analyzed data, and wrote the paper

Competing Interest Statement: The authors have no competing interests to disclose.

Classification: Biological sciences/Evolution

Keywords: genomics, transcriptomics, bioinformatic tools, eukaryotic tree of life

This PDF file includes:

Main Text
Figs. 1 to 4
Supporting information appendix

ABSTRACT

The bulk of eukaryotic diversity is microbial, with macroscopic lineages such as plant, animals and fungi nesting among a plethora of diverse lineages that include amoebae, flagellates, ciliates, and many types of algae. Our understanding of the evolutionary relationships and genome properties of microbial eukaryotes is rapidly advancing through analyses of omics (transcriptomic, genomic) data. However, phylogenomic analyses are challenging for microeukaryotes, and particularly uncultivable lineages, as single-cell approaches generate a mixture of sequence data from hosts, associated microbiomes, and contaminants. Current practices include resampling of hand-curated gene sets that can be difficult for other researchers to replicate. To address these challenges, we present PhyloToL version 6.0, a modular, user-friendly pipeline that enables effective data curation that includes phylogeny-informed contamination removal, estimation of homologous gene families, and generation of both multisequence alignments and gene trees. We provide several databases that will be of use for those interested in eukaryotic evolution: a 'Hook Database' of curated reference sequences for ~15k gene families; a database of transcriptome and genomes from 1,000 taxa with GFs assigned; and a highly-curated set of MSA and gene trees for 500 GFs in these taxa. We also demonstrate the power of a suite of stand-alone utilities that provide basic statistics on sequences, analyze compositional/codon patterns, and enable exploration of trees (e.g. clade-grabbing and efficient tip labeling). We exemplify the power of PhyloToL 6.0 in estimating eukaryotic phylogeny using the 500 conserved GFs, and set standards for curation of 'omics data for future research in the field.

Significance:

Illuminating the diversity of microbial lineages is essential for estimating the tree of life and characterizing principles of genome evolution. However, analyses of microbial eukaryotes (e.g. flagellates, amoebae) are complicated by both the paucity of reference genomes and the prevalence of contamination (e.g. by symbionts, microbiomes). PhyloToL 6.0 enables taxon-rich analyses 'on the fly' as users can choose optimal gene families for their focal taxa, and then use replicable approaches to curate data in estimating both gene and species trees. With multiple entry points and curated data from up to 15,000 gene families from 1,000 taxa, PhyloToL 6.0 provides a powerful launching point for researchers interested in the evolution of eukaryotes.

Introduction

Most of our knowledge about the nature and evolution of eukaryotic life has emerged from studies of macroscopic organisms, with a focus on 'model' lineages such as *Drosophila* and *Arabidopsis*. However, such models represent relatively narrow slices of the eukaryotic tree of life; the bulk of eukaryotic diversity is predominantly microbial (e.g. 1, 2). Insights from microbial eukaryotes (a.k.a. protists) expand our understanding of the 'rules' of evolution by their tremendous diversity of morphologies, life cycles and genome properties (3, 4). The gap in knowledge about microbial eukaryotes can be most efficiently filled through adaptable, taxon-rich phylogenomic analysis methods. However, current practices often rely on boutique datasets and decision-making (e.g. in removing contaminants and identifying orthologs) that can be challenging to replicate (e.g. 2, 5–7). To address these challenges, we developed PhyloToL 6.0, a flexible phylogenomic pipeline designed for replicable analyses of diverse eukaryotes; PhyloToL 6.0 is supplied with curated datasets plus a suite of utilities to enable efficient analyses of diverse eukaryotes.

The recent increase in molecular data (whole genome and transcriptome) and bioinformatic methods has spurred the creation of numerous pipelines to infer homology, multisequence alignments (MSAs), gene trees and species-level phylogenies (e.g. 8–11). These phylogenomic tools differ in their intentions, allowed inputs (e.g. GenBank vs. user-generated data), intended outputs (e.g. MSAs, trees); yet few include the type of curation needed for analyses of data from microbial eukaryotes given issues with contamination (i.e. from microbiomes and environmental sequences). The first step of many pipelines is collecting homologous sequences, which can be gathered directly from public databases such as GenBank (12), Pfam (13), or OrthoDB (14). Many recent pipelines rely solely on BLAST (15) or other similarity-searching algorithms (e.g. USearch (16), VSearch (17), and Diamond (18)) to infer homology. However, BLAST is based on similarity only and does not take into account biological relationships (19), and further processing is necessary to confidently establish the source of sequences as well as homology.

Alignments are challenging, particularly when dealing with data from diverse eukaryotes that span ~1.8 billion years of evolution (20), and methods of cleaning MSAs (e.g. trimming gapped or poorly aligned regions) are controversial (21). Multisequence alignment tools such as MAFFT (22), Clustal (23), and MUSCLE (24) are then used to construct gene trees with tools such as RAxML (Stamatakis, 2014) or IQ-Tree (26). To build species trees, many studies rely on a supermatrix-based approach that requires users to select orthologs (e.g. 2, 5–7), thereby removing phylogenetically informative duplication events (27).

Recent phylogenomic approaches include methods that use gene trees as inputs in inferring species level relationships (28). Such methods have been used for projects like the Open Tree of Life (29), and in studies of plants (30), animals (31) and viruses (32). An example of pipelines that follow these general steps is NovelTree, which performs homology assessment and gene tree construction, but it only accepts protein sequences as input data (8). Other examples that accept nucleotide sequences are PhyloTa (11), which focuses on homologous identification and

collection, and Sumac (10), which focuses on supermatrix building. However, they either do not include gene family assignment steps or do not offer the user flexibility to use a custom reference database for OG assignment, and none of these pipelines focus on removing contamination from the data.

Phylogenomic inference faces numerous challenges, including incongruence among loci, long branch attraction (33, 34), and lateral gene transfer, which confounds inferences (35). These issues are especially prevalent in microeukaryotes, where whole genome assemblies are still rare: moreover, many microeukaryotes possess their own microbiome, often resulting in high levels of contamination in transcriptomic samples. Such incongruences lead to conflicting and often spurious tree topologies that can be mitigated by thorough careful selection of taxa and curation of data (36, 37).

The importance of a taxon-rich dataset for estimating phylogeny accurately is well established (38, 39), and adding diverse lineages (e.g. taxonomic position, rates of evolution, levels of missing data) is likely to improve species estimates (40–43). However, even recent estimates of the eukaryotic tree of life rely on relatively few taxa (e.g., 234 taxa in Burki et al., (44), 186 taxa in Al Jewari & Baldauf (45), 158 taxa in Cerón-Romero et al., (46) and 109 taxa in Strassert et al., (47)), and many groups now resample the same genes/data matrix in generating species trees (e.g. 44, 47–49).The availability of user-friendly tools that facilitate the parallel processing of large numbers of taxa, therefore, has the power to increase the accuracy of large-scale estimates of eukaryotic phylogeny.

Here, we present PhyloToL version 6.0, substantially expanded from previous versions (50, 51) with revised scripts, enhanced data curation, a new suite of tools, and several databases (e.g. a reference database of 15,000 gene families, and curated sequence data for 1,000 taxa and 500 gene families). The core PhyloToL pipeline comprises two main components, which we refer to as PhyloToL part one and part two. PhyloToL part one takes input sequences from a whole genome or transcriptome assemblies, applies several curation steps, and provides initial homology assessment against a customizable database of reference sequences to assign GFs. Part one outputs a fasta file of curated nucleotide and amino acid sequences with gene families assigned, as well as a dataset of descriptive statistics (e.g. length, coverage, and composition) for each input sample. PhyloToL part 2 is highly modular; for a given selection of taxa and GFs, it stringently assesses homology by iterating the tool Guidance (52, 53), which outputs an MSA for each gene family. From MSAs, PhyloToL 6.0 builds gene trees (using either IQTree or RaxML), and then includes an innovative workflow for tree topology-based contamination removal. In addition to presenting the core pipeline and accompanying utility scripts, we describe the results of an analysis of 500 conserved GFs from 1,000 taxa, from which we remove contamination and construct species trees.

Results

The pipeline and accompanying scripts

PhyloToL version 6 enables efficient phylogenomic analysis of eukaryotes through phylogeny-informed curation of 'omics' data. PhyloToL is both flexible and modular, and version 6 includes updates to the core PhyloToL pipeline (50, 51) to streamline the workflow for curating data from transcriptomes/genomes (Part 1) and improve curation for producing MSAs and gene trees (Part 2). All components of the toolkit are written in Python and are available for download on GitHub (https://github.com/Katzlab/PhyloToL-6) and Zenodo (DOI:10.5281/zenodo.13323372). To demonstrate the power of PhyloToL, we present several databases (the Hook reference database for GF assignment, sequences files for 1000 taxa both before and after curation), the pipeline, and a set of utility scripts to facilitate phylogenomic analyses.

PhyloToL 6 is designed to take as input either assembled transcripts, genomic CDSs or any sequences with names matching simple criteria (i.e. a 10 digit taxon code plus a unique identifier) as input (see methods and SI Appendix). Curation steps are built into both parts of the pipeline, first enabling analysis of data within a taxon based on sequence properties (e.g. GC content; Dataset S1; Dataset S17) and second with homology assessment and phylogeny-informed identification/removal of contaminants (Fig. 1). PhyloToL part 1 allows users to either use the built-in Hook Database of ~15,000 eukaryotic gene families (GFs) or a custom reference database (see below) to produce curated amino acid and nucleotide sequences for each taxon (Fig. 1). PhyloToL Part 2 takes these files as input and constructs an MSA and gene tree for each input gene family (Fig. 1). Part 2 also includes a novel and flexible workflow for phylogeny-informed contamination removal that we refer to as the 'contamination loop', which identifies likely contaminant sequences based on the topology of the gene trees and user-determined criteria.

Beside the main pipeline, PhyloToL 6 includes a set of stand-alone utility scripts that aim to increase the power of the analysis done with or without the core PhyloToL pipeline. We divide these scripts into five main categories: basic statistics, composition tools, MSA tools, gene tree description, and contamination removal (Dataset S2).

Databases

Alongside the suite of scripts that make up the PhyloToL 6 toolkit, we provide several taxon-rich databases useful in broad analyses of eukaryotic phylogeny: 1) our Hook reference database for GF assignment, 2) amino acid and nucleotide sequence files for 1000 species with GFs assigned (called ReadyToGo files), and 3) curated MSAs and trees for 500 conserved gene families. The Hook Database is composed of 1,453,081 sequences across 15,414 GFs (Dataset S3, Dataset S24), and is used in PhyloToL part 1 as a reference database to assign either assembled transcripts or coding domain sequences to GFs. The Hook Database captures a broad diversity of eukaryotic gene families and was built using sequence data from OrthoMCL version 6.13 (54), which we sampled to select for GFs that are present across the eukaryotic

tree and/or present in under-sampled lineages of eukaryotes (see methods; Fig. S1, Fig. 2). To add value for users, we also include functional annotations for each GF in the Hook (Dataset S4; see methods in SI Appendix). Alternatively, users can insert their own Diamond-formatted database in lieu of the Hook, to target only specific genes of interest.

Performance

To assess the performance of the PhyloToL part 1, we compared the speed in processing assembled transcriptomes and genomes used for this study into ReadyToGo files. Using a personal computer (iMac Pro 2017, 64GB of RAM,10 cores) and a high performance computing cluster (Unity cluster, 128GB of RAM, 24 cores), we processed 10 and 100 transcriptomes and genomes (Dataset S5). As expected, processing genomes with the pipeline was considerably faster compared to the transcriptomes on both computers; coding domains are already called for genomes. On the personal computer it took roughly 2 hours for 10 transcriptomes (513,904 transcripts) and 24 hours for 100 transcriptomes (3,294,484 transcripts) while the same datasets took 2 and 16 hours respectively on the HPC. Processing the genomes, it took 1 hour 20 minutes and 24 hours on the personal computer for 10 (106,249 CDS) and 100 taxa (1,158,224 CDS) respectively, and 25 minutes and 21 hours on the HPC to run the same set of taxa. This demonstrates the feasibility of running PhyloToL on personal computers if an HPC is not available.

Exemplar analysis of 500 conserved gene families

To demonstrate the capabilities of PhyloToL version 6, we conducted phylogenomic analyses on 500 gene families using 1000 diverse taxa (Datasets S1-S6-S7 and S16-S18). Here our goal is to demonstrate the power of PhyloToL in terms of data curation and in providing replicable approaches in generating both gene and species trees, which includes tracking removed sequences. To this end, we estimate species trees at four different points in the contamination removal process (Fig. 4a-d): 1) before the contamination loop: after generating MSAs/trees from curated “R2G” files based on GC content at silent sites (e.g. GC3s; Datasets S1 and S16); 2) after applying sister-based rules to iteratively remove sequences determined to be potential contaminants based on user-established rules (Datasets S8 and S9); 3) after clade-grabbing by retaining only sequences for which we have the greatest confidence based on user-established expectations of taxon density (Datasets S10 and S11); 4) after endosymbiotic gene transfer (EGT) removal by removing gene families that we identify as likely affected by primary (e.g. photosynthetic eukaryotes nested among cyanobacteria (Dataset S19)) and secondary (e.g. photosynthetic eukaryotic clades interdigitated with one another (Dataset S19)) EGTs. With the goal of demonstrating the power of the approach rather than estimating the ‘correct’ eukaryotic tree of life, we provide readers with details on presence/absence of ‘supergroups’ (e.g. Amoebozoa, Archaeplastida, Opisthokonta, SAR) to orient our study in light of existing literature. All files related to this analysis, for each step of the contamination loop, can be found at on Figshare (https://doi.org/10.6084/m9.figshare.26527984) in demonstration of another powerful aspect of PhyloToL: the ability to track intermediate files.

Gene family analysis

We compiled a set of whole genome and transcriptome assemblies from 1,000 taxa (275 bacteria, 98 archaea, and 627 eukaryotes (Dataset S7)), and ran each of these samples through PhyloToL 6 part one, retaining sequences that were assigned gene families in a similarity search against the Hook Database. Following an initial round of assessment, we chose the 1000 taxa based on data quality and taxonomic diversity (see SI Appendix for details). We selected the 500 gene families based on taxonomic presence, using two criteria: 1) they are among the most shared GFs in our 1000 taxa (determined by “SharedOGs.py utility script) and 2) these GFs have low paralogy (Dataset S13).

Despite the fact that the starting OrthoMCL database is biased in terms of taxonomic availability (e.g. many parasites, relatively low diversity; (54)), the 15,414 GF Hook database assigned gene families to a broad diversity of taxa, including poorly represented taxon like *Telonema,* Centrohelidae and other orphan lineages (labeled EE for “everything else”; Fig. 2, Dataset S1). In fact, the taxonomic distribution of major clades in our ReadyToGo files is greater than in the Hook itself, with more than 75% of the GFs present in at least 4 major clades in the ReadyToGo files (Fig. 2a,b). Nevertheless, the distribution of GFs across taxa is highly variable, reflecting at least two phenomena: the differences between transcriptome and whole genome data, and the prevalence of gene loss in some lineages (e.g. fungi (55) and parasites; Fig. 2c).

Using PhyloToL utilities (CUB.py, GC_identifier.py), we further refined data based on taxon-specific GC composition ranges (SI Appendix) to produce ReadyToGo files with sequences labeled (OG6 if in GC3S range for taxon, OGG and OGA if more GC rich or AT rich, respectively). We provide the resulting ReadyToGo databases for users interested in placing their taxa into a phylogenomic context; users might choose to include all sequences knowing that those labeled OGG or OGA are outside of the GC content ranges chosen for this study (Dataset S16).

Initial MSAs and gene trees

A first result of PhyloToL part 2 is the generation of 500 MSAs and gene trees that emerge from the curated ReadyToGo data. These 500 MSAs and trees, built after five iterations of Guidance and labeled as “pre-contamination loop”, vary in taxon presence and number of sequences (Dataset S13), with a total of 15,486 sequences (Dataset S21) removed by Guidance as putative non-homologs. Notably, individual gene trees include numerous contaminants (e.g. bacterial and food source sequences from transcriptome data, Fig. 3b). To summarize the phylogenetic signal in these data, we generated a concatenated tree (Fig. 4a), which is largely consistent with the published literature. For all species tree alignments, we masked columns with ≥50% missing data and we also generated species trees using Asteroid (56) at each step (Dataset S23). The resulting trees from all analyses are discussed below.

Contamination Loop

With the goal of demonstrating the power of PhyloToL, we ran the “contamination loop” included in PhyloToL part 2 to estimate the topology of the eukaryotic tree of life in a manner that is both transparent and repeatable. This contamination loop allows both “sister/subsister” removal (i.e. removal of potential contaminants) and “clade-grabbing” (i.e. retention of sequences for which we have greatest confidence based on taxonomic density) with user-defined rules (see methods; Datasets S8 to S12). Using the set of scripts provided with the PhyloToL toolkit (e.g. ContaminationBySisters.py and CountTaxonOccurrence.py, see methods), we assessed the taxon presence and topology of each tree to establish three sets of rules: sister, subsister, and clade grabbing (Datasets S8 to S12). We make no claim to be correct in all cases, but we believe that PhyloToL’s ability to document sequence choice in a transparent manner is an improvement to best practices in the field.

