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	D1.1 – Additional metagenomic datasets added to database
Work Package	WP1
Lead	EMBL
Deliverable Due Date	29/02/2024
Submission date	29/02/2024
Author(s)	Robert Finn
Dissemination Level	Public
Version	1.0

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

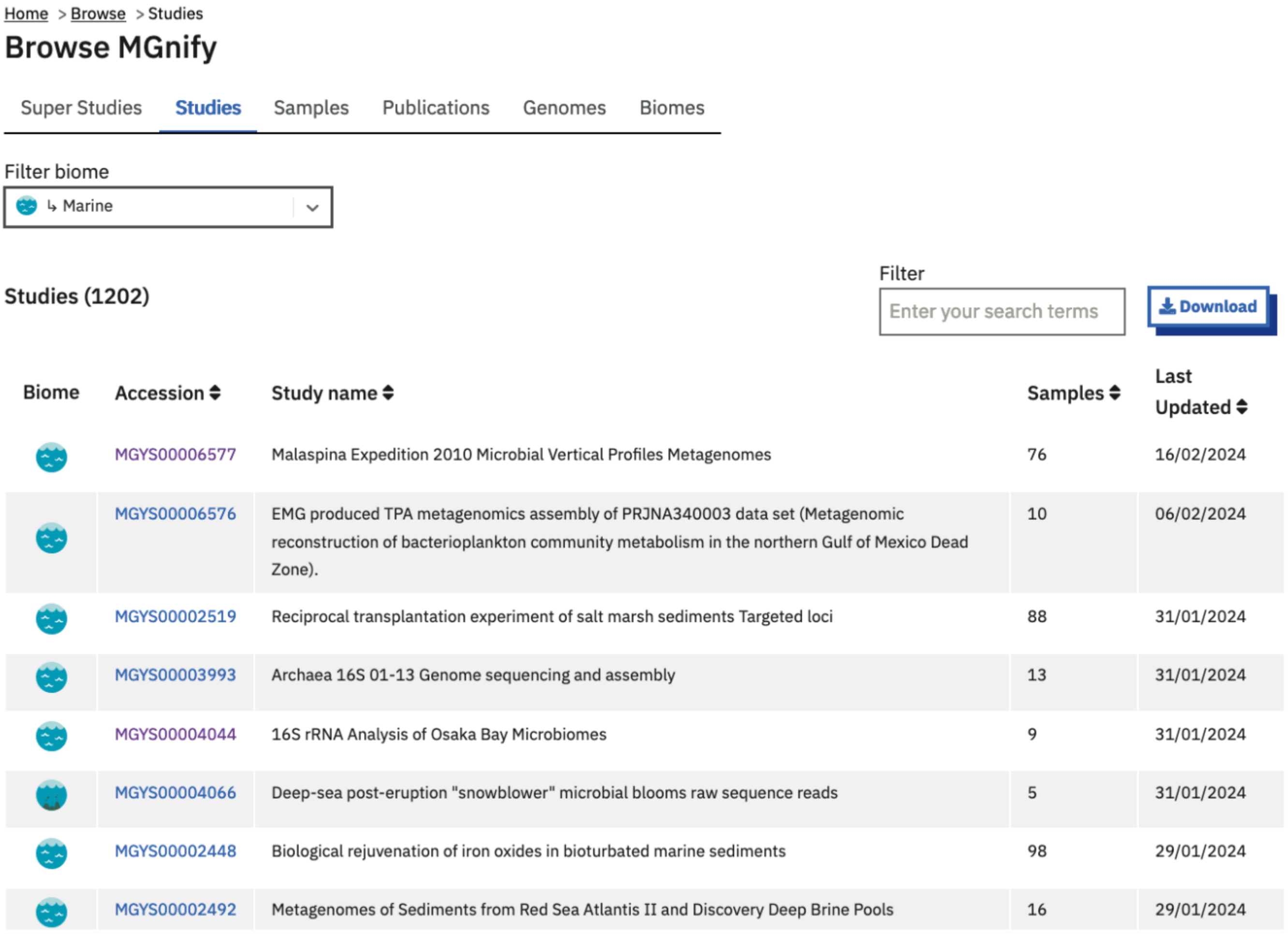
Swiss Confederation

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.

MGnify is an openly accessible, comprehensive, and worldwide hub of microbiome-derived sequence data (metagenomics, metatranscriptomics, metabarcoding). This data resource is central to the activities undertaken within the BlueRemediomics project and represents the fulcrum of the BlueRemediomics discovery platform. Beneficiary (and project coordinator) EMBL who administers the MGnify resource has focused efforts on three major different metagenomics data categories associated with the marine and aquaculture environments, namely public, private, and shotgun datasets that will be generated in the near future. Currently, all the processed datasets are short-read sequences with the vast majority produced by Illumina sequencing platforms. The primary goal is to assemble these short-read datasets into contigs, which can then be analysed for open reading frames using Prodigal and the resulting predicted proteins functionally annotated using a range of tools. As these contigs can span large genomic regions, the assembly into contigs also facilitates the identification of higher-order functional regions, such as metabolic pathways and biosynthetic gene clusters. Both public and private data types once integrated into MGnify can be shared with the BlueRemediomics consortium partners for experimental characterisation.

Public datasets - 5,860 marine metagenomics datasets that have been assembled to date, with all of these also submitted to the European Nucleotide Archive (ENA). This ensures that the assembly is appropriately associated with the raw-reads used to generate them, and the contextual metadata data is deposited with the sample. Once indexed by ENA, these assemblies are analysed using the MGnify assembly analysis pipeline (version 5.0), with all results publicly available via the MGnify website (https://www.ebi.ac.uk/metagenomics) and associated Application Programmatic Interface (API). A large fraction of these assemblies has come from the AtlantECO project (4,419 assemblies; EC-H2020 Grant Agreement No. 862923), while some of the datasets were analysed prior to the start of the AtlantECO project (1,295). As part of BlueRemediomics, we have already assembled, analysed, and uploaded 146 datasets derived from the marine environment. This represents an ongoing and continuous activity and thus, multiple other marine datasets are currently in the process of being analysed with a further 73 assembled and analysed datasets waiting to be integrated into the MGnify data stores prior to public release in MGnify. A further 202 shotgun marine metagenomics datasets are currently being assembled. All the metagenomic data from the HoloFood project (EC-H2020 Grant Agreement No. 817729) is also available via MGnify along with the assemblies from the salmon aquaculture. One of the barriers to having even greater numbers of assemblies is that many of the public datasets in ENA are mislabelled as shotgun metagenomics, when they are barcoding datasets in reality. We estimate that there may be some additional 13,000 marine datasets that can potentially be incorporated into MGnify (see below). The MGnify team will continue to identify potential datasets from relevant marine and aquaculture environments over the course of the BlueRemediomics project, making them available to both the project partners and the wider research community.



Home > Browse > Studies

Browse MGnify

Filter biome

↳ Marine

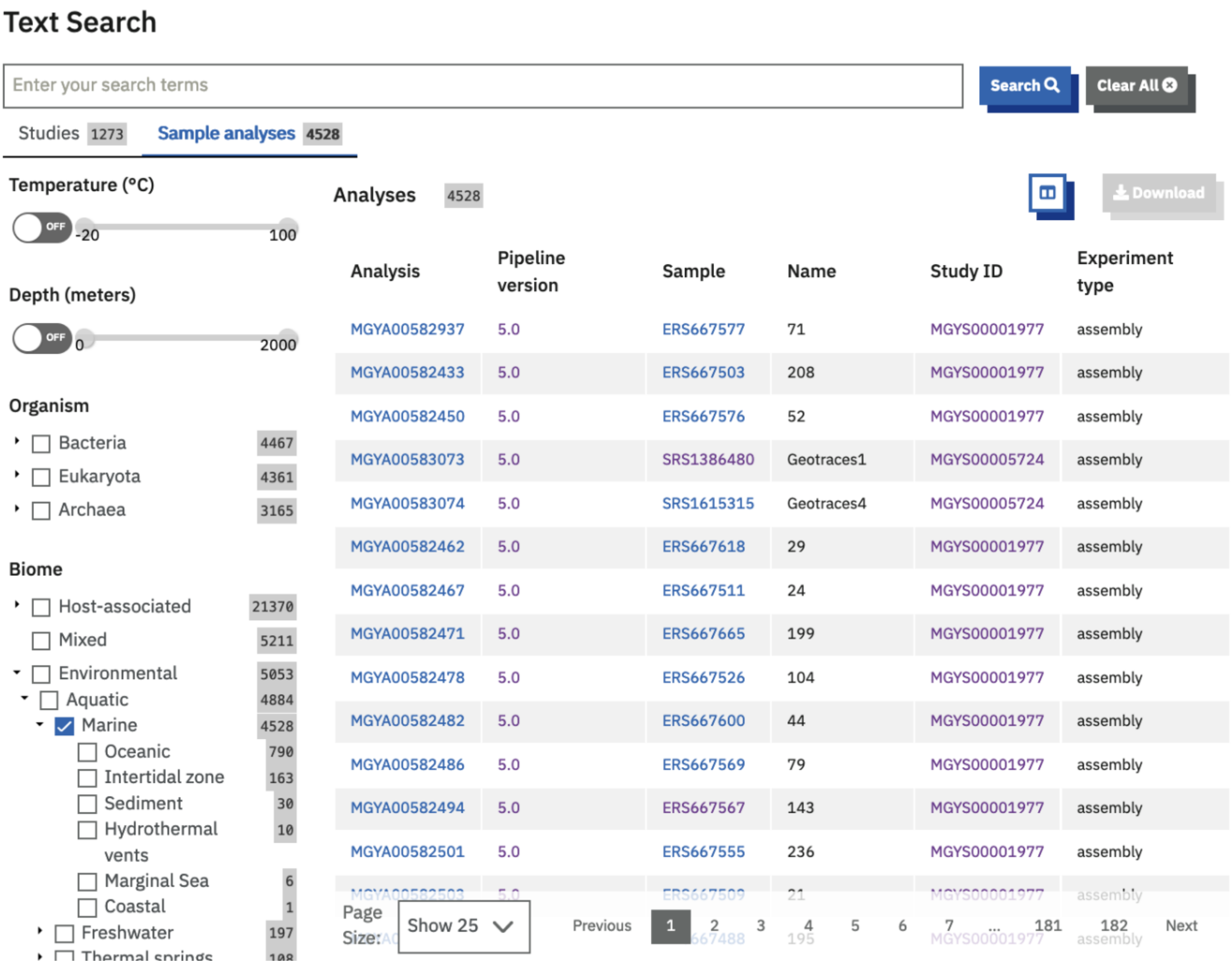
Studies (1202)

Filter

Enter your search terms

Biome	Accession ⬍	Study name ⬍	Samples ⬍	Last Updated ⬍
	MGYS00006577	Malaspina Expedition 2010 Microbial Vertical Profiles Metagenomes	76	16/02/2024
	MGYS00006576	EMG produced TPA metagenomics assembly of PRJNA340003 data set (Metagenomic reconstruction of bacterioplankton community metabolism in the northern Gulf of Mexico Dead Zone).	10	06/02/2024
	MGYS00002519	Reciprocal transplantation experiment of salt marsh sediments Targeted loci	88	31/01/2024
	MGYS00003993	Archaea 16S 01-13 Genome sequencing and assembly	13	31/01/2024
	MGYS00004044	16S rRNA Analysis of Osaka Bay Microbiomes	9	31/01/2024
	MGYS00004066	Deep-sea post-eruption "snowblower" microbial blooms raw sequence reads	5	31/01/2024
	MGYS00002448	Biological rejuvenation of iron oxides in bioturbated marine sediments	98	29/01/2024
	MGYS00002492	Metagenomes of Sediments from Red Sea Atlantis II and Discovery Deep Brine Pools	16	29/01/2024

Figure 1 – List view of the Marine studies in MGnify. The second row in the table represents an assembly produced by MGnify.



Studies 1273 | Sample analyses 4528

Temperature (°C) OFF -20 100

Depth (meters) OFF 0 2000

Organism

- □ Bacteria 4467
- □ Eukaryota 4361
- □ Archaea 3165

Biome

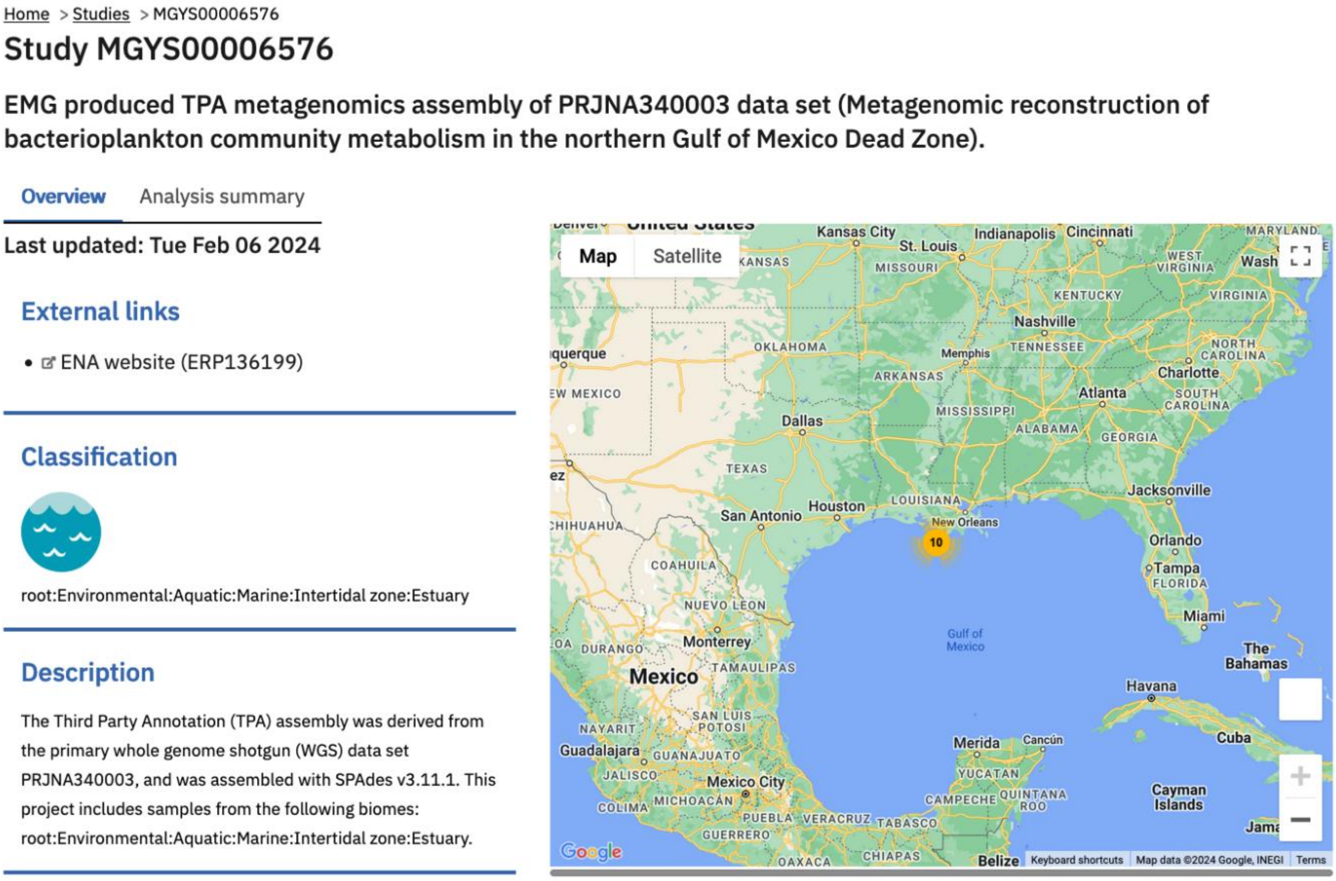
- □ Host-associated 21370
- □ Mixed 5211
- □ Environmental 5053
  - □ Aquatic 4884
    - ☑ Marine 4528
      - □ Oceanic 790
      - □ Intertidal zone 163
      - □ Sediment 30
      - □ Hydrothermal vents 10
      - □ Marginal Sea 6
      - □ Coastal 1
    - □ Freshwater 197
    - □ Thermal springs 108

