Global Metagenomics Reveals Hidden Protist Diversity

Miguel F. Romero[1]*, Arianna I. Krinos[1,2], Xyrus Maurer-Alcala[3], John A. Burns[4], Ramunas Stepanauskas[4], Tanja Woyke[1,5], Frederik Schulz[1]*

[1]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[2]Department of Earth, Environmental, and Planetary Sciences, Brown University, Providence, Rhode Island, USA
[3]American Museum of Natural History, New York, New York, USA
[4]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine, USA
[5]University of California Merced, Life and Environmental Sciences, Merced, CA 95343, USA

*Correspondence:
mfromerogutierrez@lbl.gov
fschulz@lbl.gov

Abstract

Protists, defined as unicellular eukaryotes distinct from animals, plants, and fungi, are a polyphyletic group that predominates the eukaryotic tree of life, exhibiting significant phylogenetic diversity and fulfilling critical ecological roles. Historically, research has prioritized protists associated with animals and plants, particularly those of medical significance, thereby overlooking the majority of protist diversity. Conventional molecular methods, such as 18S rRNA gene amplicon sequencing, frequently encounter limitations, including primer binding bias and PCR bias caused by gene length variations, creating a biased perspective of understudied protistan diversity. Further, most protist lineages are notoriously difficult to cultivate. Here, we apply a cultivation-independent approach in which we analyzed over 30,000 assembled metagenomes and protist single cell genomes and 21 long 18S rRNA gene amplicon data sets from various global ecosystems, including marine, freshwater, and soil environments. We recovered 157,956 18S rRNA gene sequences (≥800 bp), which clustered into 103,338 operational taxonomic units (OTUs) at 97% sequence identity and 24,438 OTUs at 85% identity. Notably, 69% of 13,238 non-singleton clusters at 85% identity consisted exclusively of environmental sequences, uncovering a wealth of yet uncultivated and unclassified novel protist diversity. A comprehensive taxonomic framework of eukaryotes based on concatenated 18S and 28S rRNA genes that incorporated most novel lineages revealed substantial underrepresentation of Amoebozoa, Discoba, and Rhizaria in reference databases, with many lacking isolate or genome sequence representation. Further, we identified 13 eukaryotic lineages with novelty on higher taxonomic ranks, such as class and phylum-level, that lack representation in public databases. The corresponding 85% OTUs were primarily affiliated with Excavata, with some branching closely to the root of the eukaryotic tree.

Comprehensive analysis of the global distribution of eukaryotes revealed uneven microbial eukaryotic diversity across supergroups and ecosystems, with notable novelty particularly in soil and marine environments. We then examined co-occurrence between protists and prokaryotes, predicting putative symbiotic or predator-prey relationships, particularly among understudied protist groups with bacteria such as Verrucomicrobia and Rickettsiales. Our results substantially enhance the understanding of protistan diversity and distribution, revealing taxonomic blind spots and laying groundwork for future studies of these organisms' ecological roles.

Introduction

The eukaryotic tree of life is predominantly composed of non-fungal unicellular organisms (protists). In addition to their remarkable diversity, protists exhibit a wide array of cellular structures and lifestyles (Foissner 2006; Adl et al. 2007). Previous research has largely focused on protists associated with animals and plants, particularly those of medical relevance (Collins and Jeffery 2007; Horn 2022; Einarsson et al. 2016). Well-characterized lineages include the TSAR group, which comprises Stramenopiles (e.g., diatoms and oomycetes), Alveolata (e.g., ciliates, dinoflagellates, and apicomplexan parasites), Rhizaria (e.g., foraminifera and radiolarians), as well as the Amoebozoa supergroup (sensu stricto amoebae) and the Discoba division (which includes heteroloboseans and euglenoids, including kinetoplastids). Even less well studied eukaryotic lineages include Metamonada, which encompass instances of symbiotic or parasitic groups like diplomonads and parabasalids (Céza et al. 2022). Recently, advances in molecular methods, such as cultivation-independent single-cell sequencing (Woyke et al. 2017), have facilitated the exploration of phylogenetic associations among selected uncultivated groups, resulting in the identification of new eukaryotic supergroups, such as Hemimastigophora (Foissner et al. 1988; Lax et al. 2018) and Provora (Tikhonenkov et al. 2022), with additional members continually being added as interest in these novel lineages expands (Eglit et al. 2024).

Molecular methods such as 18S rRNA gene amplicon sequencing have proven invaluable for assessing eukaryotic diversity in various environmental contexts. In oceanic plankton communities, studies revealed that heterotrophic protists represent the most diverse group, despite metazoans dominating in terms of abundance (de Vargas et al. 2015). A protist-centric survey in select soil environments indicated that Cercozoa are the most prevalent group, revealing correlations between protistan and bacterial taxa, including evidence of potential predator-prey interactions (Oliverio et al. 2020). At a global level, soil eukaryotic diversity has primarily been evaluated with respect to specific groups such as Fungi and testate amoebae (Fernández et al. 2016; Bahram et al. 2018). Diversity of eukaryotes in soil typically exhibits higher values in the Simpson diversity index in equatorial regions and intermediate latitudes, indicating that diversity peaks in temperate climates, likely linked to water-energy balance and present similar patterns such as plants and animals (Fernández et al. 2016; Bahram et al. 2018). A comprehensive meta-analysis that incorporated data from marine, freshwater, and topsoil ecosystems demonstrated that topsoil harbors greater protistan diversity compared to aquatic environments (Singer et al. 2021). Despite these extensive studies on eukaryotic

diversity and community structure, recent findings indicate that an overwhelming majority of protistan taxa remain undocumented. For example, a recent analysis suggested that one-third of operational taxonomic groups (OTUs) could not be assigned to known eukaryotic groups (de Vargas et al. 2015). Similarly, new lineages of protists, at both the phylum and supergroup levels, including Hemimastigophora and Provora, have only recently been discovered (Foissner et al. 1988; Lax et al. 2018; Tikhonenkov et al. 2022).

The 18S rRNA gene (also referred to as 18S rDNA) has emerged as the most frequently used marker gene for identifying eukaryotic organisms in environmental studies. This gene encodes the small subunit of the ribosomal RNA and is universally present in all nuclear genomes, comprising regions of both high conservation and high variability (Wuyts et al. 2001). Public databases, such as Silva (Quast et al. 2013) and the Protist Ribosomal Reference (PR2; Guillou et al. 2013), have served as repositories for these sequences for decades. Polymerase chain reaction (PCR)-based approaches (metabarcoding) are typically used to recover this marker gene with high sensitivity in environmental samples. However, due to the constraints of short-read DNA sequencing technologies, only a limited portion of the 18S rDNA is typically screened, predominantly the V4 to V6 and V9 regions (Hugerth et al. 2014; de Vargas et al. 2015; Oliverio et al. 2020). Unlike its prokaryotic counterpart (the 16S rRNA gene), the 18S rRNA gene exhibits extreme length variation, ranging from a minimum of 1.5 kb (in Diplomonadida) to a maximum of 6.4 kb (in Amoebozoa), with the mature 18S rRNA molecule potentially reaching lengths of up to 2.3 kb (Xie et al. 2011). This variability can be attributed to length-variable regions that frequently contain introns, some exceeding 1 kb, especially in regions commonly amplified (Xie et al. 2011). Such disparities in length, coupled with the rapid evolution of the 18S rRNA gene in certain lineages (Meyer et al. 2010), may bias PCR-based metabarcoding studies, as only a fraction of the eukaryotic diversity in the environment can be captured. Recently, long amplicon sequencing has been deployed to recover multiple markers from the rRNA gene locus within eukaryotic genomes (Jamy et al. 2022). Although this technique solves some of the challenges associated with conventional methods (Overgaard et al. 2024), primer binding bias remains a concern due to sequence divergence.

An alternative approach that is able to further mitigate limitations of amplicon-based methods is shotgun metagenomics as it does not employ specific primers for amplifying target sequences. Instead, it involves screening assembled sequences from the genomes of all taxa present in the sample for rRNA genes. Depending on sequencing depth, genomic information from rarer taxa can also be retrieved. In our study, we employed a shotgun metagenomics approach, leveraging over 30,000 assembled metagenomes to conduct a global census of eukaryotic diversity. Our primary objective was to elucidate the phylogenetic and ecosystem diversity of microbial eukaryotes on a global scale, thereby advancing the understanding of protistan diversity and distribution, including the characterization of previously unidentified lineages.

Results and discussion

Expansion of Known Eukaryotic Diversity

The screening of 32,520 assembled metagenomes resulted in the recovery of 157,956 18S rRNA genes from 18,444 sequencing datasets, all with a minimum length of 800 bp (a summary of the methods can be found in Supplementary Figure 1). Subsequently, we performed two-step clustering of these metagenome-derived sequences in conjunction with over 250,000 sequences from the curated rRNA gene databases Silva and PR2 (Quast et al. 2013; Guillou et al. 2013) and additional recent studies (Jamy et al. 2022; Tikhonenkov et al. 2022), yielding a total of 103,338 sequence clusters (OTUs) at 97% sequence identity and an additional 24,438 OTUs at 85% identity. We discarded all 11,200 85% identity clusters containing only a single sequence, retaining 13,238 clusters with two or more sequences for further analysis. As reference databases such as Silva and PR2 include sequences originating from both cultured isolates as well as environmental sources, we were able to estimate the proportion of sequences corresponding to uncultivated microbes within our dataset. Our analysis revealed that 69% of the 85% identity clusters (n=9,146) were exclusively comprised of environmental sequences, while 30% (n=4,006) included sequences from isolates and less than 1% (n=86) contained sequences derived from genomes. Clusters that contained sequences from isolates (18S rRNA genes) were generally larger, whereas the metagenomic-only clusters tended to be smaller (Figure 1A). Clusters containing genome sequences had an average of 1,334 members, greatly exceeding the average of 49 members for clusters containing sequences from isolates. Notably, the mean size of clusters containing solely environmental sequences was calculated at 10. Among the largest clusters containing only sequences from reference databases were sequences from Aconoidasida (Apicomplexa, size=1,220) and Conoidasida (Apicomplexa, size=910). Conversely, the largest clusters including metagenomic sequences comprised sequences from Malasseziales (Fungi, size=18,608) and Colpodellida (Chrompodellids, size=10,291). The largest cluster incorporating genome sequences was also affiliated with Malasseziales, while the largest cluster containing isolate-derived sequences without genome information was ascribed to Hexapoda (Metazoa, size=10,510). Notably, the largest cluster exclusively composed of environmental sequences was designated to Syndiniales Group II (Syndiniales, size=1,028). To further investigate the clusters with only reference sequences, we estimated the number of distinct taxa found per cluster. On average, clusters contained 2.4 genera and 1.7 families , suggesting that a clustering threshold of 85% sequence identity roughly corresponds to the order or family level. (Supplementary Figure 2)

Our results indicate that protists recognized as parasites of animals and humans,such as Piroplasmorida (Apicomplexa, 1,220 sequences) and Cryptosporida (Apicomplexa, 910 sequences), are the predominant representatives within reference databases, revealing a pronounced sampling bias in the representation of microbial eukaryotes (del Campo et al. 2014). Some parasitic protists, such as apicomplexans, were also found to be dominant in soil microbiomes (Mahé et al. 2017). Consequently, it is unsurprising to encounter sizable clusters

of sequences derived from parasitic protists, whether from isolates or environmental samples, with the latter exhibiting a more balanced representation of the diversity of these organisms.

To enhance the taxonomic annotation of sequences and to better reflect the phylogenetic relationships among eukaryotes, we constructed a eukaryotic reference tree utilizing 1, 438 85% OTU cluster representatives that also presented matching 28S rRNA genes. This subset of our dataset excluded most of the highly populated clusters (Figure 1B). The resulting tree broadly aligned with the established phylogeny of eukaryotes, as it effectively grouped the Amorphea together, distinctly separate from the Diaphoretickes, in agreement with observations from both rRNA and protein-based alignments (Jamy et al. 2022; Torruella et al. 2024). Further, Metamonads were positioned near the root, consistent with phylogenomic analysis of the root of the eukaryotic tree (Al Jewari and Baldauf 2023; Williamson et al. 2024) . Due to high mutation rates and pronounced length variability, certain groups such as Microsporidia (Vossbrinck et al. 1987) were excluded from the reference tree, leaving their representatives unclassified within our phylogenetic taxonomic annotation. Numerous clades grouped within Metamonada, Stramenopiles, Alveolata, Amoebozoa, Metazoa, and Fungi divisions entirely lacked sequences derived from isolates or genomic sources (Figure 1C). Clusters associated with these clades also tended to exhibit representative sequences with low blast alignment identity to reference sequences and increased average predicted primer binding bias (Figure 1C) compared to sequences from other clusters, suggesting that these represent so-called taxonomic blindspots, that have been overlooked in prior PCR amplification-based studies (Morey et al. 2024).

