Recovery of 447 Eukaryotic bins reveals major challenges for Eukaryote genome reconstruction from 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, Saxony, 04318, Germany

[2] GFZ German Research Centre for Geosciences, Section 3.7 Geomicrobiology, Telegrafenberg, Potsdam, 14473, Germany

[3] Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Videnska 1083, Praha 4, 14220, Czech Republic

Joao Pedro Saraiva: joao.saraiva@ufz.de

Alexander Bartholomäus: abartho@gfz-potsdam.de

Rodolfo Brizola Toscan: rodolfo.toscan@ufz.de

Petr Baldrian: baldrian@biomed.cas.cz

*Ulisses Nunes da Rocha: ulisses.rocha@ufz.de

* Corresponding author

Abstract

An estimated 8.7 million eukaryotic species exist on our planet. However, recent tools for taxonomic classification of eukaryotes only dispose of 734 reference genomes. As most Eukaryotic genomes are yet to be sequenced, the mechanisms underlying their contribution to different ecosystem processes remain untapped. Although approaches to recover 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 Eukaryotic genomes using 215 metagenomes from diverse environments using the EukRep pipeline. We obtained 447 eukaryotic bins from 15 classes (e.g., Saccharomycetes, Sordariomycetes, and Mamiellophyceae) and 16 orders (e.g., Mamiellales, Saccharomycetales, and Hypocreales). More than 73% of the obtained eukaryotic bins were recovered from samples whose biomes were classified as host-associated, aquatic and anthropogenic terrestrial. However, only 93 bins showed taxonomic classification to (9 unique) genera and 17 bins to (6 unique) species. A total of 193 bins contained completeness and contamination measures. Average completeness and contamination were 44.64% (σ=27.41%) and 3.97% (σ=6.53%), respectively. *Micromonas commoda* was the most frequent taxa found while *Saccharomyces cerevisia*e presented the highest completeness, possibly resulting from a more significant number of reference genomes. However, mapping eukaryotic bins to the chromosomes of the reference genomes suggests that completeness measures should consider both single-copy genes and chromosome coverage. Recovering eukaryotic genomes will benefit significantly from long-read sequencing, intron removal after assembly, and improved reference genomes databases.

Keywords: Eukaryotes, genome-resolved metagenomics, Saccharomycetales, Mamiellales, Hypocreales

Introduction

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, Probst, Grigoriev, Thomas, & Banfield, 2018). An estimated 8.7 million eukaryotic species inhabit our planet (Sweetlove, 2011), but as of 4th January 2022, only slightly more than 54000 eukaryotic reference genomes exist in RefSeq (O'Leary et al., 2016). However, recent studies have predicted more than six million species of fungi alone (Baldrian, Větrovský, Lepinay, & Kohout, 2021) which suggests that the total counts of eukaryotes greatly exceed previous estimations.

Despite current efforts, the ability to recover eukaryotic genomes is limited compared to prokaryotic genome recovery (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 natural environments (Peng et al., 2021). However, EukRep only uses 734 to perform taxonomic classification. Further, tools such as BUSCO (Waterhouse et al., 2017) and EukCC (Saary, Mitchell, & Finn, 2020) are employed to measure the quality of eukaryotic genomes (completeness and contamination). However, BUSCO only provides completeness measures and does not ascertain contamination. The reconstruction of eukaryotic genomes from whole shotgun sequencing 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, 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 multiple chromosomes and the share of repeat regions (Delmont & Eren, 2016).

Large-scale microbiome studies usually do not include the reconstruction of microbial eukaryotes (Nayfach et al., 2020; Nayfach, Shi, Seshadri, Pollard, & Kyrpides, 2019; Parks et al., 2017; Tully, Graham, & Heidelberg, 2018; Zhu et al., 2019), even in environments where they are key players. This bias towards prokaryotes may lead to incorrect or incomplete assertions on the contribution of microbes to ecosystems processes. Thus, studies including all domains of life 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 catalog, 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 natural environments; and 2) compare the quality of the best Eukaryotic metagenome-assembled genomes from this study to reference genomes.

