Received: 7 April 2022 | Revised: 23 January 2023 | Accepted: 21 February 2023

DOI: 10.1111/1755-0998.13776

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes

Joao Pedro Saraiva[1] | Alexander Bartholomäus[2] | Rodolfo Brizola Toscan[1] | Petr Baldrian[3] | Ulisses Nunes da Rocha[1]

[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research—UFZ GmbH, Leipzig, Germany

[2]GFZ German Research Centre for Geosciences, Section Geomicrobiology, Potsdam, Germany

[3]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Praha, Czech Republic

Correspondence
Ulisses Nunes da Rocha, Department of Environmental Microbiology, Helmholtz Centre for Environmental Research—UFZ GmbH, Leipzig, Saxony 04318, Germany.
Email: ulisses.rocha@ufz.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 460129525; Helmholtz-Gemeinschaft, Grant/Award Number: VH-NG-1248 Micro' Big Data'

Handling Editor: Lucie Zinger

Abstract

As most eukaryotic genomes are yet to be sequenced, the mechanisms underlying their contribution to different ecosystem processes remain untapped. Although approaches to recovering Prokaryotic genomes have become common in genome biology, few studies have tackled the recovery of eukaryotic genomes from metagenomes. This study assessed the reconstruction of microbial eukaryotic genomes using 6000 metagenomes from terrestrial and some transition environments using the EukRep pipeline. Only 215 metagenomic libraries yielded eukaryotic bins. From a total of 447 eukaryotic bins recovered 197 were classified at the phylum level. Streptophytes and fungi were the most represented clades with 83 and 73 bins, respectively. More than 78% of the obtained eukaryotic bins were recovered from samples whose biomes were classified as host-associated, aquatic, and anthropogenic terrestrial. However, only 93 bins were taxonomically assigned at the genus level and 17 bins at the species level. Completeness and contamination estimates were obtained for a total of 193 bins and consisted of 44.64% ($\sigma = 27.41\%$) and 3.97% ($\sigma = 6.53\%$), respectively. *Micromonas commoda* was the most frequent taxon found while *Saccharomyces cerevisiae* presented the highest completeness, probably because more reference genomes are available. Current measures of completeness are based on the presence of single-copy genes. However, mapping of the contigs from the recovered eukaryotic bins to the chromosomes of the reference genomes showed many gaps, suggesting that completeness measures should also include chromosome coverage. Recovering eukaryotic genomes will benefit significantly from long-read sequencing, development of tools for dealing with repeat-rich genomes, and improved reference genomes databases.

KEYWORDS
eukaryotes, genome-resolved metagenomics, Hypocreales, Mamiellales, Saccharomycetales

1 | INTRODUCTION

Microbial eukaryotes play critical roles in ecosystem processes by decomposing organic material (e.g., decomposition processes by fungi in soil) (Baldrian et al., 2012), predating on other microbes, or producing organic compounds from inorganic compounds (Bik et al., 2012; Bulan et al., 2018; Lind & Pollard, 2021; West et al., 2018). An estimated 8.7 million eukaryotic species inhabit our planet (Sweetlove, 2011), but as of 7 November 2022, only slightly more than 37,500 eukaryotic (fungi, invertebrates, plant, protozoan, mammalian vertebrates, and other vertebrates) taxonomy ids exist in RefSeq (O'Leary et al., 2016) Release 215 (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-statistics/). However, recent studies have predicted more than six million species of fungi alone (Baldrian et al., 2021) which suggests that the total counts of eukaryotes greatly exceed previous estimations.

Despite current efforts, most microeukaryotes remain difficult to isolate and sequence. Further, the recovery of their genomes from metagenomes is limited compared to prokaryotes (West et al., 2018). Nevertheless, recent tools such as EukRep (West et al., 2018) and EukDetect (Lind & Pollard, 2021) aim to improve eukaryotic genome reconstruction from environmental metagenomes (Peng et al., 2021). However, EukRep only uses a subset of the 477 single-copy genes in their database to perform taxonomic classification, while EukDetect uses 214 marker genes. Taxonomic assignment of contigs and bins generated from metagenomic libraries can also be performed by the CAT/BAT tool (von Meijenfeldt et al., 2019). However, their homology search approach is time-consuming and requires extensive databases that are currently lacking a good representation of microbial eukaryote genomes (Pronk & Medema, 2022). Other tools, such as BUSCO (Waterhouse et al., 2017) and EukCC (Saary et al., 2020) are employed to measure the quality (completeness and contamination) of microeukaryotic metagenome-assembled genomes (MAGs). However, BUSCO only provides completeness measures and does not ascertain contamination. The reconstruction of eukaryotic genomes from metagenomes also faces additional challenges compared to prokaryotes. For example, eukaryotes are present in lower abundance when compared to prokaryotes (Lind & Pollard, 2021). To reconstruct less abundant species from metagenomes, increased sequencing depths are required. Additionally, the low number of reference genomes in databases used for taxonomy assignment limits our ability to obtain a realistic overview of eukaryote diversity (Pawlowski et al., 2012). Other challenges in the recovery of eukaryotic genomes include the existence of multiploidy and the share of repeat regions (Delmont & Eren, 2016).

Large-scale metagenomic studies usually do not include the reconstruction of microbial eukaryotes (Nayfach et al., 2019, 2020; Parks et al., 2017; Tully et al., 2018; Zhu et al., 2019), even in environments where they are key players such as in soil and host-associated biomes. This bias towards prokaryotes may lead to incorrect or incomplete assertions on the contribution of microbes to ecosystem processes. Also, the study of microeukaryotes is usually associated with human (Nguyen & Kalan, 2022; Parfrey et al., 2011), animal (Kittelmann et al., 2015), and plant (Sapp et al., 2018) hosts, as well as in marine environments (Chen et al., 2017; Santi et al., 2021) and constructed ecosystems (Zahedi et al., 2019) due to their integral role in ecosystem processes. For example, in plant-associated biomes, microbial eukaryotes influence nutrient uptake (Rodriguez Jr et al., 2009), while in human-associated microbial eukaryotes influence host immune system responses via the gut microbiome (Laforest-Lapointe & Arrieta, 2018). In aquatic and anthropogenic (e.g., wastewater treatment plants) biomes, microbial eukaryotes contribute to energy production (Matsubayashi et al., 2017; Trench-Fiol & Fink, 2020). Thus, studies including all domains of life and across all biomes would provide better insights into the role and effect of microbiomes in environmental and human health. Further, the inclusion of eukaryotes would also benefit studies that aim to catalogue, at the genome level, all of Earth's microbiomes (Nayfach et al., 2020). In this study, we aim to (1) assess our ability to recover eukaryotic genomes from environmental metagenomes, and (2) compare the quality of the best Eukaryotic MAGs from this study to reference genomes.

2 | MATERIALS AND METHODS

The complete workflow used in this study is shown in Figure 1.