Our census of eukaryotic diversity revealed an uneven taxon richness across various supergroups, with Obazoa (n=3,090), Amoebozoa (n=2,998), and TSAR (n=2,081) exhibiting the highest number of OTUs. At the division level, Opisthokonta harbored the largest diversity, predominantly characterized by sequences derived from isolates. A total of 36 clusters incorporated sequences from genomes, while 2,039 sequences were drawn from isolates, and 1,015 clusters were comprised solely of environmental sequences (Figure 1D). The divisions classified under Amoebozoa exhibited a high proportion of clusters exclusive to environmental samples (95%), most of which remained unclassified at the division level (n=1,400). Similarly, Discoba, the second most diverse division, demonstrated novelty in sequences presented as a significantly larger proportion of clusters derived exclusively from metagenomic sequences (n=1,312), with only 202 clusters containing sequences from isolates and a mere 6 clusters encompassing genome-derived sequences. In the context of the TSAR supergroup, Rhizaria emerged as the most diverse division, with 73.5% of the 937 clusters comprising entirely environmental sequences. The predominance of Obazoa reflects its evolutionary success due to complex multicellularity and diverse nutritional strategies (Torruella et al. 2015), while Amoebozoa's high diversity demonstrates its adaptive radiation across soil and aquatic habitats through varied locomotion mechanisms and predatory strategies (Tekle et al. 2022), which could be true also for some groups of discobans. Similarly, Rhizaria's dominance within TSAR results from their successful rapid diversification (Burki and Pawlowski 2006), as demonstrated in marine systems, underscoring how current cultivation approaches fail to capture the true breadth of eukaryotic diversity across these major lineages.

Importantly, 2,274 (78%) of the clusters that could not be assigned to any eukaryotic supergroup were exclusively derived from metagenomic sequences, while 621 clusters contained sequences from isolates and 13 clusters harbored sequences from genomes. Among the reference-only clusters that were unassigned to any supergroup using the tree-placement method, most were represented in the reference tree according to sequence annotation, such as Cnidaria (129 clusters), or had a sister group present within the same supergroup or division, exemplified by Microsporidia (36 clusters). This finding is somewhat expected for lineages such as Microsporidia, which are characterized by short yet highly divergent 18S rRNA gene sequences (Vossbrinck et al. 1987). Challenges in supergroup assignment using tree-placement methods often stem from incomplete reference databases or rapid evolutionary rates. Despite sequence annotation indicating taxonomic placement, many organisms remain unclassified at higher taxonomic levels while having clear affiliations at lower ranks. This phenomenon reflects the evolutionary complexity of certain lineages that escape traditional classification approaches, as has been previously observed (Berger et al. 2011).

We found differences in insertion number and length among eukaryotic supergroups, with Centroplasthelida with the highest mean number of insertions (6.54), and Chlorophyta with the longest mean insertion size (665.34; Figure 1E). Even with variation within lineages of eukaryotic supergroups (Xie et al. 2011), we found differences between groups. Predicted binding bias scores also varied among groups, with those with the highest proportion of phylogenetic novelty (Figure 1D) showing also a high mean binding bias score, such as Metamonda (mean score=3.83), Evosea (3.58), Discoba (3.02), and Eukaryota.U.division (unclassified at the supergroup level, mean score=2.93). This observation could explain why some lineages within these groups might have remained as taxonomic blindspots in metabarcoding studies.

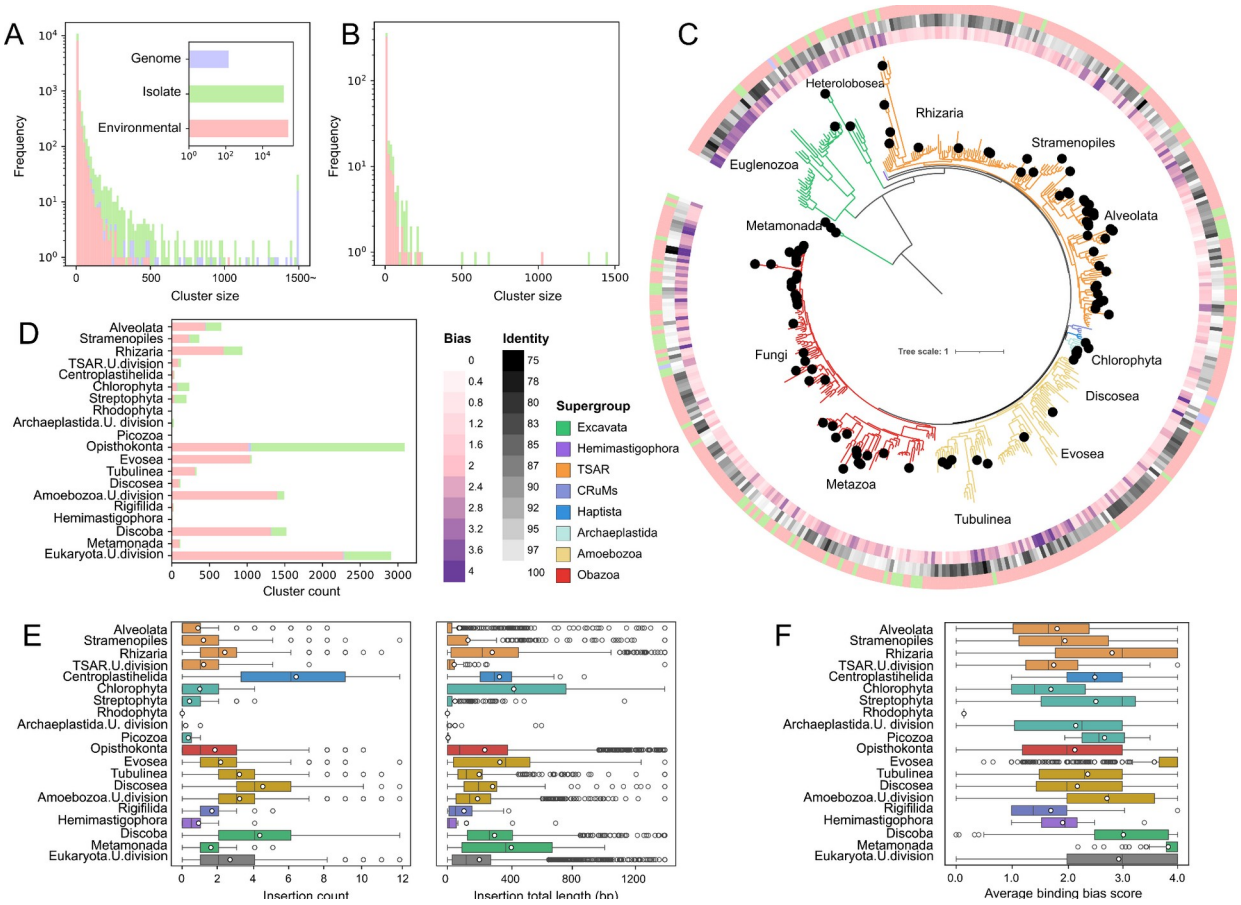


Figure 1. Phylogenetic Diversity of Eukaryotes. (A) Histogram illustrating the distribution of cluster sizes, color-coded based on the origin of the included sequences. The inset represents the number of clustered sequences from each source. Data corresponds to 85% identity clusters after discarding singletons. (B) Cluster size distribution of the OTUs that contained both 18S and 28S rRNA genes, which were employed to construct the eukaryotic reference tree. (C) Reference tree inferred from a concatenated alignment of the 18S and 28S rRNA genes derived from OTUs that had representative members with both genes present. The heat map on the innermost track indicates the cluster-wide average of the accumulated primer binding bias score of the corresponding 18S rRNA gene sequences utilizing a set of four commonly used general eukaryotic primer pairs. A score of 4 (dark purple) suggests that the sequence would not bind to any of the primer pairs, while a score of 0 (white) implies it would bind to all of them. The adjacent heatmap delineates sequence identity compared to the best hit within our combined 18S rRNA gene sequence database. The outermost track denotes the cluster composition corresponding to the legend illustrated in panel A. Black circles at the tip of the leaves indicate that the respective clusters are composed exclusively by environmental sequences extracted in this study. (D) Relative abundance of clusters for each eukaryotic division within the analyzed environmental samples. (E) Distribution of insertion count and cumulative insertion length among eukaryotic divisions. Each point shows the corresponding data for each OTU representative sequence. Insertions are defined as sequences present in each 18S rRNA gene sequence not represented in the reference covariance model. Data of insertions with length >= 10 bp is shown. White circles indicate mean values. (F) Distribution of predicted primer binding bias score for each OTU representative sequence ordered by its taxonomic classification at the division level. Mean values are shown as white circles.

The Global Distribution Of Microbial Eukaryotes

The majority of 18S rRNA gene sequences recovered (present in non-singleton clusters) were from the Integrated Microbial Genome/Metagenome (IMG/M) database (Chen et al. 2023), representing samples from six continents, encompassing longitudes and latitudes from pole to pole (Figure 2A). The main ecosystem type sampled was soil (n=5,487), followed by marine (n=4,390) and freshwater (n=3,235). Among the frequently observed subtypes, lake samples (n=1,588), peat (n=985), and oceanic samples (n=974) were predominant (Supplementary Table 1). In contrast, datasets without any identified 18S rRNA gene sequences were most frequently associated with subtypes such as phylloplane/leaf (n=408), coalbed methane well (n=319), and insect nests (n=248) (Supplementary Table 2), the presence of 16S rRNA genes in the same samples indicates that these may represent a sample preparation bias or a harsher environments for growth of microbial eukaryotes. Despite the broad spectrum from which these samples were gathered, the majority originated from latitudes ranging from 60° to 30° North, particularly within North America and Central Europe (Figure 2B). Key factors influencing protist diversity include water availability, pH, and nutrient availability (Zhao et al. 2023). The captured protistan diversity in freshwater and peat bogs is in agreement with what has been shown in previous studies (Lara et al. 2011; Krinos et al. 2024; Oliverio et al. 2020). In contrast, the absence of 18S rRNA gene sequences in habitats like coalbeds, that are characterized by low water availability and low nutrient concentrations, reflect challenging growth conditions for eukaryotes.

Next, we investigated the overall distribution of predominant and ubiquitous microbial eukaryotic orders and families. Relative abundance was defined as the average of the normalized mean sequencing depth across all contigs pertaining to each taxonomic group, while ubiquity was defined as the proportion of samples from comparable environments and latitudes in which each taxonomic group was identified. Within soil ecosystems, well-studied groups such as Exobasidiomycetes (Fungi; relative abundance = 2.73) and Insecta (Metazoa; relative abundance = 5.78) were notably abundant, whereas the crustaceans Maxillopoda (Metazoa) constituted the most abundant group within marine environments (relative abundance = 15.62). Excluding Metazoa, Fungi, and Streptophyta (plants), the most abundant group detected in marine ecosystems consisted of green algae from the family Bathycoccaceae (relative abundance = 8.65; Figure 2C), followed by diatoms (family Thalassiosiraceae; relative abundance = 8.06) and brown algae (family Phaeophyceae_XX; relative abundance = 6.37). Additionally, parasitic dinoflagellates from Digo-Group-I-Clade-1 and Dino-Group-II-Clade-7 were frequently observed with relatively high abundances (relative abundance = 3.55 and 2.76, respectively). In freshwater systems, the most abundant groups included Thalassiosiraceae (Chlorophyta; relative abundance = 11.81) and green algae (Chlamydomonadales_X; relative abundance = 5.61), with ciliates from the Oxytrichidae family being present across the largest number of freshwater samples (ubiquity = 0.22). Bathycoccaceae (Chlorophyta) emerged as the most abundant group within soil ecosystems (relative abundance = 2.85), closely followed by flagellated bacteriotrophic protists from the family Sandonidae (Cercozoa; relative abundance = 1.11). Historically, diatoms and metazoans have been classified as dominant groups within marine environments, alongside chlorophytes and parasitic dinoflagellates (de Vargas et al. 2015). Similarly, prior studies have indicated that

lineages of Cercozoa hold dominant positions within soil ecosystems, along with other groups of chlorophytes (Oliverio et al. 2020). Our findings revealed Bathycoccaceae as highly abundant protists in soil, likely due to the inclusion of diverse soil ecosystem subtypes, such as peatlands, glaciers, and wetlands, where Bathycoccaceae were most frequently observed (Supplementary Table 3). This contrasts to topsoils, which contain algae from various groups, such as Chlorophyta and Bacillariophyceae (Zancan et al. 2006). The observed differences in abundance between our findings and previous amplicon-based studies (Mahé et al. 2017; Oliverio et al. 2020) may stem from the heterogeneity of samples, enhanced gene recovery of high-abundance eukaryotes in shotgun metagenomes, and the absence of amplification and primer-binding bias. Taken together, our results provide a more unbiased perspective of abundances of microeukaryotes in global ecosystems.

Our eukaryotic census also provided a more detailed perspective on the distribution of eukaryotic groups across different latitudes (Figure 2D). Even at elevated taxonomic ranks, distinct patterns emerged across latitudes and ecosystem types. The marine environment exhibited both high ubiquity and abundance of the Stramenopiles, Alveolata, and Rhizaria (SAR) group. Notably, at the arctic tundra around 90°, representatives from all three divisions were consistently present (ubiquity = 1) and exhibited relative abundances of 5.8, 4.7, and 2 for Alveolata, Stramenopiles, and Rhizaria, respectively. Opisthokonts were also found to be ubiquitous (ubiquity = 1) and abundant (relative abundance = 7.6) in close proximity to the North Pole. Prior studies using documentary sources, such as cell counts, showed the presence of alveolates, stramenopiles, and unicellular opisthokonts within the pan-Arctic region; however, rhizarians were not observed in this context (Poulin et al. 2011). In freshwater ecosystems, Opisthokonts, particularly Metazoa, displayed the highest mean ubiquity (0.3) alongside high relative abundance at -10° and 40° (relative abundance = 17.2 and 33.2, respectively), while groups such as Discoba were also represented with mean ubiquity = 0.22 and high abundance at latitudes of 30° and 50° (relative abundance = 60.2 and 16.7, respectively), with Discoba_X (primarily composed of heteroloboseans) exhibiting a mean abundance of 5.5. Sequences lacking supergroup-level classification were most ubiquitous in freshwater (mean ubiquity = 0.3), though their abundance exhibited variability based on latitude, peaking at 64.6 at 50°. Within soil ecosystems, opisthokonts and streptophytes were ubiquitous across all latitudes (mean ubiquity = 0.5 and 0.3, respectively), whereas other groups were less frequently observed. Nevertheless, Cercozoa persisted as the most predominant eukaryotic groups in soil, with classes such as Cercozoa.U. leading with a mean relative abundance of 21.38, while Glissomonadida and Thaumatomonadida displayed lower mean abundances of 1.1 and 0.2, respectively. This is in line with previous reports that showed that Cercozoans comprise a substantial proportion of microbial biomass in soil when evaluated using metatranscriptome sequencing data (Urich et al. 2008), often represented by highly divergent 18S rRNA gene sequences even among closely related lineages, as observed from isolates (Howe et al. 2011). Our results confirm that cercozoans represent a dominant eukaryotic group in soil with variable abundance in function in of the observed latitude, even within soil environments, and further expand the prevalence of this group through many newly identified lineages.

