New data resources for marine metagenomics

Nils Peder Willassen

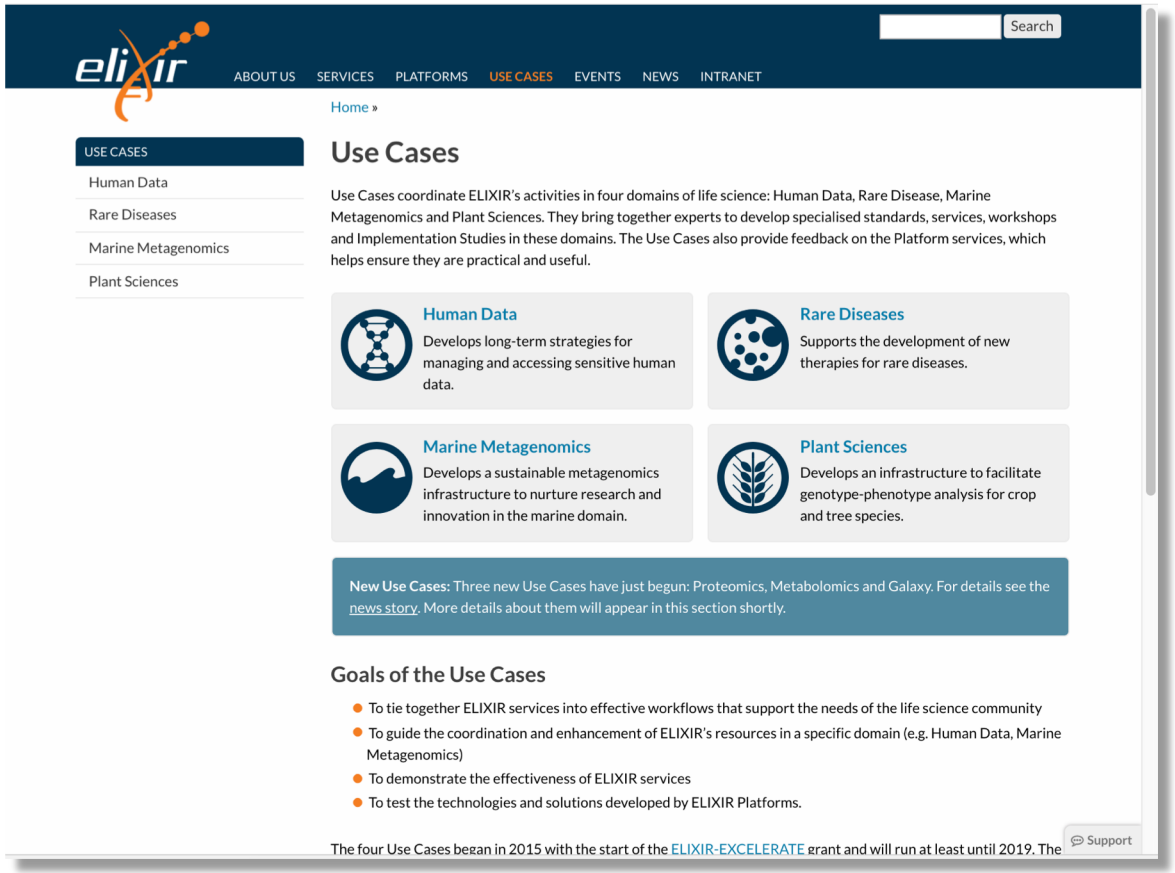
Centre for bioinformatics

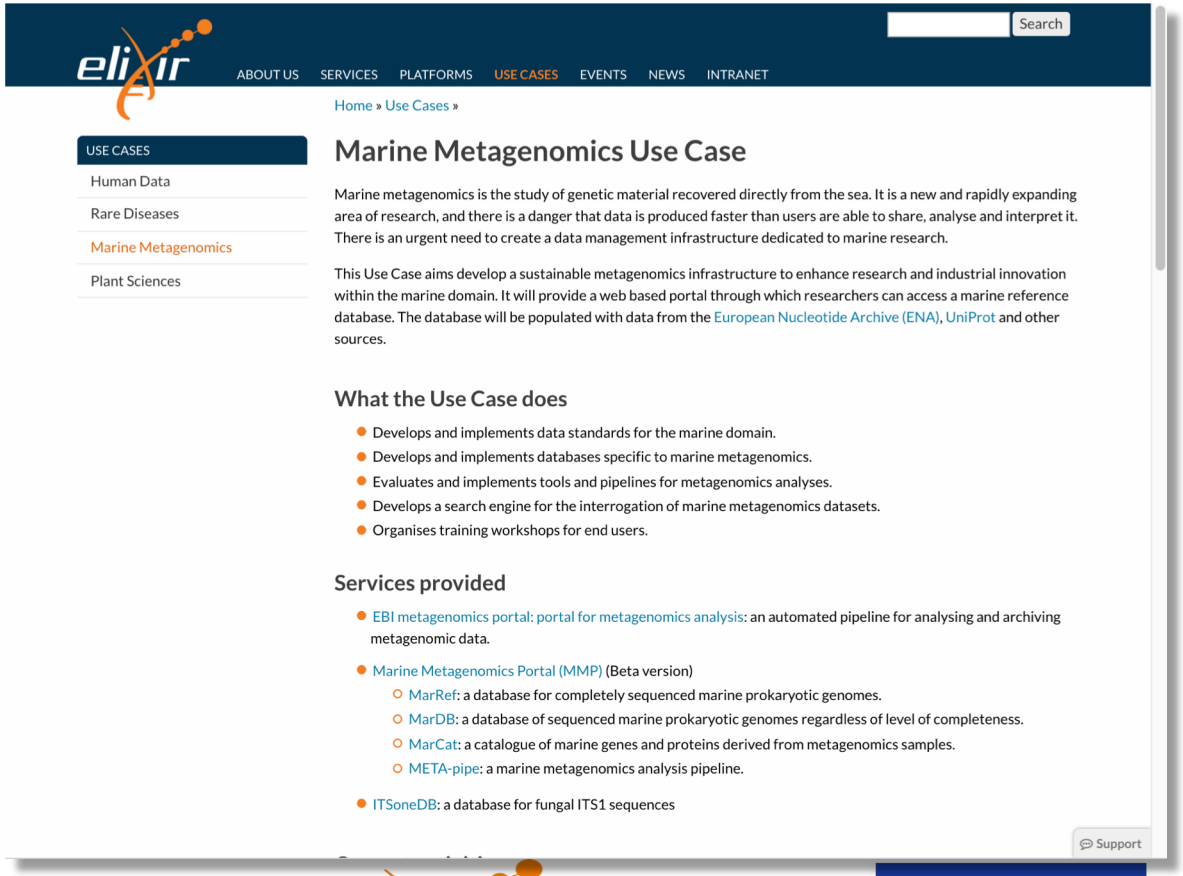
UiT The Arctic University of Norway

ELIXIR Norway

ELIXIR

Marine metagenomics Use case (Marine metagenomics Community from 2019)





ELIXIR EXCELERATE

Background

Community need for specific marine metagenomics resources

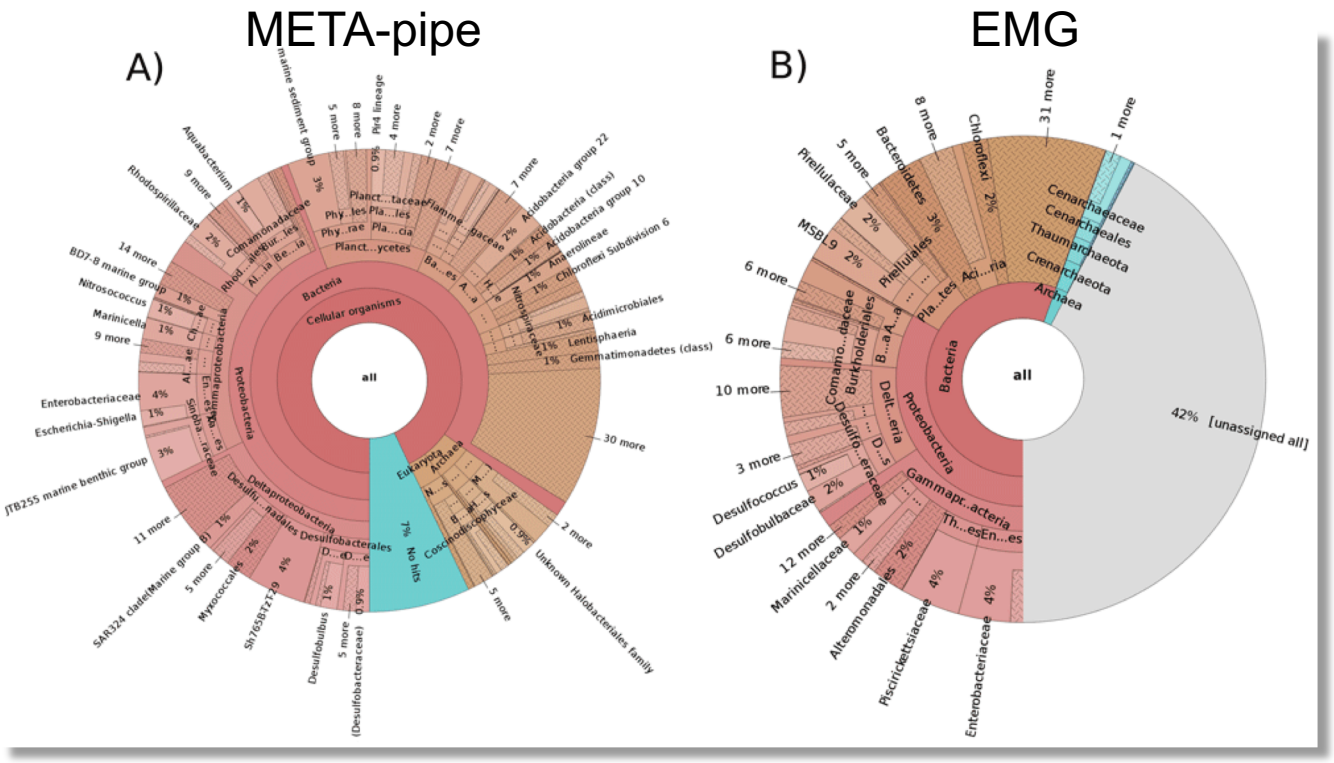
- *need to develop standards and best practice within the field*
- *need for reference databases specific for the marine environment - should include sequence and metadata (contextual)*
- *benchmark analysis tools and pipelines to optimize marine metagenomics analysis*
- *resources should follow the FAIR (Findable, Accessible, Interoperable and Reusable) principles*

ELIXIR EXCELERATE

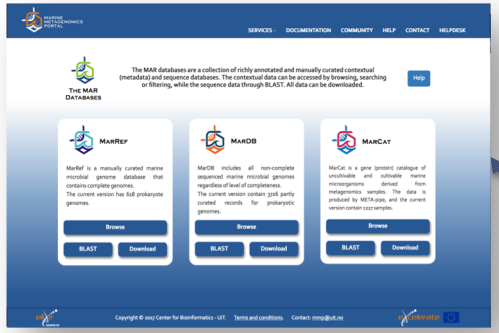
Background

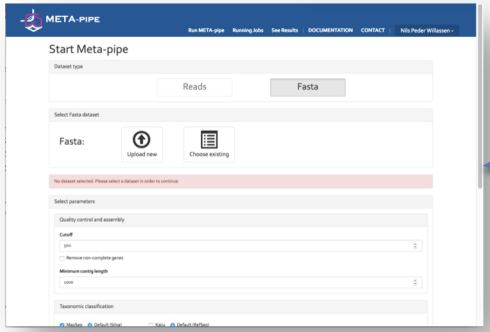
Why more (and domain-specific) resources?

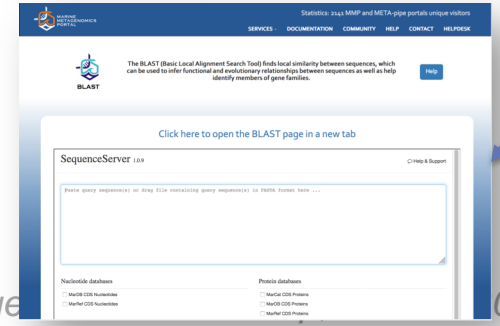
- ***need to benchmark and implement tools/pipelines for best possible output results***
- ***tools not better than the underlaying databases used***
- ***the higher quality of the databases - the better classification and annotation of metagenomic samples***

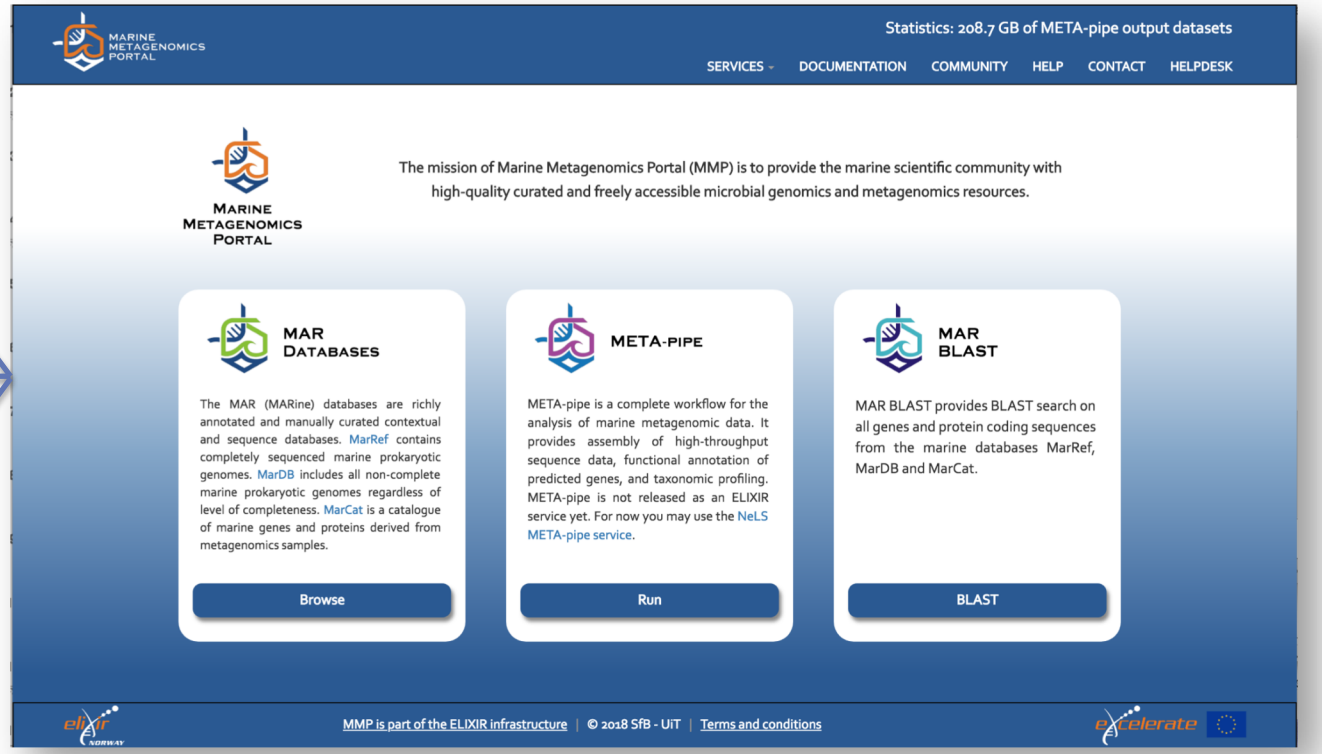


Marine Metagenomics Portal









What is a marine microbial biome?