Examples where sister and subsister rules may be helpful include cases where a taxon, or a pair of taxa, is repeatedly contaminated by a food source (e.g. ciliate sequences falling among haptophytes), or by a known host in the case of parasites (e.g. the gregarine apicomplexa whose sequences are contaminated by host metazoa; Fig. 3b). We set additional sisters/subsister rules based on our knowledge of the biology of the taxa and inspection of the resulting trees; here, for example, pairs of closely related taxa fall in aberrant places such as *Aphelidium insulamus* and *Aphelidium tribonematis* (Op_ap_Ains and Op_ap_Atri) branching together in a stramenopiles or Amoebozoa; these taxa appear contaminated by same sources. The greatest proportion of of removed sequences are from taxa within the major clade SAR (abbreviated as Sr), which includes a majority of field-caught single cell transcriptomic (Ciliates, Rhizaria; Dataset S22). At the end of the sister mode of the contamination loop, PhyloToL removed 50,903 of 565,225 sequences (Datasets S13 and S22). We again estimated a tree from concatenated alignments masked at 50% (Fig. 4b) and from Asteroid (Dataset S23), and we discuss the resulting trees with those from other stages below.

The second form of contamination removal in PhylToL6 is termed “clade-grabbing” as it relies on user-defined clade sizes for selecting sequences to retain. Here, we set minimal numbers of target taxa for clades with many representatives, noting that we likely remove a considerable amount of vertically-inherited data here given the prevalence of gene loss and missing data of some clades. We also identified a list of ‘exceptions’ (i.e. taxa with few close relatives in our analyses; for example orphan lineages as *Mantamonas* and *Hemimastix)* to retain, independent of clade size (Datasets S10-S12). We first ran clade-grabbing only for ciliates (a clade whose monophyly is not controversial in the literature) as we had a strong signal of contamination of parabasalid sequences labeled as the ciliate isolated from the digestive system of cows; here, a cluster of ciliate transcriptomes containing parabasalid sequences (Ex_pa; Fig. 3c), likely cause the ciliates to fall near parabasalids after sister/subsisters removal (Fig. 4b).

After addressing the high level of contamination of ciliate data, we deployed the “clade-grabbing” mode of the contamination loop more generally based on clade sizes determined empirically using the PhyloToL toolkit (CladeSizes.py, see methods). Aiming to retain sequences for which we have greatest confidence, we provide readers with all the

information needed to explore removed sequences and/or to implement this procedure for their own data. For example, given that we have a total of 45 diverse metazoa (Datasets S1-S6; Op_me; Opisthokonta, Metazoa), we kept only clades containing at least 11 metazoan species. We allowed contaminating sequences to interrupt clades; e.g. for metazoa, we allowed at most 10% of a clade to be non-metazoan species. We share these rules plus the resulting removed sequences in Datasets S8-S12 and S22. In the end, clade grabbing removed 129,458 of 514,272 sequences (Datasets S13 and S22). Here again, we emphasize on the importance of transparency and user defined rules for clade grabbing, as this process will most likely remove "good" sequences as well as contaminants while selecting for strongest signals in trees (e.g EGT, see below).

Given the many papers demonstrating a substantial effect of endosymbiotic gene transfer (EGT) from plastids to nuclei (reviewed in 57), we conducted a final analysis to remove gene families that had been affected by both primary and secondary EGT. To this end, we used both the contamination loop and the utility script CladeGrabbing.py to identify gene-families with a putative photosynthetic history. Gene families affected by primary EGT are those with photosynthetic lineages nested among cyanobacteria while secondary EGTs are identified as those with the greatest proportion of sequences of intermingled photosynthetic lineages (e.g. dinoflagellates nested among diatoms; Datasets S12 and S19). After removing 32 and 51 GFs determined to be affected by primary and secondary EGT, respectively, and 7 GFs putatively exhibiting both primary and secondary EGT, we ran concatenated and asteroid analyses and we compare the resulting trees below (Fig. 4d).

Comparison of trees from four stages: Pre-contamination loop, after sisters- and clade-based contamination removal, and after EGT removal.

With the goal of emphasizing the methods, we discuss the four concatenated trees and provide readers with both individual alignments/gene trees (Dataset S19) and lists of removed sequences (Dataset S22) to enhance transparency. Tree topologies reconstructed through concatenation (and Asteroid, see Dataset S23) are overall concordant from the "pre-contamination loop" through the various stages of curation for the 1,000 taxa and 500 gene families (Fig. 4a-d). Across all trees, many monophyletic clades (filled triangle) occur including dinoflagellates, Rhizaria, metazoa, green algae, plus the major clade Archaeplastida (Fig. 4a-d), but the effect of sister/subsister removal is more visible on the Asteroid trees (Dataset S23), where the "pre-contamination loop" tree shows more aberrant taxon placement resulting in non-monophyletic clades (open triangles); for example, the parasite *Piridium sociabile* (Sr_co_Psoc) falls within animals prior to running the contamination loop (Fig. 4a)

In both the pre-contamination loop species tree and the species tree built after sisters-based contamination removal, Microsporidia (a fungal lineage known to have elevated rates of evolution (58, 59), fall towards the base of the eukaryotic portion of the tree, and the major clade Amoebozoa does not emerge until the final analysis, as the Archamoeba also fall towards the base of the eukaryotic tree. Other spurious placements (e.g. the break of Alveolata (ciliates, dinoflagellates, Apicomplexa) and the placement of the stramenopile *Cantina*) also resolve

through data curation. Also noteworthy, the putatively-novel lineage Hemimastigophora either falls sister to the Ancyromonida (Fig. 4a-4c) or sister to Malawimonadida (and then Ancyromonida (Fig. 4d)), though missing data likely confounds the placements of all of these lineages (Dataset S14).

For all trees generated, we obtain a root among excavate taxa (formerly the clade 'Excavata') , though this may be due in part to a high amount of missing data in these taxa, especially metamonads; on average, metamonads have the greatest number of gaps in the concatenated alignment and contain the fewest OGs used in our analyses, followed by other excavates (Fig. S3 and Dataset S14). Interestingly, the order of appearance of clades throughout phylogeny-informed curation (Fig. 4) follows inversely the amount of missing data on average among taxa in each clade (Fig. S3), indicating that missing data may indeed play a role in generating spurious topologies. In addition to excavate lineages, other taxa with a high proportion of missing data such as the Breviata also fall near the base of eukaryotes; the placement of this and other orphan (purple) lineages like the Centrohelidae, Hemimastigophora, Malawimonads, and Ancyromonads is difficult to interpret, as these lineages lack close relatives to stabilize topologies. Only the last step of phylogeny-informed contamination removal (after the removal of most putative EGT) retrieves the monophyly of SAR, though in this version of the species tree we see numerous orphan lineages still fall towards the root of eukaryotes. Other proposed eukaryotic "supergroups" (e.g. TSAR, Crums, Obazoa, Diaphoretickes) are not recovered in any analysis.

Discussion

PhyloToL 6.0 provides a platform for the efficient curation and analysis of 'omics data from eukaryotes, using tree-based methods that enable exploration of both gene families and species relationships. Key aspects of PhyloToL are its repeatability and transparency, as users can keep a record of parameters used (e.g. in identifying contaminants) and both retained and removed sequences through every step. Analyses of diverse microbial eukaryotes, and particularly uncultivable lineages characterized by single-cell 'omics, require curation to remove contamination from the environment and possible symbionts. Numerous 'boutique' approaches that require hand-curation have been used to estimate eukaryotic phylogeny (47, e.g. 60–63), but reliance on resampling of data (i.e. choosing orthologs to match previous concatenated gene sets) can lead to issues arising from a lack of independence (reviewed in 64). While standards of curation and data quality have been developed for analyses such as genome assembly and annotation (e.g. 65, 66), analogous standards do not yet exist for phylogenomics.

PhyloToL 6.0 provides a streamlined method to include diverse data types (e.g. transcriptomes, genomes) allowing users to choose gene families for each study that maximize power in analyses and to curate data *via* both per-taxon and tree-based comparative approaches (Fig. 3). The databases and scripts require minimal effort for installation and are structured for users with modest bioinformatic skills, as accessibility and usability are key considerations in designing scientific tools (67). We provide PhyloToL 6.0 with default settings for parameters that we believe are a reasonable starting place for analyses, though all parameters are easily

customizable. Alongside the code, we provide a comprehensive manual on GitHub that describes how to use the PhyloToL toolkit (https://github.com/Katzlab/PhyloToL-6/wiki).