Our dataset of 18S rRNA gene sequences from thousands of samples presents a unique opportunity to estimate the taxonomic diversity of microbial eukaryotes on a global scale. Rarefaction curves of the number of observed OTUs as a function of number of sequencing datasets showed that microbial eukaryote OTU richness did not reach a plateau for any of the ecosystem types, indicating that more extensive sampling efforts could yield a substantial number of novel OTUs (Figure 2E). Partitioning the samples among ecosystem types according to their GOLD ecosystem subtype showed that the highest eukaryotic diversity within marine systems was observed in epipelagic, oceanic, and coastal subtypes (Figure 2F). The highest diversity in freshwater was found in river and peatland subtypes within soil. This contrasts with a prior amplicon sequencing-based study which reported soil as the most diverse ecosystem for eukaryotic taxa (Singer et al. 2021). However, this previous research focused exclusively on topsoil and planktonic samples, while our study highlights that sediment samples represent important contributors to eukaryotic diversity in aquatic environments (Supplementary Figure 3). Freshwater ecosystems might provide nutrients for trophic groups to grow and in turn allow growth of heterotrophic organisms. These ecosystem dynamics align with the crucial role of nutrient cycling in supporting diverse trophic interactions in freshwater systems (Cotner and Biddanda 2002). These results indicate that nutrient availability and a stable ecosystem can in turn produce microbial communities as diverse as others in complex ecosystems such as the soil as freshwater bacterial communities often rival soil communities in phylogenetic diversity (Newton et al. 2011). Further, environmental stability and resource availability are key drivers of microbial community complexity across ecosystem types, including both aquatic and terrestrial environments (Delgado-Baquerizo et al. 2018).

Further, we were interested in patterns of the latitudinal gradient of eukaryotic diversity, as these patterns have been studied for other microbial groups such as marine bacteria (Fuhrman et al. 2008) and is a recurrent question in ecology (Mittelbach et al. 2007). We assessed the correlation between absolute latitude and OTU diversity within each eukaryotic subdivision across the three most prevalent ecosystem types in our samples. We were able to identify distinct patterns in the correlation of the 18S rRNA diversity with absolute latitude. Most taxonomic groups displayed positive coefficients in the quadratic term, indicating that diversity peaks at the Equator and the poles (Figure 2G). This finding contrasts with previous studies, done on observation data based on richness data of observations of benthic diatoms, suggesting that protist diversity shows weak or no correlation with latitudinal gradients, in contrast to plants and animals, which exhibit pronounced responses influenced by body size (Hillebrand and Azovsky 2001). Other investigations focused on specific taxa, such as Fungi, which exhibited a higher diversity at intermediate absolute latitudes (Bahram et al. 2018), in contrast to their metazoan counterparts. This was also evident in our analysis. Further, fungal lineages have been shown to have an increase in diversity near the equator, with tendencies towards reduced diversity in polar regions (Azovsky and Mazei 2013; Lara et al. 2016). While this trend was also visible in our data, we found that the Ciliophora subdivision displayed an even higher diversity in freshwater near the equator, as did the Dinoflagellata subdivision in marine systems.

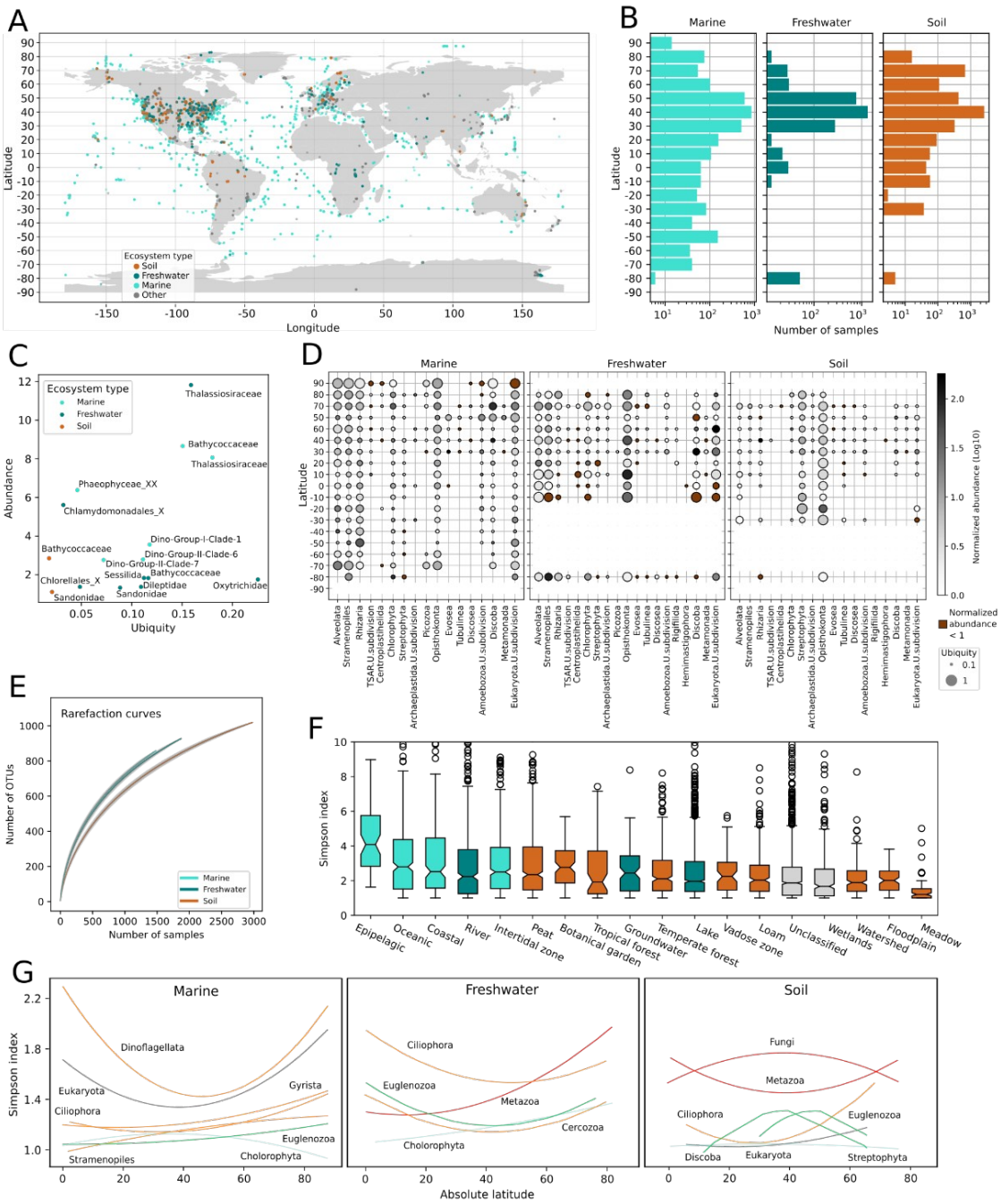


Figure 2. Global distribution of eukaryotes. A) World map illustrating the origin of the 17,371 samples from which 18S rRNA genes were recovered and geolocalization data is available. B) Total number of samples within each latitude range, partitioned by ecosystem type. C) Mean values of ubiquity and normalized abundance for eukaryotic families that exhibited abundance and ubiquity above the 80% quantile. Families belonging to Metazoa, Fungi, and Streptophyta have been excluded. D) Mean abundance and ubiquity of each eukaryotic division across latitude ranges. Rows marked with cross marks indicate no observed samples at that latitude and ecosystem type. Brown circles indicate normalized abundance lower than 1. E) Rarefaction curves representing the mean number of observed OTUs following 10,000 randomizations. Gray shades represent standard deviation values. F) Per-sample diversity distribution for each ecosystem type with at least a 1% share of the datasets. G)

Second-order polynomial fits for Simpson index values as a function of absolute latitude across different ecosystem types; only significant regressions (adjusted p-value < 0.05) are portrayed.

New Lineages Of Eukaryotes

Our comprehensive census of eukaryotic diversity has uncovered a substantial number of novel eukaryotic lineages (2,072 OTUs at 85% sequence similarity) that are solely represented by metagenomic sequences, eluding classification within any known eukaryotic lineage. In order to resolve these unclassified lineages, we expanded our robust species tree utilized for taxonomic placement by incorporating these taxa.

The expanded tree preserved the majority of eukaryotic lineages as monophyletic, while also accommodating several orphan sequences interspersed among them (black branches, n=1,670) (Figure 3A). Noteworthy is the presence of orphan branches grouping within known eukaryotic phyla positioned on extended branches. This branching pattern resembles phenomena documented in host-associated obligate intracellular bacteria (Ferla et al. 2013), potentially indicating a symbiotic lifestyle amid these novel eukaryotic lineages. Known instances of such associations within eukaryotes are rare; the most prominent example is attributed to microsporidians, a lineage of intracellular fungal-related parasites exhibiting extensively divergent 18S rRNA gene sequences (Vossbrinck et al. 1987). The existence of numerous presently unknown intracellular eukaryotes, potentially with reduced genomes, would highlight the ongoing challenge of documenting these intracellular lineages. For example, only through single cell transcriptome sequencing was it possible to identify a novel alveolate intracellular parasite closely related to perkinsids (Cooney et al. 2024). Cultivation of numerous eukaryotic lineages is notoriously challenging, leaving a considerable proportion of eukaryotic diversity unexplored on both genomic and phenotypic levels. In this study, we did not identify previously documented orphan OTUs, as our work mitigated primer binding bias. Indeed, the bias score exhibited significantly higher values (adjusted p value = 0) in orphan sequences compared to reference sequences, irrespective of their origins from isolates or environmental sequencing datasets (Figure 3B).

Notably, the majority of environmental orphan sequences (n=1,670) branched near the root of the eukaryotic tree or in proximity to discoban sequences bridging excavates and the eukaryotic root. We identified 13 clades entirely composed of orphan OTU representatives, likely indicative of major, undescribed new lineages (NLs, colored ranges). Given that 28S rRNA gene sequences were also detected within some metagenomic contigs containing orphan 18S OTU representatives, we examined whether the inclusion of additional phylogenetic information through adding the 28S rRNA gene sequences would impact their classification and placement within the reference tree. In our results the placement remained consistent for sequences in NL11, NL9, NL41, and several orphans within Fungi and TSAR (filled red circles). New lineages 5 and 11 branched outside of all eukaryotic sequences from the reference tree, suggesting they may represent early branching eukaryotes. In contrast, several of these new clades were affiliated with Excavata, specifically the Discoba division, such as NL9. Various clades branched as sister groups to sequences from the discoban phylum Heterolobosea: NL37, NL33, NL12, NL28, NL39, NL30, NL36, and NL41. The most

phylogenetically diverse group within these new lineages, NL41, formed a substantial clade that may represent a new discoban lineage closely related to Heterolobosea.

We further validated NL41 as a distinct lineage using an independent dataset of single-amplified genomes, derived from cells isolated from anoxic marine sediments in Maine. Phylogenomic analyses using a set of 317 conserved eukaryotic proteins indicated that the SAGs formed a novel clade , sister to *Percolomonas cosmopolitus* (Figure 3D). Within the novel SAG clade, we found three SAGs containing 18S rRNA gene sequences, all of which group within the NL41 clade in the 18S tree (Figure 3A,D). This phylogenetic placement, together with the observation that the best BLAST hit from the nucleotide database corresponding to these sequences was the 18S rRNA sequence of *Creneis carolina* (mean identity %=78.3), implies that NL41 belongs to Heterolobosea, potentially related to Tetramitia.

We traced the ecosystem types from which sequencing datasets belonging to the new lineages were obtained and observed that marine ecosystems contributed the largest proportion, followed by freshwater and soil (Figure 3C). Upon assessing the environmental sources of these datasets, we found that wetlands from both marine and freshwater ecosystems were the most frequently represented ecosystem subtypes (Supplementary Figure 4). Wetlands harbor diverse microbial communities, including the newly proposed bacterial phylum Candidatus Cosmopoliota (Zhang et al. 2023), which was identified through metagenome-assembled genomes (MAGs) recovered from this ecosystem.