2.1 | Metagenome data set

A total of 6000 curated metagenomes were collected from the Collaborative Multi-domain Exploration of Terrestrial metagenomes (CLUE-TERRA) consortium (https://www.ufz.de/index.php?en=47300). The first task of the curation process was to filter for true whole genome shotgun (WGS) libraries since non-metagenomic libraries in the Sequence Read Archive (SRA) can be wrongfully annotated as metagenomic. This was achieved by using PARTIE (Torres et al., 2017) with default parameters. Next, metagenomes with sequence quality scores below 70%, obtained via SRA-Tinder (https://github.com/NCBI-Hackathons/SRA_Tinder) using default parameters, were discarded. To maximize the comparability of the obtained metagenomes, only those sequenced using the Illumina sequencing platform and with a minimum of eight million paired-end reads per library were kept. Lastly, the consortium's focus on terrestrial environments excluded all libraries containing coordinates or terms for sea or ocean environments. However, our data set also includes transition environments such as rhizosphere and estuaries. Furthermore, given the dominance of eukaryotic organisms in the large-size fraction of metagenomes collected from the Tara Oceans project (Alexander et al., 2022) certain aquatic environments, such as lakes and rivers, were kept to improve our chances of recovering microbial eukaryotes. We used the definition of biomes and sample sources determined by Buttigieg and collaborators (Buttigieg et al., 2013). The exact definitions for terms used can be found in the Ontology

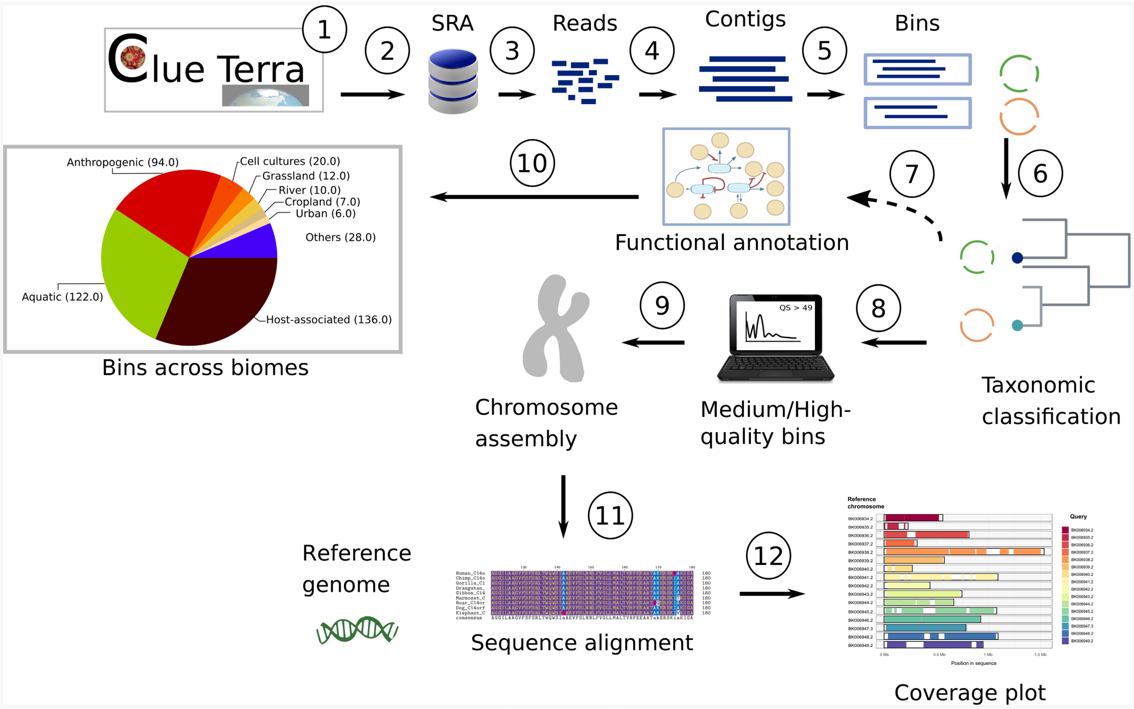

FIGURE 1 Workflow used in this study. Briefly, (1) samples were selected from the Terrestrial Metagenome Metadata Database (https://webapp.ufz.de/tmdb/), which is connected to the Collaborative multi-domain exploration of terrestrial metagenomes (CLUE-TERRA) consortium. Biomes and sample sources were defined based on the ENVO terms available at https://www.ebi.ac.uk/ols/index. (2) Next, the selected metagenomic libraries were downloaded from the Sequence Read Archive (SRA). (3) The sequencing reads were quality-controlled using metaWrap (Uritskiy et al., 2018) with default parameters. Furthermore, the reads were trimmed using TrimGalore (https://github.com/FelixKrueger/TrimGalore) with the default settings. Human host contamination was assessed and removed with bmtagger (Rotmistrovsky & Agarwala, 2011) using the human build 38 patch release 12 database (GRCh38.p12). (4) Contig assembly was performed using metaSpades (Nurk et al., 2017) with default parameters. (5) Binning of eukaryotic contigs was performed using CONCOCT (Alneberg et al., 2014) with default parameters. Bin quality was assessed using the EukCC (Saary et al., 2020) and BUSCO (Waterhouse et al., 2017) pipelines with default parameters. (6) Taxonomic classification of each generated bin was performed using taxator-tk (Dröge et al., 2015) using default parameters. (7) Genes were predicted using the GeneMark-ES model (Besemer et al., 2001) and annotated using MAKER2 (Holt & Yandell, 2011) with the RepBase gene database (Bao et al., 2015). Next, the gene sequences predicted by MAKER2 were submitted to GhostKOALA (Kanehisa et al., 2016) for functional annotation. (8) Medium and High-quality microeukaryote bins were assembled into chromosomes (9) using Chromosomer (Tamazian et al., 2016) with default parameters. (10) Microbial eukaryote genome recovery was also assessed according to biome. (11) The assembled chromosomes of the recovered microeukaryotic bins were aligned to the chromosomes of the reference genomes using Minimap2 (Li, 2018) with default parameters. (12) Lastly, the divergence rates were calculated based on the pairwise sequence alignments generated from Minimap2 using the pafr R package, with default parameters (https://rdrr.io/github/dwinter/pafr/).

Lookup Service (https://www.ebi.ac.uk/ols/index) provided by the European Molecular Biology Laboratory (EMBL).

2.2 | Preprocessing and library assembly

For each metagenomic library, the raw reads were quality-controlled using metaWrap (Uritskiy et al., 2018) with default parameters. Trimming of raw reads was performed using TrimGalore (https://github.com/FelixKrueger/TrimGalore) with the default settings. High-quality reads from each metagenomic library (using default Phred scores from TrimGalore) were aligned to potential host genomes using bmtagger (Rotmistrovsky & Agarwala, 2011) using the human build 38 patch release 12 database (GRCh38.p12). This alignment aims to remove human contamination and read pairs with only a single aligned read from the metagenomic libraries.

We used metaSpades (Nurk et al., 2017) to assemble the different samples using default parameters.

2.3 | Binning, taxonomic classification and quality assessment

Before binning, we used EukRep (West et al., 2018) to separate eukaryotic contigs from prokaryotic ones. Next, each eukaryotic assembly was binned using CONCOCT (Alneberg et al., 2014). Bins with size below 2 Mb were removed. Bin quality was assessed using the EukCC (Saary et al., 2020) and BUSCO (Waterhouse et al., 2017) pipelines. Taxonomy was assigned to each bin using taxator-tk (Dröge et al., 2015) using default parameters.

Coverage refers to the average number of reads aligned to known reference bases. Here, we calculated coverage by multiplying the number of mapped reads by the average length of reads in the libraries, then dividing by the size of the bins into base pairs (Equation 1).

$$\text{coverage} = \text{mapped reads} * \text{average read length} / \text{size of bin (bp)} \qquad (1)$$

The quality score of eukaryotic bins was assessed as determined by (Parks et al., 2017) (Equation 2) using the completeness and contamination values determined by EukCC (Saary et al., 2020).

$$\text{Quality score} = \text{completeness} - (5 * \text{contamination}) \qquad (2)$$

Eukaryotic bins were classified as high quality if their quality score was greater than 50 and presented a completeness value greater or equal to 60.

2.4 | Mapping of species-level, high-quality eukaryotic bins

To assess genome completeness, the high-quality eukaryotic bins (classified to species level) were assembled into chromosomes. This was achieved using Chromosomer (Tamazian et al., 2016) with default parameters. Next, the assembled chromosomes were aligned to the chromosomes of the reference genomes using Minimap2 (Li, 2018) with default parameters. The divergence rates were calculated based on the pairwise sequence alignments generated from Minimap2 using the pafr R package, with default parameters (https://rdrr.io/github/dwinter/pafr/) (Table 1).

2.5 | Gene prediction and functional annotation

Genes were predicted using the GeneMark-ES model (Besemer et al., 2001) and annotated using MAKER2 (Holt & Yandell, 2011) with the RepBase gene database (Bao et al., 2015). The functions of interest in this study are based on the work by Kieft and collaborators (Kieft et al., 2018) involving carbon and nitrogen cycling. Genes of the reference genomes *Bathycoccus prasinos* and *Micromonas commoda* involved in carbon fixation (Tables S1 and S2, respectively) and nitrogen metabolism (Tables S3 and S4, respectively) were extracted from Kyoto Encyclopedia of Genes and Genomes (KEGG) (release 100, 1 October 2021).

To demonstrate the potential contribution of eukaryotes to carbon fixation and nitrogen cycling, we selected the bins CTeuk-1331 (*B. prasinos*) and CTeuk-1332 (*M. commoda*) since they were recovered from the same metagenomic libraries used in Kieft and collaborators' study and presented the highest quality scores in their taxa. Next, we submitted the gene sequences predicted by MAKER2 to GhostKOALA (version 2.2) (Kanehisa et al., 2016) to determine their function and reconstruct KEGG pathways. The mapping of

TABLE 1 Gap and divergence rates of reconstructed eukaryotic bins to the reference chromosomes (number of gaps, average gap size and average per base divergence) and genomic information of the reference genomes (genome size, number of chromosomes and average chromosome size) (σ = standard deviation).

EukBin	Species	Genome size[a] (mb)	No. of chromosomes	Average chromosome size (mb)	No. of gaps[b]	Average gap size (bp)[c]	Average per base divergence[d]
CTeuk-1831	*Komagataella phaffii*	9.22	4	2.31	523	190.7266 (σ = 65.24)	0.0166
CTeuk-1331	*Bathycoccus prasinos*	14.96	19	0.79	2066	167.4468 (σ = 85.23)	0.024
CTeuk-1324	*Micromonas commoda*	20.97	17	1.23	1709	182.5044 (σ = 59)	0.151
CTeuk-1325	*Micromonas commoda*	20.97	17	1.23	1902	163.3701 (σ = 57.63)	0.154
CTeuk-1329	*Micromonas commoda*	20.97	17	1.23	1753	180.7028 (σ = 60.78)	0.154
CTeuk-1332	*Micromonas commoda*	20.97	17	1.23	1851	163.8736 (σ = 57.59)	0.152
CTeuk-1336	*Micromonas commoda*	20.97	17	1.23	1765	178.34 (σ = 60.91)	0.155
CTeuk-1341	*Micromonas commoda*	20.97	17	1.23	1696	181.01 (σ = 59.12)	0.156
CTeuk-1342	*Micromonas commoda*	20.97	17	1.23	1641	183.2572 (σ = 59.82)	0.155
CTeuk-1743	*Pichia kudriavzevii*	10.81	5	2.16	554	209.78 (σ = 766.18)	0.036
CTeuk-1741	*Saccharomyces cerevisiae*	12.07	16	0.75	452	179.8009 (σ = 54.20)	0.012
CTeuk-1822	*Saccharomyces cerevisiae*	12.07	16	0.75	607	204.3904 (σ = 33.31)	0.041
CTeuk-1829	*Saccharomyces cerevisiae*	12.07	16	0.75	1127	205.0852 (σ = 36.29)	0.033

[a] Size of the genome in megabases.

[b] Number of gaps in the genome.

[c] Average size of the gaps found in base pairs.

[d] Divergence rates of bases between the query and reference sequences.

functional orthologues (K numbers) to each gene was saved in tabular format.

3 | RESULTS

From the original 6000 metagenomes, only 215 yielded eukaryotic bins (from a total of 447 bins). The complete number of eukaryotic bins and sample attributes is shown in Tables S5 and S6. Almost 80% of all eukaryotic bins were obtained from host-associated (136), aquatic (122), and anthropogenic terrestrial (94) biomes (Table S6). Completeness and contamination measurements using EukCC (Saary et al., 2020) were only obtained for 193 bins. The average completeness and contamination were 44.64% (σ = 27.41) and 3.97 (σ = 6.53), respectively. Completeness measurements using BUSCO (Waterhouse et al., 2017) were only obtained for nine bins averaging 31.21% (σ = 37.24). Only five of the nine bins with BUSCO completeness measures also presented completeness values using EukCC. Differences in completeness values between the two pipelines ranged from 8.41% to 0.42%. The remaining four BUSCO completeness values without corresponding measures using EukCC were below ~20%. Due to BUSCO's low number of bins and average completeness values, only the results obtained with EukCC were used in further analyses.

A total of 153 eukaryotic bins were classified to family level (Table S6). Our data had a total of 51 medium/high-quality bins (Quality score ≥50) of which only 14 were classified to species level (spanning 5 unique taxa). The most frequent species-level taxonomy assigned to bins was *Micromonas commoda* (7) recovered from estuary samples. The second most frequent species-level assigned taxonomy was *Saccharomyces cerevisiae* (3) recovered from synthetic and fermentation metagenomes. Eukaryotic bins classified as *S. cerevisiae* also presented the highest genome coverage in the respective genomic libraries, ranging from ~29 to 192 times coverage. In contrast, *Bathycoccus prasinos*-classified bins showed only approximately six times coverage in their samples (Table S6). The frequencies of each taxon, at different levels as well as per biome are shown in Figure 2. The pairwise alignments of the species-level, medium/high-quality bins were reassembled into chromosomes and mapped to the chromosomes of the reference genomes as stated in the materials and methods.

The pairwise alignments for each reassembled eukaryotic bin are shown in Table S7. Assembled chromosomes with the highest divergences (per base differences between a query and target sequence) to the reference chromosomes were found in bins classified as *M. commoda* (average 0.154, σ = 0.012) (Figure 3a). In contrast, the assembled chromosomes of bins classified as *S. cerevisiae* showed the lowest divergences compared to the reference chromosomes (average 0.029, σ = 0.017) (Figure 3b). The complete set of results of divergences between assembled and reference chromosomes is shown in the Table S8. Additionally, the mapping of the bins' chromosomes to the reference genomes is shown in Figures S1–S13.

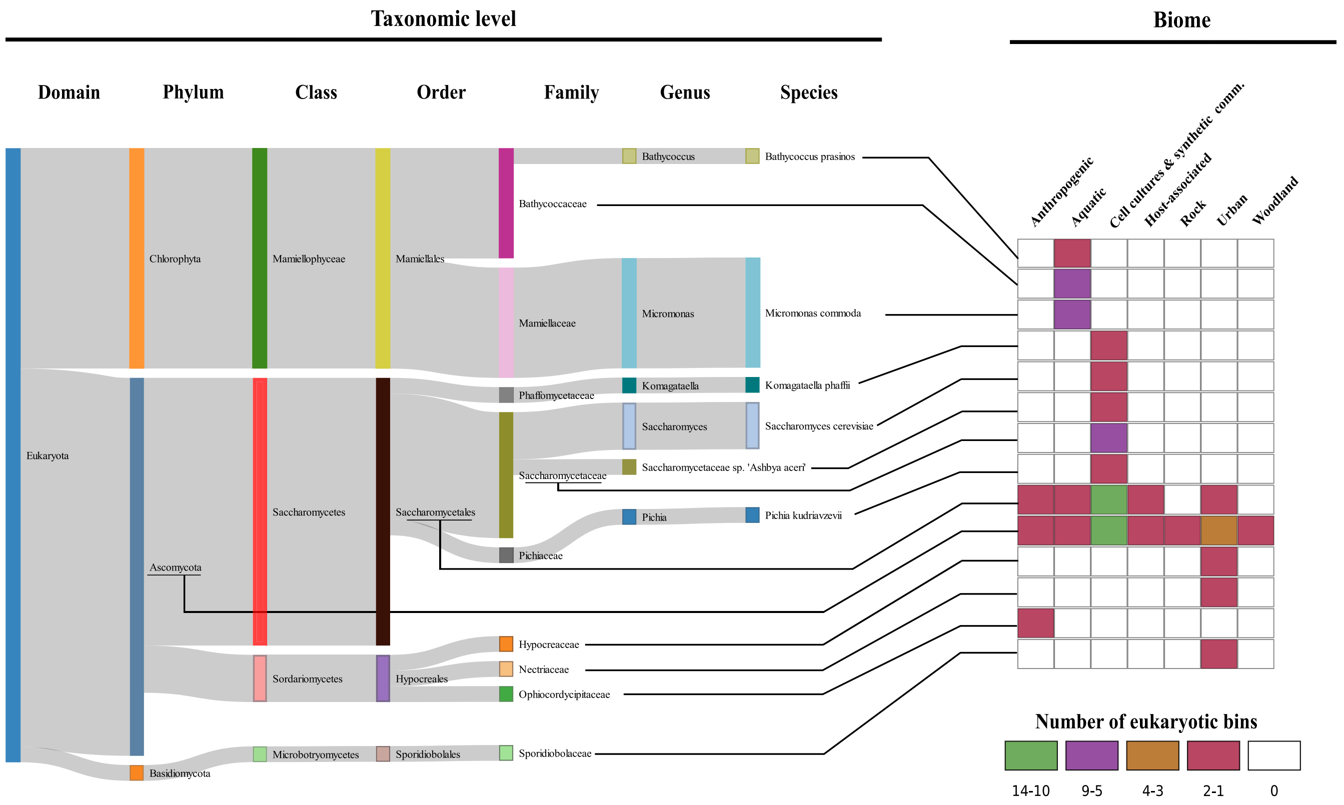

FIGURE 2 Sankey plot showing the taxonomic distribution of the recovered Eukaryotic bins and heatmap showing the number of eukaryotic bins recovered per Biome (retrieved from https://webapp.ufz.de/tmdb/ and manually curated based on the sample data).

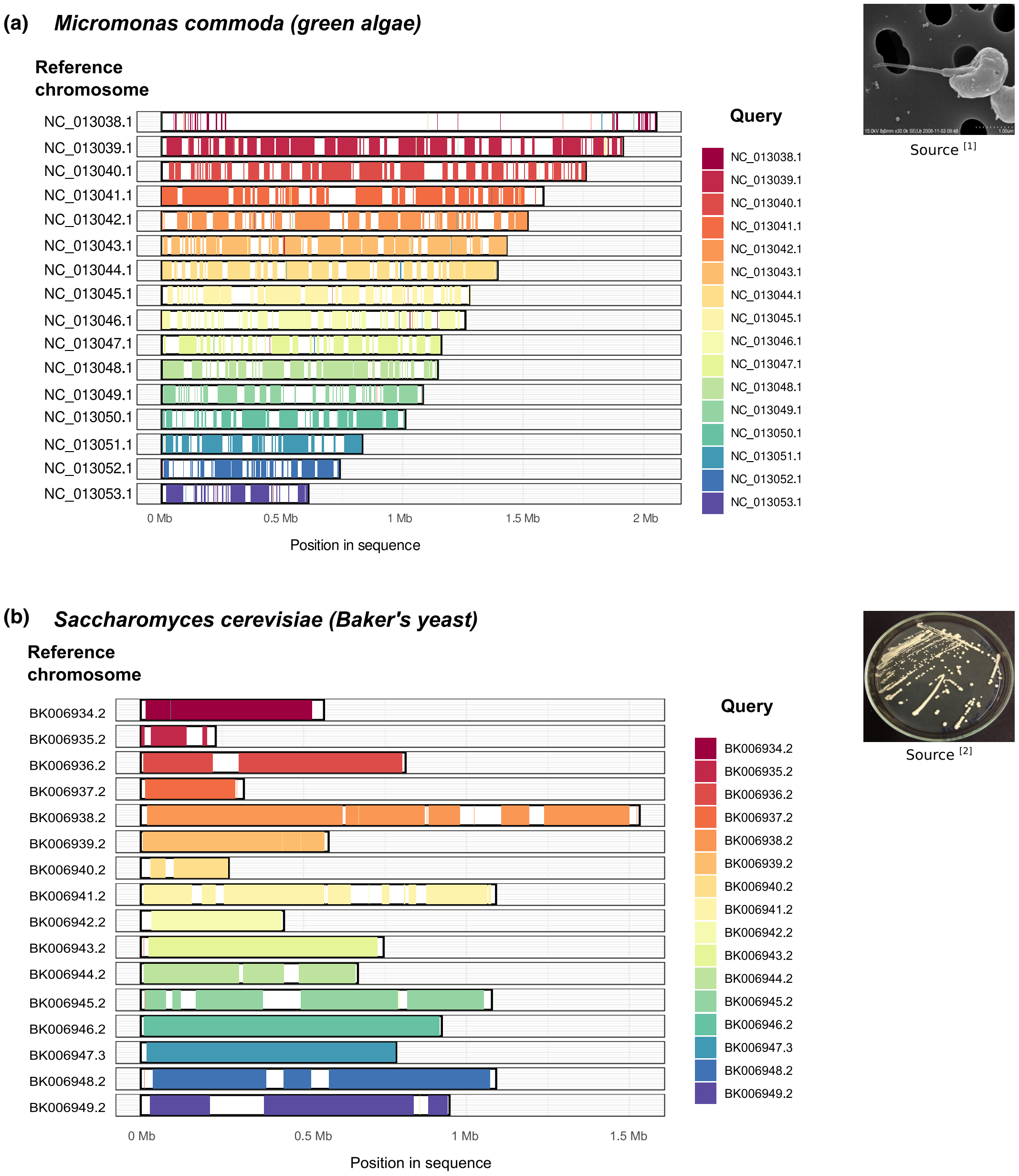

FIGURE 3 Mapping of assembled chromosomes for a eukaryotic bin (query) to the chromosomes of the reference genome. (a) *Micromonas commoda* (CTeuk-1336). (b) *Saccharomyces cerevisiae* (CTeuk-1741). [1] Vaulot D. et al., 2004, The Roscoff Culture Collection (RCC): a collection dedicated to marine picoplankton. Nova Hedwigia 79:49–70; [2] https://commons.wikimedia.org/wiki/File:Saccharomyces_cerevisiae_YGC_colonies_50.jpg

Annotation of eukaryotic bins yielded, on average, 4106, 4435, 4573, 4619, and 4150 protein-encoding genes in *S. cerevisiae*, *Komagataella phaffii*, *Pichia kudriavzevii*, *M. commoda* and *B. prasinos*, respectively. However, the predicted genes in *M. commoda* and *B. prasinos* bins only accounted for 45.57% and 52.54% of their reference genomes, respectively (Table S9).

Functional annotation of CTeuk-1331 (*B. prasinos*) revealed the presence of 10 genes involved in nitrogen metabolism and 32 genes involved in carbon fixation (Table S10). Functional annotation of CTeuk-1332 (*M. commoda*), revealed the presence of two genes involved in nitrogen metabolism and 34 genes in carbon fixation (Table S11).

Annotation of the species-level, high-quality eukaryotic bins is available in Table S12.

4 | DISCUSSION

The recovery and quality assessment of eukaryotic bins from metagenomes involve a sequence of computational steps, including the major steps of read assembly, contig binning, and gene prediction. Compared to prokaryotes, eukaryotes generally have larger genomes and a more complex gene structure (Keeling, 2019). Our results demonstrate that, despite current efforts, our ability to reconstruct high-quality Eukaryotic genomes from metagenomes is in its early stages of development. Both quality and taxonomic assignments of the metagenome-assembled eukaryotic bins are substantially lower when compared to prokaryotes. There are multiple possible reasons: First, larger eukaryotic genomes require more sequencing reads (i.e., higher sequencing depth) to obtain genome coverage (Keeling, 2019). Second, multiploidy can cause a problem for read assemblers as very similar but not identical contigs can be generated from more than one allele with various similarities (Zhang et al., 2020). Thus, multiploidy, together with eukaryotes' generally larger genome size, requires a higher sequencing depth to assemble high-quality genomes. This is supported by the observation that differences in genome recovery seem to be highly linked to genome size (Alexander et al., 2022).

The quality of reconstructed genomes from metagenomes is usually calculated by the presence and number of single-copy genes in a bin (Saary et al., 2020), which are compared to known reference genomes. The low number of reference genomes compared to prokaryotes thus also influences the quality measures (Saary et al., 2020; Waterhouse et al., 2017). Furthermore, to detect single-copy genes from bins, contigs need to be assembled and binned and genes called. The detection of single-copy genes thus relies on the quality of assembly and the quality of the subsequent calling of genes. The gene structure, that is, exon-intron sequences, in eukaryotic genomes interfere with gene calling (Roy & Penny, 2007), and gene calling is difficult due to frequent gene or genome duplication events (Kaltenegger et al., 2018). Furthermore, intron presence and number vary across eukaryotic species, making it harder to predict genes in some species accurately. For example, *Aspergillus fumigatus* has 18,293 introns compared to the 266 found in *Saccharomyces cerevisiae* (Roy & Penny, 2007).

The two genome quality assessment tools we chose for this study use single-copy genes to estimate the quality of eukaryotic genomes. However, the single-copy gene sets used by each tool differ in composition and application (e.g., BUSCO requires the user to define which sets of single-copy genes to use). The more unique and nonrepeated single-copy genes in a bin, the higher the quality determined by either tool. However, even in high-quality bins such as CTeuk-1741 (95% completeness and 0.34 contamination) (classified as *S. cerevisiae*), significant parts of each chromosome can be missing or contain misplaced reads (Figure S1), which may also be connected to its high number of chromosomes (16). The genome assembly of *P. kudriavzevii* composed of five chromosomes revealed fewer missing or misplaced reads (Figure S12). Chromosome assembly is challenging due to repeat regions, especially in the telomeres. For instance, the human reference genome is the most accurate vertebrate genome but still lacks the characterization of some chromosomes (Miga et al., 2020; Nurk et al., 2022). A study by Wang and collaborators (Wang et al., 2021), proposed a strategy for the complete assembly of two ciliates. Both studies suggest that high coverage and ultra-long nanopore sequencing may yield a better assembly of genomes.

Most genome annotations studies use short-reads and quality assessment tools such as BUSCO (Waterhouse et al., 2017) and EukCC (Saary et al., 2020). Short reads usually have high quality but also have several drawbacks concerning the assembly process. First, good coverage is necessary to assemble long contigs. Second, the presence of sequence repeats in the genome cannot be solved using short-read sequencing technologies (De Bustos et al., 2016). Nevertheless, even with long-read sequencing, repeat-rich genomes yielded more fragmented assemblies (Sevim et al., 2019). While tools have been developed specially for improved scaffolding of large, repeat-rich eukaryotic genomes (Gao et al., 2016; Miga et al., 2020; Wang et al., 2021), several long-read metagenomic data sets already exist (Corrêa et al., 2020; Kasmanas et al., 2021; Nata'ala et al., 2022). Since short reads tend to assemble into shorter contigs compared to long reads, short-read sequencing can influence the accuracy of gene predictions by not covering a gene's total length (Pearman et al., 2020), and single-copy genes may not provide a realistic measure of the completeness of complex organisms such as eukaryotes.

In this study, all metagenomes were sequenced using short reads, which might explain the low number of species-level classifications (five unique taxa) in high-quality eukaryotic bins. The pairwise alignments of the reassembled eukaryotic bins to their respective reference genomes revealed that it is possible to reconstruct a large amount of the genome using short-read sequencing when a high number of reference genomes exist (e.g., *Saccharomyces cerevisiae*). We also observed a similar relation between assemblies and reference genomes when calculating divergence rates. Genome reconstruction exhibited higher divergence rates in species with fewer reference genomes, such as *M. commoda*. Our data showed many gaps when we mapped the eukaryotic bins to the reference chromosomes, which may be linked to intron presence. Introns may also play a role in accurately predicting genes, as shown by the low number of predicted genes in *B. prasinos* and *M. commoda* eukaryotic bins (Table S4). Thus, new sequencing technologies that provide longer continuous sequences (e.g., Oxford Nanopore or PacBio sequencing) might be necessary to facilitate the recovery of high-quality Eukaryotic genomes from metagenomes (Amarasinghe et al., 2020). More work needs to be done to experimentally combine the advantages of long and short-read sequencing and develop tools that can handle both for improved assembly into long contigs. Additional improvements at the technical (e.g., improving sampling and DNA extraction kits and the use of long reads), molecular (e.g., enrichment techniques combined with molecular tagging or pulldown or hybridization, which would decrease sample complexity), and bioinformatic level (e.g., inclusion of eukaryotic taxa/branch specific detection pipelines) are also recommended.

The low taxonomic diversity is evident in Figure 2, where we exhibit the frequency counts of taxa across all libraries containing medium to high-quality eukaryotic bins. Given that current metagenomic methods rely on comparisons to known genomes, a higher number of eukaryotic reference genomes will improve eukaryotic species identification from metagenomes (e.g., Earth BioGenome Project, Human Microbiome Project, and Tara Oceans Project) (Loeffler et al., 2020). As expected, the highest number of microeukaryotes were obtained in host-associated, aquatic, and anthropogenic biomes since most studies involving microeukaryotes focus on their parasitic life form and their influence on the host rather than as free-living microbes (Aslani et al., 2022; Chen et al., 2017; Nguyen & Kalan, 2022; Parfrey et al., 2011; Santi et al., 2021; Sapp et al., 2018).

In terms of metagenomic studies, we recommend including results from eukaryotic genome recovery to avoid missing potential key players in ecosystem processes, especially because of the increased focus on interactions between microbial eukaryotes and bacteria/archaea (De Gruyter et al., 2020; Piwosz et al., 2020). Our results revealed the presence of *B. prasinos* and *M. commoda* in the metagenomes collected from estuaries (BioProject PRJNA320136). Collado-Fabbri et al. (2011) showed that *M. commoda* and *B. prasinos* had varying degrees of contribution to picophytoeukaryotic carbon biomass in upwelling ecosystems (such as estuaries) depending on the season. The carbon provided by *B. prasinos* benefits the growth rates of *M. commoda* since it increases the supply rates of ammonia to the nitrogen assimilation pathway (Cuvelier et al., 2017). Functional annotation of the *B. prasinos* and *M. commoda* bins revealed the presence of multiple genes involved in carbon fixation and nitrogen metabolism. However, *nii* genes, responsible for converting nitrite to ammonia, were missing, unlike the reference genomes. The miss-annotation of genes present in *B. prasinos* and *M. commoda* highlights the challenge of reconstructing near-complete eukaryotic genomes due to insufficient reference genomes in genome repositories. Nevertheless, including the reconstruction of eukaryotic genomes to studies involving carbon and nitrogen cycling in aquatic environments may provide a more complete picture of carbon and nitrogen cycling (Kieft et al., 2018).

A more recent study by Alexander et al. (2022) developed a new tool for eukaryotic metagenome-assembled genome recovery (EukHeist) and tested it using data from Tara Oceans (Carradec et al., 2018). EukHeist performs similarly to EukRep (West et al., 2018) in that it attempts first to separate prokaryotic from eukaryotic contigs. However, the assembly using EukHeist requires coassemblies in contrast to EukRep, which does not. The workflow employed in the study by Alexander and colleagues is similar to ours, which comprises assembly, binning, quality control, filtering, and taxonomic annotation. Their study recovered 485 eukaryotic from 94 coassemblies with a BUSCO completeness score of at least 30%. We recovered a total of 121 eukaryotic bins in our study with a minimum completeness value of 30% (using EukCC) applied. The difference in the number of recovered eukaryotic bins between both studies can be derived from the co-assembly strategy employed by Alexander and colleagues and the reported dominance of eukaryotic organisms in the Tara Oceans samples (Alexander et al., 2022). Future research should consider applying the approach proposed by Alexander and colleagues to other ecosystems that are not dominated by eukaryotes, but this was outside the objectives of this study.

To help other researchers attempt the recovery of microeukaryotes from metagenomes, we came up with recommendations and points of action to improve microeukaryote genome recovery. First, sampling strategies and extraction of high-quality DNA need improvement. Second, community efforts are needed to generate more reference genomes. Third, experimental design should consider the use of short- and long-read sequencing technology. Further, tools should be developed and improved to integrate short and long reads. In addition, alternative genome quality measures such as chromosome coverage should be considered when determining high-quality metagenome-assembled genomes.

5 | CONCLUSIONS

Our study demonstrates that performing single-domain genome reconstruction from environmental metagenomes leads to an incomplete overview of microbial communities' diversity and functional potential. To obtain accurate representations of all species present in an ecosystem, substantial efforts in tool development to identify species in all domains are still required. Eukaryotes play vital roles in ecosystems ranging from complementing the activities of other microbes to performing phototrophic and saprotrophic processes and predation (del Campo et al., 2020). Despite their importance, very few metagenomic studies attempted to reconstruct and annotate eukaryote genomes due to major methodological limitations. Increasing the number and quality of reference genomes in public databases and developing tools for intron identification may result in better genome reconstructions. Additionally, removing the identified introns may also improve gene predictions. A possible avenue to achieve this goal is to promote long-read sequencing technologies. While we did reconstruct almost 447 eukaryotic bins, only 14 were of high quality and classified to species level. Still, the identified species showed promise in adding layers of information to the original studies. Thus, reconstructing more high-quality bins will bring us closer to a more realistic overview and understanding of biodiversity and how Eukaryotes contribute to different ecosystem processes.

AUTHOR CONTRIBUTIONS

Joao Pedro Saraiva and Ulisses Nunes da Rocha developed the concept of the study. Ulisses Nunes da Rocha was the main supervisor of the study. Rodolfo Brizola Toscan downloaded all sequencing data. Rodolfo Brizola Toscan generated all data for analysis. Alexander Bartholomäus and Joao Pedro Saraiva performed all data analysis. Joao Pedro Saraiva, Alexander Bartholomäus, and Ulisses Nunes da Rocha wrote the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGEMENTS

We would like to thank Felipe Borim Correa for his advice during the selection of metagenomes. We would also like to thank Stefania Magnusdottir for her review of the manuscript and all members of the CLUE-TERRA consortium (https://www.ufz.de/index.php?en=47300) for their advice on the reconstruction of eukaryotic genomes from metagenomes. Open Access funding enabled and organized by Projekt DEAL

FUNDING INFORMATION

This work was funded by the Helmholtz Young Investigator grant VH-NG-1248 Micro "Big Data" and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project no. 460129525.

DATA AVAILABILITY STATEMENT

The metagenome-assembled genomes (MAGs) obtained in this study are available at the National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) with the BioProject accession PRJNA810309. The MAGs are available under the sample accessions SAMN26244030-SAMN26244039, SAMN26244052-SAMN26244057, SAMN26244171-SAMN26244173, SAMN26302835-SAMN26302918, SAMN26302921-SAMN26302929, SAMN26302933-SAMN26302978, SAMN26302997-SAMN26303001, SAMN26303005-SAMN26303043, SAMN26303045, SAMN26303049, SAMN26303053-SAMN26303054, SAMN26303056, SAMN26303058, SAMN26303060, SAMN26303062-SAMN26303065, SAMN26329017-SAMN26329047, SAMN26329100-SAMN26329115, SAMN26329126-SAMN26329131, SAMN26329141-SAMN26329143, SAMN26329147-SAMN26329313, SAMN26329315-SAMN26329336. The assemblies of the reference genomes used to perform pairwise alignments have been made available at the National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) under the accession identifiers GCA_000146045.2, GCA_003054445.1, GCA_000090985.2, GCA_000027005.1, and GCA_002220235.1.

BENEFIT-SHARING STATEMENT

Benefits generated: Benefits from this research accrue from the sharing of our data and results on public databases as described. Additionally, the results from this research will help guide future work in the design and execution of genome-centric studies by fostering a multi-domain approach.

ORCID

Alexander Bartholomäus https://orcid.org/0000-0003-0970-7304
Petr Baldrian https://orcid.org/0000-0002-8983-2721
Ulisses Nunes da Rocha https://orcid.org/0000-0001-6972-6692

REFERENCES

Alexander, H., Hu, S. K., Krinos, A. I., Pachiadaki, M., Tully, B. J., Neely, C. J., & Reiter, T. (2022). *Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton*. bioRxiv. https://doi.org/10.1101/2021.07.25.453713

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Aslani, F., Geisen, S., Ning, D., Tedersoo, L., & Bahram, M. (2022). Towards revealing the global diversity and community assembly of soil eukaryotes. *Ecology Letters*, *25*(1), 65–76. https://doi.org/10.1111/ele.13904

Baldrian, P., Kolařík, M., Stursová, M., Kopecký, J., Valášková, V., Větrovský, T., Zifčáková, L., Snajdr, J., Rídl, J., Vlček, C., & Voříšková, J. (2012). Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME Journal*, *6*(2), 248–258. https://doi.org/10.1038/ismej.2011.95

Baldrian, P., Větrovský, T., Lepinay, C., & Kohout, P. (2021). High-throughput sequencing view on the magnitude of global fungal diversity. *Fungal Diversity*, *114*, 539–547. https://doi.org/10.1007/s13225-021-00472-y

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, *29*(12), 2607–2618.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., & Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution*, *27*(4), 233–243. https://doi.org/10.1016/j.tree.2011.11.010

Bulan, D. E., Wilantho, A., Tongsima, S., Viyakarn, V., Chavanich, S., & Somboonna, N. (2018). Microbial and small eukaryotes associated with reefs in the upper gulf of Thailand. *Frontiers in Marine Science*, *5*, 436. https://doi.org/10.3389/fmars.2018.00436

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013). The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, *4*, 43. https://doi.org/10.1186/2041-1480-4-43

Carradec, Q., Pelletier, E., da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M. A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Chen, W., Pan, Y., Yu, L., Yang, J., & Zhang, W. (2017). Patterns and processes in marine microeukaryotic community biogeography from Xiamen coastal waters and intertidal sediments, Southeast China. *Frontiers in Microbiology*, *8*, 1912. https://doi.org/10.3389/fmicb.2017.01912

Collado-Fabbri, S., Vaulot, D., & Ulloa, O. (2011). Structure and seasonal dynamics of the eukaryotic picophytoplankton community in a wind-driven coastal upwelling ecosystem. *Limnology and Oceanography*, *56*(6), 2334–2346. https://doi.org/10.4319/lo.2011.56.6.2334

Corrêa, F. B., Saraiva, J. P., Stadler, P. F., & da Rocha, U. N. (2020). TerrestrialMetagenomeDB: A public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Research*, *48*(D1), D626–D632. https://doi.org/10.1093/nar/gkz994

Cuvelier, M. L., Guo, J., Ortiz, A. C., van Baren, M. J., Tariq, M. A., Partensky, F., & Worden, A. Z. (2017). Responses of the picoprasinophyte Micromonas commoda to light and ultraviolet stress. *PLoS One*, *12*(3), e0172135. https://doi.org/10.1371/journal.pone.0172135

De Bustos, A., Cuadrado, A., & Jouve, N. (2016). Sequencing of long stretches of repetitive DNA. *Scientific Reports*, *6*(1), 36665. https://doi.org/10.1038/srep36665

de Gruyter, J., Weedon, J. T., Bazot, S., Dauwe, S., Fernandez-Garberí, P. R., Geisen, S., de la Motte, L. G., Heinesch, B., Janssens, I. A., Leblans, N., Manise, T., Ogaya, R., Löfvenius, M. O., Peñuelas, J., Sigurdsson, B. D., Vincent, G., & Verbruggen, E. (2020). Patterns of local, intercontinental and interseasonal variation of soil bacterial and eukaryotic microbial communities. *FEMS Microbiology Ecology*, *96*(3), fiaa018. https://doi.org/10.1093/femsec/fiaa018

del Campo, J., Bass, D., & Keeling, P. J. (2020). The eukaryome: Diversity and role of microeukaryotic organisms associated with animal hosts. *Functional Ecology*, *34*(10), 2045–2054. https://doi.org/10.1111/1365-2435.13490

Delmont, T. O., & Eren, A. M. (2016). Identifying contamination with advanced visualization and analysis practices: Metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, *4*, e1839. https://doi.org/10.7717/peerj.1839

Dröge, J., Gregor, I., & McHardy, A. C. (2015). Taxator-tk: Precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics (Oxford, England)*, *31*(6), 817–824. https://doi.org/10.1093/bioinformatics/btu745

Gao, S., Bertrand, D., Chia, B. K. H., & Nagarajan, N. (2016). OPERA-LG: Efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. *Genome Biology*, *17*(1), 102. https://doi.org/10.1186/s13059-016-0951-y

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. https://doi.org/10.1186/1471-2105-12-491

Kaltenegger, E., Leng, S., & Heyl, A. (2018). The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway. *BMC Evolutionary Biology*, *18*(1), 76. https://doi.org/10.1186/s12862-018-1153-x

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, *428*(4), 726–731. https://doi.org/10.1016/j.jmb.2015.11.006

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., von Bergen, M., Stadler, P. F., Carvalho, A. C. P. L. F., & Nunes da Rocha, U. (2021). HumanMetagenomeDB: A public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research*, *49*(D1), D743–D750. https://doi.org/10.1093/nar/gkaa1031

Keeling, P. J. (2019). Combining morphology, behaviour and genomics to understand the evolution and ecology of microbial eukaryotes. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *374*(1786), 20190085. https://doi.org/10.1098/rstb.2019.0085

Kieft, B., Li, Z., Bryson, S., Crump, B. C., Hettich, R., Pan, C., Mayali, X., & Mueller, R. S. (2018). Microbial community structure–function relationships in Yaquina Bay estuary reveal spatially distinct carbon and nitrogen cycling capacities. *Frontiers in Microbiology*, *9*, 1282. https://doi.org/10.3389/fmicb.2018.01282

Kittelmann, S., Devente, S. R., Kirk, M. R., Seedorf, H., Dehority, B. A., & Janssen, P. H. (2015). Phylogeny of intestinal ciliates, including Charonina ventriculi, and comparison of microscopy and 18S rRNA gene pyrosequencing for rumen ciliate community structure analysis. *Applied and Environmental Microbiology*, *81*(7), 2433–2444. https://doi.org/10.1128/AEM.03697-14

Laforest-Lapointe, I., & Arrieta, M. C. (2018). Microbial eukaryotes: A missing link in gut microbiome studies. *MSystems*, *3*(2), e00201–e00217. https://doi.org/10.1128/mSystems.00201-17

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lind, A. L., & Pollard, K. S. (2021). Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome*, *9*(1), 58. https://doi.org/10.1186/s40168-021-01015-y

Loeffler, C., Karlsberg, A., Martin, L. S., Eskin, E., Koslicki, D., & Mangul, S. (2020). Improving the usability and comprehensiveness of microbial databases. *BMC Biology*, *18*, 37. https://doi.org/10.1186/s12915-020-0756-z

Matsubayashi, M., Shimada, Y., Li, Y. Y., Harada, H., & Kubota, K. (2017). Phylogenetic diversity and in situ detection of eukaryotes in anaerobic sludge digesters. *PLoS One*, *12*(3), e0172888. https://doi.org/10.1371/journal.pone.0172888

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Nata'ala, M., Avila Santos, A. P., Coelho Kasmanas, J., Bartholomäus, A., Saraiva, J. P., Godinho Silva, S., Keller-Costa, T., Costa, R., Gomes, N. C. M., Ponce de Leon Ferreira de Carvalho, A. C., Stadler, P. F., Sipoli Sanches, D., & Nunes da Rocha, U. (2022). MarineMetagenomeDB: A public repository for curated and standardized metadata for marine metagenomes. *Environmental Microbiome*, *17*(1), 57. https://doi.org/10.1186/s40793-022-00449-7

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I. M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2020). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *1–11*, 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510. https://doi.org/10.1038/s41586-019-1058-x

Nguyen, U. T., & Kalan, L. R. (2022). Forgotten fungi: The importance of the skin mycobiome. *Current Opinion in Microbiology*, *70*, 102235. https://doi.org/10.1016/j.mib.2022.102235

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Parfrey, L. W., Walters, W. A., & Knight, R. (2011). Microbial eukaryotes in the human microbiome: Ecology, evolution, and future directions. *Frontiers in Microbiology*, *2*, 153. https://doi.org/10.3389/fmicb.2011.00153

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., … de Vargas, C. (2012). CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, *10*(11), e1001419. https://doi.org/10.1371/journal.pbio.1001419

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long-read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*, *21*(1), 220. https://doi.org/10.1186/s12859-020-3528-4

Peng, X., Wilken, S. E., Lankiewicz, T. S., Gilmore, S. P., Brown, J. L., Henske, J. K., Swift, C. L., Salamov, A., Barry, K., Grigoriev, I. V., Theodorou, M. K., Valentine, D. L., & O'Malley, M. A. (2021). Genomic and functional analyses of fungal and bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes. *Nature Microbiology*, *6*(4), 499–511. https://doi.org/10.1038/s41564-020-00861-0

Piwosz, K., Shabarova, T., Pernthaler, J., Posch, T., Šimek, K., Porcal, P., & Salcher, M. M. (2020). Bacterial and eukaryotic small-subunit amplicon data do not provide a quantitative picture of microbial communities, but they are reliable in the context of ecological interpretations. *MSphere*, *5*(2), e00052-20. https://doi.org/10.1128/mSphere.00052-20

Pronk, L. J. U., & Medema, M. H. (2022). Whokaryote: Distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microbiology Society*, *8*, mgen000823. https://doi.org/10.1099/mgen.0.000823

Rodriguez, R. J., Jr., White, J. F., Jr., Arnold, A. E., & Redman, R. S. (2009). Fungal endophytes: Diversity and functional roles. *New Phytologist*, *182*(2), 314–330.

Rotmistrovsky, K., & Agarwala, R. (2011). *BMTagger: Best Match Tagger for removing human reads from metagenomics datasets*. Ftp://Ftp.Ncbi.Nlm.Nih.Gov/Pub/Agarwala/Bmtagger/

Roy, S. W., & Penny, D. (2007). Intron length distributions and gene prediction. *Nucleic Acids Research*, *35*(14), 4737–4742. https://doi.org/10.1093/nar/gkm281

Saary, P., Mitchell, A. L., & Finn, R. D. (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biology*, *21*(1), 244. https://doi.org/10.1186/s13059-020-02155-4

Santi, I., Kasapidis, P., Karakassis, I., & Pitta, P. (2021). A comparison of DNA metabarcoding and microscopy methodologies for the study of aquatic microbial eukaryotes. *Diversity*, *13*(5), 180. https://doi.org/10.3390/d13050180

Sapp, M., Ploch, S., Fiore-Donno, A. M., Bonkowski, M., & Rose, L. E. (2018). Protists are an integral part of the Arabidopsis thaliana microbiome. *Environmental Microbiology*, *20*(1), 30–43. https://doi.org/10.1111/1462-2920.13941

Sevim, V., Lee, J., Egan, R., Clum, A., Hundley, H., Lee, J., Everroad, R. C., Detweiler, A. M., Bebout, B. M., Pett-Ridge, J., Göker, M., Murray, A. E., Lindemann, S. R., Klenk, H. P., O'Malley, R., Zane, M., Cheng, J. F., Copeland, A., Daum, C., … Woyke, T. (2019). Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Scientific Data*, *6*, 285. https://doi.org/10.1038/s41597-019-0287-z

Sweetlove, L. (2011). Number of species on earth tagged at 8.7 million. *Nature*, *23*. https://doi.org/10.1038/news.2011.498

Tamazian, G., Dobrynin, P., Krasheninnikova, K., Komissarov, A., Koepfli, K. P., & O'Brien, S. J. (2016). Chromosomer: A reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience*, *5*(1), 38. https://doi.org/10.1186/s13742-016-0141-6

Torres, P. J., Edwards, R. A., & McNair, K. A. (2017). PARTIE: A partition engine to separate metagenomic and amplicon projects in the sequence read archive. *Bioinformatics*, *33*(15), 2389–2391. https://doi.org/10.1093/bioinformatics/btx184

Trench-Fiol, S., & Fink, P. (2020). Metatranscriptomics from a small aquatic system: Microeukaryotic community functions through the diurnal cycle. *Frontiers in Microbiology*, *11*, 1006. https://doi.org/10.3389/fmicb.2020.01006

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*, 170203. https://doi.org/10.1038/sdata.2017.203

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—A flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1

von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H., & Dutilh, B. E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, *20*(1), 217. https://doi.org/10.1186/s13059-019-1817-x

Wang, G., Wang, S., Chai, X., Zhang, J., Yang, W., Jiang, C., Chen, K., Miao, W., & Xiong, J. (2021). A strategy for complete telomere-to-telomere assembly of ciliate macronuclear genome using ultra-high coverage nanopore data. *Computational and Structural Biotechnology Journal*, *19*, 1928–1932. https://doi.org/10.1016/j.csbj.2021.04.007

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*, 543–548. https://doi.org/10.1093/molbev/msx319

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*(4), 569–580. https://doi.org/10.1101/gr.228429.117

Zahedi, A., Greay, T. L., Paparini, A., Linge, K. L., Joll, C. A., & Ryan, U. M. (2019). Identification of eukaryotic microorganisms with 18S rRNA next-generation sequencing in wastewater treatment plants, with a more targeted NGS approach required for cryptosporidium detection. *Water Research*, *158*, 301–312. https://doi.org/10.1016/j.watres.2019.04.041

Zhang, L., Zhou, X., Weng, Z., & Sidow, A. (2020). De novo diploid genome assembly for genome-wide structural variant detection. *NAR Genomics and Bioinformatics*, *2*(1), lqz018. https://doi.org/10.1093/nargab/lqz018

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., al-Ghalith, G. A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J. B., Huang, S., Salam, N., Jiao, J. Y., Wu, Z., Xu, Z. Z., Cantrell, K., Yang, Y., … Knight, R. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nature Communications*, *10*(1), 5477. https://doi.org/10.1038/s41467-019-13443-4

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Saraiva, J. P., Bartholomäus, A., Toscan, R. B., Baldrian, P., & Nunes da Rocha, U. (2023). Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes. *Molecular Ecology Resources*, *23*, 1066–1076. https://doi.org/10.1111/1755-0998.13776