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	**D7.2 – Data Management Plan - initial**
Work Package	**WP7 – Management and Coordination**
Lead	**EMBL**
Deliverable Due Date	**28 February 2023**
Submission date	**31 May 2024**
Resubmission date(s)	**21 November 2024, 10 December 2024**
Author(s)	**Robert Finn, Shriya Raj**
Dissemination Level	**Public**
Type	**DMP**
Version	**4.0**

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.

HISTORY OF CHANGES			
Version	**Publication date**	**Changes**	**Pages**
1.0	28/02/2023	▪ Initial version	N/A
2.0	31/05/2024	▪ Updated to make it more readable, specific to the project and to address the gaps from the initial version generated using the Data Wizard. Our data models are also more developed than is performed by the data wizard.	All pages have been impacted by this update.
3.0	21/11/2024	▪ Expansion of new data sets (2.1.3), inclusion of table 2 to cover all data types.	7-8 and 9-13
4.0	05/12/2024	▪ Addition of new section “Specific dataset outputs from BlueRemediomics”.	18-20

CONTENTS

1. Introduction

A key aim of the BlueRemediomics project is to establish a Discovery Platform (WP1 and 2), which connects marine microbiome datasets (WP1) to high throughput screening and assay platforms, and to the culture collections (WP2), particularly those housed by EMBRC (Beneficiaries SU, EMBRC-ERIC). The goal is to not only harmonise the data but to also link the experimental characterisation results back to the data, so as to enrich the annotations and expedite translation of data to biological based solutions for the blue circular economy (**Figure 1**).

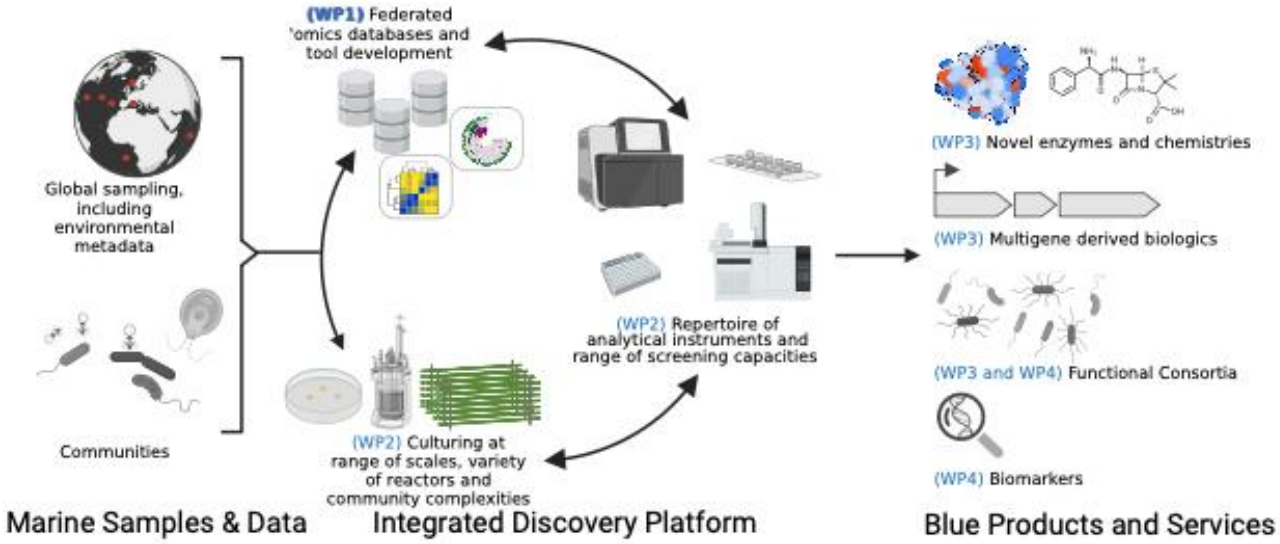

Figure 1 - BlueRemediomics Discovery Platform depicting the data flows and iterative refinement of in silico predictions and experimental validation.

As part of this effort, a major objective is to reuse microbiome sequence data that has already been generated by either project partners and/or is found in the public domain. Where possible this data will be collected, processed and analysed according to the following data management plan. Thus, there will be a significant drive to connect derived data products back to their original samples, and to ensure data provenance.

2. Data Summary

2.1. Data generation, discovery and reuse

In BlueRemediomics, the major objective is to reuse marine microbiome data that has already been generated either by the project partners and/or is found in the public domain. This data will primarily be discovered via two different databases, one an archival database for nucleotide sequences, namely the European Nucleotide Archive, ENA and the other, a knowledgebase called MGnify, which contains biological, functional, and taxonomic annotations of the microbiome derived sequence data and will form the central resource through which the data will be coalesced.

2.1.1. Sequence Data

Sequence data provide raw reads from metagenomics and metatranscriptomics experiments, as well as associated assemblies and metagenome assembled genomes, i.e. MAGs (where available and appropriate). For this project, we will predominately utilise existing marine metagenomics datasets, which are either yet to be submitted to a sequence archive, but available from the project partners (see below), or are available via public repositories, such as ENA (https://www.ebi.ac.uk/ena). To discover appropriate datasets we will utilise the ENA API to select for relevant biomes (e.g. marine, marine sediment, estuary), and ensure that these have been processed to maximise the data. For this project, there is the need to have assemblies of metagenomes and metatranscriptomes as these enable the identification of full length genes, as well as higher order structures such as operons and biosynthetic gene clusters (BGCs). Metagenomic assemblies can be further processed to produce MAGs. At the project outset, the vast majority of sequence datasets were short-read sequences (>99.9%) and a greater portion of datasets (>95%) lacked an associated assembly. While the owner of a dataset can easily attach an assembly to the Study in ENA (i.e. the BioProject), this is not the case for third-party data processors. Thus, EMBL has developed a data model to ensure all processed data can be associated with the original dataset as a third-party annotation (**Figure 2**). Raw data will be retrieved in FASTQ formats. From here, the short-read sequence data will be assembled using algorithms like metaSPAdes to produce contigs which are produced in FASTA format. The assemblies are then submitted back to ENA and associated with the original sequence data (**Figure 2, bottom right**). Each assembly file is referred to as an analysis object, as it is produced from analysing the whole sequence file. Associated with the analysis object are references to the sample(s) and sequence(s) used to generate the assembly. This model will be adopted for both metagenomic and metatranscriptomic data, which enables the association of sample metadata to the newly generated data, as well as the read coverage data. This is achieved by generating a Binary Alignment file (BAM), which gives the location of every read that is mapped to the assembly. These are large files that are not stored long term but can be easily recreated.

For BlueRemediomics beneficiaries who submit raw datasets they have generated (and own), the raw sequence data will be submitted with the same organisation as the public data, except that the analysis objects can be stored in the same study. If they are using public data, the same model will be used.

Metagenomic assemblies can then be further processed to produce binned sets of contigs that are representative of genomes, also referred to as MAGs. As before, derived sequence products will be submitted to ENA in FASTA formats, but as these employ a subset of the sequence information, they do not follow the same data model. Instead, a “derived sample” is produced that refers to the original sample, the MAG analysis object is then associated with that derived sample (**Figure 2, bottom left**). MAGs will be associated with metadata that conforms to the MiMAGs checklist and includes the quality scores and taxonomy of each MAG or bin. All of these third-party methods will be made publicly available immediately upon submission to ENA. The new metagenomics/metatranscriptomics assembly and MAG submission layers in ENA will be utilised. Metadata associated with these datasets will also be collected via ENA in XML or CSV formats. These layers will also be used for MAGs produced by the wider scientific community so that they can be included in the MAG catalogues.

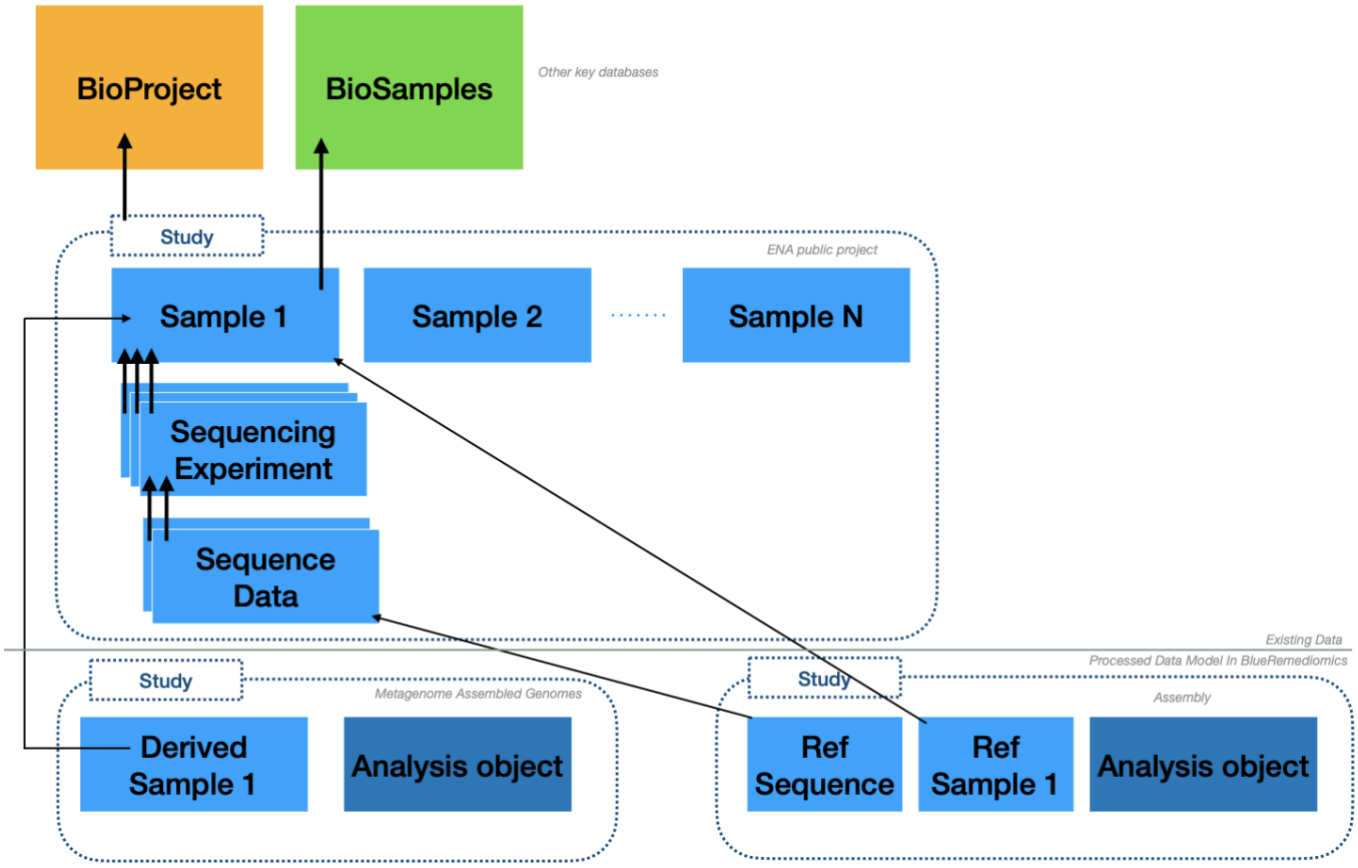

Figure 2 - Organisation of the sequence data produced by the BlueRemediomics project and their relationship to the original public raw data. The arrows indicate references to accessions within ENA (blue boxes) and cross references to other databases (green and orange).

2.1.2. Functional and taxonomic annotations

MGnify (https://www.ebi.ac.uk/metagenomics/) already houses significant numbers and volumes of microbiome related sequence data and associated annotations (summarised in **Table 1**). We will use the dataset as follows: mining for novel enzymes and secondary metabolite gene clusters. There is no need to harmonise different sources of existing data in our case, as that is already achieved through the application of standardised analysis pipelines.

Data/data product type	MGnify total	Marine-specific
Metabarcoding analysis	396,307	71,750
Metagenomic analysis	37,516	4,335

Metatranscriptomic analysis	2,242	768
Assembled metagenome analysis	35,486	4,237
Metagenomic Assembled Genomes (MAGs)	315,252	1,504
Assembled metagenomes	33,885	3,889
Predicted protein sequence (millions)	2,573	777

Table 1. Datasets already available in MGnify at the start of the project that will be utilised in the project (February 2023).

Additional datasets that are deposited by BlueRemediomics beneficiaries into ENA will then be analysed by MGnify. It is expected that MGnify will produce a few thousand additional assemblies during the project, which will also be subsequently analysed. The proteins from these analyses (estimated to be 10s of millions of sequences) will also be added to the protein database. One of the main objectives will be to establish marine MAG catalogues, which will produce representative sets of MAGs for obtaining new insights into the potential of marine microbes. Here, it is anticipated that 10,000s of MAGs will be analysed, with the resulting representative species genomes made available (~10,000 annotations).

Briefly, the following annotations are routinely produced by MGnify. The majority of these annotations are associated with recognised annotation resources, which have persistent identifiers (registered in identifiers.org) that facilitate data interoperability and reuse.

Taxonomic annotations: NCBI and GTDB identifiers.

Functional Annotations: InterPro, Pfam, KEGG, EggNOG, COG, CAZy, CRISP-Cas, AMR profiles from AMRFinder, AntiSMASH, SanntiS, GECCO.

Other annotations: Viral sequences

During the course of the project, we propose to add ChEBI identifiers to link the sequence data in MGnify to chemical reactions.

2.1.3. New data sets

While the main emphasis in BlueRemediomics is to reuse existing data, the project will also generate new data and metadata during its lifetime. This may include primary data from environmental sampling, and sequence data from new or existing isolated organisms (or samples). In addition to this raw data, there is the associated experimental and sample metadata describing the actions undertaken. Complementing this will be results from experimental assays and characterisations of isolate genomes or biogenics from work packages (WPs) 2-4. WP1 currently, and will continue to focus, extensively on generating value from existing marine metagenomics sequencing projects using derived sequence data, adopting a data layout

as indicated in **Figure 2**,. New data products, for example the pangenome outputs, will be provided in a specific resource provided by Beneficiary CEA. The construction of this data is documented and entirely reproducible. Nevertheless, we anticipate that a relatively small amount of nucleotide sequencing will be performed across WP2-4. Samples will be obtained through a number of different sequencing techniques, such as metabarcoding and whole (meta-)genome sequencing. The raw and quality-controlled files will be in the standardised fasta/fastq formats. Such datasets will be collected by individual project partners working with their own equipment or using institutional core facilities. The sequencing platforms that will be used are well established, primarily being Illumina or Oxford Nanopore Technologies (ONT). These acquired datasets will follow strict quality processes to ensure data validity. These are also subject to informatics quality control, such as the removal of poor sequences or those originating from a human source. In addition to the sequence data in WP1, genome-scale metabolic models (GEMs), will also be produced. These mathematical representations will adhere to standard formats, such as standard-GEM (an XML format), and will be made available through the partner websites and/or BioModels. WP1 will also produce protein sequences that will be accessioned as part of the MGnify protein database. Currently, since structural models produced by AlphaFold2/3 do not have a recognised archive, these will be made available as part of a publication via Zenodo (or similar). The next most common dataset is spectral data produced from mass spectrometry (MS) instruments. In this case, there is a mixture of vendor formats, which capture the raw results and the experimental metadata. Depending on the vendor and the partner involved, these will be transformed into MZXML. Similarly, the precise target database depends on the assay, with some partners preferring generalist databases, such as MetaboLights or MASSIVE, while there are other specific databases such as Global Natural Product Social Molecular Networking (GNPS). Nuclear magnetic resonance (NMR) data will be submitted to NP-MRD (the Natural Product Magnetic Resonance Database), a specific database for collecting NMR data associated with natural products. Currently, no proteomics is planned in BlueRemediomics, but in the instance such data is generated, would be submitted to PRIDE. As outlined in **Table 2**, other types of data, particularly the enzyme, bioactivity assays and antimicrobial minimal inhibitory concentration (MIC) assays do not have a natural database for archiving the results. Such data will be typically included in supplementary materials and be in the format of CSV, text files and data will also be made available via Zenodo.

In WP4, specific samples will be taken as part of the aquaculture experiments. Wherever possible, these will be collected according to checklists and utilise controlled vocabularies and ontologies. Where metadata does not fit an existing checklist, it will be submitted to the associated BioSample record as structure data, for example the histology scores and hormone concentrations. In all cases, experimental metadata will be collected and included as part of the experimental section of the checklist or captured in notebooks. While the former will be submitted to archival databases, the latter will be collected and made available via publications.

We anticipate a small number of experiments (~100) to generate multi-omics data, which will be a mixture of metagenomics, isolate genome sequencing and metabolomics all originating from the same sampling event. To retain the relationship between these very different data types, BioSamples will be used. This will utilise the latest best practices being developed as part of another Horizon Europe funded project MICRObiome Biobanking (RI) Enabler (GA Number 101094353), where a parental BioSamples accession number will enable the connection of derived sub-samples. When different omics techniques are applied to the same sub-sample, these will utilise the same BioSamples accession. In BlueRemediomics, we will ensure that this model is adopted and the BioSamples accessions are maintained in the data deposition records.

2.2. Data types, formats and sizes

The data types and associated formats, together with metadata are presented in **Table 2**. There is a large diversity of data types, as well as instruments used to collect these types and formats. In the majority of cases, the data types are recognised standards, which facilitates interoperability and downstream processing/archiving. The metadata is also captured in a heterogeneous manner. Some metadata is captured as part of the original sequence submission, where there is a derived data product (see section 3 below), while experiments capture the metadata as part of the output format. Others are captured in laboratory notebooks and/or in digital formats, such as text files. We estimate that the total data produced as part of this project will be 1-200 terabytes of data.

Work Package	Partner	Data Type	Data Source	Data Format	Metadata	Data Volume	Data Policy	Long-term Archive	Back-up Policy
WP1	CEA	Pangenome database	Software output	HDF5	No metadata, attached to original sequence input	1000GB	Publicly available	n/a	Daily
		Pangenome database	Software output	Text file	No metadata, attached to original sequence input	1GB	Publicly available	n/a	Daily
	UCL	Protein sequences	Software output	FASTA	Connected to workflow/notebooks	1GB	Publicly available	Zenodo	Daily
		Structural analysis data	Software output	PDB/ TXT	Connected to workflow/notebooks	10GB	Publicly available	Zenodo	Daily
	LBMC	Metagenome assembled genomes	Software output	FASTA	Connected to raw sequence record	1GB	Publicly available and submitted to an archive	ENA	Infrequent
	EMBL	Metatranscriptome assemblies	Software output	FASTA	Connected to raw sequence record	10GB	Publicly available and submitted to an archive	ENA/ MGnify	Daily
		Metatranscriptome annotations	Software output	TSV	Connected to raw sequence record	1000GB	Publicly available and submitted to an archive	ENA	Daily
		Metabarcoding	Software outputs	FASTA, BIOM, TSV, Darwin-Core	Metadata connected to sequence record	10GB	Publicly available and submitted to archive	MGnify	Daily

		Metagenome assembled genomes	Software output	FASTA, GFF, TSV	Connected to raw sequence record	100GB	Publicly available and submitted to an archive	ENA/ MGnify	Daily
		Metagenomic assemblies	Software output	FASTA	Connected to raw sequence record	1000GB	Publicly available and submitted to an archive	ENA/ MGnify	Daily
		Genome annotations	Software output	GFF	Connected to raw sequence record	100GB	Publicly available and submitted to an archive	MGnify	Daily
		Metabolic models	Software output	XML	No metadata	5000GB	Publicly available and submitted to an archive	BioModels	Daily
		Protein sequences	Software output	FASTA	Connected to assembly, then to the original sample	100GB	Publicly available and submitted to an archive	MGnify	Daily
	ETHZ	Metagenomic assemblies	Software output	FASTA	Connected to raw sequence record	1000GB	Publicly available and submitted to an archive	ENA	Daily
		Metagenome assembled genomes	Software output	FASTA	Connected to raw sequence record	100GB	Publicly available and submitted to an archive	ENA	Daily
WP2	SU	DNA Barcoding	Illumina sequencing platform	FASTQ	Collected according to MIxS Standard	1GB	Publicly available and submitted to an archive	ENA/ BioSamples	Infrequent
		Genome Sequencing	ONT	FAST5	Collected according to MIxS Standard	100GB	Publicly available and submitted to an archive	ENA/ BioSamples	Infrequent

	IFREMER	Metabarcoding	ONT	FAST5	Collected according to MIxS Standard	100GB	Publicly available and submitted to an archive	ENA/ BioSamples	Weekly
	UWC	Genome Sequencing	ONT	FAST5	Collected according to MIxS Standard	10GB	Private, but available to the consortium, longer-term publicly available	The genomes (and microbial strains) are held in the Biodiversity Biobank of South Africa (Microbial node) hosted in IMBM at UWC. Any published genomes from our own research from this collection are then accessible from public domains.	Infrequent
		Bioactivity	Bespoke laboratory measurements	TXT/XLS	Connected to sample	1GB	Publicly available and submitted to an archive	Figshare/ Zenodo	Infreque
	IOCB	Experimental lab data (SDS-PAGE and western blot images, results from activity measurements, graphs)	Various	Standard format: images - jpeg, png, tiff; activity data - xls; graphs - graph pad, excel.		1GB	Part of publication	Figshare/ Zenodo	Infrequent
	UCAM	Screening assays	Bespoke laboratory measurements from microfluidics	TXT/XLS	Connected to screening assay setup. No specific standards	10GB	Part of publication	Zenodo	Infrequent
	NORCE	Cultivation data	Bespoke laboratory measurements	XLS/DOC	Connected to cultivation event	<1GB	Private, but available to the consortium, longer-term publicly available	Zenodo	Weekly

WP3	UNIABDN	Mass-spectrometry spectra		Vendor format	Collected with raw data	2000GB	Publicly available and submitted to an archive	GNPS/ MASSIVE	Weekly
		Mass-spectrometry spectra		MZXML	Collected in MZXML header	2000GB	Publicly available and submitted to an archive	GNPS/ MASSIVE	Weekly
		NMR		Vendor format	Collected with raw data	2000GB	Publicly available and submitted to an archive	NP-MRD	Weekly
	SZN	Mass-spectrometry spectra		Vendor formats, .raw and .mgf	Collected in vendors format	500GB	Publicly available and submitted to an archive	GNPS	Infrequent
		Genomic Sequencing	Illumina sequencing platform	FASTQ/FASTA	Collected according to MIxS Standard	10GB	Publicly available and submitted to an archive	ENA	Infrequent
		Genomic Sequencing	ONT	FAST5/FASTA	Collected according to MIxS Standard	10GB	Publicly available and submitted to an archive	ENA	Infrequent
	UWC	Mass-spectrometry spectra		Vendor format (.raw)	Collected with raw data	100GB	Publicly available and submitted to an archive		Daily
		HPLC		Text Files / XLS / PDF	Connected to sample	1GB	Part of publication	Zenodo	Daily
		Chemical structure	Derived from Mass-spec	Text	Connected to sample	1GB	Publicly available and submitted to an archive	ChEMBL/ ChEBI	Daily
	IOCB	Peptide sequences	From WP1 and derived from software	FASTA	Connected to workflow/notebooks	1GB	Publicly available	Zenodo	Infrequent

	EMBL	Mass-spectrometry spectra		MZXML	Collected with raw data	10000GB	Publicly available and submitted to an archive	MetaboLights	Daily
	IFREMER	Genomic Sequencing	Illumina/ONT	FASTQ/FASTA	Collected according to MIxS Standard	10GB	Publicly available and submitted to an archive	ENA/ BioSamples	Infrequent
WP4	NORCE	Gene expression analyses	Illumina	FASTA/FASTQ	Connect to sample record in BioSamples, experimental metadata in ENA	10GB	Publicly available and submitted to an archive	ENA/ BioSamples	Infrequent
		histology scores		Text Files	Connect to sample, structured XML in BioSamples	<1GB	Publicly available and submitted to an archive	BioSamples	Infrequent
		hormone/protein concentration		Text Files	Connect to sample, structured XML in BioSamples	<1GB	Publicly available and submitted to an archive	BioSamples	Infrequent
		Metabolite concentration/Mass-spectrometry spectra		Vendor Formats	Collected with raw data, connected to sample	100GB	Publicly available and submitted to an archive	MetaboLights/ MASSIVE	Infrequent
	LBMC	Metagenomics Sequencing	MGI Sequencing platform	FASTQ	Collected according to MIxS Standard	10GB	Publicly available and submitted to an archive	ENA/ BioSamples	Infrequent
	CNRS	Genomic sequence profiling data (presence-absence, abundance)		SV, BIOM	Connected to samples	1000GB	Publicly available	Zenodo, EMODnet Biology, GeoNode	Weekly
		Species Distribution Models (SDM)		NetCDF	Connected to samples	100GB	Publicly available	Zenodo, EMODnet Biology, GeoNode	Weekly
		Co-occurrence graphs.		GRAPHML	Connected to samples	10GB	Publicly available	Zenodo, EMODnet Biology, GeoNode	Weekly

		Genome Sequencing*	Illumina	FASTQ/ GFF	Collected according to MIxS Standard	10GB	Publicly available and submitted to an archive	ENA/ BioSamples	Weekly
		Metagenomic Sequencing*	Illumina	FASTQ/ FASTA	Collected according to MIxS Standard	100GB	Publicly available and submitted to an archive	ENA/ BioSamples	Weekly
	HCMR/ EMBRC	Genome Sequencing*	Illumina	FASTQ/ FASTA	Collected according to MIxS Standard	10GB	Publicly available and submitted to an archive	ENA/ BioSamples	Infrequent

* **generated outside of BlueRemediomics, but part of the work package activities**

Table 2 - the breadth of data types being produced across the different WPs by BlueRemediomics partners. The table captures the source of information where possible, the format and associated metadata standards along with an indication of the expected data volumes. All of the data is eventually expected to be available publicly (see text for more details), with restrictions in virtually all cases based on publication and/or IP exploitation. The majority of data will be targeted to a long-term, recognised public archive, with Zenodo being used for data types that are out of scope for these resources. Finally, the table also captures the back-up frequency associated with the datatypes. In all cases the data is protected, but in the case of infrequent back-ups, it reflects data that is generated once and then does not change, so does not require more frequent back-ups. Where more frequent backs-ups are made, this often reflects local institutional infrastructure policies, rather than the data is constantly changing.

We do not anticipate any additional new data formats as part of the BlueRemediomics project. The data outlined above will be largely available from MGnify, ENA, BioSamples, MetaboLights, MASSIVE and PRIDE resources. All of these databases have rich application programming interfaces (API) that conform to the design principles of the representational state transfer (REST). In addition to the formats above, many of the API endpoints return JSON data. Additional data formats are available on the FTP server for MGnify Genomes, and these are described in each catalogue's README (e.g. http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/marine/v1.0/README.txt). These include Kraken formatted databases for each catalogue but are not the primary focus of this project. MGnify's protein database is distributed as FASTA and TSV files (http://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/current_release/).

2.3. Data Utility Beyond BlueRemediomics

Beyond BlueRemediomics, the data being produced is already being used by others. For example, the metabarcoding data in MGnify is being employed for biodiversity and ecosystem modelling studies. The MGnify MAGs will be used as part of two other ongoing Horizon Europe funded initiatives, namely BlueCloud-2026 (GA Number 101094227) and BIOcean5D (GA Number 101059915) as reference databases for species and trait modelling. Data from MGnify also flows into GBIF and OBIS, which help shape policy making. Sequence data in MGnify is being widely used to develop new algorithms (e.g. AlphaFold utilised the data), identifying new enzymes and understanding biodiversity. Data that is

submitted to ENA is replicated across the International Nucleotide Sequence Database Consortium (INSDC) sites. Most of the data is anticipated to be made available to the public upon acceptance of an associated publication. The data will be submitted with rich metadata concerning the sample and the experimental process, ensuring interoperability and reuse.

It is also expected that BlueTools (GA Number 101081957), the sister project funded under the same HORIZON-CL6-2022-CIRCBIO-01 initiative as BlueRemediomics, will leverage the MGnify workflows for processing and releasing their data.

3. FAIR data

3.1. Making data findable, including provisions for metadata

All data submitted to ENA, MGnify, MetaboLights and PRIDE will be assigned a persistent identifier by the corresponding resource. Within MGnify, the different data-types receive specific accessions, with MGnify metagenomic assemblies being assigned an accession in the format of MGYAXXXXXXXX, proteins as MGPYXXXXXXXX, and genomes are MGPGXXXXXXXX (where XXXXXXXX is a number). All metadata presented in MGnify relating the samples and experiments are inherited from the sequence deposition in ENA. Relevant metadata is indexed and both fetchable via the MGnify API, and can be used to filter query results within the webpage filtering and API data retrieval. Additional metadata are incorporated into the website via the Europe PMC metagenomics API supplied metadata from the literature contained in Europe PMC using a machine learning (ML) framework. Metadata pertaining to the analysis pipeline that was run (specific tools and versions etc) are all detailed and linked from the pipeline version description on MGnify (https://www.ebi.ac.uk/metagenomics/pipelines/5.0), as well as via the GitHub repository (https://github.com/EBI-Metagenomics/pipeline-v5). Furthermore, a detailed index across the analysis results facilitates users to perform a detailed faceted search, using accessions/labels, such as on functional and taxonomic terms, biome (e.g. marine), experiment type (e.g. metabarcode vs metagenomic) and analysis pipeline version. To aid the discovery of data specifically related to this project, a “Super Study” will provide a convenient entry point for the discovery of datasets produced and analysed as part of BlueRemediomics.

Research Object Crates (RO-Crates) are used to store, transfer, and display analysis results alongside their provenance metadata. MGnify’s RO-Crates follow schemas (e.g. https://www.researchobject.org/workflow-run-crate/) that allow results from workflows beyond MGnify’s standardised pipelines to be registered in and retrieved from MGnify. This standard is currently used to store and display BGC and mobilome annotations for assemblies alongside MGnify’s standard analyses.

3.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data. All of our data can become completely open immediately, but some may be retained in a private state to enable commercial exploitation as set out in the BlueRemediomics consortium agreement. Data that is not legally restrained will be released after a fixed time period (typically after two years or upon publication.), unconditionally. Metadata will be openly available, including instructions on how to get access to the data. Metadata will be available in a form that can be harvested and indexed (managed by the used repository / repositories). We have a Consortium Agreement that handles Intellectual Property. For the reference and non-reference data sets that we reuse, conditions are as follows:

- **European Nucleotide Archive** – freely available for any use (public domain).

- **MGnify** – freely available for any use (public domain or CC0).
- **MetaboLights** – freely available for any use (public domain).
- **PRIDE** – freely available for any use (public domain).

These databases are accessed via an API which largely follows the JSON:API specification (https://jsonapi.org). The API is public and can be queried directly by users, and browsed, for example https://www.ebi.ac.uk/metagenomics/api. The ENA and MGnify websites are clients for their respective API and allow non-programmatic access via search and browse interfaces (https://www.ebi.ac.uk/metagenomics).

There are several search functions for the MGnify database:

- EBI Search (a search service across EMBL-EBI resources) provides a text and facet search across MGnify samples, studies, and analyses, e.g. for finding all studies from marine biomes with a certain phrase in their title.
- API endpoint filters provide custom filtering options for lists of data, e.g. for finding samples with sampling depth metadata within a certain range.
- Genome search (based on Sourmash) provides a sketch-based comparison of query genomes against MGnify's genome catalogues.
- Genome fragment search (based on COBS) provides a kmer search of query genes/sequences against MGnify's genome catalogues.
- MGnify Sequence Search provides a HMMER-based sequence search on MGnify's protein database.

MGnify's Notebooks (https://docs.mgnify.org/src/notebooks_list.html) are an additional client, which provide examples and templates for downstream analysis beyond the features available on the website. For example, the notebooks include interactive examples for cross-study comparative metagenomics.

Some MGnify datasets are also (or only) available via the EMBL-EBI FTP server (http://ftp.ebi.ac.uk/pub/databases/metagenomics). This provides access to flat-file releases of the protein database, and the full MGnify genome catalogues, including cluster members which are not available via the MGnify API/website.

3.3. Making data interoperable

To make data and metadata interoperable, community standards, formats and methodologies will be used. MIxS standards will be used for all raw sequence data, and MiMAG standards for MAGs. As outlined in Section 2.2, a variety of well established data formats will be used. In addition to these formats and the use of established annotation databases that are highly interoperable, control vocabularies will be used, particularly ENVO (the standard ontology concerning environmental metadata) and the GO (the standard ontology for gene function). MGnify also uses the GOLD database biome hierarchy that allows different granularities of the marine environment to be described in a structured fashion. We do not anticipate generating any new standards or ontologies as part of this project.

3.4. Increase data re-use

As stated already in Section 3.2, most of our data will become completely open immediately. We will be archiving data (using *cold storage*) for long-term preservation during the project, as well as utilising established databases, which have long histories in the provision of data, as well as interoperability. Thus, the data are expected to be still understandable and reusable well beyond the lifetime of the project. We have also committed to using established databases, which have a track record for data management, and these databases will last well beyond the life-time of this project. We have developed a data model that allows the complete and transparent tracing of how derived sequence data products and analyses can be traced back to the original sequence database. To further develop the provenance of the data, we are investigating the use of RO-crates. The analysis pipeline will be extended, but a version process allows users to readily ascertain the version that has been used. During the course of the project we will adapt and improve the informatics analysis pipelines. To further validate the integrity of the results, the following steps will be carried out:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will be implementing the tools into pipelines and workflows using automated methods.
- We will run parts of the data sets repeatedly to capture unexpected changes in results.

Additional documentation will be provided in the form of published scripts, codebooks, R markdown files (e.g. ReadTheDocs) and readme files which will be deposited on GitHub or similar. This documentation will include the quality assurance processes and methodologies used for analysing the different data types. Finally, the use of BioSample accession will increase the discoverability of related multi-omic datasets.

Partners will announce to the consortium their intent to publish data and other digital assets, and allow a period of 45 working days within which the consortium can raise concerns following the project's data sharing agreement as specified in the grant agreement. The project encourages all scientists to make their data and papers publicly available using FAIR principles upon submission of a preprint of their work in a trusted preprint repository such as biorXiv (https:// https://www.biorxiv.org/) or authorea (https://authorea.com), with each data submission characterised by a unique identifier. In cases where the above-mentioned procedure is not possible, project data will be made publicly available in open source and using a CC-BY (or more permissive) licence as the project's default, as soon as the associated paper is published, or at the latest two years after the end of the project.

4. Other research outputs

In addition to data, the BlueRemediomics anticipate making multiple other research outputs, which can be reused within BlueRemediomics, as well as other parties. Briefly, we list these other data products and how they will be produced.

Software: this will be made available to BlueRemediomics partners via their version control systems, with the predominant system used by BlueRemediomics partners being Git. The software should be accompanied with documentation, lists of prerequisites and test data. The software will be made accessible via GitHub or similar. BlueRemediomics partners will be encouraged to enable the software to be installed via Pip or Conda. For complex pieces of software, partners are also encouraged to utilise containers, such as Singularity. Also released software will be versioned, to allow traceability of results.

Workflows: EMBL/MGnify already has complex workflows, which are encapsulated as Nextflow or Common Workflow Language (CWL) pipelines. These workflows are submitted to the WorflowHub (https://workflowhub.eu/), which is increasing their reuse and adaptation. For example, there are Galaxy versions of the MGnify metabarcoding pipeline that replicate the versions deposited in the WorkflowHub.

EMBL will continue to adopt this workflow release procedure (that also includes the software practices described above). This experience will also be shared with others in the BlueRemediomics project, so that additional partners can adopt this practice. Other workflow systems such as Snakemake are also encouraged. Depending on the pipeline and application, these may be made available via the WorkflowHub or via GitHub repositories.

Experimental protocols: Both wet and dry laboratory protocols will be developed as part of BlueRemediomics. For example, particularly during the development of the genome metabolic reconstructions, we anticipate that this will involve a multistep process that can not be encapsulated in an end-to-end informatics pipeline. Both types of wet/dry laboratory protocols will be made available as part of peer-review publication or via other mechanisms, e.g. through protocols.io. We will also investigate the use of Research Object (RO)-crates. This allows the capture of all of the components that went into a piece of research.

AI models: We propose developing ML/AI based tools. Where appropriate, we will use repositories such as Hugging Face to access and utilise models, as well as deposit newly generated models.

Culture collections: As part of the BlueRemediomics project, we aim to connect the data to the marine culture collections (Partners SU, EMBRC).

5. Specific dataset outputs from BlueRemediomics

5.1. Datasets

As outlined in Section 2, MGnify has generated the Marine MAG catalogue (v2.0), which produces a non-redundant set of 13,223 species that encompass 50,866 genomes, collected from both the BlueRemediomics consortium and the wider scientific community. This updated version increases the number of genomes by an order of magnitude (v1.0 contained 1,504 species). The genomes are all drawn from the ENA, and the list of ENA projects is available via the MGnify FTP site: https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/marine/v2.0/README_v2.0.txt, which also provides access to all the genomes used in the catalogue. The species representatives are functionally annotated and made available via MGnify: https://www.ebi.ac.uk/metagenomics/genome-catalogues/marine-v2-0. Other derived data-types are available from the MGnify ftp site: https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/marine/v2.0/. As the process of generating the MGnify genome collections has been published previously (https://doi.org/10.1016/j.jmb.2023.168016), there is yet to be a specific publication associated with this dataset.

5.2. Datasets associated with publications

Datasets that are associated with publications will be reported as part of the Continuous Reporting module under the “Dataset” tab in the EC funders and tenders portal. The following categories will be completed as part of this reporting process, namely:

- Type of PID specified, e.g. DOI, URL, other, or none if not available at the time of reporting.
- Description of the dataset, e.g. genome catalogue, genomic sequence, LCMS.
- PID - link where available.
- If the data underpins a publication, PID of the publication.
- URL to the repository where available.
- Open access status of the dataset.
- Provisions to make data available if required to validate conclusions of a scientific publication.
- Metadata open access status for the datasets.

During the first reporting period, the following datasets associated with publications were reported.

Publication title	Publication link	Dataset description	Dataset link
Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS	https://www.biorxiv.org/content/10.1101/2023.05.23.540769v3.full.pdf	SanntiS software	https://github.com/Finn-Lab/SanntiS https://www.ebi.ac.uk/metagenomics
		4.6M predicted biosynthetic gene clusters.	Available via MGnify (e.g. https://www.ebi.ac.uk/metagenomics/analyses/MGYA00580489?selected_contig=ERZ772995.1-NODE-1-length-272786-cov-19.766374#contigs-viewer) and included as Supplementary Table 2, also available from https://ftp.ebi.ac.uk/pub/databases/metagenomics/sanntis/supplementary_data/Supplementary_Table_1.tsv
Initial Characterization of the Viridisins' Biological Properties	https://pubs.acs.org/doi/10.1021/acsomega.4c03149	Genome sequence data LCMS data* Antimicrobial screening data* Biochemical assay data using zebrafish model*	https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000948985.2/ The raw assay/chromatogram data are recorded in lab books and as digital files which are stored in-house and available upon request*.
The IDSM mass spectrometry extension: searching mass spectra using SPARQL	https://academic.oup.com/bioinformatics/article/40/4/btae174/7638802#447451157	The dataset represents mass spectra and their metadata using the Resource Description Framework (RDF). It is based on data from the MassBank of North America (MoNA).	https://idsm.elixir-czech.cz/sparql/endpoint/idsm https://mona.fiehnlab.ucdavis.edu/
		The dataset represents mass spectra and their metadata using the Resource Description Framework (RDF). It is based on data from the In Silico Spectral Database	https://idsm.elixir-czech.cz/sparql/endpoint/idsm https://doi.org/10.5281/zenodo.8287341

		(ISDB) of natural products calculated from structures aggregated in the frame of the LOTUS Initiative.	
Linking satellites to genes with machine learning to estimate phytoplankton community structure from space	https://os.copernicus.org/articles/20/217/2024/#section7	psbO dataset.The datasets presented in this repository have been used to develop a new ocean color algorithm to derive the relative cell abundance of seven phytoplankton groups (called SOMRCA) and their contribution to total chlorophyll a.	https://zenodo.org/records/10361485 https://www.ebi.ac.uk/biostudies/studies/S-BSST761
		psbO and satellite matchups. The datasets presented in this repository have been used to develop a new ocean colour algorithm to derive the relative cell abundance of seven phytoplankton groups (called SOMRCA) and their contribution to total chlorophyll a.	https://zenodo.org/records/10361485 https://zenodo.org/records/10361485/files/Tara_Oceans_psbO_dataset_Final.xlsx?download=1
		Global HPLC pigment compiled dataset, sources and corresponding outputs of the model (SOMChlF).	https://zenodo.org/records/10361485 https://zenodo.org/records/10361485/files/Assets%20HPLC_SOMChlF.xlsx?download=1
		GlobColour dataset. The datasets presented in this repository have been used to develop a new ocean colour algorithm to derive the relative cell abundance of seven phytoplankton groups (called SOMRCA) and their contribution to total chlorophyll a.	https://zenodo.org/records/10361485 https://hermes.acri.fr/index.php
		SST CCI dataset. The datasets presented in this repository have been used to develop a new ocean colour algorithm to	https://zenodo.org/records/10361485 https://doi.org/10.48670/moi-00169

		derive the relative cell abundance of seven phytoplankton groups (called SOMRCA) and their contribution to total chlorophyll a.	

6. Allocation of resources

The costs for making available BlueRemediomics data as open access are expected to be limited to personnel costs. WP1 will play a primary role in the data management aspects of the project. The partners involved and their respective person month are as follows: EMBL (56.5 PM), CNRS (18 PM), CEA (20 PM), SZN (9 PM), IOCB (15 PM), UCL (30 PM), UWC (4 PM), ETHZ (4 PM) and LMBC (8 PM).

The work performed in WP1 can be roughly divided into 75% data analysis and 25% to ensure that the data and software are managed and made FAIR via the websites. As highlighted above, much of the effort centres on depositing data in ENA and making annotations available in MGnify. None of the repositories anticipated to be used in BlueRemediomics charges for their services. Currently, we do not require any hardware or software in addition to what is usually available in the BlueRemediomics partner institutes. If extra costs occurs within the project lifespan, costs related to open access of research data in Horizon Europe are eligible under the conditions defined in the BlueRemediomics GA Article 6 – Eligible and Ineligible Costs, such as Article 6.2.C.3 – Other goods works and services, but also other articles relevant for the cost category chosen. Project partners will be responsible for including any relevant costs in their financial statements.

Both this initial (D7.2) as well as the interim (7.3) and final (D7.4) versions of the DMP, describe all scientific data outputs. The provision of this information is the responsibility of the WP Leaders and is aggregated by the Coordinator EMBL and sent for revision to all partners before submission. No additional specialist expertise is required to execute the DMP. It will be the responsibility of the Coordinator EMBL, who is also the WP1 leader, to ensure delivery.

Publication fees are only eligible when publishing in fully open access publishing venues (venues in which the entire scholarly content is openly accessible to all) and not in hybrid venues. We have a reserved budget for the time and effort it will take to prepare the data for publication. For making data or other research outputs FAIR, we budgeted: €56,420 (Beneficiaries) and €6,000 (Associated Partners).

7. Data security

Project members will not store data or software longer-term (more than a day) on personal computers in the lab or external hard drives connected to those computers. Instead, they will ensure that data is stored on centralised storage solutions that are routinely backed-up (nightly), thereby mitigating against any data loss. Software will be committed to software version control systems, such as Git, which are also backed-up on a regular basis, as well as using community solutions, such as GitHub, to facilitate sharing across the project partners. During the course of the project, data is expected to be deposited in the appropriate public archives, ensuring long-term data access (see above). Additional data types beyond omics (e.g. annotations) will be deposited in the appropriate knowledge-bases, or deposited in open repositories, such as Zenodo (https://zendo.org).

Sensitive data or data that might be commercially exploitable will not be carried by staff involved in the

project (e.g. on laptops, USB sticks, or other external media). All data centres where project data is stored carry sufficient certifications and have processes involved to project the data. Of note are the resources that are provided by EMBL, which are backed up on a nightly basis, as well as backed up weekly into a tape-based backup pipeline running by EMBL's IT and Technical Services department (ITS). In addition, the data associated with the key databases are replicated to an offsite data centre, which is used to serve the data to the scientific community. Within this data centre, there is also data replication, with fall back servers housed at EMBL-EBI.

All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The risk of information loss in the project or organisation is acceptably low. The possible impact to the project or organisation if information is leaked is small. The possible impact to the project or organisation if information is vandalised is small. We are not using any data that requires controlled access, nor the study of humans that will require GDPR.

The archive will be stored in a remote location to protect the data against disasters. The archive needs to be protected against loss or theft. It is clear who has physical access to the archives.

8. Ethics

BlueRemediomics is committed to carry out its work in alignment with the highest ethical standards of the EU, national and international bodies as well as upholding the values of the EU in all aspects of the consortium's work. For example, we promote knowledge about access and benefit sharing (ABS) and where data has originated from a sample, we have started to introduce information about economic exclusive zones (EEZ) into MGnify, which represents the central data resource in the project. We will continue to regularly update this information as the ABS requirements are dynamic in nature. Currently data does not fall within Nagoya, but through our efforts in WP5, we are monitoring the digital sequence information policy developments and will update this DMP accordingly. We are also engaging with the policy makers to ensure that the policies are enforceable and appropriate.

The following ethical and legal issues have been identified that could impact data sharing, but the issues are perceived to be low risk:

1. Participation of non-EU countries (South Africa, Switzerland, UK), which may involve the transfer of biological material and / or personal data from and / or to the EU. This transfer of materials is monitored through a centralised BlueRemediomics system, as well as specific material transfer agreements between the parties involved.
2. The application of artificial intelligence (AI) that combines sequence data, environmental data and/or experimental results. For example, we will employ ML algorithms to optimise enzymes. However, the AI/ML techniques we will adopt do not raise ethical concerns related to human rights and values, and do not remove humans from the decision making process.

We only collect a very minimal amount of information about individuals involved in the project (name, email address, position and institute). Images of individuals or participation in interviews that will be used in the BlueRemediomics website will be obtained with informed consent. The aforementioned personal details obtained in accordance with GDPR guidelines are stored on secure internal systems and only used for the maintenance of the project.

9. Other issues

To generate the initial version of this DMP, we used the Data Stewardship Wizard with its Common DSW Knowledge Model (ID: dsw:root:2.4.4) knowledge model. Specifically, we used the https://researchers.ds-wizard.org DSW instance where the project has a direct URL: https://researchers.ds-wizard.org/projects/971f5b5f-8a7f-4f9a-a011-3fe62a9a31c7.

We have since restructured the DMP to provide a more comprehensive overview of the data outputs and management plans involved in the BlueRemediomics project. We will not be using any extra national, funder, sectorial, nor departmental policies or procedures for data management.

Beyond these points, no other issues are foreseen at this instant.