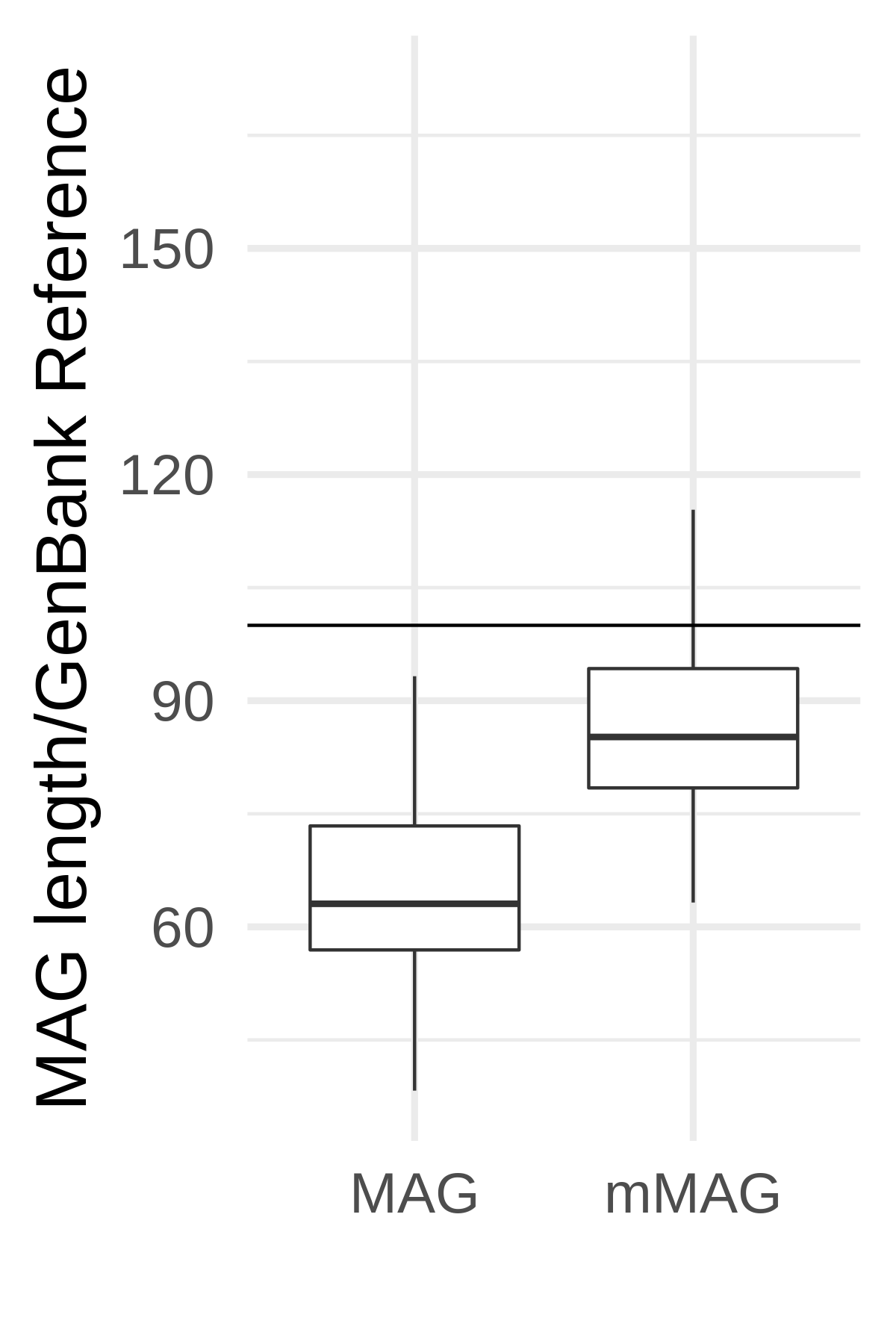
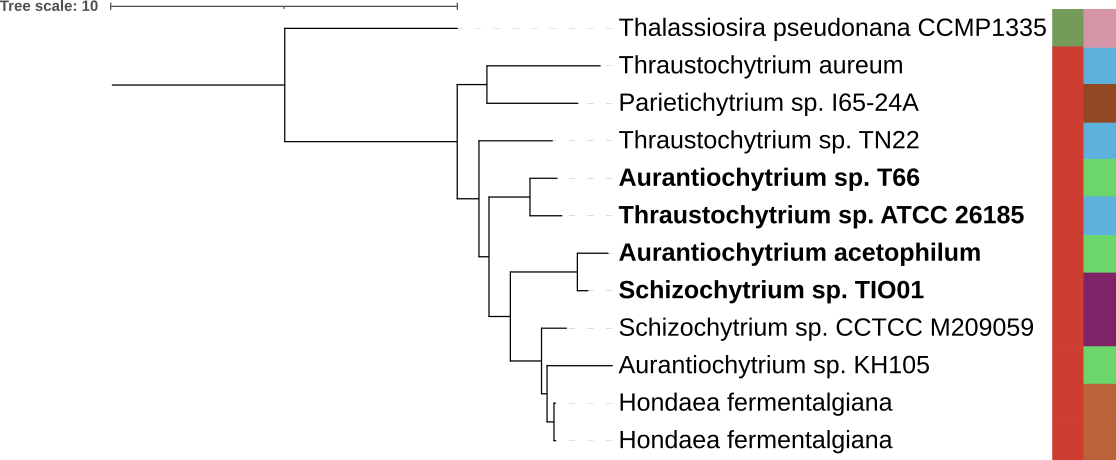
# Supporting information for “Large-scale analysis reveals the distribution of novel cellular microbes across multiple biomes and kingdoms”

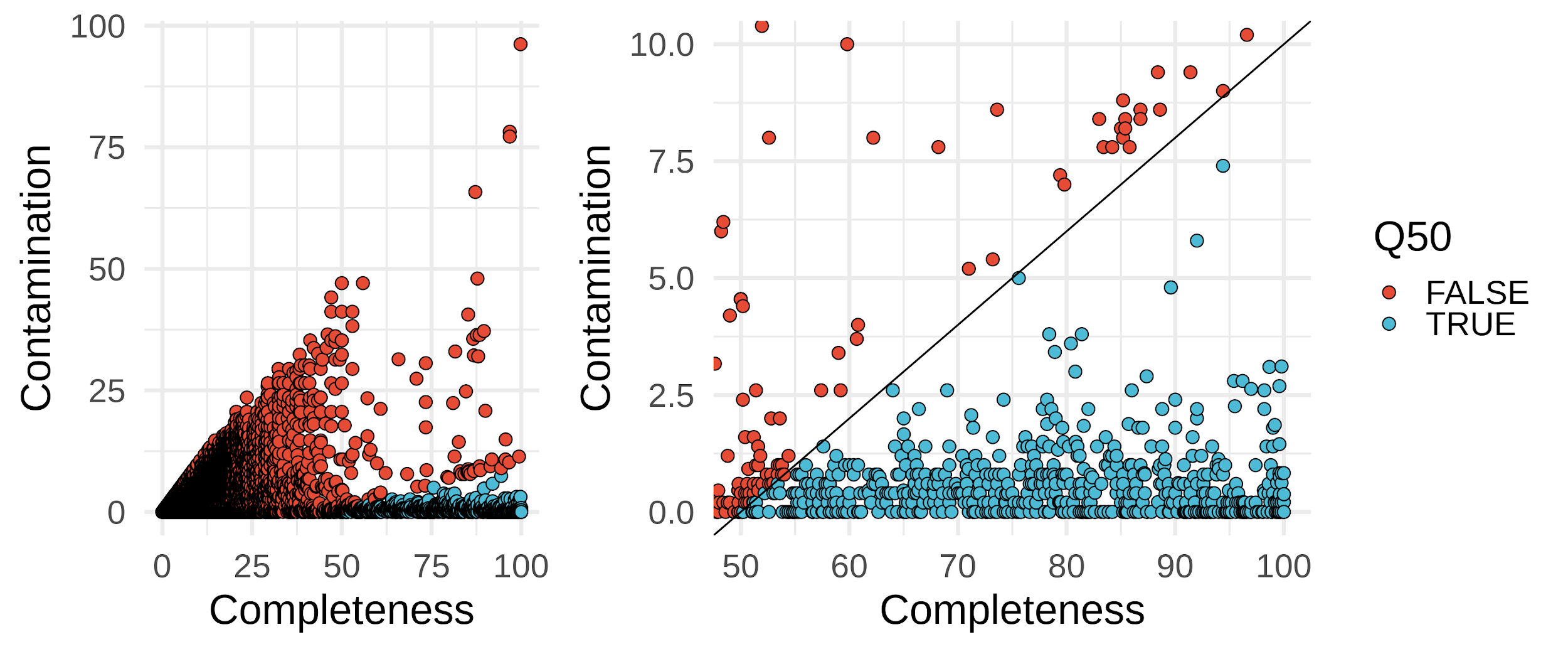
Paul Saary1, Varsha Kale1, Robert D. Finn\*1

1European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK \*Correspondence should be addressed to Robert D. Finn ([rdf@ebi.ac.uk](mailto:rdf@ebi.ac.uk))



**Figure S1:** Comparing Primary bins and mMAGs to the assigned GenBank NCBI reference genome: mMAGs are closer to the already reported genome size for that reference, and are on average not larger than the reference entries.

**Figure S2:** We used IQtree to build a phylogenetic tree for the Thraustochytriaceae family, using all entries in Genbank. To construct the tree 22 PANTHER (version 16.0) profile families were used to identify single copy genes in the genomes. We added the phylum (first, vertical annotation track) and the genus (second annotation track). In this tree we can see that *S. Sp. TIO01* and *A. acetophilum* cluster together. This supports our finding that the two genomes are closely related. In additiont, the clade also contains another anomaly with T. Sp. ATCC 26185, which seemingly closely related to A. sp 66. Aligning these genomes against another we get an alignment fraction of 42% and 99.8% with an ANI of 99.8% suggesting that the Thraustochytrium sp. ATCC 26185 genome is completely contained in the Aurantiochytrium sp. T66 genome.



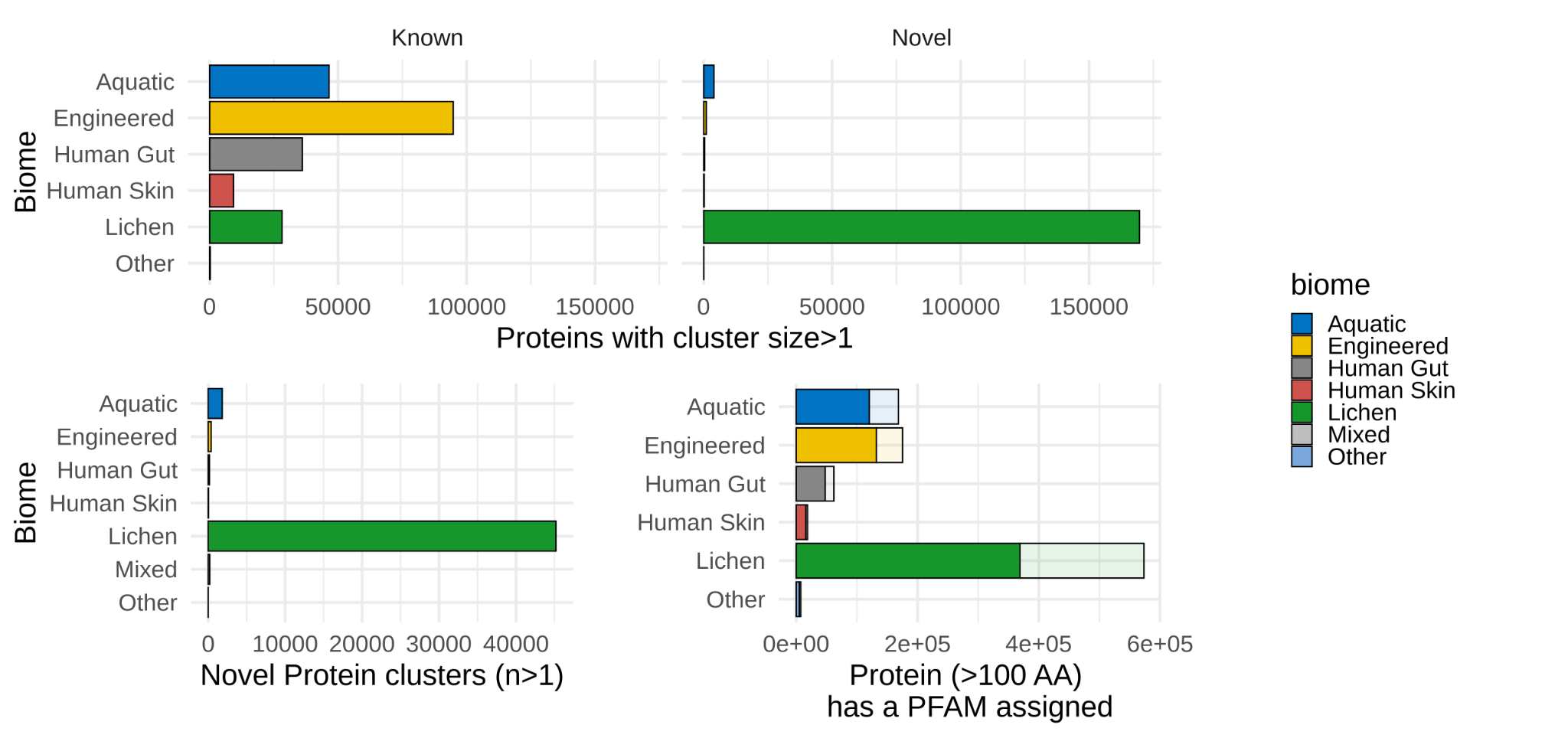
**Figure S3: Quality scores of all bins analysed using eukcc.** Using EukCC 2 we analysed 317,195 bins. For 49,130 we obtained a quality score from EukCC. We retained 504 bins with a quality score (QS50) of completeness -5\*contamination >= 50 (blue).

# 

**Figure S4**: 13 mMAGs aligned against the *M. Globosa* reference genome and contigs colored by primary (blue) and secondary (yellow) bin. Generally primary and secondary bins are not defined as distinct genomic regions of the reference genome but are dispersed across the entire reference. While some tracks show similarity no pattern across al 13 mMAGs emerges.

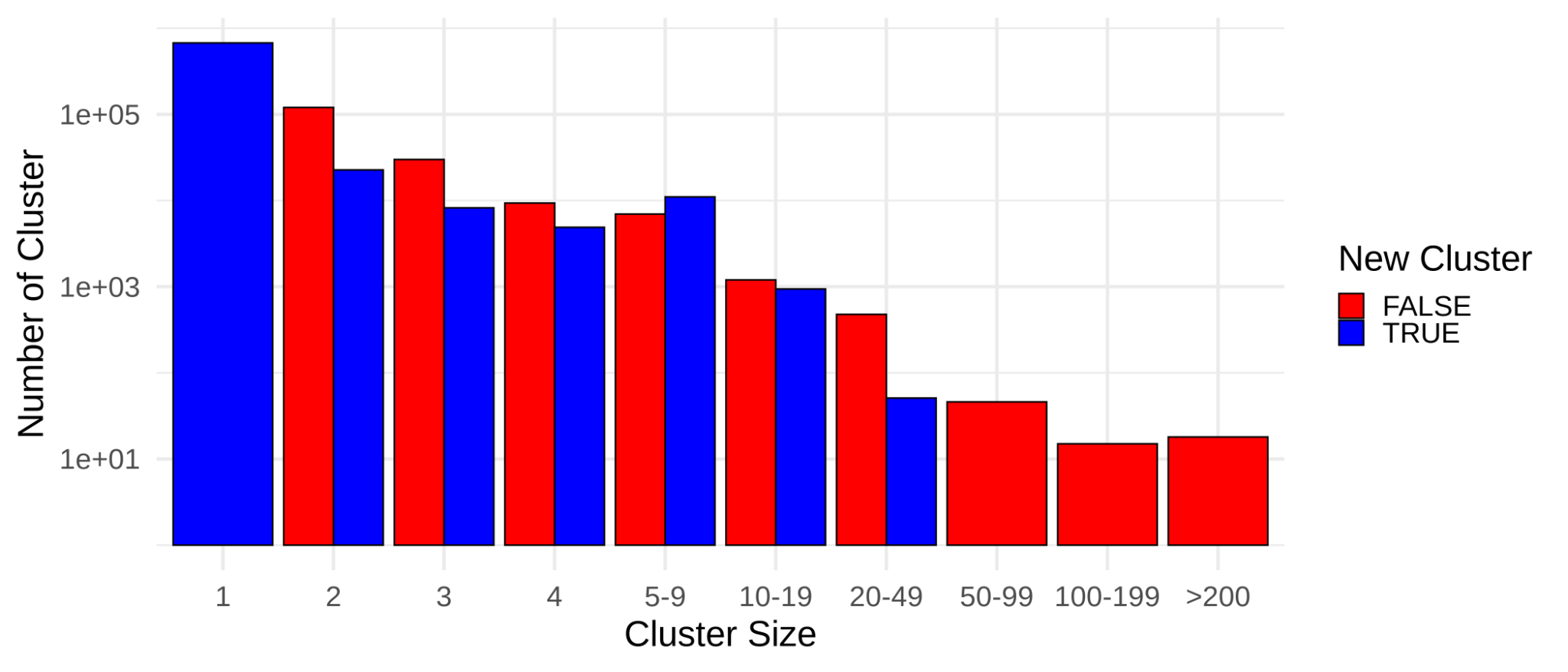


**Figure S5** A umap based on the functional modules identified in samples from different biomes clusters samples based on their biomes.

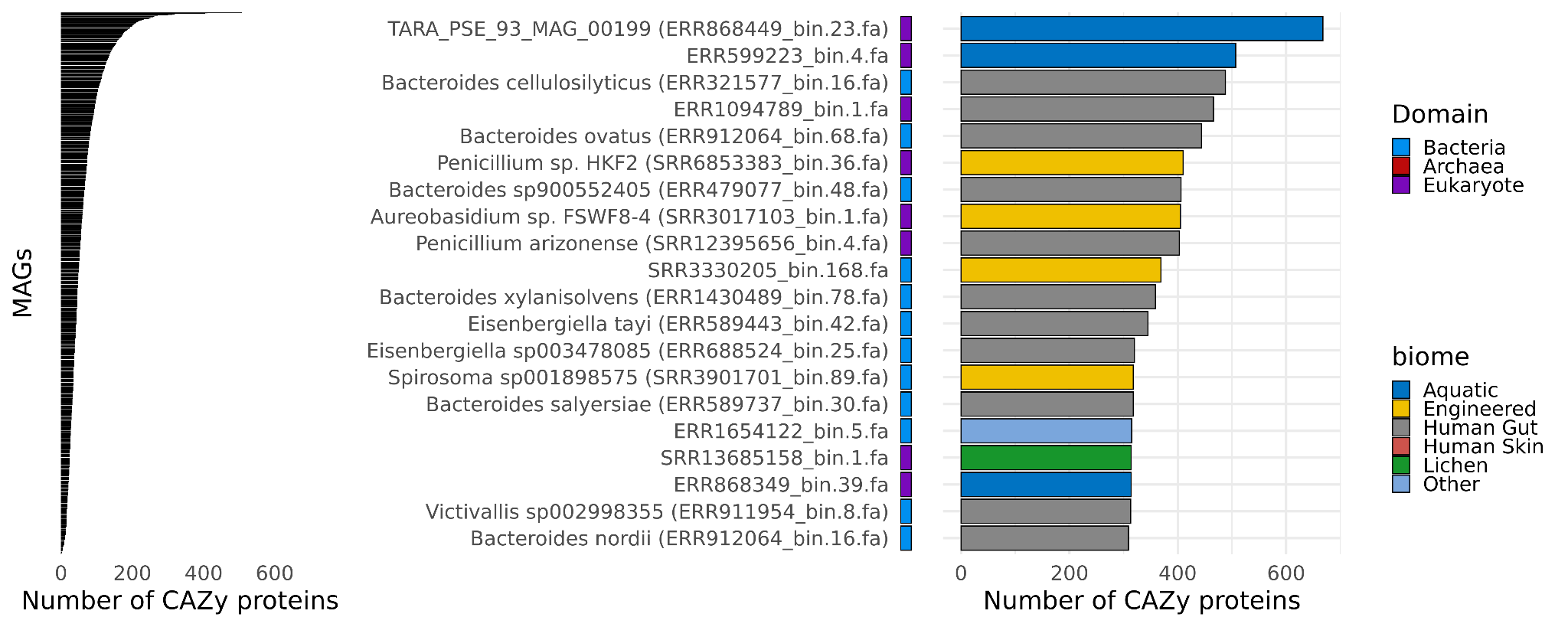


**Figure S6:** EukaryoticProteins clustered with UniRef90. Top left: Proteins clustered with UniRef90 that cluster together with at least one UniRef90 sequence, shown by biome. Top Right: Clustered proteins that do not cluster together with a UniRef90 sequence.

Bottom left: Protein clusters per biome. Bottom right: Proteins annotated with PFAM per biome.



**Figure S7:** Protein clusters by size and colored by novelty. Red clusters do contain known sequences from Uniref90. Blue clusters are comprissed entirely from new sequences. Known clusters of size 1 are excluded for clarity.



**Figure S8:** Ranking MAGs by their total number of identified CAZy proteins demonstrates an enrichment in a small number of MAGs (left) . The top 20 (right) MAGs contain 8 eukaryotes and 12 bacteria. Most bacteria are from the Human Gut. The MAGs with most CAZys are fungi from aquatic biomes, the first has been identified by Delmont et al. 2018 already, the second we report for the first time here.