Large-scale analysis reveals the distribution of novel cellular microbes across multiple biomes and kingdoms

Paul Saary
European Molecular Biology Laboratory
Varsha Kale
European Molecular Biology Laboratory
Robert Finn (rdf@ebi.ac.uk)
EMBL-EBI https://orcid.org/0000-0001-8626-2148

Article

Keywords:

Posted Date: March 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1441815/v1

Abstract

Shotgun metagenomics provides access to genetic information of microbes in a culture-independent manner. The recovery of metagenome assembled genomes (MAGs) of these organisms enables in-depth analysis of the functional potential of these, often elusive, organisms. While workflows for the recovery of prokaryotic MAGs are established, MAGs of eukaryotic species are seldom reported. To address this, we developed new approaches to improve the recovery of eukaryotic MAGs and applied these to large collections of assembled metagenomes in an effort to systematically identify both prokaryotic and eukaryotic MAGs. Using these approaches we identified 99,073 prokaryotic and 751 eukaryotic genomes, and propose a method for reducing the redundancy of single-celled eukaryotic MAGs at the species level.

Introduction

Microbes are essential members of all known ecosystems and rarely exist in isolation. Using a wide range of functions, they are able to colonise and shape environments that are too hostile for other organisms[1]. While much attention has been given to the prokaryotic species and their functional capacity[2–4], microbial eukaryotes such as fungi also perform key roles in many environments, for example by making carbon available and by creating new environments[5]. The complexity of these mixed communities can not easily be approached using traditional isolation methods, as many of these microbes are fastidious and elude cultivation. Even if cultivation were possible, isolation of all microbial species would be problematic due to the sheer number of species involved[6]. Thus shotgun metagenomics and subsequent analysis, which facilitates the recovery of the genomes of microorganisms without prior cultivation, has become a powerful tool to study the microbes found in many biomes.

With shotgun metagenomic datasets covering an ever widening range of biomes, multiple studies have been conducted with the aim of recovering the genomes of prokaryotes, viruses and more recently eukaryotes[7]. Metagenomics has also been used to address the composition and biology of holobionts, such as corals and sponges, and their associated microbiome[8,9]. These metagenomic studies have been central to our understanding of microbial communities from both functional and genomic perspectives.

Discovery of new bacterial species shows little sign of abating, demonstrating the continued need to improve the experimental methods and informatics tooling used to produce metagenome assembled genomes (MAGs) thereby allowing the full microbial diversity to be described.

While eukaryotes are often present in lower numbers, and are yet to show the same level of diversity in the microbial space, they are found in many biomes and interact with other members of those communities[10–13]. However as eukaryotic genomes are generally larger and lower in abundance in shotgun metagenomic datasets, they often go unreported. In this study we report a workflow for the recovery of eukaryotic and prokaryotic MAGs from public data covering a range of biomes. As part of this, we report an extension to our tools to improve the quality of the recovered MAGs and to enable dereplication.

Results

Over the past decade, the number of metagenome assemblies available has increased significantly. From 193 assemblies deposited in the nucleotide archives at the end of 2016 to over 78,000 at the end of 2021. One of the largest single contributions of assemblies has come via the MGnify resource accounting for over 19,000 (April 2021) submitted assemblies[14]. The assemblies created by MGnify span a wide range of biomes including the human gut, human skin, aquatic environments and engineered biomes, such as wastewater and food systems. While MGnify hosts the Unified Human Gastrointestinal Genome Catalogue (UHGG)[3] and more recently a Marine, Cow Rumen and Human Oral catalogue, the majority of assemblies have not been systematically analysed to generate MAGs.

To address this analysis gap and achieve our goal of searching for eukaryotes, we downloaded all available assemblies in April 2021 (19,800 in total) from MGnify and screened each assembly using EukRep[11], retaining only those assemblies estimated to contain at least 5Mb eukaryotic DNA (6,430 assemblies). This threshold was used to enrich for assemblies most likely to have sufficient DNA to recover a eukaryotic genome. As most datasets in MGnify pertain to experiments aimed at exploring the prokaryotic microbial fraction, the sample preparation can have excluded (e.g. due to size fractionation) or ignored features of eukaryotes that may limit their recovery (e.g. robust fungal cell walls). To enrich for experiments designed for eukaryotic organisms we further included 451 assemblies for the size fractions of protists from the TARA expedition[15] (ERP003628). Additionally we were interested to apply the same workflow to lichen data (as a holobiont) and thus assembled a collection of 72 lichen shotgun metagenomes from two studies (SRP272267, SRP305791). After removing human host contamination from all assemblies, the total dataset contained 571 Gbp from 6,181 assemblies across 272 studies. We performed genomic binning on all the assemblies using CONCOCT[16], and obtained 317,195 bins.

CONCOCT was chosen as it does not assume all genomes in the metagenome to be prokaryotic, making it suitable to bin mixed communities of prokaryotes and eukaryotes. As the assemblies in this study come from a range of different biomes, they have been categorised into six larger groupings: Human gut (n = 4,448), aquatic (n = 1,046), engineered (n = 433), human skin (n = 410), lichen (n = 72) and other (n = 472). At this stage the bins were expected to be a mixture of both prokaryotic and eukaryotic genomes.

Prokaryotic MAGs

To select prokaryotic MAGs, we applied the quality score (QS) threshold of QS50, defined by Parks et al.[4] as completeness − 5*contamination ≥ 50, based on CheckM. From the original set of 317,195 bins, 99,073 (31.1%) passed this QS threshold. As the same genome may have been recovered multiple times over, redundancy was removed using dRep to produce species level representative MAGs, based on 95% ANI and 30% AF (see methods). This process produced 7,680 species representative MAGs, of which 4,496 (58%) were categorised as near complete MAGs with > 90% completeness and < 5% contamination. As expected, only 156 of the near complete MAGs passed the MIMAG criteria for high-quality MAGs, due

to the difficulty in assembling the ribosomal operon. For the subsequent analysis, we focused on the 4,496 near complete MAGs.

To determine the fraction of species representatives that were novel, we compared our collection of MAGs with isolate genomes and other MAG collections. First, we used the GTDB and the associated toolkit (GTDB-Tk) to classify our MAGs. This demonstrated that the vast majority of all MAGs were bacterial (4,397), while 82 were classified as archaeal. 2,756 of our MAGs could be matched to known genomes using the GTDB-Tk (isolate and MAGs, Fig. 1). We then compared the remaining 1740 MAGs against the following MAG collections, chosen based on the overlap between their their source biomes with our datasets: the UHGG[3], TARA Ocean MAGs[12], the MGnify marine catalogue, a skin MAG collection[17] as well as a the Nayfach collection (MAGs from a wide range of biomes)[2]. 548 of the remaining 1,724 MAGs were assigned to one of the aforementioned collections, with an ANI of at least 95% and 30% or more of the genome aligned. 268 MAGs matched to a MAG from the Nayfach collection, 194 to a MAG from the UHGG, 72 to the Marine catalogue, 12 matching the skin MAG collection and 2 to the TARA MAG collection. The fact that 72.9% (n = 3,267) of the original representative species were matched in other catalogues, demonstrates the reproducibility of prokaryotic MAG discovery pipelines, matching both isolate and MAGs using other approaches. The remaining 1,176 (26.2%) near complete MAGs (1,157 bacteria, 19 archaea) appear to be novel species, demonstrating a huge wealth of additional prokaryotic diversity discovered in this study (Fig. 1).

We next investigated which environments contained the most prokaryotic novelty. Only 8% (163/1935) of all human gut associated MAGs, and 17% (8/45) of human skin MAGs were considered novel, an outcome that is unsurprising since both biomes have been the subject of recent, extensive studies. Conversely, 34% (235/681) and 38% (378/1001) of MAGs were novel in the aquatic and engineered clades respectively, while all reported lichen associated MAGs appear to be novel (n = 32).

Taxonomically, most novel MAGs belong to the phylum Bacteroidota (n = 286), Firmicutes_A (n = 223) and Proteobacteria (n = 205). To estimate the amount of diversity the collection of novel MAGs contributes, we computed (i) the additional branch length contributed compared to GTDB and already known MAGs by the novel MAGs per phylum and (ii) the proportion of the total phylogenetic diversity that these MAGs contribute to their assigned phylum (Fig. 1, B and C). Focusing on the top ten phyla with most gained branch length, we found that these novel MAGs increased the total branch length between 14 and 43%, while increasing the diversity between 5 and 46%. The phylum with the highest increase in total branch length was Bacteroidota, while Planctomycetota gained the most diversity. Planctomycetota play a vital role in global carbon and nitrogen cycles[18] as well as many species harbouring anaerobic ammonium oxidation. We found Planctomycetota exclusively in aquatic (marine) and engineered (bioreactors) samples, with 19 and 18 novel species recovered respectively.