PhyloToL's flexibility extends to focal gene families, allowing users to the Hook Database (Fig. 2) that includes a set of over 15,000 gene families chosen based on the breadth of their distribution among eukaryotes, or to replace this dataset with a set of gene families of interest; this flexibility extends beyond previous phylogenomic approaches (e.g. 2, 5, 6). For those interested in analyzing data from previously uncharacterized taxa, we also emphasize the importance of a taxon-rich dataset for large-scale 'omics data curation in order to capture both phylogenomic context (e.g. ensure robustness by presence of homologs in closely-related taxa and account for gene duplications) and to enable robust inferences about eukaryotic features. To this end, we provide the PhyloToL 6.0 Database, a set of curated data from 1,000 diverse transcriptomes and genomes from eukaryotes, bacteria, and archaea that we have curated (Dataset S1, Dataset S16).

Our exemplary analysis of 500 conserved gene families demonstrates the power of PhyloToL to analyze large, diverse eukaryotic datasets. Even the first concatenated tree to emerge from PhyloToL (before contamination loop but after choosing sequences based on a predetermined range of GC content at silent sites) is largely concordant with published literature (Fig. 4a), particularly for morphologically-defined clades like dinoflagellates, animals, red algae, Tubulinea, and Euglenozoa (e.g. 1, 2, 45–47). Importantly, most other published analyses rely on many fewer taxa and fail to demonstrate monophyly of so many clades. Also, all four of our analyses place the root among parabasalids and fornica, generally consistent with the hypothesis in Al Jewari and Baldauf (2023), though testing the root would require more attention to gene-family selection and to mitigating the effect of missing data. As we remove putative contaminants based on sister/subsister analysis (Fig. 4b) and then retain sequences for which we have the greatest confidence through 'clade-grabbing' (Fig. 4c) we see additional major clades supported (e.g. Opisthokonta, Alveolata (ciliates, dinoflagellates, Apicomplexa and their relatives)). Only with removal of gene families most affected by EGT do we recover SAR (stramenopiles, alveolates, Rhizaria), and we note the shift in relationships among Archaeplastida in these analyses (Fig. 4d); these analyses suggest that EGT may be a driver in inference about eukaryotes tree of life.

In summary, we have demonstrated how PhyloToL can be used as a starting place for testing evolutionary hypotheses. PhyloToL allows researchers to rapidly compare hypotheses regarding the placement of disputed lineages (e.g. Telonemia (47) or Hemimastigophora (68)) in a repeatable and transparent manner (i.e. tracking of removed sequences), while enabling taxon-rich analyses. PhyloToL also allows exploration of the placement of 'orphan' lineages on the eukaryotic tree by leveraging the ability to treat data from these taxa differently (e.g. more leniency in curation) than data from taxa belonging to better sampled clades. Moreover, because researchers can choose gene families independently for each study for up to 1000 taxa provided by this study, PhyloToL may help to mitigate the problem of recovering similar topologies across resampled datasets (e.g. 44, 47–49). PhyloToL allows for 'phylogeny on the fly' as users can reset gene families, contamination-removal rules and run the pipeline and

associated toolkit with great flexibility, modularity and transparency. In sum, PhyloToL synthesizes a broad set of tools to facilitate large phylogenomic analyses from start to finish, and most importantly sets foundations for establishing best practices in the field.

There are several important caveats to consider when using PhyloToL. While the PhyloToL pipeline is built to be generalizable, it includes stringent data-quality filters that may remove sequences of interest to certain studies, and is therefore best suited for processing data for large-scale evolutionary or population-level analyses (e.g. generating many diverse gene trees for a supertree approach to phylogeny). Similarly, some parameters which we applied universally (such as the Guidance sequence cut off) are likely not appropriate for all taxonomic groups, and there is room for improving the flexibility of parameter fitting by taxon. Users interested in the evolutionary history of specific genes should use PhyloToL with caution as fast-evolving sequences may be removed, though again assessment here is easy as PhyloToL provides intermediate files as output for inspection by users. Finally, we note that though the stochasticity associated with aligning sequences and building gene trees makes some aspects of analyses not completely replicable, the structure of PhyloToL increases transparency (i.e. by retaining user-defined rules and removed sequences) to enable streamlined and large-scale phylogenomic studies.

Synthesis

Currently, studies of microbial eukaryotes rely heavily on bioinformatics tools developed for macrobes and/or bacteria; however, such tools do not incorporate workflows that are critical for accurate analysis of eukaryotic lineages where the underlying data must be rigorously cleaned in light of contamination and non-vertical gene transfer. In light of increased attention to the importance of democratizing biology research, especially in the realm of software tools (67, 69, 70), we designed PhyloToL to be accessible to researchers with limited bioinformatic background. In its novel phylogeny-informed contamination removal methodology and especially in its flexibility in curating well-sampled lineages differently from ‘orphan’ eukaryotic lineages, PhyloToL has the potential to increase the standards and repeatability of studies of eukaryotic phylogeny.

Materials and Methods

Here we provide an overview of methods, including descriptions of taxa and gene families, the development of the PhyloToL Hook 6.6 database, brief descriptions of the functionality of PhyloToL, and details on our exemplary analyses of 500 gene families in 1000 taxa. Further details are provided in the supplemental methods section of the SI Appendix.

Taxon and gene family designations

PhyloToL 6.0 is based on carefully controlled names of both clades and species that facilitate analyses. Each transcriptome and genome in the PhyloToL 6.0 database is identified using a ten-digit code, which represents either an individual cell or GenBank accession, or a pool of

transcripts as noted in Dataset S1. The first two digits of the code identify one of eight 'major' clades as follows: Ba, Bacteria; Za, Archaea; Op, Opisthokonta; Am, Amoebozoa; Ex, excavate lineages (formerly the clade 'Excavata'); Sr, SAR (Stramenopila, Alveolata, and Rhizaria); Pl, Archaeplastida; EE, orphan lineages. The next two digits identify the taxonomy of the taxon at the 'minor' clade level (e.g. within Opisthokonta are the minor clades Op_me for Metazoa; Op_fu for Fungi; Op_ch for choanoflagellates; and Op_ic for Ichthyosporea; Dataset S6). The last four digits identify the species and, if applicable, sample ID within a species (e.g. Am_tu indicates the minor clade Tubulinea, and there are multiple samples of *Hyalosphenia papilio*, identified as Am_tu_Hp01, Am_tu_Hp02, etc.; Dataset S1). Gene families (GFs) are identified as per the notation in OrthoMCL version 6.13 (54), with the prefix OG6_ followed by a unique six number sequence (see sections on Hook database and composition-based curation below). All sequence identifiers used in PhyloToL 6.0 databases are unique and begin with the ten-digit taxon identifier, then are labeled by a unique contig/CDS ID designated either by an assembler or by annotations as downloaded from GenBank, and end with a ten-digit GF identifier.

Development of the 'Hook 6.6' database

As a starting place for evolutionary analyses of lineages sampled across the eukaryotic tree of life, we developed a 'Hook' database of 15,414 GFs selected for presence across a representative set of eukaryotes. The 'Hook' allows assignment of sequences to GFs and can easily be replaced by researchers interested in specific gene families (*e.g.* gene families involved in epigenetics, in meiosis, etc.). To develop the Hook 6.6 (Fig. S1), we started with 'core' orthologs from the OrthoMCL version 6.13 database (495,339 GFs). We then proceeded to several curation steps, with the goals of 1) reducing the size of the database as much as possible while retaining diversity within eukaryotes; 2) retaining only GFs that are present in a representative set of eukaryotes given our focus on microbial lineages (i.e. we undersample animal-specific and plant-specific GFs); and 3) removing GFs and sequences within GFs that are likely to cause sequences to be misassigned or assigned to groups of sequences without useful functional meaning (e.g. sequences that comprise only a single common domain, or chimeric sequences). To achieve these goals, we assessed the taxonomic diversity and the quality of each GF using a variety of custom scripts (DOI:10.5281/zenodo.13323372). We detail these curation steps in the SI Appendix, and reiterate the goal of generating a set of representative gene families to use in analyses of diverse eukaryotes.

PhyloToL 6 part 1

PhyloToL comprises two components: the first (PTL6p1) provides initial gene family assignment to sequences and and the second (PTL6p2) builds alignments and phylogenetic trees. Central to all of PhyloToL is the use of consistent taxon codes (see above). PTL6p1 has two versions, the first intended for use with transcriptomic data, and accepts as input assembled transcripts as produced by rnaSpades (Fig. S2). Users may use other assembly tools as long as sequence names are subsequently formatted in the style of rnaSpades (i.e. including a contig identifier, k-mer coverage and length). The second version is for use with whole genome data, and accepts as input nucleotide coding sequences (CDS). Each step can be run individually across any number of input samples, runs can be paused and resumed at any stage, and this can be

flexibly managed using a wrapper script provided in the Zenodo repository (DOI:10.5281/zenodo.13323372)

Transcriptomic pipeline

The transcriptomic pipeline requires three inputs: a fasta file of correctly named contigs (see manual), a file specifying a genetic code (if known) for each taxon, and for those interested in removing sequences misidentified due to index-hopping, a list of names of conspecifics (i.e. taxa/samples that are expected to share identical nucleotide sequences). As described in detail in the SI Appendix and in Fig. S2, PTL6p1 removes sequences based on length parameters, and optionally sequences that are likely incorrectly labeled due to index hopping (71, 72) in the same sequencing run. Next, putative rRNA sequences are moved to a separate folder and remaining sequences are labeled as possible prokaryotic contamination (ending in _P) for users to inspect downstream. To provide initial gene family assignments, Diamond (54) is used to compare sequences either to the PTL6 Hook database (described above) or a user-provided database. As the Hook database is replaceable and customizable, this step offers an opportunity to filter transcriptomic data for a group of gene families/functional groups of interest. PhyloToL then captures ORFs as both nucleotide and amino acid sequences. Finally, PTL6p1 removes putative chimeric and partial transcripts to produce “ReadyToGo” fasta files and calculates a variety of statistics for both sequences and taxa.

Genomic pipeline

The version of PTL6p1 applicable to coding domain sequences (CDSs) from whole genome assemblies is similar to the version for transcriptomic data described above and in the SI Appendix, but with some important differences. Given that coding domains are already determined, this version of PTL6p1 has no length filter, and instead immediately evaluates in-frame stop codon usage and translates the nucleotide CDSs to amino acids, at which point it uses Diamond BLASTp to assign gene families against the same reference database (in our analyses, the Hook Database). Next, the pipeline filters sequences by relative length, removing any sequence less than one third or more than 1.5 times the average the length of its gene family in the Hook Database. After some reformatting, PTL6p1 then outputs the same “ReadyToGo” files as the transcriptome version of the pipeline: a nucleotide and amino acid fasta file with gene families assigned for each taxon, a tab-separated file of BLASTp data against the Hook database, and summary statistics.

PhyloToL 6 part two

The second major component of the pipeline (PTL6p2) starts from the “ReadyToGo” files produced by part 1 (or any set of per-taxon sequences with names that match PLT6 criteria) and generates multisequence alignments and trees. Prior to running Guidance (52, 53) for homology assessment, optional filters are available in the script ‘preguidance.py’, to select the sequences to use for the analysis based on GC composition or high similarity proportions (details in the SI Appendix), on the whole dataset or on specific taxon.

Then PTL6p2 runs Guidance (52, 53) in an iterative fashion to remove non-homologous sequences defined as those that fall below the sequence score cutoff. (We note that there is

some stochasticity here given the iteration of alignments built into the method.) After inspecting a diversity of gene families, we have lowered the default sequence score cutoff from 0.6 to 0.3, though this may not be appropriate for all genes (see caveats section below). To remove regions with large gaps that can confound tree building, the resulting MSAs are then run through TrimAl (73) to remove all sites in the alignment that are at least 95% gaps (again, a parameter a user could alter). The last step of PTL6p2 before phylogeny-based contamination removal is to construct gene trees, though users can stop PhyloToL after Guidance to build trees with other softwares as they prefer. Currently PhyloToL supports RAxML (25), IQ-Tree (with the hardcoded protein LG+G model (26), and FastTree (74).

Phylogeny-based contamination removal

A key innovation in PhyloToL 6.0 is the “contamination loop”, an iterative tool to identify and remove contamination based on analyses of single gene trees. This tool incorporates two main methods of contamination assessment informed by tree topology. The first method – ‘sisters’ mode – is intended to target specific instances of contamination. It enables users to remove sequences based on cases of repeated contamination in target taxa, determined by prior assessment of trees (aided by the utility script ContaminationBySisters.py or known contaminants; Fig. 3). We provide additional details in the SI Appendix. The second method – “clade-based contamination removal” – is intended for cases when the user is interested in genes present in a group of organisms with multiple representative samples and/or species in the gene trees (Fig. 3). For a given set of target taxa, this method identifies robust, monophyletic clades containing those taxa within each gene tree (allowing a user set number of contaminants), and re-aligns and re-builds the tree excluding all sequences from the target taxa that do not fall into these robust clades. In both cases, sister and clade grabbing, a user-defined set of rules is necessary and can be built using the set of utility scripts provided with the main pipeline. Given that these methods incorporate tree-building on each iteration, users should expect some amount of stochasticity in which sequences are removed.

Ortholog selection for concatenation

PhyloToL 6 part 2 includes an option to concatenate representative sequences per GF into a supermatrix from which users can construct a species tree. This can be done as part of an end-to-end PhyloToL run, or by inputting already complete alignments and gene trees and running only the concatenation step. If a GF has more than one sequence from a taxon, PhyloToL keeps only the sequences falling in the monophyletic clade in the tree that contains the greatest number of species of the taxon’s clade as determined by its sample identifier. If multiple sequences from the taxon fall into this largest clade, then the sequence with the highest ‘score’ (defined as length times k-mer coverage for transcriptomic data with k-mer coverage in the sequence ID as formatted by rnaSpades, and otherwise just length) is kept for the concatenated alignment. If a GF is not present as a taxon, its missing data are filled in with gaps in the concatenated alignment. Along with the concatenated alignment, this part of the pipeline outputs individual alignments with orthologs selected (and re-aligned with MAFFT), in case a user wants to construct a model-partitioned or other specialized kind of species tree.

Conserved OG analysis

To demonstrate the power of PhyloToL, we conducted a phylogenetic analysis on 500 conserved gene families among 1,000 species. Selection of taxa and gene families to include in this study was based on quality of data and taxon presence. We went through several rounds of curation and selection that are detailed in the SI Appendix; the final selection of taxa is described in Datasets S1 and S7. We used PhyloToL 6.0 part 1 to produce fasta formatted CDS files (genomes) and assembled transcripts (transcriptomes) for each genomes and transcriptomes downloaded from public databases or part of our in-house dataset.

We then reran PTL6p2 with these 1000 taxa, using only the sequences labeled as 'OG6', based on GC composition (see SI Appendix for details), with five iterations of Guidance, and built trees using IQ-Tree (-m LG+G; Dataset S18). We then removed sequences identified as contaminants by the contamination loop in PTL6p2. We first ran ten iterations in 'sisters' mode, using the rules file provided in Dataset S8, followed by five iterations of 'subsisters' rules on a select number of taxa (Dataset S9). Next, we ran two separate iterations of the 'clade' mode, the first one to remove only the ciliate parasites of Parabasalids (Ex_pa) that occurred when transcriptome data were generated from co-contaminated rumen ciliates, and the second one to remove sequences from all other well-sampled taxa (see Datasets S10-S12 for rules, and SI Appendix for details).

To build species trees, we used two methods: Asteroid (56) and the concatenation option included in PhyloToL. We built species trees using all GFs, as well as a selection of GFs in which we do not detect evidence of non-vertical inheritance in photosynthetic lineages (i.e. EGT). We first used the utility CladeGrabbing.py to determine EGTs from primary endosymbiosis by identifying trees with clades of photosynthetic eukaryotes nested in cyanobacteria. Second, we captured trees affected by secondary/tertiary EGT by adjusting the rules for clade grabbing within the contamination loop (see SI Appendix for details; Dataset S12). At each step of the process, we made species trees with orthologs chosen and concatenated using the methods built into PhyloToL part two; trees were built with IqTREE (-m LG+G, Datasets S18-S20). We also used Asteroid (56) to build super trees with trees generated by PhyloToL, at each step of the contamination loop.

Data and Software Availability

The main PhyloToL pipeline and accompanying scripts, including all scripts used for this study, are available on GitHub (https://github.com/Katzlab/PhyloToL-6) and Zenodo (DOI:10.5281/zenodo.13323372). All results and outputs generated by this study, including Datasets 1 to 25 listed in the manuscript, are available on Figshare (https://doi.org/10.6084/m9.figshare.26527984).

Acknowledgements

We are grateful to Elinor Sterner and Julian Hernandez (Smith College) and Xyrus Maurer-Alcalá (AMNH) for their help in early stages of this study. Recent support for LAK has come from two NSF awards (DEB-2230391, OCE-1924570) as well as funding from the NIH (R15HG01040).

References

1. S. M. Adl, *et al.*, Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J. Eukaryot. Microbiol.* **66**, 4–119 (2019).
2. F. Burki, A. J. Roger, M. W. Brown, A. G. B. Simpson, The New Tree of Eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).
3. A. Collens, L. A. Katz, OPINION: Genetic conflict with mobile elements drives eukaryotic genome evolution, and perhaps also eukaryogenesis. *J Hered.* **112**, 140–144 (2021).
4. L. W. Parfrey, D. J. G. Lahr, L. A. Katz, The dynamic nature of eukaryotic genomes. *Mol. Biol. Evol.* **25**, 787–794 (2008).
5. L. Eme, *et al.*, Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
6. L. A. Hug, *et al.*, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
7. P. J. Keeling, F. Burki, Progress towards the Tree of Eukaryotes. *Curr. Biol.* **29**, R808–R817 (2019).
8. F. M. Celebi, S. Chou, E. McGeever, A. H. Patton, R. York, NovelTree: Highly parallelized phylogenomic inference. *Arcadia Sci.* (2023). https://doi.org/10.57844/arcadia-z08x-v798.
9. M. Drori, *et al.*, OneTwoTree: An online tool for phylogeny reconstruction. *Mol. Ecol. Resour.* **18**, 1492–1499 (2018).
10. W. A. Freyman, SUMAC: Constructing Phylogenetic Supermatrices and Assessing Partially Decisive Taxon Coverage. *Evol. Bioinforma.* **11**, EBO.S35384 (2015).
11. M. J. Sanderson, D. Boss, D. Chen, K. A. Cranston, A. Wehe, The PhyLoTA browser: Processing GenBank for molecular phylogenetics research. *Syst. Biol.* **57**, 335–346 (2008).
12. D. A. Benson, *et al.*, GenBank. *Nucleic Acids Res.* **45**, D37–D42 (2017).
13. A. Bateman, *et al.*, Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27**, 260–262 (1999).
14. D. Kuznetsov, *et al.*, OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
15. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
16. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
17. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
18. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
19. W. R. Pearson, An introduction to sequence similarity (“homology”) searching. *Curr. Protoc. Bioinforma.* **Chapter 3**, 3.1.1-3.1.8 (2013).
20. G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
21. G. Tan, *et al.*, Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst. Biol.* **64**, 778–791 (2015).
22. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
23. F. Sievers, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
24. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1–19 (2004).
25. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large

phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
26. B. Q. Minh, *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).
27. A. de Queiroz, J. Gatesy, The supermatrix approach to systematics. *Trends Ecol Evol* **22**, 34–41 (2007).
28. M. J. Sanderson, A. Purvis, C. Henze, Phylogenetic supertrees: Assembling the trees of life. *Trends Ecol. Evol.* **13**, 105–109 (1998).
29. C. E. Hinchliff, *et al.*, Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U A* (2015). https://doi.org/10.1073/pnas.1423041112.
30. S. Escobar, A. J. Helmstetter, R. Montufar, T. L. P. Couvreur, H. Balslev, Phylogenomic relationships and historical biogeography in the South American vegetable ivory palms (Phytelepheae). *Mol. Phylogenet. Evol.* **166**, 107314 (2022).
31. R. T. Kimball, *et al.*, A Phylogenomic Supertree of Birds. *Divers.-BASEL* **11**, 109 (2019).
32. T. Li, *et al.*, Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. *Sci. Rep.* **10**, 22366 (2020).
33. P. Simion, F. Delsuc, H. Philippe, “To What Extent Current Limits of Phylogenomics Can Be Overcome?” in *Phylogenetics in the Genomic Era*, C. Scornavacca, F. Delsuc, N. Galtier, Eds. (No commercial publisher | Authors open access book, 2020), p. 2.1:1-2.1:34.
34. E. Susko, A. J. Roger, Long Branch Attraction Biases in Phylogenetics. *Syst. Biol.* **70**, 838–843 (2021).
35. J. L. Steenwyk, Y. Li, X. Zhou, X.-X. Shen, A. Rokas, Incongruence in the phylogenomics era. *Nat. Rev. Genet.* 1–17 (2023). https://doi.org/10.1038/s41576-023-00620-x.
36. A. Cote-L’Heureux, X. X. Maurer-Alcalá, L. A. Katz, Old genes in new places: A taxon-rich analysis of interdomain lateral gene transfer events. *PLOS Genet.* **18**, e1010239 (2022).
37. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).
38. T. A. Heath, S. M. Hedtke, D. M. Hillis, Taxon sampling and the accuracy of phylogenetic analyses. *J. Syst. Evol.* **46**, 239–257 (2008).
39. D. J. Zwickl, D. M. Hillis, Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* **51**, 588–598 (2002).
40. A. R. Nabhan, I. N. Sarkar, The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13**, 122–134 (2012).
41. H. Philippe, *et al.*, Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLOS Biol.* **9**, e1000602 (2011).
42. B. Roure, D. Baurain, H. Philippe, Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol. Biol. Evol.* **30**, 197–214 (2013).
43. J. J. Wiens, Missing data and the design of phylogenetic analyses. *J. Biomed. Inform.* **39**, 34–42 (2006).
44. F. Burki, *et al.*, Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc Biol Sci* **283** (2016).
45. C. Al Jewari, S. L. Baldauf, An excavate root for the eukaryote tree of life. *Sci. Adv.* **9**, eade4973 (2023).
46. M. A. Cerón-Romero, M. M. Fonseca, L. de Oliveira Martins, D. Posada, L. A. Katz, Phylogenomic analyses of 2,786 Genes in 158 Lineages support a root of the eukaryotic tree of life between opisthokonts and all other lineages. *Genome Biol. Evol.* **14**, evac119 (2022).
47. J. F. H. Strassert, M. Jamy, A. P. Mylnikov, D. V. Tikhonenkov, F. Burki, New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* **36**, 757–765 (2019).
48. Y. Eglit, *et al.*, Meteora sporadica, a protist with incredible cell architecture, is related to

Hemimastigophora. *Curr. Biol.* **34**, 451-459.e6 (2024).
49. A. Yabuki, Y. Gyaltshen, A. A. Heiss, K. Fujikura, E. Kim, *Ophirina amphinema* n. gen., n. sp., a New Deeply Branching Discobid with Phylogenetic Affinity to Jakobids. *Sci. Rep.* **8**, 16219 (2018).
50. M. A. Cerón-Romero, *et al.*, PhyloToL: A taxon/gene-rich phylogenomic pipeline to explore genome evolution of diverse eukaryotes. *Mol Biol Evol* **36**, 1831–1842 (2019).
51. J. R. Grant, L. A. Katz, Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. *PLoS Curr.* **6** (2014).
52. O. Penn, *et al.*, GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* **38**, W23–W28 (2010).
53. I. Sela, H. Ashkenazy, K. Katoh, T. Pupko, GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* **43**, W7-14 (2015).
54. F. Chen, A. J. Mackey, C. J. Stoeckert, D. S. Roos, OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–D368 (2006).
55. N. Corradi, Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. *Annu. Rev. Microbiol.* **69**, 167–183 (2015).
56. B. Morel, T. A. Williams, A. Stamatakis, Asteroid: a new algorithm to infer species trees from gene trees under high proportions of missing data. *Bioinformatics* **39**, btac832 (2023).
57. S. J. Sibbald, J. M. Archibald, Genomic Insights into Plastid Evolution. *Genome Biol. Evol.* **12**, 978–990 (2020).
58. R. P. Hirt, *et al.*, Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 580–585 (1999).
59. P. J. Keeling, N. M. Fast, Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* **56**, 93–116 (2002).
60. M. W. Brown, *et al.*, Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol Evol* **10**, 427–433 (2018).
61. G. Lax, A. Cho, P. J. Keeling, Phylogenomics of novel ploeotid taxa contribute to the backbone of the euglenid tree. *J. Eukaryot. Microbiol.* **70**, e12973 (2023).
62. H. Li, TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006).
63. M. E. Schön, *et al.*, Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nat. Commun.* **12**, 6651 (2021).
64. L. W. Parfrey, *et al.*, Evaluating support for the current classification of eukaryotic diversity. *PLoS Genet* **2**, e220 (2006).
65. M. K. N. Lawniczak, *et al.*, Standards recommendations for the Earth BioGenome Project. *Proc. Natl. Acad. Sci.* **119**, e2115639118 (2022).
66. A. Whibley, J. L. Kelley, S. R. Narum, The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. *Mol. Ecol. Resour.* **21**, 641–652 (2021).
67. D. Bolchini, A. Finkelstein, V. Perrone, S. Nagl, Better bioinformatics through usability analysis. *Bioinformatics* **25**, 406–412 (2009).
68. G. Lax, *et al.*, Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).
69. K. Krampis, Democratizing Bioinformatics through easily Accessible Software Platforms for Non-Experts in the Field. *BioTechniques* **72**, 36–38 (2022).
70. B. Lawlor, R. D. Sleator, The democratization of bioinformatics: A software engineering perspective. *GigaScience* **9**, giaa063 (2020).
71. M. Costello, *et al.*, Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).

72. L. E. MacConaill, *et al.*, Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 30 (2018).
73. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
74. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *Plos One* **5**, e9490 (2010).

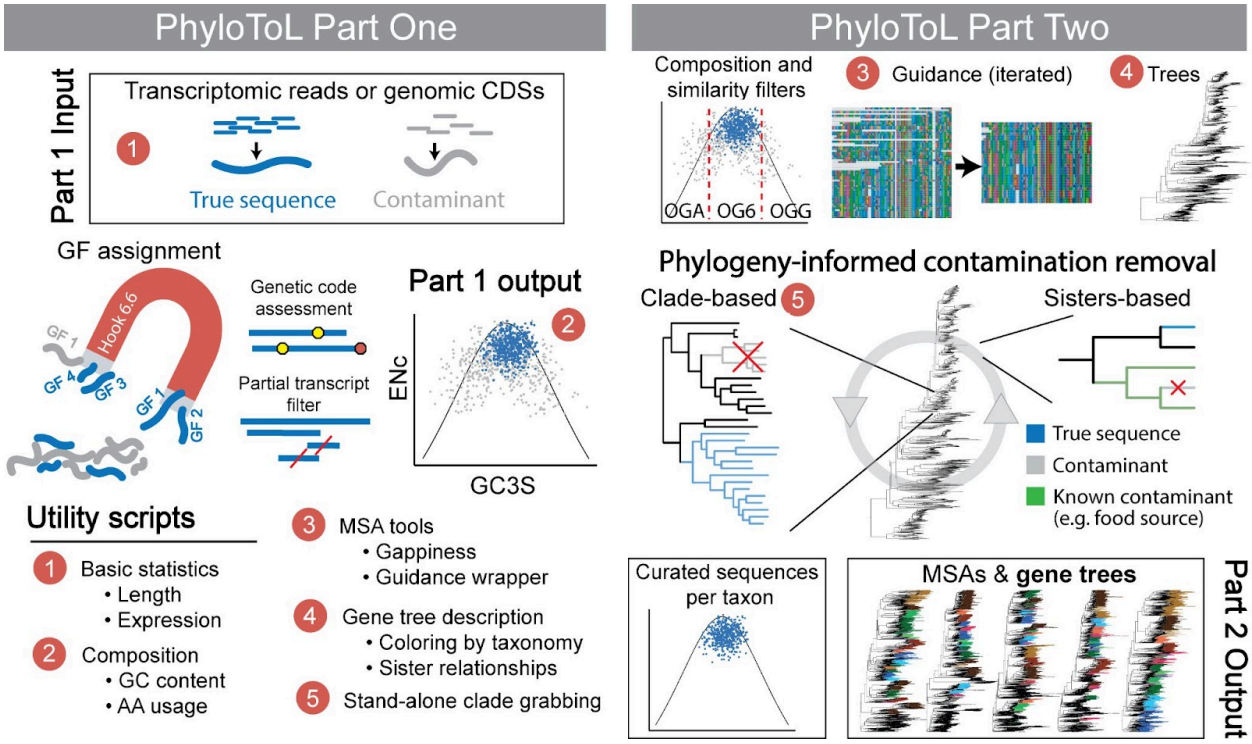

Fig. 1. A schematic of the PhyloToL 6 core pipeline. Phylotol comprises two main components, Part One and Part Two. Part One is primarily intended to apply preliminary filtration steps and assign gene families using a reference database. This reference database can be the Hook version 6.6 as described in the main text, or a custom database. Part One takes as input assembled transcripts or genomic CDS, and is able to flexibly handle a variety of genetic codes. In the graph under the “Part 1 Output” heading we show using silent-site GC content (GC3S) vs. the effective number of codons (ENc) that the true sequences from the sequenced sample (blue) tend to have similar composition, with contaminant sequences (gray) lying outside of this range. Part Two builds MSAs by iterating Guidance multiple (by default 5) times for rigorous homology assessment of each gene family, and then builds gene trees. We present a novel phylogeny-informed approach to contamination removal, where contamination is remove from trees in an iterative fashion, either by keeping only sequences in robust clades (“Clade-based”) or removing sequences sister next to known contaminants (“Sisters-based”). We also exemplify the suite of utility tools accompanying this core pipeline, identified by numbers (red circles) where the tool can be applied.

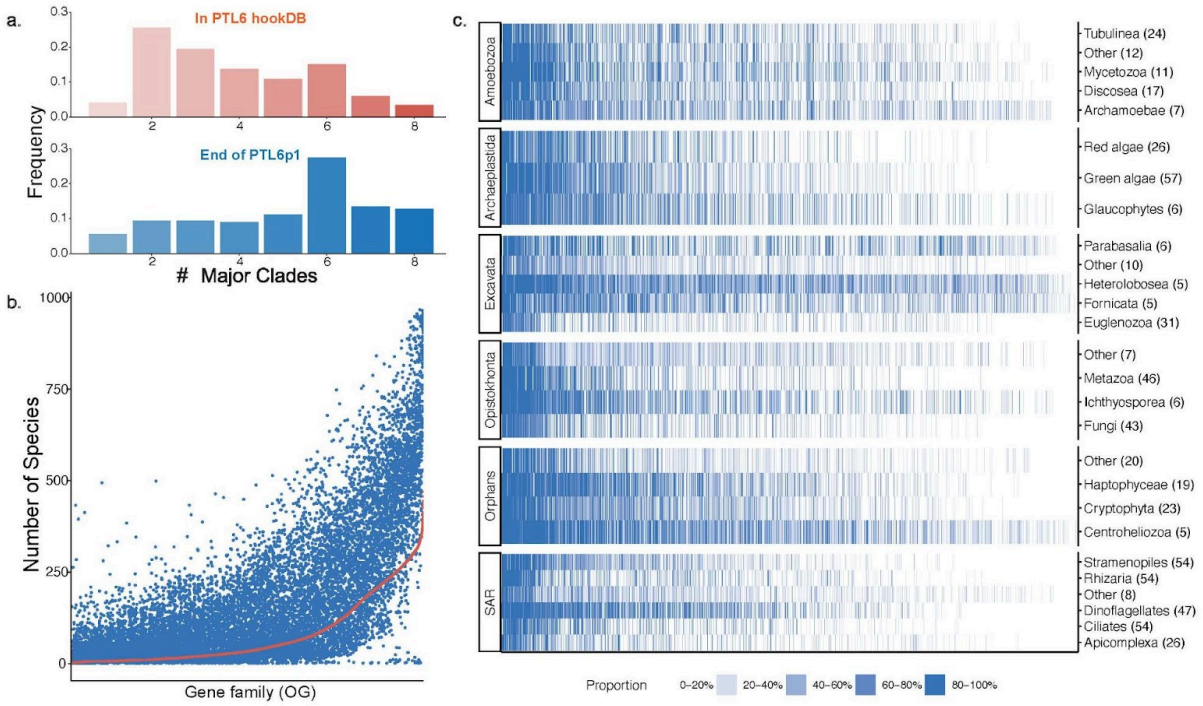

Fig. 2. The Hook reference database, which is used in PhyloToL 6 part 1, effectively captures taxonomic diversity in OG assignment. While most OGs as represented in the Hook database are only present in 2-4 major clades, once assigned to our more diverse dataset, most OGs are present in 5-8 major clades, with a mode of 6 (the number of eukaryotic major clades) (a). The number of species with an OG in the ReadyToGo files (blue) correlates with the number of species with that OG in the Hook (red), with very few OGs losing diversity (b). Panel (c) describes the proportion of species (intensity of color) in each eukaryotic minor clade (rows) in which each OG (columns) is found. OGs found in more species are on the left, and those with fewer species are on the right.

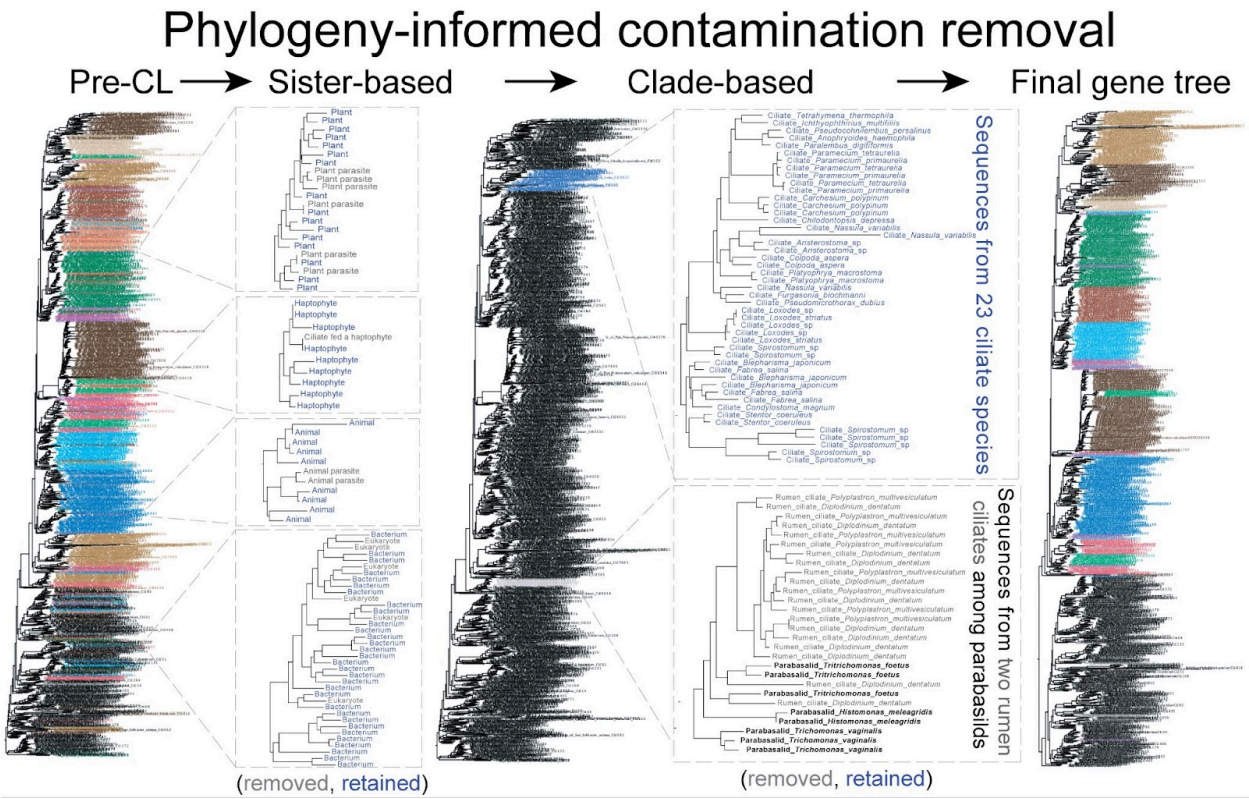

Fig. 3. A cartoon depicting phylogeny-informed contamination removal, which is a component of PhyloToL 6 part 2. Users can use the contamination loop to iteratively remove sequences based on their sister species in single-gene trees. Depicted here are sequences (gray) that are either from hosts in analyses of parasites (upper left) or bacterial sequences that come as contaminants in analyses of eukaryotic transcriptomes (lower left). In a second method of contamination removal, users can ‘grab’ (retain) sequences falling in monophyletic clades that meet user-specified robustness criteria (e.g., minimum target group species count and maximum number of non-group species). In the case depicted here, we identified substantial contamination of a subset of ciliate transcriptomes by parabasalids with which the ciliate species are known to share an environment (cow rumen). To remove this ciliate contamination, we used PhyloToL to retain only ciliate sequences falling in clades with at least 12. For clade-based contamination removal, an example of a retained clade is given in blue, and a removed clade in gray in the bottom right.

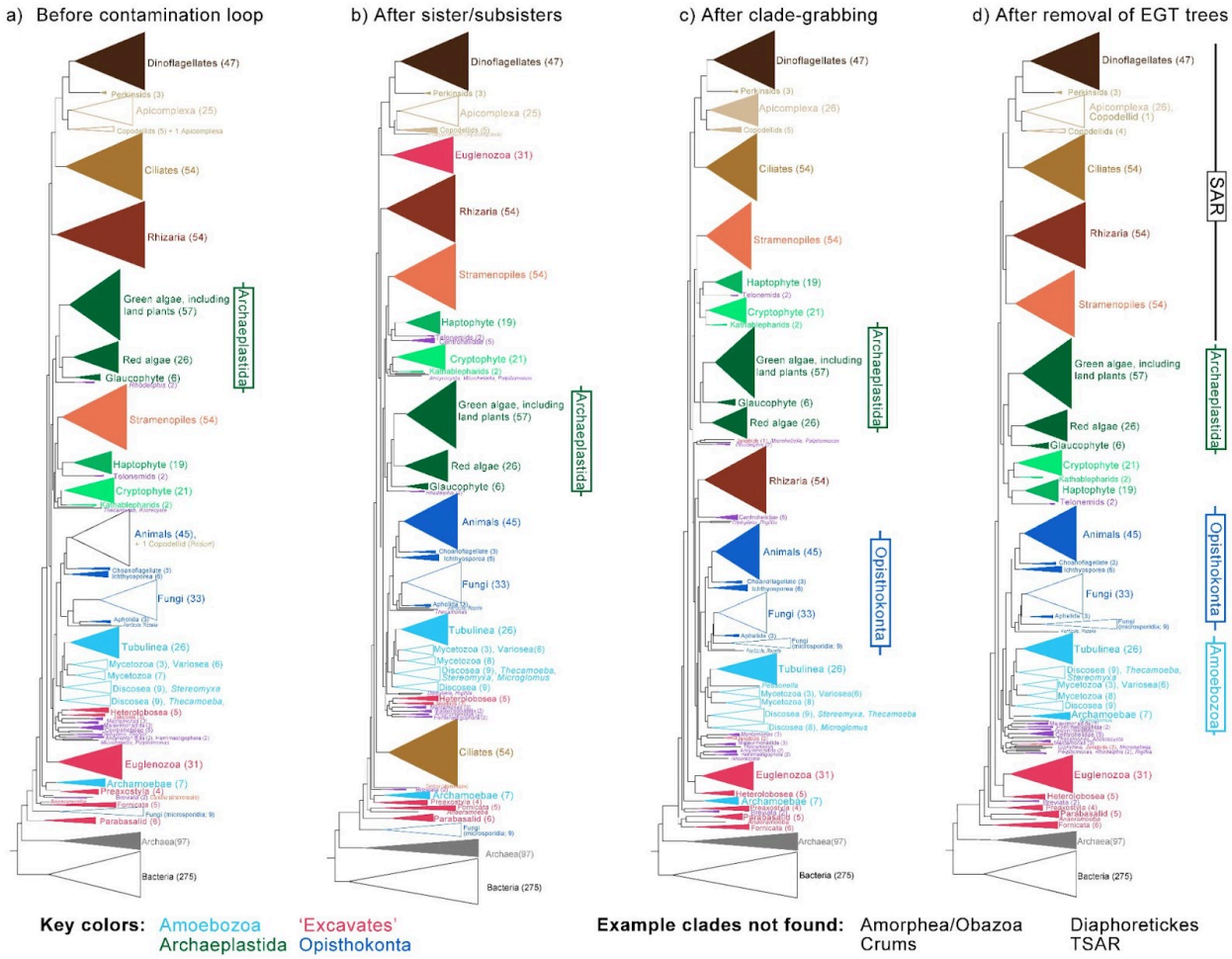

Fig. 4: Concatenated analyses at four stages in the contamination-removal process are generally concordant with published hypotheses as most morphological-defined clades (e.g. dinoflagellates, green algae, Euglenozoa, stramenopiles) and some clades supported by molecular data (Rhizaria, Archaeplastida) are recovered consistently. The four analyses are: a) before contamination loop, b) after removal of contaminants based on sister/subsister rules; c) after clade grabbing to keep 'best' sequences; and d) after removal of trees most affected by both primary and secondary endosymbiotic gene transfer (EGT). Notably, the monophyly of Opisthokonta emerges after clade grabbing (c) while the monophyly of Amoebozoa and SAR only appear after removing trees affected by EGT (d). Some "orphan" lineages (purple) are stable across trees (i.e. telonemids are always sister to haptophytes, breviates are always towards the root) while other lineages (e.g. Centrohelidae, Hemimastigophora) move position across trees; this likely reflects a combination of the effects of missing data and a lack of close relatives.