We used an example lineage for which there was limited taxonomic confidence to demonstrate that more than just the 18S rRNA gene could be extracted for these novel lineages from the metagenomes. We chose the novel lineage NL5 because of its location in the tree among discoban clades and other groups composed of unclassified sequences and better classify it in the eukaryotic tree using recovered coding genes. Thus we applied the tax-aliquots clustering approach (described in the Methods) to identify functionally annotated genes that clustered with the non-18S rRNA gene portions of contigs containing the 18S rRNA gene established through our discovery clustering pipeline (Krinos et al. 2024). This approach clusters sequences based on their *k*-mer frequencies to identify sequences from similar taxonomic lineages that cannot be aligned to existing reference sequences. We retained clustered contigs associated with protein family annotations (Pfam database) with bitscores exceeding 100, thus stringently excluding clustered sequences lacking robust protein annotations. The presence of functional genes within nucleotide sequences which clustered with the 18S rRNA gene-containing contigs provides additional evidence that the 18S rRNA genes are valid. Among sequences exclusively clustering with NL5, we identified 24 distinct Pfam annotations on clustered contigs using the hierarchical clustering approach. These Pfam annotation-containing contigs encompassed multiple membrane proteins (amino acid permease and ABC membrane transporters), a stress/heat shock-related protein (AHSA1), alpha/beta hydrolases, glycosyl transferases, carbohydrate phosphorylases, and sulfur transport proteins. We aligned these functionally-annotated proteins with proteins from the nr database from NCBI (Pruitt et al. 2006) to identify functionally-annotated proteins with similar taxonomic alignment position to NL5. Of 52 unique Pfams annotated across 68 distinct proteins identified in nucleotide-

clustered contigs, 29 of 68 (43%) were taxonomically annotated as eukaryotes and 11 (38%) of those eukaryotes had a maximum percentage identity of less than 80% assigned to the best hit to the NR database. 26 protein sequences (38.2%) did not have any BLAST hit to the NR database at all. This result aligns with the novel diversity identified via the 18S rRNA gene sequences and provides putative functional proteins that appear to have derived from the new lineages.

Previous amplicon-based studies of eukaryotic diversity in environmental samples have indicated that many protist groups contain representatives that remain virtually unknown (de Vargas et al. 2015). The classification of these new clades poses challenges, often limited to specific clades (Reboul et al. 2019; Fermani et al. 2021), primarily due to the conventional short lengths of amplicon libraries, which cover only selected regions of the 18S rRNA gene (Hugerth et al. 2014). To overcome some of these limitations, we recovered nearly complete 18S rRNA genes from assembled contigs and, if co-located, also the 28S rRNA genes of novel eukaryotic lineages, improving phylogenetic resolution. Our findings imply the presence of at least 13 major clades of eukaryotes lacking genomic representation and closely related sequences within the Silva and PR2 reference databases, which represent taxonomic blindpots. We also show how these groups can be placed in the tree of eukaryotes with confidence using single-cell sequencing methods. The detailed information on sample origins provided in this study, predominantly wetlands and sediment systems, will facilitate future targeted sampling efforts aimed at further exploring this unexpected novelty within the eukaryotic tree of life.

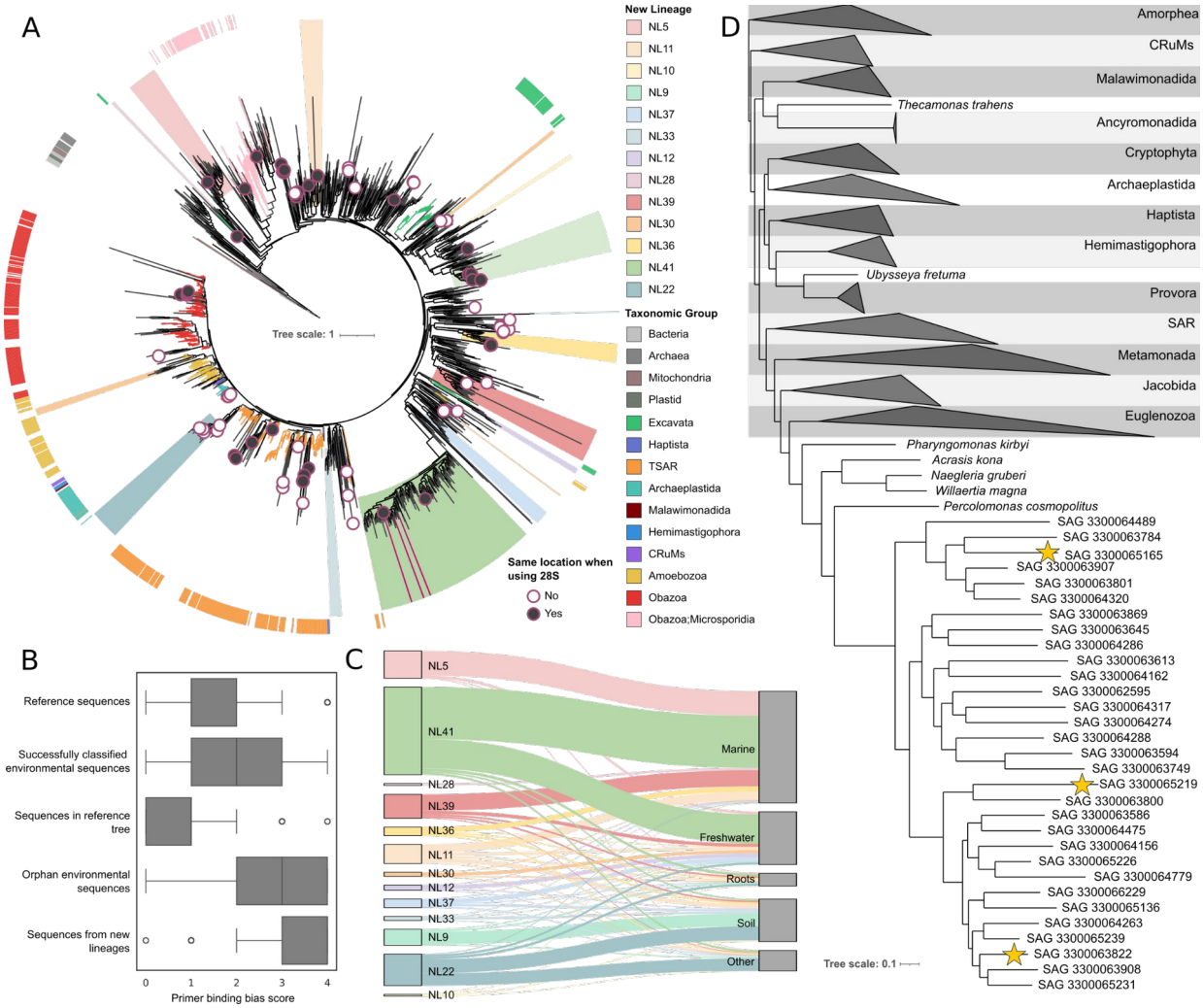


Figure 3. Distribution of new lineages of Eukaryotes. A) Maximum-likelihood tree derived from a concatenated alignment of 18S rRNA and 28S rRNArRNA genes representing reference sequences (branches colored according to division-level taxonomy) and orphan 18S rRNA gene sequences (black branches). Clades likely indicative of major new lineages (NL) are highlighted with colored segments. Dark purple circles denote orphan contigs that encompass a 28S rRNA gene sequence. Filled circles indicate conserved placements maintained upon adding the LSU sequence to the multiple sequence alignment. Purple bars within NL41 represent SAG-derived 18S rRNA gene sequences. B) Distribution of primer binding bias scores across groups of 18S rRNA gene sequences aggregated from our combined database. All comparisons exhibited an adjusted p value = 0 in paired Mann-Whitney tests. C) Diagram illustrating the GOLD Ecosystem type category assigned to the sequencing samples containing sequences from the newly identified lineages. D) Maximum likelihood tree based on a concatenated alignment of 341 protein sequences. Stars denote SAGs from which 18S rRNA gene sequences were extracted, branching within NL41 in the phylogenetic tree represented in A.

Prediction of Associations between Microbes Across the Globe

Microbial eukaryotes are documented to coexist with other microbes, some serving as hosts for bacterial symbionts, which may function as either ecto- or endosymbionts. These relationships can provide nutrient benefits but may also confer defense mechanisms and/or environmental

adaptability (Gast et al. 2009; Husnik et al. 2021). Our abundance correlation results predict associations among unclassified eukaryotes and bacteria annotated as Verrucomicrobia, ABY1, and unclassified bacteria in marine ecosystems (n=7, 5, and 5, respectively; Figure 4A). OTUs categorized within various families of the Amoebozoa supergroup demonstrated a relatively high incidence of positive correlations with prokaryotic OTUs identified as Vicinamibacteraceae (n=15), Ardenticatenales (n=25), and Bacteroidiota (n=13) in freshwater ecosystems (Figure 4B), as well as Rickettsiales (n=5) and Microgenomata (n=4) in soil environments (Figure 4C). Some of these predicted associations may reflect actual symbiotic relationships, and some of the correlations inferred here match previous reports, for example the predicted interaction between rickettsiales bacteria, a lineage that is well known to contain endosymbionts of amoebozoans, algae, and ciliates (Schulz et al. 2016; Lanzoni et al. 2019). Other associations likely reveal environmental or feeding preferences, as suggested by prior studies (Oliverio et al. 2020).

Correlations among various eukaryotic lineages within marine environments also revealed possible associations between unclassified eukaryotes at the supergroup level and unclassified sequences associated with groups such as Heterolobosea (n=6), NL39 (n=5), and NL41 (n=3; Figure 4D). Freshwater ecosystems exhibited the highest number of correlations overall (n=1,809), with sequences from Mastigamoeba and Heterolobosea representing 549 pairs (Figure 4E). The anaerobic growth preferences of many amoebae within these groups, particularly members of the Psalteriomonadidae family, coinciding with the anaerobic lifestyle of some heteroloboseans (Pánek et al. 2012) may account for this finding. We also identified 32 eukaryote-eukaryote co-occurrences across the soil samples, primarily involving unclassified groups of Heterolobosea and eukaryotes unclassified at the supergroup level (n=12; Figure 4F). The latter group was further correlated with opisthokonts, amoebozoans, and rotifers (n=4, 4, and 3, respectively). Some members of the new lineages identified in this work were found correlated with other eukaryotes, such as NL39 and NL41 correlated with unclassified eukaryotes in marine ecosystems, NL30 and NL41 with Mastigamoebidae in freshwater, and NL22 with heteroloboseans in soil. These associations could represent host-parasite interactions considering the highly divergent sequences of these new lineages, which is often a signature of an intracellular lifestyle (Vossbrinck et al. 1987). Predatory protists are known to prey upon both bacteria and other eukaryotes, and certain species of amoebozoans and heteroloboseans exhibit species-specific feeding habits (Berlinches de Gea et al. 2024). Our findings could facilitate the identification of more eukaryotic species-specific predatory relationships between members of these recognized groups and the newly identified eukaryotic lineages.

The proportion of correlated OTU pairs, whether classified to an existing supergroup, orphan, or assigned to a new lineage, varied across ecosystem types (Figure 4G). In marine and soil ecosystems, a predominance of associations occurred between classified and orphan OTUs (0.79 and 0.82, respectively), while in freshwater systems, the majority of correlated pairs comprised two successfully classified OTU representatives (0.82). This trend became even more evident when comparing the proportions of taxon pairs with correlated pairs (Figure 4H), noting a reduced ratio of taxon pairs classified at the supergroup level in freshwater (0.73).

Likewise, we found a reduced proportion of associations between orphan taxa and taxa classified at the supergroup level (0.72). Within marine ecosystems, the relationships between orphan OTUs and those classified as new lineages represented the second-largest proportion (0.21 and 0.2 for correlated OTU pair count and taxa with correlated pairs, respectively). Established routine sampling methods, often optimized for water quality monitoring and pathogen detection, seem to have led to an increased effort but also bias in the exploration of protist taxa in freshwater ecosystems (Dodds 2005). Soil has been characterized as the ecosystem type with the highest novelty regarding documented 18S rRNA gene sequences and marine the ecosystem type with even less novelty than freshwater ecosystems (Singer et al. 2021). Thus, the more frequently recovered predicted associations between known taxa in freshwater can likely be attributed to the bulk of previous work, alongside the fact that the taxa exhibiting the most associations (amoebozoans and discobans) are found to be phylogenetically diverse and may establish relationships with other eukaryotes. Conversely, the higher proportion of associations involving known and orphaned taxa that were found in this study underscores the significance of associations formed by understudied protist groups within marine and soil ecosystems.