Marine biome

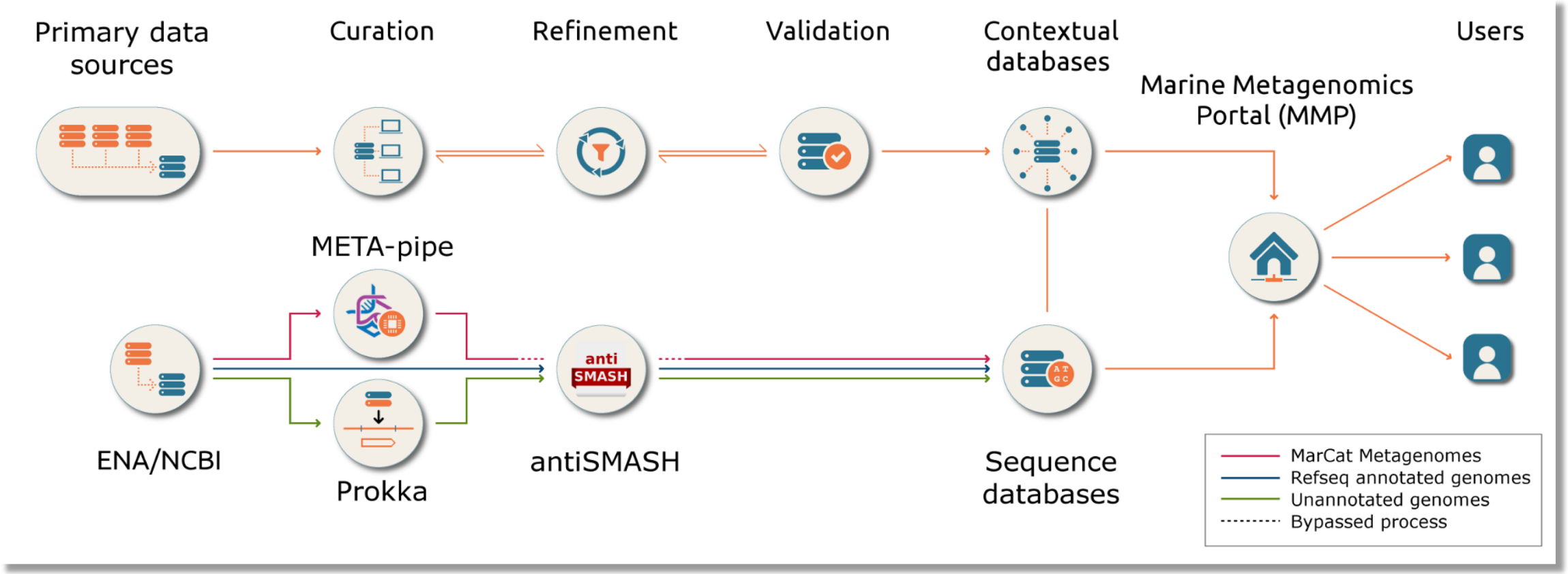
"An aquatic biome that comprises systems of open-ocean and unprotected coastal habitats, characterized by exposure to wave action, tidal fluctuation, and ocean currents as well as systems that largely resemble these. Water in the marine biome is generally within the salinity range of seawater: 30 to 38 ppt "

Marine microbial biome

"An aquatic microbial biome comprises of microbial communities from open oceans, coastal and protected habitats up to the high-water mark with salinity from 0.5 ppt (parts per thousand) as in 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"

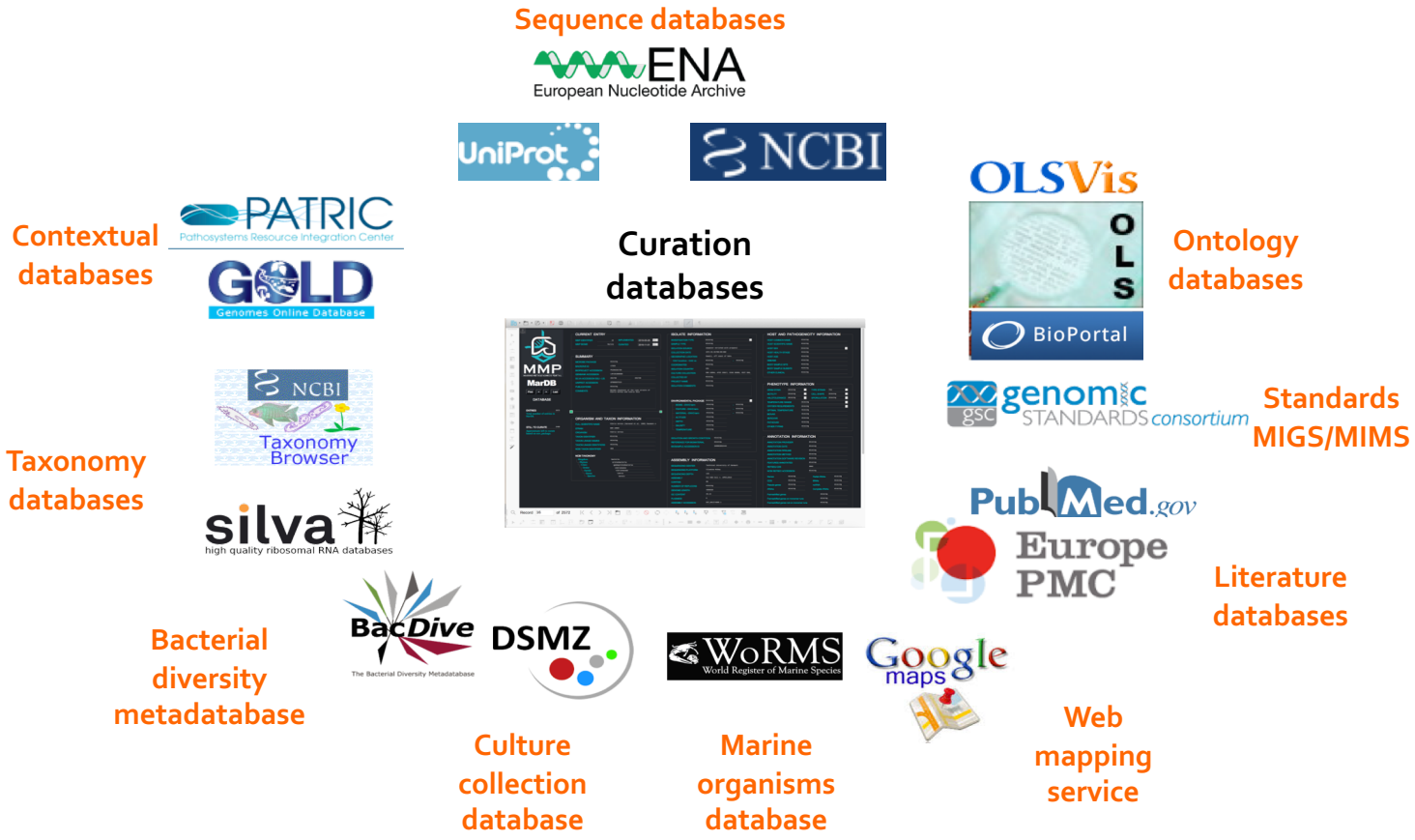
MAR databases

Implementation



MAR databases

Data resources used



MAR databases

Curation database

MAR databases

Attribute descriptions, CV and Ontologies

Structured comment name	Item	Description	Examples	Expected value	Value syntax	Preferred 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 Ex 7: 2015 Ex 4: 2003--2006 Ex 5: 2010-01--2011-03 Ex 6: 2011-05-28--2011-08-10	date and time, range	{timestamp}	-
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: harbour, 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	MIGS/MIMS/MIMARK extension for reporting of	Ex 1: Water	CV, single data entry	[Air|Host-	-

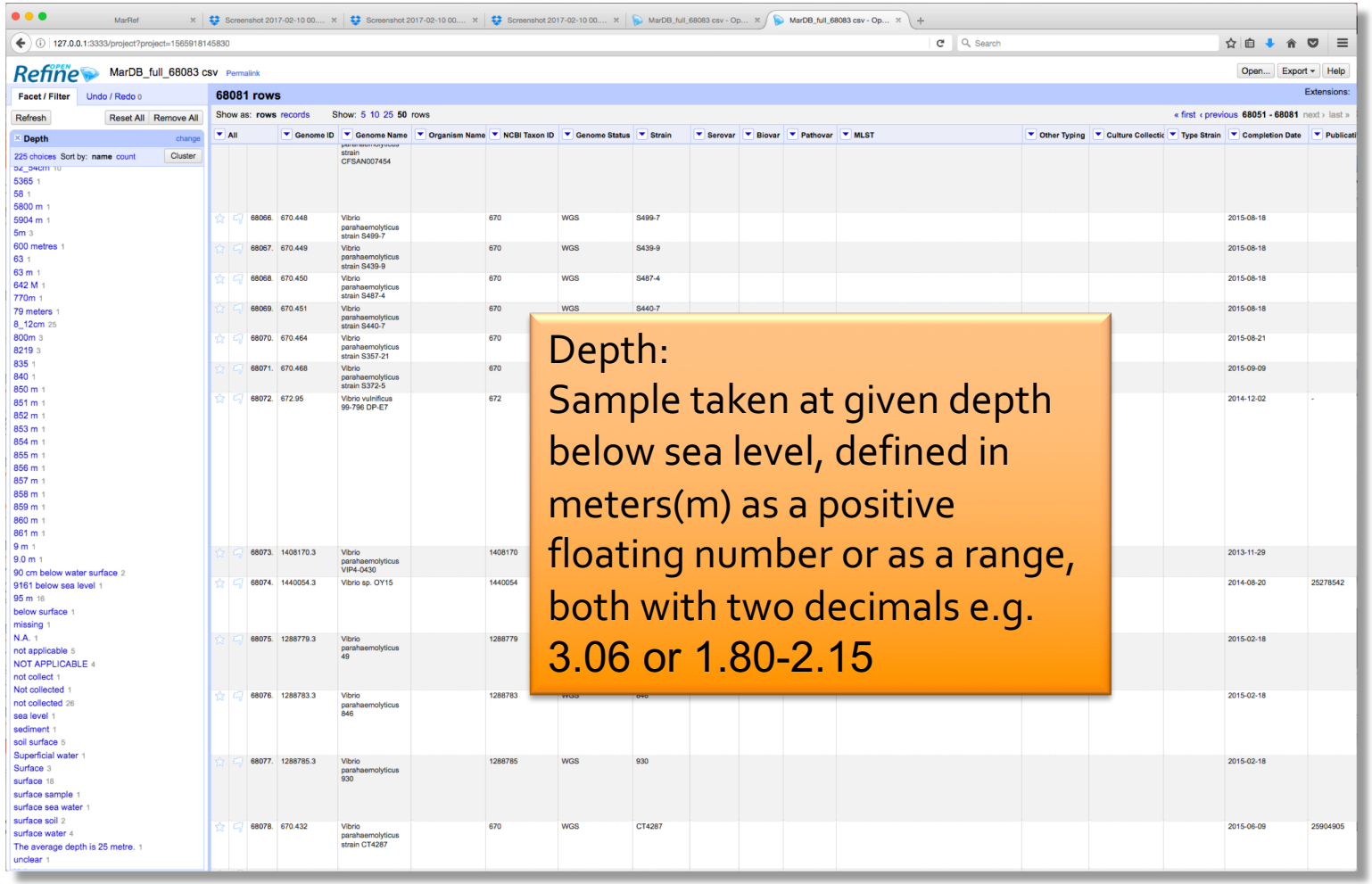
Preferred Name	Definitions	ENVO ID	Link
Marine biome	An aquatic biome that comprises systems of open-ocean and unprotected coastal habitats, characterized by exposure to wave action, tidal fluctuation, and ocean currents as well as systems that largely resemble these. Water in the marine biome is generally 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 intertidal 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 meters (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

*A **controlled vocabulary** is a set of preselected terms e.g. taxonomic domain; bacteria, archaea, eukarya*

***Ontology** is a formal naming and definition of the types, properties, and interrelationships.*

MAR databases

Refinement

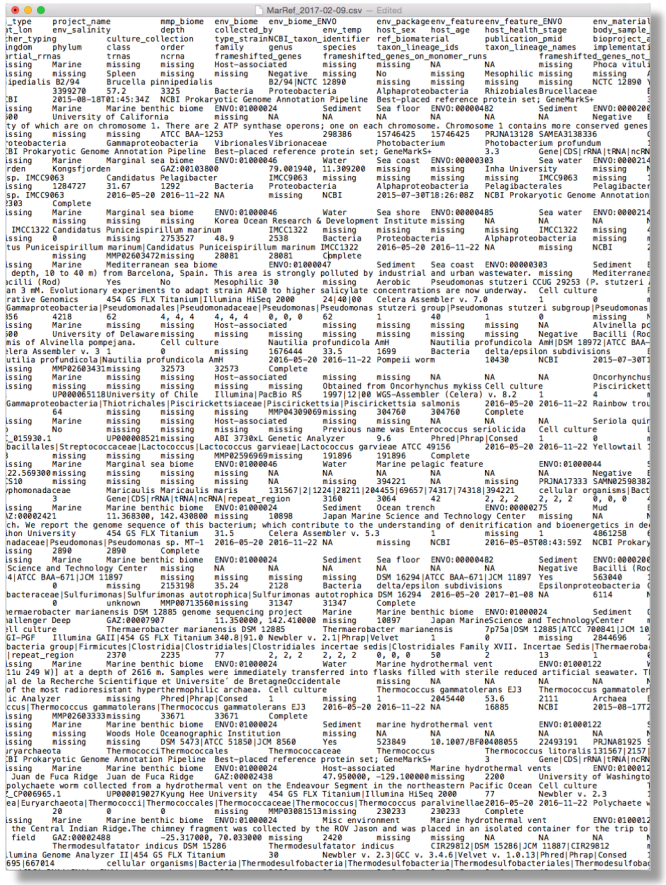


Not collected	->	missing
250 M	->	250
Not applicable	->	NA
Superficial	->	missing
-1 m	->	1
-2 m	->	2
-2901.0	->	2901
0 m.	->	0
1912 ft	->	582.80
40 mm from surface	->	0.04
0.75 m above seafloor	->	missing
700meters	->	700
Intracellular	->	missing
Surface water of 0 meter	->	0
Zero	->	0
Below surface	->	Missing

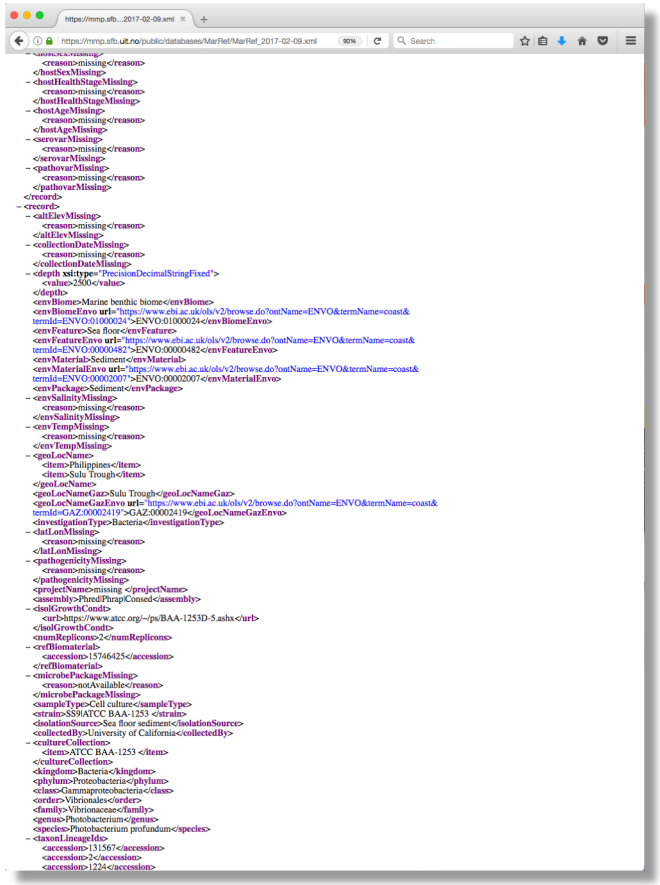
MAR databases

Conversion & validation

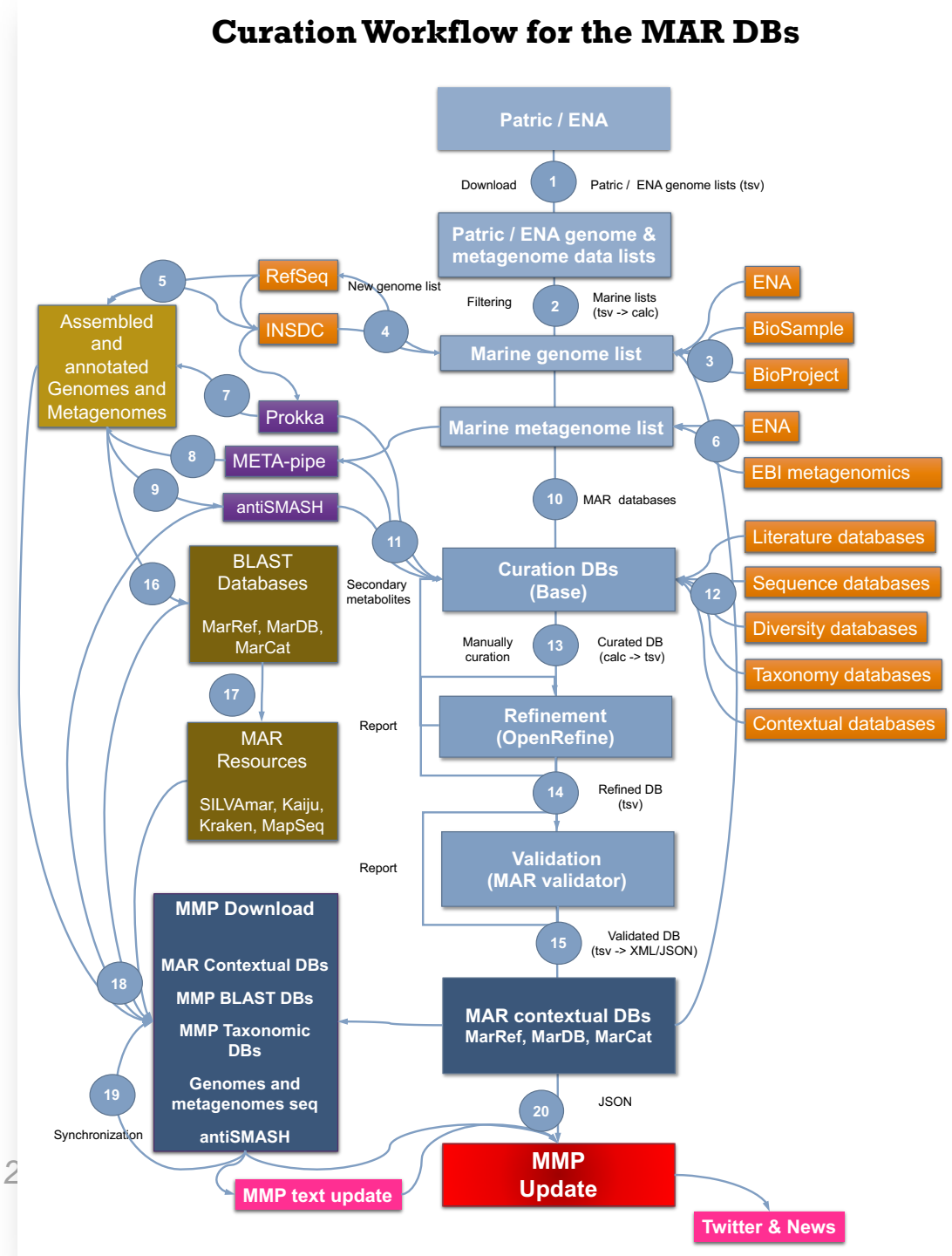
Huma-readable (TSV)



Machine-readable (XML/JSON)