Analyses 4528

Analysis	Pipeline version	Sample	Name	Study ID	Experiment type
MGYA00582937	5.0	ERS667577	71	MGYS00001977	assembly
MGYA00582433	5.0	ERS667503	208	MGYS00001977	assembly
MGYA00582450	5.0	ERS667576	52	MGYS00001977	assembly
MGYA00583073	5.0	SRS1386480	Geotraces1	MGYS00005724	assembly
MGYA00583074	5.0	SRS1615315	Geotraces4	MGYS00005724	assembly
MGYA00582462	5.0	ERS667618	29	MGYS00001977	assembly
MGYA00582467	5.0	ERS667511	24	MGYS00001977	assembly
MGYA00582471	5.0	ERS667665	199	MGYS00001977	assembly
MGYA00582478	5.0	ERS667526	104	MGYS00001977	assembly
MGYA00582482	5.0	ERS667600	44	MGYS00001977	assembly
MGYA00582486	5.0	ERS667569	79	MGYS00001977	assembly
MGYA00582494	5.0	ERS667567	143	MGYS00001977	assembly
MGYA00582501	5.0	ERS667555	236	MGYS00001977	assembly
MGYA00582503	5.0	ERS667509	21	MGYS00001977	assembly

Page Size: Show 25 | Previous 1 2 3 4 5 6 7 ... 181 182 Next

Figure 2 – List of marine assemblies in MGnify that have been analysed with the latest version of the pipeline.



Home > Studies > MGYS00006576

Study MGYS00006576

EMG produced TPA metagenomics assembly of PRJNA340003 data set (Metagenomic reconstruction of bacterioplankton community metabolism in the northern Gulf of Mexico Dead Zone).

Overview | Analysis summary

Last updated: Tue Feb 06 2024

External links

- ENA website (ERP136199)

Classification

root:Environmental:Aquatic:Marine:Intertidal zone:Estuary

Description

The Third Party Annotation (TPA) assembly was derived from the primary whole genome shotgun (WGS) data set PRJNA340003, and was assembled with SPAdes v3.11.1. This project includes samples from the following biomes: root:Environmental:Aquatic:Marine:Intertidal zone:Estuary.

Figure 3 – An example of a study that has been assembled and analysed as part of BlueRemediomics.

Private datasets – These are processed in the same fashion as described for public datasets, with the exception that they are held in a private (also termed pre-publication) state by the data generator. This enables the data generator to take advantage of MGnify using identical workflows as those that would be developed for public data yet allowing them to mine the data for new bioactives and valorise them prior to making them public. To this end, the MGnify team from Beneficiary/Coordinator EMBL continues to work closely with Associated Partner ETHZ to upload their vast marine metagenomics assemblies for subsequent exploitation within the BlueRemediomics project. This has leveraged the EMBL's data resource submission tools and MGnify team members' expertise to facilitate the upload of 12,000 assemblies into ENA. We will compare this set of assemblies with those already available in MGnify to determine the datasets currently not represented in in the database and prioritise them for inclusion.

Future datasets – As outlined in the proposal, we proposed to investigate how the BlueRemediomics discovery platform could be utilised by the other project funded within the same HORIZON-CL6-2022-CIRCBIO-01 call. To this end, we had initial discussions on the metagenomics data being generated by the BlueTools project (Horizon Europe Grant Agreement No. 101081957). Based on the outcomes of this discussion, we are actively working on a proof-of-principle dataset, namely the assemblies associated with PRJEB39821 (a biogas reactor dataset produced by a BlueTools Beneficiary as part of another research project). This will provide a template for how the data will flow from BlueTools into the public domain by leveraging the power of the BlueRemediomics discovery platform. As the datasets from BlueTools have already been assembled, the computational burden on MGnify is far lower since only the annotation pipeline needs to be executed and the results uploaded. Once this data flow is established, BlueTools plans to provide data as and when they become available.

Both BlueRemediomics and BlueTools will have their collections appropriately tagged to ensure accurate attribution to the respective projects and that project partners can also easily discover datasets associated with their projects.