

Supplementary Information

CSI-SSU: Phylogenetic contamination screening of genomic datasets, demonstrated on 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

* Contributed equally

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

This PDF file includes:

Table S1 legend

File S1 legend

Supplementary Figures S1 – S8

Supplementary dataset legends (Data S1 – S8)

Table S1. Summary of genomic-level data from the Protist 10,000 Genomes (P10K) initiative, including all represented supergroups (Amoebozoa, Archaeplastida, Excavata, Haptista, Obazoa, and TSAR). The table includes species names, sample and assembly identifiers, contamination assessment based on the CSI-SSU tool, bacterial BUSCO results, chimeric SSU identification, and basic assembly and annotation metrics. Assembly quality metrics comprise total assembly size (Mb), number of scaffolds or transcripts, N50 (bp), genome completeness (% BUSCO), predicted gene count, CDS completeness (percentage of genes with both start and stop codons), and final annotation quality level (High, Medium, Low). These metrics were not generated in the present study but are reported as provided in the P10K metadata (Gao et al., 2024). In addition to CSI-SSU results, the table includes, for Amoebozoa, the detection of SSU markers (and COI for Arcellinida) and phylogenetically informed taxonomic identification based on SSU and COI phylogenetic reconstructions.

File S1. Custom Python script used to automate local BLAST+ searches across a folder of genome or transcriptome FASTA files. For each input query sequence, the script retrieves the top *N* high-scoring hits per target file. Optionally, it can extract user-defined flanking regions around each hit, reorient the retrieved sequences to match the query strand, and save the output in FASTA format. This tool facilitates high-throughput retrieval of target regions from large-scale genomic datasets through similarity search and was used in this study to extract targeted SSU and COI sequences from assembled genomes and transcriptomes of Amoebozoa available in the P10K database.

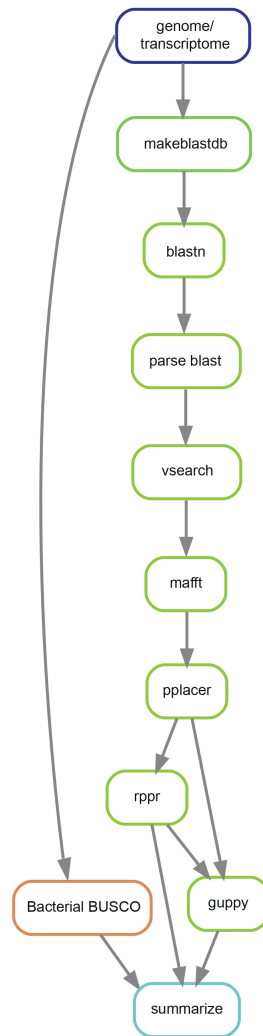


Figure S1. Overview of the CSI-SSU tool workflow. CSI-SSU is a command-line pipeline for screening SSU (18S rRNA) sequences from genomic and transcriptomic assemblies. BLAST+ v2.17.0 databases are generated from input assemblies (makeblastdb) and queried using reference SSU sequences (blastn). Hits are parsed to retrieve and orient candidate SSU sequences, which are then screened for chimeric sequences using VSEARCH v2.30.5, aligned to one of nine curated reference alignments with MAFFT v7.526, and phylogenetically placed onto reference trees using pplacer v1.1.alpha20. These reference alignments and trees were constructed from the 18S dataset of the Protist Ribosomal Reference (PR2) database and represent the nine major eukaryotic supergroups (Amoebozoa, Excavata, TSAR, Archaeplastida, Cryptista, Haptista, CRuMs, Provora, and Obazoa). Placement results are processed with guppy and rppr and summarized to generate taxonomic assignments. In parallel, bacterial contamination is assessed using BUSCO. Final outputs integrate phylogenetic placement, contamination screening, and summary statistics

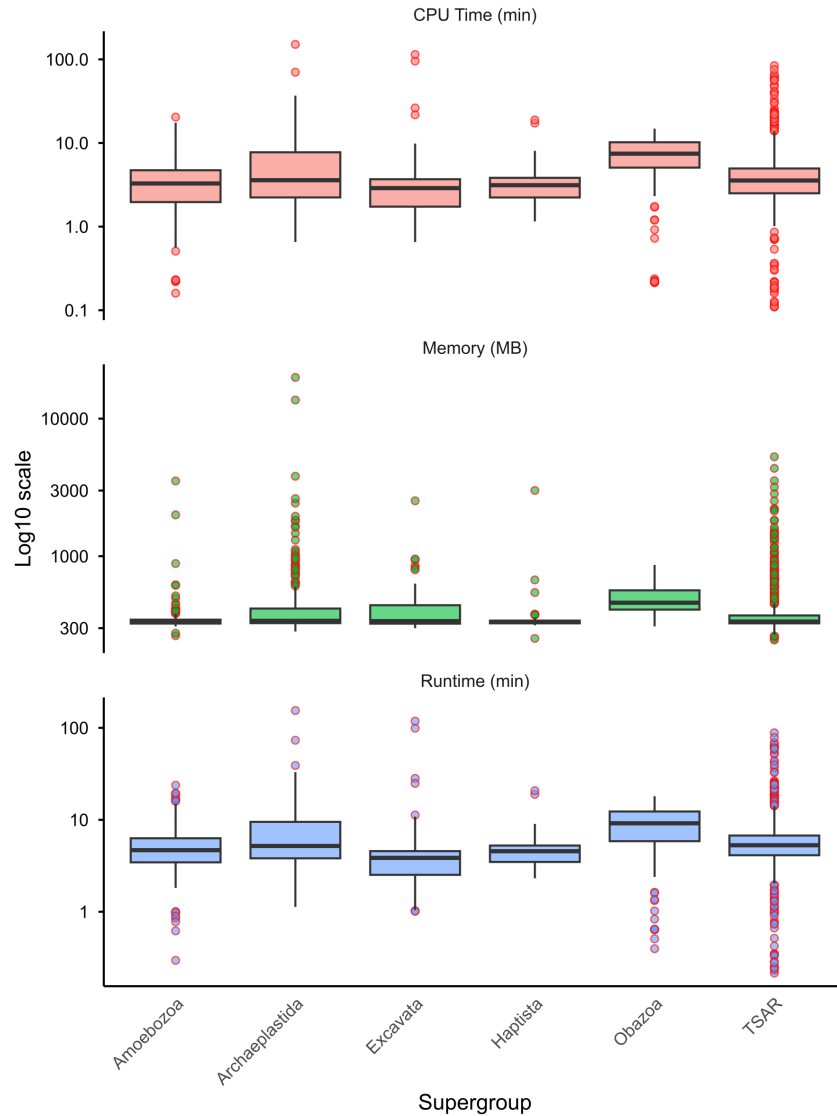


Fig. S2. Benchmarking performance of CSI-SSU across major eukaryotic supergroups. Boxplots summarize CPU time in minutes (top), memory usage in megabyte (middle), and total runtime in minutes (bottom) for analyses of 2,960 genomic-level assemblies from the P10K dataset. Values are shown on a $\log_{10}$ scale. Boxes represent the interquartile range (IQR) with median values indicated by horizontal lines; whiskers extend to $1.5 \times$ IQR. Individual points denote outliers. Across supergroups (Amoebozoa, Archaeplastida, Excavata, Haptista, Obazoa, and TSAR), most runs show low and consistent computational requirements, while a subset of analyses exhibit higher resource usage. The analyses were performed on the Quanah partition of the Texas Tech University High Performance Computing Center (HPCC) RedRaider cluster. The Quanah partition is built on Intel Xeon E5-2695 v4 processor and runs Rocky Linux 9.2/8.10, with 36 cores and 192 GB of memory available per node. All analyses were conducted using a single thread on one node of the Quanah partition.

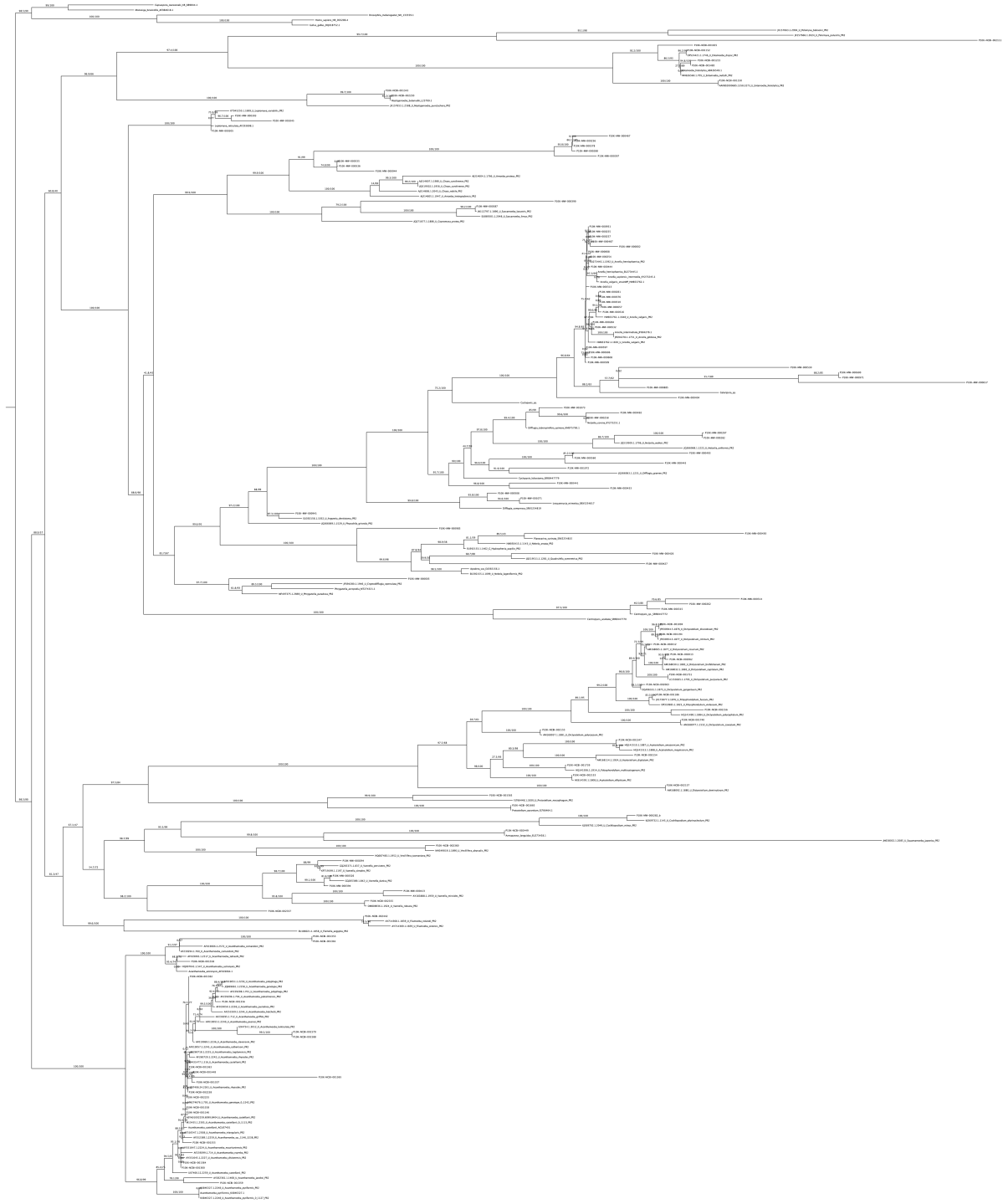


Figure S3. The maximum-likelihood phylogenetic tree of Small Subunit ribosomal RNA (SSU) inferred from a curated dataset generated in the present study, comprising SSU sequences retrieved from genomes and transcriptomes available in the P10K database, along with reference sequences from the PR² database (indicated by PR² IDs) and NCBI. The dataset includes representatives from the three major Amoebozoa groups: Tubulinea, Evosea, and Discosea. Subsets of this dataset were used to generate the trees

presented in Figures 2 and 3A, and Figures S4–S7. 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 $\geq 80/95$ considered indicative of strong support.

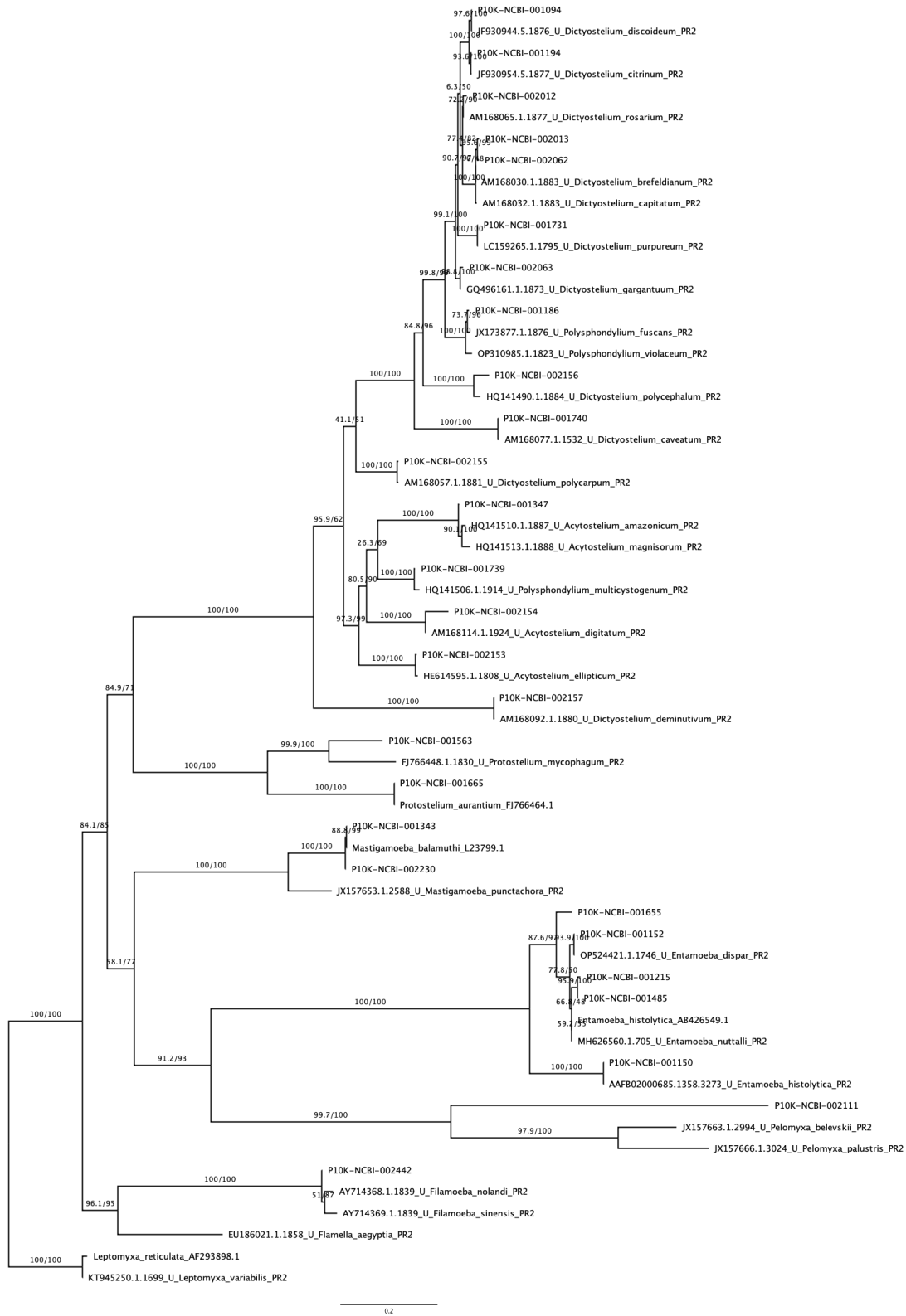


Figure S4. The maximum-likelihood phylogenetic tree of Small Subunit ribosomal RNA (SSU) inferred from a subset of the curated dataset presented in Figure S3, focuses on the major Amoebozoa group

Evosea and is shown as Figure 2A in the main text of the manuscript. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (TIM2+F+R4). 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 $\geq 80/95$ considered indicative of strong support.

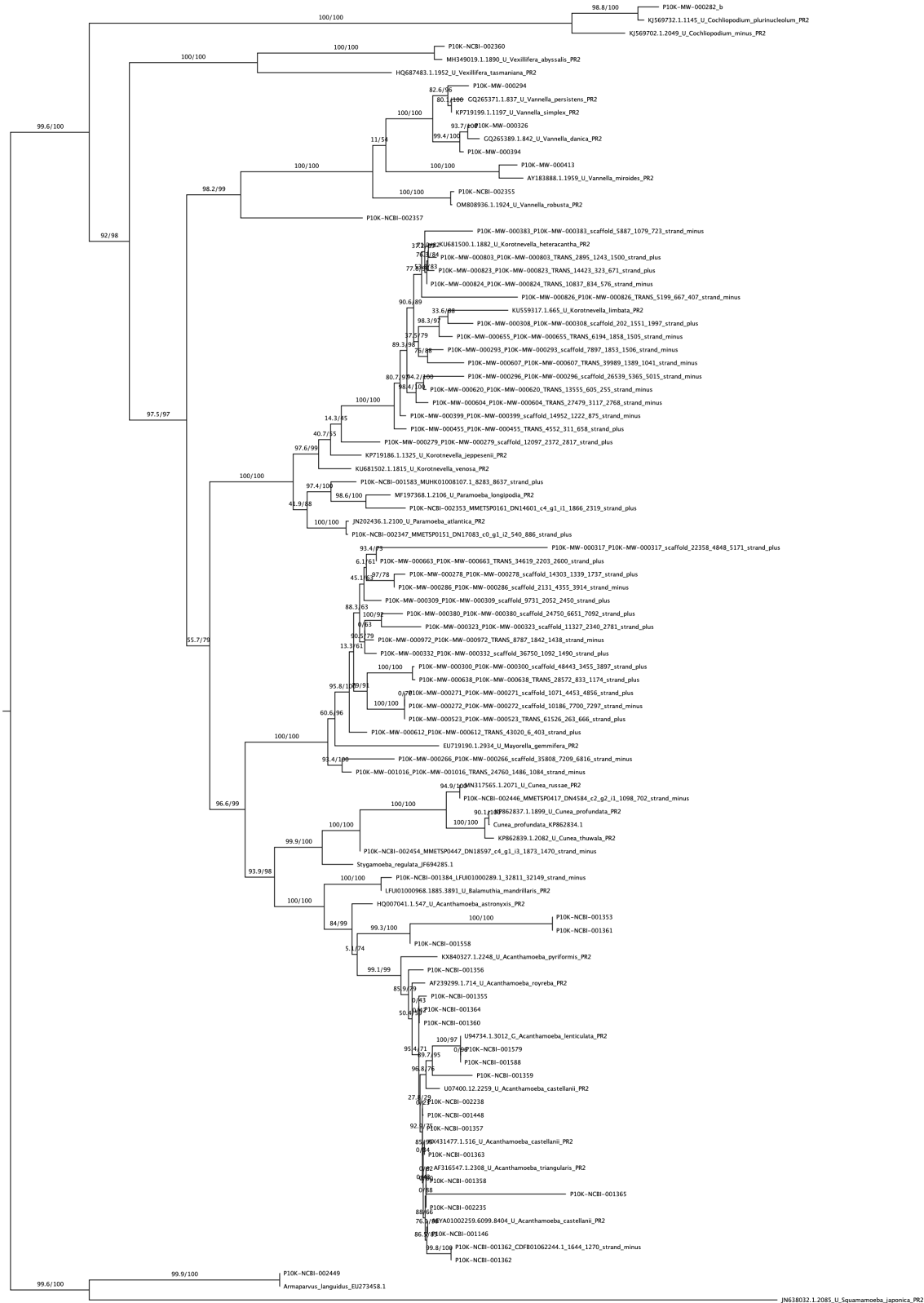


Figure S5. The maximum-likelihood phylogenetic tree of Small Subunit ribosomal RNA (SSU) inferred from a subset of the curated dataset presented in Figure S3, focuses on the major Amoebozoa group Discosea and includes Cutosea, a long-branch lineage that consistently falls outside Evosea in SSU rRNA phylogenies (Schuler and Brown 2019), and is shown as Figure 2B in the main text of the manuscript. 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 $\geq 80/95$ considered indicative of strong support.

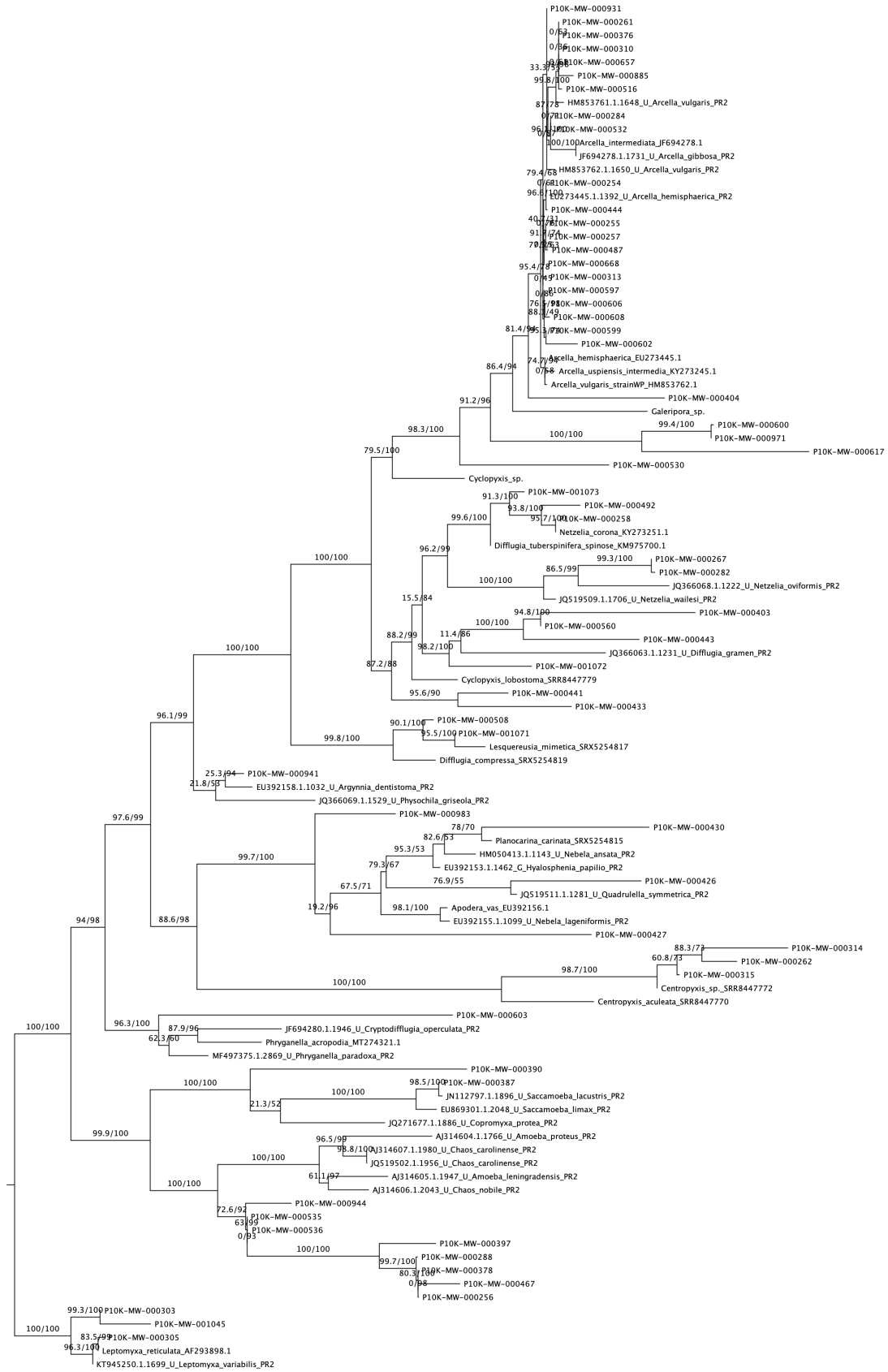


Figure S6. The maximum-likelihood phylogenetic tree of Small Subunit ribosomal RNA (SSU) inferred from a subset of the curated dataset presented in Figure S3, focuses on the major Amoebozoa group Tubulinea and is shown as Figure 3A in the main text of the manuscript. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (TIM3e+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 $\geq 80/95$ considered indicative of strong support.

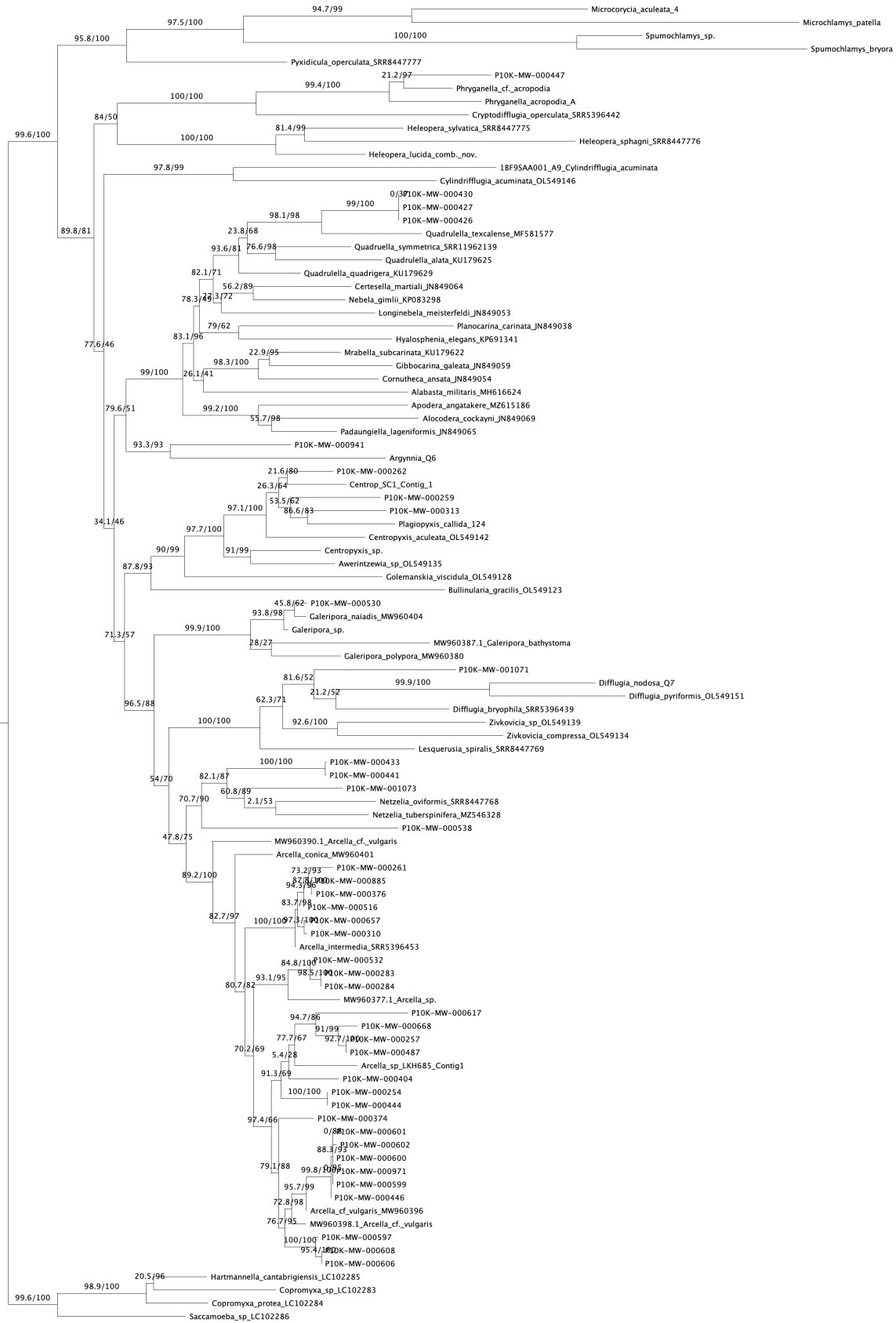


Figure S7. The maximum-likelihood phylogenetic tree of cytochrome c oxidase subunit I (COI) inferred from a curated dataset generated in the present study, focusing on Arcellinida order (Tubulinea:Amoebozoa) comprising COI sequences retrieved from genomes and transcriptomes available in the P10K database, along with reference sequences made available by previous studies and is shown as Figure 3B in the main text of the manuscript. 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 $\geq 80/95$ considered indicative of strong support.

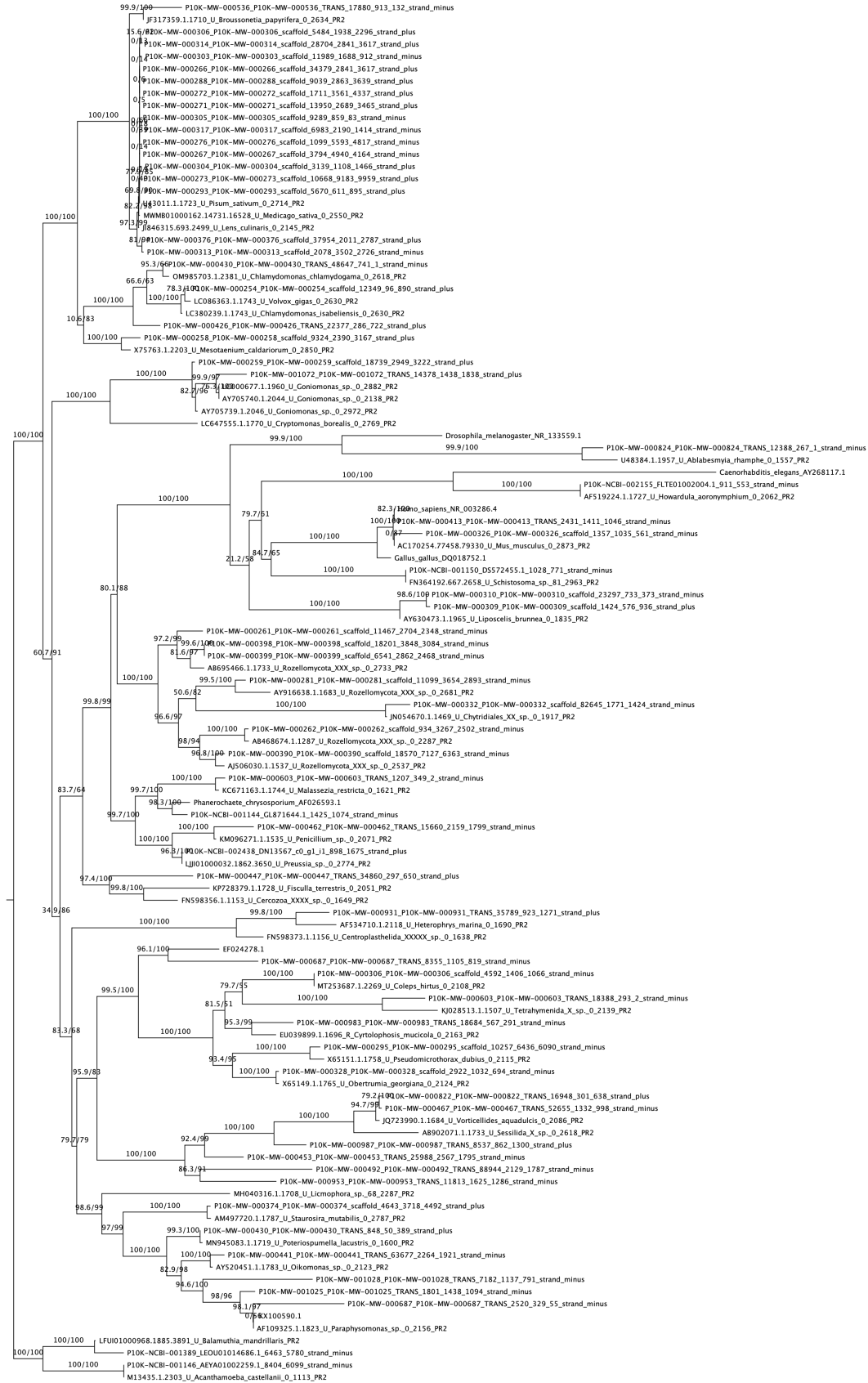


Figure S8. The maximum-likelihood phylogenetic tree of Small Subunit ribosomal RNA (SSU), inferred from a curated dataset generated in the present study, comprises SSU sequences of putative non-Amoebozoan contaminants retrieved from genomes and transcriptomes available in the P10K database, along with reference sequences from the PR² database (indicated by PR² IDs), identified through BLAST+ similarity searches using the putative contaminant SSU sequences as queries. This tree is presented as Figure 4 in the main text of the manuscript. Phylogenetic reconstruction was conducted using IQ-TREE v2.3.6, with ModelFinder identifying the best-fit substitution model (TN+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.

Supplementary dataset legends

Data S1. Representative output from running the CSI-SSU tool on the 2,960 assemblies from the P10K dataset, encompassing all supergroups considered in this study. The data include the parsed_blast output directory, containing parsed and filtered SSU sequences, the vsearch directory, containing summaries of chimeric search results, and the summary directory, containing final summary reports and phylogenetic trees with sequence placements (PDF format).

Data S2. The master SSU rDNA/rRNA dataset constructed in this study (see *Methods* section from the main text). This dataset, prior to curation (e.g., removal of duplicate and short sequences based on a preliminary phylogenetic reconstruction), includes all SSU sequences identified and retrieved from Amoebozoan genomic data available in the P10K database (indicated by P10K sample IDs), along with SSU rDNA/rRNA sequences from previous studies and the PR² database (indicated by PR² IDs) (Kang et al. 2017; Ribeiro et al. 2023; Porfirio-Sousa et al. 2024). It also incorporates reference sequences of non-amoebozoan taxa from the PR² database used for contamination screening of P10K genomic data.

Data S3. The curated SSU rDNA/rRNA master Amoebozoa dataset focusing on the major amoebozoan groups Tubulinea, Evosea, and Discosea. This is the dataset used to construct the phylogenetic tree presented in the **Figure S3** and the datasets **Data S4 – S6**.

Data S4. The curated subset of the SSU rDNA/rRNA master dataset focusing on the major Amoebozoa group Evosea. This is the dataset used to construct the phylogenetic tree presented in **Figure 2A** and **Figure S4**.

Data S5. The curated subset of the SSU rDNA/rRNA master dataset focusing on the major Amoebozoa group Discosea. This is the dataset used to construct the phylogenetic tree presented in **Figure 2B** and **Figure S5**.

Data S6. The curated subset of the SSU rDNA/rRNA master dataset focusing on the major Amoebozoa group Tubulinea. This is the dataset used to construct the phylogenetic tree presented in **Figure 3A** and **Figure S6**.

Data S7. The curated cytochrome c oxidase subunit I (COI) dataset focusing on the order Arcellinida (Tubulinea:Amoebozoa). This is the dataset used to construct the phylogenetic tree presented in **Figure 3B** and **Figure S7**.

Data S8. The curated subset of the SSU rDNA/rRNA master dataset focusing on the putative non-amoebozoan contaminants. This is the contamination screening dataset used to construct the phylogenetic tree presented in **Figure 4** and **Figure S8**.

References

- Gao Xinxin, Chen K., Xiong J., Zou D., Yang F., Ma Y., Jiang C., Gao Xiaoxuan, Wang G., Gu S., Zhang P., Luo S., Huang K., Bao Y., Zhang Z., Ma L. & Miao W. 2024. The P10K database: a data portal for the protist 10 000 genomes project. *Nucleic Acids Research*, **52**:D747–D755.
- Kang S., Tice A. K., Spiegel F. W., Silberman J. D., Pánek T., Čepička I., Kostka M., Kosakyan A., Alcântara D. M. C., Roger A. J., Shadwick L. L., Smirnov A., Kudryavtsev A., Lahr D. J. G. & Brown M. W. 2017. Between a Pod and a Hard Test: The Deep Evolution of Amoebae. *Molecular Biology and Evolution*, **34**:2258–2270.
- Porfirio-Sousa A. L., Tice A. K., Morais L., Ribeiro G. M., Blandenier Q., Dumack K., Eglit Y., Fry N. W., Gomes E Souza M. B., Henderson T. C., Kleitz-Singleton F., Singer D., Brown M. W. & Lahr D. J. G. 2024. Amoebozoan testate amoebae illuminate the diversity of heterotrophs and the complexity of ecosystems throughout geological time. *Proceedings of the National Academy of Sciences*, **121**:e2319628121.
- Ribeiro G. M., Useros F., Dumack K., González-Miguéns R., Siemensma F., Porfirio-Sousa A. L., Soler-Zamora C., Pedro Barbosa Alcino J., Lahr D. J. G. & Lara E. 2023. Expansion of the cytochrome C oxidase subunit I database and description of four new lobose testate amoebae species (Amoebozoa; Arcellinida). *European Journal of Protistology*, **91**:126013.