The largest number of Bacteroidota species were found in the human gut (n = 343), while the most novelty was identified in the mixed biome (n = 105) group (termed “other”, 25 different MGnify lineages), engineered samples (n = 98) and the aquatic biome (n = 55).

We next assessed whether the bins (that were not classified in the aforementioned section) correspond to genomes from eukaryotic microbes using EukCC[19] (version 2), which was also extended as part of this work to refine eukaryotic MAGs.

Eukaryotic MAG recovery and refinement

As with prokaryotic MAGs, we aimed to recover eukaryotic MAGs that pass the same QS50 threshold principle. For this, EukCC was run to obtain genome quality scores. During this process we noticed that many samples contained two or more bins that could be assigned to the same species, yet individually were estimated to have low completeness, i.e. scored below the QS50 threshold. This could be accounted for by recovering low quality bins (i.e. incomplete genomes) of two closely related strains in the sample or by a single genome being split across multiple bins. As eukaryotic genomes tend to be larger and have a more diverse k-mer content than bacteria, and as CONCOCT assumes a gaussian distribution of k-mers in the PCA space, we evaluated whether these eukaryotic genomes could represent a single genome. This feature has also been noticed by others, for example Delmont et al. 2020 resorted to manual curation of eukaryotic MAGs to overcome this issue[7]. While manual curation undoubtedly leads to improved bins, scaling this approach for large datasets such as this study can be problematic.

To investigate if a single eukaryotic genome has been split over multiple bins, we first manually verified by merging two bins assigned to the same taxonomy and reassessing the merged bin completeness and contamination. In the case that these two bins came from a single genome, we expected an additive increase in completeness, while maintaining contamination scores. Further validation for merging the bins was also achieved by aligning merged bins against reference genomes. Based on these initial experiments, we defined a set of rules that could be routinely applied (and have since been implemented in EukCC, version 2, for automated eukaryotic bin refinement (Fig. 2A)): Briefly, after the first round of contig binning we align paired-end reads to the entire assembly to identify reads spanning contigs found in two different bins, limited to those individual reads mapped to the first or last 1500 bp of each contig. To minimise alignment of reads across contigs from different species, we limited our search to reads aligned with at least 99% ANI (see methods). Next, we used the spanning information to identify those bins that were connected by more than 100 read pairs. This threshold of connecting reads was chosen conservatively to exclude a large number of pairwise bin comparisons (across all binned assemblies we retain ~ 2.8% of all comparisons between at least one 50% complete bin and all others from the same assembly, median links between bins: 2 reads). Those bins passing the threshold were then considered as “connected” and underwent further validation steps. To evaluate the merits of merging “connected” bins, we first identified those pairings where one bin was ≥ 50% complete and contained < 10% contamination (termed primary bin). We then took the connected bin (termed secondary bin), and evaluated the genome quality score based on the union of the primary and secondary bins. If the gained completeness was at least 10% and any increase in contamination was less than 1/5 of the completeness gain (i.e. congruent with QS50 threshold), the primary and secondary bins were considered to have originated from the same genome to form a merge MAG (mMAG) (Fig. 2A). While merging multiple secondary bins may be warranted (See Tagirdzhanova et al.[20]), instances where the addition of

multiple secondary bins passed the aforementioned criteria were infrequent, and hence our approach was to restrict the refinement to one merging event in this study (and by default to one in EukCC). Similarly, it would also be feasible to recruit additional, unbinned contigs to the MAG using a similar approach. However, both of these approaches are less likely to increase genome completeness significantly, and as smaller amounts of genomic sequence are added, it becomes harder to assess contamination. As such, these represent cases where manual curation should be undertaken. We obtained 620 bins of at least QS50 across 495 runs, using the automatic merging approach implemented in EukCC we improved the quality of several MAGs by creating 131 mMAGs.

To systematically evaluate our merging procedure, where possible, we compared the mMAGs to a published reference genome (e.g. a GenBank genome or a eukaryotic MAG) (Fig. 2). Our main source of eukaryotic reference MAGs came from Delmont et al. (2020)[7] and thus provided an opportunity to evaluate our automatic merging approach against a manually curated set. To perform the genomic comparisons, we first identified the closest reference genome entry for each primary bin from all the mMAGs using mash. Using the same reference genome, we then computed the aligned fraction (AF) and ANI using DNAdiff for the primary and secondary bin independently. To make sure the selected reference genome was of a species from the same genus, we only retained those mMAGs in this comparison for which the primary bin had at least 95% ANI with 40% of the bin aligned to the reference (n = 107, Fig. 2B-D). The median ANI of the primary bins was 99.1%, indicating that the reference genome was very similar to the recovered MAG. Similarly, when the secondary bins were compared to the reference genome, the median ANI was 98.9% (a difference of only 0.2% between the ANI), despite the reference being selected based on the primary bin only. The ANI of primary and secondary bins correlate with an R of 0.96, further supporting that the EukCC merging algorithm identifies bin pairs originating from the same genome and validating the creation of mMAGs. Finally, where possible we compared the primary bin and mMAG size (bp) to the size of the reference genome (Figure S1). This analysis was limited to those reference genomes that came from an isolate source, as the eukaryotic reference MAGs could be incomplete. The mMAGs median size is notably bigger than the primary bin, but on average mMAGs are still 15% smaller than the assigned reference genome. Normalising mMAGs and primary bin sizes by the estimated completeness, the median relative expected size is 1.07 and 1.02 of the reference respectively, again confirming that EukCC estimates the completeness correctly in MAGs and mMAGs and that creating mMAGs does not lead to artificially large MAGs. In order to test whether primary and secondary bins are defined by genomic features, 13 mMAGs that could be assigned to *Malassezia Globosa* were aligned against a highly continuous reference genome. This showed that the primary and secondary bins contain contigs that align interspersed to this reference (Figure S4).

In line with the workflow for bacterial genomes, we next investigated how to remove redundancy for these 751 microbial eukaryotic MAGs.

Delineation of microbial eukaryotes species

Defining a species is complex, often founded upon functional and/or morphological characteristics, or using marker gene trees (e.g. small subunit ribosomal RNA). However, these features are not routinely accessible in MAGs. For prokaryotes, global genomic alignments have been utilised to delineate species. Briefly, if two genomes can be aligned against each other with an AF of at least 60% and have an ANI above 95%, they are considered to be the same species and only a single representative is reported for that group. Based on this AF threshold determined for isolate genomes, Almeida et al.[3] introduced the concept of using an AF threshold of 30%, as a pair of MAGs may only be 50% complete. While this lower AF could result in the merging of two separate species into a single species, it prevents the overestimation of the number of species. To understand how genomic alignments could be used to define eukaryotic species, we generated alignments for microbial eukaryotic genomes contained in NCBI GenBank (September 2020). For this analysis, we relied on the subdivision into three different clades: fungi (n = 6793), plants (n = 1535) and protozoa (n = 945). For the plants, we excluded all Embryophyta genomes from the plant clade and will refer to the remaining plants as algae hereafter (n = 161).

Computing a comprehensive all vs. all comparison for all of GenBank is computationally expensive. To reduce the computational overhead we used the NCBI taxonomy to create a taxonomically balanced set of 2,445 pairs for fungi, 433 for protozoa and 115 for algae at a range of taxonomic distances (See methods). We then computed the ANI for all genome pairs, and plotted the ANI against the aligned fraction (AF) (Fig. 3)

In Fig. 3 both the ANI and the AF form a bimodal distribution separated by a space of lower frequency. Colouring the distributions using the NCBI taxonomy, within species comparisons (blue) are notably separated from comparisons between different species.

For the ANI values the area of low frequency spans between 83% and 96% ANI, thus any value in between those two could be used to determine if two genomes belong to the same species. A lower ANI threshold will more frequently merge two separate species, while a higher value will overestimate the number of species. Similarly the AF values between 20% and 75% are infrequently observed, with same species comparison enriched in the > 75% bracket. We chose a 95% ANI threshold as dereplication cutoff (indicated by vertical line), leaning towards species merging rather than overestimation of species. However MAGs are often incomplete, so we chose a more liberal AF threshold of 40% (indicated by horizontal line), to allow for comparison between incomplete genomes, yet still within the range supported by the bimodal distribution.

Applying these thresholds produces a very close recapitulation of the expected species division according to the NCBI taxonomy. We accurately identify 85.9% of all intra-species comparisons. For the remaining 14.1% we would not consider the genomes to have originated from the same species. These thresholds also result in some differences to the NCBI taxonomy, and include 8.7% of all inter-species comparisons from the same genus, but only 0.7% (n = 3) of inter-species comparisons across different genera. Despite the small number, we were interested to understand these more extreme “errors” in more detail: The first genome pairing where our approach brings together two species (*Nannochloropsis gaditana* and

Nannochloropsis salina) that belong to different genera, come from the clade Monodopsidaceae, with the genomes sharing an ANI of 95.38% over an AF of 85.29%. The phylogeny of this clade was first defined by Hibberd in 1981[21] and then revised using 18S sequencing information by Andersen in 1998[22]. Despite the common genus name, *N. gaditana* and *N. salina* are assigned to the genera Nannochloropsis and Microchloropsis respectively. This taxonomy contradicts the re-organisation by Fawley et al. 2015[23], who argued against the current NCBI taxonomy and annotated both species as belonging to the new proposed genus Microchloropsis, suggesting these genomes may be more closely related. The next outlier pair belongs to the family Thraustochytriaceae (slime nets) and share an ANI of 98.31% with an AF of 97.51% (*A. acetophilum*) and 98.79% (*Schizochytrium sp. TIO01*). The two genomes compared here are classified as belonging to the genera Schizochytrium and Aurantiochytrium. The entry of *Aurantiochytrium acetophilum* was only recently published by Ganuza et al. 2019[24]. Using morphology, 18S sequence information and protein comparisons, the authors describe why they classify this species as belonging to Aurantiochytrium. We found this genome to closely match *Schizochytrium sp. TIO01* species by Hu et al.[25]. However, a small phylogenetic tree for the Thraustochytriaceae family based on 22 gene profiles, places the aforementioned genomes as closely related to each other (Figure S2). The last comparison, *Rhizophagus sp. MUCL 43196*[26] and *Oehlia diaphana*, belong to the same family of Glomeraceae fungi. The entries share an AF of 87% and 75% respectively and an ANI in these aligned regions of 99.82%. The Oehlia genus was defined in 2018 by Błaszkowski et al. as a sister to Rhizophagus based on morphological and molecular markers [25]. Despite these morphological differences, the genomes are incredibly similar.

Thus, using these thresholds is a robust and reproducible approach to remove redundancy (at a species level) within collections of eukaryotic microbial genomes. Compared to the NCBI taxonomy, in ~ 15% of cases two genomes are reported for a single species (ANI < 95%, AF,40%). Whereasy in ~ 8% of cases different species from the same genus are merged using our thresholds, however rarely if at all (< 1%), species from different genera are being considered as the same species, which would result in an underestimation of novelty.

De-replication of Eukaryotic MAGs

We applied the empirically derived thresholds of 95% ANI and 40% AF parameters for removing redundancy in our microbial eukaryotic MAGs and mMAGs, and reduced our collection of 751 MAGs to 124 species representatives. Among the representative MAGs 20% (n = 26) are mMAGs, a similar proportion of mMAGs as were found in the redundant set (n = 131, 17%), indicating that mMAGs are equally likely as MAGs to be chosen as a species representative.

Our eukaryotic MAGs have a mean size of 24 Mbp, 10x larger than the median prokaryotic MAG with 2.4 Mbp. The eukaryotic MAGs range in size from 4.8 Mbp to 55.6 Mbp. Using genomic comparisons to isolate and MAG collections (see methods), 58 out of the 124 species representatives could be assigned to an already reported species, while the remaining 66 were considered to be novel eukaryotic species. Using BAT and NCBI GenBank, a putative taxonomic lineage was assigned to each MAG: 98 species are

fungi, 19 algaes and 7 protozoa. 58 representative MAGs were obtained from the lichen biome with 55 of these assigned to fungi and 3 as viridiplantae. 52 of the lichen derived representative MAGs are novel. Lichens are a symbiotic organism of at least one fungus and one photosymbiont. Beyond these two species, additional eukaryotes and prokaryotes can be found as part of the symbiosis or colonising the lichen, making the use of metagenomics to recover MAGs plausible, but arguably simpler than in environments such as soil. Besides lichen derived species, we recovered novel fungi from engineered samples (n = 4) and novel viridiplantae from the aquatic biome (n = 2). We also recovered a single novel Amoebozoa from an engineered sample (Fig. 4).

To understand the phylogenetic contribution of the species representative MAGs we used widely prevalent eukaryotic marker genes to construct a phylogenetic tree (see methods, Fig. 4). The tree groups all fungi into a single clade, but fails to correctly separate out all protozoans from the viridiplantae. Both a Diatom MAG from the Order *Bacillariales* and a MAG assigned to the *Rhizaria sp. SCN 62–66* are found within the viridiplantae clade. Measuring the branch length in each clade, the novel species recovered by us contribute 10% of the total branch length in the protist clade and more than 50% for fungi and viridiplantae.

Eukaryotic MAGs across biomes

We recovered eukaryotic MAGs from 495 out of over 6,000 assembled runs, despite them all containing an estimated 5Mbp of eukaryotic DNA. Given the redundancy of our MAGs, we searched for the presence of our representative eukaryotic species in all 6,000 runs (see methods) and found at least one species in 1,015 runs. 67 species were found in at least one additional sample to the one they were discovered in. Most species were identified only in runs from a single biome (n species = 111). For the remaining 13 species that were found in at least two biomes, the distribution was heavily skewed towards one particular biome. We found that the biomes differ by their relative occurrences of the identified fungi, algae and protozoans: we found signatures of our fungal MAGs in all analysed human skin and in most (99%) lichen samples analysed (Fig. 5A). While almost all the eukaryotes we identified in engineered samples were fungi, they were only found in 21% of engineered samples. Similarly, in aquatic biomes we mainly found algae, but only in 31% of all samples. In 7.7% (326/4192) of human gut samples we found traces of Blastocystis species (protozoans), highlighting the importance of considering eukaryotic species as part of the microbiome. Blastocystis are the most prevalent protists known to colonise the human gut. Currently there are eight representative blastocystis genomes in the NCBI GenBank catalogue. We found 56 (incl. 12 mMAGs) blastocystis MAGs passing the QS50 threshold. After de-replication, four blastocystis MAGs remained, each assigned to the reference genomes *Blastocystis sp. 2, 3, 4* and *Blastocystis sp. ATCC 50177/Nand II*. Each MAG covers at least 80% of the reference genome with an ANI of over 97% and up to 99%. While no novel blastocystis species was found in this study, we showed that we can recover half of all known blastocystis genomes with a reference free approach for the first time. Our MAG recovery approach should enable anyone working with human gut samples to access blastocystis genomes and hopefully contribute to further understanding of this organism and its global distribution.

We have demonstrated that Eukaryotes are common members of microbial communities, often overlooked in metagenomic studies. Our improved tooling has enabled us to access these eukaryotes at scales and quality previously not previously reported. This has enabled the recovery of genomes that substantially increase known diversity. To better understand the contribution to the functional capacity we investigated the proteins found in the eukaryotic MAGs and the metabolic pathways encoded in both the prokaryotic and eukaryotic genomes to provide insights to the potential contributions provided by the two kingdoms of life.

Eukaryotic MAGs are a resource of novel proteins

We obtained a set of 1,065,617 *de novo* predicted proteins from the representative eukaryotic MAGs with sizes ranging from 39 to 14,123 amino acids (aa) and a median size of 377 aa. The longest protein with 14,123 aa does not match any other known sequences in NCBI nr, but contains a single HECT domain (PF00632.25,e-value 1.1e-76) at the C-terminus and shares many characteristics to other sequences in that Pfam family. As the lichen-associated species make up 47% of the MAGs in this set, it is unsurprising that 57% of the proteins are from the lichen-associated genomes, with the rest split among genomes coming from the remaining biomes. We clustered this protein set with UniRef90 (release 2021_03, 135M entries) based on 90% sequence identity (see methods): 390,257 proteins are grouped into clusters containing two or more sequences, of which 215,031 of the proteins clustered with at least one member of the UniRef90 set, and the remaining 175,226 proteins form 47,800 novel clusters containing at least two members (12,014 of these clusters have > = 5 sequences). Among the novel clusters the single largest contributor is the lichen biome with 169,621 proteins contributing a total of 45,193 protein clusters (See suppl. Figure S6).

Metabolic potential of metagenomic communities

Similar to prokaryotes, eukaryotic communities are varyt between environments: in some environments we obtained mostly fungi (human skin, engineered samples), while in others we recovered mostly viridiplantae (marine) (Fig. 5A). To understand the potential metabolic contribution made by these different eukaryotes to the overall metabolic landscape we predicted KEGG[28] and CAZy[29] modules in all eukaryotic genomes and near complete prokaryotic MAGs (See methods). We then removed KEGG modules that were kingdom-specific (e.g. using keywords such as fungi or Prokaryotic) and determined the sets of unique and overlapping modules per kingdom in all biomes. Using this approach, we identified 185 KEGG modules in bacteria, 62 in archaea and 102 in eukaryotes. Similarity we identified 327 CAZy modules in bacteria, 77 in archaea and 207 in eukaryotes. 43.2% of these modules are common between kingdoms, 44.6% unique to prokaryotes and 12.3% unique to eukaryotes.

Including eukaryotic genomes leads to a mean gain in KEGG and CAZy modules per sample between 8.4 and 40.7%, depending on the biome, and up to 50–75% in some samples (Fig. 5C). Using this data we can show that each community of microorganisms is uniquely adapted to its environment, with a distinct metabolic fingerprint that clusters by biome (Figure S5).

Among the complete set of all species we find 70 modules unique to eukaryotic species (23 KEGG modules, 47 CAZy modules). The 47 eukaryotic specific CAZy modules are 24 Glycosyl Transferases (GTs), 8 Auxiliary Activities (AAs), 7 Glycoside Hydrolases (GHs), 6 Carbohydrate-Binding Modules (CBMs) and a single Carbohydrate Esterase (CE) and Polysaccharide Lyase (PL). Separating the CAZy modules per biome, between 18.9 to 30.9% of GTs are unique to eukaryotes in the aquatic (30.9%), engineered (30.0%), human gut (24.3%) and human skin biome (18.9%) (Fig. 5C).

It is known that marine fungi can contain a large number of CAZy functions and even be able to grow on pure cellulose and plant waste[30–32]. This metabolic capacity makes marine fungi interesting candidates for biotechnology applications. We sorted all MAGs, prokaryotic and eukaryotic, by their total number of CAZy proteins: the top 10 MAGs exhibiting the largest number of CAZy functions (n > = 369), were mostly eukaryotes (n = 6) and human gut bacteria (n = 3) (Figure S8). Two out of four marine fungi recovered in this study, could be found among these top 10, occupying the two top spots, with more than 500 identifiable CAZy proteins each. The two fungi producing this large number of CAZys belong to the orders Helotiales (ERR868449_bin.23) and Pleosporales (ERR599223_bin.4), one of which is novel to this study. To break down cellulose, fungi require GH's of the groups 5,6,7,12 and 45, which were all found in both aforementioned fungi, highly indicative of their capability to utilise cellulose as a carbon source[32]. The prokaryote with the largest number of CAZys was *Bacteroides cellulosilyticus* recovered from the human gut. It is known to degrade cellulose and grow on multiple carbon sources[33], functions aligning well with the high number of carbohydrate associated functions detected.

Further fungi found in the top 10 were recovered from the human gut and engineered samples: two human gut associated fungal species were a *Pseudogymnoascus* associated with a 5300-year-old mummy sample (Acc: ERP012908) and a *Penicillium arizonense* from the human gut (Acc: SRR12395656). Fungi from the engineered biome were obtained from the International Space Station (ISS) (*Penicillium sp. HKF2*) and involved in wine fermentation (*Aureobasidium sp. FSWF8-4*), showing the large metabolic potential and broad biome distribution fungi have.

Discussion

We recovered over 4,500 high quality MAGs from across the prokaryotic and the eukaryotic kingdoms from published datasets covering a wide range of biomes. We show that even assembled, published, and publicly available data, contain a large number of novel species (1,242 out of 4,620 reported here). Numerous biological and experimental factors account for the uneven distribution of the novelty we discovered here, such as the extensively studied nature of the human gut. On the other hand, engineered samples cover a diverse range of environments, stretching from ISS derived metagenomes to fermented beverages. We identified a large number of novel prokaryotic MAGs from these samples (n = 378), which partly can be attributed to the relative undersampling of these biomes and lack of MAG characterisation.

We have revealed that eukaryotic genomes can become fragmented during binning, but that this can be overcome through our unsupervised approach implemented in EukCC to recover larger and more

complete eukaryotic MAGs. Using whole genome alignments against isolate genomes and manually curated MAGs, we showed that our mMAGs capture more complete genomes without impacting contamination.

To remove the redundancy of eukaryotic MAGs, we have demonstrated that ANI and AF can be used to select species representatives that closely resemble the NCBI taxonomy. While the dataset that we used to obtain these thresholds is heavily biassed towards fungi, the data that we have from other clades suggests this threshold can be generalised to all clades of microbial eukaryotes. With more MAGs and isolate genomes being sequenced every month, the validity of these parameters beyond fungi will continue to be assessed in the future. For now this approach provides an appropriate mechanism for removing redundancy in eukaryotic MAGs, and will present a new standard in the field.

Using our eukaryotic genome recovery methods we obtained over 750 MAGs, with representatives for over 100 eukaryotic species. Recovering eukaryotes from any environment is always limited by the sequence coverage, as sequence reads from eukaryotic genomes often make up only a small fraction of the total set of reads. This may be biologically driven as eukaryotes tend to only comprise a small portion of the total biomass, but also can be a symptom of the library preparation, which is often optimised to gain access to the prokaryotes. Nonetheless, it is still worth searching public datasets for evidence of eukaryotic MAGs to make the most out of the work that has been put into the collection and sequencing of the samples. These MAGs provide invaluable insights into the proteins and functional potential these organisms contribute to the communities in which they are found.

Overall, we recovered 124 species of eukaryotes and over 4,500 species of prokaryotes. Despite this large numeric discrepancy, the microbial eukaryotic genomes contribute a noteworthy proportion of novel metabolic function, not provided by the prokaryotic fraction of organisms across most biomes. Even excluding kingdom-specific pathways from this analysis, it is clear that eukaryotes are key members of microbial ecosystems. We also show, especially in the cases of lichens, that the metagenomic assembly of eukaryotic genomes leads to the discovery of a wealth of novel proteins.

Conclusion

Eukaryotes play a vital role in the communities all around us. They are part of microbial communities and can be recovered from existing and new shotgun-metagenomic datasets. Including eukaryotic MAGs can explain additional metabolic processes that would be missed when only considering prokaryotes. Our tools and approaches facilitate large-scale recovery of eukaryotic MAGs, providing new insights into these larger and more complex microbial genomes. We expect that the recovery of eukaryotic MAGs will increasingly become easier once long-read data becomes more widely adopted, but as of now our approach can already reproducibly recover MAGs from many clades of microbial eukaryotes.

Methods

Bin recovery from assemblies

Metagenomic assemblies and matching reads were downloaded from ENA and scanned using EukRep with default settings. We retained any assembly that contained at least 5Mb of eukaryotic DNA. Additionally we obtained assemblies for ERP003628 [15] and created assemblies using metaSPAdes (version 3.13.0) for SRP272267 and SRP305791. Resulting assemblies were subsequently binned using CONCOCT[16] (version 1.1) as part of metaWRAP[34] after trimming reads using fastp[35] (version 0.20). To recover prokaryotic MAGs we used CheckM v1.1.3[36] to estimate completeness and contamination and retained 99,073 bins with at least QS50 (completeness - 5* contamination >=50). All bins not passing the prokaryotic-based QS50 were then processed additionally with EukCC[19] (version 2.0) and retained if they matched QS50 using the EukCC completeness and contamination estimates.

Prokaryotic MAG processing

To remove the redundancy (and select species representatives) within the 99,073 prokaryotic MAGs passing the QS50 threshold, and to group MAGs into quality classes, we followed a similar process as outlined by Almeida et al. 2019[37]. Briefly, MAGs were de-replicated using dRep (version 3) in multiple steps. Firstly, bins in each study were de-replicated, reducing the set to 22,277 MAGs. We then split the 22,000 MAGs into groups of 5000 and de-replicated all 5 groups individually. The resulting set of MAGs was de-replicated a final time to create a set of 7680 non-redundant species MAGs. In all MAGs with at least 90% completeness and at most 5% contamination (near complete MAGs) we annotated ncRNA to identify high quality MAGs following the MIMAG standard[38]. We searched for ribosomal RNAs (rRNAs) using infernal's cmsearch[39] (-Z 1000 –hmmonly –cut_ga –cpu 4 –noali) with the Rfam[40] covariance models for 5S, 16S and 23S. rRNAs genes were counted as present if the sum of non-overlapping hits was at least 80% of the gene length. We identified transfer RNAs (tRNAs) with tRNAscan-s.e. v.2.0[41] using the bacterial tRNA model (-B -Q). MAGs were then classified as either high quality (>=90% completeness, <=5% contamination, presence of 5S, 16S and 23S rRNA genes, and at least 18 tRNAs), near-complete MAGs (>=90% completeness, <=5% contamination, not matching MIMAG standard for rRNAs and tRNAs) and medium completeness (completeness - 5*contamination>=50). The set of MAGs was classified using GTDB-Tk (v1.5.0) using default parameters. Additionally MAGs were compared using mash[42] and mummer dnadiff[43] (version 3.23) against three large MAG collections from Nayfach et al. 2020[2], the unified human gut MAG catalogue (UHGG)[3] and a collection of skin MAGs[17]. For this we ran dnadiff with the top scoring mash hit for each MAG. MAGs are considered novel if they were not classified to a known species by GTDB-Tk and could not be assigned an already reported MAG with an ANI>=95% and AF>=30%.

Eukaryotic bin refinement

For eukaryotic bin refinement, reads were aligned to all bins using bwa mem. Only reads aligned with a Quality score (QS) >=20 were retained. Paired-reads were then reduced to those aligned in the first or last 1500 bp of a given contig and counted using the $b \in l \in ks.\,py$ script of EukCC. EukCC was then run on each assembly (eukcc folder –improve_percent 10 –n_combine 1 –threads 4 –improve_ratio 5 –links links.csv –min_links 100 –db $DB –out out –prefix $NAME_merged. binfolder): Firstly, EukCC calculated the quality scores for all bins. Bins between 50-90% completeness were marked as potential primary bins if they were connected to a second bin of less than 50% completeness with at least 100 paired reads. A merged MAG was created by merging the primary and secondary bins and retained if completeness increased by at least 10% and contamination at most by 1/5 of the gained completeness. Both QSs for the created mMAGs and all other bins were collected. To verify the genome quality of the merged bins, we compared each primary bin to genomes in NCBI or published eukaryotic MAGs using mash (dist -s 10000). We then used dnadiff to align the primary and secondary bin to the best hit, retaining only those merged bins if the primary bin and reference bin had at least 40% alignment fraction and an ANI>=95%.

13 mMAGs assigned to *Malassezia Globosa* (GCA_010232095.1) were identified (ANI>95%) and each primary and secondary bin aligned against the GenBank reference genome using minimap2 (–secondary=no -ax asm5). Aligned contigs were extracted using bedtools and visualised using circlize[44].

Eukaryotic species comparisons

To compute genome-to-genome comparisons of ANI, mash distance and amino acid identity (AAI), summary tables were obtained from NCBI GenBank for fungi, protozoa and plants. All Embryophyta entries were removed from the plant submissions. To compute comparisons between genomes, and avoid the computational overhead of all-vs-all comparisons, pairs of genomes were determined: in each clade (e.g. fungi, protozoa, plant) all representative species were identified first, which reduces the entries to a single entry per species. Each representative entry was then linked to up to five (subject to availability of entries) representative entries of the same genus but different species. Subsequently it was also paired with up to five (subject to availability of entries) representative entries of the same family but different genus. Finally for species to species comparisons the representative entry was paired with other non-representative species entries from NCBI GenBank. Pairs were only formed if the genome size discrepancy was smaller than 50% of the smaller entry and each deposited genome was larger than 5 Mbp of DNA. In total 2,445 genome pairs for fungi, 433 for protozoa and 115 plant pairs were determined for which ANI, Mash distance and AAI were to be computed. Thus for each pair the Mash distance was computed using mash dist -s 10000, the ANI was computed using dnadiff with default parameters and to compute AAI proteins were first predicted in each genome using GeneMark-ES which were used by compareM (comparem aai, default parameters).

Eukaryotic phylogenetic tree

To create a phylogenetic tree for all recovered representative MAGs, proteins were predicted with MetaEuk (version=4)[15] using the MetaEuk database bundled with the EukCC (version 2) database. Proteins were then annotated with HMMER (v 3.2.1)[45] using the same set of 49 widely-present single-copy PANTHER families (version 16) as EukCC's base set. The best matching protein was then aligned across all MAGs using FAMSA[46] (version 1.6.1) and trimmed with trimAl[47] (version 1.4.rev15, -gappyout). Alignments were then concatenated and low-complexity regions removed using trimAl (-gt 0.8 -cons 60 -w 3). IQ-TREE[48] (version 2.0.3) was subsequently used to build a maximum likelihood tree under the LG+R7 model. The tree was annotated and visualised in iTol (version 5)[49]. All MAGs were taxonomically annotated using CAT[50] (v. 5.1.2, CAT bins) using the NCBI nr database (fetched 14. Feb. 2020). Using Mash, we assigned each MAG to either a GenBank representative genome from the clades fungi, plant and protozoa fetched using the ncbi-genome-download tool (June 2021), or to 1151 eukaryotic and potentially-eukaryotic MAGs. This MAG collection consists of MAGs published by Delmont et al. 2018[12], Saary et al. 2020 as well as a subset of MAGs published by Delmont et al. 2020[7]. To create the subset we ran EukCC on all MAGs from Delmont et al. 2020 and de-repliacted all QS50 MAGS using drep (ANI 95, AF 40). We compared our MAGs with the top mash hit using dnadiff with default settings. MAGs are defined as novel if neither a MAG nor an isolate genome could be aligned with at least 40% AF and 95% ANI.

Eukaryotic Protein Comparison

Proteins in eukaryotic MAGs were predicted *de novo* using GeneMark-ES[51] with pygmes. We clustered all proteins together with UniRef90 (release 2021_03) using mmseqs2 linclust (easy-linclust $fasta $result tmp –threads 16 –min-seq-id 0.9 –cov-mode 1 -c 0.8, version: 602689c1). We then used hmmsearch (-E 1e-5 –cpu 2 –noali –tblout, version 3.2.1) to annotate all proteins using Pfam 34.0. Figures were created in R (version 4.1.1) using tidyverse[52,53].

Functional annotation of samples

We sketched all reads using sourmash compute (–track-abundance –scaled=2000 –merge). And all bacterial MAGs and all eukaryotic MAGs were sketched into their respective collection (sourmash compute –track-abundance –scaled 2000). We then computed containment using sorumash gather (k=31). An OTU table was constructed for all runs based on the median abundance if the fraction matched was at least 25%. We predicted proteins in all high-quality prokaryotic MAGs using prokka, and in all eukaryotic MAGs using GeneMark-ES with pygmes. Proteomes were then annotated using dbCAN2 [54] (CAZy) and kofamscan.py (KEGG). KEGG pathways were only retained if they were at least 90% complete. We removed KEGG modules matching keywords related to a single kingdom (e.g Bacteria, Eukaryotes, Plants etc.). dbCAN2 allows users to identify CAZy modules using HMMER, DIAMOND, and Hotpep. We only retained CAZys that were predicted by at least two of these methods. Subsequent analysis was then carried out using R and ggplot2.

Declarations

Data availability

MAGs have been submitted to ENA under the study accession PRJEB51083. Assemblies of the lichen metagenomes have been submitted under the accessions PRJEB50944 and PRJEB50945. EukCC version 2 is available on https://github.com/Finn-Lab/EukCC/.

Conflict of interest

No conflict of interest declared.

Funding

Work described in this publication was supported by the European Molecular Biology Laboratory core funds.

Acknowledgements

We thank Lorna Richardson for her editorial help with the manuscript.

References

1. Merino, N. *et al.* Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Front. Microbiol.* **10**, (2019).

2. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 1–11 (2020) doi:10.1038/s41587-020-0718-6.

3. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).

4. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

5. Kumar, V. *et al.* Ecology and Evolution of Marine Fungi With Their Adaptation to Climate Change. *Front. Microbiol.* **12**, (2021).

6. Louca, S., Mazel, F., Doebeli, M. & Parfrey, L. W. A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biol.* **17**, e3000106 (2019).

7. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv* 2020.10.15.341214 (2020) doi:10.1101/2020.10.15.341214.

8. Littman, R., Willis, B. L. & Bourne, D. G. Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. *Environ. Microbiol. Rep.* **3**, 651–660 (2011).

9. Kennedy, J., Marchesi, J. R. & Dobson, A. D. W. Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Appl. Microbiol. Biotechnol.* **75**, 11–20 (2007).

10. Vargas, C. de *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).

11. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* gr.228429.117 (2018) doi:10.1101/gr.228429.117.

12. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **3**, 804–813 (2018).

13. Deng, L., Wojciech, L., Gascoigne, N. R. J., Peng, G. & Tan, K. S. W. New insights into the interactions between Blastocystis, the gut microbiota, and host immunity. *PLOS Pathog.* **17**, e1009253 (2021).

14. Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* **46**, D726–D735 (2018).

15. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).

16. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144–6 (2014).

17. Saheb Kashaf, S. *et al.* Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nat. Microbiol.* **7**, 169–179 (2022).

18. Wiegand, S., Jogler, M. & Jogler, C. On the maverick Planctomycetes. *FEMS Microbiol. Rev.* **42**, 739–760 (2018).

19. Saary, P., Mitchell, A. L. & Finn, R. D. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol.* **21**, 244 (2020).

20. Tagirdzhanova, G. *et al.* Predicted input of uncultured fungal symbionts to a lichen symbiosis from metagenome-assembled genomes. *Genome Biol. Evol.* (2021) doi:10.1093/gbe/evab047.

21. Hibberd, D. J. Notes on the taxonomy and nomenclature of the algal classes Eustigmatophyceae and Tribophyceae (synonym Xanthophyceae). *Bot. J. Linn. Soc.* **82**, 93–119 (1981).

22. Andersen, R. A., Brett, R. W., Potter, D. & Sexton, J. P. Phylogeny of the Eustigmatophyceae Based upon 18S rDNA, with Emphasis on Nannochloropsis. *Protist* **149**, 61–74 (1998).

23. Fawley, M. W., Jameson, I. & Fawley, K. P. The phylogeny of the genus Nannochloropsis (Monodopsidaceae, Eustigmatophyceae), with descriptions of N. australis sp. nov. and Microchloropsis gen. nov. *Phycologia* **54**, 545–552 (2015).

24. Ganuza, E., Yang, S., Amezquita, M., Giraldo-Silva, A. & Andersen, R. A. Genomics, Biology and Phylogeny Aurantiochytrium acetophilum sp. nov. (Thraustrochytriaceae), Including First Evidence of Sexual Reproduction. *Protist* **170**, 209–232 (2019).

25. Hu, F., Clevenger, A. L., Zheng, P., Huang, Q. & Wang, Z. Low-temperature effects on docosahexaenoic acid biosynthesis in Schizochytrium sp. TIO01 and its proposed underlying mechanism. *Biotechnol. Biofuels* **13**, 172 (2020).

26. Morin, E. *et al.* Comparative genomics of Rhizophagus irregularis, R. cerebriforme, R. diaphanus and Gigaspora rosea highlights specific genetic features in Glomeromycotina. *New Phytol.* **222**, 1584–1598 (2019).

27. Błaszkowski, J., Kozłowska, A., Niezgoda, P., Goto, B. T. & Dalpé, Y. A new genus, Oehlia with Oehlia diaphana comb. nov. and an emended description of Rhizoglomus vesiculiferum comb. nov. in the Glomeromycotina. *Nova Hedwig.* 501–518 (2018) doi:10.1127/nova_hedwigia/2018/0488.

28. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

29. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).

30. Balabanova, L. A. *et al.* Polysaccharide-Degrading Activity in Marine and Terrestrial Strains of Mycelial Fungi. *Russ. J. Bioorganic Chem.* **44**, 431–437 (2018).

31. Faten, A. M. & Abeer, A. A. E. A. Enzyme activities of the marine-derived fungus Alternaria alternata cultivated on selected agricultural wastes. *J. Appl. Biol. Sci.* **7**, 39–46 (2013).

32. Balabanova, L., Slepchenko, L., Son, O. & Tekutyeva, L. Biotechnology Potential of Marine Fungi Degrading Plant and Algae Polymeric Substrates. *Front. Microbiol.* **9**, 1527 (2018).

33. Robert, C., Chassard, C., Lawson, P. A. & Bernalier-Donadille, A. 2007. Bacteroides cellulosilyticus sp. nov., a cellulolytic bacterium from the human gut microbial community. *Int. J. Syst. Evol. Microbiol.* **57**, 1516–1520.

34. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

36. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

37. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

38. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

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

40. Kalvari, I. *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).

41. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).

42. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

43. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

44. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

45. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).

46. Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.* **6**, 33964 (2016).

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

48. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

49. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab301.

50. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).

51. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).

52. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

53. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).

54. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

Figures

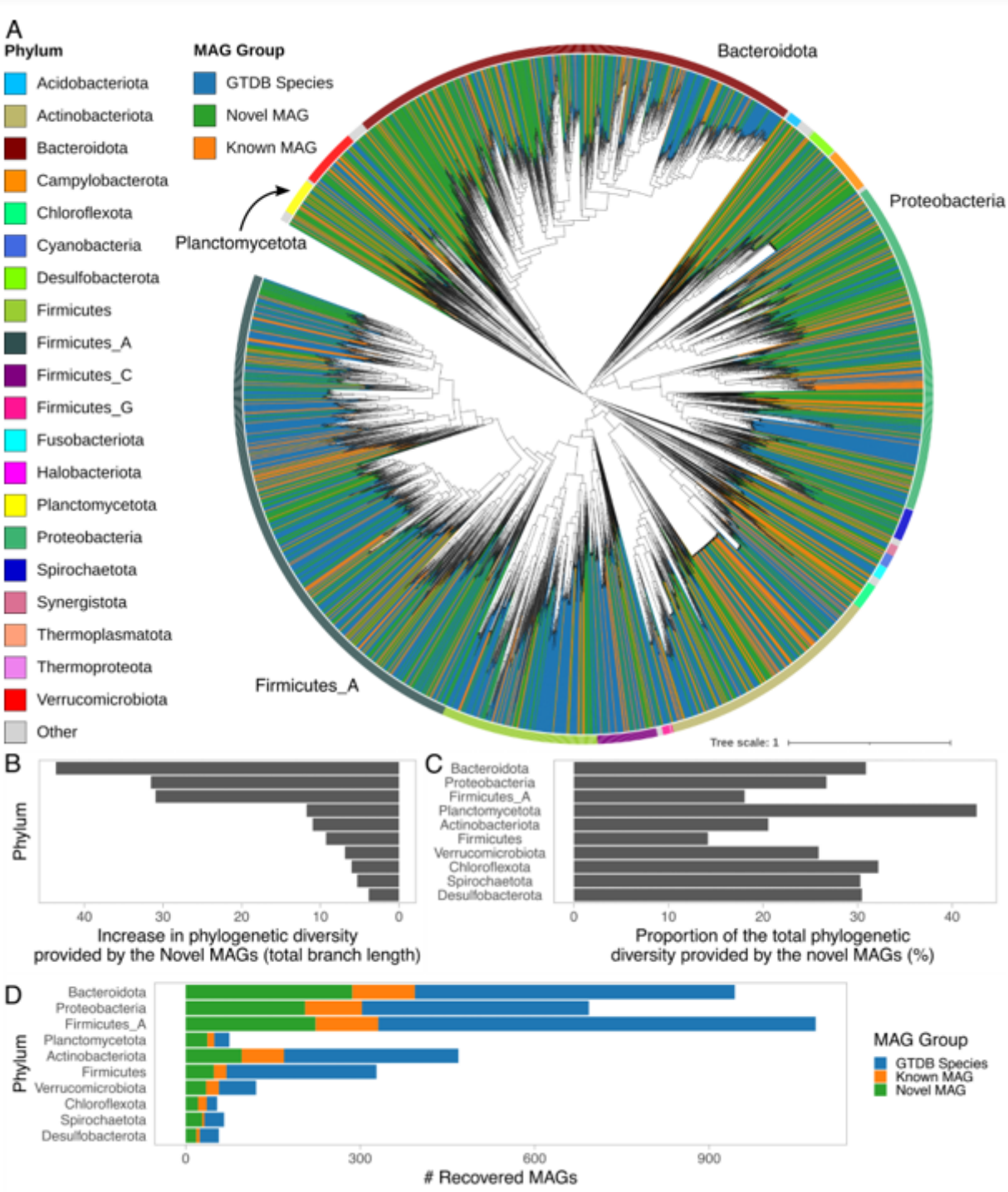

Figure 1

Bacterial MAGs. A) Phylogenetic tree based on the GTDB-Tk alignment of all bacterial high quality MAGs. Annotated with the GTDB phylum for the 20 largest phyla. MAGs of novel species are colored in orange. **B)** The ten phyla with the most gained total branch length, as contributed by the novel MAGs, are shown. The increase in total branch length per phylum is largest for Bacteroidota, which are found in the human gut, engineered samples and the aquatic biome. **C)** The relative amount of contributed branch length per clade is shown here and varies for these phyla between 10 and over 40% of the clade branch length. **D)** Number of MAGs in each of the clades, colour coded for GTDB species, known MAGs and novel MAGs.

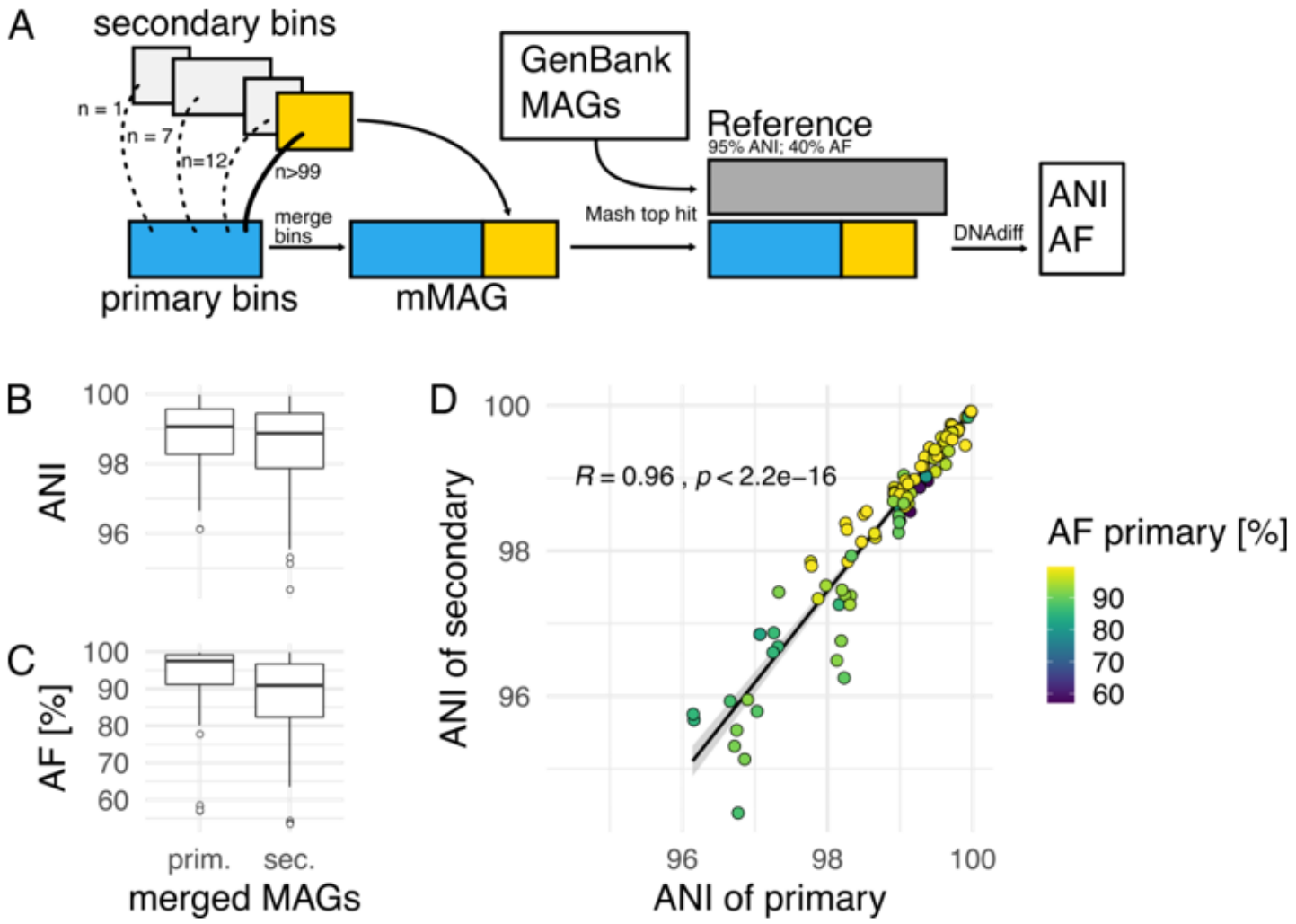

Figure 2

A) We identified linked secondary bins (at least 100 paired read links) for incomplete bins of at least 50% completeness (primary bin). Merging bins *in silico* we retained only such combinations where the gained completeness was at least 10% and five times higher than the increase in contamination. This method resulted in 131 mMAGs. Using mash, we assigned the best reference MAG or GenBank genome, including manually refined genomes from Delmont et al.[7], to each primary bin and used mummer DNAdiff to compute the alignment fraction and ANI between the assigned reference and the primary and secondary bin. We retained only comparisons if the primary bin shared at least 95% ANI and 40 % AF with the reference (107/131). **B)** Comparing the ANI between the primary and secondary bin we see no substantial difference. While secondary bins are smaller they still all have an ANI of over 95%, meaning they are the same species as their assigned reference genome. **C)** The aligned fraction of secondary bins is slightly smaller but still on average above 90%. **D)** looking at the ANI of primary and secondary bins we note that they correlate with a pearson correlation of .95. It is also observable that high ANI over 98% of the primary bin links to a high aligned fraction of both the primary (shown) and the secondary (not shown).

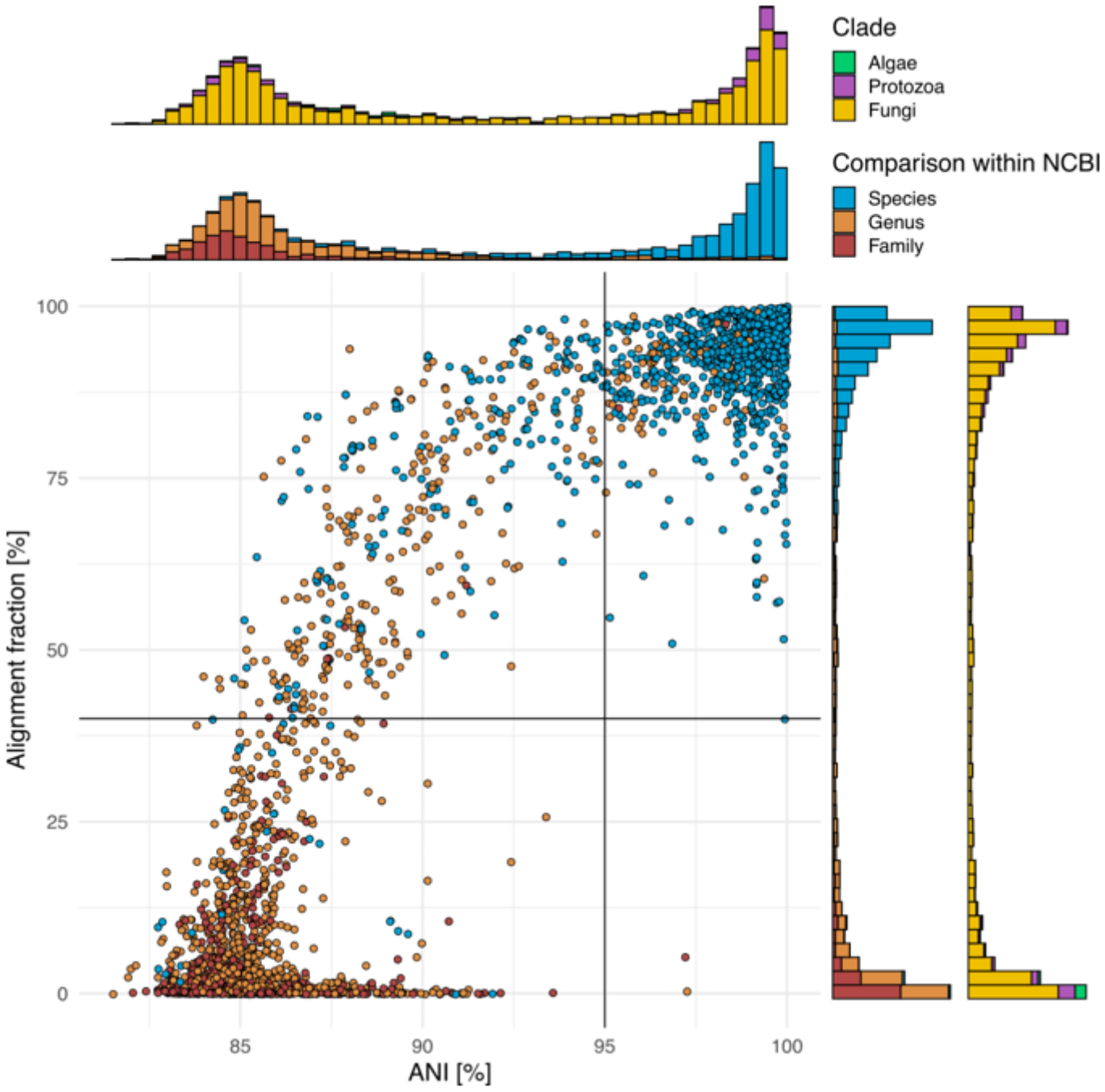

Figure 3

Marginal plot showing the Aligned Fraction (AF) and the Average Nucleotide Identity (ANI) for 2,118 comparisons in fungi (n=1795), protozoa (n=253) and algae (n=70). Both measurements show a bimodal distribution. Indicated is the 40% AF and 95% ANI thresholds used to dereplicate MAGs of eukaryotic origin at the species level.

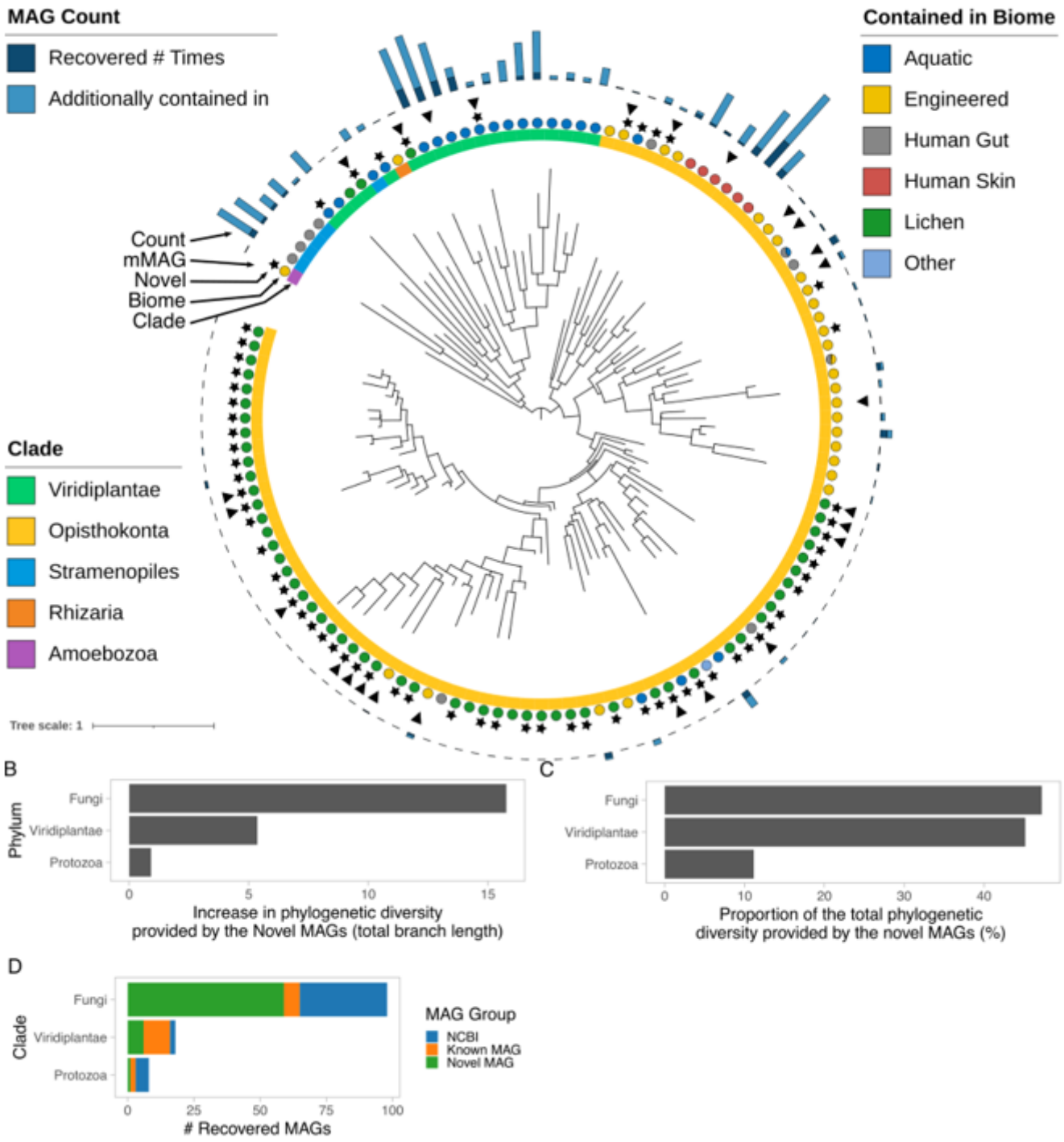

Figure 4

Phylogenetic species tree of the recovered eukaryotic MAGs. (A) Tree based on 71 marker genes found in almost all MAGs as a single copy. After aligning proteins using FAMSA the tree was constructed using IQtree under the LG+R7 model, we rooted the tree between fungi and all other clades, which was done for visualisation purposes only. The first annotation track shows the phylogenetic clade the genome belongs to. More than half of all MAGs (n=98) are of fungal origin. The next annotation track shows the biome in which these MAGs can be found, noticeably most MAGs are found in a single biome. Two MAGs were found in two biomes, with the second biome making up more than ¼ of all detected runs. The stars

indicate those MAGs that are newly reported here while the triangle indicates those MAGs created using bin merging. The outer bar chart shows the number of samples that contained the MAG and the number of times we recovered that MAG with a QS50 score. **(B+C)** We computed the total and relative proportion of branch length added by the inclusion of novel MAGs compared to species previously described. For protozoa over 10% of novel branch length is contributed by the novel MAGs from this study, whereas for fungi and viridiplantae over 40% of branch length is added. **(D)** counts of MAGs matched to either an NCBI reference species, a known MAG or deemed novel.

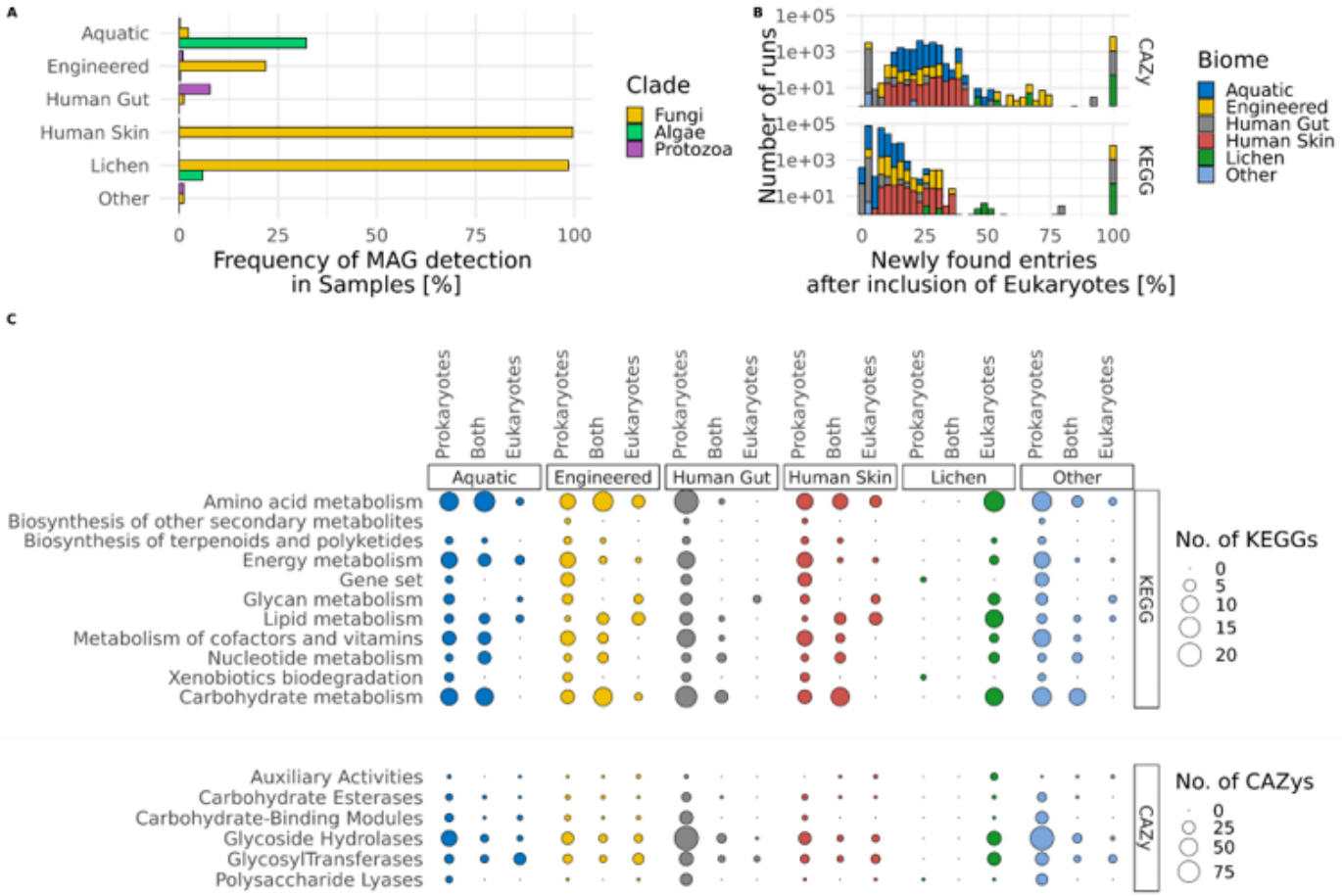

Figure 5

KEGG Modules gained when including eukaryotic MAGs in KEGG prediction. Using our MAG catalogue we determined the community composition of all samples using sourmash coverage of at least 25%. In each metagenome we subsequently predicted KEGG modules and CAZYs. **(A)** Different compositions of eukaryotes are identified in different biomes. In virtually all samples of human skin and lichen at least one fungus from this catalogue was identified. In the aquatic biome algae are found in more samples than fungi. **(B)** The fraction of novel metabolic modules found after inclusion of eukaryotes in addition to prokaryotes shows, that especially for the human skin, the ocean and most engineered samples between

10 and 50% novel functions can be identified. Lichen communities are small and due to low coverage in some samples only a single fungus was detected, leading to an overestimate of novelty in this analysis. **(C)** Metabolic modules (KEGGs and CAZy) are shown in rows. Columns show the number of functions unique to prokaryotes, eukaryotes and those found in both for each biome. In agreement with (B) most functions are either provided by the prokaryotic fraction or found in both. Notable additions to the metabolic potential can be seen in the aquatic, engineered, human gut and human skin biome for Glyocyl Transferases (GT).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx