PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes

Mario A. Cerón-Romero,[1,2] Xyrus X. Maurer-Alcalá,[†,1,2] Jean-David Grattepanche,[‡,1] Ying Yan,[1] Miguel M. Fonseca,[3] and L.A. Katz*[,1,2]

[1]Department of Biological Sciences, Smith College, Northampton, MA

[2]Program in Organismic and Evolutionary Biology, University of Massachusetts Amherst, Amherst, MA

[3]CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal

[†]Present address: Institute of Cell Biology, University of Bern, Bern, Switzerland

[‡]Present address: Biology Department, Temple University, Philadelphia, PA

***Corresponding author:** E-mail: lkatz@smith.edu.

Associate editor: Fabia Ursula Battistuzz

Abstract

Estimating multiple sequence alignments (MSAs) and inferring phylogenies are essential for many aspects of comparative biology. Yet, many bioinformatics tools for such analyses have focused on specific clades, with greatest attention paid to plants, animals, and fungi. The rapid increase in high-throughput sequencing (HTS) data from diverse lineages now provides opportunities to estimate evolutionary relationships and gene family evolution across the eukaryotic tree of life. At the same time, these types of data are known to be error-prone (e.g., substitutions, contamination). To address these opportunities and challenges, we have refined a phylogenomic pipeline, now named PhyloToL, to allow easy incorporation of data from HTS studies, to automate production of both MSAs and gene trees, and to identify and remove contaminants. PhyloToL is designed for phylogenomic analyses of diverse lineages across the tree of life (i.e., at scales of >100 My). We demonstrate the power of PhyloToL by assessing stop codon usage in Ciliophora, identifying contamination in a taxon- and gene-rich database and exploring the evolutionary history of chromosomes in the kinetoplastid parasite *Trypanosoma brucei*, the causative agent of African sleeping sickness. Benchmarking PhyloToL's homology assessment against that of OrthoMCL and a published paper on superfamilies of bacterial and eukaryotic organellar outer membrane pore-forming proteins demonstrates the power of our approach for determining gene family membership and inferring gene trees. PhyloToL is highly flexible and allows users to easily explore HTS data, test hypotheses about phylogeny and gene family evolution and combine outputs with third-party tools (e.g., PhyloChromoMap, iGTP).

Key words: **phylogenomic pipeline, high-throughput sequencing data, contamination removal, genome evolution, chromosome mapping.**

Introduction

An important way to study biodiversity is through phylogenomics, which uses the generation of multiple sequence alignments (MSAs), gene trees, and species trees (e.g., Katz and Grant 2015; Hug et al. 2016). During the last two decades, advances in DNA sequencing technology (e.g., 454, Illumina, Nanopore, and PacBio) have led to the rapid accumulation of data (transcriptomes and genomes) from diverse lineages across the tree of life, greatly expanding the opportunities for phylogenomic studies (e.g., Katz and Grant 2015; Burki et al. 2016; Brown et al. 2018; Heiss et al. 2018). Such approaches are powerful by using increasingly large molecular data sets to reduce the discordance between gene and species trees. Indeed, studies relying on a small number of genes are often impacted by lateral gene transfer, gene duplication and loss, and incomplete lineage sorting (e.g., Maddison 1997; Tremblay-Savard and Swenson 2012; Mallo and Posada 2016). Large-scale phylogenomic analyses allow for the exploration of deep evolutionary relationships (e.g., dos Reis et al. 2012; Wickett et al. 2014; Katz and Grant 2015; Hug et al. 2016), but such analyses require data-intensive computing methods. As a result, numerous laboratories have developed custom phylogenomic pipelines proposing different methods to efficiently process and analyze massive gene and taxon databases (Sanderson et al. 2008; Wu and Eisen 2008; Smith et al. 2009; Kumar et al. 2015).

In general, phylogenomic pipelines are composed of three steps: 1) collection of homologous gene data sets from various input sources (e.g., whole genome sequencing, transcriptome analyses, PCR-based studies), 2) production of MSAs, and 3) generation of gene trees and sometimes a species tree. Phylogenomic pipelines typically put more effort in the first two steps (collecting homologous genes and MSA curation) to ensure a more accurate tree inference. For instance, pipelines such as PhyLoTA (e.g., Sanderson et al. 2008) and BIR (Kumar et al. 2015) focus on the identification and collection of homologous genes from public databases such as GenBank (Benson et al. 2017). However, pipelines such as AMPHORA

Article

(Wu and Eisen 2008) and Mega-phylogeny (Smith et al. 2009) focus on the construction and refinement of robust alignments rather than the collection of homologs. A recently published tool, SUPERSMART (Antonelli et al. 2017), incorporates more efficient methods for data mining than PhyLoTA (Sanderson et al. 2008). SUPERSMART includes methods for tree inference using a multilocus coalescent model, which benefits biogeographical analyses. Although these pipelines incorporate sophisticated methods for data mining, alignment, and tree inference, a major issue is that they are optimized for either a relatively narrow taxonomic sampling (e.g., plants) or relatively narrow sets of conserved genes/gene markers.

A major problem for phylogenomic analyses using public sequence data, including GenBank and EMBL (Baker et al. 2000), is the inherent difficulty in identifying and removing annotation errors and contamination (e.g., data from food sources, symbionts, or organelles). Additional errors are introduced when nonprotein coding regions (e.g., pseudogenes, promoters, and repeats) are inferred as open reading frames (ORFs) by gene-prediction tools such as GENESCAN (Burge and Karlin 1997), SNAP (Korf 2004), AUGUSTUS (Stanke and Morgenstern 2005), and MAKER (Cantarel et al. 2008). Similarly, some public databases are more prone to annotation errors than others depending on how much effort they invest in manual curation of public submissions. For instance, data from GenBank NR, TrEMBL (Bairoch and Apweiler 2000), and KEGG (Kanehisa and Goto 2000) may have very high rates of these kind of errors, whereas curated resources like Gene Ontology (Ashburner et al. 2000) and SwissProt (Bairoch and Apweiler 2000) are more likely to have low to moderate rates of such errors (Schnoes et al. 2009). The misidentification errors in these databases often stem from problems surrounding accurate taxonomic identification of sequences from high-throughput sequencing (HTS) data sets, as contamination by other taxa can be frequent, particularly of organisms that cannot be cultured axenically (Shrestha et al. 2013; Lusk 2014; Parks et al. 2015). Hence, a crucial element of any phylogenomic pipeline that relies on public databases is the ability to identify and exclude annotation errors and contaminants from its analyses.

At the same time, the availability of curated databases and third-party tools provide considerable power and efficiency for phylogenomic analyses. We rely on OrthoMCL, a database generated initially to support analyses of the genome of *Plasmodium falciparum* and other apicomplexan parasites (Li et al. 2003; Chen et al. 2006), for the initial identification of homologous gene families (GFs). We also incorporate GUIDANCE V2.02 (Penn et al. 2010; Sela et al. 2015) for assigning statistical confidence MSA scores using guide-trees and a bootstrap approach. GUIDANCE allows an efficient identification and removal of potentially nonhomologous sequences (i.e., sequences having very low scoring values) and unreliably aligned columns and residues under various parameters (Privman et al. 2012; Hall 2013; Vasilakis et al. 2013). This flexibility is critical—whereas concepts such as homology and paralogy have clear definitions in textbooks, when it comes to deploying phylogenomic tools on inferences at the scale of >100 My, they become working definitions that depend on parameters and sampling of both genes and taxa. Finally, we have chosen RAxML V8 (Stamatakis et al. 2005; Stamatakis 2014) for tree inference as its efficient algorithms allow for robust estimation of maximum likelihood trees [though users can aexport the MSAs from our pipeline for analyses with other software].

Our original phylogenomic pipeline aimed to explore the eukaryotic tree of life using multigene sequences available in GenBank from diverse taxa (Grant and Katz 2014a; Katz and Grant 2015). This first version generated a collection of ~13,000 GFs from ~800 species distributed among Eukaryota, Bacteria, and Archaea, and included a suite of methods to process gene alignments and trees. The 800 species were a subset of available taxa, picked to represent, more or less evenly, the main eukaryotic lineages with no more than two species per genus. Moreover, although the focus was on eukaryotes, bacteria and archaea were also included to allow detection of contamination, lateral gene transfer events, and/or for exploring phylogenetic relationships that include all cellular life. GFs originally defined by OrthoMCL were used as seeds to search more homologous sequences from additional taxa. Then, the enriched GFs pass through an additional quality-check step that re-evaluates homology. This step includes applying a combination of methods that can remove alleles, non-homologous genes, and highly divergent sequences based on pairwise comparisons with Needle (Rice et al. 2000), with robust alignments produced with MAFFT (Katoh and Standley 2013) that were then filtered with GUIDANCE. These refined high-quality MSAs were used to produce gene trees with RAxML. An additional option was to identify orthologs based on their position in gene trees, which can be used to generate concatenated alignments for species tree inference (see Grant and Katz 2014a).

This new version, which we name PhyloToL (Phylogenomic Tree of Life), incorporates significant improvements over Grant and Katz (2014a), including a more efficient method to capture HTS data, a more robust homology detection approach, a novel tree-based method for contamination removal, and substantially more efficient scripts and improved databases. PhyloToL contains a database of 13, 103 GFs that include up to 627 eukaryotes (58 generated in our lab), 312 bacteria and 128 archaea. Here, we describe our updated approaches, providing examples of stop codon usage assessment in Ciliophora and detection of contamination produced by many HTS studies (including our own). We also illustrate the potential of PhyloToL by depicting the evolutionary history of the genes on the chromosomes of the human parasite *Trypanosoma brucei*, the causative agent of African sleeping sickness, and we also benchmark against published studies.

New Approaches

PhyloToL (https://github.com/Katzlab/PhyloTOL; last accessed May 2, 2019) is divided into four major components: 1) gene family assessment per taxon, 2) refinement of homologs and gene tree reconstruction, 3) tree-based

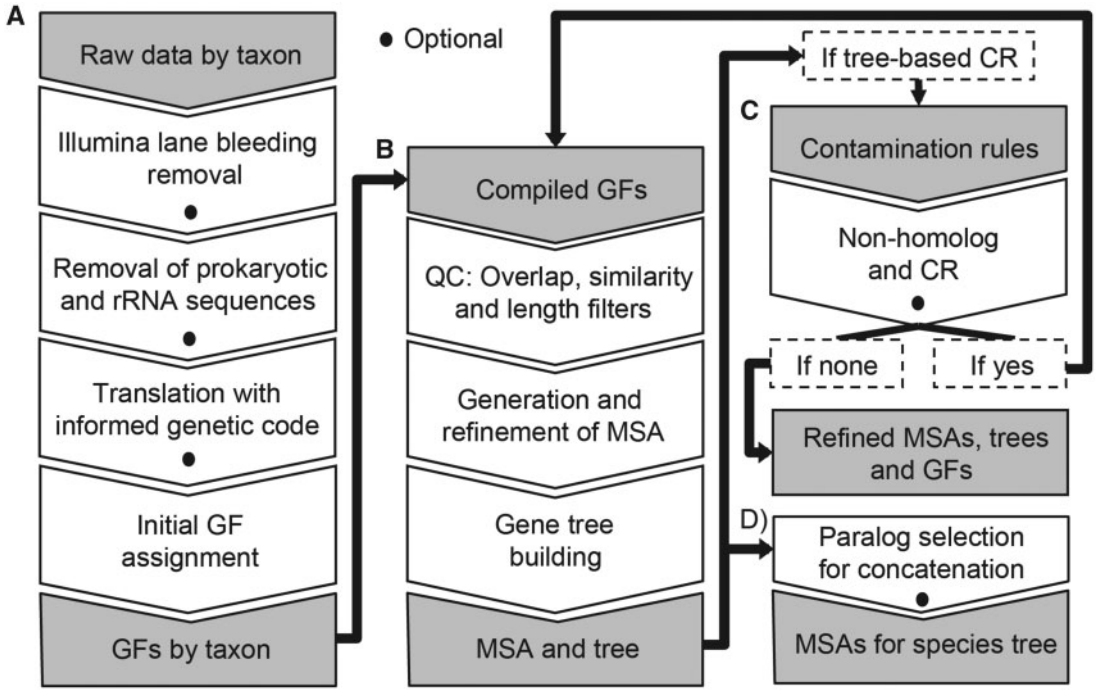

FIG. 1. The four components of PhyloToL: A) gene family assessment per taxon, B) refinement of homologs and gene tree reconstruction, C) tree-based contamination removal, and D) generation of a supermatrix for species tree inference. GF = Gene Family, QC = Quality Control, CR = Contamination Removal. (*A*) The first component processes and classifies raw data from different sources (e.g., transcriptomes, genomes, and protein data) into a collection of GFs. In the initial step, transcriptomes produced in-lab are processed to identify and remove sample bleeding (Mitra et al. 2015) in an Illumina lane (cross-contamination). Then, prokaryotic sequences and rRNA sequences are removed from transcriptomes. Finally, transcriptomic and genomic sequences are translated using informed genetic codes. (*B*) The second component compiles all GFs by taxon in the gene family database, refines an MSA, and produces a phylogenetic tree for each gene family. (*C*) The third component (optional) detects contaminant sequences using gene trees and predefined contamination rules, and also detects nonhomologous sequences after the MSA refinement process. Contaminants and nonhomologs are identified and removed from the gene family database iteratively. (*D*) The fourth component (optional) identifies orthologous sequences using a tree-based approach for removing paralogs. Alignments of orthologs can be concatenated to produce a species tree.

contamination removal, and 4) generation of a supermatrix for species tree inference (i.e., concatenation). The first component starts with data from either public databases or those generated by the users' 'omics projects, and categorizes sequences into a collection of candidate GFs. This part of PhyloToL includes steps for removing bacterial contamination (given our focus on eukaryotes) and translating sequences using the most appropriate inferred genetic code (fig. 1*A*). The second component includes a series of steps to assess homology in the candidate GFs based on sequence similarity, sequence overlap, and refinement of MSAs prior to reconstructing phylogenies (fig. 1*B*). The third component includes a novel method that iterates the second component (refinement of homologs and gene tree reconstruction) to remove contamination inferred from phylogenetic trees (fig. 1*C*), which is critical given the high frequency of contamination in many HTS data sets. Although the combination of methods in the first three components identifies homologs within GFs (see Materials and Methods section), the distinction between paralogous and orthologous sequences occurs only in the optional fourth component. This component detects orthologous sequences based on their position in phylogenetic trees and concatenates them into a supermatrix for species tree inference (fig. 1*D*); this last component has not been modified since the last published version of the pipeline (Grant and Katz 2014a, 2014b; Katz and Grant 2015), and users can explore other tools for concatenation (e.g., Leigh et al. 2008; Narechania et al. 2012; Drori et al. 2018; Vinuesa et al. 2018) using the single gene MSAs generated by PhyloToL, or alternative methods to concatenation (e.g., iGTP, Chaudhary et al. 2010; *Guenomu*, De Oliveira Martins et al. 2016) using outputs from the third component.

Additional to the primary goal of PhyloToL, which was reconstructing the evolutionary history of eukaryotes, this new version emphasizes the flexibility to allow studies of GFs evolution as well as phylogenomics with varying parameters and taxon/gene inclusion. Though there are many other tools out there for phylogenomic analyses (e.g., OneTwoTree [Drori et al. 2018], SUPERSMART [Antonelli et al. 2017], and PhyLoTA [Sanderson et al. 2008]), we believe PhyloToL is distinctive because of its combination of: 1) inclusion of both database and user-inputted data; 2) focus on broad taxon inclusion for "deep" events (e.g., $\geq$100 My); and 3) flexibility for exploration of multiple hypotheses and parameters (supplementary table S1, Supplementary Material online).

Results and Discussion

The overall structure of PhyloToL was improved over Grant and Katz (2014a) by dividing the pipeline into four major components (fig. 1) allowing different modes to execute these components depending on the type of study. PhyloToL also includes new methods to incorporate data from more sources (in component 1, fig. 1A), refine MSAs from GFs (in component 2, fig. 1B), and to remove contaminant sequences (in component 3, fig. 1C). Here, we explain the improvements on the overall structure of PhyloToL and benchmark the performance of new methods by analyses of ancient GFs.

Pipeline Structure

Although PhyloToL is designed for phylogenomic analyses of diverse lineages across the tree of life, it can also be deployed in different ways for a variety of purposes such as phylogenomic chromosome mapping (Cerón-Romero et al. 2018), gene discovery, or metatranscriptome analyses. For instance, the GF assessment per taxon, refinement of GFs and gene tree reconstruction (i.e., first and second components of PhyloToL) can be run independently, and the tree-based contamination removal and generation of a supermatrix (third and fourth components) are optional. Moreover, the user can also run the second component in two alternative modes: 1) only quality control (QC) for GFs and 2) without generating a gene tree. Running the second component of PhyloToL only for QC for GFs is helpful when the primary aim is to collect sequences for candidate GFs (QC involves filtering sequences by length, overlap and similarity, see Materials and Methods section) or for exploring taxonomic diversity within each gene family. Likewise, running the second component of PhyloToL without generating gene trees is useful for inspecting regions of homology (motif searching), trying alternative methodologies (i.e., those other than RAxML V8, which is incorporated into PhyloToL) for phylogenetic tree inference and to simply create a curated database of aligned homologous proteins (i.e., having sequences with divergence levels above the defined threshold removed by GUIDANCE). Our approach for determining homology is through generation of MSAs using GUIDANCE V2.02 (Penn et al. 2010; Sela et al. 2015) with sequence and column cutoff 0.3 and 0.4, respectively, to determine which sequences meet criteria for retention. These GUIDANCE parameters were chosen based on inspection of early runs of our data because the default parameters in GUIDANCE are geared for shallower levels of diversity and tend to exclude many of our focal taxa. Indeed, GUIDANCE scores are alignment dependent and so cutoffs are empirically defined. As described in our manual (Supplementary Material online) users can change these parameters for their own data sets to explore homology more deeply.

Performance of PhyloToL in GF Estimation per Taxon

To exemplify outputs of the first component of PhyloToL, GF assessment per taxon, we provide data from RNA-seq studies of the ciliates *Blepharisma japonicum* (MMETSP1395) and *Strombidium rassoulzadegani* (MMETSP0449_2). Each of these two data sets starts with >20,000 assembled transcripts, from which ~1% are contamination from rRNAs, bacterial and archaeal sequences that are removed (table 1). The final data sets after running through PhyloToL (only the GF assessment per taxon component) contain between 5,000 and 10,000 transcripts assigned to eukaryotic GFs and representing ~20% of the initial set of sequences (table 1). PhyloToL also allows us to assess that *B. japonicum* potentially uses the "*Blepharisma*" genetic code (i.e., UAR as stop codon, UGA is translated into tryptophan; Lozupone et al. 2001; Sugiura et al. 2012) and *S. rassoulzadegani* uses the "ciliate" genetic code (i.e., only use UGA as stop codon, and UAR is reassigned to glutamine; Caron and Meyer 1985).

Table 1. Summary of representative Gene Family Assessment per Taxon.

Sequences (contigs)	*Blepharisma japonicum*	*Strombidium rassoulzadegani*
Original assembly	**45,231**	**24,810**
Removed rRNA	**114**	**33**
Removed prokaryotic	**453**	**290**
Assigned to PhyloToL GF	**10,060**	**4,764**

Contigs not assinged to PhyloToL GFs are likely taxon-specific.

We evaluated the importance of PhyloToL's inspection of putative stop codons for these two taxa by also processing the transcriptomic data forcing translation with the universal and the "ciliate" genetic codes (fig. 2A). Here we found that when using PhyloToL's inferred alternative genetic code, transcripts were substantially longer than when forced to be processed with universal or ciliate genetic codes (fig. 2A), which suggests that using the carefully assessed genetic code allows the user to retrieve a larger proportion of each transcript.

Performance of PhyloToL in Tree-Based Contamination Removal

We then tested the third component of PhyloToL (i.e., tree-based contamination removal) using a data set of 152 GFs that includes up to 167 taxa distributed among eukaryotes, bacteria, and archaea (Supplementary Material online). To give the user a sense of the time involved, using a computer with 128 GB of RAM and 10 cores, the analyses took 86 h and 5 iterations of contamination removal. However, 79% of the contaminant sequences were removed in the first iteration, which took ~50% of the total time (fig. 2B).

Contaminant sequences detected often originated from food sources or endosymbionts (at least ~50% and ~40% of the total contaminants, respectively; Supplementary Material online). For instance, sequences from the amoeba *Neoparamoeba* are often nested within Euglenozoa (in 14 GFs; fig. 3A) likely because some of its data are actually from a (past or present) kinetoplastid endosymbiont as previously reported by Tanifuji et al. (2011). Likewise, sequences from the foraminifera *Sorites*, which hosts a dinoflagellate endosymbiont (Langer and Lipps 1995), are sometimes nested within dinoflagellate sequences (37 GFs; fig. 3B). However, sequences from the Katablepharid *Roombia truncata*

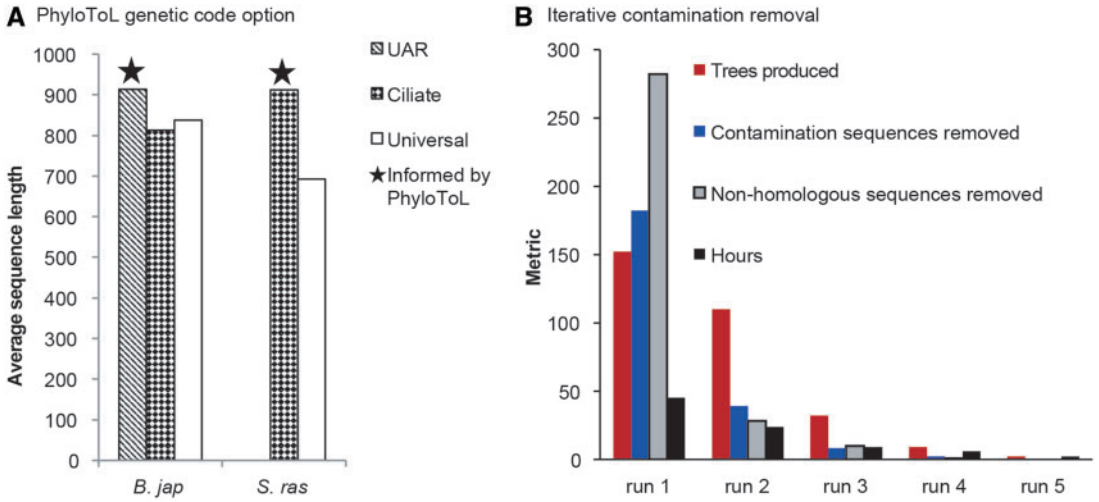

FIG. 2. Evaluation of performance of the first and second component of PhyloToL (fig. 1A and B). (A) Gene family assessment per taxon performance using the informed genetic code (indicated with a star) and the ciliate and universal genetic codes for the ciliates *Blepharisma japonicum* (B. jap) and *Strombidium rassoulzadegani* (S. ras). The length of the inferred sequences is higher when using the informed genetic code because it will not terminate the sequences at potentially reassigned in-frame stop codons. (B) Example of contamination removal using our test data set, containing 152 GFs with up to 167 taxa. Overall five iterations were required to remove all contaminant and nonhomologous sequences. with most of the sequence removal occurring during the first iteration.

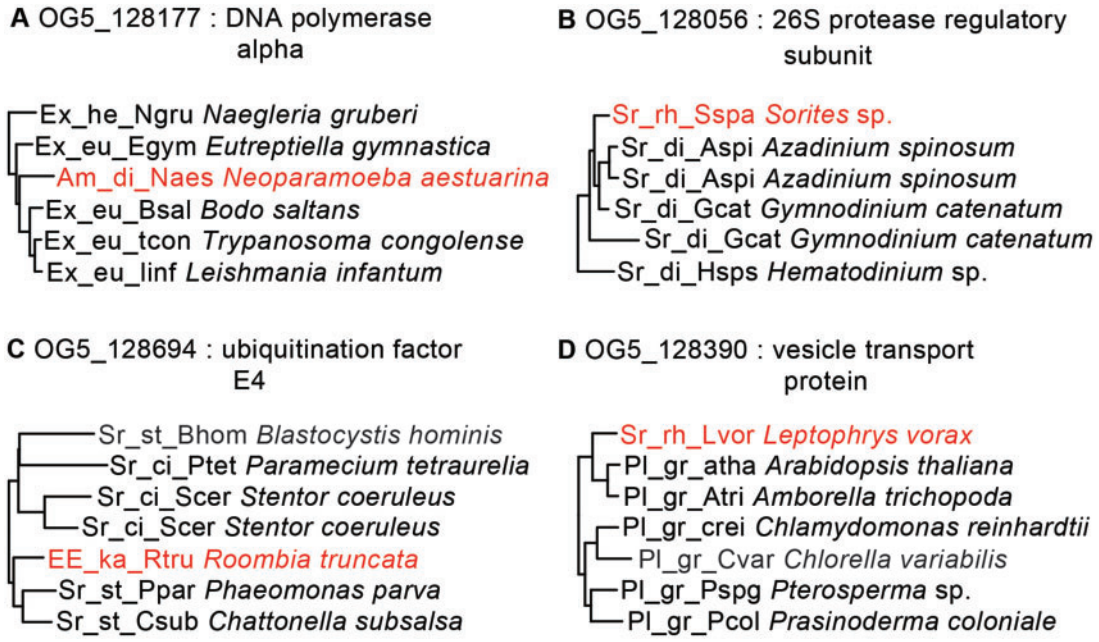

FIG. 3. Examples of contamination within gene trees that are used to define rules for the contamination removal loop of component 3 of PhyloToL (see fig. 1C). All sequences are named by major clade (Am = Amoebozoa, EE = everything else, Ex = Excavata, Pl = Archaeplastida, Sr = SAR), "minor" clade (di = Dinophyceae, he = Heterolobosea, eu = Euglenozoa, st = Stramenopile, ci = Ciliophora, ka = Katablepharidophyta, gr = green algae, rh = Rhizaria), and a four-digit code unique to each species (e.g., Ngru = *Naegleria gruberi*). (A) Possible case of contamination in *Neoparamoeba aestuarina* by an endosymbiotic excavate. (B) Possible case of contamination in *Sorites* by an endosymbiotic dinoflagellate. (C) Possible case of contamination from *Roombia truncata*'s diatom food source. (D) Possible case of contamination in *Leptophrys vorax* from its green algal food source. Each image is a small part of a much larger tree.

are sometimes nested among the SAR clade as sister to Stramenopila (in 3 GFs; fig. 3C); these sequences are potentially from diatoms that are used for feeding *R. truncata* (Okamoto et al. 2009). Finally, sequences from the Rhizaria *Leptophrys vorax*, which feeds on green algae, are often nested among green algal clades (38 GFs; fig. 3D).

Using the methods developed here, users can identify sources of contamination in individual taxa and then remove contaminating sequences in PhyloToL's contamination loop. This step is critical because sequence contamination is a common problem in HTS data of public databases (Merchant et al. 2014; Kryukov and Imanishi 2016). Indeed, previous

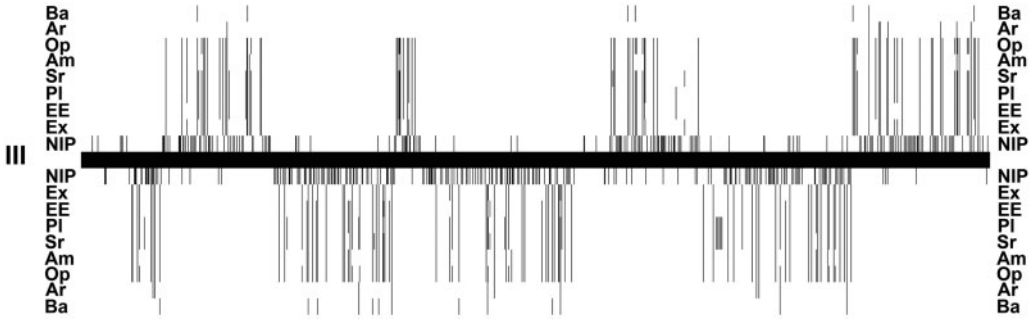

FIG. 4. Example of phylogenomic map of the polycistronic regions of chromosome III of *Trypanosoma brucei* generated by combining PhyloToL and PhyloChromoMap (Cerón-Romero et al. 2018). Horizontal line represent chromosome III of *T. brucei* and bars above/below reflect levels of conservation along each strand. First row from the bottom (NIP, "not in pipeline") indicates ORFs that do not match our criteria for tree inference (i.e., likely *Trypanosoma*-specific, highly divergent, and/or misannotated ORFs). The remaining rows (bottom to top) reflect the presence or absence of the gene in the major clades Excavata (Ex), orphans (EE, "everything else"), Archaeplastida (Pl), SAR (Sr), Amoebozoa (Am), Opisthokonta (Op), Archaea (Ar), and Bacteria (Ba). Genes are organized in polycistronic gene clusters (PGC) with variable gene density as described in Results and Discussion section.

studies have demonstrated that sequence contamination is one of the most important obstacles for evolutionary studies (Laurin-Lemay et al. 2012; Struck 2013; Philippe et al. 2017).

Implementation for Phylogenomic Chromosome Mapping

To exemplify an implementation of PhyloToL, we combined outputs with our tool PhyloChromoMap (Cerón-Romero et al. 2018) to explore the evolutionary history of chromosomes in the kinetoplastid parasite that causes African sleeping sickness, *Trypanosoma brucei gambiense* DAL972 (assembly ASM21029v1). Combining these tools, with PhyloChromoMap for mapping genes along each strand separately, we generated a map that displays the evolutionary history of 9,755 genes across both strands of the *T. brucei gambiense* chromosomes (fig. 4 and supplementary fig. S1, Supplementary Material online).

Previous studies have shown that karyotypes of kinetoplastid parasites have large syntenic polycistronic gene clusters (PGC), where genes are sequentially arranged on the same strand of DNA and expressed as multigene transcripts (Berriman et al. 2005; El-Sayed et al. 2005; Daniels et al. 2010; Martinez-Calvillo et al. 2010). We observed that almost all genes matching our GFs fall in PGCs and have a wide distribution throughout all 11 chromosomes, with variable gene density among chromosomes (fig. 4 and supplementary fig. S1, Supplementary Material online). Besides the presence of PGCs in *T. brucei*, previous studies proposed that large subtelomeric arrays of species-specific genes might serve as breakpoints for ectopic recombination in the nuclear membrane (Berriman et al. 2005; El-Sayed et al. 2005), a phenomenon that is also described in the apicomplexan parasite *P. falciparum* (Freitas-Junior et al. 2000; Scherf et al. 2001; Hernandez-Rivas et al. 2013; Cerón-Romero et al. 2018). However, whereas young and highly recombinant subtelomeric regions of at least 58 Mbp (up to 218 Mbp) are present in all *P. falciparum* chromosomes (Cerón-Romero et al. 2018), in *T. brucei gambiense* this pattern is only evident in chromosomes 3 and 9 (supplementary fig. S1, Supplementary Material online). This indicates that although ectopic recombination of subtelomeric regions can play a role in the karyotype evolution of *T. brucei*, it may not be as crucial to the success of this parasite as compared with *P. falciparum*.

We also explored the level of evolutionary conservation of genes in *T. brucei gambiense* based on their phylogenetic distribution as estimated by PhyloToL. Here, we detected that genes tend to be either very conserved or very divergent, with few genes of intermediate conservation (χ^2, $P < 0.05$; supplementary fig. S2, Supplementary Material online). About 73% of the published genes in the *T. brucei gambiense* DAL972 (assembly ASM21029v1) genome lacked homologs to any of our ancient GFs and thus may be *Trypanosoma-*

Table 2. Summary of Conservation of Genes in *Trypanosoma brucei*.

Description	Number of Genes[b]
Total in *T. brucei*	**9,755**
Recent (NIP): not in PhyloToL[a]	**7,125**
Older (IP): in PhyloToL[a]	**2,630**
Distribution	
Only in eukaryotes	
One major clade	**39**
Two major clades	**85**
Three major clades	**113**
Four major clades	**190**
Five major clades	**385**
All major clades (including EE)	**1,150**
In eukaryotes and prokaryotes	
Eukarya, Archaea, and Bacteria[c]	**205**
Eukarya and Archaea[c]	**207**
Eukarya and Bacteria[c]	**185**
Excavata and either Bacteria or Archaea	**2**

[a]NIP = did not meet the requirement of ≥4 sequences (from the 167 taxa that were chosen for this study) to produce a tree, and are therefore likely either very divergent or misannotated.
[b]A gene is considered to be present in a major clade only if it is present in at least 25% of the clades from the next taxonomic rank (e.g., Euglenozoa in Excavata, Apicomplexa in SAR, Animals, or Fungi in Opisthokonta); sequences in only a few lineages may be contaminants or the result of gene transfers.
[c]In at least five eukaryotic major clades: Excavata (Ex), Archaeplastida (Pl), SAR (Sr), Amoebozoa (Am), and Opisthokonta (Op). For every tree the root was placed in between Bacteria and Archaea + Eukaryotes when there were Bacteria; between Archaea and Eukaryotes when there were not Bacteria; or in Opisthokonta when there were not prokaryotes (Katz and Grant 2015).

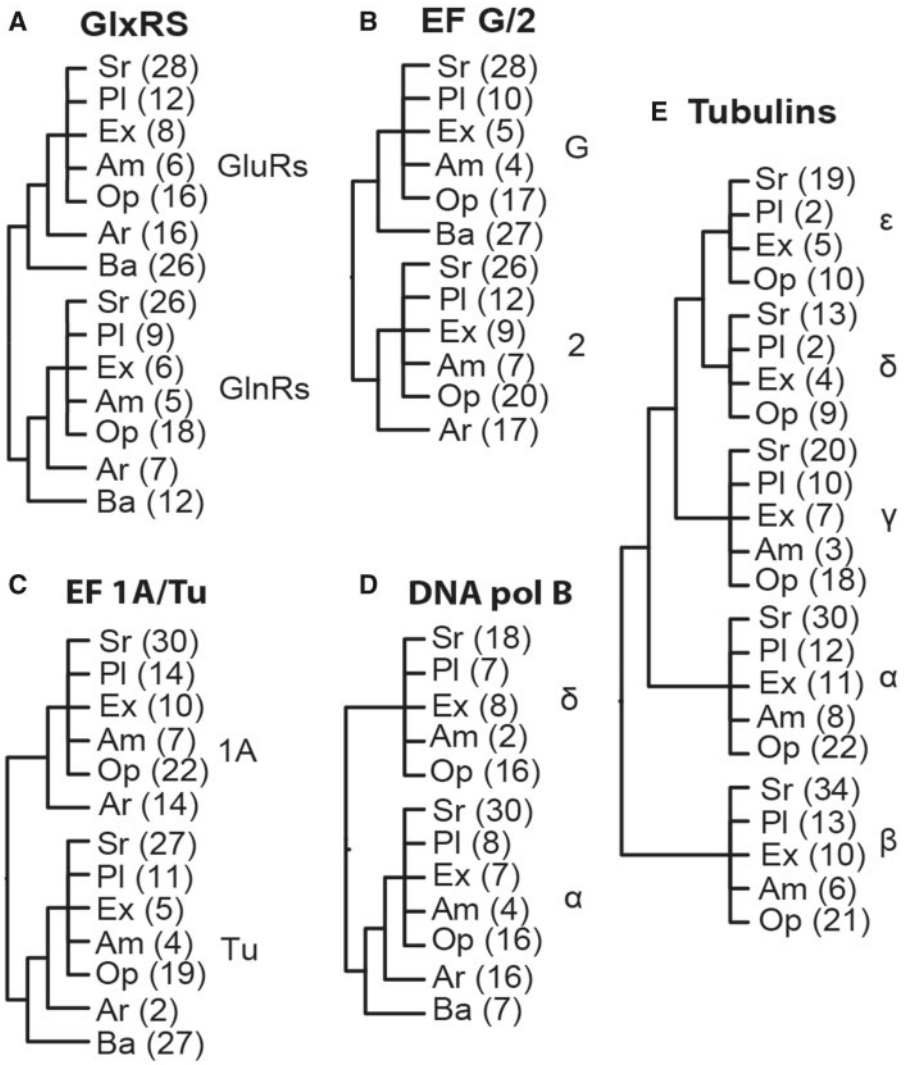

FIG. 5. PhyloToL homology assessment for well-known GFs that duplicated prior to the last eukaryotic common ancestor (LECA) or the last universal common ancestor (LUCA). Subfamilies of these ancient GFs are often categorized in different orthologous groups by OrthoMCL. The cartoon trees show the reconstruction of the phylogeny of 5 of the 8 analyzed ancient GF by PhyloToL. (*A*) glutamyl- and glutaminyl-tRNA synthetases, (*B*) elongation factors Tu/1a, (*C*) elongation factors G/2, (*D*) family B DNA polymerase, (*E*) Tubulins. Ar = Archaea, Ba = Bacteria, Op = Opisthokonta, Am = Amoebozoa, Ex = Excavata, Pl = Archaeplastida, Sr = SAR. The number in every tip represents the number of species per major clade. Full trees for the eight analyzed ancient GFs are found as Newick strings in supplementary fig. S3, Supplementary Material online.

specific genes and/or mis-annotations (table 2). Of the remaining 27% of genes that match conserved eukaryotic GFs, ~44% are conserved among all the major eukaryotic clades, ~8% are shared between all major eukaryotic clades and Archaea and ~8% are conserved among all major eukaryotic clades, Archaea and Bacteria (table 2).

Test of Homology Assessment

To benchmark the homology assessment in PhyloToL, we compared reconstructions of ancient (i.e., present in bacteria, archaea, and eukaryotes) GFs originally estimated in OrthoMCL. Members of ancient GFs tend to be categorized in different orthologous groups in OrthoMCL (e.g., α-tubulin is group OG5_126605 and β-tubulin is group OG5_126611). We analyzed eight ancient GFs that were likely present in the last eukaryotic common ancestor (LECA) or the last universal common ancestor (LUCA): ATPases, family B DNA polymerase, elongation factors Tu/1a, elongation factors G/2, glutamyl- and glutaminyl-tRNA synthetases, RNA polymerase subunit A, RNA polymerase subunit B, and tubulins (Brown and Doolittle 1997; Mulkidjanian et al. 2007; Nureki et al. 2010; Findeisen et al. 2014). Overall, our recovery of the homology of these ancient GFs was robust to our taxon-rich analyses (fig. 5 and supplementary fig. S3, Supplementary Material online). For four of the eight GFs (i.e., glutaminyl-tRNA synthetases, RNA polymerase subunit A, RNA polymerase subunit B, and tubulins) there were a few cases (<0.05%) where sequences were misclassified in the earlier steps of PhyloToL, likely due to the limited taxon sampling in the OrthoMCL-based "seeds" for BLAST analyses (supplementary fig. S3, Supplementary Material online).

We also benchmarked PhyloToL against the reconstruction of GFs of bacterial and eukaryotic organelle outer membrane pore-forming proteins as proposed by Reddy and Saier

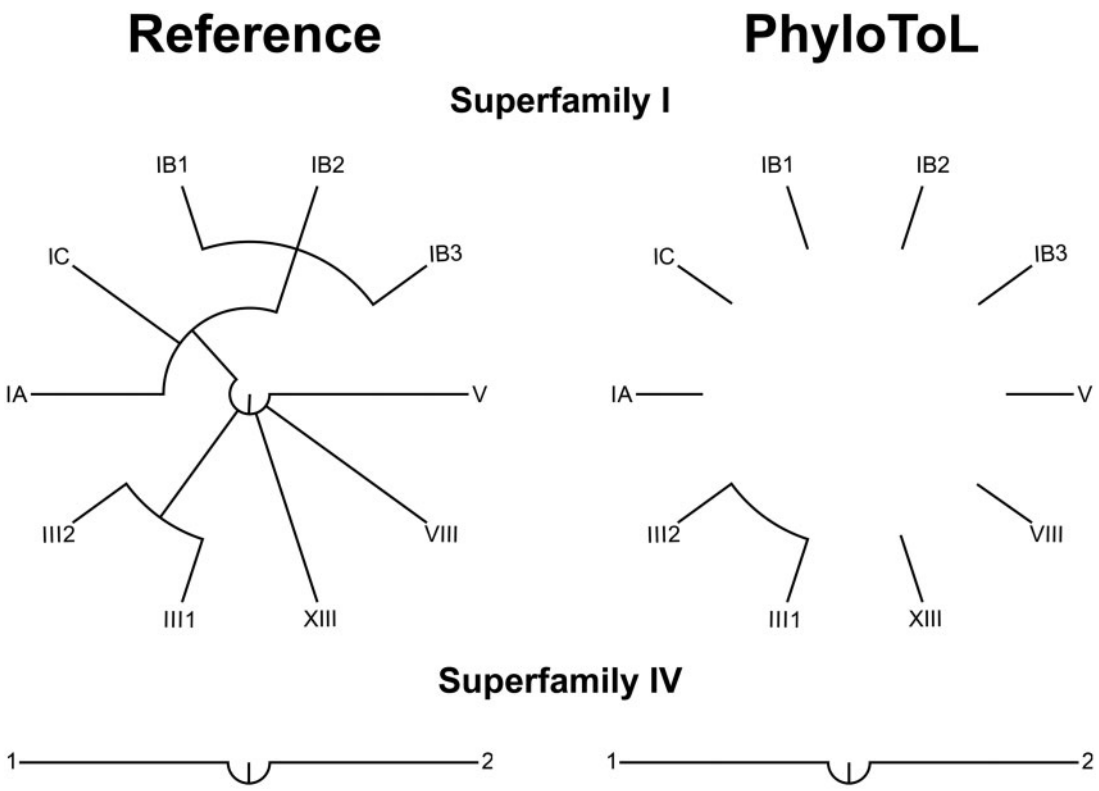

FIG. 6. PhyloToL homology assessment for candidate superfamilies (S) of outer membrane pore-forming proteins as proposed by Reddy and Saier (2016). The left hand "Reference" columns show the proposed superfamilies SI and SIV whereas the right hand "PhyloToL" column shows the surviving homologs (i.e., those connected by lines). Only cluster III of SI and the two GFs of SIV are homologous based on PhyloToL's default parameters (i.e., GUIDANCE V2.02: sequences cutoff = 0, 3, column cutoff = 0.4, 5 iterations).

(2016). Reddy and Saier (2016) combined 76 GFs among 5 superfamilies of varying size. To compare their homology statements to inferences from PhyloToL, we focused on the 12 GFs already included in the PhyloToL databases that fall into two superfamilies, the prokaryotic superfamily I (SFI) and eukaryotic superfamily IV (SFIV). Under PhyloToL's default parameters (i.e., GUIDANCE V2.02 sequence cutoff = 0.3, column cutoff = 0.4, number of iterations = 5), many SFI members (i.e., different GFs) determined by Reddy and Saier (2016) do not meet our criteria for homology: when running the full set of sequences of SFI in PhyloToL, only sequences of the largest GF survive, indicating that the other GFs are too dissimilar to be included in a MSA under our parameters (supplementary table S2, Supplementary Material online). We then re-ran PhyloToL to test homology in every cluster and sub-cluster of GFs that form SFI, but in the end only cluster III meets our conservative criteria for homology (fig. 5 and supplementary table S1, Supplementary Material online). In contrast to SFI, both members of the eukaryotic SFIV are retained under default parameters in PhyloToL (fig. 6 and supplementary table S2, Supplementary Material online). We then forced the GFs determined by Reddy and Saier (2016) to align, and found limited evidence of homology (e.g., conserved columns in MSAs). In sum, our estimation of homology is more stringent than in Reddy and Saier (2016), and our exploration of this question took ~3 h on a computer with four threads, highlighting the flexibility of PhyloToL for users.

Materials and Methods

There are four components in PhyloToL's algorithm: 1) GF assessment per taxon, 2) refinement of GFs and gene tree reconstruction, 3) tree-based contamination removal, and 4) generation of a supermatrix for species tree inference. The GF assessment per taxon includes features such as translation using informed genetic codes. The refinement of GFs and gene tree reconstruction filters and asserts homology in the GFs comparing sequences by length, overlap, similarity, and MSA. The component tree-based contamination removal detects and removes contaminant sequences based on predefined contamination rules and the position of the sequences in gene trees. Finally, the component generating a supermatrix for species tree inference chooses orthologs and discards paralogs based on tree topology to concatenate MSAs for species tree inference.

Naming Sequences

PhyloToL uses standardized names that are compatible with the third-party tools incorporated into the pipeline (e.g., GUIDANCE, RAxML). Although the users are free to assign different codes to the taxa at their convenience, PhyloToL requires that every taxon is named using a ten-digit code that can broadly reflect taxonomy (see Supplementary Material online for our suggested codes); this code is divided into three components, a major clade (e.g., Op = Opisthokonta), a "minor" clade (e.g., Op_me = Metazoa), and a species

name (e.g., Op_me_hsap for *Homo sapiens*). For each sequence, the ten-digit code is followed by the sequence identifier such as the GenBank accession or Ensembl ID (e.g., Op_me_hsap_ENSP00000380524). This naming system allows an easy control of names when handling alignments and trees.

GF Assessment per Taxon

The first component of PhyloToL (i.e., GF assessment per taxon; fig. 1A) allows the inclusion of a large number of data sources from online repositories (e.g., GenBank) or from the user's lab, and of different types (e.g., transcriptomes, proteins or annotated proteins from genomic sequences [e.g., 454, Illumina, ESTs]). The first steps aim to accurately assign sequences to homologous GFs, with improvements to the efficiency of these processes as compared with our original pipeline (Grant and Katz 2014a, 2014b; Katz and Grant 2015). To exemplify methods, we focus on the inclusion of Illumina transcriptome data, though the structure can easily be adapted for other sources. PhyloToL uses a pipeline (https://github.com/Katzlab/PhyloTOL/tree/master/AddTaxa; last accessed May 2, 2019) for passing assembled transcripts through a variety of steps for: segregation of short contigs (at a user-defined length), segregation of putative contaminants [from ribosomal RNAs (rRNA), bacteria and archaea], and assess GFs. To segregate rRNA sequences, we rely on BLAST, comparing each sequence against a database of diverse rRNA sequences sampled from across the tree of life (75 bacteria, 26 archaea, and 77 eukaryotes; Supplementary Material online). This is followed by the identification and removal of bacterial/archaeal transcripts through USEARCH V10 (Edgar 2010), which compares data against both a database of diverse bacterial + archaeal proteins and another database of diverse eukaryotic proteins, retaining all nonbacterial/archaeal transcripts (i.e., those with strong matches to eukaryotes, and those remaining unassigned). With this pruned data set, USEARCH is again used to bin these eukaryotic-enriched sequences into OrthoMCL GFs whereas rRNA and bacterial/archaeal transcripts are saved in a different location for easy retrieval if desired.

With growing evidence for the diversity of stop codon reassignments across the eukaryotic tree of life (e.g., Keeling and Doolittle 1997; Lozupone et al. 2001; Keeling and Leander 2003; Heaphy et al. 2016; Swart et al. 2016; Pánek et al. 2017), we include an optional step to evaluate potential alternatives to conventional stop codon usage (frequent in frame unconventional stop codons). This step is essential for some clades such as Ciliophora, where there are at least eight unconventional genetic codes (i.e., not all three traditional stop codons terminate translation). Using the most appropriate genetic code, each nucleotide sequence is then translated into the corresponding amino acid ORF.

Given the imperfect nature of HTS data, we take a conservative approach to avoid inflating the number of paralogs for each taxon and, therefore, we remove nearly identical sequences. These nearly identical sequences generated by HTS can represent an unknown mixture of alleles, recent paralogs and more importantly sequencing and/or assembly errors, which can be problematic for the comparative aspects of PhyloToL. To avoid this issue, for every taxon we remove nearly identical sequences at the nucleotide level (>98% nucleotide identity across ≥70% of their length).

An additional step is available to address the well-known phenomenon of sample bleeding (also known as index switching; Mitra et al. 2015; Larsson et al. 2018) that occurs during Illumina sequencing. Based on the observation that some of our taxa were contaminated by one another during Illumina sequencing, we developed a method to remove low read coverage contigs that are identical to higher read coverage contigs. To this end, we performed a USEARCH ("BLAST") all versus all of the nucleotide ORFs (at a minimum identity of 98% across ≥70% of their length). Those sequences that form clusters of hits to other taxa represent potential cross-contaminants. Next, those sequences with a substantially high read coverage compared with the mean (e.g., 10× more than the mean) are retained and low-read coverage sequences are excluded. In ambiguous cases (i.e., all are low read number), the entire group of sequences is discarded. Although this step is highly dependent on transcriptional state and sequencing depth, this conservative approach impacts <5% of transcripts for a given taxon using our own Illumina data.

Refinement of Homologs and Gene Tree Reconstruction

In the second component of PhyloToL (i.e., refinement of homologs and gene tree reconstruction; fig. 1B), GFs pass through a procedure to assess homology and then to produce gene trees. The procedure starts with a QC step that includes two filters: an overlap filter and a similarity filter. The overlap filter aims to remove nonhomologous sequences, which are sequences substantially longer than putative homologs (e.g., those with only shared motifs), or atypically short (i.e., those with insufficient overlap). Such sequences will confound paralog assessment and can negatively impact the alignments. To proceed, we start by identifying a "master sequence" as the putative homolog. This sequence has the lowest *E* value from the GF assignment and is also ≤150% the average length of the members from the reference GF data set. We then retain all sequences that have a pairwise local alignment overlap that includes at least 35% of the length of the master sequence. In contrast, the optional similarity filter allows the user to remove alleles and recent paralogs (i.e., too similar sequences) at a user-defined cutoff to improve efficiency. The similarity filter uses an iterative process in which the next longest sequence acts as the "master sequence" to remove highly similar sequences, and repeats until there are no more sequences that can be assigned as a "master sequence."

For the next part of the procedure, to assess homology within each GF, PhyloToL relies on GUIDANCE V2.02 scores, and using a user-specified number of iterations, identifies, and removes unreliably aligned and potentially nonhomologous sequences (fig. 1B). Then, GUIDANCE is used to filter the final alignment using preset cutoffs for sequences and columns

(default parameters or empirically defined, in our case 0.3 for sequences and 0.4 for columns). In contrast to the previous version of the pipeline that relied on only two iterations of GUIDANCE, one for removing poorly aligned sequences and another for removing poorly aligned columns, PhyloToL iterates the sequence-removal step for either a user-defined number of iterations or until all unreliable sequences have been removed. Only then the columns are removed based on the user-specified confidence threshold score (the default number of bootstrap replicates for each GUIDANCE run is 10). Residues with low confidence scores, based on a settable residue score cutoff, can be masked in the alignment with an "X" (turned off in our defaults). Finally, in PhyloToL, GUIDANCE uses more accurate MAFFT V7 parameters, including an iterative refinement method (E-INS-i algorithm, and up to 1,000 iterations). The E-INS-i algorithm was chosen because it makes the smallest number of assumptions of the three iterative refinement methods implemented in MAFFT and is recommended if the nature of sequences is less clear.

Tree-Based Contamination Removal

The third component of PhyloToL (i.e., tree-based contamination removal; fig. 1C) includes a method to identify and remove contaminants based on their location within the phylogenetic trees, though user scrutiny of results is required. If inspection of gene trees reveals sequences from a given taxon frequently nested among distantly related lineages, the user can create a set of "rules for contamination removal" and then run the tree-based contamination removal that will detect and remove potential contaminants from the alignments and subsequent trees (fig. 1C). To help users to define their rules for contamination removal, PhyloToL also generates a report (summary_contamination.csv) containing the frequency of every sister clade per lineage ignoring those with significantly longer branches than the average branch length of the tree, which allows the users to differentiate contamination (e.g., food, symbionts, and other sources) from fast evolving taxa that were incorrectly placed in trees. This component of PhyloToL iterates the refinement of homologs and gene tree reconstruction (i.e., second component) using the predefined rules to identify sequences of contamination and removing them for the next iteration. This continues until no more "contaminant" sequences are identified. The component tree-based contamination removal also produces a full list of contaminant sequences that can be removed from the permanent databases. To run the tree-based contamination removal more efficiently, potentially nonhomologs (i.e., sequences discarded by GUIDANCE) are also removed in every iteration.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank members of the Katz lab for comments on earlier versions of the manuscript. This work was supported by National Institutes of Health grant number 1R15GM113177-01, and National Science Foundation grant numbers DEB-1541511 and DEB-1651908 to L.A.K. Two anonymous reviewers are gratefully acknowledged for insightful comments on an earlier version of this manuscript.

References

Antonelli A, Hettling H, Condamine FL, Vos K, Nilsson RH, Sanderson MJ, Sauquet H, Scharn R, Silvestro D, Topel M, et al. 2017. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. *Syst Biol*. 66(2):152–166.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet*. 25(1):25–29.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 28(1):45–48.

Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA. 2000. The EMBL nucleotide sequence database. *Nucleic Acids Res*. 28(1):19–23.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2017. GenBank. *Nucleic Acids Res*. 45(D1):D37–42.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309(5733):416–422.

Brown JR, Doolittle WF. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol Mol Biol Rev. 61(4):456–502.

Brown MW, Heiss AA, Kamikawa R, Inagaki Y, Yabuki A, Tice AK, Shiratori T, Ishida KI, Hashimoto T, Simpson AGB, et al. 2018. Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol Evol*. 10(2):427–433.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 268(1):78–94.

Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc Biol Sci*. 283(1823):20152802.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 18(1):188–196.

Caron F, Meyer E. 1985. Does *Paramecium primaurelia* use a different genetic code in its macronucleus? *Nature* 314(6007):185–188.

Cerón-Romero MA, Nwaka E, Owoade Z, Katz LA. 2018. PhyloChromoMap, a tool for mapping phylogenomic history along chromosomes, reveals the dynamic nature of karyotype evolution in *Plasmodium falciparum*. *Genome Biol Evol*. 10(2):553–561.

Chaudhary R, Bansal M, Wehe A, FernûÀndez-Baca D, Eulenstein O. 2010. iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioinform. 11:574.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*. 34(90001):D363–368.

Daniels JP, Gull K, Wickstead B. 2010. Cell biology of the trypanosome genome. *Microbiol Mol Biol Rev*. 74(4):552–569.

De Oliveira Martins L, Mallo D, Posada D. 2016. A Bayesian supertree model for genome-wide species tree reconstruction. *Syst Biol*. 65(3):397–416.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang ZH. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc Biol Sci*. 279(1742):3491–3500.

Drori M, Rice A, Einhorn M, Chay O, Glick L, Mayrose I. 2018. OneTwoTree: an online tool for phylogeny reconstruction. *Mol Ecol Resour*. 18(6):1492–1499.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.
El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al. 2005. Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309(5733):404–409.
Findeisen P, Mû¥hlhausen S, Dempewolf S, Hertzog J, Zietlow A, Carlomagno T, Kollmar M. 2014. Six subgroups and extensive recent duplications characterize the evolution of the eukaryotic tubulin protein family. *Genome Biol Evol*. 6(9):2274–2288.
Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* 407(6807):1018–1022.
Grant JR, Katz LA. 2014a. Building a phylogenomic pipeline for the eukaryotic tree of life—addressing deep phylogenies with genome-scale data. *PLoS Curr*. 6. pii: ecurrents.tol.c24b6054aebf3602748ac042ccc8f2e9.
Grant JR, Katz LA. 2014b. Phylogenomic study indicates widespread lateral gene transfer in Entamoeba and suggests a past intimate relationship with parabasalids. *Genome Biol Evol*. 6(9):2350–2360.
Hall BG. 2013. Building phylogenetic trees from molecular data with MEGA. *Mol Biol Evol*. 30(5):1229–1235.
Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. 2016. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. *Mol Biol Evol*. 33(11):2885–2889.
Heiss AA, Kolisko M, Ekelund F, Brown MW, Roger AJ, Simpson A. 2018. Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *R Soc Open Sci*. 5(4):171707.
Hernandez-Rivas R, Herrera-Solorio AM, Sierra-Miranda M, Delgadillo DM, Vargas M. 2013. Impact of chromosome ends on the biology and virulence of *Plasmodium falciparum*. *Mol Biochem Parasitol*. 187(2):121–128.
Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. *Nat Microbiol*. 1:16048.
Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28(1):27–30.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):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. *Syst Biol*. 64(3):406–415.
Keeling PJ, Doolittle WF. 1997. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. *Proc Natl Acad Sci U S A*. 94(4):1270–1275.
Keeling PJ, Leander BS. 2003. Characterisation of a non-canonical genetic code in the oxymonad *Streblomastix strix*. *J Mol Biol*. 326(5):1337–1349.
Korf I. 2004. Gene finding in novel genomes. *BMC Bioinform*. 5:59.
Kryukov K, Imanishi T. 2016. Human contamination in public genome assemblies. *PLoS One* 11(9):e0162424.
Kumar S, Krabberod AK, Neumann RS, Michalickova K, Zhao S, Zhang X, Shalchian-Tabrizi K. 2015. BIR pipeline for preparation of phylogenomic data. *Evol Bioinform Online* 11:79–83.
Langer MR, Lipps JH. 1995. Phylogenetic incongruence between dinoflagellate endosymbionts (Symbiodinium) and their host foraminifera (Sorites): small-subunit ribosomal RNA gene sequence evidence. *Mar Micropaleontol*. 26(1–4):179–186.
Larsson AJM, Stanley G, Sinha R, Weissman IL, Sandberg R. 2018. Computational correction of index switching in multiplexed sequencing libraries. *Nat Methods* 15(5):305–307.
Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr Biol*. 22(15):R593–594.
Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. *Syst Biol*. 57(1):104–115.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13(9):2178–2189.
Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. *Curr Biol*. 11(2):65–74.
Lusk RW. 2014. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One* 9(10):e110808.
Maddison WP. 1997. Gene trees in species trees. *Syst Biol*. 46(3):523–536.
Mallo D, Posada D. 2016. Multilocus inference of species trees and DNA barcoding. *Philos Trans R Soc Lond B Biol Sci*. 371(1702):20150335.
Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE. 2010. Gene expression in trypanosomatid parasites. *J Biomed Biotechnol*. 2010:525241.
Merchant S, Wood DE, Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675.
Mitra A, Skrzypczak M, Ginalski K, Rowicka M. 2015. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using Illumina platform. *PLoS One* 10(4):e0120520.
Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV. 2007. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Microbiol*. 5(11):892–899.
Narechania A, Baker RH, Sit R, Kolokotronis SO, DeSalle R, Planet PJ. 2012. Random Addition Concatenation Analysis: a novel approach to the exploration of phylogenomic signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. *Genome Biol Evol*. 4(1):30–43.
Nureki O, O'Donoghue P, Watanabe N, Ohmori A, Oshikane H, Araiso Y, Sheppard K, Söll D, Ishitani R. 2010. Structure of an archaeal non-discriminating glutamyl-tRNA synthetase: a missing link in the evolution of Gln-tRNAGln formation. *Nucleic Acids Res* 38(20): 7286–7297.
Okamoto N, Chantangsi C, Horak A, Leander BS, Keeling PJ. 2009. Molecular phylogeny and description of the Novel Katablepharid *Roombia truncata* gen. et sp nov., and Establishment of the Hacrobia Taxon nov. *PLoS One* 4(9):e7080.
Pánek T, Žihala D, Sokol M, Derelle R, Klimeš V, Hradilová M, Zadrobílková E, Susko E, Roger AJ, Čepička I, et al. 2017. Nuclear genetic codes with a different meaning of the UAG and the UAA codon. *BMC Biol*. 15(1):8.
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 25(7):1043–1055.
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(Web Server):W23–28.
Philippe H, Vienne D, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. *Eur J Tax*. 283:1–25.
Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol*. 29(1):1–5.
Reddy BL, Saier MH Jr. 2016. Properties and phylogeny of 76 families of bacterial and eukaryotic organellar outer membrane pore-forming proteins. *PLoS One* 11(4):e0152733.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 16(6):276–277.
Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A. 2008. The PhyLoTA browser: processing GenBank for molecular phylogenetics research. *Syst Biol*. 57(3):335–346.
Scherf A, Figueiredo LM, Freitas-Junior LH. 2001. Plasmodium telomeres: a pathogen's perspective. *Curr Opin Microbiol*. 4(4):409–414.
Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*. 5(12):e1000605.

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(W1):W7–14.

Shrestha PM, Nevin KP, Shrestha M, Lovley DR. 2013. When is a microbial culture "pure"? Persistent cryptic contaminant escapes detection even with deep genome sequencing. *mBio* 4(2):e00591-12.

Smith SA, Beaulieu JM, Donoghue MJ. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol Biol.* 9:37.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Stamatakis A, Ott M, Ludwig T. 2005. RAxML-OMP: an efficient program for phylogenetic inference on SMPs. *Lect Notes Comput Sci.* 3606:288–302.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33(Web Server issue):W465–467.

Struck TH. 2013. The impact of paralogy on phylogenomic studies—a case study on annelid relationships. *PLoS One* 8(5):e62892.

Sugiura M, Tanaka Y, Suzaki T, Harumoto T. 2012. Alternative gene expression in type I and type II cells may enable further nuclear changes during conjugation of *Blepharisma japonicum*. *Protist* 163(2):204–216.

Swart EC, Serra V, Petroni G, Nowacki M. 2016. Genetic codes with no dedicated stop codon: context-dependent translation termination. *Cell* 166(3):691–702.

Tanifuji G, Kim E, Onodera NT, Gibeault R, Dlutek M, Cawthorn RJ, Fiala I, Lukes J, Greenwood SJ, Archibald JM. 2011. Genomic characterization of *Neoparamoeba pemaquidensis* (Amoebozoa) and its Kinetoplastid Endosymbiont. *Eukaryot Cell* 10(8):1143–1146.

Tremblay-Savard O, Swenson KM. 2012. A graph-theoretic approach for inparalog detection. *BMC Bioinform.* 13(19 Suppl):S16.

Vasilakis N, Forrester NL, Palacios G, Nasar F, Savji N, Rossi SL, Guzman H, Wood TG, Popov V, Gorchakov R, et al. 2013. Negevirus: a proposed new taxon of insect-specific viruses with wide geographic distribution. *J Virol.* 87(5):2475–2488.

Vinuesa P, Ochoa-Sanchez LE, Contreras-Moreira B. 2018. GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus *Stenotrophomonas*. *Front Microbiol.* 9:771.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A.* 111(45):E4859–4868.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9(10):R151.