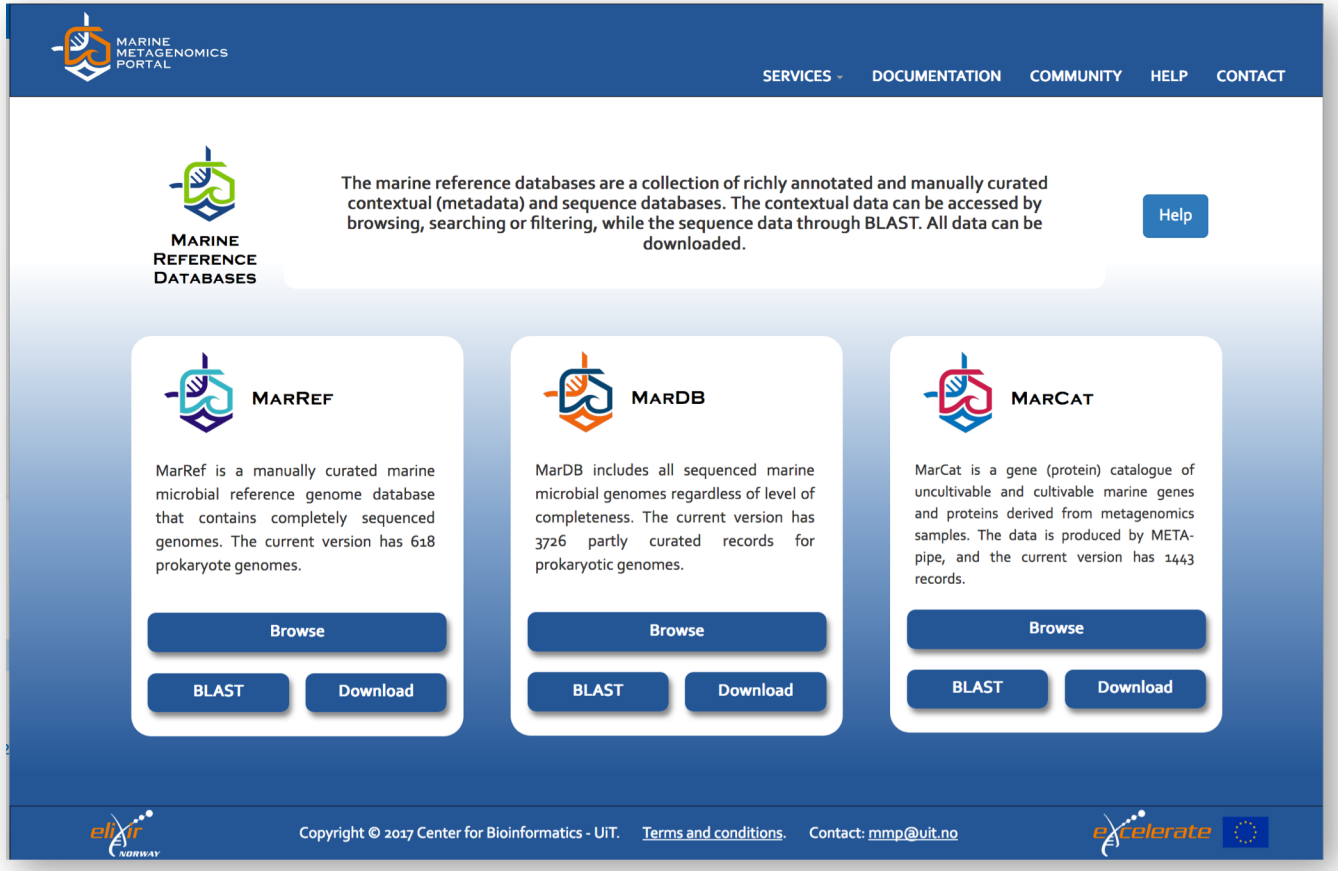


MAR databases



MAR databases

Marine reference databases

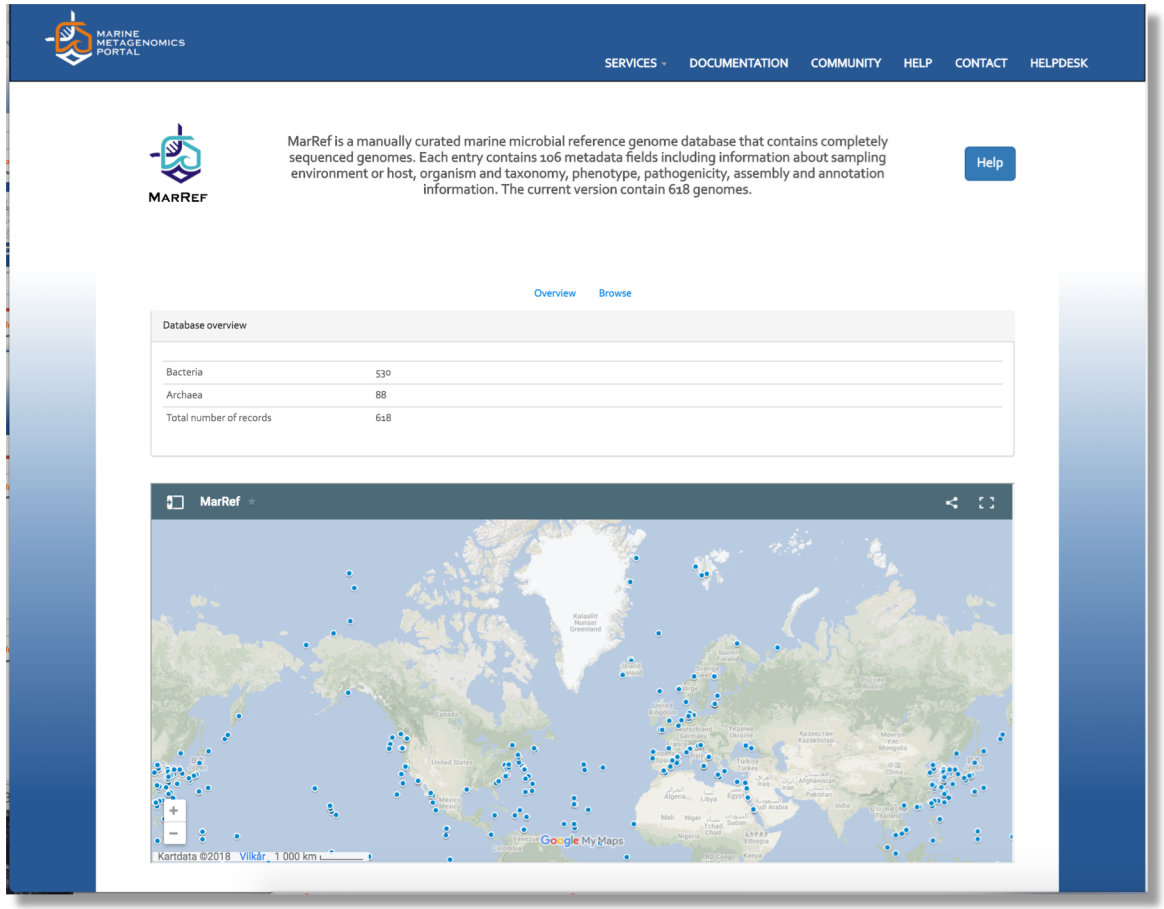


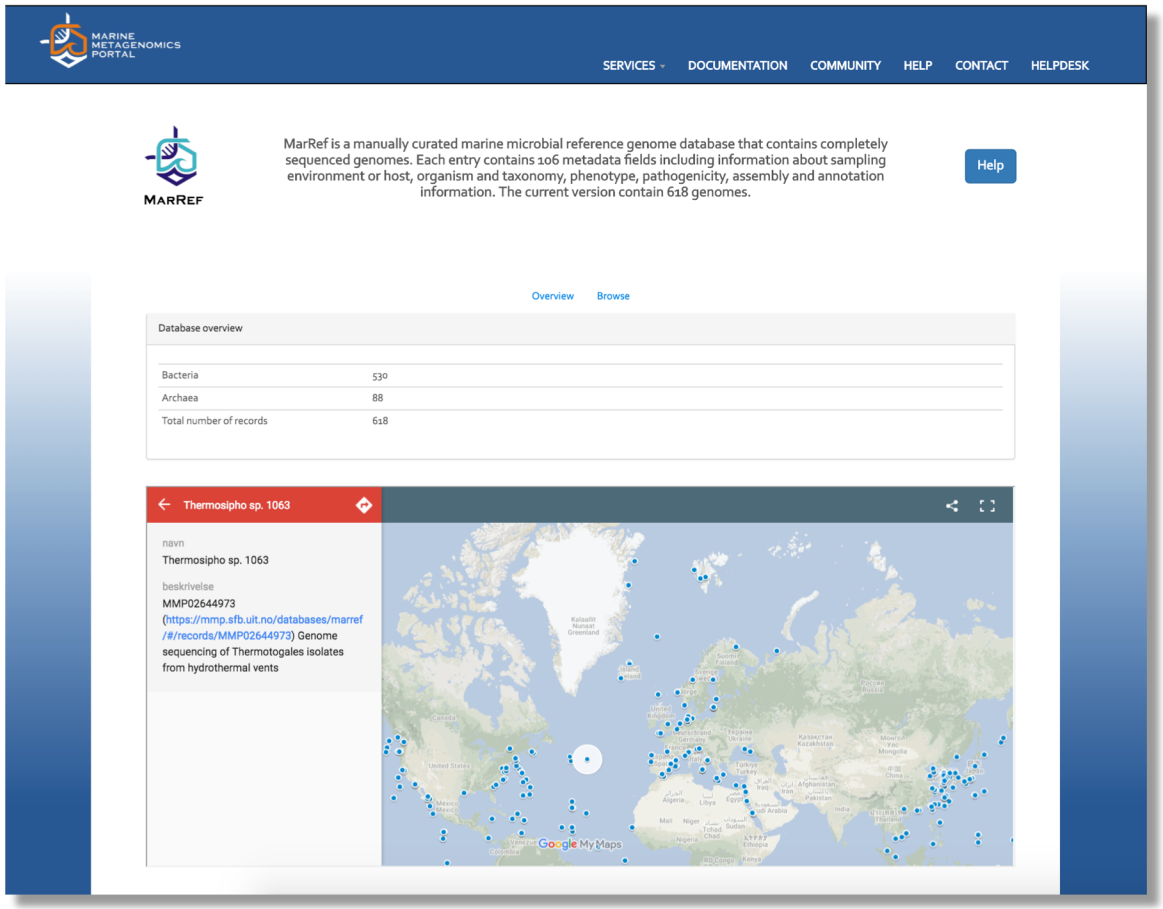
Accessible from the Marine Metagenomics Portal (MMP)
https://mmp.sfb.uit.no/

- MarRef, MarDB and MarCat
- Currently 771, 8649 and 1227 entries
- Contextual and sequence databases
- 122/55 metadata fields
- Manually expert curated
- Ontologies and controlled vocabularies implemented
- Biannually update
- Advanced search and filtering
- BLAST
- Open access and downloadable

MAR databases

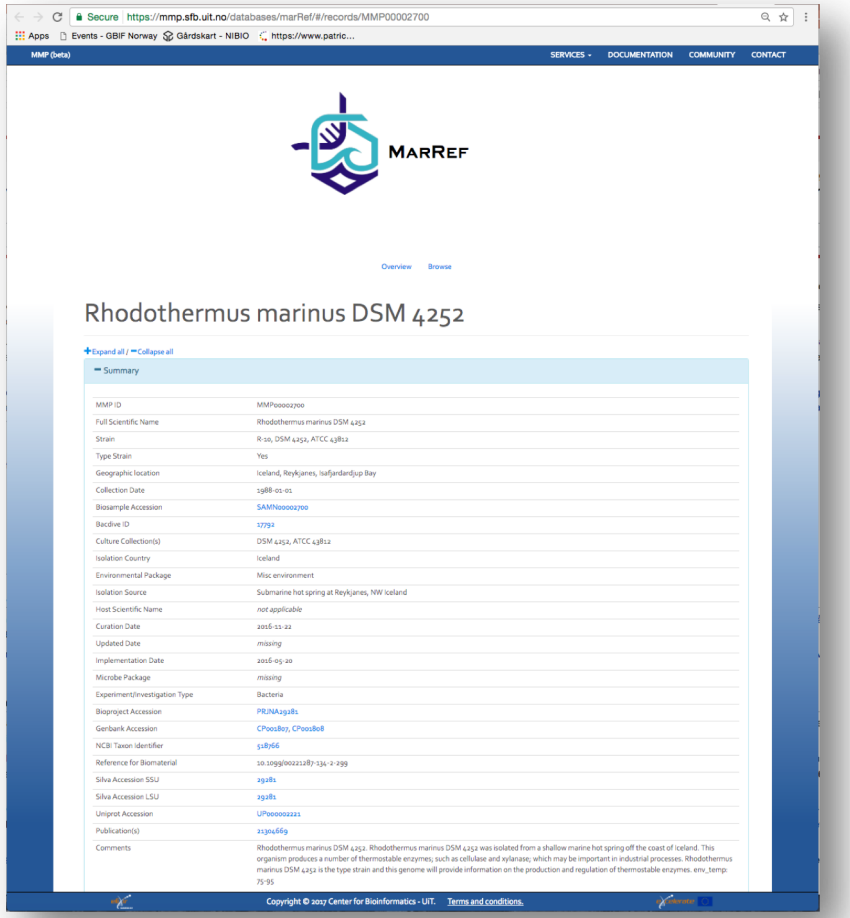
Browse and filtering

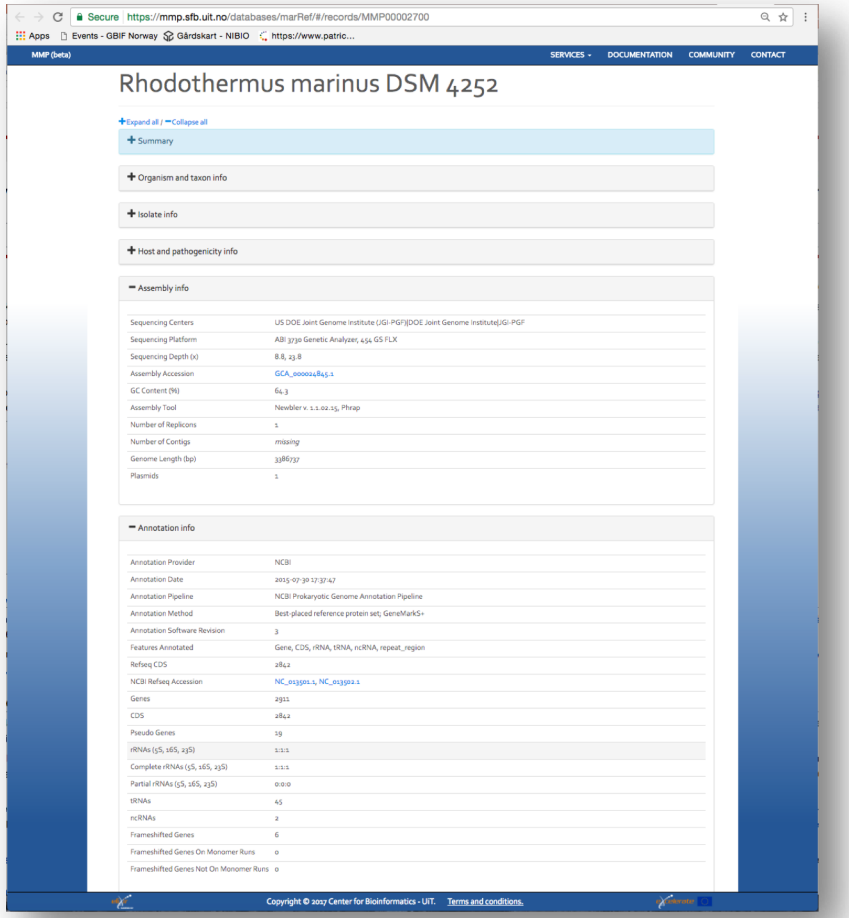




MAR databases

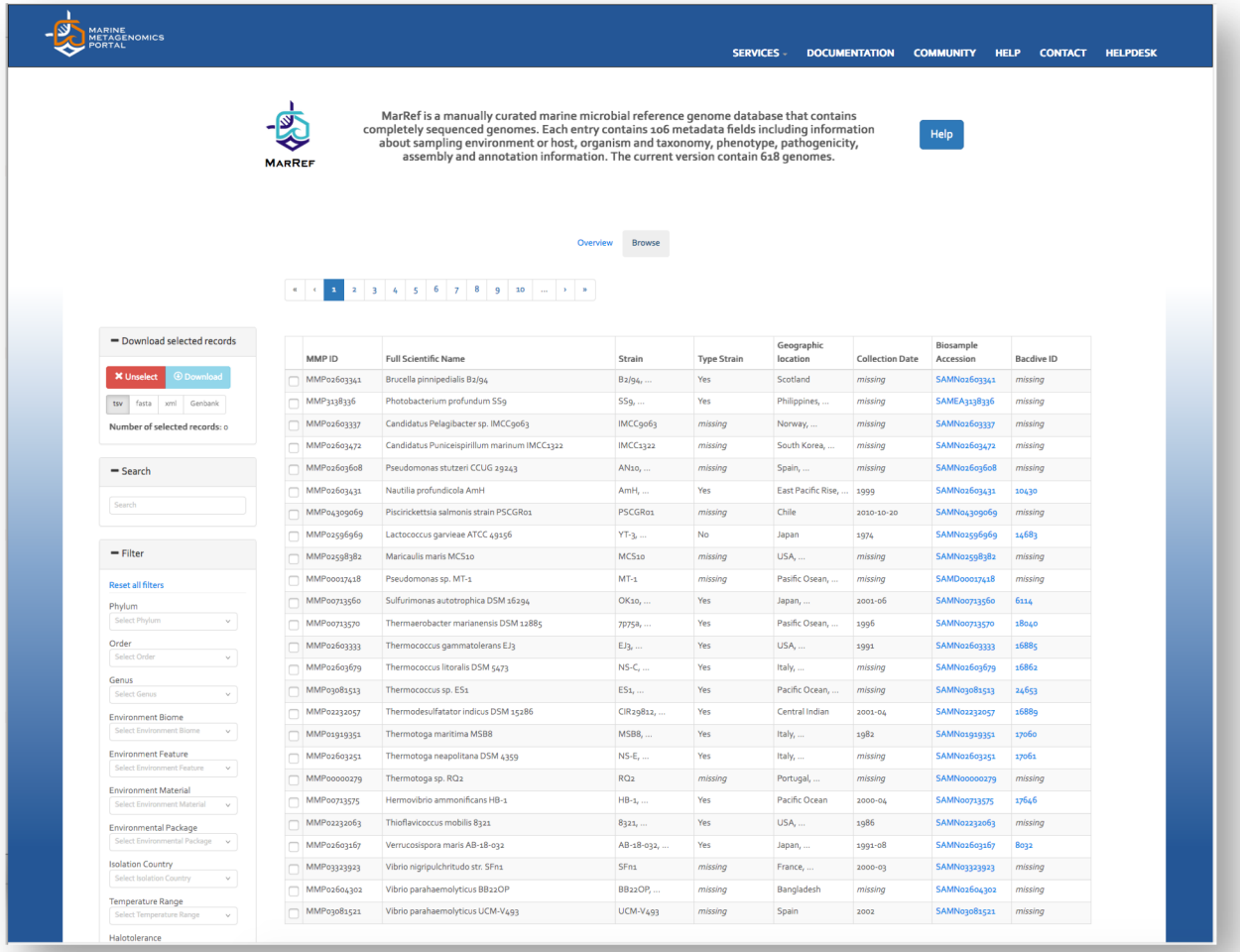
Browse and filtering

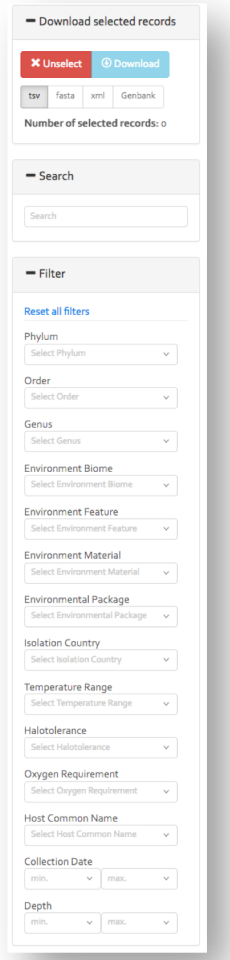




MAR databases

Browse and filtering





MAR databases

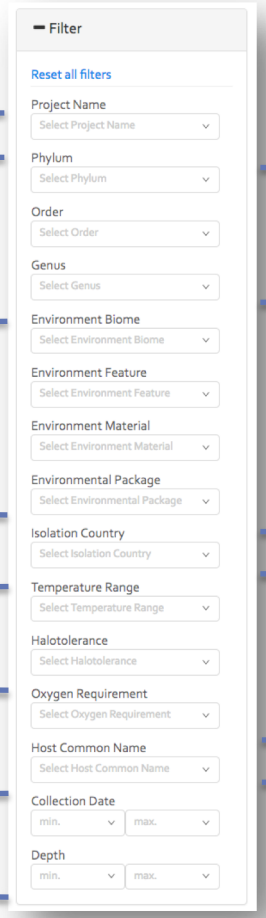
Browse and filtering

Filtering based on project name

Filtering based on sampling environment; biome, feature & material

Filtering based on phenotypic trait

Filtering based on collection date/depth



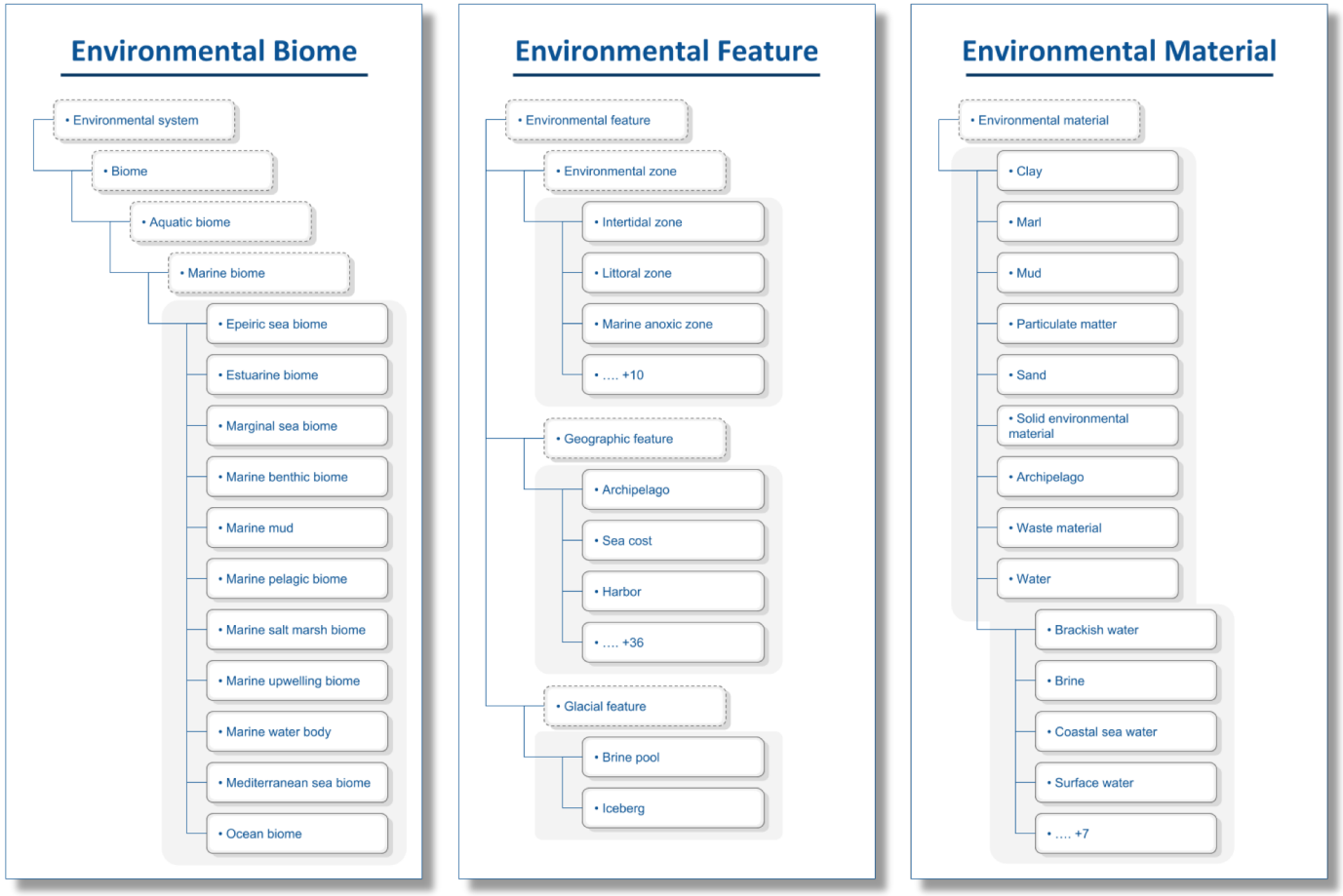
Taxonomic filtering based on 3 levels; Phylum, Order & Genus

Filtering based on region or country

Filtering based on host (pathogen or host associate)

MAR databases

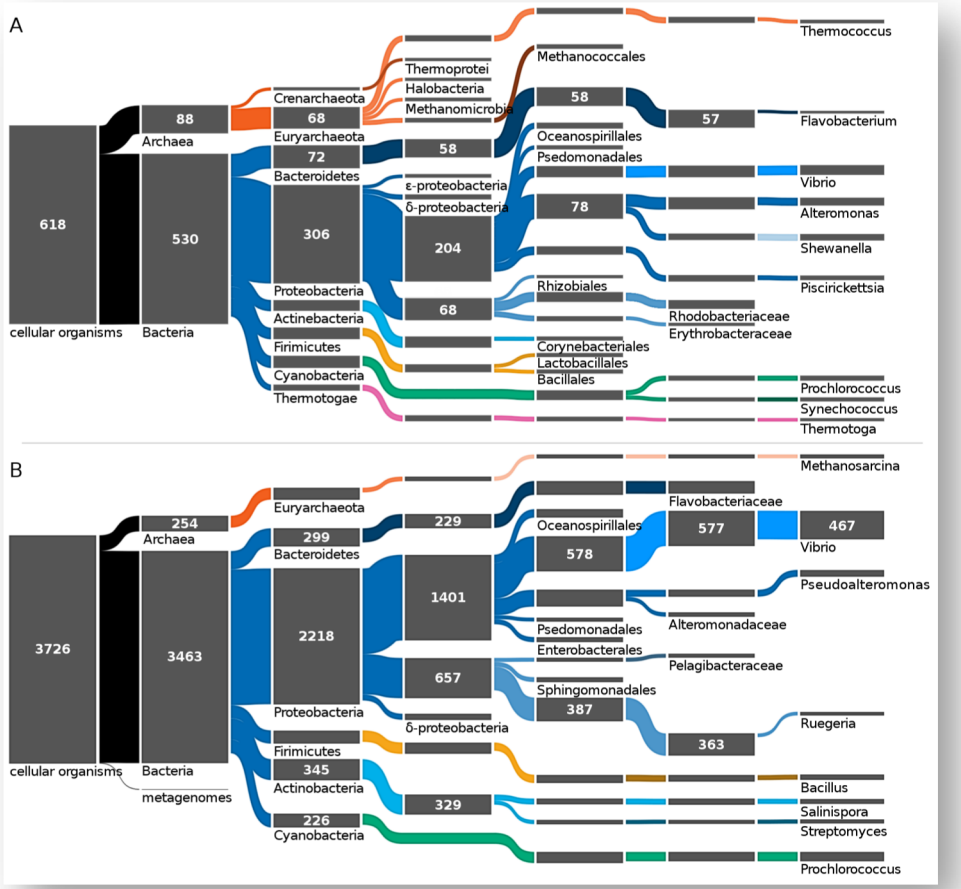
ENVO ontologies

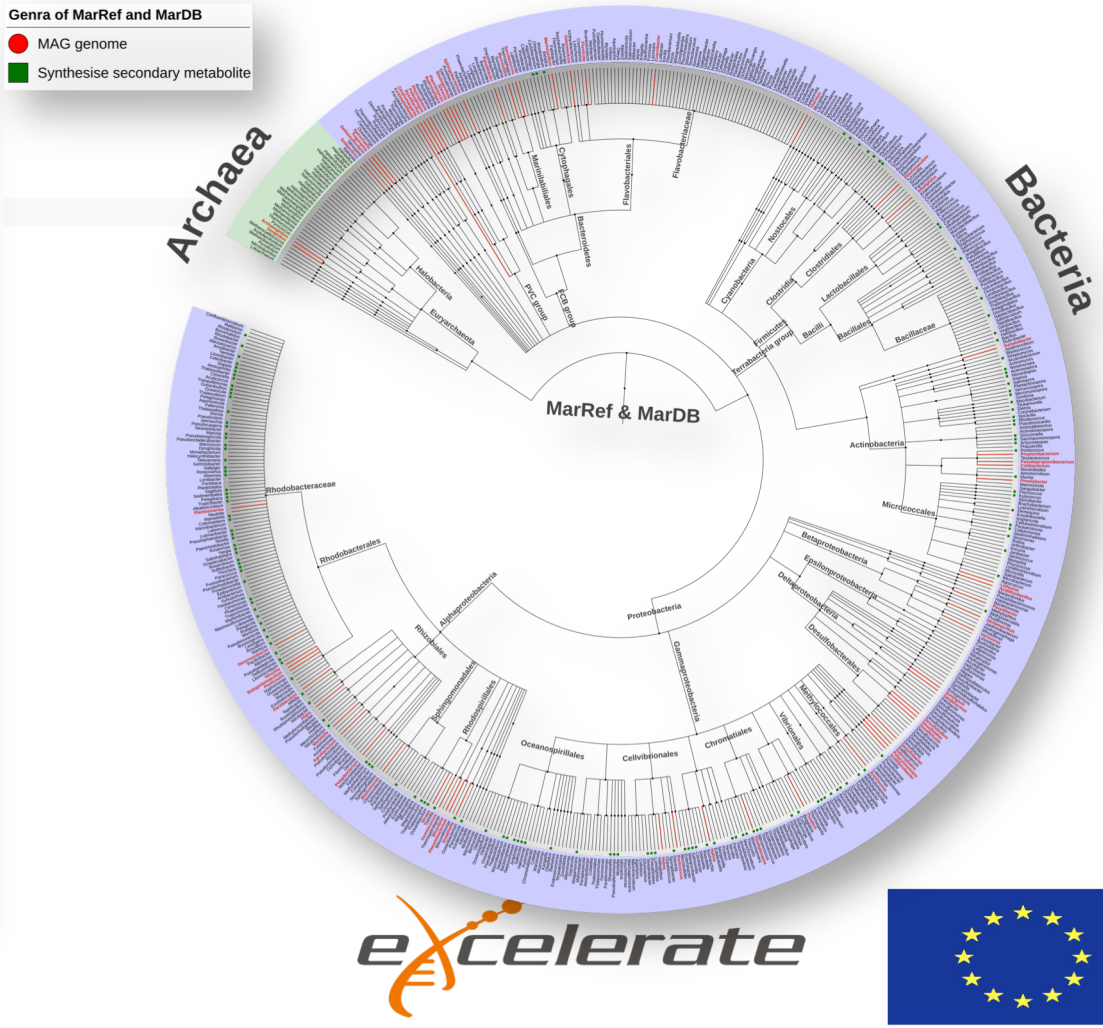


MAR databases

Taxonomic distribution

V1 - MarRef: 618, MarDB: 3726 V2 - MarRef:771, MarDB: 8649 V3 - MarRef:852, MarDB: 10896

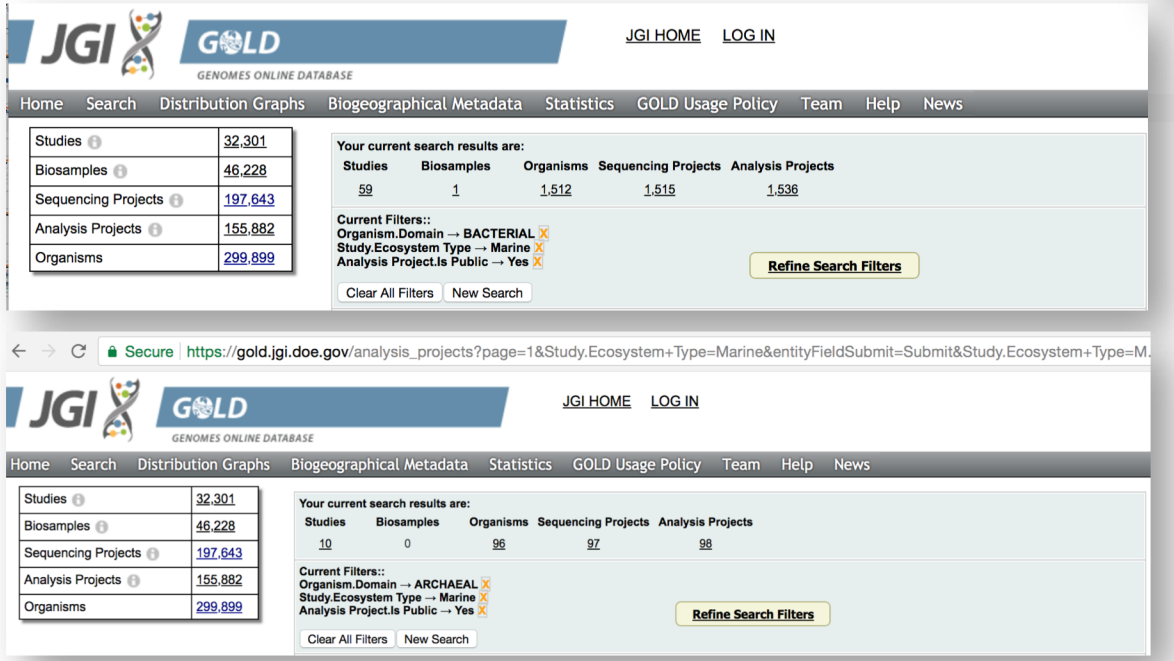




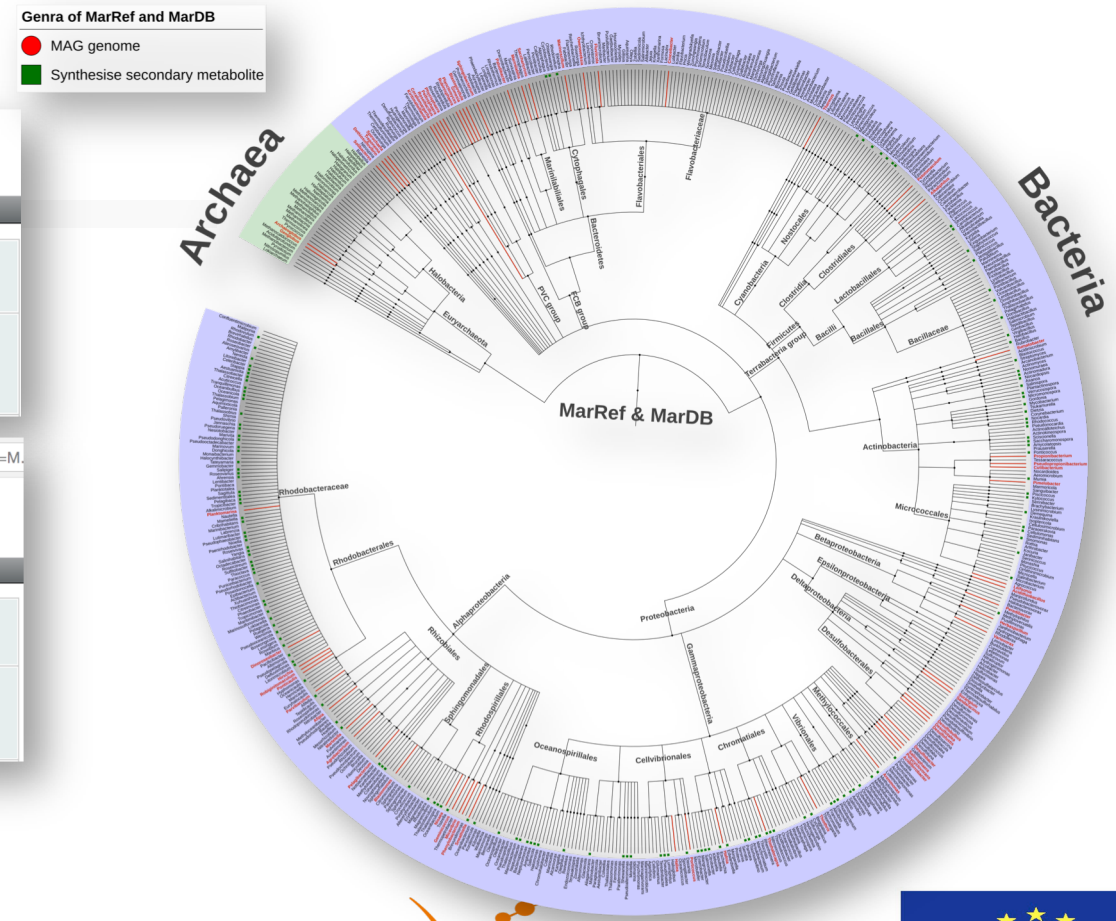
MAR databases

Taxonomic distribution

JGI Gold database: 1608

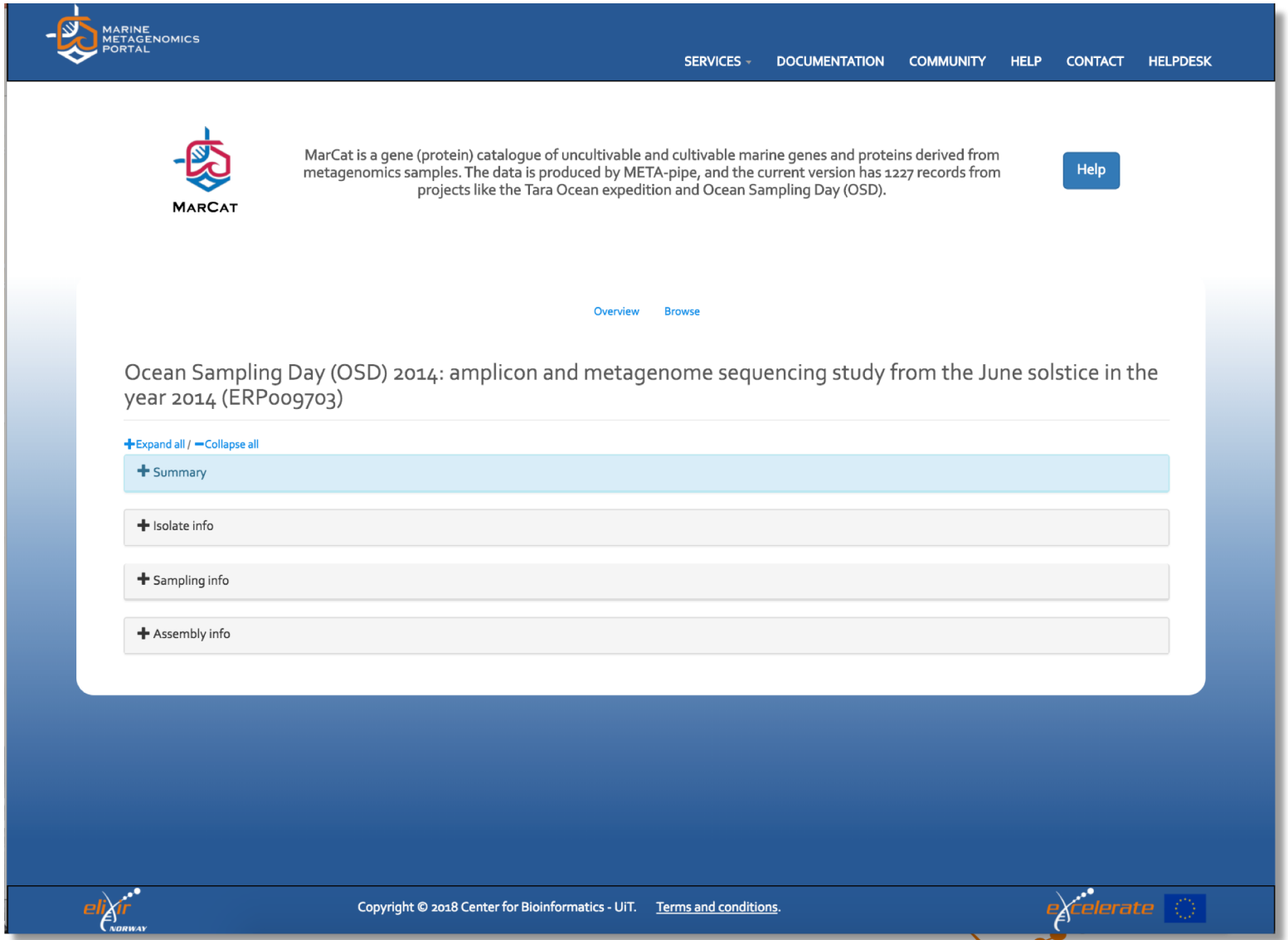


Version 2 - 9457



excelerate

MarCat

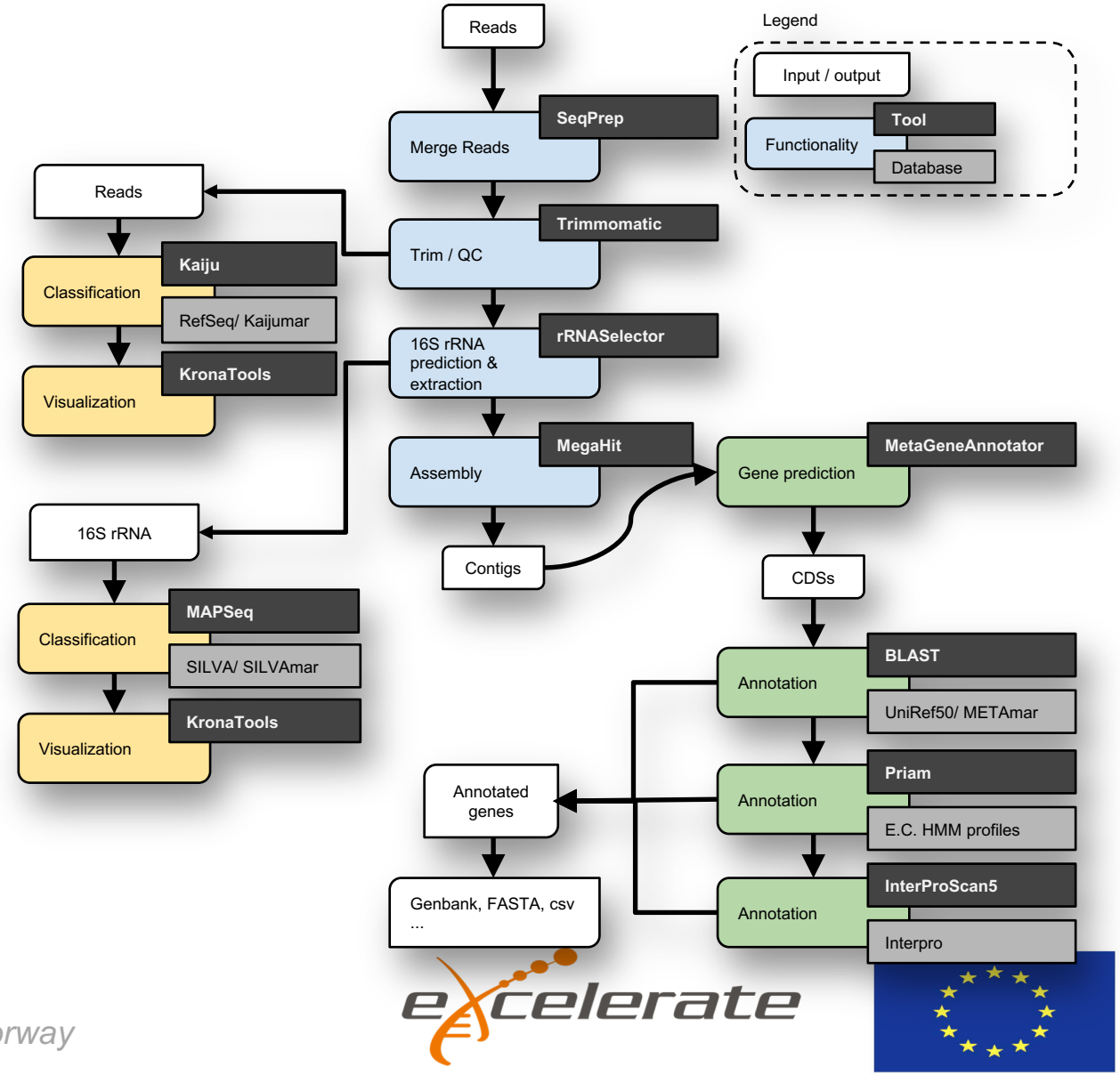

MARINE METAGENOMICS PORTAL
SERVICES
DOCUMENTATION
COMMUNITY
HELP
CONTACT
HELPDESK
MARCAT
MarCat is a gene (protein) catalogue of uncultivable and cultivable marine genes and proteins derived from metagenomics samples. The data is produced by META-pipe, and the current version has 1227 records from projects like the Tara Ocean expedition and Ocean Sampling Day (OSD).
Help
Overview
Browse
Ocean Sampling Day (OSD) 2014: amplicon and metagenome sequencing study from the June solstice in the year 2014 (ERP009703)
+Expand all / −Collapse all
+ Summary
+ Isolate info
+ Sampling info
+ Assembly info
Copyright © 2018 Center for Bioinformatics - UiT.
Terms and conditions.

MAR databases

Sequence databases

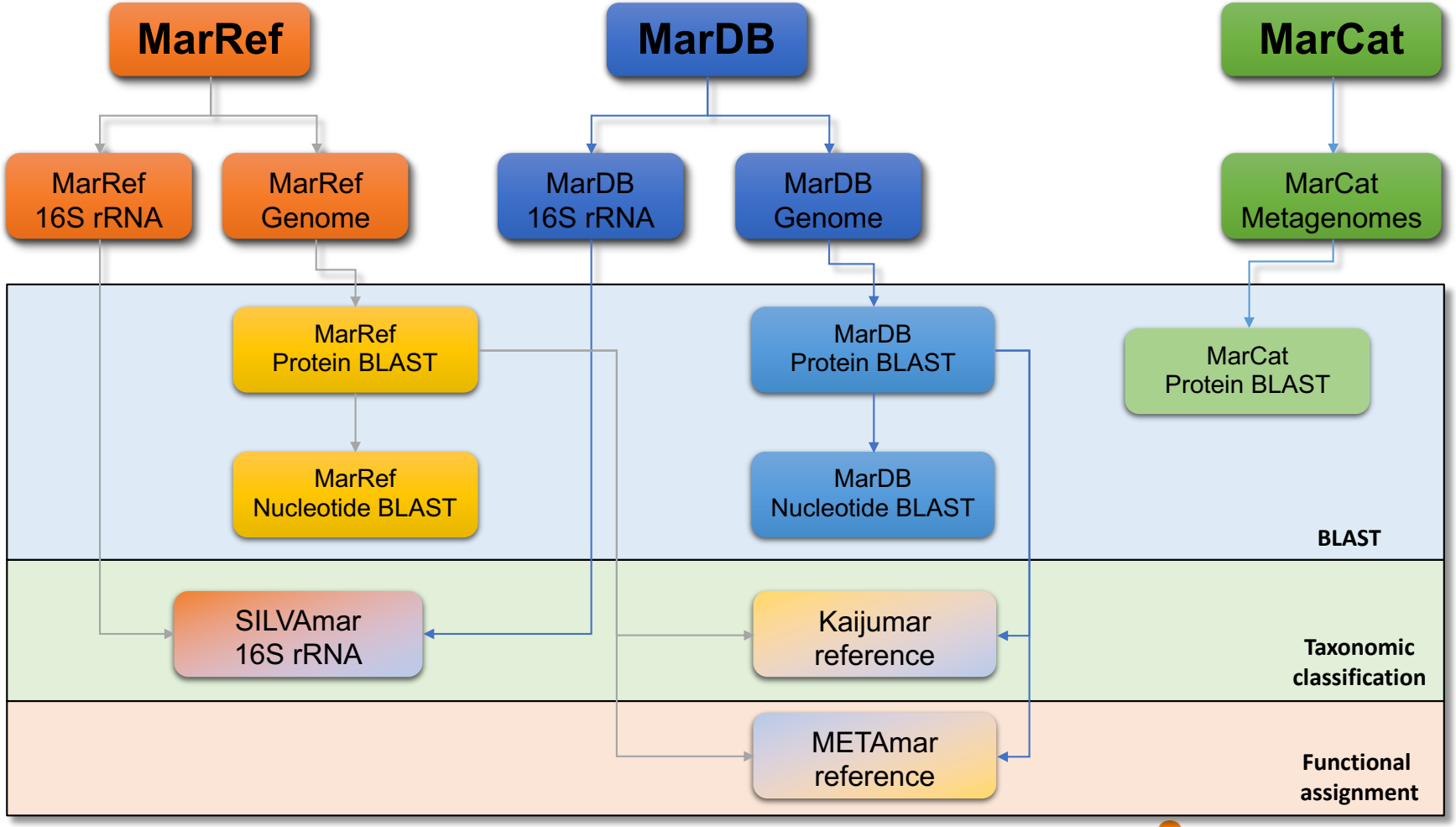
- ***MarRef and MarDB***
  - Based on RefSeq
  - Genomes annotated using NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP)
  - 20% of MarDB annotated using Prokka

- ***MarCat***
  - Metagenome samples filtered, assembled and annotated using META-pipe



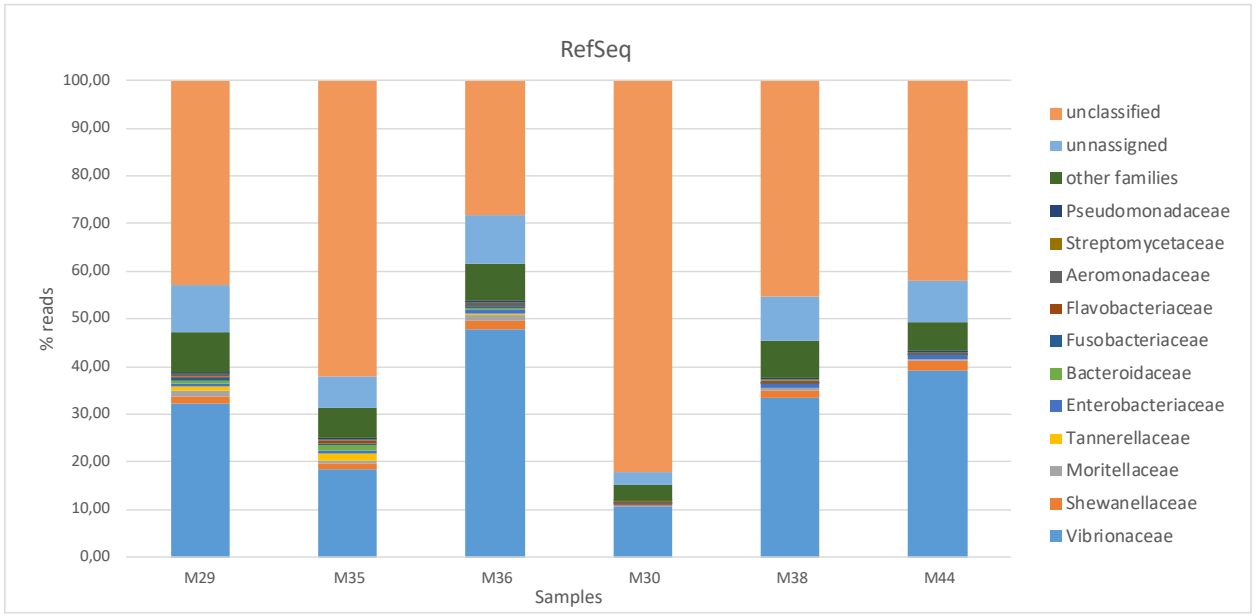
MAR databases

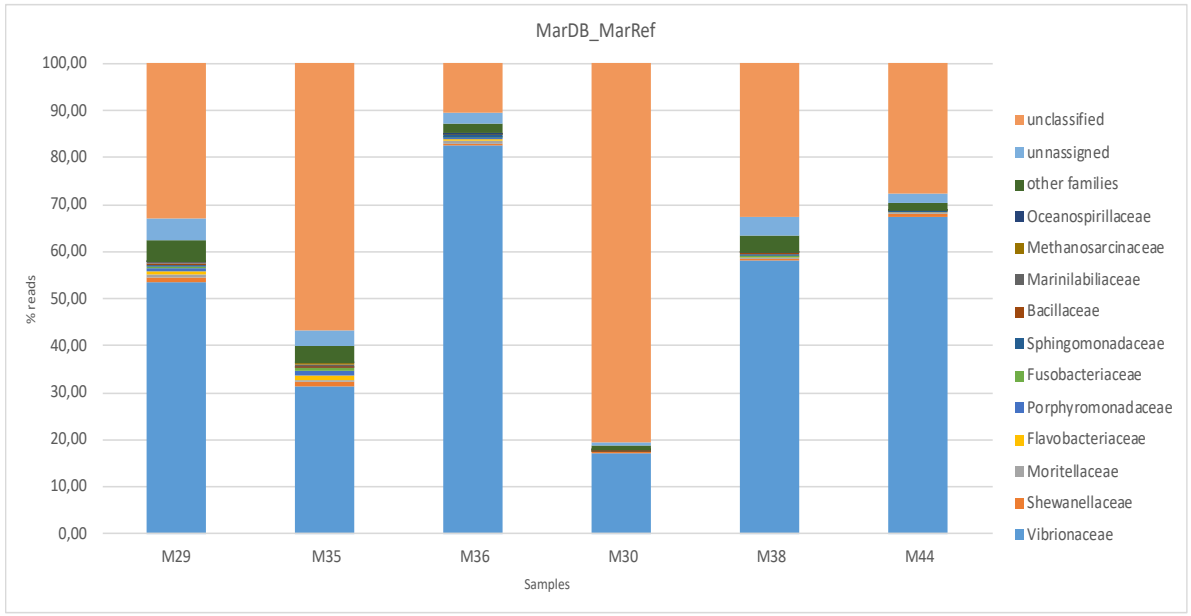
Sequence databases



MAR databases

Sequence databases

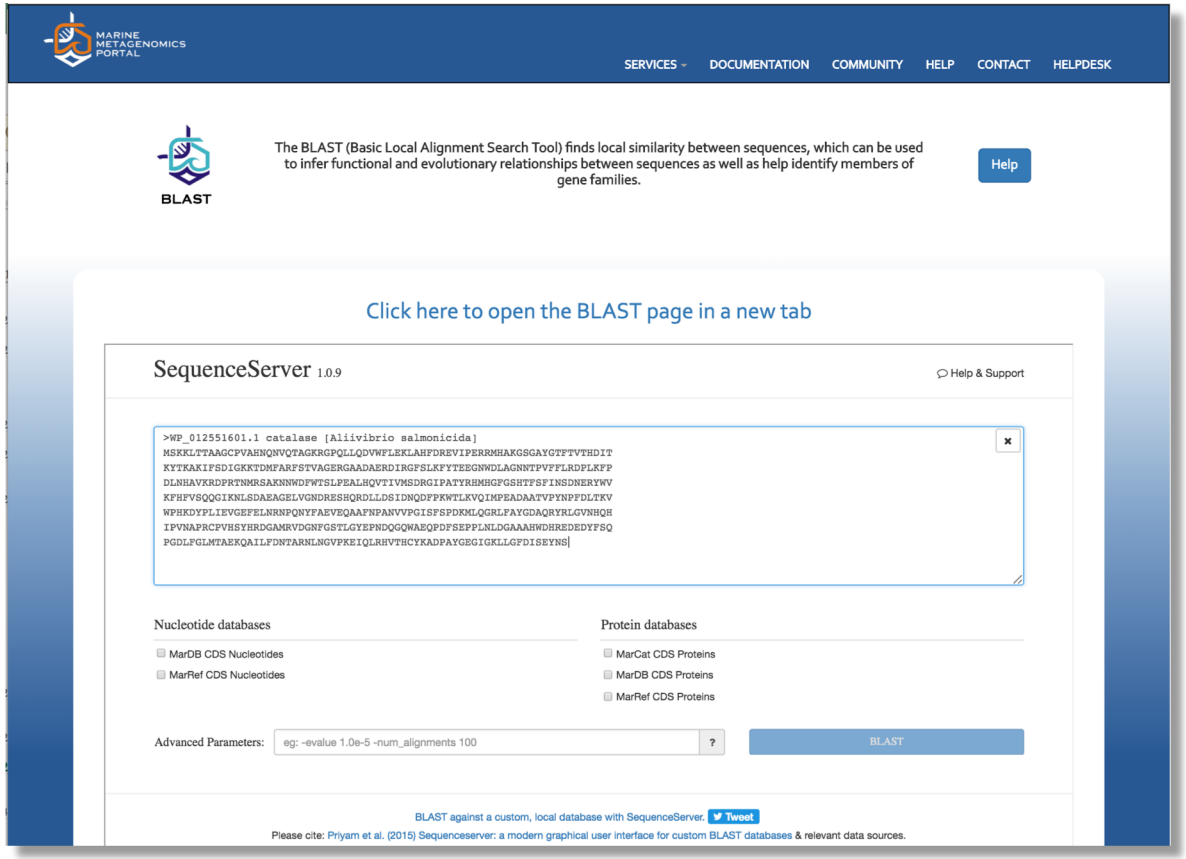


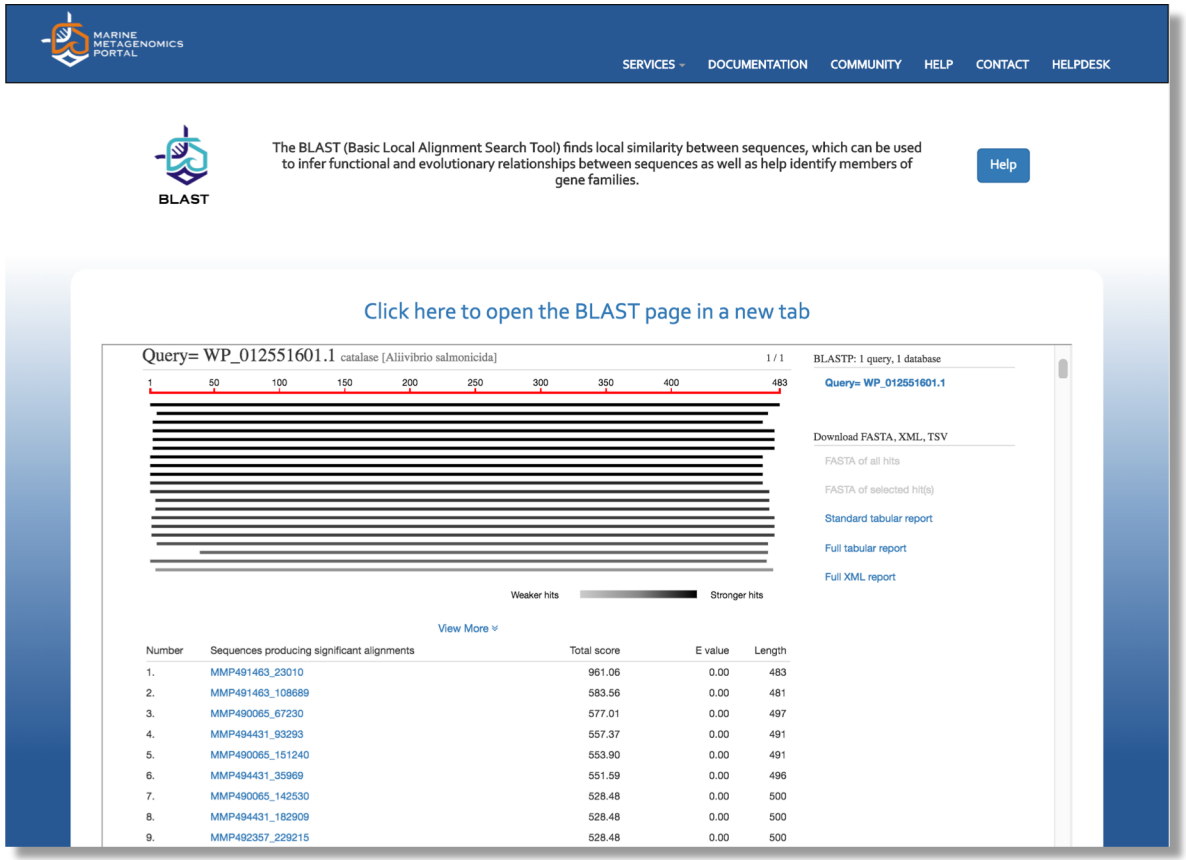


RefSeq: 52073 complete genomes
MarRef/DB: 9602 marine genomes incl. 4314 MAGs

MAR databases

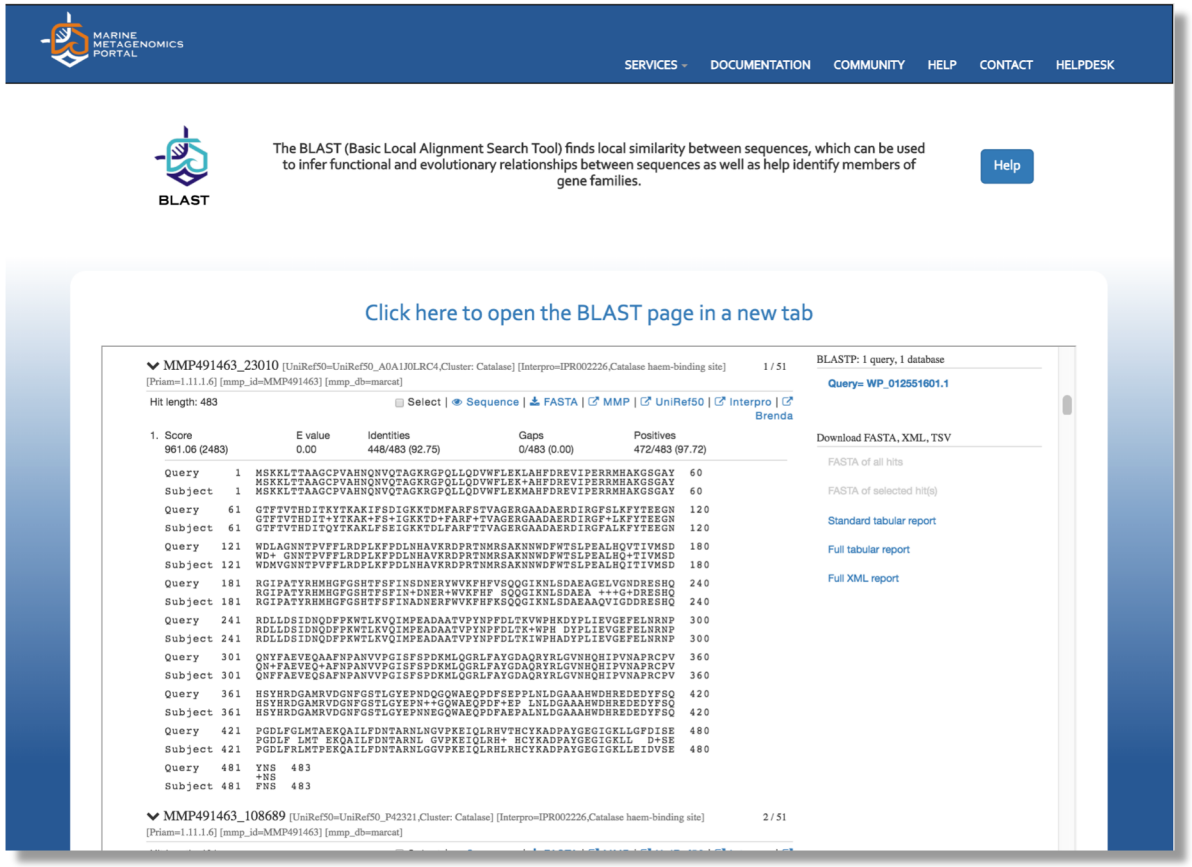
BLAST

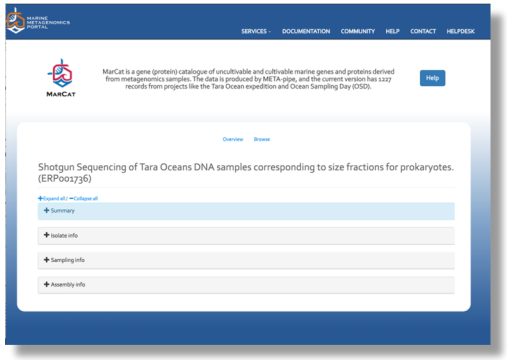


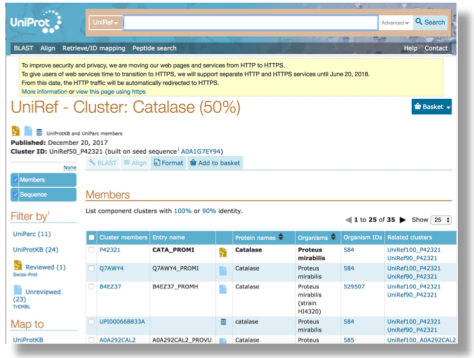


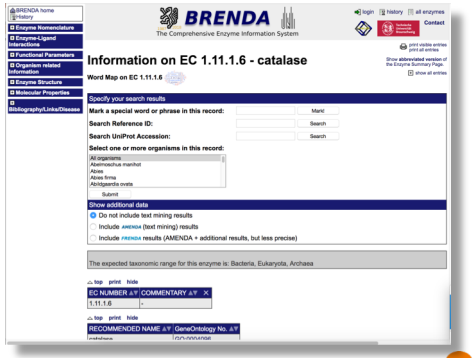
MAR databases

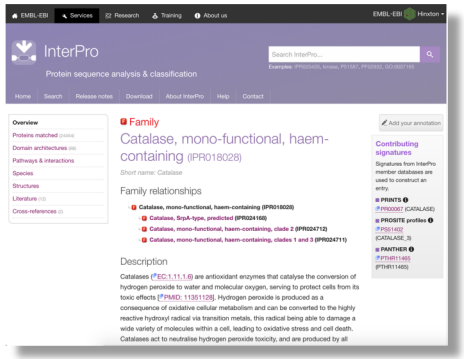
BLAST





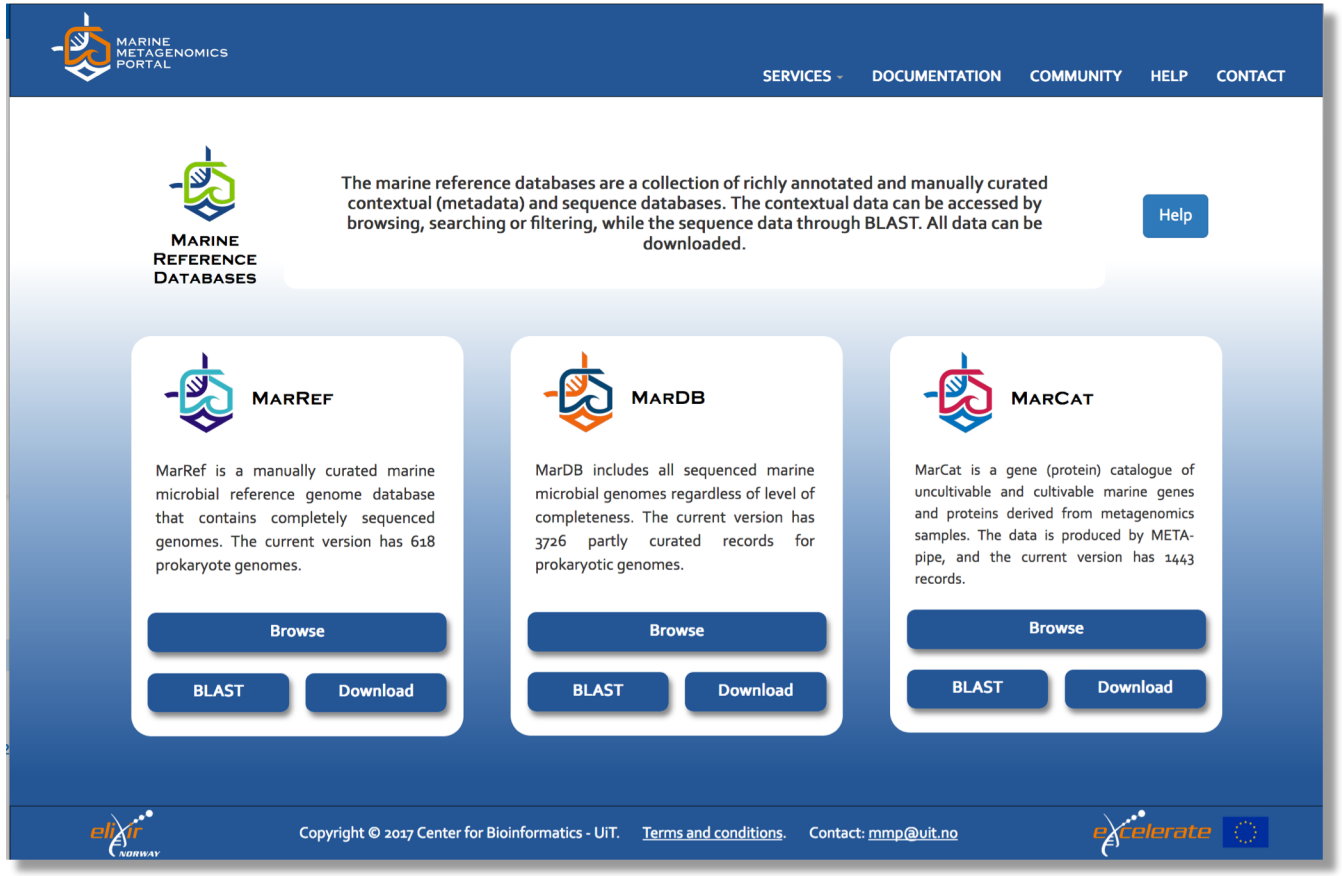






MAR databases

Marine reference databases



Nucleic Acids Research, 2017 1
doi: 10.1093/nar/gkx1036

The MAR databases: development and implementation of databases specific for marine metagenomics

Terje Klemetsen[1], Inge A. Raknes[1], Juan Fu[1], Alexander Agafonov[1], Sudhagar V. Balasundaram[1], Giacomo Tartari[1,2], Espen Robertsen[1] and Nils P. Willassen[1,*]

[1]Centre for Bioinformatics, Faculty of science and technology, UiT The Arctic University of Norway, PO Box 6050 Langnes, TromsøN-9037, Norway and [2]Department of Information Technology, UiT The Arctic University of Norway, PO Box 6050 Langnes, TromsøN-9037, Norway

Received August 14, 2017; Revised October 12, 2017; Editorial Decision October 13, 2017; Accepted October 18, 2017

ABSTRACT

We introduce the marine databases; *MarRef*, *MarDB and MarCat* (https://mmp.sfb.uit.no/databases/), which are publicly available resources that promote marine research and innovation. These data resources, which have been implemented in the Marine Metagenomics Portal (MMP) (https://mmp.sfb.uit.no/), are collections 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 incomplete sequenced prokaryotic genomes regardless level of completeness. The last database, *MarCat*, represents a gene (protein) catalog of uncultivable (and cultivable) marine genes and proteins derived from marine metagenomics samples. The first versions of *MarRef* and *MarDB* contain 612 and 3726 records, respectively. Each record is built up of 106 metadata fields including attributes for sampling, sequencing, assembly and annotation in addition to the organism and taxonomic information. Currently, *MarCat* contains 1227 records with 55 metadata fields. Ontologies and controlled vocabularies are used in the contextual databases to enhance consistency. The user-friendly web interface lets the visitors browse, filter and search in the contextual databases and perform BLAST searches against the corresponding sequence databases. All contextual and sequence databases are freely accessible and downloadable from https://s1.sfb.uit.no/public/mar/.

INTRODUCTION

Microorganisms are ubiquitous in the marine environment, where they play key roles in many global and local biogeochemical processes such as nutrient recycling (1). These microorganisms and the communities they form, drive and respond to changes in the environment and alterations in the marine environment (2). With an estimated 10^4 to 10^6 cells per milliliter seawater and totally over 10^{29} bacterial cells in open sea, the marine microorganisms provide the grounds for immense genetic diversity (3).

Since the first complete bacterial genome published in 1995 (4), the number of sequenced microbial genomes has increased dramatically. Currently, more than 103 000 prokaryotic genomes are available in the National Center for Biotechnology Information (NCBI) Genome microbial database (https://www.ncbi.nlm.nih.gov/genome/microbes/). Originally sequencing efforts were prioritized to study cultured microbes. However, it is well established that the vast majority of bacterial and archaeal taxa remain uncultivated *in vitro* (5). Recently, cultivation-independent methods such as single cell genomics and genomes reconstructed from metagenomic deep sequencing, have begun to yield complete or near-complete genomes from many novel lineages (5–7). Metagenomics, the study of genetic material recovered directly from environmental samples, is a powerful tool for surveying the diversity of marine microbes, which are important for the study of marine sciences. Prominent examples of metagenomics studies in the marine field include the Sorcerer II expeditions (8), Malaspina expedition (9), Global Ocean Sampling (GOS) campaign (10) and Tara Oceans expedition (11). Most of these data as well as other marine metagenomic data are stored in publicly available metagenomic databases such as iMicrobe (https://www.imicrobe.us/), Viral Informatics Resource for Metagenome Exploration (VIROME) (12), EBI metagenomics (13), Integrated Microbial Genomes and Microbiomes (IMG/M) (14) and Metagenomics Rapid Annota-

*To whom correspondence should be addressed. Tel: +47 7764 4651; Email: nils-peder.willassen@uit.no

Accessible from the Marine Metagenomics Portal (MMP)
https://mmp.sfb.uit.no/

MMP- team:

Terje Klementsen
Juan Fu
Aleksander Agafonov
Sudhagar Veerabadran Balasundaram
Espen Robertsen
Erik Hjerde
Espen Åberg
Nils Peder Willassen

Web-site:

http://mmp.sfb.uit.no

Contact:

mmp@uit.no

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe