Extracting microbial genomes from coral genomic data

MSc in Bioinformatics

Master Thesis

Author: Laura Moscat Martínez

Thesis supervisor: Javier del Campo

Academic tutor: Jaime Martinez Urtaza

Submission date: July 2024

Faculty of Biosciences

Department of Genetics and Microbiology

ACKNOWLEDGEMENTS

I would like to thank Javier del Campo, the project supervisor, for his dedication and disposition to help me, as well as for dedicating part of his time, advising me and guiding me until the last moment.

I also thank Anthony Bonacolta for developing the MAGs Factory scripts, without which I would not have been able to carry out this work. I thank him for the time he took to meet with me remotely from Florida to guide me through the MAGs extraction process.

In addition, I thank Professor Miquel Àngel Senar for helping me with access to the UAB cluster and for his immediate responses to assist me with subsequent problems. I also thank Professor Javier Panadero for teaching me to program from zero and making it seem simple, as well as for his dedication in the explanations, which only ended when we really understood what he was explaining.

Finally, I would like to thank Francesc Torradeflot, administrator of the Port d'Informació Científica, for his quick responses to help us with the problems that occurred during the execution of the scripts.

Abstract

Although the taxonomy of coral microbiomes is well documented, their functional role in coral ecosystems remains poorly understood. Access to the genomes of microbial communities can enhance our understanding of their metabolism, capabilities, and interactions with corals and their environment. We downloaded raw coral (meta)genomes from the NCBI-SRA database, including genetic information from the coral host but also from its associated microbiome. Microbiome information is commonly discarded as contaminants and never used. Employing the MAGs Factory pipeline, we extracted eukaryotic and prokaryotic genomes from the downloaded coral (meta)genomes. This workflow assembles raw reads into contigs and groups them into Metagenome-Assembled Genomes (MAGs). To filter complete MAGs, we used BUSCO's completeness metrics, the implications of which are discussed. We obtained 36 complete eukaryotic MAGs, all classified as corals, and 31 prokaryotic MAGs classified within the Pseudomonadota and SAR324 phyla. We placed them in a broad phylogenomic context using for that the bacterial genomes available in the GTDB reference database. We identified bacterial species previously not known to be associated with corals. The availability of their genomes is crucial for advancing our understanding of their roles within the coral holobiont.

INDEX

1. INTRODUCTION

Corals are the structural basis of many benthic habitats named coral gardens. Coral's three-dimensional structure creates a complexity that serves as refugee or nursery to many other associated species (Miller et al., 2012). These habitats contain a wide diversity, included many species of commercial interest, and play a significant role in biogeochemical cycles (Bo et al., 2015; Chimienti et al., 2020; Terzin et al., 2021). Moreover, coral gardens provide crucial ecosystem services to society (Brenner et al., 2010), as tourism attraction, food supply or water filtering.

Corals belong to the phylum cnidaria. They live in colonies in which each individual is called polyp (Fig. 1.1) and is function-specialised. In most cases, they are covered by a self-made skeleton, made of organic compounds, such as gorgonin or calcium carbonate. Corals thrive in a wide range of conditions, from shallow coastal waters (Agarwal et al., 2024) to greater depths that can reach the 3000 m (Freiwald & Roberts, 2005; Lumsden et al., 2007), where temperature and pressure are extreme. They are also found in tropical (Andrello et al., 2022) and polar latitudes (Pierrejean et al., 2020).

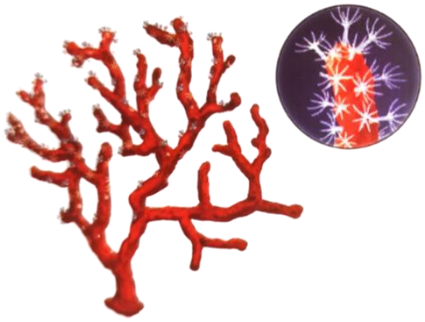

Figure 1.1. Illustration of the coral *Corallium rubrum* and its polyps. The top right circular illustration represents a zoom of the illustration. Each white arborescent element is a polyp. (Ballesteros and Sagarra, 2015).

Each coral is a holobiont, meaning it is not only the cnidarian species but also includes an associated microbiome (Hou et al., 2015). These microorganisms play an important role in the metabolism of holobionts (Siboni et al., 2008) and can be essential, as seen in the case of coral bleaching (Van Oppen & Lough, 2018). Massive coral mortalities have occurred due to the loss of photosynthetic coral-associated microorganisms caused by rising temperatures (Coker et al., 2012; Van Oppen & Lough, 2018; Lishchenko et al., 2024).

Phylogenetic surveys have shown that the dominant coral microbiome organisms belong to the taxon Proteobacteria (particularly Gamma- and Alphaproteobacteria) as well as Actinobacteria, Bacteroidetes (especially Flavobacteria), and Cyanobacteria (Blackall et al., 2015; Littman et al., 2009; Rohwer et al., 2002; Sunagawa et al., 2010). While the diversity of coral-associated microorganisms is well-documented (Bourne et al., 2016; Mohamed et al., 2023), the specific roles of each microorganism in the holobiont's functioning remain insufficiently studied (Glasl et al., 2020). Understanding the genomes of these associated microorganisms is crucial to know their functions and interactions with the corals' host and the environment.

To obtain the genomes of the coral microbiome, A. Bonacolta and J. del Campo (colleagues from the research group), developed the MAGs Factory pipeline (Fig. 1.2). It is based on the assumption that a raw coral genome is actually a metagenome, as it includes the genomes of its microbiome. This is why we refer to coral genomes as (meta)genomes. The MAGs Factory pipeline groups the sequences of a raw coral (meta)genome into Metagenome Assembled Genomes (MAGs), which belong to the holobiont microbiome. Since we obtained the coral (meta)genomes from the Sequence Read Archive of the National Center for Biotechnology Information (NCBI-SRA), we have been able to reuse samples and leverage previous sapling efforts, making this a low-cost project.

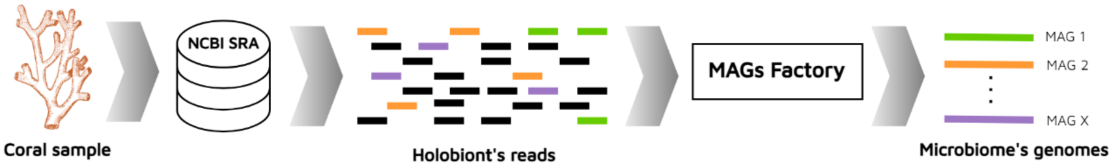

Figure 1.2. General process of the MAGs extraction. Raw coral (meta)genomes are uploaded to NCBI-SRA database. Then we downloaded them. Using the MAGs Factory pipeline, we assembled the raw reads into contigs, and we grouped them into MAGs.

2. OBJECTIVES

Coral (meta)genomes were randomly split in 12 batches. Our objectives are:

- Obtain MAGs processing coral meta(genomes) from batches one and two with the MAGs Factory pipeline.
- Assign taxonomy to the resulting MAGs and place them in a phylogenetic tree.
- Build a script that integrates the MAGs Factory steps to automate the process of extracting MAGs.

3. MATERIALS AND METHODS

3.1 Data obtention

On February 2022, we downloaded all (meta)genomes from the NCBI-SRA database that matched with “coral genome” search (Fig. 3.1.A). We obtained 6000 coral (meta)genomes that belong to 215 BioProjects of 50 different coral species. As we acquired the (meta)genomes from NCBI-SRA database, we get the raw sequencing reads.

3.2 MAGs Factory: from metagenomes to genomes

We did not remove the coral genomes from the raw (meta)genomes because it would’ve required a super clean coral genome specific to each species we processed. Additionally, most coral genomes are not contamination-free and thus extracting with them would extract microbe reads too.

Adapter trimming (fastp)

We trimmed reads adapters using the Adapter Trimming module of fastp (S. Chen et al., 2018) for paired-end data (Fig. 3.1.B). It searches for the overlap of each pair and considers the non-overlapped bases as adapter content to remove. The advantage of this tool is that it can trim adapters with even only one base in the tail.

Metagenomes assembly (megahit)

Clean reads were assembled into contigs using Megahit (D. Li et al., 2015), a NGS *de novo* assembler (Fig. 3.1.C). Megahit uses succinct *de Bruijint* graphs (SdBG; Bowe et al., 2012) to represent k-mers overlapping in a compressed way. Then, the contigs are generated by traversing paths through the graph. Megahit follows a multiple k-mer size strategy, so multiple SdBGs are generated iteratively from a small k to a large k. In each iteration, corrections are made. While small k-mer SdBGs are useful to filter erroneous edges and fill gaps, large k-mers SdBGs allow resolving repeats.

SdBG construction is done with a parallel algorithm that exploits the parallelism of a graphics processing unit (GPU, CUDA-enabled) by adapting the recent BWT-construction algorithm CX1 (Liu et al., 2014).

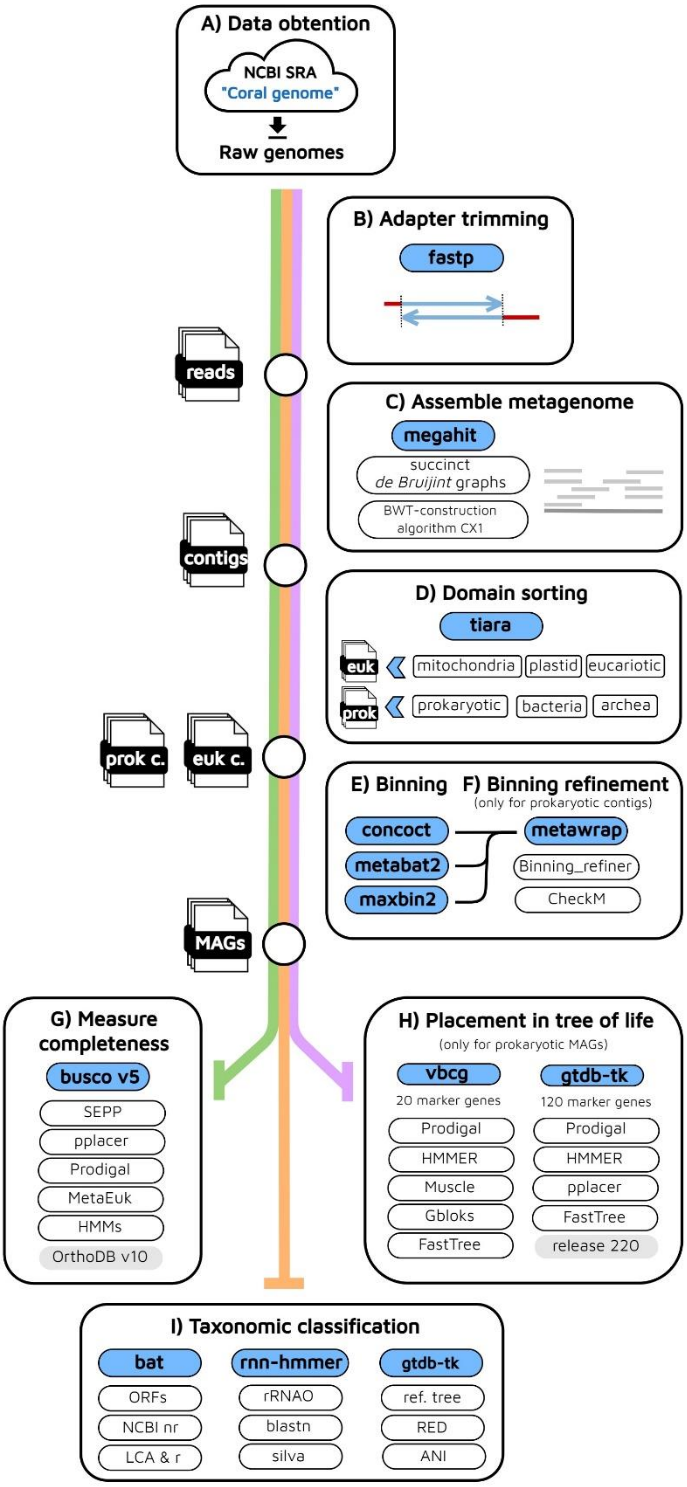

Figure 3.1. Workflow diagram. Prok c.: prokaryote contigs; euk c.: eukaryote contigs.

Domain sorting (tiara)

We separated the eukaryotic contigs from the prokaryotic ones using Tiara (Karlicki et al., 2022) (Fig. 3.1.D), a deep-learning-based tool. Contigs splitting is done in two steps with two distinct two-layer feed-forward neural network architectures. First, contigs are separated in six classes: bacteria, archaea, prokaryote, eukaryote, organelle or unknown. Second, contigs classified as organelle are differentiated between mitochondria, plastids or unknown. For the following steps, we considered archaea, bacteria and prokaryote contigs as prokaryotic and eucaryote, mitochondria and plastid contigs as eucaryotic.

The Tiara classification process is the same in both steps. Initially, contigs are divided into 5kb fragments, each of which is classified into a class based on the training dataset. Then, an average probability for each class is calculated for each contig, based on their fragment's classification. If the probability of a class is higher than a threshold (we used the default value 0.65), the contig is assigned to that class. However, if no individual probability exceeds the threshold, the contig is classified as unknown, unless the cumulative probabilities of bacteria and archaea together surpass it, then is classified as prokaryote.

Binning (metaWrap)

The binning step consists of grouping the contigs that belong to the same genome in the same 'bin', so each 'bin' contains a MAG. This task is especially difficult in metagenomics studies because there’s not a reference genome that could be used as a template to assemble the contigs.

Prokaryotic contigs binning was conducted in two stages. Initially, due to the absence of consensus on the "best" binner, three binning versions were created using CONCOCT (Alneberg et al., 2013), Metabat2 (Kang et al., 2019), and MaxBin2 (Wu et al., 2016) (Fig. 3.1.E). Subsequently, the three versions were refined into more complete bins running the metaWRAP-Bin_refinement module (Uritskiy et al., 2018) (Fig. 3.1.F). However, for the eukaryotic contigs binning, only the first stage was applied because previous tests indicated that the metaWRAP-Bin_refinement module required a significant amount of time for eukaryotic contigs. Despite CONCOCT, Metabat2 and Maxbin2 being independent tools, we executed them using the metaWRAP-Binning module.

CONCOCT uses Gaussian mixture models (GMMs) (Fraley, 1998) to cluster contigs into genomes based on sequence composition (k-mer frequencies) and coverage across multiple samples. It determines the number of different genomes performing a model selection using the Bayesian information criteria (BIC) (Fraley, 1998).

MaxBin2 estimates the number of bins using single-copy marker genes. It also calculates tetranucleotide frequencies and contigs coverage levels. With all this information, an expectation-maximization (EM) algorithm calculates the probability that a contig belongs to a bin and makes the bins.

MetaBAT2 computes for each contig pair, the abundance distance probability (ADP) and the tetranucleotide frequency distance probability (TDP), which is calculated from a distribution model from 1,414 reference genomes. Both probabilities are then combined, and the resulting distance is used to build a distance matrix, based on which, bins are iteratively made.

The metaWRAP-Bin_refinement module produces a different bin set based on the original ones using Binning_refiner (Song & Thomas, 2017). Then, CheckM (Parks et al., 2015) identifies single-copy genes expected in the genome of each bin's taxonomy, both in the original set and the refined bins. Based on this data, completion[1] and contamination[2] metrics are computed. We set thresholds of 50% for minimum completion and 10% for maximum contamination to determine output bins.

Completeness estimation (BUSCO)

Once we had the MAGs (or bins), we estimated their completeness using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Manni et al., 2021) (Fig. 3.1.G). It encompasses 193 odb10 (OrthoDB v10) data sets. When a sequence is provided, a phylogenetic approximation is performed to select the most appropriate dataset. This process is carried out using the 'auto-lineage' option, which places the input sequence within a set of precomputed phylogenetic trees using SEPP (Warnow, 2015) and pplacer (Matsen et al., 2010). Subsequently, if the input sequence is classified as bacterial, archaeal, or viral, Prodigal (Hyatt et al., 2010) performs gene prediction. For eukaryotic sequences, MetaEuk (Levy Karin et al., 2020) is used for this task. Finally, HMM (Eddy, 2011) compares the predicted genes with those in the dataset, classifying matches as unique, duplicated, fragmented or missing. Then, calculates the completeness by calculating the ratio of unique and duplicated matches to the total number of genes in the assigned set.

We defined a criterion to select the complete MAGs based on the BUSCO completeness metric. Prokaryote complete MAGs were those with a completeness above 70% and eukaryote ones were those with a completeness greater than 50%.

3.3 MAGs taxonomical classification

We classified complete MAGs with three different methods to compare the classification results (Fig. 3.1.H).

[1] Completion: “percentage of expected single-copy genes that is found in a bin” (Uritskiy et al., 2018)

[2] Contamination: “percentage of single-copy genes that are found in duplicate” (Uritskiy et al., 2018)

Bin Annotation Tool (BAT)

Initially, BAT predicts Opening Reading Frames (ORFs) for each MAG using Prodigal. Next, DIAMOND (Buchfink et al., 2015) queries the predicted ORFs protein translations to the National Center for Biotechnology Information (NCBI) non-redundant protein database (nr) (Sayers et al., 2022). Then, each ORF is classified by determining the Last Common Ancestor (LCA) of all hits within a specified range of the top hit (parameter r), and the top-hit bit-score is registered for the subsequent step. Finally, each MAG is classified based on all its ORFs' classifications and its corresponding bit-score, evaluating total bit-score evidence to assign each MAG to a taxon.

rRNA-based classification (rRNA-BC)

Using rrn-hmmer, a tool developed by members of our group, we extracted the rRNA operons from each MAG. This tool utilizes an hmm profile constructed using rRNA sequences to identify similar patterns in the provided MAG rRNA sequences. It's important to note that the hmm profile differs between eukaryotes and prokaryotes due to variations in their rRNA sequences. The tool also filters out rRNA sequences that are shorter than 4500 bases in eukaryotes and 3500 bases in prokaryotes.

For MAGs classification, we submitted prokaryotic MAGs rRNA sequences to SILVA, a specialized database for rRNA. SILVA compared these sequences against its database, as well as those in the European Nucleotide Archive (ENA), Genome Taxonomy Database (GTDB), All-Species Living Tree (LTP), and Ribosomal Database Project (RDP), providing classifications from each database. The MAGs rRNA sequences of eukaryotic bacteria were submitted to BLAST (Basic Local Alignment Search Tool), which compared them with the sequences in the NCBI database and provided taxonomic classifications accordingly.

GTDB-Tk *classify workflow*

The GTDB-Tk (Chaumeil et al., 2020) classification tool places MAGs in a reference tree and classifies each MAG based on the tree's topology. When rank assignments are ambiguous, RED (Parks et al., 2018) value is used to resolve them. In addition, species assignments are established using ANI, as calculated with FastANI (Jain et al., 2017).

To place the MAGs in the reference tree, GTDB-Tk first predicts and identifies genes using Prodigal and HMM. Subsequently, each genome is assigned to the domain with the highest proportion of identified marker genes. The selected domain-specific markers are aligned with HMMER, concatenated into a single multiple sequence alignment, and trimmed with the ~5000-column bacterial or archaeal mask used by GTDB-Tk. Then, FastTree (Price et al., 2009, 2010) with pplacer places the genomes into the domain-specific reference tree, which contains the reference genomes of the GTDB-Tk release 220.

3.4 MAGs phylogenomic placement

As we didn’t find any eukaryotic microorganism, we only did a phylogenomic placement for prokaryotic MAGs.

We built two phylogenomic trees with the obtained MAGs using two different tools, VBCG (Tian & Imanian, 2023) and GTDB-Tk (Fig. 3.1.I). While VBCG tool builds the tree only using the provided MAGs and just 20 bacterial core genes, GTDB-Tk tool also includes reference genomes from the GTDB database and uses 120 bacterial and 122 archaeal marker genes.

GTDB-Tk *de novo* workflow

GTDB-Tk places provided genomes within the reference tree using *de novo* workflow. It places them as explained before (3.3: MAGs taxonomical classification; GTDB-Tk classify workflow)**.** Given that our bacterial tree comprised 107,262 sequences and it was very difficult to edit, we repeated the procedure, this time filtering the reference genomes aligned with HMM and selecting those closely placed to the provided MAGs.

VBCG

VBCG first uses Prodigal to predict the gene and protein sequences of the provided MAGs. Then, HMM is used to identify the 20 validated bacterial core genes (VBCG). Muscle (Edgar, 2004) aligns them and Gblocks (Castresana, 2000) select conserved blocks. After, alignments are concatenated and MAGs that have four or less unique marker genes are removed. Finally, FastTree builds the tree with the concatenated alignments.

3.5 MAGs Factory script

We automated the MAGs extraction in a Python script (Code S1). It integrates the steps from metagenome assembly (Fig. 3.1.C) to binning refinement (Fig. 3.1.F).

From lines 001 to 010 we import all the necessary modules to execute the script. From lines 012 to 049 we define all the path variables. Note that all of them depend on the MOTHER_DIR variable, which contains the path to the directory where all the other directories are placed. From lines 053 to 286 we define the 18 functions that later we call. Finally, from lines 294 till the end, eight steps are carried out to extract the MAGs.

The first step consists of submitting the megahit jobs to Condor (lines 294 to 299). To avoid the execution of next steps before the end of these jobs, we made a loop that only ends when none of the job IDs is in the condor queue (lines 301-318). In the second step, we execute a bash script that groups in the same file all the contigs that belong to the same (meta)genome (lines 321-325).

The third step consists of submitting the Tiara jobs to Condor (lines 327-331). Next, we use the same strategy as in Megahit to wait until the end of these jobs (lines 333-350).

In the fourth step, we run a bash script that moves the contigs into 'euk' or 'prok' directories based on their Tiara classification (lines 356-359).

All the remaining steps involve metawrap tasks (lines 364-399). In the fifth step, we run a bash script that renames the original genomes and move them to another folder (line 366), which will be a metawrap input. Then, in the sixth step, with the execution of other bash script the file names of these new folder are written in a file called datasets.txt (line 370), which also will be a metawrap input. The seventh step consists of submitting the metawrap jobs to condor (line 374). Here, we wait again for the jobs end with the same loop as before (lines 377-393). Finally, in the eight step, with the MAGs already build, the last bash script is executed to rename the MAGs and assign them an unique ID (line 398).

4. RESULTS

4.1. Metagenomes

We obtained MAGs from all processed (meta)genomes (Table 4.1), with a mean number of 5.7 MAGs per (meta)genome. Although we found prokaryotic MAGs in all the (meta)genomes, complete ones were only present in those (meta)genomes that had a genomic source. Conversely, complete eukaryotic MAGs were only found in (meta)genomes with a transcriptomic source.

Table 4.1 | (Meta)genomes metadata. All (meta)genomes platform is Illumina. Numbers between brackets in "Prok MAGs" and "Euk MAGs" represent the number of complete MAGs, that is the number of MAGs that had a BUSCO completeness above 70% in prokaryotes and above 50% in eukaryotes. Bold (meta)genomes are the ones, who's MAGs are placed in tree of Fig. 4.2.

Genome run	Sample name	Submiter	Organism	Sampling location	Instrument	Strategy	Source	Contigs (M)	Bases (G)	Tissue	Prod. MAGs	Prok. MAGs	Euk. MAGs
DRR235373	SAMD00231992	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	34.0	10.2	ovaries	5	2 (0)	3 (2)
DRR235374	SAMD00231993	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	ovaries	6	2 (0)	4 (2)
DRR235375	SAMD00231994	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.9	10.2	ovaries	5	2 (0)	3 (1)
DRR235376	SAMD00231995	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.5	10.0	ovaries	5	2 (0)	3 (2)
DRR235377	SAMD00231996	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.5	10.0	ovaries	5	2 (0)	3 (1)
DRR235378	SAMD00231997	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	34.1	10.2	ovaries	6	2 (0)	4 (2)
DRR235379	SAMD00231998	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	ovaries	5	2 (0)	3 (2)
DRR235380	SAMD00231999	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	ovaries	5	2 (0)	3 (1)
DRR235381	SAMD00232000	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	ovaries	5	2 (0)	3 (2)
DRR235382	SAMD00232001	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	ovaries	6	2 (0)	4 (1)
DRR235383	SAMD00232002	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.5	10.0	ovaries	5	1 (0)	4 (2)
DRR235384	SAMD00232003	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.7	10.1	testes	6	2 (0)	4 (3)
DRR235385	SAMD00232004	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.7	10.1	testes	4	2 (0)	2 (2)
DRR235386	SAMD00232005	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	testes	6	2 (0)	4 (1)
DRR235387	SAMD00232006	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	34.0	10.2	testes	5	2 (0)	3 (1)
DRR235388	SAMD00232007	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.8	10.1	testes	6	2 (0)	4 (2)
DRR235389	SAMD00232008	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	34.0	10.2	testes	5	1 (0)	4 (1)
DRR235390	SAMD00232009	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.4	10.0	testes	5	2 (0)	3 (1)
DRR235391	SAMD00232010	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.5	10.1	testes	5	1 (0)	4 (1)
DRR235392	SAMD00232011	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.6	10.1	testes	5	1 (0)	4 (2)
DRR235393	SAMD00232012	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.6	10.1	testes	5	2 (0)	3 (2)
DRR235394	SAMD00232013	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.8	10.1	testes	5	2 (0)	3 (2)
DRR235395	SAMD00232014	AORI	*F. ancora*	Kenting National Park, Taiwan	HiSeq X Ten	RNA-Seq	T	33.8	10.1	testes	5	2 (0)	3 (1)
ERR2188869	SAMEA104366359	REFUGE2020	*A. tenuis*	-	Il. HiSeq 2000	WGS	G	242.7	48.5	-	6	1 (0)	5 (2)
ERR2190369	**SAMEA104368292**	**REFUGE2020**	***F. fungites***	**Great Barrier reef/Red Sea**	**Il. HiSeq 2000**	**WGS**	**G**	**165.7**	**49.7**	**Sperm**	**5**	**5 (2)**	**0 (0)**
ERR2191367	**SAMEA104368919**	**REFUGE2020**	***G. fascicularis***	**Great Barrier reef/Red Sea**	**Il. HiSeq 1500**	**WGS**	**G**	**170.0**	**51.0**	**Sperm**	**3**	**3 (1)**	**0 (0)**
ERR2192880	SAMEA104372105	REFUGE2020	*P. speciosa*	Australia	Il. HiSeq 2500	RNA-Seq	T	238.7	47.7	-	7	2 (0)	5 (2)
ERR2216064	**SAMEA104421631**	**Penn State**	***C. cruxmelitensis***	**Chimney Rock, England**	**NextSeq 500**	**WGS**	**G**	**32.4**	**9.0**	**-**	**6**	**6 (5)**	**0 (0)**
ERR2216065	**SAMEA104421631**	**Penn State**	***C. cruxmelitensis***	**Chimney Rock, England**	**NextSeq 500**	**WGS**	**G**	**33.7**	**9.5**	**-**	**8**	**8 (5)**	**0 (0)**
ERR2216066	**SAMEA104421631**	**Penn State**	***C. cruxmelitensis***	**Chimney Rock, England**	**NextSeq 500**	**WGS**	**G**	**33.6**	**9.3**	**-**	**7**	**7 (4)**	**0 (0)**
ERR2216067	**SAMEA104421631**	**Penn State**	***C. cruxmelitensis***	**Chimney Rock, England**	**NextSeq 500**	**WGS**	**G**	**33.3**	**9.4**	**-**	**8**	**8 (5)**	**0 (0)**
ERR571458	SAMEA2673766	REFUGE2020	*P. lutea*	Orpheus Island, Australia	Il. HiSeq 2500	WGS	G	74.7	22.4	Sperm	1	0 (0)	1 (0)
ERR571459	**SAMEA2673766**	**REFUGE2020**	***P. lutea***	**Orpheus Island, Australia**	**Il. HiSeq 2500**	**WGS**	**G**	**147.1**	**44.1**	**Sperm**	**13**	**11 (6)**	**2 (0)**
ERR571463	**SAMEA2673766**	**REFUGE2020**	***P. lutea***	**Orpheus Island, Australia**	**Il. HiSeq 2500**	**WGS**	**G**	**99.2**	**29.8**	**Sperm**	**9**	**7 (3)**	**2 (0)**
ERR571494	SAMEA2675461	REFUGE2020	*P. lutea*	Orpheus Island, Australia	Il. HiSeq 2000	EST	T	146.2	29.2	-	8	1 (0)	7 (2)

AORI: Atmosphere and Ocean Research Institute (The University of Tokyo), REFUGE2020: Reef Future Genomics 2020 consortium, Penn State: Pennsylvania State University, Prok. MAGs: Prokaryotic MAGs, Euk. MAGs: Eukaryotic MAGs, *F. ancora: Fimbriaphyllia ancora, A. tenuis: Acropora tenuis*, *F. fungites: Fungia fungites*, *G. fascicularis: Galaxea fascicularis*, *P. speciosa: Pachyseris speciosa*, *C. cruxmelitensis*: *Calvadosia cruxmelitensis*, *P. lutea*: *Porites lutea*, Il.a: Illumina, RNA-Seq: RNA sequencing, WGS: Whole Genome Sequencing, EST: Expressed sequence tag, T: transcriptomic, G: genomic.

4.2. MAGs

From the 35 coral metagenomes analysed (Table 4.1), 201 MAGs were extracted (Table 4.2). According to BUSCO classification, 101 were bacterial and 100 were eukaryotic. Thirty-one prokaryotic MAGs had a BUSCO completeness above 70% (Table 4.3), while thirty-six eukaryotic MAGs had a BUSCO completeness above 50% (Table S1).

Table 4.2 | MAGs count. The number of metagenomes samples analysed of each coral specie is shown in brackets.

	Eukaryote			Prokaryote			Total
	uncompl.	compl.	total	uncompl.	compl.	total	
***Acropora tenuis* (1)**	3	2	5	1	0	1	6
***Calvadosia cruxmelitensis* (4)**	0	0	0	10	19	29	29
***Fimbriaphyllia ancora* (23)**	48	30	78	42	0	42	120
***Fungia fungites* (1)**	0	0	0	3	2	5	5
***Galaxea fascicularis* (1)**	0	0	0	2	1	3	3
***Pachyseris speciosa* (1)**	3	2	5	2	0	2	7
***Porites lutea* (4)**	10	2	12	10	9	19	31
Total (35)	64	36	100	70	31	101	201

uncompl.: uncomplete MAGs (under BUSCO completeness threshold).
compl.: complete MAGs (above BUSCO completeness threshold).

From the complete MAGs, eukaryotes (Table S1) had significantly more contigs than prokaryotes (Table 4.3). The average number of contigs in eukaryote complete MAGs was 25071 contigs/genome, while in prokaryotes was 900 contigs/genome. However, N50 contig length was bigger in prokaryote complete MAGs (39337 bp) than in eukaryote ones (2292 bp).

Table 4.3 shows an overview of the complete prokaryotic MAGs metrics. MAGs names are composed by its genome run name, an abbreviation of their coral host, and a unique number that acts as the MAG id, so it's unique for each MAG. Note that all bacterial complete MAGs that exceed the completeness threshold belong to *G. fascicularis, P. lutea, F. fungites* or *C. cruxmelitensis* corals.

Table 4.3 | Metrics of complete prokaryotic MAGs. GTDB column shows if the MAG is placed in the reference tree (Y) or not (N) and in brackets the number of single marker genes identified of the 120 marker genes set.

MAG name	Coral specie	No. of contigs	Contig N50 (bp)	Completeness (%)	Single BMG (%)	Duplicated BMG (%)	Sin./Dupl. BMG	No. BMG	GTDB (No. single MG of 120)
ERR2191367_Gfas_01	*G. fascicularis*	393	25031	96.1	96.1	0.0	-	688	Y (114)
ERR2190369_Ffun_02	*F. fungites*	1574	9462	83.2	79.1	4.1	19.3	688	Y (64)
ERR571459_Plut_03	*P. lutea*	124	68262	94.1	94.0	0.1	940.0	688	Y (110)
ERR571459_Plut_04	*P. lutea*	452	68262	98.9	98.8	0.1	988.0	688	Y (116)
ERR2190369_Ffun_05	*F. fungites*	679	13019	90.2	89.2	1.0	89.2	688	Y (100)
ERR2216064_Ccrux_06	*C. cruxmelitensis*	39	61118	74.7	74.5	0.2	372.5	432	Y (101)
ERR2216067_Ccrux_07	*C. cruxmelitensis*	34	64630	76.6	76.6	0.0	-	124	Y (96)
ERR2216066_Ccrux_08	*C. cruxmelitensis*	31	62155	77.4	77.4	0.0	-	124	Y (96)
ERR2216065_Ccrux_09	*C. cruxmelitensis*	42	49347	83.1	82.9	0.2	414.5	432	Y (116)
ERR2216065_Ccrux_10	*C. cruxmelitensis*	711	23579	87.5	87.3	0.2	436.5	432	Y (120)
ERR2216067_Ccrux_11	*C. cruxmelitensis*	91	64595	87.2	87.0	0.2	435.0	432	Y (119)
ERR2216064_Ccrux_12	*C. cruxmelitensis*	1575	2514	100.0	82.3	17.7	4.6	124	Y (74)
ERR571463_Plut_13	*P. lutea*	1148	3758	72.1	69.3	2.8	24.8	833	Y (102)
ERR571459_Plut_14	*P. lutea*	118	71781	78.0	76.7	1.3	59.0	833	Y (80)
ERR571459_Plut_15	*P. lutea*	117	71516	97.3	94.8	2.5	37.9	833	Y (108)
ERR571459_Plut_16	*P. lutea*	923	5518	83.4	83.2	0.2	416.0	833	Y (105)
ERR571459_Plut_17	*P. lutea*	745	5748	81.6	81.4	0.2	407.0	833	Y (106)
ERR571463_Plut_18	*P. lutea*	682	9896	90.9	90.4	0.5	180.8	833	Y (107)
ERR571463_Plut_19	*P. lutea*	510	10582	91.0	90.6	0.4	226.5	833	Y (111)
ERR2216065_Ccrux_20	*C. cruxmelitensis*	160	58182	90.3	90.3	0.0	-	124	Y (113)
ERR2216067_Ccrux_21	*C. cruxmelitensis*	57	71870	89.5	89.5	0.0	-	124	Y (113)
ERR2216067_Ccrux_22	*C. cruxmelitensis*	153	63741	89.5	89.5	0.0	-	124	Y (113)
ERR2216065_Ccrux_23	*C. cruxmelitensis*	61	67978	88.7	88.7	0.0	-	124	Y (111)
ERR2216066_Ccrux_24	*C. cruxmelitensis*	55	73460	72.6	72.6	0.0	-	124	Y (99)
ERR2216066_Ccrux_25	*C. cruxmelitensis*	181	71186	89.5	88.7	0.8	110.9	124	Y (110)
ERR2216064_Ccrux_26	*C. cruxmelitensis*	164	51364	89.5	88.7	0.8	110.9	124	Y (112)
ERR2216064_Ccrux_27	*C. cruxmelitensis*	65	60463	87.1	86.3	0.8	107.9	124	Y (106)
ERR2216064_Ccrux_28	*C. cruxmelitensis*	4169	2639	100.0	24.2	75.8	0.3	124	N (3)
ERR2216065_Ccrux_29	*C. cruxmelitensis*	4392	2548	100.0	21.0	79.0	0.3	124	N (0)
ERR2216066_Ccrux_30	*C. cruxmelitensis*	4288	2544	100.0	22.6	77.4	0.3	124	N (2)
ERR2216067_Ccrux_31	*C. cruxmelitensis*	4156	2705	100.0	21.0	79.0	0.3	124	N (1)

G. fascicularis: Galaxea fascicularis, P. lutea: Porites lutea, F. fungites: Fungia fungites, C. cruxmelitensis: Calvadosia cruxmelitensis, BMG: BUSCO Marker Genes, MG: marker genes

4.3. MAGs taxonomic assignment

Prokaryotic MAGs

According to taxonomic classifications, we split the 31 MAGs into eight groups (Fig. 4.1).

All successful classifications for MAGs 1-5 were agree[1] (Fig. 4.1). GTDB-Tk provided the most specific classification, it classified all of them till the specie *Comamonas acidovorans* (Table S2). ENA, RDP and SLV did it in the same way for three of the five MAGs but only till Delftia genus, which is equivalent to Comamonas (Bilgin et al., 2015). BAT classified all of them till the order Burkholderiales.

MAGs 6-12 classifications were agreed till Alphaproteobacteria class (Fig. 4.1). While GTDB-Tk was able to classify all of them till family range, BAT did it only till phylum, except in one case that also specified the class (Table S2). RDP and SLV classified three MAGS till genus and family respectively. GTDB-Tk and RDP and SLV differ in the order classification, GTDB-Tk classified them inside UBA8366 and RDP and SLV inside Rhodospirillales order. However, RDP and SLV classified them in different families. RDP did it inside Rhodospirillaceae and SLV inside Terasakiellaceae family.

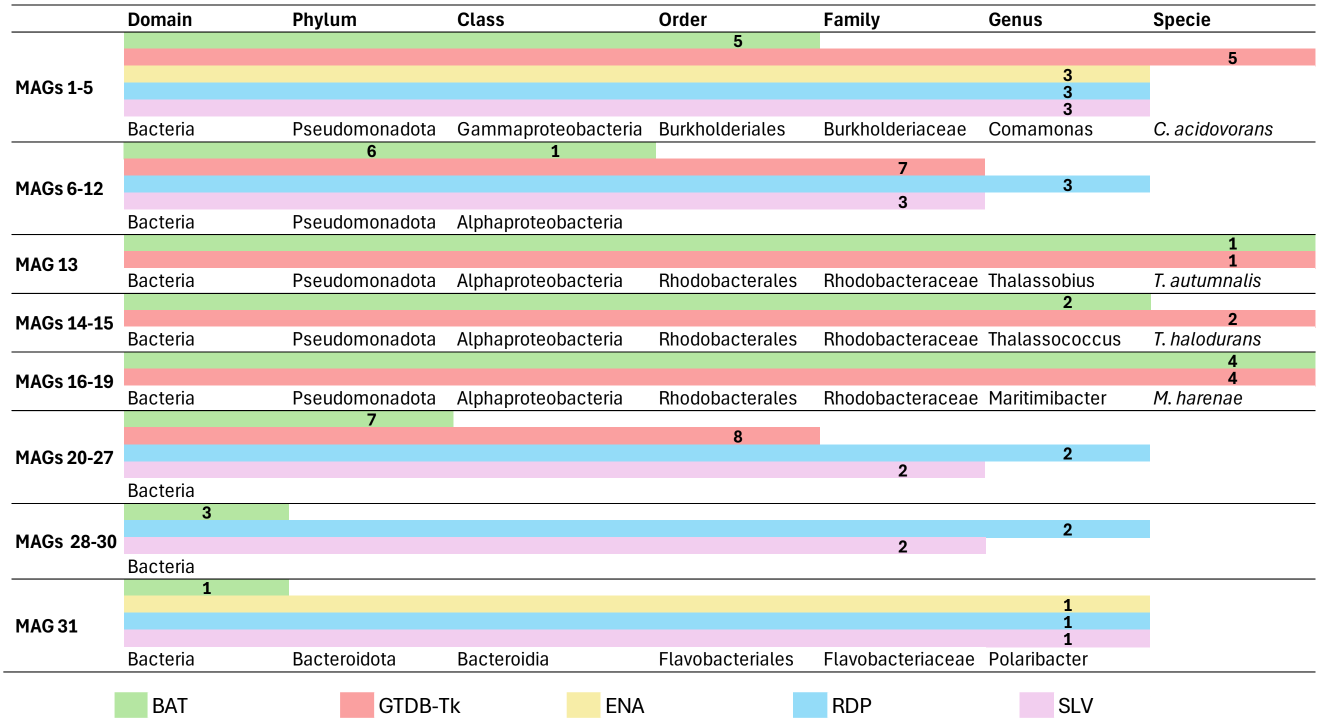

Figure 4.1. MAGs taxonomical classification comparison between different tools. Colour bars represent the maximum range where each tool has classified the MAGs. Numbers inside colour bars represent the number of MAGs of the groups that has been classified till this range. Behind colour bars is shown the classification in which tools has not contradictions. It does not show the failed classifications, where MAG was not classified.

[1] Table S2, which shows classification of each MAGs, shows discrepancies in the class and assignation. This is because Betaproteobacteria class, within BAT, ENA and RDP classified MAGs 1-5, is included in Gammaproteobacteria class in GTDB and SLV databases (Parks et al., 2018).

MAGs 13-19, were all classified successfully by BAT and GTDB-Tk, and both classifications agreed (Fig. 4.1). MAGs 13 and 16-19 were classified by BAT and GTDB-Tk till specie, as *Thalassobius autumnalis* and *Maritimibacter harenae* respectively (Table S2). Whereas MAGs 14 and 15 were classified by BAT till genus and by GTDB-Tk till *Thalassococcus halodurans* specie. MAGs 20-27 were all classified by BAT till phylum Proteobateria (also named Pseudomonadota), however, GTDB-Tk classified all of them inside XYD2-FULL-50-16 order, which is inside SAR324 phylum. RDP classified two MAGs inside Aestuariispira genus, and SLV classified the same two inside Terasakiellaceae family. Both taxa are inside of Proteobacteria phylum.

BAT did not assign to the remaining MAGs any phylum classification (Fig. 4.1), it only indicated that they were inside Bacteria domain (Fig. 4.1). GTDB-Tk did not assign any classification to those MAGs because in the align step less than 4 marker unique genes were identified. MAG number 31 was classified by ENA, RDP and SLV inside Polaribacter genus (Table S2), which is inside Bacteroidota phylum. Finally, two of the three remaining MAGs were classified by RDP inside Aestuariispira genus and by SLV inside Terasakiellaceae family.

The GTDB and LTP databases did not provide any classification for the rRNA sequences of the submitted MAGs.

Eukaryotic MAGs

All classifications for eukaryotic MAGs corresponded to coral genomes (Table S3). BAT classified all the eukaryotic complete MAGs inside the Scleractinia order, except two that were unclassified. BLAST only found results for 15 of the 36 eukaryotic MAGs (Table S3). It classified the MAGs coming from *A. tenuis* and the MAGs coming from *P. lutea* as the same specie coming from. Nine *F. ancora* MAGs as the 6th *G. fascicularis* chromosome, and the *P. speciosa* MAGs as *Siderastrea sidereal* 13th chromosome.

4.4. Phylogenomic placement of prokaryotic MAGs

Twenty-seven prokaryotic MAGs were successfully placed in the GTDB reference tree (Fig. 4.2). The remaining four were excluded by the tool because in the align step less than 4 marker unique genes were identified (Table 4.3).

In https://github.com/moscatlauramartinez/TFM_MAGs_Factory is the GTDB tree including our MAGs. Since we could not edit this tree, we present a smaller version in Figure 4.2. This version was generated using the same tool but only includes the GTDB genomes that were closest to our MAGs in the original.

The analysed MAGs were placed into two different phyla: 19 within Pseudomonadota and 8 within SAR324. Five Pseudomonadota MAGs (1-5) were placed inside Comamonas genus, which belongs to Gammaproteobacteria class, order Burkholderiales and Burkholderiaceae B family. The nearest specie of these five MAGs was *C. acidovorans*.

The remaining 14 Pseudomonadota MAGs were assigned to the Alphaproteobacteria class. Of these, seven were placed in the Rhodobacterales order and Rhodobacteraceae family. Within this group, one MAG (13) was placed in the Thalassobius genus, with *T. autumnalis* being the nearest specie. Two MAGs (14,15) were placed next to Thalassococcus genus, with *T. halodurans* as the closest species, and four MAGs (16-19) were placed in the Maritimibacter genus, with *M. harenae* as the nearest specie. The remaining seven Alphaproteobacteria MAGs (7-12) could not be assigned to any specific order, but their closest sequence was identified as JAQWUY01 sp..

All MAGs in the SAR324 phylum (20-17) were placed under the SAR324 class and XYD2-FULL-50-16 order. These MAGs could not be placed within any family or genus, but their nearest sequence was JAERRU01 sp., a contig isolated from a gas hydrate in the Gulf of Mexico that is classified under the JAERRU01 family and genus.

VBCG tree (Fig. S1) placed 25 of the 31 provided MAGs, grouping them in the same way as GTDB-Tk. The seven unplaced MAGs did not have enough marker genes.

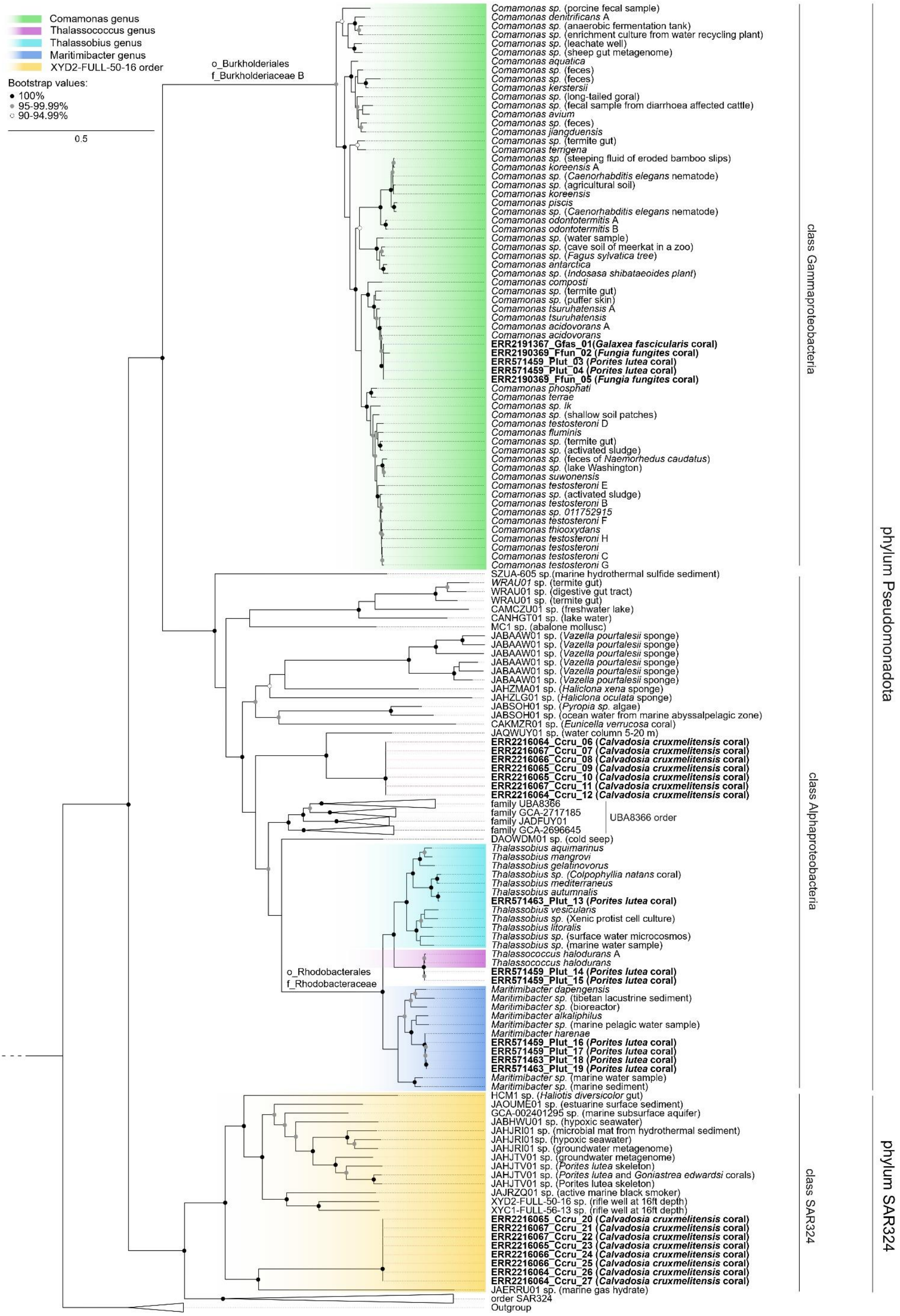
Comamonas genus
Thalassococcus genus
Thalassobius genus
Maritimibacter genus
XYD2-FULL-50-16 order
Bootstrap values:
● 100%
● 95-99.99%
○ 90-94.99%
0.5
o_Burkholderiales
f_Burkholderiaceae B
Comamonas sp. (porcine fecal sample)
Comamonas denitrificans A
Comamonas sp. (anaerobic fermentation tank)
Comamonas sp. (enrichment culture from water recycling plant)
Comamonas sp. (leachate well)
Comamonas sp. (sheep gut metagenome)
Comamonas aquatica
Comamonas sp. (feces)
Comamonas sp. (feces)
Comamonas kerstersii
Comamonas sp. (long-tailed goral)
Comamonas sp. (fecal sample from diarrhoea affected cattle)
Comamonas avium
Comamonas sp. (feces)
Comamonas jiangduensis
Comamonas sp. (termite gut)
Comamonas terrigena
Comamonas sp. (steeping fluid of eroded bamboo slips)
Comamonas koreensis A
Comamonas sp. (Caenorhabditis elegans nematode)
Comamonas sp. (agricultural soil)
Comamonas koreensis
Comamonas piscis
Comamonas sp. (Caenorhabditis elegans nematode)
Comamonas odontotermitis A
Comamonas odontotermitis B
Comamonas sp. (water sample)
Comamonas sp. (cave soil of meerkat in a zoo)
Comamonas sp. (Fagus sylvatica tree)
Comamonas antarctica
Comamonas sp. (Indosasa shibataeoides plant)
Comamonas composti
Comamonas sp. (termite gut)
Comamonas sp. (puffer skin)
Comamonas tsuruhatensis A
Comamonas tsuruhatensis
Comamonas acidovorans A
Comamonas acidovorans
ERR2191367_Gfas_01(Galaxea fascicularis coral)
ERR2190369_Ffun_02 (Fungia fungites coral)
ERR571459_Plut_03 (Porites lutea coral)
ERR571459_Plut_04 (Porites lutea coral)
ERR2190369_Ffun_05 (Fungia fungites coral)
Comamonas phosphati
Comamonas terrae
Comamonas sp. lk
Comamonas sp. (shallow soil patches)
Comamonas testosteroni D
Comamonas fluminis
Comamonas sp. (termite gut)
Comamonas sp. (activated sludge)
Comamonas sp. (feces of Naemorhedus caudatus)
Comamonas sp. (lake Washington)
Comamonas suwonensis
Comamonas testosteroni E
Comamonas sp. (activated sludge)
Comamonas testosteroni B
Comamonas sp. 011752915
Comamonas testosteroni F
Comamonas thiooxydans
Comamonas testosteroni H
Comamonas testosteroni
Comamonas testosteroni C
Comamonas testosteroni G
SZUA-605 sp.(marine hydrothermal sulfide sediment)
WRAU01 sp. (termite gut)
WRAU01 sp. (digestive gut tract)
WRAU01 sp. (termite gut)
CAMCZU01 sp. (freshwater lake)
CANHGT01 sp. (lake water)
MC1 sp. (abalone mollusc)
JABAAW01 sp. (Vazella pourtalesii sponge)
JABAAW01 sp. (Vazella pourtalesii sponge)
JABAAW01 sp. (Vazella pourtalesii sponge)
JABAAW01 sp. (Vazella pourtalesii sponge)
JABAAW01 sp. (Vazella pourtalesii sponge)
JABAAW01 sp. (Vazella pourtalesii sponge)
JAHZMA01 sp. (Haliclona xena sponge)
JAHZLG01 sp. (Haliclona oculata sponge)
JABSOH01 sp. (Pyropia sp. algae)
JABSOH01 sp. (ocean water from marine abyssalpelagic zone)
CAKMZR01 sp. (Eunicella verrucosa coral)
JAQWUY01 sp. (water column 5-20 m)
ERR2216064_Ccru_06 (Calvadosia cruxmelitensis coral)
ERR2216067_Ccru_07 (Calvadosia cruxmelitensis coral)
ERR2216066_Ccru_08 (Calvadosia cruxmelitensis coral)
ERR2216065_Ccru_09 (Calvadosia cruxmelitensis coral)
ERR2216065_Ccru_10 (Calvadosia cruxmelitensis coral)
ERR2216067_Ccru_11 (Calvadosia cruxmelitensis coral)
ERR2216064_Ccru_12 (Calvadosia cruxmelitensis coral)
family UBA8366
family GCA-2717185
family JADFUY01
family GCA-2696645
UBA8366 order
DAOWDM01 sp. (cold seep)
Thalassobius aquimarinus
Thalassobius mangrovi
Thalassobius gelatinovorus
Thalassobius sp. (Colpophyllia natans coral)
Thalassobius mediterraneus
Thalassobius autumnalis
ERR571463_Plut_13 (Porites lutea coral)
Thalassobius vesicularis
Thalassobius sp. (Xenic protist cell culture)
Thalassobius litoralis
Thalassobius sp. (surface water microcosmos)
Thalassobius sp. (marine water sample)
Thalassococcus halodurans A
Thalassococcus halodurans
ERR571459_Plut_14 (Porites lutea coral)
ERR571459_Plut_15 (Porites lutea coral)
o_Rhodobacterales
f_Rhodobacteraceae
Maritimibacter dapengensis
Maritimibacter sp. (tibetan lacustrine sediment)
Maritimibacter sp. (bioreactor)
Maritimibacter alkaliphilus
Maritimibacter sp. (marine pelagic water sample)
Maritimibacter harenae
ERR571459_Plut_16 (Porites lutea coral)
ERR571459_Plut_17 (Porites lutea coral)
ERR571463_Plut_18 (Porites lutea coral)
ERR571463_Plut_19 (Porites lutea coral)
Maritimibacter sp. (marine water sample)
Maritimibacter sp. (marine sediment)
HCM1 sp. (Haliotis diversicolor gut)
JAOUME01 sp. (estuarine surface sediment)
GCA-002401295 sp. (marine subsurface aquifer)
JABHWU01 sp. (hypoxic seawater)
JAHJRI01 sp. (microbial mat from hydrothermal sediment)
JAHJRI01sp. (hypoxic seawater)
JAHJRI01 sp. (groundwater metagenome)
JAHJTV01 sp. (groundwater metagenome)
JAHJTV01 sp. (Porites lutea skeleton)
JAHJTV01 sp. (Porites lutea and Goniastrea edwardsi corals)
JAHJTV01 sp. (Porites lutea skeleton)
JAJRZQ01 sp. (active marine black smoker)
XYD2-FULL-50-16 sp. (rifle well at 16ft depth)
XYC1-FULL-56-13 sp. (rifle well at 16ft depth)
ERR2216065_Ccru_20 (Calvadosia cruxmelitensis coral)
ERR2216067_Ccru_21 (Calvadosia cruxmelitensis coral)
ERR2216067_Ccru_22 (Calvadosia cruxmelitensis coral)
ERR2216065_Ccru_23 (Calvadosia cruxmelitensis coral)
ERR2216066_Ccru_24 (Calvadosia cruxmelitensis coral)
ERR2216066_Ccru_25 (Calvadosia cruxmelitensis coral)
ERR2216064_Ccru_26 (Calvadosia cruxmelitensis coral)
ERR2216064_Ccru_27 (Calvadosia cruxmelitensis coral)
JAERRU01 sp. (marine gas hydrate)
order SAR324
Outgroup
class Gammaproteobacteria
class Alphaproteobacteria
class SAR324
phylum Pseudomonadota
phylum SAR324

Figure 4.2. GTDB phylogenetic tree. It contains 27 of the 31 prokaryote MAGs placed.

5. DISCUSSION

5.1. MAGs production

As seen in Table 4.1, all processed (meta)genomes produced MAGs. However, we can see a clear relationship between the domain of the complete MAGs produced and the sequencing strategy. This distinction is due to the method of the sequencing strategy. RNA sequencing (RNA-seq) involves sequencing mRNA by anchoring the adapter to the poly(A) tail of mRNA (Yu et al., 2020). The fact is that this tail is only present in eukaryotes' mRNAs, leading to exclusive sequencing of eukaryote cells. Therefore, we did not obtain prokaryotic complete MAGs from (meta)genomes sequenced by this method because as any prokaryotic cell had mRNA with a poly(A) tail. On the contrary, WGS sequences all DNAs, so the lack of complete eukaryotic MAGs in WGS sequenced (meta)genomes could be due to the predominance of prokaryotic cells in the coral holobiont.

As each coral specie is sequenced with one technique only, in any case we have obtained complete MAGs from both domains, except in *P. lutea*. In this coral, we obtained complete prokaryotic MAGs from the WGS sequenced (meta)genomes and one eukaryotic MAG from the (meta)genome sequenced by Expressed Sequence Tags (EST).

5.2. Prokaryotic MAGs BUSCO's completeness

According to BUSCO's completeness values (Table 4.3), MAGs 28-31 are the most complete, achieving 100% completeness. However, they have not been included in the GTDB reference tree due to the lack of marker genes, which appears to be contradictory.

Note that the proportion of single/duplicate BUSCO marker genes in the unplaced MAGs (28-31) is significantly lower than in the included ones. In fact, metawrap considers duplicated single-copy genes a contamination metric. Additionally, the unplaced MAGs (28-31) had a higher number of contigs and a much lower N50 contig length, indicating a high level of fragmentation. For future selection of complete MAGs, we recommend evaluating not only MAG completeness but also single/duplicated proportion, number of contigs, and N50 length.

5.3. MAG's discrepant classifications

As seen in Figure 4.1, Table S2, and Table S3, different tools sometimes provided different classifications for the same MAG. These discrepancies occur because each method uses different parts of the MAG for classification. We used three different methods to classify MAGs: BAT, GTDB-Tk, and rRNA-BC. BAT classifies genomes based on ORFs, rRNA-BC uses rRNA sequences, and GTDB-Tk relies on a set of 120 marker genes. Since the sequences selected for comparison with the database differ in each case, the classifications do not always match. In fact, these tools classify only the selected sequences from the MAG, not the entire MAG itself.

Additionally, the same sequences can provide different classifications when compared with different databases. This happened with prokaryote MAGs 6-12 (Table S2); ENA did not classify these MAGs, RDP classified them as belonging to the Aestuariispira genus, and SLV classified them as part of the Terasakiellaceae family. A possible explanation is that each database compares the provided rRNA sequences with their reference sequences in different ways.

5.4. Absence of microeukaryotic organisms

The fact that we did not find eukaryotic MAGs is not because there are none. It is probably because there was much more genetic material from the coral than from the associated microeukaryotes.

Since we did not remove the coral genome from the downloaded meta(genome) sample, it makes sense that we found coral genomes in the eukaryote MAGs (Table S3). BAT classified all MAGs within the Scleractinia order, which is logical because all the corals these MAGs belong to are from this order.

BLAST classified the MAGs of A. tenuis and P. lutea as the same species. This means that the rRNA of these two species is well differentiated from others. Furthermore, BLAST classified *F. ancora* MAGs as *G. fascicularis*, both of which belong to the Euphylliidae family. This indicates that the rRNA of these two species is not very different, which is not surprising since they belong to the same family. However, *P. speciosa* MAGs were classified as *S. siderea* coral, which is also a scleractinian coral, but belongs to a different family.

5.5. Phylogenetic trees

All GTDB-Tk MAGs classifications are consistent with their placement in the GTDB tree, except in MAGs 6-12 classifications. These are classified by GTDB-Tk inside the Alphaproteobacteria class, UBA8366 order and GCA-2696645 family (Table S2). However, in the GTDB reference tree, they are placed outside the UBA8366 order (Fig. 4.2). It may be that these discrepancies are due to small differences between the GTDB-Tk classifying workflow and the GTDB-Tk *de novo* workflow.

GTDB-Tk was able to place 27 of the 31 complete prokaryotic MAGs (Fig. 4.2), two more than VCBG, which placed 25 (Fig. S1). Both tools grouped the MAGs in the same way. This could be due to the fact that 19 of the 20 marker genes used by VBCG, are also included in the 120 marker gene set used by GTDB-Tk.

We gave more importance to the GTDB tree because the reference genomes provided more information about the MAGs context. Furthermore, its bootstrap values were much higher than the VBCG ones, which provided us with more confidence. However, VBCG is much faster, using one node and twelve tasks per node, it lasts 4 min 30 s in building the tree, while GTBD lasts 36 h and 22 min under the same conditions. We consider that VBCG can be useful to have a first and quick approximation of a genome set clustering and its relationships.

MAGs context

Most of the genomes placed near MAGs 1-5, classified as *C. acidovorans* (Fig. 4.2), were not associated with the marine environment. Even so, some of them were isolated from animal samples.

MAGs 6-19, placed within Alphaproteobacteria class, had marine-related genomes near them. In addition, some of them were associated with marine animals, such as *Colpophyllia natans* or *Eunicella verrucosa* corals.

All genomes placed within the order XYD2-FULL-50-16, where MAGs 20-27 are, are related with the marine environment. Those genomes belong to a wide range of conditions. Note that, near MAGs 20-27, there are genomes associated with corals, such as *P. lutea.*

5.6. Coral microbiome diversity

In *C. cruxmelitensis* metagenome we found seven MAGs (6-12) classified inside Alphaproteobacteria class and eight more (20-27) inside XYD2-50-16 order (Fig. 5.1). Note that the four unplaced genomes (28-31) also belong to this coral species.

In *P. lutea* genomes we found five of the six groups of bacteria placed in the GTDB reference tree (Fig. 5.1). There are the four MAGs (16-19) classified as *M. harenae*, the two MAGs (14, 15) placed next to *T. halodurans*, two more MAGs (3,4) classified as *C. acidovorans* and the one (13) classified as *T. autumnalis*.

In F*. fungites* and *G. fascicularis* genomes we only found two (2, 5) and one (1) *C. acidovorans* MAGs respectively.

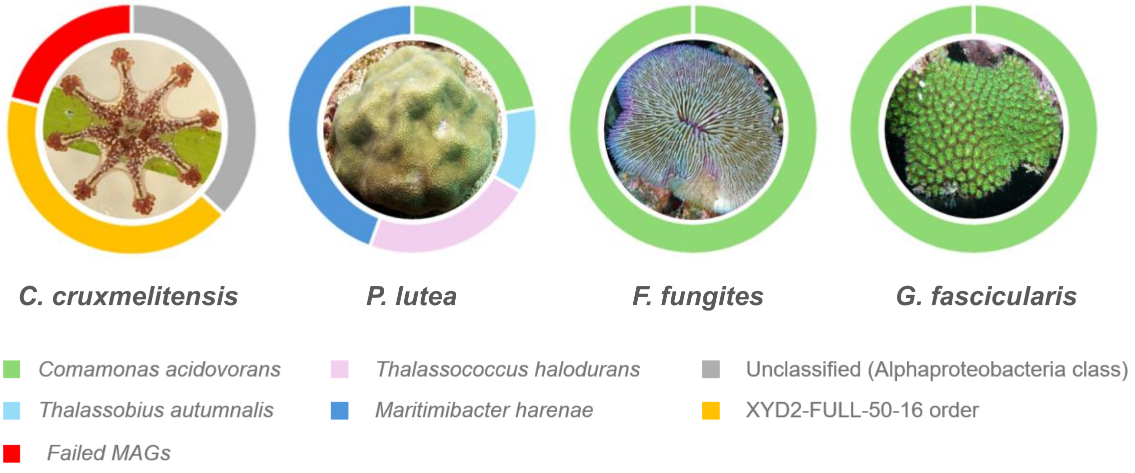

Figure 5.1. Coral prokaryotic diversity. The photo inside each circle is of the coral it represents. Photos authors: David Fenwik (*C. cruxmelitensis*), Emre Turak (*P. lutea*), Zarinah Waheed (*F. fungites*), Larry Basch (*G. fascicularis*).

C. acidovorans

In reviewing the literature, no data were found on the association between the bacterium *C. acidovorans* and the corals where we found it, *P. lutea*, *F. fungites*, or *G. fascicularis*. However, previous reports have shown *C. acidovorans* associated with the coral *Montastraea franksi* (Rohwer et al., 2001, 2002). Additionally, a recent study identified bacteria of the genus *Comamonas* in *P. lutea* and *G. fascicularis* corals (R.-W. Chen et al., 2024), and Li et al. (2013) reported bacteria of the order *Burkholderiales* associated with *P. lutea* and *G. fascicularis*.

C. acidovorans is a Gram-negative and motile bacteria widely distributed (Wen et al., 1999). It can be found causing human infections (Yildiz et al., 2019) as in the marine environment (Sindhu & Potty, 2015). Comamonas genus bacteria associated with corals are capable of metabolizing DMSP/DMS (J. Li et al., 2013). These molecules, produced by marine phytoplankton, promote cloud formation when they are vaporized within water into the atmosphere, reducing the solar radiation that reaches the earth's surface (Hegg et al., 1991).

Rhodobacteraceae family

T. halodurans, *T. autumnalis*, and *M. harenae* belong to Rhodobacteraceae family, which is highly associated with corals (Pujalte et al., 2014; Rubio-Portillo et al., 2016; Luo et al., 2021). Becker at al. (2021) report the relation between Thalassobius and Thalassococcus genus with the Stony Coral Tissue Loss Disease (SCTLD).

While *T. halodurans* was also found Pootakham et al. (2019) in *P. lutea*, we did not find any study that reports the presence of *T. autumnalis* or *M. harenae* in corals. Even so, other near bacteria of the Thalassobius genus (Fig 4.2), such as *Thalassobius mediterraneus*, were found in *P. lutea* (Pootakham et al., 2019) and other coral species (Sekar et al., 2006). However, Maritimibacter genus is not related with any coral specie, but it does with some fishes (Messyasz et al., 2021).

T. halodurans was isolated for the first time from the marine sponge *Halichondria panicea,* in Friday Harbor, San Juan Island, WA, USA (Lee et al., 2007). It is a bacteria Gram-negative, with gliding motility, strictly aerobic and halophilic. It is Oxidase- and catalase-positive and it can reduce nitrate to nitrite but not to nitrogen gas (N_2) (Lee et al., 2007).

T. autumnalis also is a Gram-negative bacteria non-motile and aerobic. Is chemoorganotroph and oxidase and catalase positive (Pujalte et al., 2018). It was isolated from coastal seawater surrounding cultivated oysters, in the Mediterranean Sea, at Vinaroz coast, Spain. We did not find any report apart from Pujaltre et al. (2018) that describes this bacterium.

M. harenae is Gram-negative, strictly aerobic, and nonmotile. It is positive for glucose fermentation, nitrate reduction, and catalase and oxidase activities (Khan et al., 2021). It was isolated from sea sand in South Korea by Khan et al. (2021). Apart from this study, there's no literature published that expands the knowledge about this bacterium.

XYD2-FULL-50-16 order

No reports are referring to the order XYD2-FULL-50-16, nor the class SAR324 in the genus *Calvadosia*. However, some studies have indicated the presence of bacteria from the class SAR324 in the water surrounding corals (Doyle et al., 2022). For instance, Jansen et al. identified this class involved in methane oxidation within the plankton of deep-water coral reefs at depths of 260–350 m. Additionally, Glasl et al. (2020) found these bacteria in the water surrounding shallower corals (less than 10 m depth). Furthermore, bacteria from the SAR324 phylum were found within *P. lutea* and *G. edwardsi* corals, where they are involved in carbon fixation (Cárdenas et al., 2022).

6. CONCLUSIONS

I. Whole Genome Sequencing (WGS) is an effective sequencing technique for obtaining prokaryotic Metagenome-Assembled Genomes (MAGs).

II. The completeness metric provided by BUSCO is not ideal for selecting MAGs, as it includes duplicated single-copy genes, which indicate contamination. Instead, unique single-copy genes offer a more accurate measure of MAG completeness.

III. Our results demonstrate that the MAGs Factory pipeline is well-suited for obtaining prokaryotic MAGs from coral (meta)genomes. However, it is insufficient for constructing eukaryotic MAGs from the coral microbiome.

IV. The classification tools in this study sometimes provided varying classifications for the same MAG because they compared different sequences. BAT classified MAGs based on ORFs, rRNA-BC used rRNA sequences, and GTDB-Tk relied on a set of 120 marker genes. Although GTDB-Tk offered the most accurate classifications, BAT successfully classified almost all MAGs.

V. We reported, for the first time, the presence of *C. acidovorans* in the corals *P. lutea*, *F. fungites*, and *G. fasciularis*, as well as the association of *T. autumnalis* and *M. harenae* with corals, specifically in *P. lutea*. Additionally, we identified the class SAR324 as part of the coral microbiome.

VI. In addition to the taxonomic novelties of the coral microbiome, we obtained 31 bacterial genomes. All of them have a completeness greater than 70% according to BUSCO, and 28 have a unique single-copy gene percentage higher than 72%, indicating a contamination level below 30%. Analysing these genomes can help identify genes that enhance our understanding of the role of these bacteria in the coral holobiont and the functioning of coral gardens.

7. REFERENCES

Agarwal, M., Lamb, R. W., Smith, F., & Witman, J. D. (2024). Distribution and ecology of shallow-water black corals across a depth gradient on Galápagos rocky reefs. *Coral Reefs*, *43*(3), 733-745. https://doi.org/10.1007/s00338-024-02497-6

Andrello, M., Darling, E. S., Wenger, A., Suárez-Castro, A. F., Gelfand, S., & Ahmadia, G. N. (2022). A global map of human pressures on tropical coral reefs. *Conservation Letters*, *15*(1), e12858. https://doi.org/10.1111/conl.12858

Bilgin, H., Sarmis, A., Tigen, E., Soyletir, G., & Mulazimoglu, L. (2015). Delftia acidovorans: A rare pathogen in immunocompetent and immunocompromised patients. *The Canadian Journal of Infectious Diseases & Medical Microbiology*, *26*(5), 277-279.

Blackall, L. L., Wilson, B., & Oppen, M. J. H. (2015). Coral—The world's most diverse symbiotic ecosystem. *Molecular Ecology*, *24*(21), 5330-5347. https://doi.org/10.1111/mec.13400

Bo, M., Bavestrello, G., Angiolillo, M., Calcagnile, L., Canese, S., Cannas, R., Cau, A., D'Elia, M., D'Oriano, F., Follesa, M. C., Quarta, G., & Cau, A. (2015). Persistence of Pristine Deep-Sea Coral Gardens in the Mediterranean Sea (SW Sardinia). *PLOS ONE*, *10*(3), e0119393. https://doi.org/10.1371/journal.pone.0119393

Bourne, D. G., Morrow, K. M., & Webster, N. S. (2016). Insights into the Coral Microbiome: Underpinning the Health and Resilience of Reef Ecosystems. *Annual Review of Microbiology*, *70*(Volume 70, 2016), 317-340. https://doi.org/10.1146/annurev-micro-102215-095440

Bowe, A., Onodera, T., Sadakane, K., & Shibuya, T. (2012). Succinct de Bruijn Graphs. En B. Raphael & J. Tang (Eds.), *Algorithms in Bioinformatics* (pp. 225-235). Springer. https://doi.org/10.1007/978-3-642-33122-0_18

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Cárdenas, A., Raina, J.-B., Pogoreutz, C., Rädecker, N., Bougoure, J., Guagliardo, P., Pernice, M., & Voolstra, C. R. (2022). Greater functional diversity and redundancy of coral endolithic microbiomes align with lower coral bleaching susceptibility. *The ISME Journal*, *16*(10), 2406-2420. https://doi.org/10.1038/s41396-022-01283-y

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, *17*(4), 540-552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925-1927. https://doi.org/10.1093/bioinformatics/btz848

Chen, R.-W., Li, Z., Huang, J., Liu, X., Zhu, W., Li, Y., Wang, A., & Li, X. (2024). The community stability of Symbiodiniaceae and bacteria of different morphological corals and linkages to coral susceptibility to anthropogenic disturbance. *Coral Reefs*, *43*(2), 467-481. https://doi.org/10.1007/s00338-024-02475-y

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Chimienti, G., De Padova, D., Mossa, M., & Mastrototaro, F. (2020). A mesophotic black coral forest in the Adriatic Sea. *Scientific Reports*, *10*(1), 8504. https://doi.org/10.1038/s41598-020-65266-9

Coker, D. J., Pratchett, M. S., & Munday, P. L. (2012). Influence of coral bleaching, coral mortality and conspecific aggression on movement and distribution of coral-dwelling fish. *Journal of Experimental Marine Biology and Ecology*, *414-415*, 62-68. https://doi.org/10.1016/j.jembe.2012.01.014

Doyle, S. M., Self, M. J., Hayes, J., Shamberger, K. E. F., Correa, A. M. S., Davies, S. W., Santiago-Vázquez, L. Z., & Sylvan, J. B. (2022). Microbial Community Dynamics Provide Evidence for Hypoxia during a Coral Reef Mortality Event. *Applied and Environmental Microbiology*, *88*(9), e00347-22. https://doi.org/10.1128/aem.00347-22

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195-e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792-1797. https://doi.org/10.1093/nar/gkh340

Fraley, C. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, *41*(8), 578-588. https://doi.org/10.1093/comjnl/41.8.578

Freiwald, A., & Roberts, J. M. (2005). *Cold-water corals and ecosystems*.

Glasl, B., Robbins, S., Frade, P. R., Marangon, E., Laffy, P. W., Bourne, D. G., & Webster, N. S. (2020). Comparative genome-centric analysis reveals seasonal variation in the function of coral reef microbiomes. *The ISME Journal*, *14*(6), Article 6. https://doi.org/10.1038/s41396-020-0622-6

Hegg, D. A., Radke, L. F., & Hobbs, P. V. (1991). Measurements of Aitken nuclei and cloud condensation nuclei in the marine atmosphere and their relation to the DMS-Cloud-climate hypothesis. *Journal of Geophysical Research*, *96*(D10), 18727-18733. https://doi.org/10.1029/91JD01870

Hou, X.-M., Xu, R.-F., Gu, Y.-C., Wang, C.-Y., & Shao, C.-L. (2015). Biological and Chemical Diversity of Coral-Derived Microorganisms. *Current Medicinal Chemistry*, *22*(32), 3707-3762.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2017). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *bioRxiv*. https://doi.org/10.1101/225342

Karlicki, M., Antonowicz, S., & Karnkowska, A. (2022). Tiara: Deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, *38*(2), 344-350. https://doi.org/10.1093/bioinformatics/btab672

Khan, S., Jung, H., Park, H., & Jeon, C. (2021). Maritimibacter harenae sp. nov. and Sneathiella litorea sp. nov.: Members of Alphaproteobacteria isolated from sea sand. *Antonie van Leeuwenhoek*, *114*. https://doi.org/10.1007/s10482-021-01559-x

Lee, O. O., Tsoi, M. M. Y., Li, X., Wong, P.-K., & Qian, P.-Y. (2007). Thalassococcus halodurans gen. Nov., sp. Nov., a novel halotolerant member of the Roseobacter clade isolated from the marine sponge Halichondria panicea at Friday Harbor, USA. *International Journal of*

Systematic and Evolutionary Microbiology, *57*(8), 1919-1924. https://doi.org/10.1099/ijs.0.64801-0

Levy Karin, E., Mirdita, M., & Söding, J. (2020). MetaEuk—Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, *8*(1), 48--48. https://doi.org/10.1186/s40168-020-00808-x

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674-1676. https://doi.org/10.1093/bioinformatics/btv033

Li, J., Chen, Q., Zhang, S., Huang, H., Yang, J., Tian, X.-P., & Long, L.-J. (2013). Highly Heterogeneous Bacterial Communities Associated with the South China Sea Reef Corals Porites lutea, Galaxea fascicularis and Acropora millepora. *PLoS ONE*, *8*(8), e71301. https://doi.org/10.1371/journal.pone.0071301

Lishchenko, F. V., Burmistrova, Y. A., Petrochenko, R. A., Nguyen, T. H., & Britayev, T. A. (2024). Seasonal bleaching and partial mortality of Pocillopora verrucosa corals of the coast of central Vietnam. *Frontiers in Marine Science*, *11*. https://doi.org/10.3389/fmars.2024.1338464

Littman, R. A., Willis, B. L., Pfeffer, C., & Bourne, D. G. (2009). Diversities of coral-associated bacteria differ with location, but not species, for three acroporid corals on the Great Barrier Reef. *FEMS Microbiology Ecology*, *68*(2), 152-163. https://doi.org/10.1111/j.1574-6941.2009.00666.x

Liu, C.-M., Luo, R., & Lam, T.-W. (2014). *GPU-Accelerated BWT Construction for Large Collection of Short Reads* (arXiv:1401.7457). arXiv. http://arxiv.org/abs/1401.7457

Luo, D., Wang, X., Feng, X., Tian, M., Wang, S., Tang, S.-L., Ang, J., Yan, A., & Luo, H. (2021). Population differentiation of Rhodobacteraceae along with coral compartments. *The ISME Journal*, *15*(11), 3286-3302. https://doi.org/10.1038/s41396-021-01009-6

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1997). The RDP (Ribosomal Database Project). *Nucleic Acids Research*, *25*(1), 109-110. https://doi.org/10.1093/nar/25.1.109

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647-4654. https://doi.org/10.1093/molbev/msab199

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538

Messyasz, A., Maher, R. L., Meiling, S. S., & Thurber, R. V. (2021). Nutrient Enrichment Predominantly Affects Low Diversity Microbiomes in a Marine Trophic Symbiosis between Algal Farming Fish and Corals. *Microorganisms*, *9*(9), Article 9. https://doi.org/10.3390/microorganisms9091873

Miller, R. J., Hocevar, J., Stone, R. P., & Fedorov, D. V. (2012). Structure-Forming Corals and Sponges and Their Use as Fish Habitat in Bering Sea Submarine Canyons. *PloS One*, *7*(3), e33885-e33885. https://doi.org/10.1371/journal.pone.0033885

Mohamed, A. R., Ochsenkühn, M. A., Kazlak, A. M., Moustafa, A., & Amin, S. A. (2023). The coral microbiome: Towards an understanding of the molecular mechanisms of coral–

microbiota interactions. *FEMS Microbiology Reviews*, *47*(2), fuad005. https://doi.org/10.1093/femsre/fuad005

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996-1004. https://doi.org/10.1038/nbt.4229

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043-1055. https://doi.org/10.1101/gr.186072.114

Pierrejean, M., Grant, C., Neves, B. de M., Chaillou, G., Edinger, E., Blanchet, F. G., Maps, F., Nozais, C., & Archambault, P. (2020). Influence of Deep-Water Corals and Sponge Gardens on Infaunal Community Composition and Ecosystem Functioning in the Eastern Canadian Arctic. *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.00495

Pootakham, W., Mhuantong, W., Yoocha, T., Putchim, L., Jomchai, N., Sonthirod, C., Naktang, C., Kongkachana, W., & Tangphatsornruang, S. (2019). Heat-induced shift in coral microbiome reveals several members of the Rhodobacteraceae family as indicator species for thermal stress in Porites lutea. *MicrobiologyOpen*, *8*(12), e935. https://doi.org/10.1002/mbo3.935

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, *26*(7), 1641-1650. https://doi.org/10.1093/molbev/msp077

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Pujalte, M. J., Lucena, T., Rodrigo-Torres, L., & Arahal, D. R. (2018). Comparative Genomics of Thalassobius Including the Description of Thalassobius activus sp. Nov., and Thalassobius autumnalis sp. Nov. *Frontiers in Microbiology*, *8*, 2645-2645. https://doi.org/10.3389/fmicb.2017.02645

Pujalte, M. J., Lucena, T., Ruvira, M. A., Arahal, D. R., & Macián, M. C. (2014). The Family Rhodobacteraceae. En E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes* (pp. 439-512). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30197-1_377

Rohwer, F., Breitbart, M., Jara, J., Azam, F., & Knowlton, N. (2001). Diversity of bacteria associated with the Caribbean coral Montastraea franksi. *Coral Reefs*, *20*(1), 85-91. https://doi.org/10.1007/s003380100138

Rohwer, F., Seguritan, V., Azam, F., & Knowlton, N. (2002). Diversity and distribution of coral-associated bacteria. *Marine Ecology Progress Series*, *243*, 1-10. https://doi.org/10.3354/meps243001

Rubio-Portillo, E., Santos, F., Martínez-García, M., de los Ríos, A., Ascaso, C., Souza-Egipsy, V., Ramos-Esplá, A. A., & Anton, J. (2016). Structure and temporal dynamics of the bacterial communities associated to microhabitats of the coral Oculina patagonica.

Environmental Microbiology, *18*(12), 4564-4578. https://doi.org/10.1111/1462-2920.13548

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20-D26. https://doi.org/10.1093/nar/gkab1112

Sekar, R., Mills, D. K., Remily, E. R., Voss, J. D., & Richardson, L. L. (2006). Microbial Communities in the Surface Mucopolysaccharide Layer and the Black Band Microbial Mat of Black Band-Diseased Siderastrea siderea. *Applied and Environmental Microbiology*, *72*(9), 5963-5973. https://doi.org/10.1128/AEM.00843-06

Siboni, N., Ben-Dov, E., Sivan, A., & Kushmaro, A. (2008). Global distribution and diversity of coral-associated Archaea and their possible role in the coral holobiont nitrogen cycle. *Environmental Microbiology*, *10*(11), 2979-2990. https://doi.org/10.1111/j.1462-2920.2008.01718.x

Sindhu, M. R., & Potty, V. P. (2015). Biodegradation of Cashew Nut Shell Liquid by Delftia Acidovorans and Pseudomonas Aeruginosa Isolated from Marine Environment. *International Journal of Agriculture Environment & Biotechnology*, *8*(4), 837-846. https://doi.org/10.5958/2230-732X.2015.00094.7

Song, W.-Z., & Thomas, T. (2017). Binning_refiner: Improving genome bins through the combination of different binning programs. *Bioinformatics*, *33*(12), 1873-1875. https://doi.org/10.1093/bioinformatics/btx086

Sunagawa, S., Woodley, C. M., & Medina, M. (2010). Threatened Corals Provide Underexplored Microbial Habitats. *PLoS ONE*, *5*(3), e9554. https://doi.org/10.1371/journal.pone.0009554

Terzin, M., Paletta, M. G., Matterson, K., Coppari, M., Bavestrello, G., Abbiati, M., Bo, M., & Costantini, F. (2021). Population genomic structure of the black coral Antipathella subpinnata in Mediterranean Vulnerable Marine Ecosystems. *Coral Reefs*, *40*(3), 751-766. https://doi.org/10.1007/s00338-021-02078-x

Tian, R., & Imanian, B. (2023). VBCG: 20 validated bacterial core genes for phylogenomic analysis with high fidelity and resolution. *Microbiome*, *11*(1), 247. https://doi.org/10.1186/s40168-023-01705-9

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

Van Oppen, M. J. H., & Lough, J. M. (2018). *Coral bleaching: Patterns, processes, causes and consequences* (Second edition). Springer.

Warnow, T. (2015). SATe-Enabled Phylogenetic Placement. En K. E. Nelson (Ed.), *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools* (pp. 619-621). Springer US. https://doi.org/10.1007/978-1-4899-7478-5_711

Wen, A., Fegan, M., Hayward, C., Chakraborty, S., & Sly, L. I. (1999). Phylogenetic relationships among members of the Comamonadaceae, and description of Delftia acidovorans (den Dooren de Jong 1926 and Tamaoka et al. 1987) gen. Nov., comb. Nov. *International Journal of Systematic and Evolutionary Microbiology*, *49*(2), 567-576. https://doi.org/10.1099/00207713-49-2-567

Yildiz, H., Sünnetçioğlu, A., Ekin, S., Baran, A. İ., Özgökçe, M., Aşker, S., Üney, İ., Turgut, E., & Akyüz, S. (2019). Delftia acidovorans pneumonia with lung cavities formation. *Colombia Medica (Cali, Colombia)*, *50*(3), 215-221. https://doi.org/10.25100/cm.v50i3.4025

Yu, F., Zhang, Y., Cheng, C., Wang, W., Zhou, Z., Rang, W., Yu, H., Wei, Y., Wu, Q., & Zhang, Y. (2020). Poly(A)-seq: A method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLOS ONE*, *15*(6), e0234696. https://doi.org/10.1371/journal.pone.0234696

8. SUPPLEMENTARY MATERIALS

Code S1. Script that manages and executes Mehagit (Fig. 3.1.C) to Metawap (Fig. 3.1.F) programs.

```
#!/usr/bin/env python
import time
import glob
import argparse
import gzip
import os
from pathlib import Path
import re
import subprocess
import numpy as np
import pandas as pd
import htcondor

# GENERAL_DIRS:
MOTHER_DIR = '/data/coral/scratch/laura_workflow/test10/'
DATA_DIRECTORY = MOTHER_DIR + 'data/'
OUTPUT_DIRECTORY = MOTHER_DIR + 'data/metagenomes/'
COMPRESS_RATIO = 4
WALLTIME_RATIO = 6/2.5 # hours/GB
MAXWALLTIME = 24 # hours
ALLOWED_CPUS = [2**i for i in range(4)]
PROGRAM = MOTHER_DIR + 'scripts/'
LOG_FOLDER = MOTHER_DIR + 'processing/'

# MEGAHIT_DIRS:
MEGAHIT_DAT_DIR = DATA_DIRECTORY + 'raw_mod'
MEGAHIT_OUT_DIR = OUTPUT_DIRECTORY + 'megahit'
MEGAHIT_PROGRAM = PROGRAM + '1-megahit/ngs_assembly.sh'
MEGAHIT_LOG_DIR = LOG_FOLDER + 'megahit/logs'

# NAMES DIR
NAMES_PROGRAM = PROGRAM + 'others/names.sh'

# TIARA DIRS
TIARA_INP_DIR = Path(f'{DATA_DIRECTORY}metagenomes/megahit/batch')
TIARA_OUT_DIR = Path(f'{DATA_DIRECTORY}metagenomes/tiara')
TIARA_PROCESSING_FOLDER = Path(__file__).parent.resolve()
TIARA_PROGRAM = (TIARA_PROCESSING_FOLDER / 'run_tiara.sh')
TIARA_LOG_DIR = LOG_FOLDER + 'tiara/logs'

# LOOP DIR
LOOP_PROGRAM = PROGRAM + 'others/post_tiara.sh'


# METAWRAP
MW_raw_DIR = DATA_DIRECTORY + 'raw'
MW_datasets_DIR = PROGRAM + '3-metawrap'
MW_1_STEP = PROGRAM + '3-metawrap/first_step.sh'
MW_2_STEP = PROGRAM + '3-metawrap/second_step.sh'
MW_3_STEP = PROGRAM + '3-metawrap/metawrap.sub'
MW_4_STEP = PROGRAM + '3-metawrap/fourth_step.sh'

# FUNTIONS

def def_parser():

    parser = argparse.ArgumentParser()

    subparsers = parser.add_subparsers(dest='command')

    # add subparser to get file info
    parser_a = subparsers.add_parser('get_file_info', help='a help')
    parser_a.add_argument('--output', '-o', help='csv to store the file information to')

    # add subparser to get dataset resources need
    parser_b = subparsers.add_parser('resources')
    parser_b.add_argument('--input', '-i', help='csv with file information')
    parser_b.add_argument('--output', '-o', help='csv with dataset resources information')

    # add subparser to get dataset resources need
    parser_c = subparsers.add_parser('submit')
    parser_c.add_argument('--input', '-i', help='csv with dataset and resources information')

    return parser


def get_fastq_file_info(fastq_file):

    # get the first 4 lines of the file containing one read
    # assuming the structure for each read is the same
    with gzip.open(fastq_file,'rt') as f:
        lines = [f.readline() for i in range(4)]

        # If we want to count the number of lines in the file we can do this
        # But it takes loooooong
        #n_lines = 4
        #for l in f:
        #    n_lines += 1

    # Get read length
    # assuming all reads are equal length
    read_length = len(lines[1]) - 1

    # Quick and dirty approximation to get the number of reads in a file
    # assuming fixed compression ratio
    n_bytes_read = sum(map(len, lines))
    n_bytes_file = os.path.getsize(fastq_file)
    n_reads = int(n_bytes_file*COMPRESS_RATIO/n_bytes_read)

    return((fastq_file, read_length, n_reads))


def get_file_info(output_file):

    l = glob.glob(f'{MEGAHIT_DAT_DIR}/*fastq*')
    file_info = [get_fastq_file_info(fastq_file) for fastq_file in l]
    df = pd.DataFrame.from_records(file_info, columns=['filename', 'read_length', 'n_reads'])
    df.sort_values(by=['filename', 'n_reads'], inplace=True)
    df.to_csv(output_file, index=False)

def compute_dataset_resources(file_info, file_resources):

    df_in = pd.read_csv(file_info)
    df_in['dataset'] = df_in.filename.str.extract(MEGAHIT_DAT_DIR + r'/(.*)_[12]\.trim\.fastq\.gz')
    df_in['data_GB'] = df_in.read_length*df_in.n_reads*2/1024/1024/1024

    df_resources = pd.DataFrame(df_in.groupby('dataset').data_GB.sum())
    df_resources['request_memory_GB'] = np.clip(2**np.round(np.log2(df_resources.data_GB)), 4, 64)
    df_resources['cputime_h'] = df_resources['data_GB']*WALLTIME_RATIO

    df_resources['request_cpu'] = np.clip(2**np.ceil(np.log2(df_resources.cputime_h / MAXWALLTIME)), 1, 8)
    df_resources['walltime_h'] = (df_resources.cputime_h/df_resources.request_cpu).astype('int')

    df_resources.to_csv(file_resources)


def build_default_resources():
    '''Build a dataframe with default resources (8 cores, 64GB RAM)
    to request for each megahit job'''

    # Identify the datasets in the data directory
    # by listing the files and getting the identifiers of each sample
    data_path = Path(MEGAHIT_DAT_DIR)
    fastq_list = data_path.glob('*fastq')
    extract_dataset = lambda x: re.match(r'(.*)_[12]\.trim\.fastq', x.name)[1]
    datasets = list(set(map(extract_dataset, fastq_list)))

    # Build the dataframe with default resources
    df = pd.DataFrame({'dataset': datasets,
        'request_memory_GB': 64,
        'request_cpu': 8
        })

    return df


def submit(dataset_resources):
    ids_list = []
    if dataset_resources:
        df = pd.read_csv(dataset_resources)
    else:
        df = build_default_resources()

    for ind, row in df.iterrows():

        dataset = row['dataset']

        # If the output folder already exists raise a warning and continue
        # it means that the dataset has already been processed
        output_folder = os.path.join(MEGAHIT_OUT_DIR, dataset)
        if os.path.exists(output_folder):
            print(f'Output folder for dataset {dataset} already exists, skipping it')
            continue

        id1 = submit_single(row)
        ids_list.append(id1)
    #print(ids_list)
    return ids_list

def submit_single(job_description):

    ds = job_description['dataset']
    output_folder = os.path.join(MEGAHIT_OUT_DIR, ds)
    ngs_assembly_job = htcondor.Submit({
        "executable": MEGAHIT_PROGRAM,
        "output": f"{MEGAHIT_LOG_DIR}/megahit_{ds}_$(ClusterId).out",
        "error": f"{MEGAHIT_LOG_DIR}/megahit_{ds}_$(ClusterId).err",
        "log": f"{MEGAHIT_LOG_DIR}/megahit_{ds}_$(ClusterId).log",
        "request_cpus": job_description['request_cpu'],
        "request_memory": f"{job_description['request_memory_GB']}GB",
        "+experiment": '"coral\"',
        "+flavour": '"long"',
        "environment": f'DATASET={ds}'
    })

    schedd = htcondor.Schedd()
    submit_result = schedd.submit(ngs_assembly_job)
    print(submit_result.cluster())
    id = submit_result.cluster()
    return id


def main(args):

    if args.command == 'get_file_info':
        get_file_info(args.output)
    elif args.command == 'resources':
        compute_dataset_resources(args.input, args.output)
    elif args.command == 'submit':
        ids_list2 = submit(args.input)
        #print(ids_list2)
        return ids_list2


def exec_bash(script_path):
    try:
        resultado = subprocess.run(['/bin/bash', script_path], check=True, text=True, capture_output=True)
        print(f"Bash:\n{resultado.stdout}")

    except subprocess.CalledProcessError as e:
        print(f"Error Bash: {e}")
        print(f"Error:\n{e.stderr}")


def exec_bash_1arg(programa, argumento1):
    try:
        resultado = subprocess.run([programa, argumento1], check=True, text=True, capture_output=True)
        print(f"Bash:\n{resultado.stdout}")

    except subprocess.CalledProcessError as e:
        print(f"Error Bash: {e}")
        print(f"Error:\n{e.stderr}")

def exec_bash_2arg(programa, argumento1, argumento2):
    try:
        resultado = subprocess.run([programa, argumento1, argumento2], check=True, text=True, capture_output=True)
        print(f":Bash:\n{resultado.stdout}")

    except subprocess.CalledProcessError as e:
        print(f"Error Bash: {e}")
        print(f"Error:\n{e.stderr}")


def sub_condor_submit(script_path):
    try:
        resultado = subprocess.run(['condor_submit', script_path], check=True, text=True, capture_output=True)
        #print(f"Salida del script de Bash:\n{resultado.stdout}")

    except subprocess.CalledProcessError as e:
        print(f"Error Bash: {e}")
        print(f"Error:\n{e.stderr}")

    frase = resultado.stdout
    match = re.search(r'cluster (\d+)\.', frase)
    job_id = match.group(1)
    return (job_id)

def tiara_submit():
    tiara_ids_list = []
    for input_file in TIARA_INP_DIR.glob('*contigs_fixed.fa'):

        # check if the file has already been processed
        contigs_file_name = input_file.name
        base_name = input_file.name.split('.')[0]
        if len(list(TIARA_OUT_DIR.glob(f'{base_name}*'))) > 0:
            print(f'file {contigs_file_name} has already been processed')
            continue
        id = tiara_submit_single(input_file)
        tiara_ids_list.append(id)
    #print(tiara_ids_list)
    return(tiara_ids_list)

def tiara_submit_single(input_file):

    ds = input_file.name.split('.')[0]
    contigs_file_name = input_file.name

    tiara_job = htcondor.Submit({
        "executable": str(TIARA_PROGRAM),
        "output": f"{TIARA_LOG_DIR}/{ds}_$(ClusterId).out",
        "error": f"{TIARA_LOG_DIR}/{ds}_$(ClusterId).err",
        "log": f"{TIARA_LOG_DIR}/{ds}_$(ClusterId).log",
        "request_cpus": 1.0,
        "request_memory": "8.0GB",
        "+experiment": '"coral"',
        "+flavour": '"long"',
        "environment": f'CONTIGS_FILE={contigs_file_name}'
    })

    #print(tiara_job)
    #return

    schedd = htcondor.Schedd()
    submit_result = schedd.submit(tiara_job)
    #print(submit_result.cluster())
    return(submit_result.cluster())


# CODE


# MEGAHIT

if __name__ == '__main__':
    print('>>i MEGAHIT_1/2 (main) has started (1/8): assembling reads into contigs')
    parser = def_parser()
    args = parser.parse_args()
    ids_list3 = main(args)
    #print(ids_list3)

i = 'running jobs'
while i == 'running jobs':
    queue = subprocess.run(['condor_q'], capture_output=True, text=True)
    queue_out = queue.stdout
    i = 'jobs finished'
    for id in ids_list3:
        if str(id) in queue_out:
            i = 'running jobs'
    if i == 'running jobs':
        print('not all jobs finished')
        print(queue_out)
        print(ids_list3)
        time.sleep(5)
        continue
    else:
        print('all jobs finished')
        print('>>f MEGAHIT_1/2 (main) has finished')
        break


# NAMES
if __name__ == "__main__":
    print('>>i MEGAHIT_2/2 (names) has started (2/8): transfering files from megahit to batch folder')
    exec_bash(NAMES_PROGRAM)
    print('>>f MEGAHIT_2/2 (names) has finished')

# TIARA
print('>>i TIARA_1/2 (main) has started (3/8): domain classification')
if __name__ == '__main__':
    tiara_ids_DEFlist = tiara_submit()
    print(tiara_ids_DEFlist)

i = 'running jobs'
while i == 'running jobs':
    queue = subprocess.run(['condor_q'], capture_output=True, text=True)
    queue_out = queue.stdout
    i = 'jobs finished'
    for id in tiara_ids_DEFlist:
        if str(id) in queue_out:
            i = 'running jobs'
    if i == 'running jobs':
        print('not all jobs finished')
        print(queue_out)
        print(tiara_ids_DEFlist)
        time.sleep(5)
        continue
    else:
        print('all jobs finished')
        print('>>f TIARA_1/2 (main) has finished')
        break


# LOOP WORKFLOW


if __name__ == "__main__":
    print('>>i TIARA_2/2 (loop) has started (4/8): preparing files for metawrap assignment')
    exec_bash(LOOP_PROGRAM)
print('>>f TIARA_2/2 (loop) has finished')


# METAWRAP

if __name__ == "__main__":
    print('>>i METAWRAP_1/4 has started (5/8): coping files from raw_mod to raw folders')
    exec_bash_1arg(MW_1_STEP, DATA_DIRECTORY)
    print('>>f METAWRAP_1/4 has finished')
    print('----------------------------------')
    print('>>i METAWRAP_2/4 has started (6/8): building datasets.txt file')
    exec_bash_2arg(MW_2_STEP, MW_raw_DIR, MW_datasets_DIR)
    print('>>f METAWRAP_2/4 has finished')
    print('----------------------------------')
    print('>>i METAWRAP_3/4 has started (7/8): binning contigs')
    mw_job_id = sub_condor_submit(MW_3_STEP)

import time
i = 'running jobs'
while i == 'running jobs':
    queue = subprocess.run(['condor_q'], capture_output=True, text=True)
    queue_out = queue.stdout
    i = 'jobs finished'
    if mw_job_id in queue_out:
        i = 'running jobs'
    if i == 'running jobs':
        print('not all jobs finished')
        print(queue_out)
        print(mw_job_id)
        time.sleep(5)
        continue
    else:
        print('all jobs finished')
        print('>>f METAWRAP_3/4 has finished')
        break


if __name__ == "__main__":
    print('>>i METAWRAP_4/4 has started (8/8): organising MAGs in All_Bins directory')
    exec_bash_1arg(MW_4_STEP, OUTPUT_DIRECTORY)
    print('>>f METAWRAP_4/4 has finished')

print('ALL DONE !!!')
```

Table S1 | Metrics of complete eukaryotic MAGs.

MAG name	Coral specie	No. of contigs	Contig N50 (bp)	Completeness (%)	Single BMG (%)	Duplicated BMG (%)	Sin./Dupl. BMG	No. BMG
DRR235375_Fanc_01	*F. ancora*	23789	3050	88.5	88.1	0.4	220.3	954
DRR235383_Fanc_02	*F. ancora*	19871	3162	88.4	87.9	0.5	175.8	954
DRR235381_Fanc_03	*F. ancora*	22497	3072	88.3	88.1	0.2	440.5	954
DRR235373_Fanc_04	*F. ancora*	23960	3115	88.2	87.6	0.6	146.0	954
DRR235378_Fanc_05	*F. ancora*	22578	3049	88.2	87.6	0.6	146.0	954
DRR235382_Fanc_06	*F. ancora*	21399	3033	88	87.3	0.7	124.7	954
DRR235379_Fanc_07	*F. ancora*	22141	3071	86.9	86.4	0.5	172.8	954
DRR235376_Fanc_08	*F. ancora*	24898	2985	86.8	86.2	0.6	143.7	954
DRR235390_Fanc_09	*F. ancora*	25279	3009	86.3	85.1	1.2	70.9	954
DRR235387_Fanc_10	*F. ancora*	25861	3035	86.1	85.3	0.8	106.6	954
DRR235380_Fanc_11	*F. ancora*	21919	3072	85.8	85.5	0.3	285.0	954
DRR235392_Fanc_12	*F. ancora*	26376	2952	85.8	85.3	0.5	170.6	954
DRR235377_Fanc_13	*F. ancora*	24078	3101	85.7	85.2	0.5	170.4	954
DRR235393_Fanc_14	*F. ancora*	20699	2957	85.6	85.3	0.3	284.3	954
DRR235389_Fanc_15	*F. ancora*	25794	2948	85.4	85.1	0.3	283.7	954
DRR235384_Fanc_16	*F. ancora*	35180	2448	84.9	82.7	2.2	37.6	954
DRR235395_Fanc_17	*F. ancora*	23870	2945	84.9	84.2	0.7	120.3	954
DRR235374_Fanc_18	*F. ancora*	22803	3024	84.6	84	0.6	140.0	954
DRR235391_Fanc_19	*F. ancora*	25231	2944	84.5	83.9	0.6	139.8	954
DRR235386_Fanc_20	*F. ancora*	25740	2905	84.4	83.9	0.5	167.8	954
DRR235388_Fanc_21	*F. ancora*	24944	2927	84.2	83.5	0.7	119.3	954
DRR235394_Fanc_22	*F. ancora*	29610	2590	84.1	82.8	1.3	63.7	954
DRR235385_Fanc_23	*F. ancora*	25608	2964	84	83.4	0.6	139.0	954
ERR571494_Plut_24	*P. lutea*	71378	2303	79.2	71.4	7.8	9.2	255
ERR571494_Plut_25	*P. lutea*	71378	2303	79.2	71.4	7.8	9.2	255
DRR235373_Fanc_26	*F. ancora*	17088	3105	68.7	68.4	0.3	228.0	954
DRR235394_Fanc_27	*F. ancora*	17779	2805	67.5	66.8	0.7	95.4	954
DRR235384_Fanc_28	*F. ancora*	15240	2999	67.2	67	0.2	335.0	954
DRR235385_Fanc_29	*F. ancora*	14857	3328	64.8	64.5	0.3	215.0	954
DRR235388_Fanc_30	*F. ancora*	17212	2882	62.9	62.5	0.4	156.3	954
DRR235374_Fanc_31	*F. ancora*	15891	2957	62.6	62.1	0.5	124.2	954
DRR235393_Fanc_32	*F. ancora*	13908	2868	59.6	59.3	0.3	197.7	954
ERR2192880_Pspe_33	*P. speciosa*	57367	2487	57.8	56.4	1.4	40.3	954
ERR2192880_Pspe_34	*P. speciosa*	57367	2487	57.8	56.4	1.4	40.3	954
ERR2188869_Aten_35	*A. tenius*	47158	2652	54.5	52.5	2	26.3	255
ERR2188869_Aten_36	*A. tenius*	47158	2652	54.5	52.5	2	26.3	255

F. ancora: Fimbriaphyllia_ancora, P. lutea: Porites lutea, P. speciosa: Pachyseris speciosa, A. tenuis: Acropora tenuis, BMG: BUSCO Marker Genes.

Table S2 | All individual classifications for complete prokaryotic MAGs.

	Domain	Phylum	Class	Order	Family	Genus	Specie
MAG	**ERR2191367_Gfas_01**						
BAT	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales			
GTDB	Bacteria	Pseudomonadota	Gammaproteobacteria	Burkholderiales	Burkholderiaceae B	Comamonas	*Comamonas acidovorans*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2190369_Ffun_02**						
BAT	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales			
GTDB	Bacteria	Pseudomonadota	Gammaproteobacteria	Burkholderiales	Burkholderiaceae B	Comamonas	*Comamonas acidovorans*
ENA	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
RDP	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
SLV	Bacteria	Proteobacteria	Gammaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
MAG	**ERR571459_Plut_03**						
BAT	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales			
GTDB	Bacteria	Pseudomonadota	Gammaproteobacteria	Burkholderiales	Burkholderiaceae B	Comamonas	*Comamonas acidovorans*
ENA	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
RDP	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
SLV	Bacteria	Proteobacteria	Gammaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
MAG	**ERR571459_Plut_04**						
BAT	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales			
GTDB	Bacteria	Pseudomonadota	Gammaproteobacteria	Burkholderiales	Burkholderiaceae B	Comamonas	*Comamonas acidovorans*
ENA	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
RDP	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
SLV	Bacteria	Proteobacteria	Gammaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
MAG	**ERR2190369_Ffun_05**						
BAT	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales			
GTDB	Bacteria	Pseudomonadota	Gammaproteobacteria	Burkholderiales	Burkholderiaceae B	Comamonas	*Comamonas acidovorans*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216064_Ccrux_06**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216067_Ccrux_07**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216066_Ccrux_08**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216065_Ccrux_09**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216065_Ccrux_10**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216067_Ccrux_11**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		

	Domain	Phylum	Class	Order	Family	Genus	Specie
MAG	**ERR2216064_Ccrux_12**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria				
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	UBA8366	GCA-2696645		
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571463_Plut_13**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Thalassobius	*Thalassobius autumnalis*
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Thalassobius	*Thalassobius autumnalis*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571459_Plut_14**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Thalassococcus	
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Thalassococcus	*Thalassococcus halodurans*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571459_Plut_15**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Thalassococcus	
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Thalassococcus	*Thalassococcus halodurans*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571459_Plut_16**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571459_Plut_17**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571463_Plut_18**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR571463_Plut_19**						
BAT	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
GTDB	Bacteria	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae	Maritimibacter	*Maritimibacter harenae*
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216065_Ccrux_20**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216067_Ccrux_21**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216067_Ccrux_22**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	-						
SLV	-						

	Domain	Phylum	Class	Order	Family	Genus	Specie
MAG	**ERR2216065_Ccrux_23**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216066_Ccrux_24**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216066_Ccrux_25**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	-						
SLV	-						
MAG	**ERR2216064_Ccrux_26**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216064_Ccrux_27**						
BAT	Bacteria	Proteobacteria					
GTDB	Bacteria	SAR324	SAR325	XTD2-FULL-50-16			
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216064_Ccrux_28**						
BAT	Bacteria						
GTDB	-						
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216065_Ccrux_29**						
BAT	Bacteria						
GTDB	-						
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216066_Ccrux_30**						
BAT	Bacteria						
GTDB	-						
ENA	-						
RDP	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Aestuariispira	
SLV	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Terasakiellaceae		
MAG	**ERR2216067_Ccrux_31**						
BAT	Bacteria						
GTDB	-						
ENA	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Polaribacter	
RDP	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Polaribacter	
SLV	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Polaribacter	

Table S3 | All individual classifications for complete eukaryotic MAGs.

MAG name	BAT classification	BLAST classification
DRR235375_Fanc_01	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235383_Fanc_02	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235381_Fanc_03	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235373_Fanc_04	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235378_Fanc_05	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235382_Fanc_06	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235379_Fanc_07	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235376_Fanc_08	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235390_Fanc_09	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235387_Fanc_10	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235380_Fanc_11	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235392_Fanc_12	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235377_Fanc_13	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235393_Fanc_14	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235389_Fanc_15	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235384_Fanc_16	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235395_Fanc_17	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235374_Fanc_18	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235391_Fanc_19	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235386_Fanc_20	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235388_Fanc_21	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Galaxea fascicularis genome assembly, chromosome: 6
DRR235394_Fanc_22	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235385_Fanc_23	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
ERR571494_Plut_24	N/A	Porites lutea genome assembly, chromosome: 11
ERR571494_Plut_25	N/A	Porites lutea genome assembly, chromosome: 11
DRR235373_Fanc_26	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235394_Fanc_27	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235384_Fanc_28	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235385_Fanc_29	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235388_Fanc_30	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235374_Fanc_31	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
DRR235393_Fanc_32	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	-
ERR2192880_Pspe_33	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Siderastrea siderea genome assembly, chromosome: 13
ERR2192880_Pspe_34	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia	Siderastrea siderea genome assembly, chromosome: 13
ERR2188869_Aten_35	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia, Astrocoeniina, Acroporidae, Acropora	Acropora palmata genome assembly, chromosome: 14
ERR2188869_Aten_36	Eukaryota, Opisthokonta, Eumetazoa, Anthozoa, Hexacorallia, Scleractinia, Astrocoeniina, Acroporidae, Acropora	Acropora palmata genome assembly, chromosome: 14

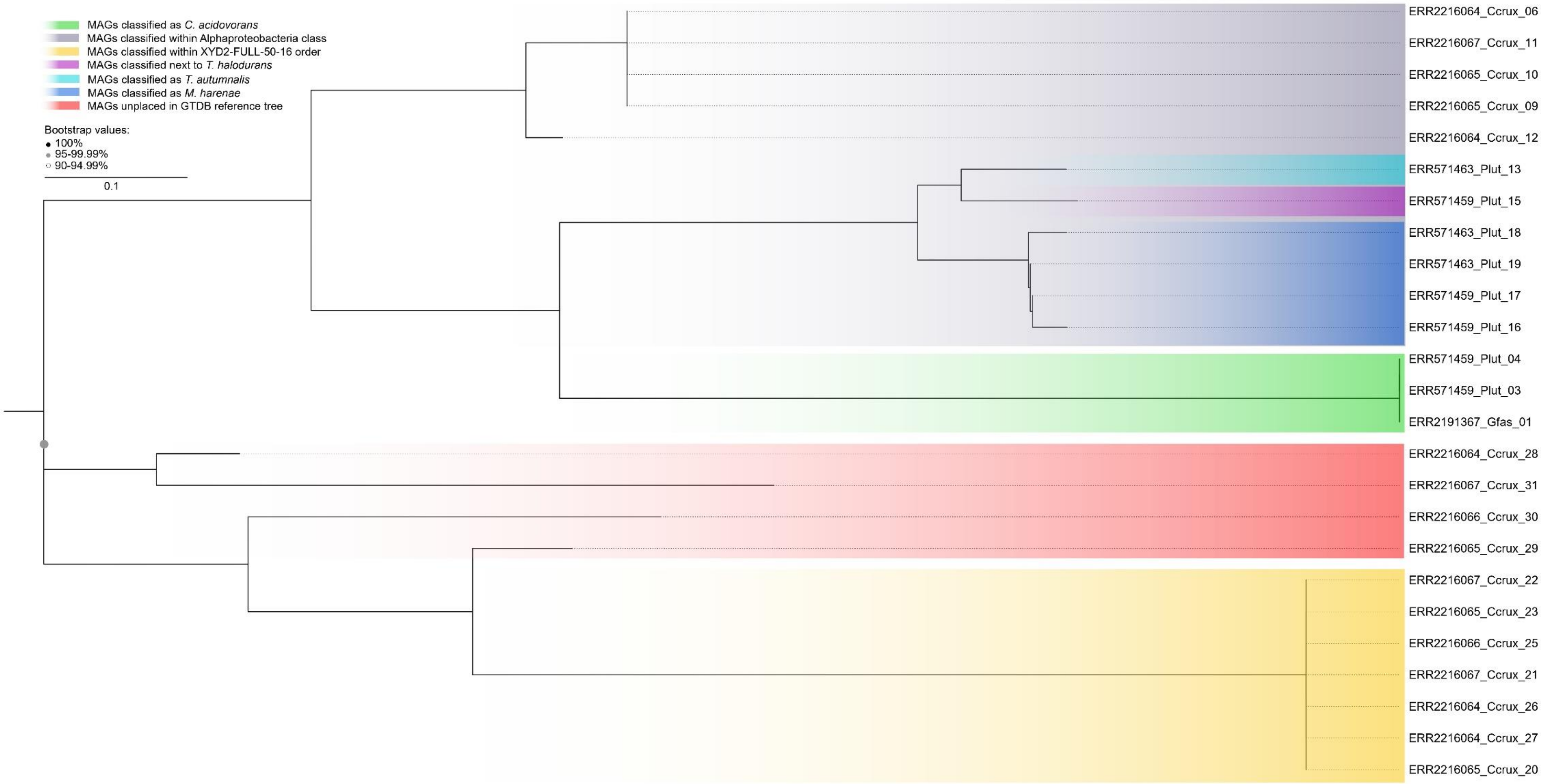

Figure S1. Phylogenetic tree built by VBCG.