BlueRemediomics

Project Number	**101082304**
Project Acronym	**BlueRemediomics**
Project Title	**BlueRemediomics: Harnessing the marine microbiome for novel sustainable biogenics and ecosystem services**
Funding Programme	**Horizon Europe**
Instrument	**RIA**
Project Start Date	**01/12/2022**
Duration of the Project	**48 months**
Deliverable Number and Name	**D1.3 – MGnify database-initial**
Work Package	**WP1**
Lead	**EMBL**
Deliverable Due Date	**31/05/2024**
Submission date	**30/05/2024**
Author(s)	**Robert Finn, Martin Beracochea, Lorna Richardson, Ekaterina Sakharova**
Dissemination Level	**Public**
Type	**Data**

Funded by the European Union

BlueRemediomics has received funding from the European Union's Horizon Europe Programme under Grant Agreement No. 101082304. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

UK Research and Innovation

UK Partners on **BlueRemediomics** are supported by UK Research and Innovation (UKRI) under the UK Government's Horizon Europe funding guarantee Grant No. IFS 10061678 (University College London); IFS 10055633 (The Chancellors Masters and Scholars of the University of Cambridge); IFS 10057167 (University of Aberdeen).

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

The Swiss Partner (Eidgenoessische Technische Hochschule Zuerich) on **BlueRemediomics** has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) under Contract No. 22.00384.

CONTENTS

MGnify Genomes marine catalogue v2.0

Building on Deliverable D1.1 which identified, assembled and analysed additional marine datasets, we utilised these assemblies as part of a focussed effort to derive new metagenome-assembled genomes (MAGs) from those and other datasets to enhance the diversity of the MGnify Genomes Marine catalogue. We employed a recently-developed bioinformatics pipeline (https://workflowhub.eu/workflows/884) (developed as part of the EU H2020-funded AtlantECO project; Grant Agreement Number 862923) for the generation of both prokaryotic and eukaryotic genomes at scale from metagenomic datasets. In total, 72 studies (pertaining to 3804 samples) were processed for MAG-generation by this pipeline resulting in 3279 prokaryotic genomes and 31 eukaryotic genomes that passed the quality threshold of completeness > 50% and contamination < 5%. All genomes generated from this pipeline have been submitted to the ENA MAG-layer (a specific deposition point for the archiving of MAGs), tagged with the relevant quality measurements in accordance with the GSC specified MiMAG checklist. The MAGs are associated with the underlying raw-reads, assembled contigs, and samples that they were generated from, thus allowing them to inherit the relevant environmental metadata, making them findable and reusable as a data resource for the research community. This data linkage, together with the submitted workflow promotes reproducible science, allowing others to either recreate these MAGs or to follow the same procedure.

In addition to these MGnify-generated MAGs, community-assembled marine MAGs were sourced from the ENA MAG-layer, such as the large study “Biosynthetic potential of the global ocean microbiome” submitted by BlueRemediomics partner ETHZ. Furthermore, we included a curated set of 13,338 marine genomes (comprising both isolates and MAGs) from MarDB, with the aim of generating an updated version of the marine catalogue that represents the fullest extent possible of the existing community knowledge on marine microbes.

As a result of these efforts, v2.0 of the MGnify Genomes marine catalogue was released, containing genomes from 1628 studies. This includes genomes from major sampling expeditions such as TARA, Malaspina, GO-Ship, and Geotraces, amongst others, resulting in a broad geographical distribution of samples represented in the catalogue (see Figure 1). In total 50,866 genomes (50,634 MAGs and 232 isolates) were included, which were in turn clustered into 13,223 species-level clusters, each represented by a cluster representative genome (see Figure 2). All contributing genomes were subject to a threshold of QS50 (QS, quality score, defined as completeness - 5 x contamination) to ensure the quality of the data represented in the catalogue (Figure 3). The catalogue was generated using the latest release (v2.3.0) of the MGnify genomes catalogue pipeline (available in GitHub and WorkflowHub).

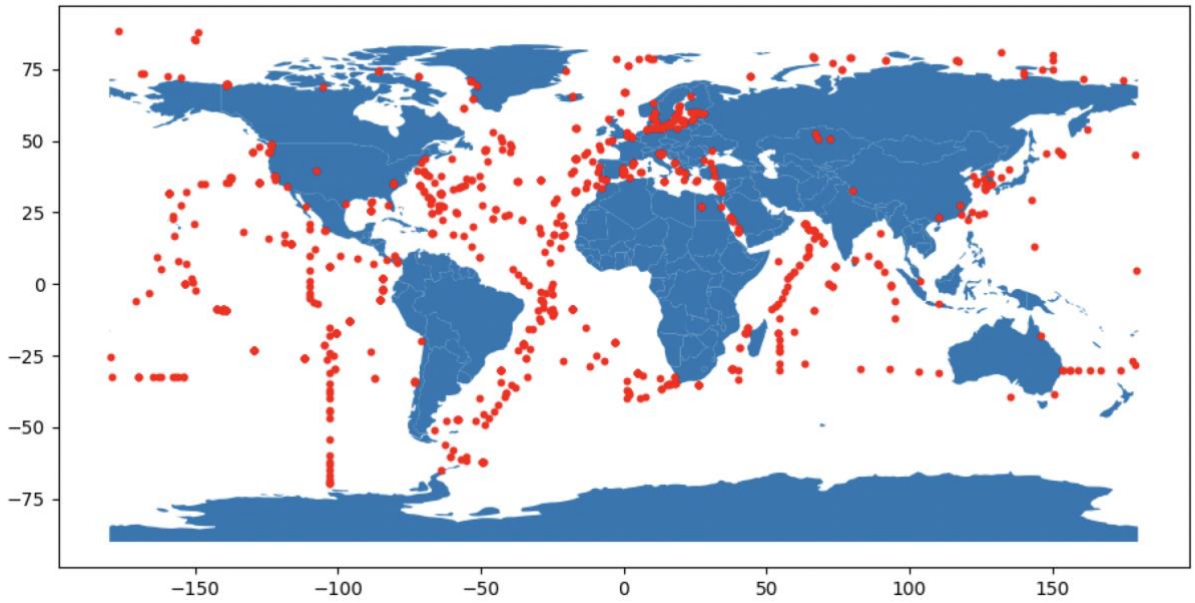
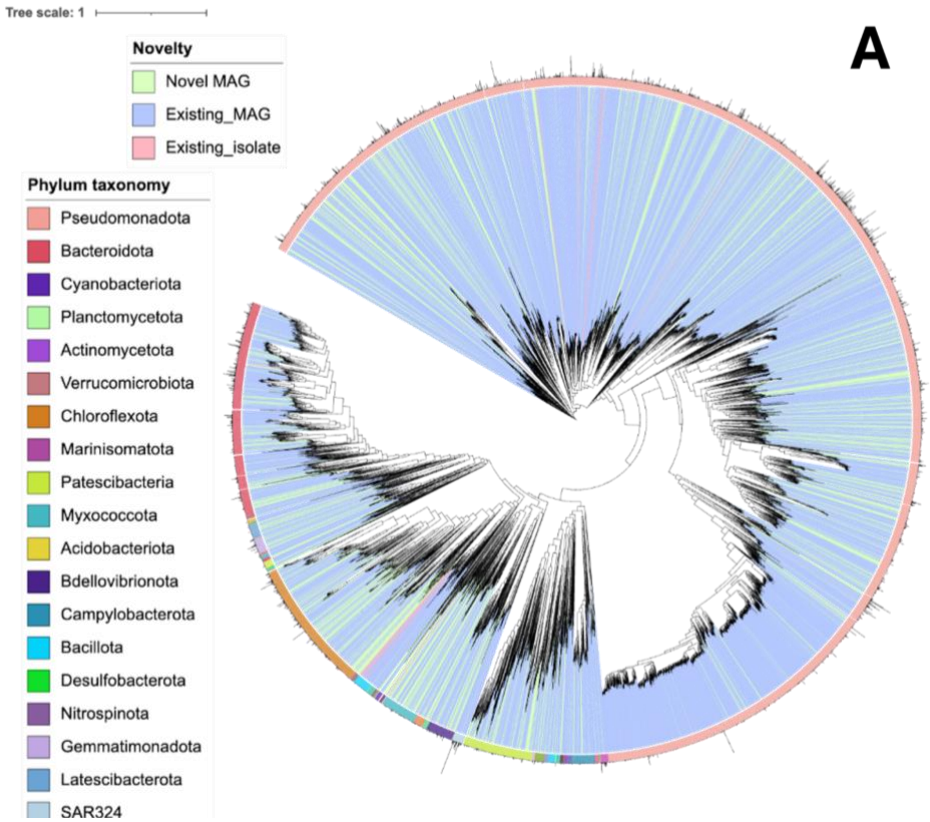


Figure 1: Geographical map of samples from which the genomes in Marine Catalogue v2 were derived. The resolution is such that multiple samples in close proximity may appear as a single data point. Those data points within land masses are as a result of errors in the metadata associated with the submitted sequence data.



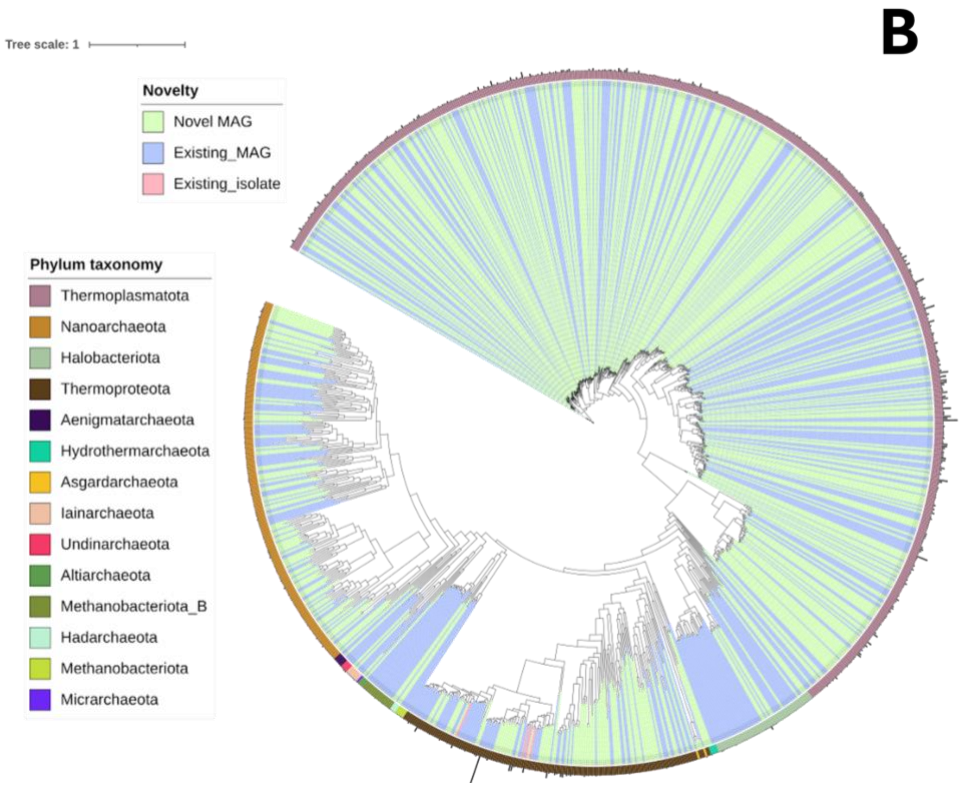


Figure 2: Phylogenetic trees of (A) bacterial and (B) archaeal species representatives. Branch colours represent novel MAGs, existing MAGs and existing isolates. Phylum taxonomy is indicated by the coloured ring around the outside. Bar plots on the edge correspond to the number of genomes in each cluster.

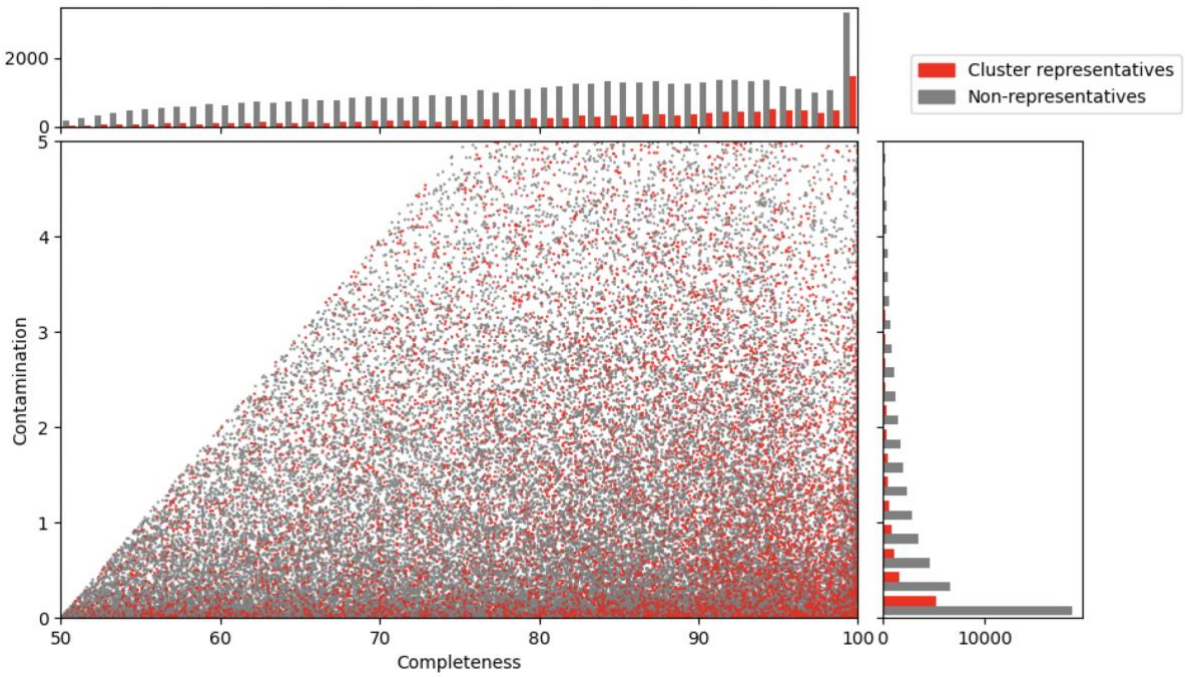


Figure 3: Scatter plot showing the completeness and contamination scores for all genomes in the catalogue. Cluster reps are shown in red; all other genomes are shown in grey. Histograms for score

ranges are shown at the top and right to help interpret the areas with high density of data points on the scatter plot.

Catalogue statistics

Genomes in catalogue	50,866
Isolate genomes in catalogue	232
MAGs in catalogue	50,634
Species-level clusters	13,223
Bacterial species-level clusters	12,133
Archeal species-level clusters	1,087
Pangenome clusters	7,095
Bacterial clusters novel wrt GTDB	5,260
Archaeal clusters novel wrt GTDB	536
Non-redundant sequences in protein catalogue	52,607,289
Protein sequence clusters at 90% amino acid identity	25,747,277

Multi-kingdom data

While we have generated (and continue to generate) eukaryotic MAGs at scale (albeit generating significantly smaller numbers of MAGs compared to prokaryotic MAGs), and those MAGs have been submitted to the ENA MAG-layer, we are yet to release the genome annotations of those MAGs. Existing approaches for gene calling on microbial eukaryotes are limited, and vary in quality. Currently, there is the choice of using *ab initio* approaches or approaches that use other data to identify genes. However, *ab initio* approaches are particularly error prone, especially in the relatively novel genomic space that these MAGs relate to. Conversely, other approaches to annotation of eukaryotic genomes rely on associated transcriptomic data, which does not exist for metagenome-derived genomes. However, there are a number of approaches we are actively pursuing with a view to improving gene-prediction in eukaryotic MAGs and releasing a marine eukaryotic catalogue in the near future. For example, transcriptomic resources (such as https://metdb.sb-roscoff.fr/metdb/) may provide a source of species-matched transcriptomic data that could be employed in the gene-calling pipelines. In preparation for this eukaryotic MAG dataset we are also prioritising MAG-generation from studies likely to yield a higher proportion of eukaryotic MAGs, for example those samples size-fractionated to enrich for protists.

Catalogue availability

This latest version of the MGnify Genomes marine catalogue (v2.0) is available to access and query both through the MGnify website as well as via the MGnify API. The catalogue can be browsed by list of

genomes and taxonomic tree, with the ability to filter on many metadata fields, allowing the user to access individual cluster representative genome records. Within individual genome records there are comprehensive genome statistics, summaries of annotations, and an interactive genome browser allowing interrogation of the various annotation tracks and their genomic context. All results files can be downloaded via FTP from the catalogue directory. In addition to browse-based access, there are two sequence-based search approaches that can be carried out via the website or API. The first is a COBS (Compact Bit-Sliced Signature Index)-based query for searching gene sequences against the catalogue. The second is a kmer-based search using Sourmash to allow querying of whole genomes or sets of genomes against the catalogue. This permits the catalogue to be used as a reference to determine novelty of the query genome(s) with respect to the catalogue.

Future updates to the marine catalogue

While this version 2.0 of the marine catalogue represents a significant increase in data with respect to version 1.0, the ongoing increase in available marine data means that future updates will be required. There are a number of MAG datasets generated by BlueRemediomics partners that are either already submitted but currently private (such as those produced by ETHZ) or soon to be submitted such as those from CNRS and SZN. These (along with others) will be scheduled for inclusion in the future version 3 of the marine catalogue, in a bid to ensure the catalogue remains representative of the marine data available.