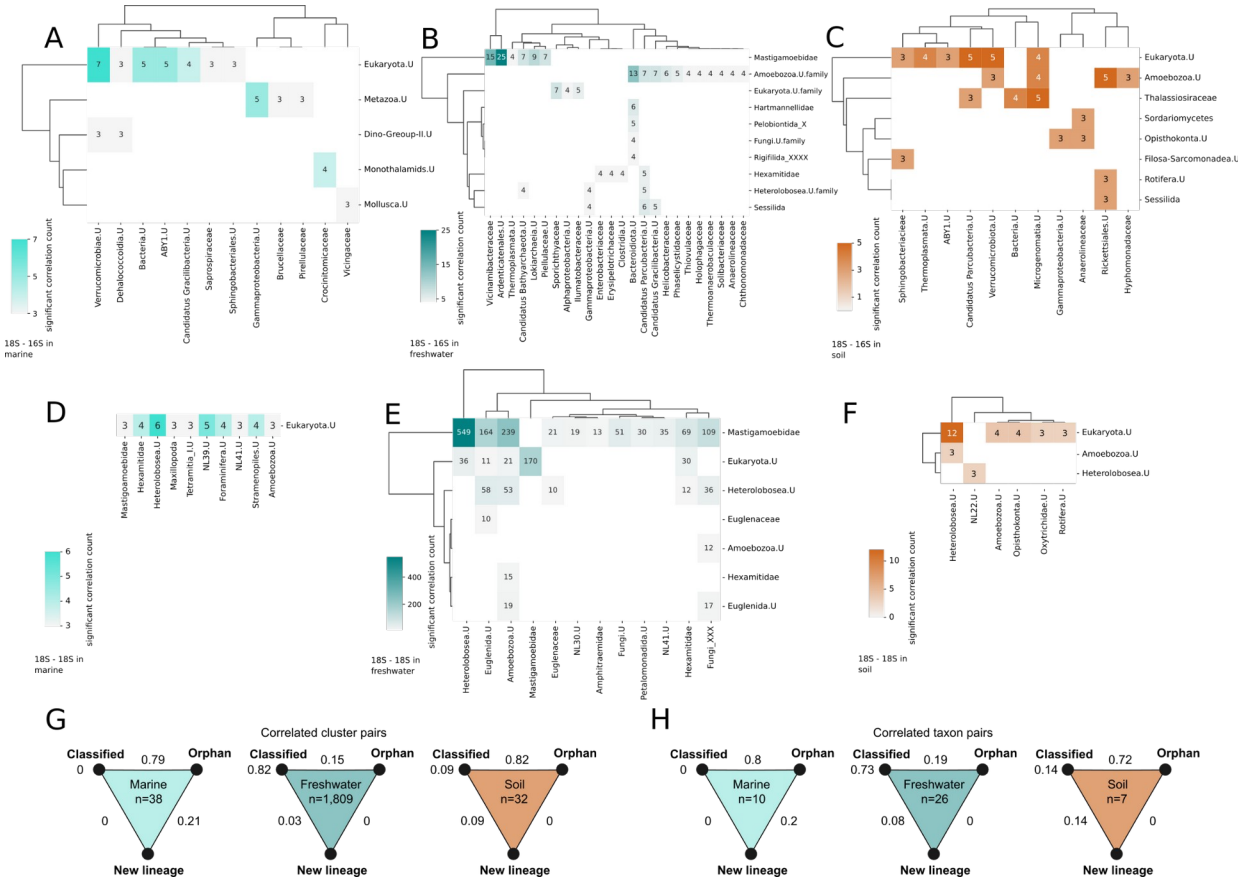


Figure 4. Association between abundances of eukaryotes and prokaryotes across ecosystem types. Heatmaps illustrating the count of significant positive correlations between 18S and 16S rRNA gene sequences (A, B, and C) as well as between 18S and 16S rRNA sequences (D, E, and F) aggregated at the family level. Values are represented for family pairs exhibiting a minimum of three significant (adjusted p-value < 0.05) correlations. G. Proportions of the total number of correlated pairs

(n) between each cluster type according to their taxonomic classification (Classified, Orphan, New Lineage) for each ecosystem type regarding abundance correlations among eukaryotic sequences. H. Proportion of the total number of taxa manifesting correlated pairs between each cluster type across each ecosystem type.

Conclusion

Our comprehensive, metagenome-based survey of eukaryotic diversity substantially expands the known breadth of protistan lineages and reveals previously unseen branches across the eukaryotic tree. By leveraging assembled metagenomes to circumvent the inherent biases of amplicon-based approaches, we captured novel lineages belonging to taxonomic blind spots that often lack cultured representatives and inhabit a broad range of global environments. The identified protists include deeply divergent groups within Excavata and previously uncharacterized taxa branching near the eukaryotic root. We further uncovered likely associations between understudied protists and bacteria, hinting at intricate symbiotic or predator-prey relationships that shape community structure. Together, these findings set the stage for more targeted cultivation efforts and future genomic characterization, and ecological studies aimed at illuminating the roles of these enigmatic lineages and their partners. Ultimately, our findings reshape our understanding of eukaryotic diversity and underscore the potential for future discoveries in unexplored microbial habitats.

Methods

rRNA Sequence Search

We compiled assembled contig files corresponding to 32,520 public datasets derived from metagenome, metatranscriptome, and single-cell sort sequencing projects marked as both public and unrestricted within the Integrated Microbial Genome/Metagenome (IMGM) database. The cmsearch program from the Infernal (Nawrocki and Eddy 2013) package was utilized to search the contig files for sequence matches to each of the following small subunit rRNA (SSU rRNA) models from the Rfam database (Kalvari et al. 2018): eukaryotic 18S rRNA (RF01960), microsporidial 18S rRNA gene (RF02542), archaeal 16S rRNA gene (RF01959), and bacterial 16S rRNA gene (RF00177). Employing the --anytrunc option in cmsearch allowed for the inclusion of long SSU rRNA sequences in our analysis, enabling us to parse the output for contigs displaying sequential and multiple non-overlapping alignments to the covariance models. Furthermore, long subunit rRNA (LSU rRNA) sequences were screened using the eukaryotic 28S rRNA gene (RF02543), archaeal 23S rRNA gene (RF02540), and bacterial 23S rRNA gene (RF02541) covariance models from Rfam, utilizing cmsearch with default parameters.

Preliminary Annotation

We searched the retrieved SSU rRNA sequences against a database that contained all SSU rRNA sequences from both Silva (Quast et al. 2013) and PR2 (Guillou et al. 2013) databases using the BLAST+ (Camacho et al. 2009) program. Sequences with no hits to the database were added to a reference SSU rRNA tree to confirm their phylogenetic affiliation at the domain

level. The reference tree was built as follows: SSU rRNA from the Silva and PR2 databases were clustered at 75% identity to group higher groups and furthr reduce redundancy using Usearch (Edgar 2010) and the resulting centroids were aligned with Mafft (Katoh and Standley 2013) using the Linsi algorithm. Tree building was done with IQtree (Nguyen et al. 2015) with default parameters. Sequences that branched within a cluster that belonged to a different domain were considered misanotations and subsequently removed from the alignment and the tree then rebuilt with the filtered alignment. Sequences confirmed by tree building to belong to the domain Eukarya were added to the 18S rRNA gene sequence database together with those that had a best hit with 18S rRNA gene sequences in Silva and PR2. Our search resulted in a total of 157,956 18S rRNA gene sequences with a length of at least 800 bp from 18,612 different datasets.

rRNA SSU Sequence Clustering

We assembled a database inclusive of 18S rRNA gene sequences from the Silva database (Quast et al. 2013), the PR2 database (Guillou et al. 2013), sequences referenced in the description of the Provora supergroup (Tikhonenkov et al. 2022), long amplicon sequences from a recent study (Jamy et al. 2022), sequences extracted from eukaryotic genomes housed in GenBank (Sayers et al. 2022) represented within the EukProt database (Richter et al. 2022), as well as environmental sequences sourced from IMG/M in February 2022 (Chen et al. 2023). Two-step clustering was conducted using the Usearch program with the -cluster_fast option, initially targeting 97% identity and subsequently at 85% identity to account for the high variability of 18S rRNA geen sequences in related taxa. Following the removal of clusters containing only a single member, we obtained 13,238 sequence clusters (OTUs), wherein the largest cluster comprised 18,608 sequences.

Concurrently, SSU rRNA sequences extracted from the IMG/M database, those recognized as having a 16S rRNA gene as their best hit during BLAST searches, were clustered at 97% identity and subsequently at 85% together with the 16S rRNA gene sequences from Silva and PR2 databases. Clusters with only a single member were excluded from analysis. Taxonomic annotations for each sequence were inferred from the cluster representative's best hit and reformatted to be compatible with the NCBI Taxonomy structure. This procedure produced a total of 839,222 16S rRNA gene sequences divided among 23,233 clusters, representing Bacteria, Mitochondria, Archaea, and Plastids (Supplementary Figure 5).

Primer Binding Analysis

The software primer prospector was employed to search for four different primer pairs (Hugerth et al. 2014; Amaral-Zettler et al. 2009; Nolte et al. 2010; Stoeck et al. 2010) across each 18S rRNA gene sequence from the non-singleton 85% identity clusters. Sequences with an overall weighted score exceeding 1 for any forward or reverse primer were classified as undetected by respective primer pairs. We aggregated the weighted score of all four primer pairs to calculate the binding bias score and subsequently summarized the binding bias score at the cluster level, determining the mean bias score of all sequences within each cluster. We compared the

distribution of primer binding bias scores across different dataset groups: reference sequences (originating from Silva, PR2, Provora, Genbank, and long amplicons), successfully classified environmental sequences, orphan environmental sequences (lacking classification at the supergroup level), and orphan sequences representing potential new lineages. To assess significance, we performed a Kruskal-Wallis test, followed by paired Mann-Whitney tests across pairs, adjusting p-values via the Benjamini-Hochberg procedure using the scipy.stats module (Virtanen et al. 2020). A p value < 0.05 was deemed statistically significant.

18S and 28S rRNA Gene Tree Construction

We selected a subset of 1,247 cluster representative sequences based on specific inclusion criteria: 1) both the 18S and 28S rRNA gene sequences could be extracted from the same contig; 2) 18S sequences exhibited lengths of >= 1,200 bp; 3) 28S sequences displayed lengths of >= 2,800 bp; and 4) centroids were sourced from environmental samples from IMG or from long amplicon sequencing (Jamy et al. 2022). Employing the cmalign program and the RF01960 model, we aligned 18S rRNA gene sequences, and utilized the RF02543 model to align 28S rRNA gene sequences, employing the --match-only option to restrict alignment sites to those represented within the model. We concatenated sequences from both rRNA genes utilizing AMAS (Borowiec 2016) and constructed a tree through IQTree applying the GTR+F+R10 model.

To enhance the tree, we removed terminal branches exhibiting lengths > 0.5. We then recovered the original versions of the remaining sequences and separately aligned the 28S and 28S rRNA gene sequences using mafft-linsi software. Subsequently, we concatenated the alignments of both genes and employed the trimAl program (Capella-Gutiérrez et al. 2009) with the -gt 0.1 option to exclude sites with significant indel content. Afterwards, we used the IQtree program with the GTR+F+R10 model along with 1,000 ultrafast bootstrap replicates to construct a phylogenetic tree. The resultant tree underwent additional manual inspection in which sequences branching within a cluster dominated by sequences from differing supergroups were excluded from the concatenated alignment. The final tree included 577 sequences.

Phylogenetic Placement Based Annotation

We selected an outgroup comprising 38 bacterial, archaeal, and organellar genomes from GenBank (Supplementary Table 4), aligning the respective SSU and LSU rRNA gene sequences with the RF01960 and RF02543 covariance models utilizing the cmalign program with the --match-only option. A phylogenetic tree was developed using sequences from the eukaryotic 18S and 28S tree alongside outgroup SSU and LSU rRNA sequences, constructed by IQtree employing the GTR+F+R10 model and incorporating 1,000 ultrafast bootstrap replicates. All cluster representatives of 18S rRNA gene sequences were aligned with the RF01960 covariance model using the --match-only option, subsequently appending a tail of gaps (-) to each sequence to account for the absence of the corresponding 28S rRNA gene sequence. We utilized the EPA-NG software to place aligned centroid sequences into the

reference tree with the inclusion of the added outgroup, assigning taxonomy to the placed sequences based on the reference tree taxonomy via the Gappa software. The taxonomy of eukaryotic groups was reformatted for compatibility with the PR2 database.

Global Distribution Analysis

We extracted longitude and latitude data from the IMG/M database metadata and visualized it through the geopandas module for Python. Latitude data was digitized into a discrete scale with a step size of 10 degrees. Distribution was visualized using the geopandas module (Jordahl et al. 2022) for Python3. Abundance was defined as the average of the normalized mean sequencing depth across all contigs pertaining to each taxonomic group, while ubiquity was defined as the proportion of samples from comparable environments and latitudes in which each taxonomic group was identified. For the evaluation of eukaryotic family distribution, we focused exclusively on those groups attaining ubiquity and abundance at or above the 80th percentile to emphasize those groups that were highly abundant and ubiquitous.

Sequence Read Depth

We retrieved data on per-contig average read depths across metagenomic datasets whenever available from IMG/M, employing this information to estimate the sample-wide average read depth via the following expression:

$$Sample\ average\ depth = \frac{\sum_{i}^{n} \left(contig_i\ average\ read\ depth \times contig_i\ length \right)}{\sum_{i}^{n} \left(contig_i\ length \right)}$$

The sample average depth was subsequently employed to compute the normalized average read depth for every contig harboring an SSU rRNA gene in the dataset.

Ecosystem Diversity Analysis

For this analysis, we exclusively considered samples from the three most prevalent GOLD ecosystem types represented in our dataset: Marine, Freshwater, and Soil. Given that the metrics utilized for analyzing diversity rely on count data, we simulated read counts based on the normalized average read depth data, employing the expression:

$$Simulated\ read\ count = \frac{Normalized\ average\ read\ depth}{Simulated\ read\ length}$$

With a simulated read depth of 100 bp, read counts were rounded to the nearest integer. Contigs with a simulated read count of 0 were subsequently removed. To account for sequencing disparity, we executed a random rarefaction employing the rrarefy function from R

version (R Core Team, 2013) utilizing the vegan package (Oksanen et al. 2013) to align read counts with that of the smallest sample.

To visualize the 18S rRNA gene OTU sampling effort, we employed a per-sample basis rarefaction curve for each distinct ecosystem type, estimating the mean observed OTUs per new sample across 10,000 randomizations. We then fitted our accumulation curves to various linear and nonlinear models, as outlined below:

$$Simple\,linear: y = bX$$
$$Logarithmic: y = \log(X)$$
$$Power: y = X^b$$
$$Exponential: y = b^X$$
$$Rational: y = (a + bX)/(1 + cX)$$

We used the python ols function from the statsmodels module (Seabold and Perktold 2010) and the curve_fit function from the scipy module (Virtanen et al. 2020) for model fitting. The best-fitting model for the accumulation curve was determined using the Akaike Information Criterion (AIC). The rarefaction curves for the Freshwater and Marine ecosystem types were comparable, exhibiting overlaps. We fitted the rarefied count data to five distinct models, inclusive of an asymptotic rational model. Across all assessments, the power function constituted the best fit, confirming a non-asymptotic relationship.

Subsequently, we utilized the python3 ecopy module (https://github.com/Auerilas/ecopy) to quantitate Shannon H, Simpson, and Inverse Simpson diversity indices on the rarefied OTU table. Second-order polynomial linear regressions were applied to model the relationship between diversity (measured by the Simpson index) and absolute latitude. For this analysis, we segmented the OTU table for each taxon across varying taxonomic levels: supergroup, division, and subdivision, limiting our focus to taxonomic groups present in a minimum of 10 samples. To account for data partitioning, we adjusted p-values for polynomial fits using the Benjamini-Hochberg procedure.

Search For Potential New Lineages

We constructed a phylogenetic tree utilizing OTU representative sequences classified under the eukaryotic domain and categorized unknown supergroups. These sequences were combined with data from the reference tree employed for phylogenetic placement analysis. Using concatenated alignments for the 18S and 28S genes based on the RF01960 and RF02543 covariance models, we constructed the tree employing the IQtree program with the GTR+F+R10 model and 1,000 ultrafast bootstrap replicates. Clades comprising at least five leaves classified as eukaryotic with unknown supergroup status and belonging to clusters containing only metagenomic sequences, along with maximum bootstrap support values, were defined as candidates for potential new lineages. The environmental origins of samples

harboring sequences from these new lineages were visualized employing the plotting capabilities of the plotly python3 module (https://github.com/plotly/plotly.py).

Phylogenomic Analysis

To further validate the taxonomic affiliation of the potential new lineages, we used an independent dataset of single-amplified genomes derived from cells isolated from anoxic marine sediments in Maine. We performed a phylogenetic analysis expecting a robust classification based on several marker genes. Initial filtering included the removal of SAGs with fewer than 100 putatively annotated eukaryotic proteins. The peptides from the remaining SAGs were incorporated into broadly conserved gene families using PhyloToL (Cerón-Romero et al. 2019), identifying 345 gene families well represented among the remaining SAGs. Subsequently, we excluded SAGs with low representation within this pool of gene families (i.e., present in <80 gene families), thereafter generating single-gene alignments, conducting homology assessments, and selecting orthologs with PhyloToL. For each alignment, columns with more than 90% missing data were removed with trimAl (Capella-Gutiérrez et al. 2009). Following alignment processing, we concatenated the 341 single-gene alignments, utilizing IQTree2 (v2.2; Minh et al. 2019) for phylogenomic reconstruction, applying the LG+C60+G+F model under the PMSF approximation with 100 bootstrap iterations.

Tax-aliquots Clustering Method

The tax-aliquots clustering method (Krinos et al. 2024) was employed to revisit the original metagenomes, recovering contigs akin to the non-18S rRNA gene portions of the 18S rRNA gene-containing contigs used for identifying new lineages. We utilized mmseqs2 (version 13.45111) (Steinegger and Söding 2017) to precluster all contigs from samples from which a new lineage was identified, establishing a threshold coverage value of 0.01 and a percent sequence identity of 0.7, utilizing coverage mode 1. All contigs within clusters containing at least one sequence inclusive of an 18S rRNA gene-containing contig were retained, thus minimizing the total input size of sequence content while limiting sequence representatives to a manageable number by restricting the search to contigs already exhibiting some degrees of similarity to the rRNA gene-containing contigs. Subsequent to obtaining this original clustered assembly, we differentiated sequences associated with the original 18S rRNA gene-containing contigs from those that did not incorporate these genes. With the coordinates of the rRNA gene within the 18S rRNA gene-containing contig, we masked the 18S rRNA gene sequences, recombining the files as input for tax-aliquots. The reference content consisted of the masked 18S rRNA gene-containing contigs, while query content comprised other contigs remaining subsequent to the initial mmseqs2 filtering (employing SeqIO from Biopython for sequence masking; (Cock et al. 2009). We subsequently clustered this sequence content utilizing mmseqs2 (Steinegger and Söding 2017) with a coverage of 0.2 and percent sequence identity of 0.2, retaining clusters that encompassed at least one of the masked rRNA-gene containing contigs and at least one additional contig that did not contain an 18S rRNA gene. The resulting assembly was subject to further clustering with sequence information parsed into k-mers of length 3.

Correlation Between SSU rRNA Sequence Read Depth

For each of the most frequently represented ecosystem types in our dataset (Soil, Freshwater, and Marine), we evaluated the sequences within each 18S rRNA gene cluster, retaining only those with a minimum of three sequences across different samples from specific environments, resulting in the acquisition of ecosystem-filtered 18S rRNA gene clusters. An analogous procedure was conducted to retrieve ecosystem-filtered 16S clusters, from which we calculated Spearman's rank correlation coefficients for each pair of 18S and 16S rRNA gene filtered clusters whenever a minimum of three samples contained sequences from both clusters. Adjustments for p-values were made using the Benjamini-Hochberg method for each GOLD ecosystem type independently, with a replication of this process for correlations among 18S rRNA gene sequence clusters, incorporating the additional step of excluding autocorrelations.

Acknowledgments

The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Data availability

Sequence data including SSU rRNA sequences and SAG assemblies, together with tables containing the full metadata from sequencing datasets from IMGM and sequence cluster information are available at: https://portal.nersc.gov/cfs/nelli/eukcensus/

References

Adl, S. M., Leander, B. S., Simpson, A. G. B., Archibald, J. M., Anderson, O. R., Bass, D., Bowser, S. S., Brugerolle, G., Farmer, M. A., Karpov, S., Kolisko, M., Lane, C. E., Lodge, D. J., Mann, D. G., Meisterfeld, R., Mendoza, L., Moestrup, Ø., Mozley-Standridge, S. E., Smirnov, A. V., & Spiegel, F. (2007). Diversity, nomenclature, and taxonomy of protists. *Systematic Biology*, *56*(4), 684–689.

Al Jewari, C., & Baldauf, S. L. (2023). An excavate root for the eukaryote tree of life. *Science Advances*, *9*(17), eade4973.

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS One*, *4*(7), e6372.

Azovsky, A., & Mazei, Y. (2013). Do microbes have macroecology? Large-scale patterns in the diversity and distribution of marine benthic ciliates. *Global Ecology and Biogeography: A Journal of Macroecology*, *22*(2), 163–172.

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., … Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, *560*(7717), 233–237.

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology*, 60(3), 291-302.

Berlinches de Gea, A., Geisen, S., Grootjans, F., Wilschut, R. A., & Schwelm, A. (2024). Species-specific predation determines the feeding impacts of six soil protist species on bacterial and eukaryotic prey. *European Journal of Protistology*, *94*(126090), 126090.

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660.

Burki, F., & Pawlowski, J. (2006). Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts. *Molecular biology and evolution*, 23(10), 1922-1930.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* , *25*(15), 1972–1973.

Cerón-Romero, M. A., Maurer-Alcalá, X. X., Grattepanche, J.-D., Yan, Y., Fonseca, M. M., & Katz, L. A. (2019). PhyloToL: A Taxon/Gene-Rich Phylogenomic Pipeline to Explore Genome Evolution of Diverse Eukaryotes. *Molecular Biology and Evolution*, *36*(8), 1831–1842.

Céza, V., Kotyk, M., Kubánková, A., Yubuki, N., Šťáhlavský, F., Silberman, J. D., & Čepička, I. (2022). Free-living Trichomonads are Unexpectedly Diverse. *Protist*, *173*(4), 125883.

Chen, I.-M. A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S. J., Webb, C., Wu, D., Varghese, N. J., Reddy, T. B. K., Mukherjee, S., Ovchinnikova, G., Nolan, M., Seshadri, R., Roux, S., Visel, A., Woyke, T., … Ivanova, N. N. (2023). The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Research*, *51*(D1), D723–D732.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–1423.

Collins, W. E., & Jeffery, G. M. (2007). Plasmodium malariae: parasite and disease. *Clinical Microbiology Reviews*, *20*(4), 579–592.

Cooney, E. C., Leander, B. S., & Keeling, P. J. (2024). Hidden syndinian and perkinsid infections in dinoflagellate hosts revealed by single-cell transcriptomics. *The ISME Journal*, *18*(1).

Cotner, J. B., & Biddanda, B. A. (2002). Small players, large role: microbial influence on biogeochemical processes in pelagic aquatic ecosystems. *Ecosystems*, 5, 105-121.

Creators Kelsey Jordahl1 Joris Van den Bossche Martin Fleischmann Jacob Wasserman James McBride Jeffrey Gerard Jeff Tratner Matthew Perry Adrian Garcia Badaracco Carson Farmer2 Geir Arne Hjelle3 Alan D. Snow4 Micah Cochran5 Sean Gillies6 Lucas Culbertson Matt Bartos7 Nick Eubank8 maxalbert Aleksey Bilogur Sergio Rey9 Christopher Ren Dani Arribas-Bel Leah Wasser10 Levi John Wolf Martin Journois11 Joshua Wilson12 Adam Greenhall Chris Holdgraf13 Filipe François Leblanc14 Show affiliations 1. Planet Labs 2. @textileio 3. Nextbridge Analytics 4. @corteva 5. City of Athens, Alabama, USA 6. Mapbox 7. @kLabUM , @ESIPFed 8. Academia 9. Center for Geopatial Sciences, University of California, Riverside 10. Earth Lab-University of Colorado, Boulder 11. MICHELIN solutions 12. @fernleafinteractive 13. UC Berkeley 14. @shopify. (n.d.). *geopandas/geopandas: v0.8.1*. https://doi.org/10.5281/zenodo.3946761

del Campo, J., Sieracki, M. E., Molestina, R., Keeling, P., Massana, R., & Ruiz-Trillo, I. (2014). The others: our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, *29*(5), 252–259.

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., ... & Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science*, 359(6373), 320-325.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., … Karsenti, E. (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605.

Dodds, W. K. (2005). Freshwater Microbiology: Biodiversity and Dynamic Interactions of Microorganisms in the Aquatic Environment. By David C Sigee. Chichester (United Kingdom) and Hoboken (New Jersey): John Wiley & Sons. 180.00 (hardcover); 65.00 (paper). xix + 524 p; ill.; index. ISBN: 0–471–48528–4 (hc); 0–471–48529–2 (pb). 2005. *The Quarterly Review of Biology*, *80*(3), 355–356.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* , *26*(19), 2460–2461.

Eglit, Y., Shiratori, T., Jerlström-Hultqvist, J., Williamson, K., Roger, A. J., Ishida, K.-I., & Simpson, A. G. B. (2024). Meteora sporadica, a protist with incredible cell architecture, is related to Hemimastigophora. *Current Biology: CB*, *34*(2), 451–459.e6.

Einarsson, E., Ma'ayeh, S., & Svärd, S. G. (2016). An up-date on Giardia and giardiasis. *Current Opinion in Microbiology*, *34*, 47–52.

Ferla, M. P., Thrash, J. C., Giovannoni, S. J., & Patrick, W. M. (2013). New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. *PloS One*, *8*(12), e83383.

Fermani, P., Metz, S., Balagué, V., Descy, J. P., Morana, C., Logares, R., Massana, R., & Sarmento, H. (2021). Microbial eukaryote assemblages and potential novel diversity in four tropical East African Great Lakes. *FEMS Microbiology Ecology*, *97*(9).

Fernández, L. D., Fournier, B., Rivera, R., Lara, E., Mitchell, E. A. D., & Hernández, C. E. (2016). Water–energy balance, past ecological perturbations and evolutionary constraints shape the latitudinal diversity gradient of soil testate amoebae in south-western South America. *Global Ecology and Biogeography: A Journal of Macroecology*, *25*(10), 1216–1227.

Foissner, W. (2006). Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozoologica*, *45*, 111–136.

Foissner, W., Blatterer, H., & Foissner, I. (1988). The hemimastigophora (Hemimastix amphikineta nov. gen., nov. spec.), a new protistan phylum from gondwanian soils. *European Journal of Protistology*, *23*(4), 361–383.

Foissner, W., Foissner, H., & Blatterer. (1988). The hemimastigophora (Hemimastix amphikineta nov. gen., nov. spec.), a new protistan phylum from gondwanian soils. *European Journal of Protistology*, *23*(4), 361–383.

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., & Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences*, 105(22), 7774-7778.

Gast, R. J., Sanders, R. W., & Caron, D. A. (2009). Ecological strategies of protists and their symbiotic relationships with prokaryotic microbes. *Trends in Microbiology*, *17*(12), 563–569.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., Del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(Database issue), D597–D604.

Hillebrand, H., & Azovsky, A. I. (2001). Body size determines the strength of the latitudinal diversity gradient. *Ecography*, *24*(3), 251–256.

Horn, D. (2022). A profile of research on the parasitic trypanosomatids and the diseases they cause. *PLoS Neglected Tropical Diseases*, *16*(1), e0010040.

Howe, A. T., Bass, D., Scoble, J. M., Lewis, R., Vickerman, K., Arndt, H., & Cavalier-Smith, T. (2011). Novel cultured protists identify deep-branching environmental DNA clades of cercozoa: New Genera Tremula, Micrometopion, Minimassisteria, Nudifila, Peregrinia. *Protist*, *162*(2), 332–372.

Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., Wilmes, P., & Andersson, A. F. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PloS One*, *9*(4), e95567.

Husnik, F., Tashyreva, D., Boscaro, V., George, E. E., Lukeš, J., & Keeling, P. J. (2021). Bacterial and archaeal symbioses with protists. *Current Biology: CB*, *31*(13), R862–R877.

Jamy, M., Biwer, C., Vaulot, D., Obiol, A., Jing, H., Peura, S., Massana, R., & Burki, F. (2022). Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nature Ecology & Evolution*, *6*(10), 1458–1470.

Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., & Petrov, A. I. (2018). Non-Coding RNA Analysis Using the Rfam Database. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *62*(1), e51.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Krinos, A. I., Bowers, R. M., Rohwer, R. R., McMahon, K. D., Woyke, T., & Schulz, F. (2024). Time-series metagenomics reveals changing protistan ecology of a temperate dimictic lake. *Microbiome*, *12*(1), 133.

Krinos, A. I., Mars Brisbin, M., Hu, S. K., & Cohen, N. R. (2024). Missing microbial eukaryotes and misleading meta-omic conclusions. *Nature Communications*.

Lanzoni, O., Sabaneyeva, E., Modeo, L., Castelli, M., Lebedeva, N., Verni, F., Schrallhammer, M., Potekhin, A., & Petroni, G. (2019). Diversity and environmental distribution of the cosmopolitan endosymbiont “Candidatus Megaira.” *Scientific Reports*, *9*.

Lara, E., Mitchell, E. A. D., Moreira, D., & López García, P. (2011). Highly diverse and seasonally dynamic protist community in a pristine peat bog. *Protist*, *162*(1), 14–32.

Lara, E., Roussel-Delif, L., Fournier, B., Wilkinson, D. M., & Mitchell, E. A. D. (2016). Soil microorganisms behave like macroscopic organisms: patterns in the global distribution of soil euglyphid testate amoebae. *Journal of Biogeography*, *43*(3), 520–532.

Lax, G., Eglit, Y., Eme, L., Bertrand, E. M., Roger, A. J., & Simpson, A. G. B. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, *564*(7736), 410–414.

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., Singer, D., Mayor, J., Bunge, J., Sernaker, S., Siemensmeyer, T., Trautmann, I., Romac, S., Berney, C., Kozlov, A., Mitchell, E. A. D., Seppey, C. V. W., Egge, E., Lentendu, G., … Dunthorn, M. (2017).

Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution*, *1*(4), 91.

Meyer, A., Todt, C., Mikkelsen, N. T., & Lieb, B. (2010). Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity. *BMC Evolutionary Biology*, *10*, 70.

Minh, B. Q., Trifinopoulos, J., Schrempf, D., Schmidt, H. A., & Lanfear, R. (2019). IQ-TREE version 2.0: tutorials and Manual Phylogenomic software by maximum likelihood. *URL Http://www. Iqtree. Org*.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., ... & Turelli, M. (2007). Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecology letters*, 10(4), 315-331.

Morey, K. C., Myler, E., Hanner, R., & Tetreault, G. (2024). Taxonomic blind spots: A limitation of environmental DNA metabarcoding-based detection for Canadian freshwater fishes. *Environmental DNA (Hoboken, N.J.)*, *6*(6), e70054.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* , *29*(22), 2933–2935.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and molecular biology reviews*, 75(1), 14-49.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Nolte, V., Pandey, R. V., Jost, S., Medinger, R., Ottenwälder, B., Boenigk, J., & Schlötterer, C. (2010). Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular Ecology*, *19*(14), 2908–2915.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., & Oksanen M. (2013). Package "vegan." *Community Ecology Package, Version*, *2*(9), 1–295.

Oliverio, A. M., Geisen, S., Delgado-Baquerizo, M., Maestre, F. T., Turner, B. L., & Fierer, N. (2020). The global-scale distributions of soil protists and their contributions to belowground systems. *Science Advances*, *6*(4), eaax8787.

Overgaard, C. K., Jamy, M., Radutoiu, S., Burki, F., & Dueholm, M. K. D. (2024). Benchmarking long-read sequencing strategies for obtaining ASV-resolved rRNA operons from environmental microeukaryotes. *Molecular Ecology Resources*, e13991.

Pánek, T., Silberman, J. D., Yubuki, N., Leander, B. S., & Cepicka, I. (2012). Diversity, evolution and molecular systematics of the Psalteriomonadidae, the main lineage of anaerobic/microaerophilic heteroloboseans (excavata: discoba). *Protist*, *163*(6), 807–831.

Poulin, M., Daugbjerg, N., Gradinger, R., Ilyash, L., Ratkova, T., & von Quillfeldt, C. (2011). The pan-Arctic biodiversity of marine pelagic and sea-ice unicellular eukaryotes: a first-attempt assessment. *Marine Biodiversity: A Journal of the Senckenberg Research Institute / Senckenberg Forschungsinstitut Und Naturmudeum*, *41*(1), 13–28.

Pruitt, K., Tatusova, T., & Maglott, D. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*, D61–D65.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596.

Reboul, G., Moreira, D., Bertolino, P., Hillebrand-Voiculescu, A. M., & López-García, P. (2019). Microbial eukaryotes in the suboxic chemosynthetic ecosystem of Movile Cave, Romania. *Environmental Microbiology Reports*, *11*(3), 464–473.

Richter, D. J., Berney, C., Strassert, J. F. H., Poh, Y.-P., Herman, E. K., Muñoz-Gómez, S. A., Wideman, J. G., Burki, F., & de Vargas, C. (2022). EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal*, *2*(e56).

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26.

Schulz, F., Martijn, J., Wascher, F., Lagkouvardos, I., Kostanjšek, R., Ettema, T. J. G., & Horn, M. (2016). A Rickettsiales symbiont of amoebae with ancient features. *Environmental Microbiology*, *18*(8), 2326–2342.

Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. *SciPy*, *7*(1).

Singer, D., Seppey, C. V. W., Lentendu, G., Dunthorn, M., Bass, D., Belbahri, L., Blandenier, Q., Debroas, D., de Groot, G. A., de Vargas, C., Domaizon, I., Duckert, C., Izaguirre, I., Koenig, I., Mataloni, G., Schiaffino, M. R., Mitchell, E. A. D., Geisen, S., & Lara, E. (2021). Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. *Environment International*, *146*, 106262.

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly

complex eukaryotic community in marine anoxic water. *Molecular Ecology*, *19 Suppl 1*, 21–31.

Team, R. C., & Others. (2013). R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria*.

Tekle, Y. I., Wang, F., Wood, F. C., Anderson, O. R., & Smirnov, A. (2022). New insights on the evolutionary relationships between the major lineages of Amoebozoa. *Scientific reports*, 12(1), 11173.

Tikhonenkov, D. V., Mikhailov, K. V., Gawryluk, R. M. R., Belyaev, A. O., Mathur, V., Karpov, S. A., Zagumyonnyi, D. G., Borodina, A. S., Prokina, K. I., Mylnikov, A. P., Aleoshin, V. V., & Keeling, P. J. (2022). Microbial predators form a new supergroup of eukaryotes. *Nature*, *612*(7941), 714–719.

Torruella, G., De Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M. A., Del Campo, J., ... & Ruiz-Trillo, I. (2015). Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Current Biology*, 25(18), 2404-2410.

Torruella, G., Galindo, L. J., Moreira, D., & López-García, P. (2024). Phylogenomics of neglected flagellated protists supports a revised eukaryotic tree of life. In *bioRxiv* (p. 2024.05.15.594285)

Urich, T., Lanzén, A., Qi, J., Huson, D. H., Schleper, C., & Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PloS One*, *3*(6), e2527.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Vossbrinck, C. R., Maddox, J. V., Friedman, S., Debrunner-Vossbrinck, B. A., & Woese, C. R. (1987). Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature*, *326*(6111), 411–414.

Williamson, K., Eme, L., Baños, H., McCarthy, C., Susko, E., Kamikawa, R., Orr, R. J. S., Muñoz-Gómez, S. A., Simpson, A. G. B., & Roger, A. J. (2024). A robustly rooted tree of eukaryotes reveals their excavate ancestry. In *bioRxiv*.

Woyke, T., Doud, D. F. R., & Schulz, F. (2017). The trajectory of microbial single-cell sequencing. *Nature Methods*, *14*(11), 1045–1054.

Wuyts, J., Van de Peer, Y., & De Wachter, R. (2001). Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Research*, *29*(24), 5017–5028.

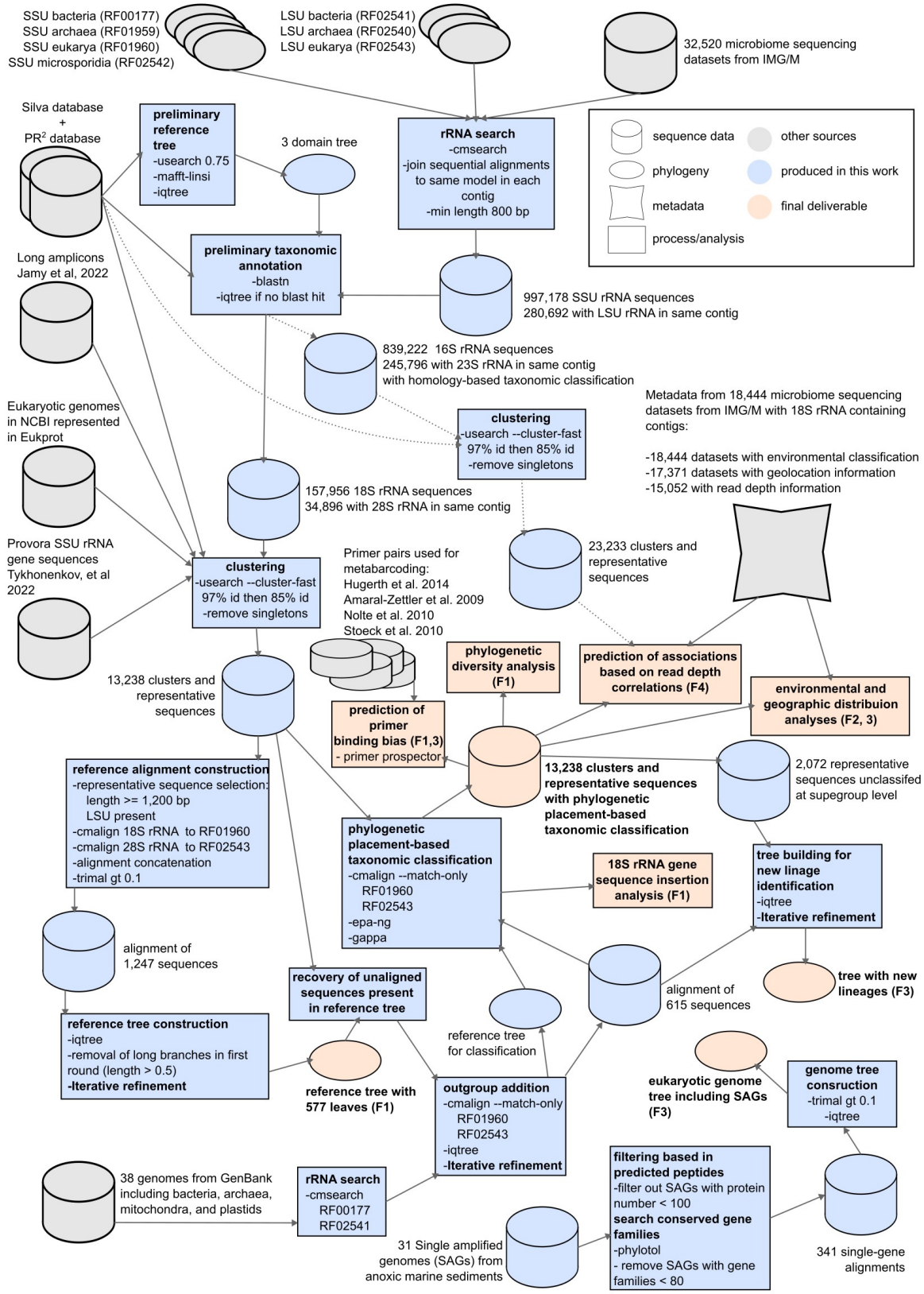
Xie, Q., Lin, J., Qin, Y., Zhou, J., & Bu, W. (2011). Structural diversity of eukaryotic 18S rRNA and its impact on alignment and phylogenetic reconstruction. *Protein & Cell*, *2*(2), 161–170.

Zancan, S., Trevisan, R., & Paoletti, M. G. (2006). Soil algae composition under different agro-ecosystems in North-Eastern Italy. *Agriculture, Ecosystems & Environment*, *112*(1), 1–12.

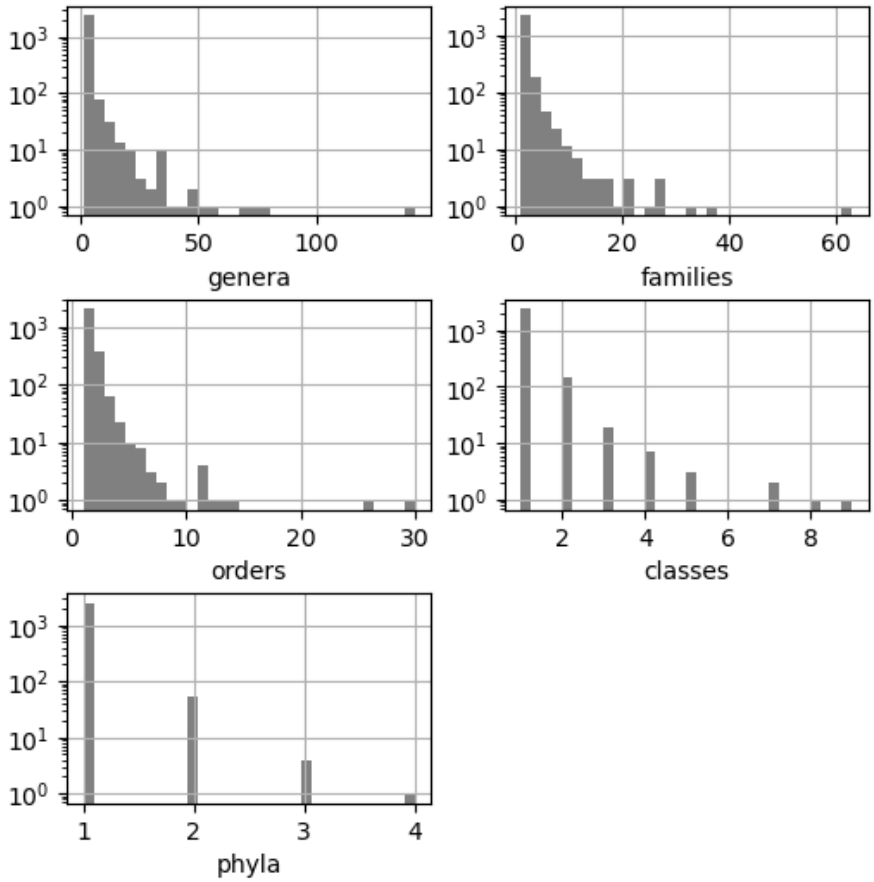
Zhang, Z.-F., Liu, L.-R., Pan, Y.-P., Pan, J., & Li, M. (2023). Long-read assembled metagenomic approaches improve our understanding on metabolic potentials of microbial community in mangrove sediments. *Microbiome*, *11*(1), 188.

Zhao, J., Fan, D., Guo, W., Wu, J., Zhang, X., Zhuang, X., & Kong, W. (2023). Precipitation drives soil protist diversity and community structure in dry grasslands. *Microbial Ecology*, *86*(4), 2293–2304.

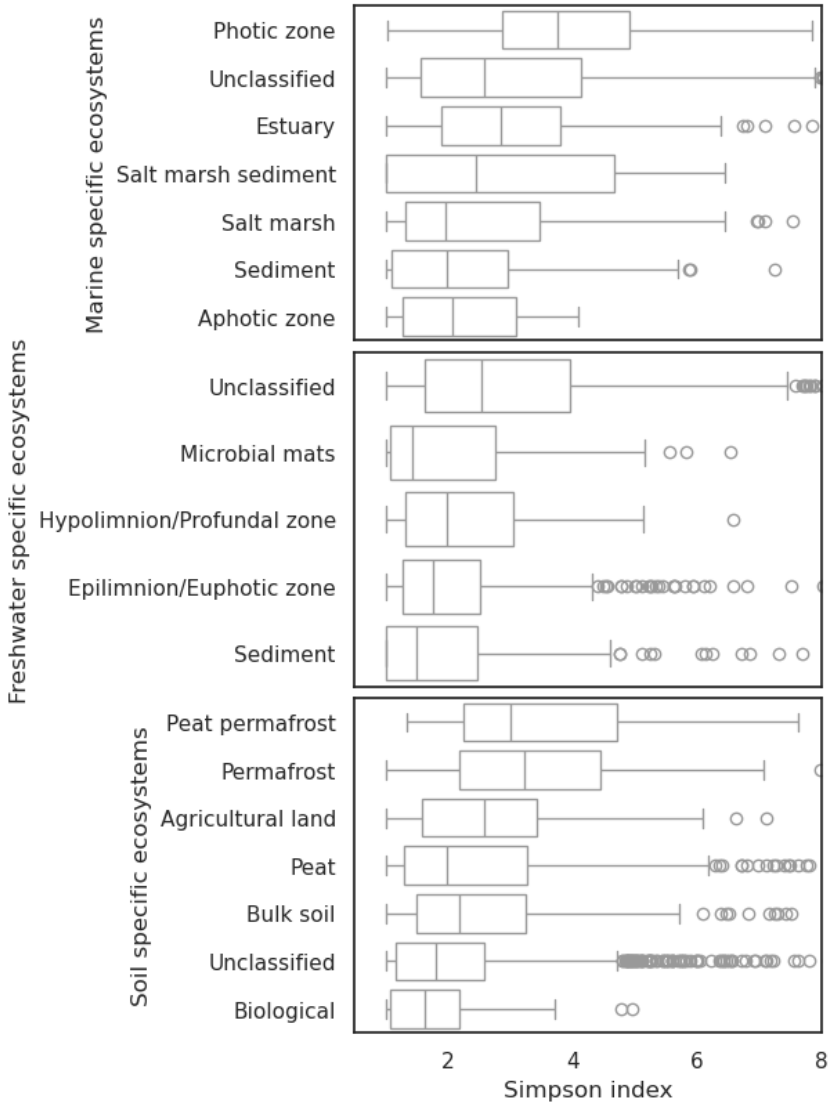
Supplementary Figures



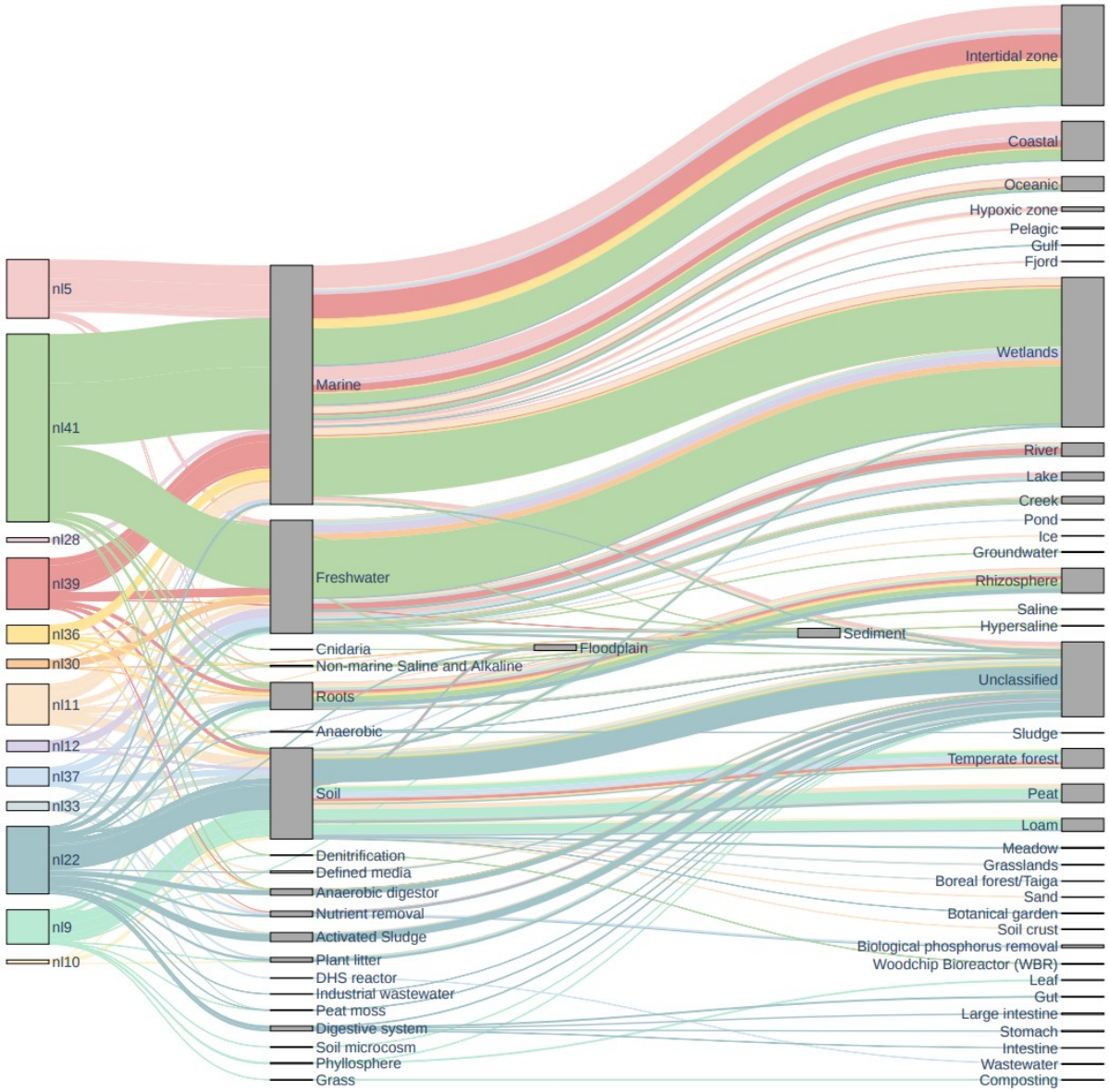
Supplementary figure 1. Summary of the methods used in this work. Cylinders correspond to sequence data, while the ovals indicate phylogenies, stars indicate metadata, and squares represent analyses. Some final deliverables include an F and one or two numbers in parenthesis, which indicate the figure of the main text in which they are included. “Iterative refinement” refers to the process of manually inspecting the tree and removing sequences representing rouge taxa and long branches that could produce artifacts in the topology from the seed alignment and rebuilding the tree.



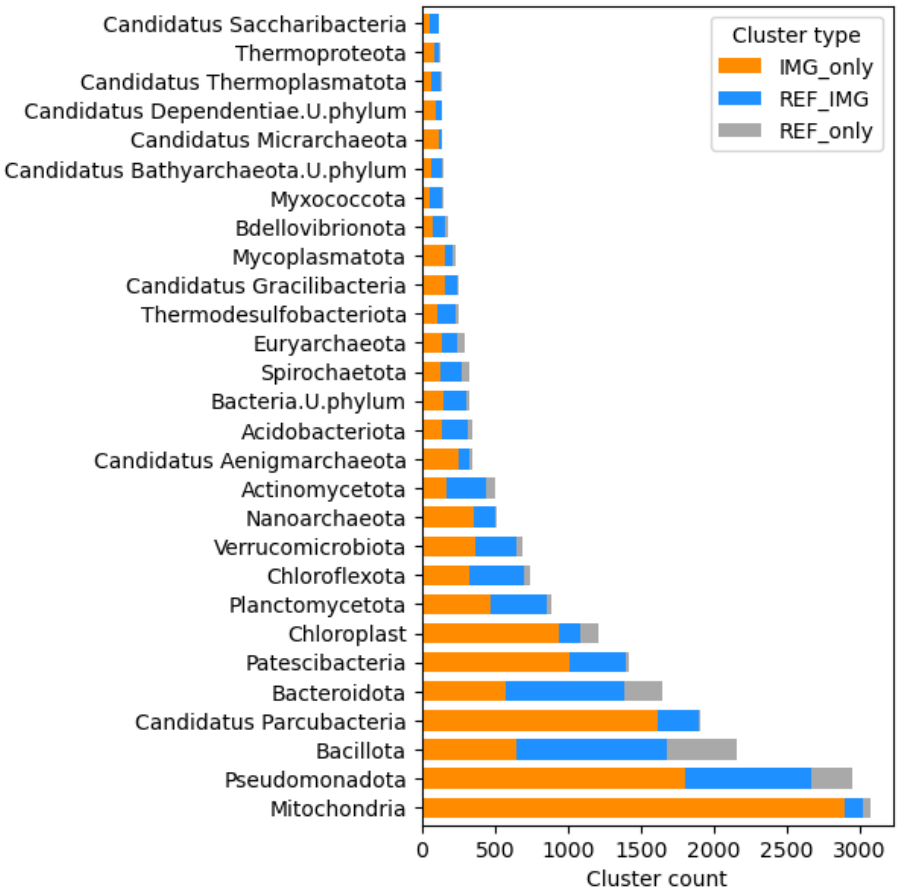
Supplementary figure 2. Distribution of clusters comprising any given number of different taxa within a single reference sequence-only cluster.



Supplementary figure 3. Per-sample diversity distribution for each specific ecosystem with at least 1% share of the datasets separated by ecosystem types.



Supplementary figure 4. Contribution of each ecosystem subtype and ecosystem type to the observed pool of 18S rRNA gene sequences attributable to the 13 new lineages.



Supplementary figure 5. Abundance of clusters for each prokaryotic taxonomic group representing at least 0.5% of the total number of clusters within the analyzed environmental samples. Bars are partitioned according to their share of cluster types as defined by the origin of the member sequences. REF_IMG: includes both sequences from IMG and from reference databases; REF_only: includes sequences from reference databases exclusively.

Supplementary Tables

Supplementary Table 1. Number of sequencing datasets with observed 18S rRNA sequences by Ecosystem Subtype.

Supplementary Table 2. Counts of samples with no observed 18S rRNA sequences by Ecosystem Subtype.

Supplementary Table 3. Mean ubiquity, abundance and weighted abundance data of 18s rRNA sequences among ecosystem types at the family level.

Supplementary Table 4. Information on selected genomes of archaea, bacteria, mitochondria and plastids used as outgroup for the tree built for phylogenetic placement-based taxonomic classification.