EXCELERATE Deliverable D6.1

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	Specific marine databases	
WP No.	6	
Lead Beneficiary:	UiT	
WP Title	Marine metagenomic infrastructure as a driver for research and industrial innovation	
Contractual delivery date:	17 02 2017	
Actual delivery date:	17 02 2017	
WP leader:	Rob Finn, EMBL-EBI Nils Peder Willassen, UiT	1: EMBL-EBI; 24: UiT
Partner(s) contributing to this deliverable:	UiT, EMBL-EBI	

Authors and Contributors:

Inge Alexander Raknes (UiT), Terje Klementsen (UiT), Sudhagar Veerabadran (UiT), Balasundaram, Giacomo Tartari (UiT), Juan Fu (UiT), Espen Robertsen (UiT), Lars Ailo Bongo (UiT), Nils Peder Willassen (UiT), Guy Cochrane (EMBL-EBI), Rob Finn (EMBL-EBI)

Table of content

1. Executive Summary

The marine databases; ***MarRef***, ***MarDb***, and ***MarCat***, are public available resources that promotes marine research and innovation.
The marine resources, which have been implemented in the Marine Metagenomics Portal (MMP), are a collection of richly annotated and manually curated contextual (metadata) and sequence databases representing three tiers of accuracy. While *MarRef* is a database for completely sequenced marine prokaryotic genomes, which represent a marine prokaryote reference genome database, *MarDb* includes all sequenced marine prokaryotic genomes regardless of level of completeness. *MarCat* represent a gene (protein) catalogue of uncultivable (and cultivable) marine genes and proteins derived from metagenomics samples.
The first versions of *MarRef* and *MarDb* contain 484 and 2557 entries, respectively. Each record is build up of 104 metadata fields including attributes for sampling, sequencing, assembly and annotation in addition to organism and taxonomic information. For *MarRef* and *MarDb*, data from various sources, such as sequence, contextual, taxonomy and literature databases, in addition to data from bacterial diversity metadata and culture collection databases has been curated and integrated to produce robust databases. The corresponding genome, gene and protein sequence databases has been build by downloading the individual entries from ENA (European Nucleotide Archive).
MarCat contains currently the Tara Ocean samples containing 1433 entries. In *MarCat* each record contains 103 metadata fields. As for *MarRef* and *MarDb* each entries has been manually curated and enriched with taxonomical annotation, assembly and functional annotation data. The corresponding DNA, gene and protein databases were generated using META-pipe, a pipeline for taxonomic classification and functional annotation of metagenomics sample.
To generate the contextual databases, controlled vocabularies and ontologies are used, which allow a more streamlined curation, better consistency of the data, enhanced quality control (QC) and not least data to be more easily aggregated and analysed. The manual curation of the data produces more robust, richly annotated datasets with highly accurate and detailed information.

The contextual and sequence databases has been incorporated into the Marine Metagenomics Portal (MMP) and are available at https://mmp.sfb.uit.no/

Horizon 2020 grant n. 676559

2. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Development and implementation of selected standards for the marine domain. (Task 6.1)		x
2	Development and implementation of databases specific for the marine metagenomics. (Task 6.2)	x	
3	Evaluation and implementation of tools and pipelines for metagenomics analysis. (Task 6.3)		x
4	Development of a search engine for interrogation of marine metagenomics datasets and establish training workshops for end users. (Task 6.4)		x

3. Delivery and schedule

The delivery is delayed: ☑ Yes √ No

4. Adjustments made

No adjustments was made

Horizon 2020 grant n. 676559

5. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

Work package number	6	Start date or starting event:	month 1
Work package title	Use Case A: Marine metagenomic infrastructure as driver for research and industrial innovation		
Lead	Nils Peder Willassen (NO) and Rob Finn (EMBL-EBI)		
Participant number and person months per participant P1: EMBL-EBI (28PM) - P17: FCG (2PM) - P20: CCMAR (11PM) – P24 UiT (36PM) – P27: CNRS (10PM) - P31: CNR (10 PM)			
Objectives The main objective for this Use Case is to develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain before M36 of the ELIXIR-EXCELERATE project. The main objective will be achieved by the following specific objectives: • Development and implementation of selected standards for the marine domain. (Task 6.1) • Development and implementation of databases specific for the marine metagenomics. (Task 6.2) • Evaluation and implementation of tools and pipelines for metagenomics analysis. (Task 6.3) • Development of a search engine for interrogation of marine metagenomics datasets and establish training workshops for end users. (Task 6.4) ***Description of work*** Metagenomics has the potential to provide unprecedented insight into the structure and function of heterogeneous communities of microorganisms and their vast biodiversity. Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems. They can be exploited e.g. to identify novel biocatalysts for production of fuels or chemicals (bioprospecting), make functional feed for aquaculture species, and for environmental monitoring. However, in order to expand the potential further for the research community and biotech industry, especially within the marine domain, the metagenomics methodologies need to overcome a number of challenges			

Horizon 2020 grant n. 676559

related to standardization, development of relevant databases and bioinformatics tools. New and emerging sequencing technologies, integration of metadata gives an extra burden to the development of future databases and tools. The Use Case "Marine metagenomic infrastructure as driver for research and industrial innovation" will contribute to the overall objectives of the ELIXIR-EXCELERATE project by developing research infrastructure and service provision specific for the marine domain in order to enable metagenomic approaches responding to societal and industrial needs. The outcome of the proposed Use Case will meet the major needs expresses by the marine domain (e.g. ESF Marine board Position Paper 17 "Marine Microbial Diversity and its role in Ecosystem Functioning and Environmental Change" and Position Paper 15 "Marine Biotechnology: A New Vision and Strategy for Europe").

Task 6.1: Development and implementation of a comprehensive metagenomics data standards environment for the marine domain (12 PM)

To maximise the impact and long term utility and discoverability of metagenomics datasets, it is essential the experimental methods and data acquisition/storage protocols be established. In Task 6.1, we will bring together a comprehensive metagenomics data standards environment in collaboration with marine experimental scientists, data providers, end users and the existing communities involved in marine standards development. The environment will bring together three components:

- Data format conventions and standards will address the various data types for which sharing is required, that will include contextual data (e.g. sample information, expedition-related data), metadata (e.g. provenance and tracking information, descriptions of experimental configurations and bioinformatics tools in use) and data (e.g. raw sequence data, aligned reads, taxonomic identifications, gene calls).
- Reporting standards will address community-accepted thresholds for richness/precision that are required to make data useful, including depth of raw machine data, such as resolution of sequence quality scoring, conventions for references to reference assemblies and minimal reporting requirements for contextual data.
- Validation tools will address the automated validation of compliance with conventions and standards and the meeting of minimal reporting expectations for given datasets in preparation by the marine research community. In this task, we will bring together components that exist already – in particular the contextual data and metadata reporting standards we have developed under the Micro B3 project (EU FP7), data standards and conventions developed around our European Nucleotide Archive (ENA) programme, such as CRAM, FASTQ conventions, work existing in the biodiversity and molecular ecology domains (such as tabular data conventions and BIOM matrices) – and construct new components as required. The major output of this work will be a set of well described and navigable elements to aid the marine community in the preparation, sharing, dissemination and publication of highly interoperable and comprehensive metagenomics datasets.

Partners: EMBL-EBI, NO

Horizon 2020 grant n. 676559

Task 6.2. Establishment of marine specific data resources (20PM)
Due to the data biases of existing reference databases, only about one quarter of sequences are annotated, and this fraction diminishes further when more diverse samples such as soil and marine are analyzed. To improve the characterization of marine metagenomic samples, this task involves the construction of sustainable public data resources for the marine microbial domain. Task 6.2 will be achieved by establishing marine microbial databases including reference genomes, nucleotide and protein databases. The established databases, based on the standards developed in Task 6.1, will enhance the precision and accuracy of biodiversity and function analysis. The reference databases will be non-redundant datasets generated from sequences acquired from ENA (as part of the International Nucleotide Sequence Database Collaboration), UniProt and other publicly available datasets. In particularly, we will use some of the higher-coverage and higher quality sequence outputs from the Tara Oceans and Ocean Sampling Day metagenomic projects, to build high quality marine specific reference databases. All datasets will be checked with respect to quality, consistency, and interoperability, and in compliance with standards developed in Task 6.1. The respective knowledge-enhanced databases will be the cornerstone for sustainable analysis of marine metagenomics sequence data. The databases will be developed in collaboration with members of the ESFRI infrastructures EMBRC and MIRRI and made publicly available through ELIXIR.

Partners: NO, EMBL-EBI, IT

Task 6.3: Gold-standards for metagenomics analysis (58PM)
The majority of existing metagenomics analysis platforms, while providing insights into the prokaryotic taxonomic diversity and functional potential for individual samples, but lack the tools that enable discoverability across samples and industrial innovation. This task will focus on the evaluation and implementation of new tools and pipelines in order to accelerate research, discoverability and innovation, reducing time to market for new products. In combination with new standards and databases developed in Task 6.1 and Task 6.2, respectively, new tools for community structure (microbial biodiversity), genetic and functional potential will be evaluated and implemented for environmental applications. For industrial application tools and pipelines for the identification of gene products (e.g. enzymes and drug targets) and pathways will be implemented and made publicly available. The evaluation and implementation will be performed in near collaboration with end-users (research groups, environmental centers, biotech companies) to ensure usability for the end user community in order to improve [ELIXIR-EXCELERATE]
quality, productivity and functionality, as well as reduction of costs for the end-users. New tools and pipelines will be made publicly available through the e.g. META-pipe (ELIXIR-NO), EBI Metagenomics Portal (EMBL EBI) and/or EMBL Embassy cloud technology. Technical requirements will be mapped by WP3 and implemented to meet the requirements of the ELIXIR community. The continued advancement of sequencing technologies and the growing number of public marine metagenomics projects means that it is becoming increasingly difficult to mine these vast datasets. In this task, initially a web-based search engine will be developed for the interrogation of marine metagenomics results available from the EBI Metagenomics Portal, based on combinations of queries to our web services (already in existence, or to be built as part of existing projects outside ELIXIR-EXCELERATE) for the

Horizon 2020 grant n. 676559

discovery of data through metadata, taxonomic and functional fields. This will extend the back-end search functionality that is to be developed as part of on-going efforts. In addition to being downloadable, we will enable search results to flow into an expanded comparison tool (currently limited to gene ontology terms from samples in the same project), to allow more in-depth analysis of a user selected datasets, allowing functional and taxonomic comparisons. In the second phase of this task, the search engine will build upon the data exchange formats in Task 6.1, and federate the search across different pipeline results sets (e.g. META-pipe), so that different results based on the same underlying dataset, can be amalgamated into a single search. This will dramatically enhance the discoverability across different marine datasets, allowing the identification of common trends and/or differences. These tools will be developed using user-experience testing and in collaboration with end users to ensure they are fit for purpose.

Partners: NO, EMBL-EBI, IT, FR, PT

Task 6.4: Training workshops for end users (7PM)
In this task training workshops will be established, in collaboration with WP11 "ELIXIR Training Programme", for end-users with the aim to facilitate accessibility, by training European researchers and industry to more effectively exploit the data, tools and pipelines, and compute infrastructure provided by the ELIXIR marine metagenomics infrastructure. These training workshops and materials will be converted to online training resources, extending the reach of the workshop.

Partners: PT, NO

Appendix 1:

Development and implementation of databases specific for the marine metagenomics.

Horizon 2020 grant n. 676559

Development and implementation of databases specific for marine metagenomics.

Summary

The marine databases; ***MarRef***, ***MarDb***, and ***MarCat***, are public available resources that promotes marine research and innovation.

The marine resources, which have been implemented in the Marine Metagenomics Portal (MMP), are a collection of richly annotated and manually curated contextual (metadata) and sequence databases representing three tiers of accuracy. While *MarRef* is a database for completely sequenced marine prokaryotic genomes, which represent a marine prokaryote reference genome database, *MarDb* includes all sequenced marine prokaryotic genomes regardless of level of completeness. *MarCat* represent a gene (protein) catalogue of uncultivable (and cultivable) marine genes and proteins derived from metagenomics samples.

The first versions of *MarRef* and *MarDb* contain 484 and 2557 entries, respectively. Each record is build up of 104 metadata fields including attributes for sampling, sequencing, assembly and annotation in addition to organism and taxonomic information. For *MarRef* and *MarDb*, data from various sources, such as sequence, contextual, taxonomy and literature databases, in addition to data from bacterial diversity metadata and culture collection databases has been curated and integrated to produce robust databases. The corresponding genome, gene and protein sequence databases has been build by downloading the individual entries from ENA (European Nucleotide Archive).

MarCat contains currently the Tara Ocean samples containing 1433 entries. In *MarCat* each record contains 103 metadata fields. As for *MarRef* and *MarDb* each entries has been manually curated and enriched with taxonomical annotation, assembly and functional annotation data. The corresponding DNA, gene and protein databases were generated using META-pipe, a pipeline for taxonomic classification and functional annotation of metagenomics sample.

To generate the contextual databases, controlled vocabularies and ontologies are used, which allow a more streamlined curation, better consistency of the data, enhanced quality control (QC) and not least data to be more easily aggregated and analysed. The manual curation of the data produces more robust, richly annotated datasets with highly accurate and detailed information.

The contextual and sequence *Mar* databases and are available at https://mmp.sfb.uit.no/

Horizon 2020 grant n. 676559

Background

Microorganisms are ubiquitous in the marine environment, where they play key roles in many biogeochemical processes. With recent advances in community DNA shotgun sequencing (metagenomics) and computational analysis, it is now possible to access the taxonomic and genomic content (microbiome) of marine communities and, thus, to study their diversity, structural patterns, and functional potential. These microorganisms, and the communities they form, drive and respond to changes in the environment and alterations in ocean stratification and currents. With an estimated 10^4 to 10^6 cells per milliliter seawater and totally 10^{29} bacterial cells they also provide the grounds for immense genetic diversity.

All research and innovation is based on comparison to existing knowledge and information. Therefore, sustainable and highly accurate data resources, which are easy to excess, browse and retrieve data from, are vital for performing high-class and beyond the state of art research and innovation.

Up to now, no dedicated data resources exist for the marine metagenomics domain, which not only hamper the utilization of the vast genetic resources for biotechnology research and innovation (biosprospecting), but also impede the development of sustainable tools and resources for example for environmental monitoring, monitoring of fish and shellfish pathogens and development of sustainable feed for marine aquaculture.

Due to the data biases of existing generic reference databases, only about one quarter of sequences in a metagenomics samples are annotated, and this fraction diminishes further when more diverse samples such as marine water and sediment samples are analysed.

Task 6.2 was established in order to construct non-redundant contextual and sequence databases, including genomes/metagenomes, nucleotide and protein databases to improve annotation of marine metagenomic samples.

Overview and status

Definition of marine microorganism

To define "marine microorganism" or "microbial marine biome " is not straightforward since there are many habitats, which are in the borderline such as sandy shores and near river deltas. According to the definition set by the Environmental Ontology[1] a "*marine biome*" is defined as "*An aquatic biome that comprises systems of open-ocean and unprotected coastal habitats, characterized by exposure to wave action, tidal fluctuation, and ocean currents as well as systems that largely resemble these. Water in the marine biome is generally within the salinity range of seawater: 30 to 38 ppt*". This definition does not fit to our mission to ensemble sequence and metadata data of marine microorganisms from the marine environment, since the definition does not include protected costal habitats such as harbours, and estuaries environments.

In order to establish the marine resources we have chosen to define "*marine microbial biome*" as "*An aquatic microbial biome comprises of microbial communities from open-oceans, costal and protected habitats up to the high water mark with salinity from 0.5 ppt as in*

[1] http://purl.obolibrary.org/obo/ENVO_00000447

Horizon 2020 grant n. 676559

estuaries (brackish water) environments to above 100 ppt as in sea ice brine. The biome also includes marine microbial communities obtained from marine species associated with these habitats". We have chosen to include soil samples from sandy shores, intertidal zone, salt marshes (coastal salt marsh or a tidal marsh), mudflats and estuaries, in addition to habitats such as seawater saltern, sea ice brine, black smokers (hydrothermal vents) where the salinity can we extreme high or low compared to seawater. Microorganisms associated with marine species, as defined by the World Register of Marine Species, WoRMS[2], have also been defined as marine. This includes microorganisms associated with or causing diseases in marine animals or plants for example coral, shellfish and fish.

Short description of MarRef, MarDb and MarCat

The construction of the marine sequence databases (BLAST) and their corresponding contextual databases are shown in Fig.1.

The *MarRef*, MarDb and MarCat sequence databases are based on non-redundant genome and metagenome data sets obtained from ENA (European Nucleotide Archive)[3] and/or NCBI (The National Center for Biotechnology Information)[4].

MarRef is a database for completely sequenced marine prokaryotic genomes. MarDb includes all sequenced marine prokaryotic genomes regardless the level of completeness. Each genome assigned as marine microbial biome according to our definition was incorporated in the MarRef and *MarDb* contextual databases, respectively.

***Figure 1**. Database construction procedures in MarRef, MarDb and MarCat.*

MarCat represent a gene (protein) catalogue of uncultivable (and cultivable) marine genes derived from marine metagenomics samples. Metagenomics sequences were obtained from ENA and their corresponding gene and protein feature annotations unique to each sample were generated using META-pipe, a pipeline for taxonomic classification and functional annotation of metagenomics sample.

The corresponding contextual databases, supports the International community-driven standards of the Genomics Standards Consortium[5] and is fully compliant with its recommendations for Minimum Information about any (x) Sequence (MIxS) standards, including MIGS (Minimum Information about a Genome Sequence) and MIMS (Minimal Information of Metagenome sequence). The databases also include the proposed standards for provenance of analysis developed in Task 6.1.

[2] http://www.marinespecies.org/
[3] http://www.ebi.ac.uk/ena
[4] https://www.ncbi.nlm.nih.gov/
[5] http://gensc.org/

Contextual databases

Data collection

The MarRef, MarDb and MarCat contextual databases are built by compiling data from a number of public available sequence, taxonomy and literature databases in a semi-automatic fashion. Other databases or resources such as bacterial diversity and culture collections databases, web mapping service and ontology databases were used extensively for curation of metadata. Resources used in generation of the marine curation databases are shown in Fig. 3 (See also Appendix Table 1).

Figure 2. *Public data resources utilized for construction of MarRef, MarDb and MarCat*

Curation

The imported data files were compiled, converted to Tab separated Value files (.tsv) format and imported into Base, a full-featured desktop database front end, provided by LibreOffice[6]. *MarRef* and *MarDb* contain in total 484 and 2557 entries with 106 metadata fields out of which 30 are represented by controlled vocabularies (CV) and the remaining are free text or numeric fields (Appendix Table 2). These 106 metadata fields include information about sampling environment or host, organism and taxonomy, phenotype, pathogenicity, assembly and annotation information (See Fig. 3). The use of CV and ontologies can shortly be described by the following example. The three environmental metadata fields used for describing the sampling site of the microorganisms; environmental *biome*, *feature* and *material* are controlled by 104 CV terms. The environmental *biome* metadata field contains 11 controlled environmental ontology (ENVO) terms covering environments such as Estuarine biome (ENVO:01000020), Marginal sea biome (ENVO:01000046), Marine benthic

Figure 3. *Base Interface for manually curated of entries.*

[6] https://no.libreoffice.org/

Horizon 2020 grant n. 676559

biome (ENVO:01000024), Marine mud (ENVO:00005795), Marine pelagic biome (ENVO:01000023), Marine water body (ENVO:00001999), and Ocean biome (ENVO:01000048). These ontologies used in the environmental *biome*, *feature* and *material* fields are all well defined and described (http://www.environmentontology.org/) and allows consistency across the datasets.

This first version of *MarCat* contains 1433 entries from the Tara Ocean expedition and each record contains 103 metadata fields. We have included metadata fields for provenance of analysis includes metagenomics analysis metadata such as filtering, assembly, taxonomy, gene prediction and functional assignment.

The databases have links to other public databases. For example in *MarRef* sixteen of the metadata fields have active links to other databases such as the literature databases PubMed and PMC Europe, ontologies such as ENVO and GAZ, sequence databases such as UniProt and ENA, taxonomy databases such as NCBI taxon and Silva, and DMSZ culture collection. These links allows the site visitors to easily access other site in order to obtain more information about each microorganism. For *MarRef* all metadata fields has been manually curated to ensure consistency across the datasets, which allow the end user to easily search and browse entries.

For microorganisms, which have been completely sequenced, more information can be found in databases compared to partially or draft sequenced microorganisms. While MarRef is thoroughly curated, *MarDb* and *MarCat* are only partly curated - a focus has been on curating taxonomic and environmental biome, feature and material.

Entries in the marine databases, *MarRef, MarDb* and *MarCat* follow the MIxS standard guidelines developed by the Genomic Standard Consortium, in addition to other ontologies such as Environmental ontologies (ENVO). Links to other external recourses will such as culture collections, literature and other secondary databases is provided if available. A list of resources is shown in Appendix Table 1.

Refinement

OpenRefine[7] was used for refining the metadata fields by cleaning, trimming of leading and trailing whitespace, transforming data from one format into another and extending it with web services and external data.

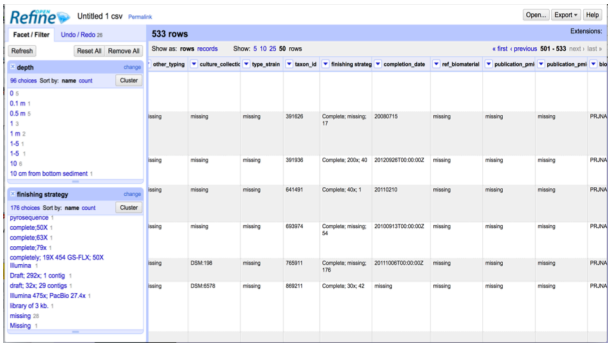

Figure 4. *Refinement of metadata fields*

Validation

A validator was developed to convert Tab Separated Value files (TSV) to Extensible Markup Language files (XML) and from TSV to XML to link the source TSV curation databases to the XML database. The validator

[7] http://openrefine.org/

defines a set of rules for the conversion - warnings and errors during conversion are reported.

Sequence databases

The *MarRef*, *MarDb* and *MarCat* sequence databases are based on non-redundant genome and metagenome data sets obtained mainly from ENA (European Nucleotide Archive) and/or NCBI (The National Centre for Biotechnology Information).

MarRef and MarDb

While *MarRef* is a database for completely sequenced marine prokaryotic genomes, *MarDb* includes all sequenced marine prokaryotic genomes regardless the level of completeness. Each genome assigned as marine microbial biome according to our definition was incorporated in the *MarRef* and *MarDb* contextual databases, respectively. Each annotated genome represents a set of gene and protein feature annotations that are unique to that genome were downloaded from the RefSeq database (NCBI) and used in their respective sequence databases. All RefSeq archaeal and bacterial genomes are annotated using NCBI's prokaryotic genome annotation pipeline PGAAP[8], which improves consistency across the datasets. However, approx. 20% of all entries in *MarDb* did not contain any gene and protein information. These genomes were annotated using Prokka, a command line software tool, for annotation of prokaryote genomes.[9]

MarCat

MarCat represent a gene (protein) catalogue of uncultivable (and cultivable) marine genes derived from metagenomics samples. Metagenomics sequences were obtained from ENA and their corresponding gene and protein feature annotations unique to each sample were generated using META-pipe, a pipeline for taxonomic classification and functional annotation of metagenomics sample[10]. As a start we used the high-coverage and high quality sequence outputs from the Tara Oceans and Ocean Sampling Day metagenomic projects, to build the high quality marine specific reference databases.

Implementation and user interface

The marine reference databases provide login free access to all publicly available data. The reference databases has been incorporated into the Marine Metagenomics Portal (MMP) and implemented using the Hugo static website engine[11]. The website engine reads the reference databases from XML files allows the site visitors to:

- *Browse* each of the databases that allow the user view all database entries.
- *Select* attributes to be visible in the table.

[8] http://www.ncbi.nlm.nih.gov/genome/annotation_prok/
[9] https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu153
[10] https://f1000research.com/articles/6-70/v1
[11] https://gohugo.io/

- *Filter* entries to be visible in the table based on the most important record attribute, such as environmental ontologies (biome, feature and material) and taxonomy (phylum, order and genus).
- *Advanced filtering* allows the site visitor to (i) add one or more filters; (ii) refine current filters by adding new filters or removing already applied filters and (iii) remove all filters and launch a new search.

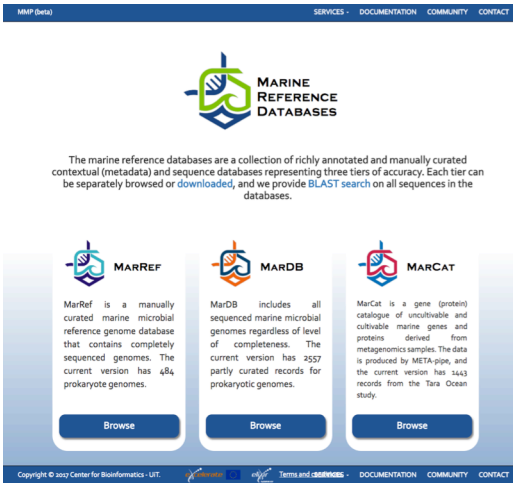

Figure 5. *Access to MarRef, MarDB and MarCat*

The BLAST databases provide similarity search against all nucleotide and protein sequences entries included in *MarRef*, *MarDb* and MarCat. While the MarRef and MarDB databases were build on sequences obtained from the RefSeq database or annotated using the Prokka software package, the *MarCat* database was generated using META-pipe. All metagenomics samples were assembled and annotated - only full length protein coding (CDS) was included in *MarCat*.

The contextual and sequence databases has been incorporated into the Marine Metagenomics Portal (MMP) and are available at https://mmp.sfb.uit.no/

Further plans

The further plans can be classified into five broadly categories; (i) Acquisition of data, (ii) Including viral, eukaryote microbial genomes and transcriptome samples, (iii) Controlled vocabularies, (iv) Implementation of Bioschemas, (v) Downloading of data and (v) User interface.

Acquisition of sequence and contextual data

Collection of data from public available resources will continue. However, due to increasing amount of genomic and metagenomic sequence- and metadata, development of automatic and semi-automatic import scripts that generates data for the curation database will be improved in order to build more efficient import pipelines.

Including viral, eukaryote microbial genomes and transcriptome data

In this first version of the MarRef and MarDb databases only prokaryote genomes has been included. In the future we aim to include virus, eukaryote microbial genomes and transcriptome data. In addition we aim to include more metagenomics and metatranscriptomics data to enhance the quality of the MarCat databases.

Horizon 2020 grant n. 676559

Controlled vocabularies and ontologies

In order to enhance the curation efficiency and provide better consistency of the datasets, the number of metadata fields in the contextual databases will be increased with controlled vocabularies and ontologies. This effort will streamline the curation and data will be more robust, easier aggregated and analysed.

Implementing schema.org markup

To improve data interoperability we intend to implement schema.org markup, so that MMP websites and services contain more structured information. This structured information will make it easier for the end user to discover, collate and analyze our data.

User interface

Site visitors interact with the databases through user interfaces for browsing, filtering and downloading data. We will implement better search and filtering features by including more metadata fields with controlled vocabularies.

Downloading of data

In the first version of the databases the site visitor can only download all sequence (BLAST) and contextual databases. For the contextual databases the data can be downloaded either in TSV or XML format. We aim to implement better systems for downloading single entries or entries selected by searching or filtering of the datasets.

Funding

The work has been conducted as a part of H2020 ELIXIR EXCELERATE project (Grant no. 676559), with support from Research Council of Norway (Grant no 208481/F50) and UiT The Arctic University of Norway.

Appendix

Table 1. Resources used to generate the marine reference databases

Resources	Description	Links
ENA (European Nucleotide Archive)	A comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.	http://www.ebi.ac.uk/ena
NCBI (The National Center for Biotechnology Information)	NCBI houses a series of databases relevant to biotechnology and biomedicine and an important resource for bioinformatics tools and services including GenBank, BLAST and PubMed.	https://www.ncbi.nlm.nih.gov/
UniProt	A comprehensive, high-quality and resource of protein sequence and function	http://www.uniprot.org/
BacDive	BacDive - the Bacterial Diversity Metadatabase merges detailed strain-linked information on the different aspects of bacterial and archaeal biodiversity.	http://bacdive.dsmz.de/
DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH)	The DSMZ (Deutsche sammlung von Mikroorganismen und Zellkulturen GmbH) is one of the largest biological resource centers worldwide, including about 27,000 different bacterial, 4,000 fungal strains and 13,000 different types of bacterial genomic DNA.	https://www.dsmz.de/
OLS	OLS (Ontology Lookup Service) - a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions.	https://www.ebi.ac.uk/ols/index
OLSVis	Visual browser for OLS	http://ols.wordvis.com/
BioPortal	The world's most comprehensive repository of biomedical ontologies	http://bioportal.bioontology.org/
PubMed	A bibliographic database for the biomedical literature.	https://www.ncbi.nlm.nih.gov/pubmed/
Europe PMC	Europe PMC is a repository, providing access to worldwide life sciences articles, books, patents and clinical guidelines.	https://europepmc.org/
Google Maps	Google Maps is a web mapping service developed by Google.	https://www.google.com/maps
Silva	High quality ribosomal RNA database	https://www.arb-silva.de/
NCBI taxonomy browser	A curated classification and nomenclature database for all of the organisms in public sequence databases	https://www.ncbi.nlm.nih.gov/taxonomy
Patric	An information system designed to support the biomedical research community on bacterial infectious diseases via integration of vital pathogen information with rich data and analysis tools.	https://www.patricbrc.org/
Gold	Genomes Online Database - a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata.	https://gold.jgi.doe.gov/index
WoRMS	World Register of Marine Species (WoRMS) provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy.	http://www.marinespecies.org/

Table 2. Attributes included in MarDb, MarRef and MarCat

Structured comment name	Item	Description	Examples	Expected value	Value syntax	Prefered units / suffix
alt_elev	Geographic location (altitude/elevation)	Sample taken at given elevation above sea level, defined in meters(m) as a positive floating number with two decimals.	Ex 1: 3.06 Ex 2: 1.80-2.15	-	{float} or {range}	meters (m)
collection_date	Collection date	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated.	Ex 1: 2008-01-23T19:23:10+00:00 Ex 2: 2011-11-10 Ex 3: 2001-12	date and time, range	{timestamp}	-

Horizon 2020 grant n. 676559

			Ex 7: 2015 Ex 4: 2003--2006 Ex 5: 2010-01--2011-03 Ex 6: 2011-05-28--2011-08-10			
depth	Depth	Please refer to the definitions of depth in the environmental packages. Water: Sample taken at given depth below sea level, defined in meters(m) as a positive floating number or as a range, both with two decimals.	Ex 1: 355.20 Ex 2: 2.00-5.00	-		meters (m)
env_biome	Environment (biome)	In environmental biome level are the major classes of ecologically similar communities of plants, animals, and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Examples include: desert, taiga, deciduous woodland, or coral reef. EnvO (v1.53) terms listed under environmental biome can be found from the link:(http://www.environmentontology.org/Browse-EnvO)	Ex 1: coral reef Ex 2: tropical	EnvO	{free text}	-
env_biome_ENVO	Environment (biome_id)	Corresponding ENVO identifier related to the term name of Environment (biome).	Ex 1: ENVO:00000150 Ex 2: ENVO:01000204	EnvO	{accession}	-
env_feature	Environment (feature)	Environmental feature level includes geographic environmental features. Examples include: harbor, cliff, or lake. EnvO (v1.53) terms listed under environmental feature can be found from the link:(http://www.environmentontology.org/Browse-EnvO)	Ex 1: coast Ex 2: ocean floor	EnvO	{term}	-
env_feature_ENVO	Environment (feature_id)	Corresponding ENVO identifier related to the term name of Environment (feature).	https://www.ebi.ac.uk/metagenomics/projects/SRP000183/samples/SRS000447	EnvO	{accession}	-
env_material	Environment (material)	The environmental material level refers to the matter that was displaced by the sample, prior to the sampling event. EnvO (v1.53) terms listed under environmental matter can be found from the link:(http://www.environmentontology.org/Browse-EnvO)	Ex 1: sea water Ex 2: ice	EnvO	{term}	-
env_material_ENVO	Environment (material_id)	Corresponding ENVO identifier related to the term name of Environment (material).	Ex 1: ENVO:00002149 Ex 2: ENVO:01000277	EnvO	{accession}	-
env_package	Environmental package	MIGS/MIMS/MIMARK extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported.	Ex 1: Water Ex 2: Host-associated	CV, single data entry	[Air|Host-associated|Microbial mat/biofilm|Misc environment|Plant-associated|Sediment|Soil|Wastewater/sludge|water]	-
env_salinity	Environmental salinity	Please refer to the definitions of salinity in the environmental packages. Water: Salinity measurement, given in practical salinity units (psu) as a positive number with two decimals.	39.14	-	{float} {unit}	psu
env_temp	Environmental temperature	Package defined temperature. Temperature in degrees Celsius of the sample at time of sampling, given in degrees Celsius as positive or negative numbers with two decimals.	16.25	-	{float} {unit}	°C
geo_loc_name	Geographic location (country and/or sea, region)	The geographical origin of the sample as defined by the country or sea name followed by specific region name. Country or sea names should be chosen from the INSDC country list here. Source: (http://insdc.org/country.html)	Japan:Kochi Prefecture:Cape Muroto	country or sea name (INSDC):region:specific location name	{string}:{string}:{string}	-

Horizon 2020 grant n. 676559

geo_loc_name_GAZ	Geographic location (GAZ)	The GAZ ontology (v1.446) may also be used to specify the location of sampling. Source: (http://purl.bioontology.org/ontology/GAZ)	Goto Islands	gaz name	{term}	-
geo_loc_name_GAZ_ENVO	Geographic location (GAZ_id)	Corresponding Gazetteer (GAZ) identifier related to the term name of Geographic location. LinkOut to EMBL Ontology Lookup Service, see example for GAZ:00045749.	https://www.ebi.ac.uk/ols/ontologies/gaz/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FGAZ_00045749	gaz number	{accession}	-
investigation_type	Investigation type	Nucleic Acid Sequence Report is the root element of all MIGS/MIMS compliant reports as standardized by Genomic Standards Consortium. This field is either eukaryote,bacteria,virus,plasmid,organelle, metagenome, miens-survey or miens-culture.	bacteria	CV, single data entry	[Eukaryote|Bacteria|Archaea|Plasmid|Virus|Organelle|Metagenome|Mimarks-survey|Mimarks-specimen|NA|unknown|missing]	-
lat_lon	Coordinates (latitude and longitude)	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees, positive and negative numbers, and according to the WGS84 system. This can be found using google maps.	69.692320, 18.973551	decimal degrees	{string}:{string}	-
pathogenicity	Known pathogenicity	To what is the entity pathogenic. Ex: human, animal, plant, fungi, bacteria	Animal	Free text	{string}	-
project_name	Project name	Name of the project within which the sequencing was organized.	-	Free text	{text}	-
assembly	Assembly	How was the assembly done (e.g. with a text based assembler like phrap or a flowgram assembler); estimated error rate associated with the finished sequences (e.g. error rate of 1 in 1000 bp); and the method of calculation. Source: (https://www.ncbi.nlm.nih.gov/assembly)	Newbler assembler v. 2.5	Free text	{text}	-
isol_growth_condt	Isolation and growth condition	Publication reference in the form of pubmed ID (pmid), digital object identifier (doi) or url for isolation and growth condition specifications of the organism/material. PubMed ID or DOI ID. Source: (https://www.ncbi.nlm.nih.gov/pubmed)	Ex 1: 15393697 Ex 2: 10.1007/BF00210994 Ex 1 LinkOut: https://www.ncbi.nlm.nih.gov/pubmed/?term=15393697 Ex 2 LinkOut: https://doi.org/10.1007/BF00210994	PMID,DOI, URL or PMCID, single data entry or a list	{string} or {list}	-
num_replicons	Number of replicons	Reports the number of replicons in a nuclear genome of eukaryotes, in the genome of a bacterium or archaea or the number of segments in a segmented virus. Always applied to the haploid chromosome count of a eukaryote. Source: (https://www.ncbi.nlm.nih.gov/assembly)	3	for eukaryotes and bacteria: chromosomes (haploid count); for viruses: segments	{integer}	-
ref_biomaterial	Reference for biomaterial	Primary publication if isolated before genome publication; otherwise, primary genome report. PubMed ID or DOI ID. Source: (https://www.ncbi.nlm.nih.gov/pubmed)	Ex 1: 15393697 Ex 2: 10.1007/BF00210994 Ex 1 LinkOut: https://www.ncbi.nlm.nih.gov/pubmed/?term=15393697 Ex 2 LinkOut: https://doi.org/10.1007/BF00210994	PMID,DOI, URL or PMCID, single data entry or a list	{string} or {list}	-
microbe_package	Microbe	A package represents a type of BioSample (NCBI) and specifies the list of attributes by which it should be described. Use for bacteria or other unicellular microbes when it is not appropriate or advantageous to use MIxS, Pathogen or Virus packages. See: https://www.ncbi.nlm.nih.gov/biosample/docs/packages/ See: http://www.ncbi.nlm.nih.gov/biosample/SAMN02911891	Microbe version 1.0	Free text	{string}	-

sample_type	Sample type	Sample type, such as cell culture, mixed culture, tissue sample, whole organism, single cell, metagenomic assembly. Source: (https://www.ncbi.nlm.nih.gov/biosample/)	Enrichment culture	Free text	{string}	-
strain	Strain	Microbial or eukaryotic strain name. Microbial designated/original strain name followed by sequenced strain from culture collection. Multiple values allowed separated by ':'. Source: (http://bacdive.dsmz.de)	Ex 1: N-2927 Ex 2: Och 323:DSM 19469:CIP 107377 Ex 3: 7-ME:DSM 22347:LMG 26961	Free text	{text}	-
isolation_source	Isolation source	Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived.	Mid-Okinawa Trough hydrothermal sediments	Free text	{text}	-
collected_by	Collected by	Name of persons or institute who collected the sample	R. Smith	Free text	{text}	-
culture_collection	Culture collection	Name of source institute and unique culture identifier. See the description for the proper format and list of allowed institutes, http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier. The sample represented by the entry can be stored in one or several culture collections like DSMZ, ATCC and JCM. Source: (http://bacdive.dsmz.de)	CIP 109004:DSM 16917:KCTC 32106	Culture collection abrevation followed by ID number. Single data entry or a list.	{string} or {list}	-
bacdive_id	BacDive ID {LinkOut}	The given ID number of the entry in BacDive meta-database. Source: (http://bacdive.dsmz.de)	http://bacdive.dsmz.de/index.php?search=2546	BacDive ID accession number	URL{integer}	-
curation_date	Curation date	The date in which the entry curation was completed.	2016-11-28	YYYY-MM-DD	{timestamp}	-
implementation_date	Implementation date	The date in which the entry was added to the database.	2016-11-28	YYYY-MM-DD	{timestamp}	-
updated_date	Updated	The date in which the entry was updated with more recent information than originally curated.	2016-12-05	YYYY-MM-DD	{timestamp}	-
mmp_biome	-	Marked as marine according to the MMP standard.	Marine	-	(string)	-
base_ID	-	Individual numeric ID for MarRef, MarDB and MarCat, which is only needed for identification in LObase.	1	-	{integer}	-
mmp_ID	-	Individual numeric ID for MarRef, MarDB and MarCat that corresponds to the numeric system in biosample_accession. Incase of duplicate BioSample number we use ".1, .2 ..." as an aditional integer.	Ex 1: MMP02256433 Ex 2: MMP02954345.1 Ex 3: MMP02954345.2	-	{accession}	-
silva_accession_SSU	Silva accession ID {LinkOut}	Link out to the SILVA Small Subunit rRNA Database (SSU). Number in the BioProject identifier is used to lookup the entry in SILVA. See example for how the BioProject id PRJEA30703 is used.	https://www.arb-silva.de/search/show/ssu/insdc/30703	URL followed by the BioProject number in the ID.	{accession}	-
silva_accession_LSU	Silva accession ID {LinkOut}	Link out to the SILVA Large Subunit rRNA Database (SSU). Number in the BioProject identifier is used to lookup the entry in SILVA. See example for how the BioProject id PRJEA30703 is used.	http://www.arb-silva.de/search/show/lsu/insdc/30703	URL followed by the BioProject number in the ID.	{accession}	-
uniprot_accession	UniProt proteome ID {LinkOut}	The given ID number of the entry in UniProt Proteomes. Source: (http://www.uniprot.org/proteomes/)	http://www.uniprot.org/proteomes/UP000027362	-	{accession}	-
assembly_accession	ENA Assembly accession ID	The given accession number of the entry in NCBI/ENA Assembly.	http://www.ebi.ac.uk/ena/data/view/GCA_0004002	Assembly accession	{accession}	-

	{LinkOut}	Source: (https://www.ncbi.nlm.nih.gov/assembly)	45.1	number		
bioproject_accession	ENA BioProject accession ID {LinkOut}	The given accession number of the entry in NCBI/ENA BioProject. Source: (https://www.ncbi.nlm.nih.gov/bioproject/)	http://www.ebi.ac.uk/ena/data/search?query=PRJNA40879	Bioproject accession number	{accession}	-
biosample_accession	ENA BioSample accession ID {LinkOut}	The given accession number of the entry in NCBI/ENA BioSample. Source: (https://www.ncbi.nlm.nih.gov/biosample/)	http://www.ebi.ac.uk/biosamples/samples/SAMN02589594	Biosample accession number	{accession}	-
genbank_accession	ENA GenBank accession ID {LinkOut}	The given accession number of the entry in NCBI/ENA GenBank. Source: (https://www.ncbi.nlm.nih.gov/genbank/)	http://www.ebi.ac.uk/ena/data/view/JQOJ01000000	Genbank accession number	{accession}	-
NCBI_refseq_accession	NCBI Refseq accession ID {LinkOut}	The given accession number of the entry in NCBI RefSeq. Source: (https://www.ncbi.nlm.nih.gov/refseq/)	https://www.ncbi.nlm.nih.gov/nuccore/NZ_BACE00000000	NCBI refseq accession number	{accession}	-
NCBI_taxon_identifier	NCBI Taxon identifier {LinkOut}	The given taxon ID number of the entry in NCBI Taxonomy. Source: (https://www.ncbi.nlm.nih.gov/taxonomy)	https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1454200	ID number	{accession}	-
annotation_provider	Annotation provider	-	NCBI	-	{string}	-
annotation_date	Annotation date	-	Ex 1: 2008-01-23T19:23:10+00:00 Ex 2: 2011-11-10 Ex 3: 2001-12	date and time	{timestamp}	-
annotation_pipeline	Annotation pipeline	Name of the pipeline	NCBI Prokaryotic Genome Annotation Pipeline	-	{string}	-
annotation_method	Annotation method	Databases and tools integrated in the pipeline	Best-placed reference protein set:GeneMarkS	List of databases and tools integrated in the given pipeline.	{string}:{string}	-
annotation_software_revision	Annotation software revision	-	3.0	-	{float}	-
features_annotated	Features annotated	-	Gene:CDS:rRNA:tRNA:ncRNA:repeat_region	List of features annotated	{string}:{string}:{string}:{string}:{string}:{string}	-
genes	Genes	-	3717	-	{integer}	-
cds	CDS	-	3625	-	{integer}	-
pseudo_genes	Pseudo genes	-	24	-	{integer}	-
rrnas	rRNAs	The defined number of 5S, 16S and 23S predicted. Field can not be empty.	2, 1, 2; 0, 0, 0	Integer list containing a constant of three numbers that represents 5S, 16S and/or 23S sequences.	{integer}, {integer}, {integer}	5S, 16S, 23S
complete_rrnas	Complete rRNAs	The defined number of 5S, 16S and 23S predicted. Field cannot be empty.	2, 3, 5	Integer list containing a constant of three numbers that represents 5S, 16S and/or 23S sequences.	{integer}, {integer}, {integer}	5S, 16S, 23S
partial_rrnas	Partial rRNAs	The defined number of 5S, 16S and 23S predicted. Field cannot be empty.	0, 3, 0	Integer list containing a	{integer}, {integer},	5S, 16S, 23S

				constant of three numbers that represents 5S, 16S and/or 23S sequences.	{integer}	
trnas	tRNAs	-	57	-	{integer}	-
ncrna	ncRNA	-	1	-	{integer}	-
frameshifted_genes	Frameshifted genes	-	12	-	{integer}	-
frameshifted_genes_on_monomer_runs	Frameshifted genes on monomer runs	-	2	-	{integer}	-
frameshifted_genes_not_on_monomer_runs	Frameshifted genes not on monomer runs	-	5	-	{integer}	-
host_common_name	Host common name	The common name of the organism e.g. Human.	Salmon	Organism common name	{text}	-
host_scientific_name	Host	The natural (as opposed to laboratory) host to the organism from which the sample was obtained. Use the full taxonomic name, e.g. "*Homo sapiens*".	Salmo salar	Host scientific name	{text}	-
organism	Organism	The most descriptive organism name for this sample (to the species, if relevant but without strain or culture collection name).	Mobilicoccus pelagius	Text	{text}	-
full_scientific_name	Full Scientific Name	Scientific name with authority, strain or culture collection appended. Source: (http://bacdive.dsmz.de)	Stigmatella aurantiaca Berkeley and Curtis 1875	Species name followed by authority or designated strain/culture collection.	{text}	-
disease	Disease	Relevant disease related to the sample. We restrict our consideration to two parent MeSH headings: Gram-Positive Infections, and Gram-Negative Infections. Source: (https://www.ncbi.nlm.nih.gov/mesh)	Botulism	Free text	{text}	-
publication_pmid	Publication {LinkOut}	Publication(s) including the genome data. PMID, DOI or URL. Source: (https://www.ncbi.nlm.nih.gov/pubmed)	Ex 1: 15393697 Ex 2: 10.1007/BF00210994 Ex 1 LinkOut: https://www.ncbi.nlm.nih.gov/pubmed/?term=15393697 Ex 2 LinkOut: https://doi.org/10.1007/BF00210994	PMID,DOI, URL or PMCID, single data entry or a list	{string} or {list}	-
isolation_country	Isolation country	Same as geo_loc_name	Philippines	Free text	{text}	-
isolation_comments	Isolation comments	PATRIC	isolated at a depth of 2500 m from the Sulu Trough	Free text	{text}	-
comments	Comments	PATRIC	Genomic DNA from Strain IMCC9063 from SAR11 group 3 isolated from an Arctic Envirionment.	Free text	{text}	-
sequencing_centers	Sequencing center	PATRIC	Zhejiang University	-	{text}	-
body_sample_site	Body sample site	PATRIC	kidney	Free text	{text}	-

body_sample_subsite	Body sample subsite	PATRIC	-	Free text	{text}	-
other_clinical	Other clinical	PATRIC	host_health_state:diseased	Free text	{text}	-
gram_stain	Gram stain	PATRIC http://trace.ddbj.nig.ac.jp/bioproject/submission_e.html#Gram	Positive	CV, single data entry	[Positive|Negative|unknown|NA|missing]	-
cell_shape	Cell shape	PATRIC, For more information see: http://trace.ddbj.nig.ac.jp/bioproject/submission_e.html#Shape	Bacilli	CV, single data entry	[Bacilli (Rod)|Cocci (Spherical)|Spirilla (Spiral)|Coccobacilli (Elongated coccal)|Filamentous (Long threds)|Vibrios (Slightly curved rod)|Fusobacteria (Spindle)|Square Shaped|Curved Shaped|Tailed|Oval|unknown|NA|missing]	-
motility	Motility	PATRIC http://trace.ddbj.nig.ac.jp/bioproject/submission_e.html#Motility	Yes	CV, single data entry	[Yes|No|unknown|NA|missing]	-
sporulation	Sporulation	PATRIC http://trace.ddbj.nig.ac.jp/bioproject/submission_e.html#Endospores	No	CV, single data entry	[Yes|No|unknown|NA|missing]	-
temperature_range	Temperature range	Bacteria growth temperature tolerance.	Psychrophilic	CV, single data entry	[Cryophilic|Psychrophilic|Psychrotolerant|Mesophilic|Thermophilic|Hyperthermophilic|unknown|NA|missing]	-
optimal_temperature	Optimal growth temperature	The optimum growth temperature of the bacteria. Given as a single positive or negative number or as a range with two decimals.	Ex 1: 37.23 Ex 2: 15.00-18.50	-	{float}	°C
halotolerance	Halotolerance	Slight halophiles prefer 0.3 to 0.8 M (1.7 to 4.8% — seawater is 0.6 M or 3.5%), moderate halophiles 0.8 to 3.4 M (4.7 to 20%), and extreme halophiles 3.4 to 5.1 M (20 to 30%) salt content. Halophiles require sodium chloride (salt) for growth, in contrast to halotolerant organisms, which do not require salt but can grow under saline conditions. http://trace.ddbj.nig.ac.jp/bioproject/submission_e.html#Salinity	Extreme halophilic	CV, single data entry	[Halophilic|Halotolerant|Non Halophilic|Moderate Halophilic|Extreme Halophilic|Euryhaline|Stenohaline||unknown|NA|missing]	-
oxygen_requirement	Oxygen requirement	PATRIC http://trace.ddbj.nig.ac.jp/bioproject/submission_e.html#OxygenReq	Aerobe	CV, single data entry	[Aerobic|Microaerophilic|Facultative|Anaerobic|unknown|NA|missing]	-
plasmids	Plasmids	PATRIC	4	-	{integer}	-
genome_length	Genome length	PATRIC	1518636	-	{integer}	bp
gc_content	GC content	Float with two decimal numbers.	29.84	-	{float}	%
refseq_cds	Refseq CDS	0 = missing	1490	-	{integer}	-
biovar	Biovar	A biovar is a variant prokaryotic strain that differs physiologically and/or biochemically from other strains in a particular species.	Ex 1: Biovar 1 (parvo) Ex 2: Biovar 2 (T960)	-	{text}	-

other_typing	Other typing	PATRIC	genotype:Wild type	typing:term	{typing}:{term}	-
type_strain	Type strain	PATRIC	Yes	CV, single data entry	[Yes|No|unknown|NA|missing]	-
sequencing_platform	Sequencing Platform	Sequencing method used; e.g. Sanger, pyrosequencing, ABI-solid. Data source: (https://www.ncbi.nlm.nih.gov/assembly) Options: http://trace.ddbj.nig.ac.jp/dra/submission_e.html#Instrument	Hi Seq Illumina:Ion PGM:PacBio	free text, single data entry or a list	[454 GS|454 GS 20|454 GS FLX|454 GS FLX+|454 GS FLX Titanium|454 GS Junior|Illumina Genome Analyzer|Illumina Genome Analyzer II|Illumina Genome Analyzer IIx|Illumina HiSeq|Illumina HiSeq 1000|Illumina HiSeq 1500|Illumina HiSeq 2000|Illumina HiSeq 2500|Illumina HiSeq 3000|Illumina HiSeq 4000|Illumina MiSeq|Illumina HiScanSQ|HiSeq X Five|HiSeq X Ten|NextSeq 500|NextSeq 550|Helicos HeliScope|AB SOLiD System|AB SOLiD System 2.0|AB SOLiD System 3.0|AB SOLiD 3 Plus System|AB SOLiD 4 System|AB SOLiD 4hq System|AB SOLiD PI System|AB 5500 Genetic Analyzer|AB 5500xl Genetic Analyzer|AB 5500xl-W Genetic Analysis System|Complete Genomics|MinION|GridION|PromethION|PacBio RS|PacBio RS II|Sequel|Ion Torrent|Ion Torrent PGM|Ion Torrent Proton|AB|AB Genetic Analyzer|AB 377 Genetic Analyzer|AB 310 Genetic Analyzer|AB 3130 Genetic Analyzer|AB 3130xL Genetic Analyzer|AB 3500 Genetic Analyzer|AB 3500xL Genetic Analyzer|AB 3730 Genetic Analyzer|AB 3730xL Genetic Analyzer|Sanger sequencing]	-

sequencing_depth	Sequencing depth	-	25.3:102.8	-	{float}	x:x
contigs	Contigs	-	125	-	{integer}	-
genome_status	Genome status	Complete, Draft	Complete	-	[Complete|Draft]	-
host_sex	Host sex	Physical sex of the host	male	CV, single data entry	[Male|Female|Neuter|Hermaphrodite|Not determined|NA|unknown|missing]	-
host_health_stage	Host health	Health state of host	diseased	Host condition at the time of sampling	{text}	-
host_age	Host age	Age of host at the time of sampling; relevant scale depends on species and study, e.g. could be seconds for amoebae or centuries for trees.	Ex 1: 5 years Ex 2: 26 days Ex 3: 2.5-3.0 hours	Life stage or length of life	{float}	minutes, hours, days, weeks, years, decades, centuries
serovar	Serovar	Epidemiologic classification of organisms to the sub-species level based on their cell surface antigens. May be sub-classifications below biovar.	O4:K8	-	{text}	-
pathovar	Pathovar	The term pathovar is used to refer to strains with similar features that are differentiated at the subspecies level on the basis of differences in plant host range and types of symptoms, and additionally by biochemical profiles. http://www.isppweb.org/about_tppb_naming.asp	Pseudomonas syringae pv. lachrymans abbreviated - P.s. pv. lachrymans	-	{text}	-
kingdom	Kingdom	NCBI taxonomy	Bacteria	Free text	{string}	-
phylum	Phylum	NCBI taxonomy	Bacteroidetes	Free text	{string}	-
class	Class	NCBI taxonomy	Flavobacteriia	Free text	{string}	-
order	Order	NCBI taxonomy	Flavobacteriales	Free text	{string}	-
family	Family	NCBI taxonomy	Flavobacteriaceae	Free text	{string}	-
genus	Genus	NCBI taxonomy	Algibacter	Free text	{string}	-
species	Species	NCBI taxonomy	Altererythrobacter marensis	Free text	{string}	-
taxon_lineage_ids	Taxon lineage ids	The complete identifier lineage to the most specific taxon of the organism.	131567:2;1224:28211:204457:335929:361177:543877	List	{string}:{string}:{string}:{string}:{string}:{string}:{string}	-
taxon_lineage_names	Taxon lineage names	The complete name lineage from cellular organisms to the most specific taxon of the organism.	cellular organisms:Bacteria:Proteobacteria:Alphaproteobacteria:Sphingomonadales:Erythrobacteraceae:Altererythrobacter:Altererythrobacter marensis	List	{string}:{string}:{string}:{string}:{string}:{string}:{string}	-

Horizon 2020 grant n. 676559

Table 3. Controlled vocabularies for Environmental biome

Preferred Name	Definitions	ENVO ID	Link
Marine biome	An aquatic biome that comprises systems of open-ocean and unprotected coastal habitats, characterized by exposure to wave action, tidal fluctuation, and ocean currents as well as systems that largely resemble these. Water in the marine biome is generally within the salinity range of seawater: 30 to 38 ppt.	ENVO:00000447	http://purl.obolibrary.org/obo/ENVO_00000447
Epeiric sea biome	The epeiric sea (also known as an epicontinental sea) biome comprises of a shallow seas that extend over part of a continent. Epeiric seas are usually associated with the marine transgressions of the geologic past, which have variously been due to either global eustatic sea level changes, local tectonic deformation, or both, and are occasionally semi-cyclic.	ENVO:01000045	http://purl.obolibrary.org/obo/ENVO_01000045
Estuarine biome	Expressions of the estuarine biome occur at wide lower courses of rivers where they flow into a sea. Estuaries experience tidal flows and their water is a changing mixture of fresh and salt.	ENVO:01000020	http://purl.obolibrary.org/obo/ENVO_01000020
Marginal sea biome	The marginal sea biome comprises parts of an ocean partially enclosed by land such as islands, archipelagos, or peninsulas. Unlike Mediterranean seas, marginal seas have ocean currents caused by ocean winds. Many marginal seas are enclosed by island arcs that were formed from the subduction of one oceanic plate beneath another.	ENVO:01000046	http://purl.obolibrary.org/obo/ENVO_01000046
Marine benthic biome	The marine benthic biome (benthic meaning 'bottom') encompasses the seafloor and includes such areas as shores, littoral or intertidal areas, marine coral reefs, and the deep seabed.	ENVO:01000024	http://purl.obolibrary.org/obo/ENVO_01000024
Marine mud	A liquid or semi-liquid mixture of water and some combination of soil, silt, and clay.	ENVO:00005795	http://purl.obolibrary.org/obo/ENVO_00005795
Marine pelagic biome	The marine pelagic biome (pelagic meaning open sea) is that of the marine water column, from the surface to the greatest depths.	ENVO:01000023	http://purl.obolibrary.org/obo/ENVO_01000023
Marine salt marsh biome	The marine salt marsh biome comprises marshes that are transitional intertidals between land and salty or brackish marine water (e.g.: sloughs, bays, estuaries). It is dominated by halophytic (salt tolerant) herbaceous plants. The daily tidal surges bring in nutrients, which tend to settle in roots of the plants within the salt marsh. The natural chemical activity of salty (or brackish) water and the tendency of algae to bloom in the shallow unshaded water also allow for great biodiversity.	ENVO:01000022	http://purl.obolibrary.org/obo/ENVO_01000022
Marine upwelling biome	A marine biome which contains communities adapted to living in an environment determined by an upwelling process.	ENVO:01000858	http://purl.obolibrary.org/obo/ENVO_01000858
Marine water body	A significant accumulation of water which is part of a marine biome. Ideas like "significant" are fuzzy and need to be modelled more accurately. The definition is a candidate for review.	ENVO:00001999	http://purl.obolibrary.org/obo/ENVO_00001999
Mediterranean sea biome	The Mediterranean sea biome comprises mostly enclosed seas that have limited exchange of deep water with outer oceans and where the water circulation is dominated by salinity and temperature differences rather than winds.	ENVO:01000047	http://purl.obolibrary.org/obo/ENVO_01000047
Ocean biome	The ocean biome comprises major bodies of saline water, principal components of the hydrosphere. Approximately 71% of the Earth's surface is covered by ocean, a continuous body of water that is customarily divided into several principal oceans and smaller seas. More than half of this area is over 3,000 metres (9,800 ft) deep. Average oceanic salinity is around 35 parts per thousand (ppt) (3.5%), and nearly all seawater has a salinity in the range of 30 to 38 ppt.	ENVO:01000048	http://purl.obolibrary.org/obo/ENVO_01000048

Table 3. Controlled vocabularies for Environmental feature

Preferred Name	Definitions	ENVO ID	Link
Environmental feature	A material entity determines an environmental system when its removal would cause the collapse of that system. For example, a seamount determines a seamount environment, acting as its 'hub'. This class is currently being aligned to the Basic Formal Ontology. Following this alignment, its definition and the definitions of its subclasses will be revised. A material entity, which determines an environmental system. Includes environmental zones, geographic features, glacial feature, microscopic physical objects, volcanic feature and organic feature.	ENVO:00002297	http://purl.obolibrary.org/obo/ENVO_00002297

Environmental zone or environmental area	An environmental zone is an environmental feature whose extent is determined by the presence or influence of one or more material entities or processes. An environmental zone may, itself, assume the role of an environmental feature. For example, a intertidal zone is that part of a coast which is exposed to air and water due to tidal processes. It determines the intertidal zone environment. This class is experimental and not suitable for annotation. Includes: circalittoral zone (ENVO:01000412), field (ENVO:01000352), high tide zone (ENVO:00000318), infralittoral zone (ENVO:01000411), intertidal zone (ENVO:00000316), iron-reducing zone of petroleum contaminated sediment (ENVO:00002178), littoral zone (ENVO:01000407), low tide zone (ENVO:00000319), marine anoxic zone (ENVO:01000066),marine eulittoral zone (ENVO:01000410), marine oligotrophic desert (ENVO:01000073), marine sub-littoral zone (ENVO:01000126), marine supra-littoral zone (ENVO:01000124).	ENVO:01000408	http://purl.obolibrary.org/obo/ENVO_01000408
Geographic feature	May appear on a map. Includes: archipelago (ENVO:00000220), volcanic arc (ENVO:00000351), Arrugado (ENVO:00000538), Cost (ENVO:01000687), sea cost (ENVO:00000303), brackish estuary (ENVO:00002137), estuarine water (ENVO:01000301, estuarine mud (ENVO:00002160), saline wedge estuary (ENVO:00000226), lagoon (ENVO:00000038), mangrove swamp (ENVO:00000057), Mudflat (ENVO:00000192), tidal mudflat (ENVO:00000241), Sea beach (ENVO:00000092), Sea cliff (ENVO:00000088), Sea shore (ENVO:00000485), harbor (ENVO:00000463), marine current (ENVO:01000067). Further hydrographic feature (ENVO:00000012) such as algal bloom (ENVO:2000004), ocean floor (ENVO:00000426), sea floor (ENVO:00000482) and Marine pelagic feature (ENVO:01000044) such as mid-ocean ridge (ENVO:00000406), Ocean current (ENVO:00000147), Ocean basin (ENVO:00002450), Ocean trench, (ENVO:00000275), marine reef (ENVO:01000143), island (ENVO:00000098), atoll (ENVO:00000166), marine hydrothermal vent chimney (ENVO:01000129), oil seep (ENVO:00002063), oil reservoir (ENVO:00002185), peninsula (ENVO:00000305), continental margin (ENVO:01000298), continental rise (ENVO:00000274), riffle (ENVO:00000148), saline evaporation pond (ENVO:00000055), tidal pool (ENVO:00000317),	ENVO:00000000	http://purl.obolibrary.org/obo/ENVO_00000000
Glacial feature	A hydrographic feature characterized by the dominance of snow or ice. Including Iceberg (ENVO:00000298), brine pool (ENVO:01000060),	ENVO:00000131	http://purl.obolibrary.org/obo/ENVO_00000131

Table 4. Controlled vocabularies for Environmental material

Preferred Name	Definitions	ENVO ID	Link
Environmental material	Everything under this parent must be a mass noun. All subclasses are to be understood as being composed primarily of the named entity, rather than restricted to that entity. For example, "ENVO:water" is to be understood as "environmental material composed primarly of some CHEBI:water". This class is currently being aligned to the Basic Formal Ontology. Following this alignment, its definition and the definitions of its subclasses will be revised. A portion of environmental material is a fiat object, which forms the medium or part of the medium of an environmental system.	ENVO:00010483	http://purl.obolibrary.org/obo/ENVO_00010483
Clay	A group of hydrous aluminium phyllosilicate (phyllosilicates being a subgroup of silicate minerals) minerals (see clay minerals), that are typically less than 2micrometres in diameter. Clay consists of a variety of phyllosilicate minerals rich in silicon and aluminium oxides and hydroxides, which include variable amounts of structural water.	ENVO:00002982	http://purl.obolibrary.org/obo/ENVO_00002982
Marl	Marl is a mass of calcium carbonate derived from mollusk shells and mixed with silt and clay.	ENVO:01000853	http://purl.obolibrary.org/obo/ENVO_01000853
Mud	A liquid or semi-liquid mixture of water and some combination of soil, silt, and clay. Includes: anaerobic mud, arsenic-rich mud, estuarine mud, marine mud	ENVO:01000001	http://purl.obolibrary.org/obo/ENVO_01000001
Particulate matter	Particulate material is an environmental material, which is composed of microscopic portions of solid or liquid material suspended in another environmental material.	ENVO:01000060	http://purl.obolibrary.org/obo/ENVO_01000060
Sand	A naturally occurring granular material composed of finely divided rock and mineral particles. Includes acid dune sand, beach sand, rocky sand and sea sand	ENVO:01000017	http://purl.obolibrary.org/obo/ENVO_01000017
Sediment	Sediment is an environmental substance comprised of any particulate matter that can be transported by fluid flow and which eventually is deposited as a layer of solid particles on the bed or bottom of a body of water or other liquid. Includes anaerobic sediment, biogenous sediment, boulder sediment, clay sediment, cobble sediment, colloidal sediment, contaminated sediment, granular sediment, hydrogenous sediment, hyperthermophilic sediment,	ENVO:00002007	http://purl.obolibrary.org/obo/ENVO_00002007

	intertidal sediment, marine sediment, mesophilic sediment, pebble sediment, sandy sediment, silty sediment, stream sediment, suspended sediment, terrigenous sediment, thermophilic sediment.		
Solid environmental material	An environmental material, which is in a solid state. This is a defined class: its subclasses will not be asserted, but filled by inference. Includes mineral material, rock, water ice and wood.	ENVO:01000814	http://purl.obolibrary.org/obo/ENVO_01000814
Waste material	A material which is not the desired output of a process and which is typically the input of a process which removes it from its producer (e.g. a disposal process). Includes biological waste material, industrial waste material and waste water.	ENVO:00002264	http://purl.obolibrary.org/obo/ENVO_00002264
Water	An environmental material primarily composed of dihydrogen oxide in its liquid form. Includes: contaminated water (ENVO:00002186), eutrophic water (ENVO:00002224), hydrothermal fluid (ENVO:01000134), muddy water (ENVO:00005793), oil field production water (ENVO:00002194), saline water or salt water (ENVO:00002010), sea water or ocean water (ENVO:00002149), brackish water (ENVO:00002019), estuarine water (ENVO:01000301), hypersaline water (ENVO:00002012), brine (ENVO:00003044), coastal sea water (ENVO:00002150), surface water (ENVO:00002042)	ENVO:00002006	http://purl.obolibrary.org/obo/ENVO_00002006