Materials and Methods

Metagenome dataset

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, Edwards, & McNair, 2017). Next, metagenomes with sequence quality scores

below 70%, obtained via SRA-Tinder (https://github.com/NCBI-Hackathons/SRA_Tinder), were discarded. To allow for comparative studies, only metagenomes sequenced using the Illumina sequencing platform and with a minimum of eight million paired-end reads per library were kept. Lastly, given the consortium's focus on terrestrial environments, all libraries containing coordinates or terms for sea environments were excluded.

Pre-processing and Library assembly

The raw reads were quality controlled using metaWrap (Uritskiy, DiRuggiero, & Taylor, 2018) with default parameters. Trimming of raw reads was performed using TrimGalore (https://github.com/FelixKrueger/TrimGalore) with the default settings. High-quality reads (using default Phred scores from TrimGalore) were aligned to potential host genomes using bmtagger (Rotmistrovsky & Agarwala, 2011). The goal of this alignment is to remove host contamination and read pairs with only a single aligned read from the metagenomic libraries. Read Quality control was performed using FASTQC ("Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data," n.d.).

We used metaSpades (Nurk, Meleshko, Korobeynikov, & Pevzner, 2017) to assemble the different samples using default parameters.

Binning

Before binning, we used EukRep (West et al., 2018) to separate eukaryotic contigs from prokaryotic ones. Next, binning of eukaryotic assemblies was performed 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.

Taxonomic classification

Taxonomy was assigned using taxator-tk (Dröge, Gregor, & McHardy, 2015).

Coverage calculation

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 in base pairs (Equation 1).

Equation 1: coverage = mapped reads * average read length / size of bin (bp)

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). Next, the assembled chromosomes were aligned to the chromosomes of the reference genomes using Minimap2 (Li, 2018, p. 2). 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).

Gene prediction and Functional annotation

Genes were predicted using the GeneMark-ES model (Besemer, Lomsadze, & Borodovsky, 2001) and annotated using MAKER2 (Holt & Yandell, 2011, p. 2) with RepBase gene database (Bao, Kojima, & Kohany, 2015). The functions of interest in this work 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 (Supplementary data – Table S1 and Table S2, respectively) and nitrogen metabolism

(Supplementary data – Table S3 and Table S4, respectively) were extracted from Kyoto Encyclopedia of Genes and Genomes (KEGG).

To demonstrate the potential contribution of eukaryotes to carbon fixation and nitrogen cycling, we selected EukBins CTeuk-1331 (*B. prasinos*) and CTeuk-1332 (*M. commoda*) since they were recovered from the same metagenomic libraries used in Kieft and collaborator's study and presented the highest quality scores in their taxa. Next, we submitted the gene sequences predicted by MAKER2 to GhostKOALA (Kanehisa, Sato, & Morishima, 2016)**.** The mapping of K numbers to each gene was saved in tabular format.

Results

We recovered 447 Eukaryotic bins (EukBins) from 215 terrestrial samples. Completeness and contamination measurements using EukCC (Saary et al., 2020) were only obtained for 193 EukBins. 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 9 EukBins averaging 31.21% (σ=37.24). 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 EukBins were classified to family level (Supplementary Table S5). For subsequent analyses, we filtered EukBins with quality scores above or equal to 53 as well as those without EukCC completeness and contamination values (Supplementary Table S5). Our data had a total of 47 medium/high quality EukBins (Quality score ≥ 53) of which only 13 were classified to species level (spanning 5 unique taxa). The most frequent species-level taxonomy assigned to EukBins, 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. EukBins 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 EukBins showed only approximately six times coverage in their samples (Supplementary data – Table S5). The frequencies of each taxon, at different levels as well as per biome is shown in Figure 1. The species-level, medium/high quality EukBins were reassembled into chromosomes using Chromosomer (Tamazian et al., 2016) and mapped to the chromosomes of the reference genomes using Minimap2 (Li, 2018, p. 2). The pairwise alignments for each reassembled Eukbin are shown in the Supplementary data (Table S6). Assembled chromosomes with the highest divergences (per base differences between a query and target sequence) to the reference chromosomes were found in EukBins classified as *M. commoda* (average 0.154, σ=0.012) (Figure 2A). In contrast, assembled chromosomes of EukBins classified as *S. cerevisiae* showed the lowest divergences when compared to the reference chromosomes (average 0.029, σ=0.017) (Figure 2B). The complete set of results of divergences between assembled and reference chromosomes is shown in the Supplementary data - Table S7. Additionally, the mapping of the chromosomes of the EukBins to the chromosomes of the reference genomes is shown in the Supplementary data – Figures S1-S13.

