Research article

Title: Phylogenetic placement and contamination screening of Amoebozoa genomic data from the Protist 10,000 Genomes (P10K) Database

Authors: Alfredo L. Porfirio-Sousa[a,1], Robert E. Jones[a], Matthew W. Brown[b], Daniel J. G. Lahr[c], Alexander K. Tice[a,1]

[a] Department of Biological Sciences, Texas Tech University, Lubbock, TX, USA

[b] Department of Biological Sciences, Mississippi State University, Mississippi State, MS, USA

[c] Institute of Biosciences, University of São Paulo, São Paulo 055080-090, Brazil

[1] corresponding authors: alex.tice@ttu.edu (A. K. Tice) and alfredolpsousa@gmail.com (A. L. Porfirio-Sousa)

ORCID

ALPS: https://orcid.org/0000-0001-7490-158X

REJ: https://orcid.org/0000-0001-5227-4773

MWB: https://orcid.org/0000-0002-1254-0608

DJGL: https://orcid.org/0000-0002-1049-0635

AKT: https://orcid.org/0000-0002-3128-1867

Abstract

Background: Genomic data are essential for uncovering the evolutionary history, ecological roles, and diversity of life. Yet, microbial eukaryotes like Amoebozoa, an ancient and morphologically diverse lineage, remain critically underrepresented in genomic repositories. This has limited our ability to address fundamental questions in eukaryotic evolution. The Protist 10,000 Genomes (P10K) initiative seeks to fill this gap by generating and compiling genome- and transcriptome-level data for a wide range of microbial eukaryotes. To ensure the reliability of these resources, accurate taxonomic identification and contamination screening are vital. In this study, we aimed to assess the taxonomic consistency and integrity of the P10K database with a phylogenetic-based approach using Amoebozoa as a case study.

Results: Through SSU rDNA/rRNA and COI phylogenetic reconstructions this study confirmed several initial taxonomic identifications provided in the P10K database, resolved ambiguities at higher taxonomic levels, and corrected misassignments among morphologically similar but phylogenetically distant taxa. Moreover, the contamination screening using SSU rDNA/rRNA revealed several amoebozoan data that are contaminated by sequence from other eukaryotic taxa, representing contaminated genomic assemblies.

Conclusion: Phylogenetic placement coupled with contamination screening enabled us to distinguish the higher-quality Amoebozoa datasets currently available in the P10K database from those requiring decontamination or additional sequencing before downstream use. These findings serve as a reference for the future use of these data and as a guide for further sequencing efforts aimed at expanding the taxonomic diversity of Amoebozoa represented at the genomic level. By applying a phylogenetic survey to the Amoebozoa data, we present a framework that can be extended to other microbial eukaryote lineages. Addressing imprecise taxonomic identifications and contamination in certain P10K datasets, as well as data reproducibility, will further enhance the value of this unprecedented genomic resource for protists, with significant potential to illuminate the evolution and diversification of eukaryotic life.

Keywords: Amoebozoa, Arcellinida, Eukaryotic diversity, Comparative genomics, Genomic database curation

Introduction

Genomic-level data are essential for advancing our understanding of the evolution of life on Earth [1, 2]. High-quality genome and transcriptome sequences enable comparative analyses that reveal patterns of genomic evolution, including gene family expansions, horizontal gene transfers, and changes in genome organization and regulatory systems [3, 4]. Such data also clarify phylogenetic relationships through multi-gene and genome-scale reconstructions [5, 6]. Moreover, genomic analyses uncover ecological interactions by identifying metabolic pathways, symbiotic associations, and genetic adaptations to specific environmental conditions [7]. Consequently, genomics has become central to studying life's diversity and complexity.

While genomics has progressed rapidly for major eukaryotic groups such as plants, animals, fungi, and other traditional model systems, most microbial eukaryotes (commonly referred to as protists) remain vastly underrepresented [1, 2]. Although they comprise most of eukaryotic diversity, the majority of lineages within this highly diverse paraphyletic assemblage still lack genomic data [2, 5]. Over the past decade, several initiatives have contributed to fill this gap, such as the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) for marine microbial genomics [8], the Tree of Life Programme for eukaryotic genome sequencing [9], and the One Thousand Plant Transcriptomes initiative (1KP) for plants, including single-celled algae [10]. Similarly, the Protist 10,000 Genomes (P10K) initiative aims to address the underrepresentation of microbial eukaryotes in genomic databases by generating new genomic data (i.e., genomes and transcriptomes) and compiling previously available data from a wide array of lineages, coupled with taxonomic identification and decontamination procedures [2]. This large-scale effort, consolidated in the P10K database, represents an unprecedented genomic resource for protists and potentially provides a valuable foundation for achieving a comprehensive and integrative view of eukaryotic evolution.

However, to fully exploit the potential of this newly available genomic resource for protists, three key challenges must be considered: (1) accurate taxonomic identification, (2) the presence of contaminated genomic data resulting from non-target eukaryotic contamination, and (3) data interpretation and reproducibility [11, 12]. Specifically, the taxonomic identification of the P10K database relies on small subunit ribosomal DNA (SSU rDNA) retrieval and BLAST similarity searches against curated SSU databases, including SILVA, PR^2, and NCBI's NT (nucleotide sequence) database

[2]. Although informative, a BLAST-based approach is less accurate than phylogeny-based methods for taxonomic identification and may lead to a less precise classification of the target groups. Regarding contamination, protists typically inhabit environments with high eukaryotic microbial diversity, several prey on other organisms, many host endosymbionts, and most are often difficult to isolate as monoeukaryotic cultures or single cells [13–15]. As a result, genomic-level sequencing efforts might produce contaminated assemblies that contain not only the genome of the target organism but also sequences from other eukaryotic organisms [16, 17]. While P10K performs a decontamination strategy to remove bacterial, archaeal, viral, fungal, and other eukaryotic contaminants from ciliate data; for other protist groups, only bacterial, archaeal, and viral contaminants are filtered out [2]. Finally, the P10K database currently lacks photo-documentation and detailed information, for instance, about the sequencing platforms used for each sample (i.e., Illumina, Oxford Nanopore, or PacBio Sequel II), which are essential for deeper interpretation and reproducibility of the data. In this context, accurate taxonomic identification, assessment of eukaryotic contamination, and the availability of more detailed sample information are most essential for the effective downstream use of data in the P10K database.

Here, we present a survey of the taxonomic identification and contamination screening of the P10K database using Amoebozoa as a case study. Amoebozoa is an ancient clade of heterotrophic organisms, estimated to have originated around 1.5 billion years ago, and it exhibits remarkable morphological and ecological diversity [13, 15, 18]. Because Amoebozoa are heterotrophic and often share habitats with a wide range of eukaryotic organisms, making them particularly prone to sequence contamination, they serve as an ideal model group for evaluating the accuracy of taxonomic identification and contamination assessment in the P10K database. In short, we retrieved all publicly available Amoebozoa genomes from the P10K database, representing the three major amoebozoan clades; Tubulinea, Evosea, and Discosea [13]. From these data, we conducted phylogenetic analyses using two commonly employed molecular markers: small subunit ribosomal DNA/RNA (SSU rDNA/rRNA or 18S) and cytochrome c oxidase subunit I (COI, also known as CO1 or cox1). This phylogenetic investigation enables the phylogenetic placement and contamination screening of Amoebozoa genomic data present in the P10K database, leading to the identification of higher-quality data, while also highlighting those that require decontamination or additional sequencing prior to downstream analyses. Additionally, we present some strategies that can mitigate contamination during sample preparation and discuss how the availability of more detailed metadata directly associated to

each sample can improve the interpretation, usability, and reproducibility of P10K data. Ultimately, this study serves as a proof of concept for using a phylogenetic framework to improve taxonomic identification and contamination assessment within the P10K dataset, as an approach that can be extended to other taxa represented in the database.

Methods

Datasets construction

We retrieved genomic-level assemblies for the 201 amoebozoans from the P10K database based on the taxonomic annotations provided on the platform (**Supplementary Information - Table S1**). We constructed a small subunit ribosomal DNA/RNA (SSU rDNA/rRNA) dataset considering all 201 of these assemblies. Further, we constructed a cytochrome c oxidase subunit I (COI) dataset focusing only on the testate amoebae order Arcellinida (Tubulinea), since this marker is well sampled for this lineage as it has been traditionally used for phylogenetics in arcellinids. To construct datasets for the small subunit ribosomal DNA/RNA (SSU rDNA/rRNA) and cytochrome c oxidase subunit I (COI) markers, we extracted sequences from these assemblies using similarity searches implemented in BLAST+ v2.16.0+ and a custom Python script (**Supplementary Information - File S1**). For each marker, the script automated the creation of BLAST databases using makeblastdb and performed local blastn searches for each query sequence in a multi-FASTA file against each Amoebozoa genomic data. Specifically, the script executed the commands *makeblastdb -in P10KID.fasta -dbtype nucl* and *blastn -query marker_query -db P10K.fasta -outfmt [script_default_choice]* (**Supplementary Information - File S1**). The script retrieved the top five hits per genomic data file, extracting each aligned region along with 1000 bp of upstream and downstream flanking sequence. This approach enabled recovery of extended SSU and COI regions suitable for downstream phylogenetic analyses and compatible with Amoebozoa SSU and COI data available in the PR^2 database [19] and NCBI. The orientation of each retrieved sequence was assessed, and sequences were reverse complemented when necessary to match the strand of the original query. For the query sequences, we used SSU rDNA/rRNA data from the PR^2 database for three representative species of the major amoebozoan clades: *Arcella vulgaris* WP (Tubulinea; GenBank: HM853762.1), *Dictyostelium discoideum* (Evosea; GenBank: AM168040.1),

and *Acanthamoeba castellanii* Neff (Discosea; GenBank: U07416.1). As the COI query sequence, we considered the sequence of *Arcella uspiensis* (SRR5396453).

To build a phylogenetically informative datasets, we combined the retrieved sequences with previously published SSU rDNA/rRNA and COI datasets for Amoebozoa, as well as sequences from the PR^2 database for SSU [13, 15, 19, 20]. For SSU, we curated a non-redundant dataset broadly representative of the major lineages in Amoebozoa (Tubulinea, Evosea, and Discosea) and for COI a dataset broadly representative of the major lineages in Arcellinida, for both markers excluding environmental sequences. This strategy ensured a robust and interpretable dataset for our phylogenetic framework. From a preliminary phylogenetic reconstruction, we curated the SSU and COI datasets used to generate the main trees in this study. This initial analysis allowed us to identify and remove identical or highly similar sequences that resulted from retrieving the top five BLAST+ hits per genome. It also enabled visual inspection of the tree to detect SSU sequences that either failed to cluster within Amoebozoa or formed unusually long branches. Many of these were short sequences (<200 bp) and were excluded from downstream analyses, while some long branches corresponded to full-length sequences that likely represented contaminants. These were retained for further contamination screening.

For contamination screening, we focused on the SSU rDNA/rRNA dataset. After the initial phylogenetic reconstruction (see *Phylogenetic reconstructions* section), any SSU sequences from P10K genome assemblies that represented long branches or did not branch within the Amoebozoa clade were selected for further analysis. These sequences were subjected to additional BLAST+ searches against the PR^2 database, a curated SSU resource representing eukaryotic diversity, using our custom script and the same parameters described above (**Supplementary Information - File S1**). To perform this search locally, we downloaded the complete PR^2 database and used it as the reference for the BLAST+ similarity search. This approach allowed us to retrieve SSU sequences from the PR^2 database that were similar to those of the putative contaminant eukaryotes and to assign their taxonomic affiliations through subsequent phylogenetic analyses.

Phylogenetic reconstructions

All phylogenetic reconstructions in this study were based on multiple sequence alignments (MSAs) generated using MAFFT v7.490 with the E-INS-I algorithm and 1000 refinement iterations. Alignments were produced with the following command: *mafft --genadpair --maxiterate 1000 input.fasta > output_aligned.fasta*. Automated alignment trimming was performed with trimAl v1.2, using the command *trimal -in input_aligned.fasta -out output_aligned_trimmed.fasta -keepheader -gt [threshold]*, where the gap threshold was set to 0.3 for SSU rDNA/rRNA and 0.5 for COI. Phylogenetic trees were inferred from the trimmed alignments using the maximum likelihood method implemented in IQ-TREE v2.3.6, with ModelFinder for model selection and node support assessed via 1,000 ultrafast bootstrap replicates and 1,000 SH-aLRT tests. The analysis was executed with the command *iqtree2 -s aligned_trimmed.fasta -alrt 1000 -bb 1000 -m TEST*.

Results and discussion

Phylogeny-based taxonomic identification of P10K amoebozoan data

Amoebozoan small subunit ribosomal DNA/RNA (SSU rDNA/rRNA) sequences were successfully retrieved from 151 of the 201 genomic datasets available for Amoebozoa in the P10K database (**Supplementary Information - Table S1**). Given the established use of the Cytochrome c oxidase subunit I (COI) marker in Arcellinida phylogenetics, we also specifically targeted this marker for arcellinid taxa. Arcellinid COI sequences were successfully recovered from 40 of the 59 genomic data available for Arcellinida (**Supplementary Information - Table S1**). Phylogenetic analyses based on SSU rDNA/rRNA and COI were largely congruent, supporting most of the original taxonomic assignments in the P10K database (**Figs. 1 and 2**). Moreover, they enabled a more precise identification for 43 taxa, including a refined classification at the genus and family levels for taxa initially classified only at the family or higher levels (**Figs. 1 and 2**; **Supplementary Information - Table S1**). As species-level identification typically requires extensive morphological and morphometric data in addition to molecular evidence, we adopted a conservative approach and assigned identifications at the genus level based on our phylogenetic results. However, it is worth noting that several genomic data from the P10K database originate from well-established cultures of widely used and shared strains, some originally available in the NCBI database, for which detailed morphological data are available in

the literature, allowing confident species-level identifications (**Figs. 1 and 2**; **Supplementary Information - Table S1**).

Notably, seven original taxonomic assignments from the P10K database were not corroborated by our phylogenetic reconstructions (**Figures 1 and 2**; **Supplementary Information - Table S1**). This was most apparent within Arcellinida, where misidentifications primarily involved closely related or morphologically similar genera and families. Although Arcellinida are well known for their test (shell), which provides informative taxonomic characters, convergent evolution has led to similar shell morphologies across distantly related lineages. For example, species of *Difflugia* (Difflugiidae), *Hyalosphenia* (Hyalospheniidae), and *Netzelia* (Netzelliidae) often possess rounded, ovoid, or elongated shells, which can lead to misidentification if other shell features (e.g., aperture shape, composition) or cellular characteristics are not considered [21, 22]. Consistent with this, our analyses revealed that several taxa initially assigned to Difflugiidae (infraorder Longithecina) belong to Hyalospheniidae (Hyalospheniformes) or Netzelliidae (Sphaerothecina) (**Supplementary Information - Table S1**). These infraorders are distantly related, with their last common ancestor estimated to have lived over 500 million years ago [15, 23], making such misassignments evolutionarily significant. Outside Arcellinida, only one notable case of misidentification was observed: *Pessonella* sp. PRA-29, a culture originally submitted to the American Type Culture Collection (ATCC) under the genus *Pessonella*, was later described as a new genus and species, *Armaparvus languidus*, representing the correct taxonomic identification for this organism [24].

Contamination screening of the P10K amoebozoan data

Non-amoebozoan SSU sequences were retrieved from 58 of the 201 genomic datasets available for Amoebozoa in the P10K database (**Supplementary Information - Table S1**). The phylogenetic analysis of these non-amoebozoan SSU and SSU sequences from the PR^2 database reveals a widespread contamination of the P10K amoebozoan dataset by a taxonomically diverse set of eukaryotic lineages, mirroring the ecological complexity of the environments these protists inhabit (**Figure 3 and Supplementary Information - Table S1**). Among the contaminant groups were fungi and metazoans, including sequences from arthropods and nematodes, likely introduced via soil

particles, organic debris, or sample handling procedures (**Figure 3**). SSU sequences affiliated with ciliates were particularly abundant, with representatives spanning multiple clades such as Vorticellidae, *Coleps*, and *Pseudomicrothorax* (**Figure 3**). Additional contaminants included lineages within the Stramenopiles such as *Paraphysomonas* and *Poterioochromonas*, as well as centrohelid heliozoans (Haptista) and cercozoans (Rhizaria) (**Figure 3**). Several photosynthetic eukaryotes represented another major group of contaminants, including land plants, especially angiosperms of the Fabaceae family, likely introduced through pollen or plant debris, as well as green algae from the Chlorophyceae (e.g., *Chlamydomonas*, *Hyalomonas*) and Zygnematophyceae, the closest relatives of land plants (**Figure 3**). Collectively, these findings demonstrate that contamination in amoebozoan assemblies is both frequent and taxonomically widespread, encompassing multiple branches of the eukaryotic tree.

Quality assessment of the P10K amoebozoan data

Based on the results from single-marker retrieval, phylogenetic inference, and quality assessment, we were able to evaluate the current state of Amoebozoa genomic data in the P10K database. As a relative measure, we can consider samples to be of relatively higher quality if they contain the SSU marker (and COI for arcellinids), show no signs of contamination based on SSU phylogenetic reconstruction, and have a BUSCO completeness score of at least 50% (**Figure 4; Supplementary Information - Table S1**). These samples are more likely to yield meaningful results in downstream analyses, including phylogenomics and comparative genomics (**Figure 4; Supplementary Information - Table S1).** It is important to note, however, that the SSU-based contamination screening we used as an exploratory assessment of the amoebozoan P10K data has limitations and cannot, on its own, detect all sources of potential contamination. Therefore, incorporating additional markers for further screening is advisable before using the data in downstream analyses. Similarly, while we chose a relatively low BUSCO threshold (≥50%) to include a broader range of potentially useful genomic data. More stringent completeness cutoffs, as well as additional quality metrics such as genome contiguity, N50 values, and assembly size, are required depending on the downstream applications of the data and should be considered accordingly. On the other hand, the P10K samples lacking key markers, flagged as contaminated based on SSU phylogenetic inference, or with BUSCO scores below 50%

represent more incomplete and lower-quality genomic data (**Figure 4; Supplementary Information - Table S1**). In this context, several of the amoebozoan genomic data available in the P10K database require decontamination or further sequencing prior to downstream analysis (**Figure 4D; Supplementary Information - Table S1**). Finally, this quality assessment of the P10K genomic resource highlights that several major Amoebozoa lineages remain unsampled at the genomic level and serves as a useful guide for targeted sampling efforts aimed at expanding the taxonomic diversity sampled for genomic data.

Widespread Contamination in Protist Genomes: Sources, Impacts, and Mitigation Strategies

While the widespread contamination identified in amoebozoan genome assemblies is certainly undesirable, it is consistent with the natural ecological context not only of Amoebozoa but of microbial eukaryotes in general. Protists typically inhabit complex microbial communities, including soil, biofilms, freshwater and marine sediments, mosses, and decaying organic matter, where they coexist and interact with a wide diversity of organisms [13, 21, 25]. Like many free-living protists, amoebozoans are predatory and feed on bacteria, algae, fungi, and other eukaryotic cells, or form close physical associations with them, such as transient or stable endosymbiotic relationships [13, 21]. These ecological interactions, combined with the technical difficulty of isolating single amoebozoan cells free from other microbial associates, make the presence of contaminant sequences in genome assemblies not only possible but likely [16]. Many species cannot be maintained in long-term axenic or monoeukaryotic cultures, and even those that can often require extensive purification efforts to eliminate co-cultured organisms [26].

Contaminant sequences can significantly impact genomic-level downstream analyses. They may compromise gene prediction, reveal artifactual patterns of gene family evolution, mislead functional annotations, and introduce biases in comparative genomic studies [4, 27, 28]. More specifically, contamination can lead to overestimation of genomic complexity and distorts analyses of gene family evolution by introducing homologs or paralogs from unrelated lineages, which can result in artificial expansions or contractions of gene families and misrepresentation of evolutionary trajectory studies [4, 27, 28]. Functional annotation is similarly affected, as contaminant sequences may be erroneously

assigned to the target genome, leading to inaccurate inferences about metabolic capabilities, signaling pathways, or ecological roles [4, 27, 28]. Similarly, assemblies that include sequences from diverse eukaryotic contaminants may cluster incorrectly in phylogenomic datasets, leading to artifactual phylogenetic trees [29]. Technically, contamination can also affect genome completeness metrics by artificially increasing the number of expected genes detected. This may create the false impression of high assembly quality and completeness, even when substantial portions of the assembly are derived from non-target organisms [16]. Ultimately, unaddressed contamination undermines efforts to draw biologically meaningful conclusions about evolution, diversity, and functional biology from genomic data.

Given these challenges, several strategies that are not reported to be used by the P10K initiative have been successfully used to reduce contamination in protist genomic-level data generation efforts, including for Amoebozoa [13, 15, 23]. For taxa that can be cultivated, growing cultures through multiple generations can help eliminate contaminant organisms that were initially co-isolated with the target taxon [13, 15, 23]. Other effective practices include visual inspection of cultures to detect fungi or small eukaryotes, filtration of culture media, and rigorous sterile handling during DNA and RNA extraction. For species that must be isolated as single cells from environmental samples, useful techniques include repeated transfers of the cell through filtered sterile water followed by overnight starvation in sterile medium [13, 15, 23]. These procedures allow cleaning of the cells and digestion of prey items, reducing the risk of capturing genetic material from non-target organisms derived from their environment or food source.

Even when applying these methods, it is not possible to guarantee that genomic data will be completely free from contamination, particularly when working with environmentally derived specimens. Therefore, comprehensive screening of genome assemblies remains essential. When contamination is detected, identifying the phylogenetic affinities of non-target sequences can guide decisions on data curation and inform subsequent analyses [29]. As corroborated in this study, combining single marker-based phylogenetic screening with genome-level examination provides a powerful and generalizable strategy for distinguishing genuine genomic content from artifactual sequences, especially SSU that have been traditionally sequenced from diverse eukaryotes and for which comprehensive curated databases like PR^2 are available. Importantly, this strategy couples the strengths of likelihood-based phylogenetics, which outperform similarity-based approaches such as BLAST by incorporating explicit

models of molecular evolution and statistically grounded inference [30, 31]. These features allow for more accurate reconstruction of evolutionary relationships, particularly among divergent or closely related taxa, where mere sequence similarity may be misleading [31, 32]. This approach serves as a guide for data curation and ensures that genomic data accurately represent the biology and evolutionary histories of the target protist lineages.

Reproducibility of the P10K data

Currently, the P10K database lacks photo-documentation and detailed metadata directly associated with each sample, which impairs both the reproducibility and deeper interpretation of the data. In particular, the absence of voucher images of specimens or cultures from which the data were derived prevents taxonomic confirmation and revision. Without such reference material, especially for uncultivable organisms, it becomes impossible to attempt re-isolation of the target taxa for additional sequencing efforts aimed at generating more complete genomic data. Moreover, the integration of morphological documentation with phylogenetic analyses would not only improve taxonomic accuracy and robustness, but also enable more comprehensive, integrative discoveries that combine molecular and morphological information. Ideally, the database could include voucher photographs and, when possible, images of multiple specimens of the same taxon from environmental samples or cultures. This would facilitate further morphometric analyses and potentially contribute to the utility of the P10K dataset for researchers who study protists using both molecular and morphological approaches. Another limitation affecting data accessibility and reproducibility is the lack of clear information on the specific sequencing platforms used for each sample (e.g., Illumina with short or long-insert libraries, Oxford Nanopore, or PacBio Sequel II). Since each platform has characteristic error profiles and biases, this metadata is critical for downstream analyses and informed interpretation of the genomic data. Thus, the incorporation of photo-documentation and detailed metadata for each sample could substantially contribute to the scientific value, reproducibility, and long-term impact of the P10K database.

Conclusions

This study presents a comprehensive phylogenetic assessment of taxonomic assignments and contamination across Amoebozoa genomic datasets in the P10K database. By using SSU rDNA/rRNA

and COI markers, we confirmed many of the original classifications, refined others, and identified multiple cases of misidentification within morphologically similar lineages. Additionally, we uncovered widespread contamination by diverse eukaryotic lineages, including fungi, metazoans, green algae, and other protists. These findings highlight the ecological complexity of protist-associated environments and the inherent challenges of obtaining contamination-free genomic data for target lineages. Despite these challenges, our results demonstrate that single-marker phylogenetic screening, particularly using SSU, provides a reliable and scalable strategy for verifying taxonomic identity and for an exploratory detection of contamination. By improving the taxonomic resolution and reliability of available genome assemblies, this work contributes to downstream evolutionary, ecological, and functional genomic studies of Amoebozoa enabled by the genomic resources available in the P10K database. More broadly, our framework offers a practical and generalizable approach for curating the growing volume of genomic data from protists. As genomic resources for microbial eukaryotes continue to expand, phylogenetically aware data curation efforts, alongside the strategies to minimize contamination and improve data reproducibility discussed here, will be critical to ensuring data accuracy. Accordingly, the P10K project already envisions improving the reliability of its genomic data in future developments by providing, for instance, bioinformatic tools for multiple sequence alignment and phylogenetic analysis, made available through the P10K database. Ultimately, this will maximize the impact of the genomic data available through the P10K initiative in addressing major biological questions, including the origins of complex traits, symbioses, multicellularity, and the diversification of eukaryotic life on Earth.

Availability of data and materials

All compiled and curated SSU and COI datasets associated with this manuscript are publicly available on FigShare at https://doi.org/10.6084/m9.figshare.29814947

Competing interests

The authors declare that they have no competing interests

Funding

A.L.P-S and R.E.J were supported by startup funds provided to A.K.T. by Texas Tech University. This work was supported by the National Science Foundation Division of Environmental Biology (2100888) awarded to M.W.B., D.J.G.L. is supported by a FAPESP award #2019/22815-2.

Authors' contributions

Alfredo L. Porfirio-Sousa: Conceptualization, Data curation, Visualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Robert E. Jones**: Investigation, Data curation, Writing – review & editing. **Matthew W. Brown:** Investigation, Writing – review & editing. **Daniel Lahr:** Investigation, Writing – review & editing. **Alexander K. Tice:** Conceptualization, Data curation, Visualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Resources, Supervision, Funding acquisition.

Acknowledgments

The authors acknowledge the High Performance Computing Center (HPCC) at Texas Tech University for providing computational resources that have contributed to the research results reported within this paper. URL: http://www.hpcc.ttu.edu

Declaration of generative AI in scientific writing

During the preparation of this work the authors used ChatGPT GPT-4o to improve the readability and language of the first draft of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

References

1. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences. 2018;115:4325–33. https://doi.org/10.1073/pnas.1720115115.

2. Gao X, Chen K, Xiong J, Zou D, Yang F, Ma Y, et al. The P10K database: a data portal for the protist 10 000 genomes project. Nucleic Acids Research. 2024;52:D747–55. https://doi.org/10.1093/nar/gkad992.

3. Brown MW, Tice AK. A genetic toolbox for marine protists. Nat Methods. 2020;17:469–70. https://doi.org/10.1038/s41592-020-0794-z.

4. Schoenle A, Francis O, Archibald JM, Burki F, Vries J de, Dumack K, et al. Protist genomics: key to understanding eukaryotic evolution. Trends in Genetics. 2025;0. https://doi.org/10.1016/j.tig.2025.05.004.

5. Burki F, Roger AJ, Brown MW, Simpson AGB. The New Tree of Eukaryotes. Trends in Ecology & Evolution. 2020;35:43–55. https://doi.org/10.1016/j.tree.2019.08.008.

6. Williamson K, Eme L, Baños H, McCarthy CGP, Susko E, Kamikawa R, et al. A robustly rooted tree of eukaryotes reveals their excavate ancestry. Nature. 2025;640:974–81. https://doi.org/10.1038/s41586-025-08709-5.

7. López-García P, Moreira D. The Syntrophy hypothesis for the origin of eukaryotes revisited. Nat Microbiol. 2020;5:655–67. https://doi.org/10.1038/s41564-020-0710-4.

8. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biology. 2014;12:e1001889. https://doi.org/10.1371/journal.pbio.1001889.

9. The Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. Proceedings of the National Academy of Sciences. 2022;119:e2115642118. https://doi.org/10.1073/pnas.2115642118.

10. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574:679–85. https://doi.org/10.1038/s41586-019-1693-2.

11. Lahr DJ. An emerging paradigm for the origin and evolution of shelled amoebae, integrating advances from molecular phylogenetics, morphology and paleontology. Mem Inst Oswaldo Cruz. 2021;116:e200620. https://doi.org/10.1590/0074-02760200620.

12. Ribeiro GM, Lahr DJG. Survival in a Changing World: The role of transcriptomics and the urgent need for genomes to understand Arcellinida's adaptive capabilities. Acta Protozoologica. 2025;2024 Volume 63, Special Issue / Early View.

13. Kang S, Tice AK, Spiegel FW, Silberman JD, Pánek T, Čepička I, et al. Between a Pod and a Hard Test: The Deep Evolution of Amoebae. Molecular Biology and Evolution. 2017;34:2258–70. https://doi.org/10.1093/molbev/msx162.

14. Onsbring H, Tice AK, Barton BT, Brown MW, Ettema TJG. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on Giardia intestinalis cells. BMC Genomics. 2020;21:448. https://doi.org/10.1186/s12864-020-06858-7.

15. Porfirio-Sousa AL, Tice AK, Morais L, Ribeiro GM, Blandenier Q, Dumack K, et al. Amoebozoan testate amoebae illuminate the diversity of heterotrophs and the complexity of ecosystems throughout geological time. Proceedings of the National Academy of Sciences. 2024;121:e2319628121. https://doi.org/10.1073/pnas.2319628121.

16. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55. https://doi.org/10.1101/gr.186072.114.

17. Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. Genome Biol. 2024;25:60. https://doi.org/10.1186/s13059-024-03198-7.

18. Eme L, Sharpe SC, Brown MW, Roger AJ. On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. Cold Spring Harb Perspect Biol. 2014;6:a016139. https://doi.org/10.1101/cshperspect.a016139.

19. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research. 2013;41:D597–604. https://doi.org/10.1093/nar/gks1160.

462 20. Ribeiro GM, Useros F, Dumack K, González-Miguéns R, Siemensma F, Porfírio-Sousa AL, et al.
463 Expansion of the cytochrome C oxidase subunit I database and description of four new lobose testate
464 amoebae species (Amoebozoa; Arcellinida). European Journal of Protistology. 2023;91:126013.
465 https://doi.org/10.1016/j.ejop.2023.126013.

466 21. Kosakyan A, Gomaa F, Lara E, Lahr DJG. Current and future perspectives on the systematics,
467 taxonomy and nomenclature of testate amoebae. European Journal of Protistology. 2016;55:105–17.
468 https://doi.org/10.1016/j.ejop.2016.02.001.

469 22. González-Miguéns R, Todorov M, Blandenier Q, Duckert C, Porfirio-Sousa AL, Ribeiro GM, et al.
470 Deconstructing *Difflugia*: The tangled evolution of lobose testate amoebae shells (Amoebozoa:
471 Arcellinida) illustrates the importance of convergent evolution in protist phylogeny. Molecular
472 Phylogenetics and Evolution. 2022;175:107557. https://doi.org/10.1016/j.ympev.2022.107557.

473 23. Lahr DJG, Kosakyan A, Lara E, Mitchell EAD, Morais L, Porfirio-Sousa AL, et al. Phylogenomics
474 and Morphological Reconstruction of Arcellinida Testate Amoebae Highlight Diversity of Microbial
475 Eukaryotes in the Neoproterozoic. Current Biology. 2019;29:991-1001.e3.
476 https://doi.org/10.1016/j.cub.2019.01.078.

477 24. Schuler GA, Brown MW. Description of Armaparvus languidus n. gen. n. sp. Confirms
478 Ultrastructural Unity of Cutosea (Amoebozoa, Evosea). Journal of Eukaryotic Microbiology.
479 2019;66:158–66. https://doi.org/10.1111/jeu.12640.

480 25. Burki F, Sandin MM, Jamy M. Diversity and ecology of protists revealed by metabarcoding.
481 Current Biology. 2021;31:R1267–80. https://doi.org/10.1016/j.cub.2021.07.066.

482 26. Kosakyan A. Towards testate amoebae genomics and beyond, a wish list…. Acta Protozoologica.
483 2025;2024 Volume 63, Special Issue:41–7.

484 27. Francois CM, Durand F, Figuet E, Galtier N. Prevalence and Implications of Contamination in
485 Public Genomic Resources: A Case Study of 43 Reference Arthropod Assemblies. G3 (Bethesda).
486 2020;10:721–30. https://doi.org/10.1534/g3.119.400758.

487 28. Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. Genome Biol.
488 2022;23:60. https://doi.org/10.1186/s13059-022-02619-9.

489 29. Tice AK, Žihala D, Pánek T, Jones RE, Salomaki ED, Nenarokov S, et al. PhyloFisher: A
490 phylogenomic package for resolving eukaryotic relationships. PLOS Biology. 2021;19:e3001365.
491 https://doi.org/10.1371/journal.pbio.3001365.

492 30. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet.
493 2012;13:303–14. https://doi.org/10.1038/nrg3186.

494 31. Emms DM, Kelly S. SHOOT: phylogenetic gene search and ortholog inference. Genome Biology.
495 2022;23:85. https://doi.org/10.1186/s13059-022-02652-8.

496 32. Smith SA, Pease JB. Heterogeneous molecular processes among the causes of how sequence
497 similarity scores can fail to recapitulate phylogeny. Brief Bioinform. 2017;18:451–7.
498 https://doi.org/10.1093/bib/bbw034.

499

500

501

502

503

Figures

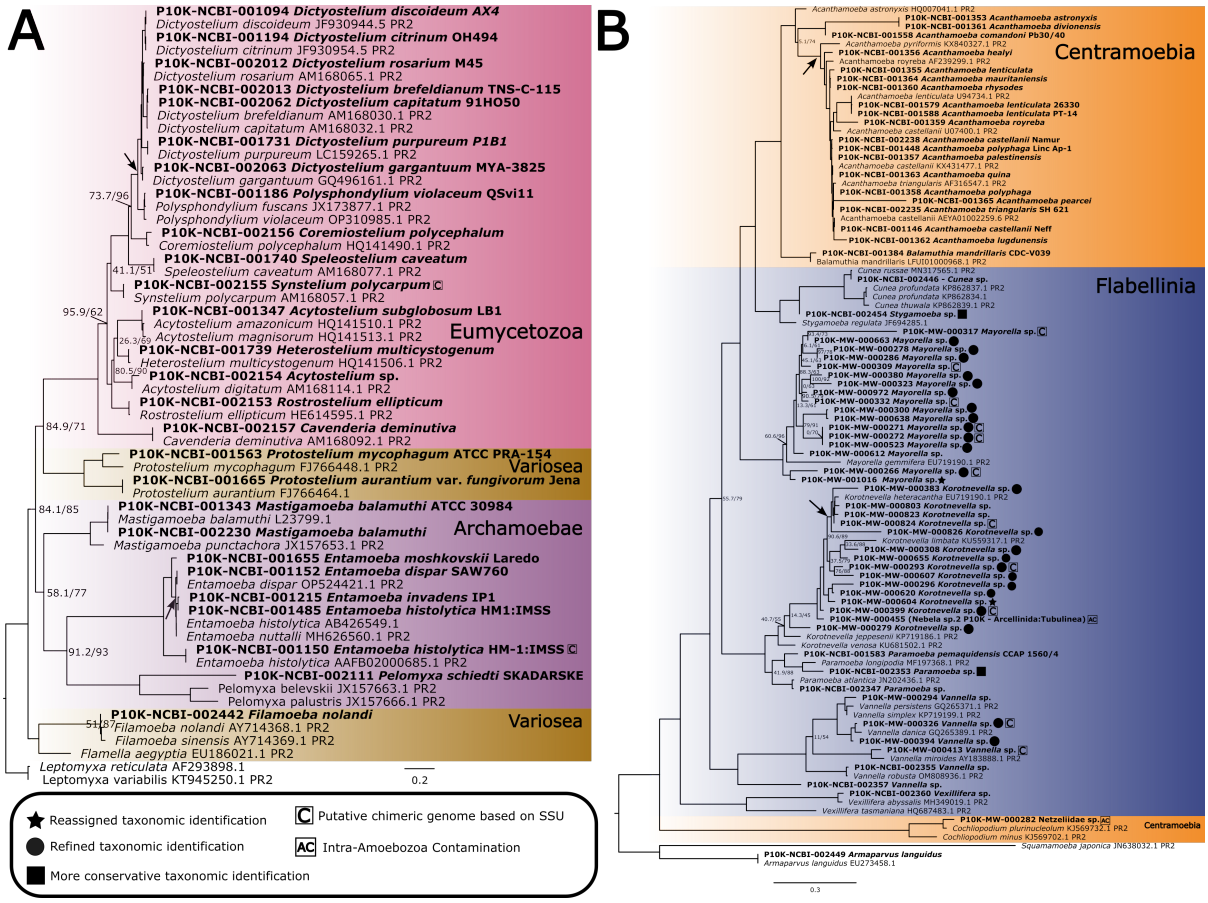


Figure 1. The phylogenetic placement of Amoebozoa genomic data from the P10K focused on the major groups Evosea and Discosea. Maximum-likelihood phylogenetic trees constructed from the Small Subunit ribosomal RNA (SSU) inferred from a subset of the curated dataset presented in Figure S1. **A.** Focuses on the major Amoebozoa group Evosea. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (TIM2+F+R4). **B.** Focuses on the major Amoebozoa group Discosea. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (GTR+F+G4). Node support was assessed using both ultrafast bootstrap (UFBoot) and the Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT). Support values are reported as SH-aLRT / UFBoot, with values ≥80/95 considered indicative of strong support. For clarity, high-support values are omitted in this figure, as well as support values for nodes above the nodes indicated by the arrows, which are represented mostly by flat branches. The complete tree, including all support values, is shown in Figures S2 and S3. Stars indicate genomic data reassigned to a different taxonomic identity than reported in the P10K database. Filled circles mark indicate cases with refined taxonomic resolution. Filled squares denote more conservative identifications (e.g., genus or family level) rather than the more specific genus or species-level identification originally provided in the P10K database. 'C' indicates putative chimeric genomes containing sequences from multiple eukaryotes and 'AC' denotes intra-Amoebozoa contamination.

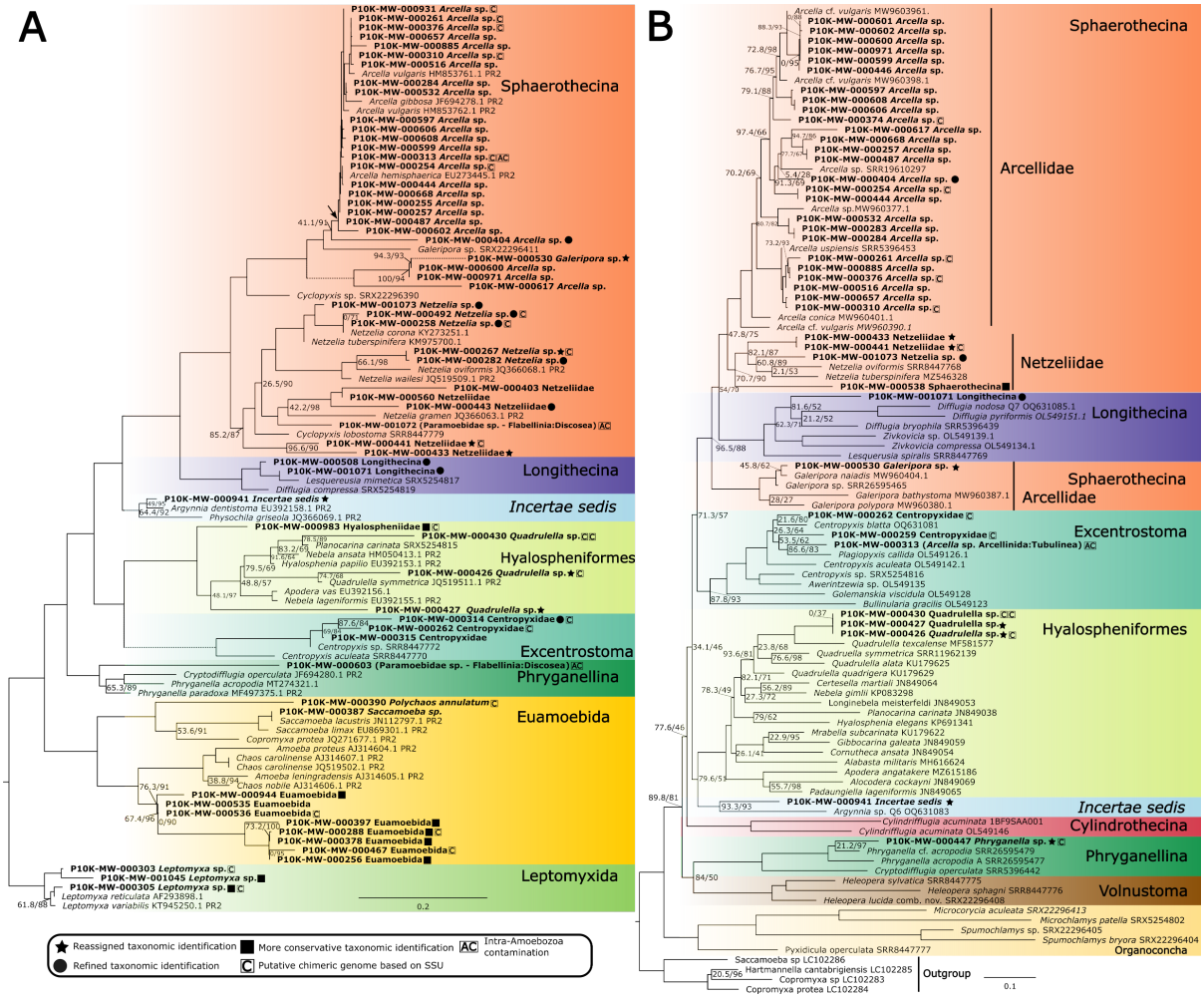


Figure 2. The phylogenetic placement of Amoebozoa genomic data from the P10K focused on the major group Tubulinea. A. The maximum-likelihood phylogenetic tree constructed from the Small Subunit ribosomal RNA (SSU) inferred from a subset of the curated dataset presented in Figure S1, focuses on the major Amoebozoa group Tubulinea. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (TIM3e+G4). The length of branches depicted as dashed lines have been reduced by 50% for presentation purposes. **B.** The maximum-likelihood phylogenetic tree of cytochrome c oxidase subunit I (COI) inferred from a curated dataset generated in the present study, focusing on Arcellinida order (Tubulinea:Amoebozoa) comprising COI sequences retrieved from genomes and transcriptomes available in the P10K database, along with reference sequences made available by previous. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (GTR+F+I+G4). Node support was assessed using both ultrafast bootstrap (UFBoot) and the Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT). Support values are reported as SH-aLRT / UFBoot, with values ≥80/95 considered indicative of strong support. For clarity, high-support values are omitted in this figure, as well as support values for nodes above the one indicated by the arrow, which are represented mostly by flat branches. The complete tree, including all support values, is shown in Figures S4 and S5. Stars indicate genomic data reassigned to a different taxonomic identity than reported in the P10K database. Filled circles mark indicate cases with refined taxonomic resolution. Filled squares denote more conservative identifications (e.g., genus or family level) rather than the more specific genus or species-level identification originally provided in the P10K database. ‘C’ indicates putative chimeric genomes containing sequences from multiple eukaryotes and ‘AC’ denotes intra-Amoebozoa contamination.

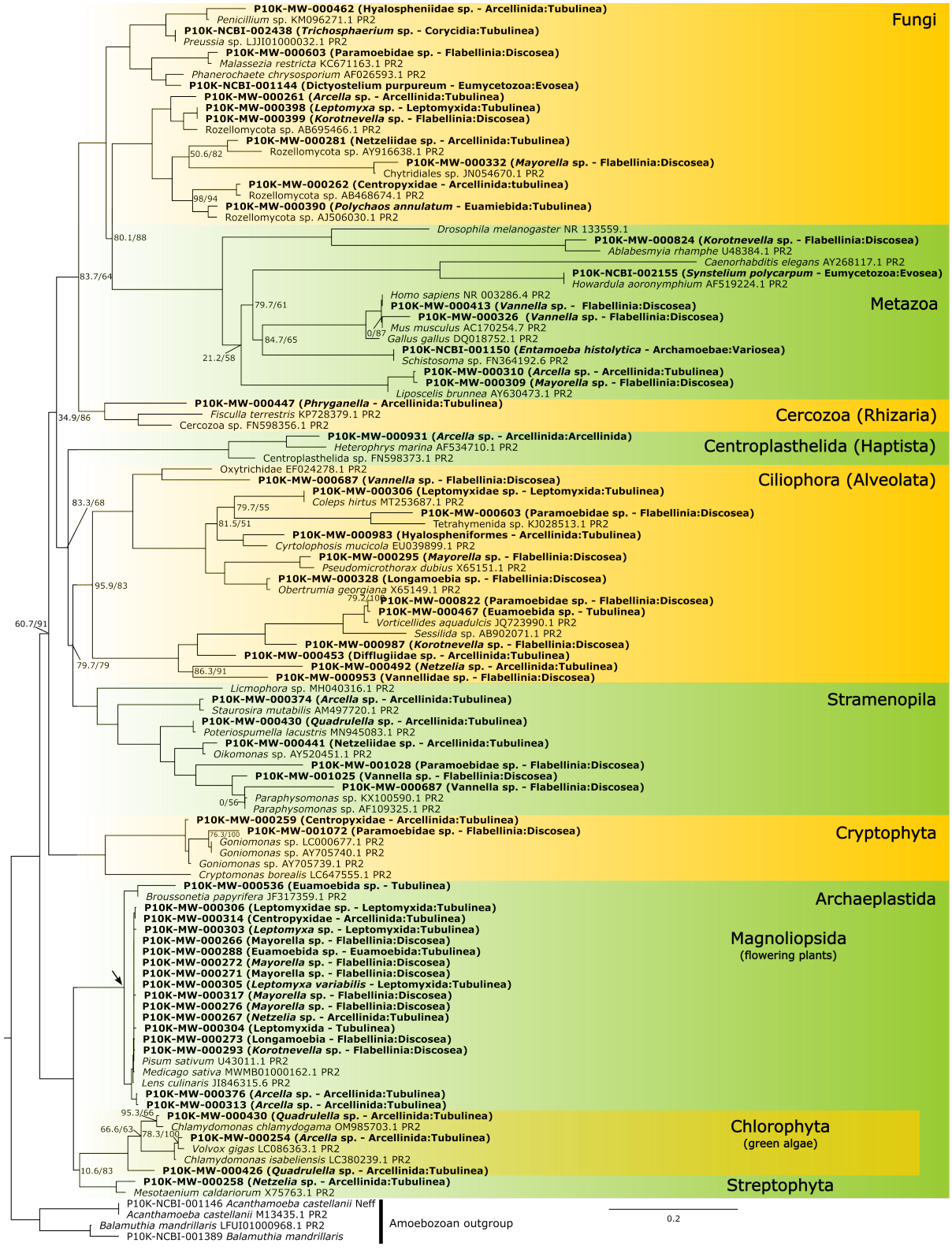


Figure 3. Contamination-screening phylogenetic tree. The maximum-likelihood phylogenetic tree constructed from the Small Subunit ribosomal RNA (SSU), inferred from a curated dataset generated in the present study, comprises SSU sequences of putative non-Amoebozoan contaminants retrieved from genomes and transcriptomes available in the P10K database, along with reference sequences from the PR^2 database (indicated by PR^2 IDs), identified through BLAST+ similarity searches using the putative contaminant SSU sequences as queries. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (TN+F+I+G4). Node

559 support was assessed using both ultrafast bootstrap (UFBoot) and the Shimodaira–Hasegawa
560 approximate likelihood ratio test (SH-aLRT). Support values are reported as SH-aLRT / UFBoot, with
561 values ≥80/95 considered indicative of strong support. For clarity, high-support values are omitted in
562 this figure, as well as support values for nodes above the one indicated by the arrow, which are
563 represented mostly by flat branches. The complete tree, including all support values, is shown in Figure
564 S6. The taxonomic identification of Amoebozoa taxa corresponding to the displayed P10K ID codes is
565 shown in parentheses.

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

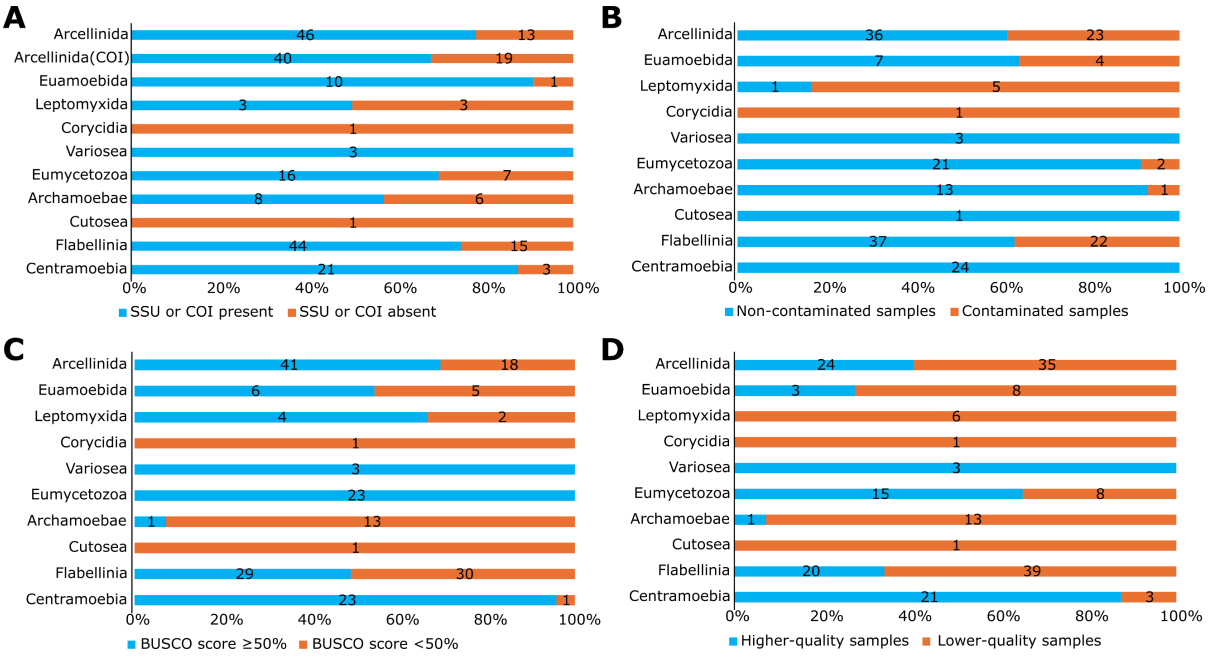
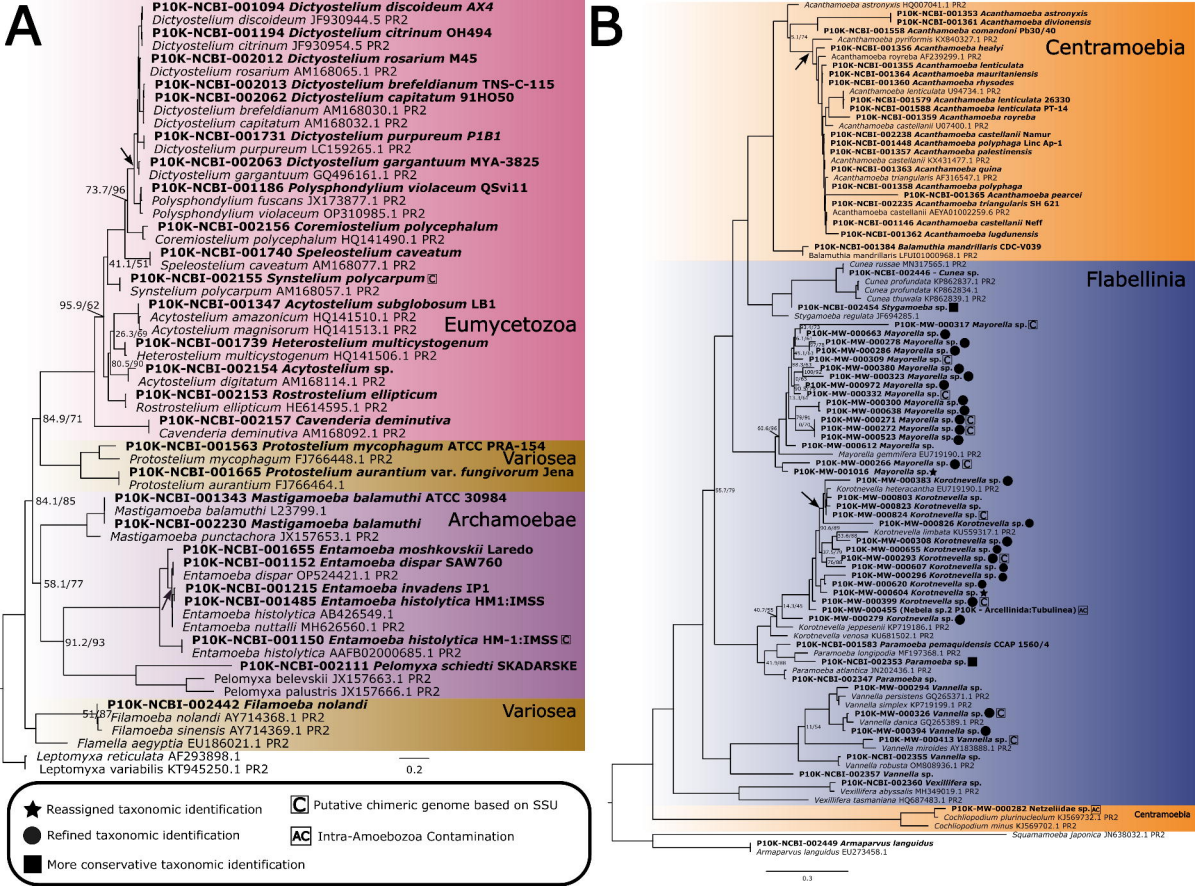
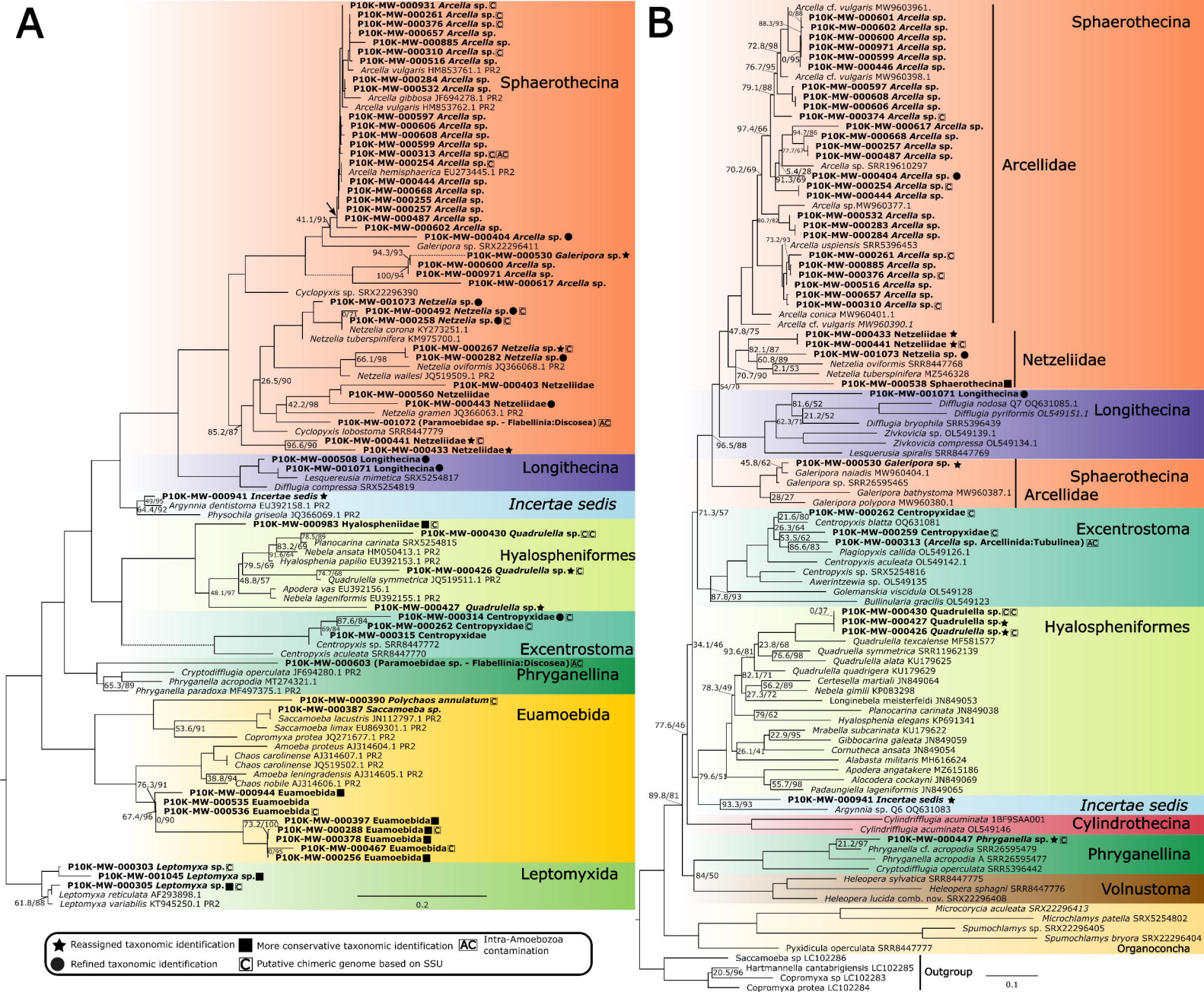
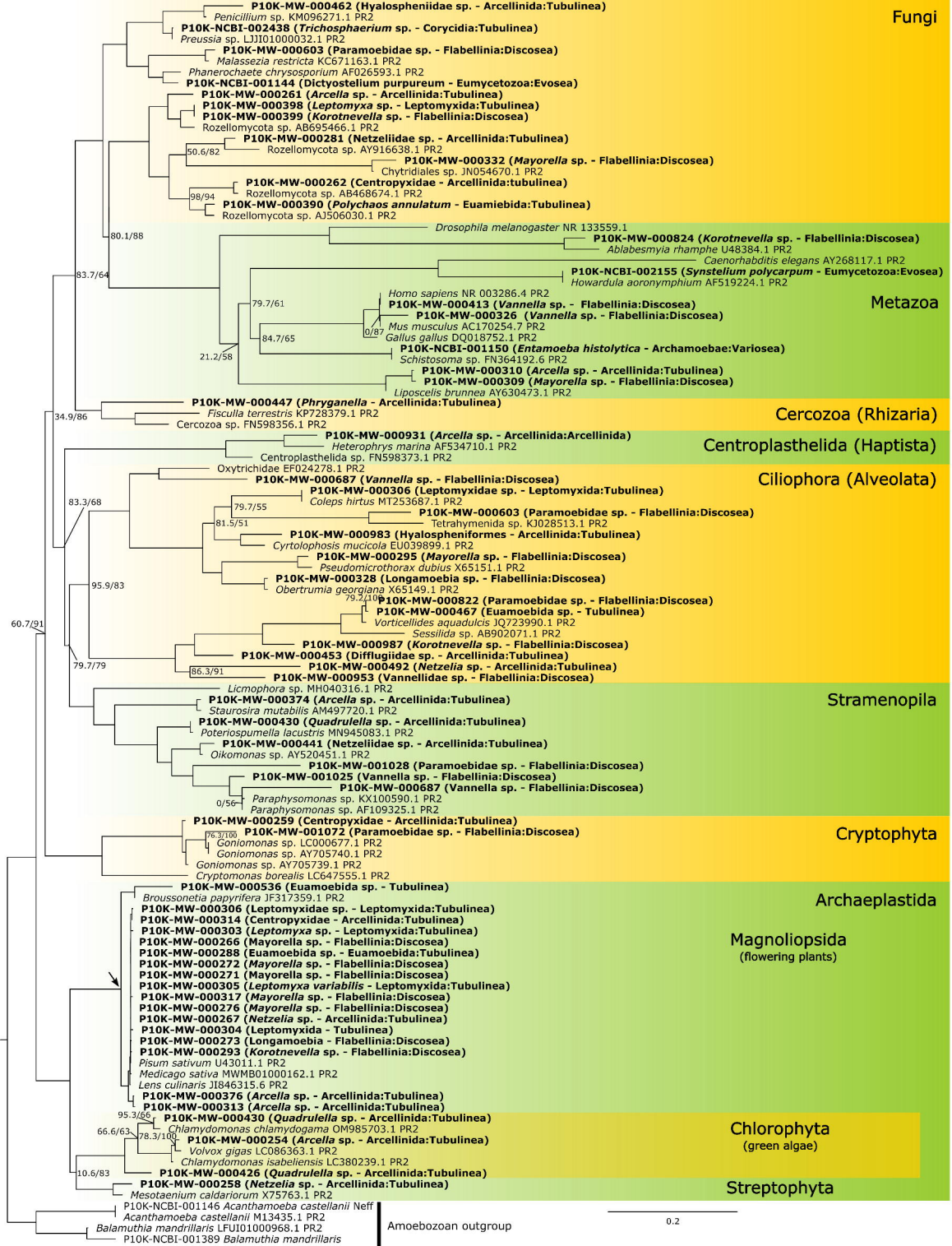
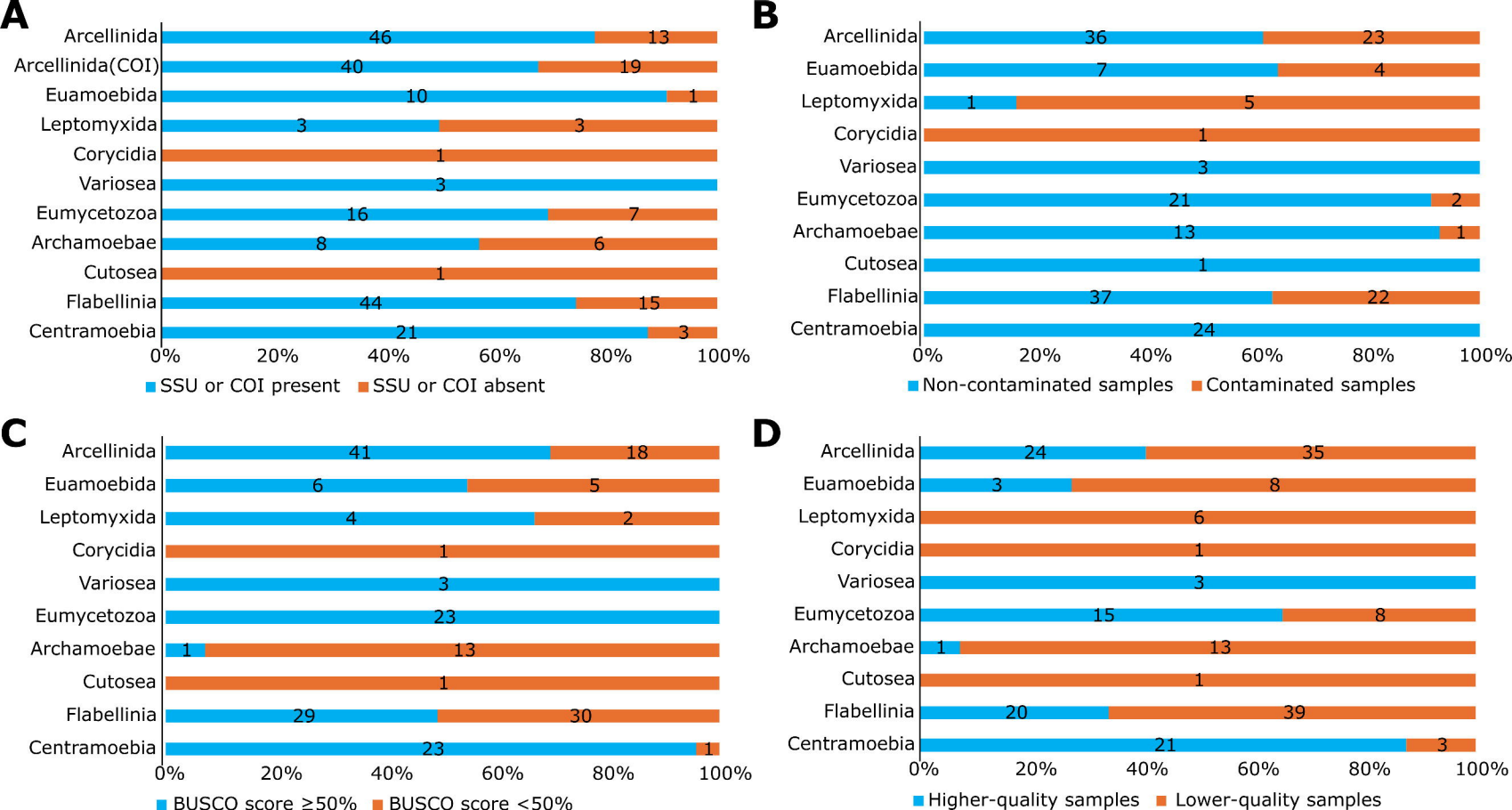


Figure 4. Summary of Amoebozoa P10K database sample counts by taxonomic group, focusing on the key parameters used to evaluate the data. A. presence of SSU or COI. **B.** BUSCO completeness score. We considered the BUSCO score of each sample as originally provided in the P10K database and reported in Gao et al. (2024). **C.** contamination by another eukaryotic lineage. **D.** sample quality assessment. Higher-quality samples are defined as those with SSU (and COI in the case of Arcellinida), a BUSCO score ≥50%, and no contamination identified based on the SSU. Lower-quality samples are those that require decontamination or further sequencing prior to reliable downstream analysis. The bars represent the total number of Amoebozoa samples in the P10K database available for each taxonomic group. The samples considered as higher quality are highlighted in the Supplementary Information – Table S1.


A
P10K-NCBI-001094 Dictyostelium discoideum AX4
Dictyostelium discoideum JF930944.5 PR2
P10K-NCBI-001194 Dictyostelium citrinum OH494
Dictyostelium citrinum JF930954.5 PR2
P10K-NCBI-002012 Dictyostelium rosarium M45
Dictyostelium rosarium AM168065.1 PR2
P10K-NCBI-002013 Dictyostelium brefeldianum TNS-C-115
P10K-NCBI-002062 Dictyostelium capitatum 91HO50
Dictyostelium brefeldianum AM168030.1 PR2
Dictyostelium capitatum AM168032.1 PR2
P10K-NCBI-001731 Dictyostelium purpureum P1B1
Dictyostelium purpureum LC159265.1 PR2
P10K-NCBI-002063 Dictyostelium gargantuum MYA-3825
Dictyostelium gargantuum GQ496161.1 PR2
73.7/96
P10K-NCBI-001186 Polysphondylium violaceum QSvi11
Polysphondylium fuscans JX173877.1 PR2
Polysphondylium violaceum OP310985.1 PR2
P10K-NCBI-002156 Coremiostelium polycephalum
Coremiostelium polycephalum HQ141490.1 PR2
P10K-NCBI-001740 Speleostelium caveatum
41.1/51
Speleostelium caveatum AM168077.1 PR2
P10K-NCBI-002155 Synstelium polycarpum C
Synstelium polycarpum AM168057.1 PR2
95.9/62
P10K-NCBI-001347 Acytostelium subglobosum LB1
Acytostelium amazonicum HQ141510.1 PR2
126.3/69
Acytostelium magnisorum HQ141513.1 PR2
Eumycetozoa
P10K-NCBI-001739 Heterostelium multicystogenum
Heterostelium multicystogenum HQ141506.1 PR2
80.5/99
P10K-NCBI-002154 Acytostelium sp.
Acytostelium digitatum AM168114.1 PR2
P10K-NCBI-002153 Rostrostelium ellipticum
Rostrostelium ellipticum HE614595.1 PR2
84.9/71
P10K-NCBI-002157 Cavenderia deminutiva
Cavenderia deminutiva AM168092.1 PR2
P10K-NCBI-001563 Protostelium mycophagum ATCC PRA-154
Protostelium mycophagum FJ766448.1 PR2
Variosea
P10K-NCBI-001665 Protostelium aurantium var. fungivorum Jena
Protostelium aurantium FJ766464.1
84.1/85
P10K-NCBI-001343 Mastigamoeba balamuthi ATCC 30984
Mastigamoeba balamuthi L23799.1
P10K-NCBI-002230 Mastigamoeba balamuthi
Archamoebae
Mastigamoeba punctachora JX157653.1 PR2
P10K-NCBI-001655 Entamoeba moshkovskii Laredo
P10K-NCBI-001152 Entamoeba dispar SAW760
Entamoeba dispar OP524421.1 PR2
58.1/77
P10K-NCBI-001215 Entamoeba invadens IP1
P10K-NCBI-001485 Entamoeba histolytica HM1:IMSS
Entamoeba histolytica AB426549.1
Entamoeba nuttalli MH626560.1 PR2
91.2/93
P10K-NCBI-001150 Entamoeba histolytica HM-1:IMSS C
Entamoeba histolytica AAFB02000685.1 PR2
P10K-NCBI-002111 Pelomyxa schiedti SKADARSKE
Pelomyxa belevskii JX157663.1 PR2
Pelomyxa palustris JX157666.1 PR2
51/87
P10K-NCBI-002442 Filamoeba nolandi
Filamoeba nolandi AY714368.1 PR2
Filamoeba sinensis AY714369.1 PR2
Variosea
Flamella aegyptia EU186021.1 PR2
Leptomyxa reticulata AF293898.1
Leptomyxa variabilis KT945250.1 PR2
0.2
Reassigned taxonomic identification
Refined taxonomic identification
More conservative taxonomic identification
C Putative chimeric genome based on SSU
AC Intra-Amoebozoa Contamination
B
Acanthamoeba astronyxis HQ007041.1 PR2
P10K-NCBI-001353 Acanthamoeba astronyxis
P10K-NCBI-001361 Acanthamoeba divionensis
P10K-NCBI-001558 Acanthamoeba comandoni Pb30/40
5.1/74
Acanthamoeba pyriformis KX834271.1 PR2
Centramoebia
P10K-NCBI-001356 Acanthamoeba healyi
Acanthamoeba royreba AF239299.1 PR2
P10K-NCBI-001355 Acanthamoeba lenticulata
P10K-NCBI-001364 Acanthamoeba mauritaniensis
P10K-NCBI-001360 Acanthamoeba rhysodes
Acanthamoeba lenticulata U94734.1 PR2
P10K-NCBI-001579 Acanthamoeba lenticulata 26330
P10K-NCBI-001588 Acanthamoeba lenticulata PT-14
P10K-NCBI-001359 Acanthamoeba royreba
Acanthamoeba castellanii U07400.1 PR2
P10K-NCBI-002238 Acanthamoeba castellanii Namur
P10K-NCBI-001448 Acanthamoeba polyphaga Linc Ap-1
P10K-NCBI-001357 Acanthamoeba palestinensis
Acanthamoeba castellanii KX431477.1 PR2
P10K-NCBI-001363 Acanthamoeba quina
Acanthamoeba triangularis AF3165471 PR2
P10K-NCBI-001358 Acanthamoeba polyphaga
P10K-NCBI-001365 Acanthamoeba pearcei
P10K-NCBI-002235 Acanthamoeba triangularis SH 621
Acanthamoeba castellanii AEYA01002259.6 PR2
P10K-NCBI-001146 Acanthamoeba castellanii Neff
P10K-NCBI-001362 Acanthamoeba lugdunensis
P10K-NCBI-001384 Balamuthia mandrillaris CDC-V039
Balamuthia mandrillaris LFU01000968.1 PR2
Cunea rossae MN317565.1 PR2
P10K-NCBI-002446 – Cunea sp.
Cunea profundata KP862837.1 PR2
Cunea profundata KP862834.1
Cunea thuwala KP862839.1 PR2
P10K-NCBI-002454 Stygamoeba sp.
Stygamoeba regulata JF694285.1
Flabellinia
P10K-MW-000317 Mayorella sp. C
P10K-MW-000663 Mayorella sp.
P10K-MW-000278 Mayorella sp.
P10K-MW-000286 Mayorella sp.
P10K-MW-000309 Mayorella sp. C
P10K-MW-000380 Mayorella sp.
P10K-MW-000323 Mayorella sp.
P10K-MW-000972 Mayorella sp.
P10K-MW-000332 Mayorella sp. C
P10K-MW-000300 Mayorella sp.
P10K-MW-000638 Mayorella sp.
P10K-MW-000271 Mayorella sp. C
P10K-MW-000272 Mayorella sp. C
P10K-MW-000523 Mayorella sp.
P10K-MW-000612 Mayorella sp.
Mayorella gemmifera EU719190.1 PR2
P10K-MW-000266 Mayorella sp. C
P10K-MW-001016 Mayorella sp.
P10K-MW-000383 Korotnevella sp.
Korotnevella heteracantha EU719190.1 PR2
P10K-MW-000803 Korotnevella sp.
P10K-MW-000823 Korotnevella sp.
P10K-MW-000824 Korotnevella sp. C
P10K-MW-000826 Korotnevella sp.
Korotnevella limbata KU559317.1 PR2
P10K-MW-000308 Korotnevella sp.
P10K-MW-000555 Korotnevella sp.
P10K-MW-000293 Korotnevella sp. C
P10K-MW-000607 Korotnevella sp.
P10K-MW-000298 Korotnevella sp.
P10K-MW-000620 Korotnevella sp.
P10K-MW-000604 Korotnevella sp.
P10K-MW-000399 Korotnevella sp. C
P10K-MW-000455 (Nebela sp.2 P10K - Arcellinida:Tubulinea) AC
P10K-MW-000279 Korotnevella sp.
Korotnevella jeppesenii KP719186.1 PR2
Korotnevella venosa KU681502.1 PR2
P10K-NCBI-001583 Paramoeba pemaquidensis CCAP 1560/4
Paramoeba longipodia MF197368.1 PR2
P10K-NCBI-002353 Paramoeba sp.
Paramoeba atlantica JN224436.1 PR2
P10K-NCBI-002347 Paramoeba sp.
P10K-MW-000294 Vannella sp.
Vannella persistens GQ265371.1 PR2
Vannella simplex KP719199.1 PR2
P10K-MW-000326 Vannella sp. C
Vannella dania GQ265389.1 PR2
P10K-MW-000394 Vannella sp.
P10K-MW-000413 Vannella sp. C
Vannella mirabiles AY183888.1 PR2
P10K-NCBI-002355 Vannella sp.
Vannella robusta OM808936.1 PR2
P10K-NCBI-002357 Vannella sp.
P10K-NCBI-002360 Vexillifera sp.
Vexillifera abyssalis MH349019.1 PR2
Vexillifera tasmaniana HQ687483.1 PR2
P10K-MW-000282 Netzelitiidae sp. AC
Centramoebia
Cochliopodium pluriunucleolum KJ569732.1 PR2
Cochliopodium minus KJ569702.1 PR2
Squamamoeba japonica JN638032.1 PR2
P10K-NCBI-002449 Armaparvus languidus
Armaparvus languidus EU273458.1
0.3


A
B
Sphaerothecina
Arcellidae
Netzeliidae
Longithecina
Sphaerothecina
Arcellidae
Excentrostoma
Hyalospheniformes
Incertae sedis
Cylindrothecina
Phryganellina
Volnustoma
Organoconcha
Outgroup
Hyalospheniformes
Incertae sedis
Euamoebida
Leptomyxida
Reassigned taxonomic identification
Refined taxonomic identification
More conservative taxonomic identification
Intra-Amoebozoa contamination
Putative chimeric genome based on SSU


P10K-MW-000462 (Hyalospheniidae sp. - Arcellinida:Tubulinea)
Penicillium sp. KM096271.1 PR2
P10K-NCBI-002438 (Trichosphaerium sp. - Corycidia:Tubulinea)
Preussia sp. LJJI01000032.1 PR2
P10K-MW-000603 (Paramoebidae sp. - Flabellinia:Discosea)
Malassezia restricta KC671163.1 PR2
Phanerochaete chrysosporium AF026593.1 PR2
P10K-NCBI-001144 (Dictyostelium purpureum - Eumycetozoa:Evosea)
P10K-MW-000261 (Arcella sp. - Arcellinida:Tubulinea)
P10K-MW-000398 (Leptomyxa sp. - Leptomyxida:Tubulinea)
P10K-MW-000399 (Korotnevella sp. - Flabellinia:Discosea)
Rozellomycota sp. AB695466.1 PR2
P10K-MW-000281 (Netzeliidae sp. - Arcellinida:Tubulinea)
50.6/82
Rozellomycota sp. AY916638.1 PR2
P10K-MW-000332 (Mayorella sp. - Flabellinia:Discosea)
Chytridiales sp. JN054670.1 PR2
P10K-MW-000262 (Centropyxidae - Arcellinida:tubulinea)
98/94
Rozellomycota sp. AB468674.1 PR2
P10K-MW-000390 (Polychaos annulatum - Euamiebida:Tubulinea)
Rozellomycota sp. AJ506030.1 PR2
Fungi
Drosophila melanogaster NR 133559.1
80.1/88
P10K-MW-000824 (Korotnevella sp. - Flabellinia:Discosea)
Ablabesmyia rhamphe U48384.1 PR2
Caenorhabditis elegans AY268117.1 PR2
83.7/64
P10K-NCBI-002155 (Synstelium polycarpum - Eumycetozoa:Evosea)
Howardula aoronymphium AF519224.1 PR2
Homo sapiens NR 003286.4 PR2
79.7/61
P10K-MW-000413 (Vannella sp. - Flabellinia:Discosea)
P10K-MW-000326 (Vannella sp. - Flabellinia:Discosea)
Mus musculus AC170254.7 PR2
0/87
84.7/65
Gallus gallus DQ018752.1 PR2
21.2/58
P10K-NCBI-001150 (Entamoeba histolytica - Archamoebae:Variosea)
Schistosoma sp. FN364192.6 PR2
P10K-MW-000310 (Arcella sp. - Arcellinida:Tubulinea)
P10K-MW-000309 (Mayorella sp. - Flabellinia:Discosea)
Liposcelis brunnea AY630473.1 PR2
Metazoa
P10K-MW-000447 (Phryganella - Arcellinida:Tubulinea)
34.9/86
Fisculla terrestris KP728379.1 PR2
Cercozoa sp. FN598356.1 PR2
Cercozoa (Rhizaria)
P10K-MW-000931 (Arcella sp. - Arcellinida:Arcellinida)
Heterophrys marina AF534710.1 PR2
Centroplasthelida sp. FN598373.1 PR2
Centroplasthelida (Haptista)
Oxytrichidae EF024278.1 PR2
P10K-MW-000687 (Vannella sp. - Flabellinia:Discosea)
P10K-MW-000306 (Leptomyxidae sp. - Leptomyxida:Tubulinea)
83.3/68
Coleps hirtus MT253687.1 PR2
79.7/55
P10K-MW-000603 (Paramoebidae sp. - Flabellinia:Discosea)
81.5/51
Tetrahymenida sp. KJ028513.1 PR2
P10K-MW-000983 (Hyalospheniformes - Arcellinida:Tubulinea)
Cyrtolophosis mucicola EU039899.1 PR2
P10K-MW-000295 (Mayorella sp. - Flabellinia:Discosea)
Pseudomicrothorax dubius X65151.1 PR2
95.9/83
P10K-MW-000328 (Longamoebia sp. - Flabellinia:Discosea)
Obertrumia georgiana X65149.1 PR2
79.2/100
P10K-MW-000822 (Paramoebidae sp. - Flabellinia:Discosea)
P10K-MW-000467 (Euamoebida sp. - Tubulinea)
60.7/91
Vorticellides aquadulcis JQ723990.1 PR2
Sessilida sp. AB902071.1 PR2
P10K-MW-000987 (Korotnevella sp. - Flabellinia:Discosea)
P10K-MW-000453 (Difflugiidae sp. - Arcellinida:Tubulinea)
79.7/79
P10K-MW-000492 (Netzelia sp. - Arcellinida:Tubulinea)
86.3/91
P10K-MW-000953 (Vannellidae sp. - Flabellinia:Discosea)
Ciliophora (Alveolata)
Licmophora sp. MH040316.1 PR2
P10K-MW-000374 (Arcella sp. - Arcellinida:Tubulinea)
Staurosira mutabilis AM497720.1 PR2
P10K-MW-000430 (Quadrulella sp. - Arcellinida:Tubulinea)
Poteriospumella lacustris MN945083.1 PR2
P10K-MW-000441 (Netzeliidae sp. - Arcellinida:Tubulinea)
Oikomonas sp. AY520451.1 PR2
P10K-MW-001028 (Paramoebidae sp. - Flabellinia:Discosea)
P10K-MW-001025 (Vannella sp. - Flabellinia:Discosea)
P10K-MW-000687 (Vannella sp. - Flabellinia:Discosea)
0/56
Paraphysomonas sp. KX100590.1 PR2
Paraphysomonas sp. AF109325.1 PR2
Stramenopila
P10K-MW-000259 (Centropyxidae - Arcellinida:Tubulinea)
76.3/100
P10K-MW-001072 (Paramoebidae sp. - Flabellinia:Discosea)
Goniomonas sp. LC000677.1 PR2
Goniomonas sp. AY705740.1 PR2
Goniomonas sp. AY705739.1 PR2
Cryptomonas borealis LC647555.1 PR2
Cryptophyta
P10K-MW-000536 (Euamoebida sp. - Tubulinea)
Broussonetia papyrifera JF317359.1 PR2
P10K-MW-000306 (Leptomyxidae sp. - Leptomyxida:Tubulinea)
P10K-MW-000314 (Centropyxidae - Arcellinida:Tubulinea)
P10K-MW-000303 (Leptomyxa sp. - Leptomyxida:Tubulinea)
P10K-MW-000266 (Mayorella sp. - Flabellinia:Discosea)
P10K-MW-000288 (Euamoebida sp. - Euamoebida:Tubulinea)
P10K-MW-000272 (Mayorella sp. - Flabellinia:Discosea)
P10K-MW-000271 (Mayorella sp. - Flabellinia:Discosea)
P10K-MW-000305 (Leptomyxa variabilis - Leptomyxida:Tubulinea)
P10K-MW-000317 (Mayorella sp. - Flabellinia:Discosea)
P10K-MW-000276 (Mayorella sp. - Flabellinia:Discosea)
P10K-MW-000267 (Netzelia sp. - Arcellinida:Tubulinea)
P10K-MW-000304 (Leptomyxida - Tubulinea)
P10K-MW-000273 (Longamoebia - Flabellinia:Discosea)
P10K-MW-000293 (Korotnevella sp. - Flabellinia:Discosea)
Pisum sativum U43011.1 PR2
Medicago sativa MWMB01000162.1 PR2
Lens culinaris JI846315.6 PR2
P10K-MW-000376 (Arcella sp. - Arcellinida:Tubulinea)
P10K-MW-000313 (Arcella sp. - Arcellinida:Tubulinea)
Archaeplastida
Magnoliopsida
(flowering plants)
95.3/66
P10K-MW-000430 (Quadrulella sp. - Arcellinida:Tubulinea)
66.6/63
Chlamydomonas chlamydogama OM985703.1 PR2
78.3/100
P10K-MW-000254 (Arcella sp. - Arcellinida:Tubulinea)
Volvox gigas LC086363.1 PR2
Chlamydomonas isabeliensis LC380239.1 PR2
10.6/83
P10K-MW-000426 (Quadrulella sp. - Arcellinida:Tubulinea)
P10K-MW-000258 (Netzelia sp. - Arcellinida:Tubulinea)
Mesotaenium caldariorum X75763.1 PR2
Chlorophyta
(green algae)
Streptophyta
P10K-NCBI-001146 Acanthamoeba castellanii Neff
Acanthamoeba castellanii M13435.1 PR2
Balamuthia mandrillaris LFUI01000968.1 PR2
P10K-NCBI-001389 Balamuthia mandrillaris
Amoebozoan outgroup
0.2


A
Arcellinida
Arcellinida(COI)
Euamoebida
Leptomyxida
Corycidia
Variosea
Eumycetozoa
Archamoebae
Cutosea
Flabellinia
Centramoebia
46
13
40
19
10
1
3
3
1
3
16
7
8
6
1
44
15
21
3
0%
20%
40%
60%
80%
100%
SSU or COI present
SSU or COI absent
B
Arcellinida
Euamoebida
Leptomyxida
Corycidia
Variosea
Eumycetozoa
Archamoebae
Cutosea
Flabellinia
Centramoebia
36
23
7
4
1
5
1
3
21
2
13
1
1
37
22
24
0%
20%
40%
60%
80%
100%
Non-contaminated samples
Contaminated samples
C
Arcellinida
Euamoebida
Leptomyxida
Corycidia
Variosea
Eumycetozoa
Archamoebae
Cutosea
Flabellinia
Centramoebia
41
18
6
5
4
2
1
3
23
1
13
1
29
30
23
1
0%
20%
40%
60%
80%
100%
BUSCO score ≥50%
BUSCO score <50%
D
Arcellinida
Euamoebida
Leptomyxida
Corycidia
Variosea
Eumycetozoa
Archamoebae
Cutosea
Flabellinia
Centramoebia
24
35
3
8
6
1
3
15
8
1
13
1
20
39
21
3
0%
20%
40%
60%
80%
100%
Higher-quality samples
Lower-quality samples