Annotation of EukBins yielded, on average, 4106, 4435, 4573, 4619, and 4150 protein-encoding genes in *Saccharomyces cerevisiae*, *Komagataella phaffii*, *Pichia kudriavzevii*, *Micromonas commoda* and *Bathycoccus prasinos*, respectively. However, the number of predicted genes in *M. commoda* and *B. prasinos* EukBins only accounted for 45.57% and 52.54% of their reference genomes, respectively (Supplementary data – Table S8).

Functional annotation of *B. prasinos* CTeuk-1331 revealed the presence of 10 genes involved in nitrogen metabolism and 32 genes involved in carbon fixation (Supplementary data - Table S9). Functional annotation of CTeuk-1332 (*M. commoda*), revealed the presence of two genes involved in nitrogen metabolism and 34 genes in carbon fixation (Supplementary data - Table S10).

Annotation of the species-level, high-quality EukBins is available in the Supplementary data – Table S11.

Discussion

Our results demonstrate that, despite current efforts, our ability to reconstruct high-quality Eukaryotes genomes is in its early stages of development. Both quality and taxonomic assignments of the metagenome-assembled eukaryotic bins (EukBins) are substantially lower when compared to prokaryotes which can be attributed mainly to the lower availability of reference genomes and marker genes (Saary et al., 2020; Waterhouse et al., 2017). Another factor that may influence our results is the difficulty of assemblers dealing with diploid taxa given that genes may exist in two alleles with various similarities (Zhang, Zhou, Weng, & Sidow, 2020). The lower taxonomic diversity is evident in Figure 1, where we exhibit the frequency counts of taxonomies across all libraries containing EukBins. Given that current metagenomic methods rely on comparisons to known genomes, a higher number of eukaryotic reference genomes would help to identify species in complex communities (Loeffler et al., 2020). For example, in the reference database used by taxator-tk, *S. cerevisiae* was represented by more than 550 sequences while *M. commoda* was only represented by two. The quality of reconstructed genomes from metagenomes is usually calculated by the presence and numbers of Single Copy Genes (SCGs) (Saary et al., 2020). EukCC (Saary et al., 2020) and BUSCO

(Waterhouse et al., 2017) both use SCGs to estimate the quality of eukaryotic genomes. However, the SCG sets used by each tool differ in composition and application (e.g., BUSCO requires the user to define which sets of SCGs to use). The more unique and non-repeated SCGs in a bin, the higher the quality determined by either tool. However, even in high-quality EukBins such as CTeuk-1741 (95.96% completeness and 0.34 contamination) (classified as *S. cerevisiae*), significant parts of each chromosome can be missing and/or contain misplaced reads (Supplementary Figure S1). In our study, we also calculated the coverage of each EukBin. This measure allowed us to infer the relative dominance of the draft genome within a given sample. Our results showed an elevated coverage of *S. cerevisiae* EukBins. This is not surprising as the metagenomic libraries were generated from actively grown culture samples from wine barrels (PRJNA390460). In addition, the identification of *M. commoda* and *B. prasinos* EukBins in studies not accounting for eukaryotic presence suggest an incomplete overview of the entire community in ecosystem processes.

The genomic structure of eukaryotic genomes may also help to explain the differences in taxonomic classification. Eukaryotic genomes are usually differentiated from prokaryotic genomes by size and complexity (Keeling, 2019). For example, introns in eukaryotic genomes interfere with gene calling (Roy & Penny, 2007), and gene calling is difficult due to frequent gene or genome duplication events (Kaltenegger, Leng, & Heyl, 2018). Furthermore, intron presence and number vary across eukaryotic species, making it harder to accurately predict genes in some species. For example, *Aspergillus fumigatus* has 18293 introns compared to the 266 found in *Saccharomyces cerevisiae* (Roy & Penny, 2007).

Most studies involving genome annotations use short-reads and quality assessment tools such as BUSCO (Waterhouse et al., 2017) and EukCC (Saary et al., 2020). The use of short reads can

influence the accuracy of gene predictions since short reads may not cover a gene's total length (Pearman, Freed, & Silander, 2020), and SCGs 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 Eukbins. The pairwise alignments of the reassembled EukBins 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 a low number of reference genomes such as *M. commoda*. Our data showed many gaps in the mapping of the EukBins 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* EukBins (Supplementary data - Table S4). Thus, the use of 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).

In terms of microbiome studies, we recommend including results from eukaryotic genome recovery to avoid missing potential key players in ecosystem processes. For example, Kieft and collaborators (Kieft et al., 2018) studied the relationship between microbial community structure and function in carbon and nitrogen cycling in estuaries. Our results revealed the presence of *B. prasinos* and *M. commoda*. *B. prasinos* is responsible for generating almost 50% of photosynthetic picoeukaryote carbon (Vaulot et al., 2012). The carbon provided by *B. prasinos*

245 benefits the growth rates of *M. commoda* since it increases the supply rates of ammonia to the
246 nitrogen assimilation pathway (Cuvelier et al., 2017). Functional annotation of the *B. prasinos*
247 and *M. commoda* EukBins, revealed the presence of multiple genes involved in carbon fixation
248 and nitrogen metabolism. However, *nii* genes, which are responsible for converting nitrite to
249 ammonia, were missing, unlike the reference genomes. The miss-annotation of genes present in
250 *B. prasinos* and *M. commoda* highlights the challenge of reconstructing near-complete
251 eukaryotic genomes due to insufficient reference genomes in genome repositories. Still,
252 including the reconstruction of eukaryotic genomes to the study by Kieft and collaborators (Kieft
253 et al., 2018) would provide a more complete picture of carbon and nitrogen cycling in aquatic
254 environments.

255 **In summary,** performing single domain genome reconstruction from natural environments leads
256 to an incomplete overview of microbial communities' diversity and functional potential. To
257 obtain accurate representations of all species present in an ecosytem, substantial efforts in tool
258 development to identify species in all domains are still required. Eukaryotes play vital roles in
259 ecosystems ranging from complementing the activities of other microbes to performing
260 phototrophic and saprotrophic processes and predation (del Campo, Bass, & Keeling, 2020).
261 Despite their importance, very few microbiome studies include the reconstruction and analysis of
262 eukaryotes. The major difficulty in their inclusion lies in our ability to reconstruct their genome
263 and perform genome annotation accurately. Increasing the number and quality of reference
264 genomes in public databases coupled with developing tools for intron identification and removal
265 from eukaryotic genomes may result in better genome reconstructions and gene predictions. A
266 possible avenue to achieve this goal is to promote long-read sequencing technologies. While we
267 did reconstruct almost 950 eukaryotic genomes, only 29 were of high quality and classified to

268 species level. Still, the identified species showed promise in adding layers of information to the
269 original studies. Thus, the reconstruction of more high-quality genomes will bring us closer to a
270 more realistic overview and understanding of biodiversity as a whole and how Eukaryotes
271 contribute to different ecosystem processes.

272 **Acknowledgments**

273 We want to thank Felipe Borim Correa for his advice during the selection of metagenomes. We
274 would also like to thank all members of the CLUE-TERRA consortium
275 (https://www.ufz.de/index.php?en=47300) for their advice in the reconstruction of eukaryotic
276 genomes from metagenomes.

277 **References**

278 Alneberg, J., Bjarnason, B. S., Bruijn, I. de, Schirmer, M., Quick, J., Ijaz, U. Z., … Quince, C.
279 (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*,
280 *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103
281 Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020).
282 Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*,
283 *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5
284 Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data.
285 (n.d.). Retrieved October 25, 2021, from
286 https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
287 Baldrian, P., Kolařík, M., Štursová, M., Kopecký, J., Valášková, V., Větrovský, T., …
288 Voříšková, J. (2012). Active and total microbial communities in forest soil are largely
289 different and highly stratified during decomposition. *The ISME Journal*, *6*(2), 248–258.
290 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*. 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 & 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

Cuvelier, M. L., Guo, J., Ortiz, A. C., Baren, M. J. van, 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

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

313 Delmont, T. O., & Eren, A. M. (2016). Identifying contamination with advanced visualization
314 and analysis practices: Metagenomic approaches for eukaryotic genome assemblies.
315 *PeerJ*, *4*, e1839. https://doi.org/10.7717/peerj.1839
316 Dröge, J., Gregor, I., & McHardy, A. C. (2015). Taxator-tk: Precise taxonomic assignment of
317 metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*
318 *(Oxford, England)*, *31*(6), 817–824. https://doi.org/10.1093/bioinformatics/btu745
319 Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database
320 management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1),
321 491. https://doi.org/10.1186/1471-2105-12-491
322 Kaltenegger, E., Leng, S., & Heyl, A. (2018). The effects of repeated whole genome duplication
323 events on the evolution of cytokinin signaling pathway. *BMC Evolutionary Biology*,
324 *18*(1), 76. https://doi.org/10.1186/s12862-018-1153-x
325 Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG
326 Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal*
327 *of Molecular Biology*, *428*(4), 726–731. https://doi.org/10.1016/j.jmb.2015.11.006
328 Keeling, P. J. (2019). Combining morphology, behaviour and genomics to understand the
329 evolution and ecology of microbial eukaryotes. *Philosophical Transactions of the Royal*
330 *Society B: Biological Sciences*, *374*(1786), 20190085.
331 https://doi.org/10.1098/rstb.2019.0085
332 Kieft, B., Li, Z., Bryson, S., Crump, B. C., Hettich, R., Pan, C., … Mueller, R. S. (2018).
333 Microbial Community Structure–Function Relationships in Yaquina Bay Estuary Reveal
334 Spatially Distinct Carbon and Nitrogen Cycling Capacities. *Frontiers in Microbiology*, *9*,
335 1282. https://doi.org/10.3389/fmicb.2018.01282

336 Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford,*
337 *England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
338 Lind, A. L., & Pollard, K. S. (2021). Accurate and sensitive detection of microbial eukaryotes
339 from whole metagenome shotgun sequencing. *Microbiome*, *9*(1), 58.
340 https://doi.org/10.1186/s40168-021-01015-y
341 Loeffler, C., Karlsberg, A., Martin, L. S., Eskin, E., Koslicki, D., & Mangul, S. (2020).
342 Improving the usability and comprehensiveness of microbial databases. *BMC Biology*,
343 *18*, 37. https://doi.org/10.1186/s12915-020-0756-z
344 Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., … Eloe-Fadrosh, E.
345 A. (2020). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, 1–11.
346 https://doi.org/10.1038/s41587-020-0718-6
347 Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from
348 uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510.
349 https://doi.org/10.1038/s41586-019-1058-x
350 Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new
351 versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834.
352 https://doi.org/10.1101/gr.213959.116
353 O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K.
354 D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic
355 expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745.
356 https://doi.org/10.1093/nar/gkv1189
357 Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., …
358 Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes

359 substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542.
360 https://doi.org/10.1038/s41564-017-0012-7
361 Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., … Vargas, C. de. (2012).
362 CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal,
363 Plant, and Fungal Kingdoms. *PLOS Biology*, *10*(11), e1001419.
364 https://doi.org/10.1371/journal.pbio.1001419
365 Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and
366 disadvantages of short- and long- read eukaryotic metagenomics using simulated reads.
367 *BMC Bioinformatics*, *21*(1), 220. https://doi.org/10.1186/s12859-020-3528-4
368 Peng, X., Wilken, S. E., Lankiewicz, T. S., Gilmore, S. P., Brown, J. L., Henske, J. K., …
369 O'Malley, M. A. (2021). Genomic and functional analyses of fungal and bacterial
370 consortia that enable lignocellulose breakdown in goat gut microbiomes. *Nature*
371 *Microbiology*, *6*(4), 499–511. https://doi.org/10.1038/s41564-020-00861-0
372 Rotmistrovsky, K., & Agarwala, R. (2011). BMTagger: Best Match Tagger for removing human
373 reads from metagenomics datasets. *Ftp://Ftp.Ncbi.Nlm.*
374 *Nih.Gov/Pub/Agarwala/Bmtagger/*.
375 Roy, S. W., & Penny, D. (2007). Intron length distributions and gene prediction. *Nucleic Acids*
376 *Research*, *35*(14), 4737–4742. https://doi.org/10.1093/nar/gkm281
377 Saary, P., Mitchell, A. L., & Finn, R. D. (2020). Estimating the quality of eukaryotic genomes
378 recovered from metagenomic analysis with EukCC. *Genome Biology*, *21*(1), 244.
379 https://doi.org/10.1186/s13059-020-02155-4
380 Sweetlove, L. (2011). Number of species on Earth tagged at 8.7 million. *Nature*.
381 https://doi.org/10.1038/news.2011.498

382 Tamazian, G., Dobrynin, P., Krasheninnikova, K., Komissarov, A., Koepfli, K.-P., & O'Brien, S.
383 J. (2016). Chromosomer: A reference-based genome arrangement tool for producing draft
384 chromosome sequences. *GigaScience*, *5*(1), 38. https://doi.org/10.1186/s13742-016-
385 0141-6
386 Torres, P. J., Edwards, R. A., & McNair, K. A. (2017). PARTIE: A partition engine to separate
387 metagenomic and amplicon projects in the Sequence Read Archive. *Bioinformatics*,
388 *33*(15), 2389–2391. https://doi.org/10.1093/bioinformatics/btx184
389 Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft
390 metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*, 170203.
391 https://doi.org/10.1038/sdata.2017.203
392 Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for
393 genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158.
394 https://doi.org/10.1186/s40168-018-0541-1
395 Vaulot, D., Lepère, C., Toulza, E., Iglesia, R. D. la, Poulain, J., Gaboyer, F., … Piganeau, G.
396 (2012). Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling.
397 *PLOS ONE*, *7*(6), e39648. https://doi.org/10.1371/journal.pone.0039648
398 Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., …
399 Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction
400 and phylogenomics. *Molecular Biology and Evolution*.
401 https://doi.org/10.1093/molbev/msx319
402 West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-
403 reconstruction for eukaryotes from complex natural microbial communities. *Genome*
404 *Research*, *28*(4), 569–580. https://doi.org/10.1101/gr.228429.117

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., … 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

Data Accessibility and Benefit-Sharing

Data Accessibility 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 are available at 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.

Author contributions

JPS and UNR developed the concept of the study. UNR is the main supervisor of the study. RBT downloaded all sequencing data. RBT generated all data for analysis. AB and JPS performed all data analysis. JPS, ABS and UNR wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

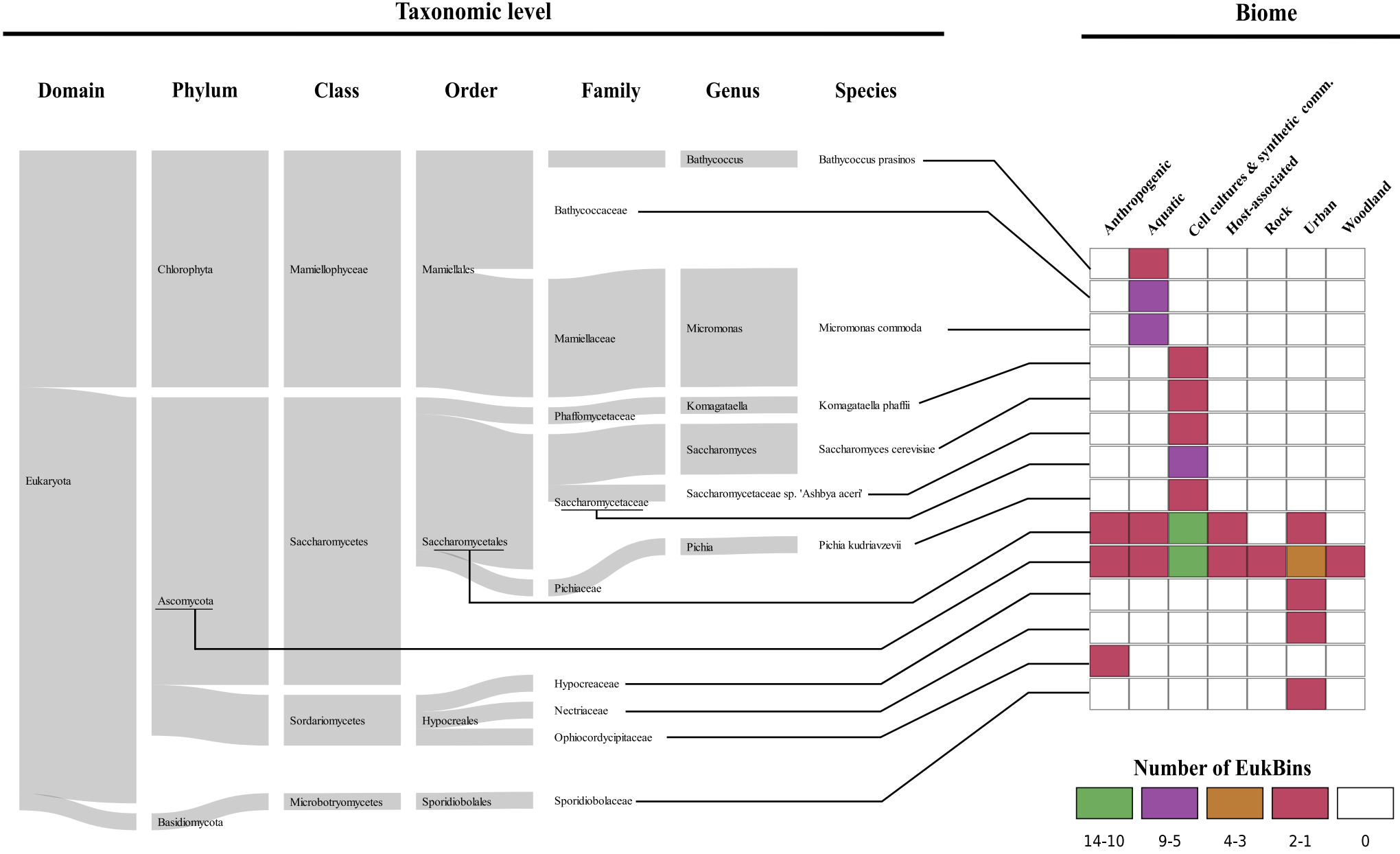
This work was supported by the Helmholtz Young Investigator grant VH-NG-1248 Micro' Big Data'.

Figure Legends

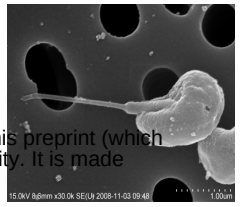
Figure 1 - Sankey plot showing taxonomic distribution of the recovered Eukaryotic bins (EukBins) and heatmap showing the number of EukBins recovered per Biome (retrieved from https://webapp.ufz.de/tmdb/ and manually curated based on the sample data).

Figure 2 - Mapping of assembled chromosomes for an 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

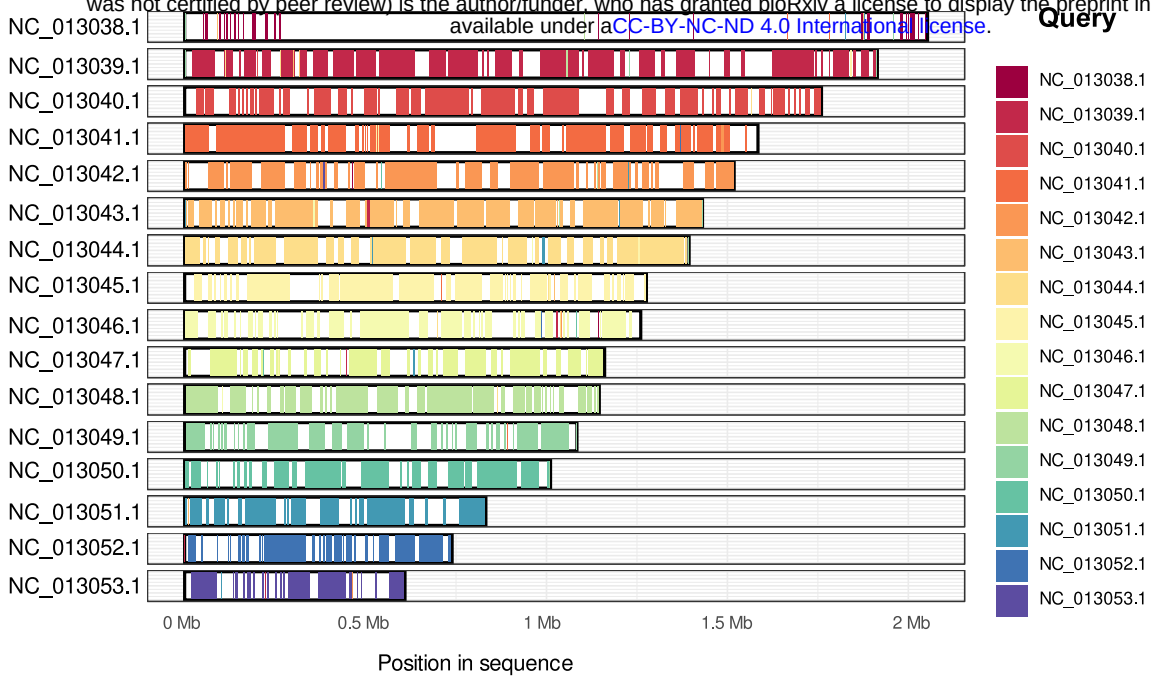
446


Taxonomic level
Biome
Domain
Phylum
Class
Order
Family
Genus
Species
Eukaryota
Chlorophyta
Ascomycota
Basidiomycota
Mamiellophyceae
Saccharomycetes
Sordariomycetes
Microbotryomycetes
Mamiellales
Saccharomycetales
Hypocreales
Sporidiobolales
Bathycoccaceae
Mamiellaceae
Phaffomycetaceae
Saccharomycetaceae
Pichiaceae
Hypocreaceae
Nectriaceae
Ophiocordycipitaceae
Sporidiobolaceae
Bathycoccus
Micromonas
Komagataella
Saccharomyces
Pichia
Bathycoccus prasinos
Micromonas commoda
Komagataella phaffii
Saccharomyces cerevisiae
Saccharomycetaceae sp. 'Ashbya aceri'
Pichia kudriavzevii
Anthropogenic
Aquatic
Cell cultures & synthetic comm.
Host-associated
Rock
Urban
Woodland
Number of EukBins
14-10
9-5
4-3
2-1
0

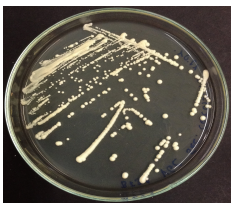
A ***Micromonas commoda (green algae)***



Source [1]



B ***Saccharomyces cerevisiae (Baker's yeast)***



Source [2]