



# *New data resources for marine metagenomics*

*Nils Peder Willassen*

*Centre for bioinformatics*

*UiT The Arctic University of Norway*

*ELIXIR Norway*



*European Life Sciences Infrastructure for Biological Information*  
[www.elixir-europe.org](http://www.elixir-europe.org)



## Marine metagenomics Use case (Marine metagenomics Community from 2019)

The screenshot shows the 'Use Cases' page on the ELIXIR website. The navigation bar includes 'ABOUT US', 'SERVICES', 'PLATFORMS', 'USE CASES', 'EVENTS', 'NEWS', and 'INTRANET'. A search bar is located on the right. The main content area is titled 'Use Cases' and includes a sub-header 'Home » Use Cases'. A sidebar on the left lists 'USE CASES' with links to 'Human Data', 'Rare Diseases', 'Marine Metagenomics', and 'Plant Sciences'. The main text explains that Use Cases coordinate ELIXIR's activities in four domains: Human Data, Rare Disease, Marine Metagenomics, and Plant Sciences. Below this are four icons representing each domain with brief descriptions of their goals. A blue box highlights 'New Use Cases' that have just begun: Proteomics, Metabolomics, and Galaxy. At the bottom, 'Goals of the Use Cases' are listed, and a note states that the four Use Cases began in 2015 with the start of the ELIXIR-EXCELERATE grant.

**USE CASES**

- Human Data
- Rare Diseases
- Marine Metagenomics
- Plant Sciences

**Use Cases**

Use Cases coordinate ELIXIR's activities in four domains of life science: Human Data, Rare Disease, Marine Metagenomics and Plant Sciences. They bring together experts to develop specialised standards, services, workshops and Implementation Studies in these domains. The Use Cases also provide feedback on the Platform services, which helps ensure they are practical and useful.

**Human Data**  
Develops long-term strategies for managing and accessing sensitive human data.

**Rare Diseases**  
Supports the development of new therapies for rare diseases.

**Marine Metagenomics**  
Develops a sustainable metagenomics infrastructure to nurture research and innovation in the marine domain.

**Plant Sciences**  
Develops an infrastructure to facilitate genotype-phenotype analysis for crop and tree species.

**New Use Cases:** Three new Use Cases have just begun: Proteomics, Metabolomics and Galaxy. For details see the [news story](#). More details about them will appear in this section shortly.

**Goals of the Use Cases**

- To tie together ELIXIR services into effective workflows that support the needs of the life science community
- To guide the coordination and enhancement of ELIXIR's resources in a specific domain (e.g. Human Data, Marine Metagenomics)
- To demonstrate the effectiveness of ELIXIR services
- To test the technologies and solutions developed by ELIXIR Platforms.

The four Use Cases began in 2015 with the start of the [ELIXIR-EXCELERATE](#) grant and will run at least until 2019. The

The screenshot shows the 'Marine Metagenomics Use Case' page on the ELIXIR website. The navigation bar is the same as the previous page. The main content area is titled 'Marine Metagenomics Use Case' and includes a sub-header 'Home » Use Cases »'. A sidebar on the left lists 'USE CASES' with links to 'Human Data', 'Rare Diseases', 'Marine Metagenomics', and 'Plant Sciences'. The main text defines marine metagenomics as the study of genetic material recovered directly from the sea. It states that this Use Case aims to develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain. Below this, 'What the Use Case does' is listed with five bullet points. 'Services provided' are also listed with five bullet points, including the EBI metagenomics portal, Marine Metagenomics Portal (MMP), MarRef, MarDB, MarCat, META-pipe, and ITSoneDB.

**USE CASES**

- Human Data
- Rare Diseases
- Marine Metagenomics
- Plant Sciences

**Marine Metagenomics Use Case**

Marine metagenomics is the study of genetic material recovered directly from the sea. It is a new and rapidly expanding area of research, and there is a danger that data is produced faster than users are able to share, analyse and interpret it. There is an urgent need to create a data management infrastructure dedicated to marine research.

This Use Case aims develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain. It will provide a web based portal through which researchers can access a marine reference database. The database will be populated with data from the [European Nucleotide Archive \(ENA\)](#), [UniProt](#) and other sources.

**What the Use Case does**

- Develops and implements data standards for the marine domain.
- Develops and implements databases specific to marine metagenomics.
- Evaluates and implements tools and pipelines for metagenomics analyses.
- Develops a search engine for the interrogation of marine metagenomics datasets.
- Organises training workshops for end users.

**Services provided**

- **EBI metagenomics portal: portal for metagenomics analysis:** an automated pipeline for analysing and archiving metagenomic data.
- **Marine Metagenomics Portal (MMP) (Beta version)**
  - **MarRef:** a database for completely sequenced marine prokaryotic genomes.
  - **MarDB:** a database of sequenced marine prokaryotic genomes regardless of level of completeness.
  - **MarCat:** a catalogue of marine genes and proteins derived from metagenomics samples.
  - **META-pipe:** a marine metagenomics analysis pipeline.
- **ITSoneDB:** a database for fungal ITS1 sequences

# 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*

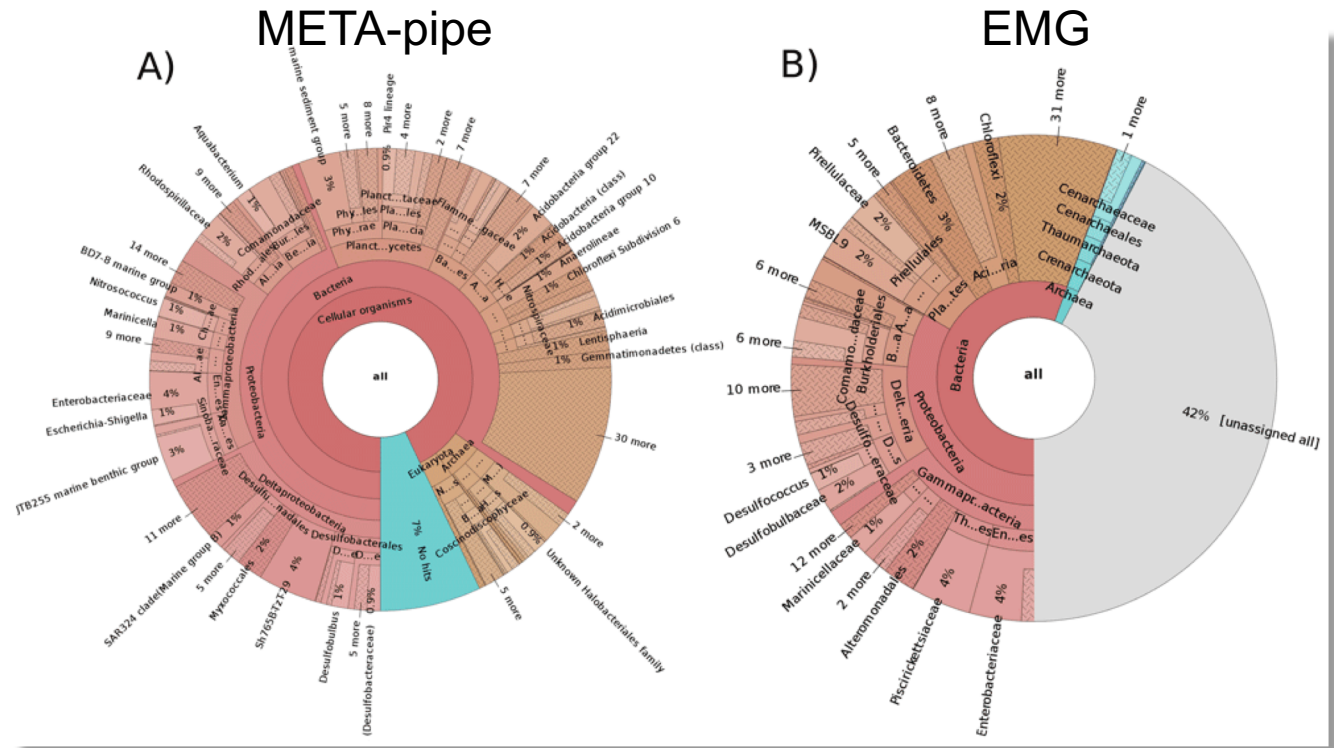


# ELIXIR EXCELERATE

## Background

### Why more (and domain-specific) resources?

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





# Marine Metagenomics Portal

The MAR databases are a collection of richly annotated and manually curated contextual (metadatas) and sequence databases. The contextual data can be accessed by browsing, searching or filtering, while the sequence data through BLAST. All data can be downloaded.

- MARREF**: MarRef is a manually curated marine microbial genome database that contains complete genomes. The current version has 688 prokaryotic genomes.
- MARDB**: MarDB includes all non-complete sequenced marine microbial genomes regardless of level of completeness. The current version contains 708 partly curated records for prokaryotic genomes.
- MARCAT**: MarCat is a gene oriented catalogue of uncultivable and culturable marine microorganisms - derived from metagenomics samples. The data is provided by META-pipe, and the current version contains 100 samples.

Start Meta-pipe

Dataset type: Reads (selected) | Fasta

Select Fasta dataset

Fasta: Upload file | Choose existing

Select parameters

Quality control and assembly

Reads: 1000000000

Minimum overlap length: 100

Taxonomic classification

Statistics: 208.7 GB of META-pipe output datasets

SERVICES | DOCUMENTATION | COMMUNITY | HELP | CONTACT | HELPDESK

**MARINE METAGENOMICS PORTAL**

The mission of Marine Metagenomics Portal (MMP) is to provide the marine scientific community with high-quality curated and freely accessible microbial genomics and metagenomics resources.

- MAR DATABASES**: The MAR (MARine) databases are richly annotated and manually curated contextual and sequence databases. MarRef contains completely sequenced marine prokaryotic genomes. MarDB includes all non-complete marine prokaryotic genomes regardless of level of completeness. MarCat is a catalogue of marine genes and proteins derived from metagenomics samples. **Browse**
- META-PIPE**: META-pipe is a complete workflow for the analysis of marine metagenomic data. It provides assembly of high-throughput sequence data, functional annotation of predicted genes, and taxonomic profiling. META-pipe is not released as an ELIXIR service yet. For now you may use the NeLS META-pipe service. **Run**
- MAR BLAST**: MAR BLAST provides BLAST search on all genes and protein coding sequences from the marine databases MarRef, MarDB and MarCat. **BLAST**

MMP is part of the ELIXIR infrastructure | © 2018 SF6 - UIT | Terms and conditions

Statistics: 2443 MMP and META-pipe unique visitors

The BLAST (Basic Local Alignment Search Tool) finds local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Click here to open the BLAST page in a new tab

SequenceServer 1.0.9

Nucleotide databases: MarRef CDS Nucleotides, MarDB CDS Nucleotides, MarCat CDS Nucleotides

Protein databases: MarRef CDS Proteins, MarDB CDS Proteins, MarCat CDS Proteins

# What is a marine microbial biome?

## **Marine biome**

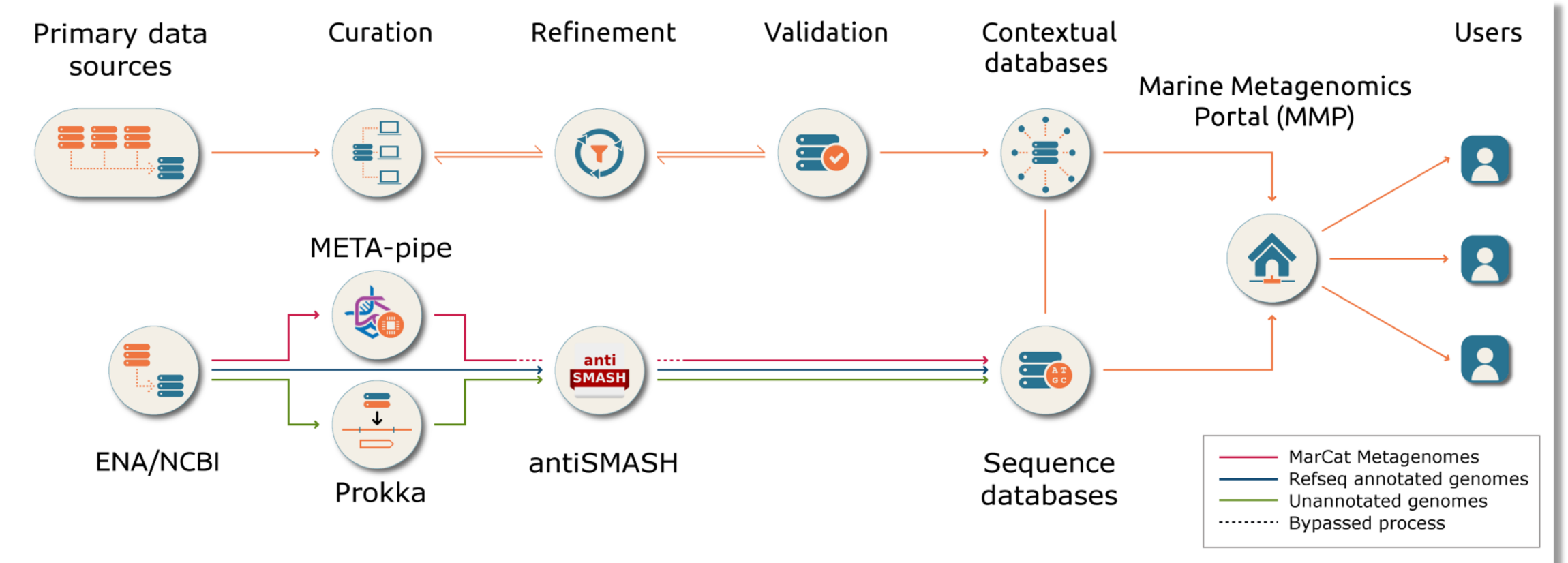
*"An aquatic biome that comprises systems of open-ocean and unprotected coastal habitats, characterized by exposure to wave action, tidal fluctuation, and ocean currents as well as systems that largely resemble these. Water in the marine biome is generally within the salinity range of seawater: 30 to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

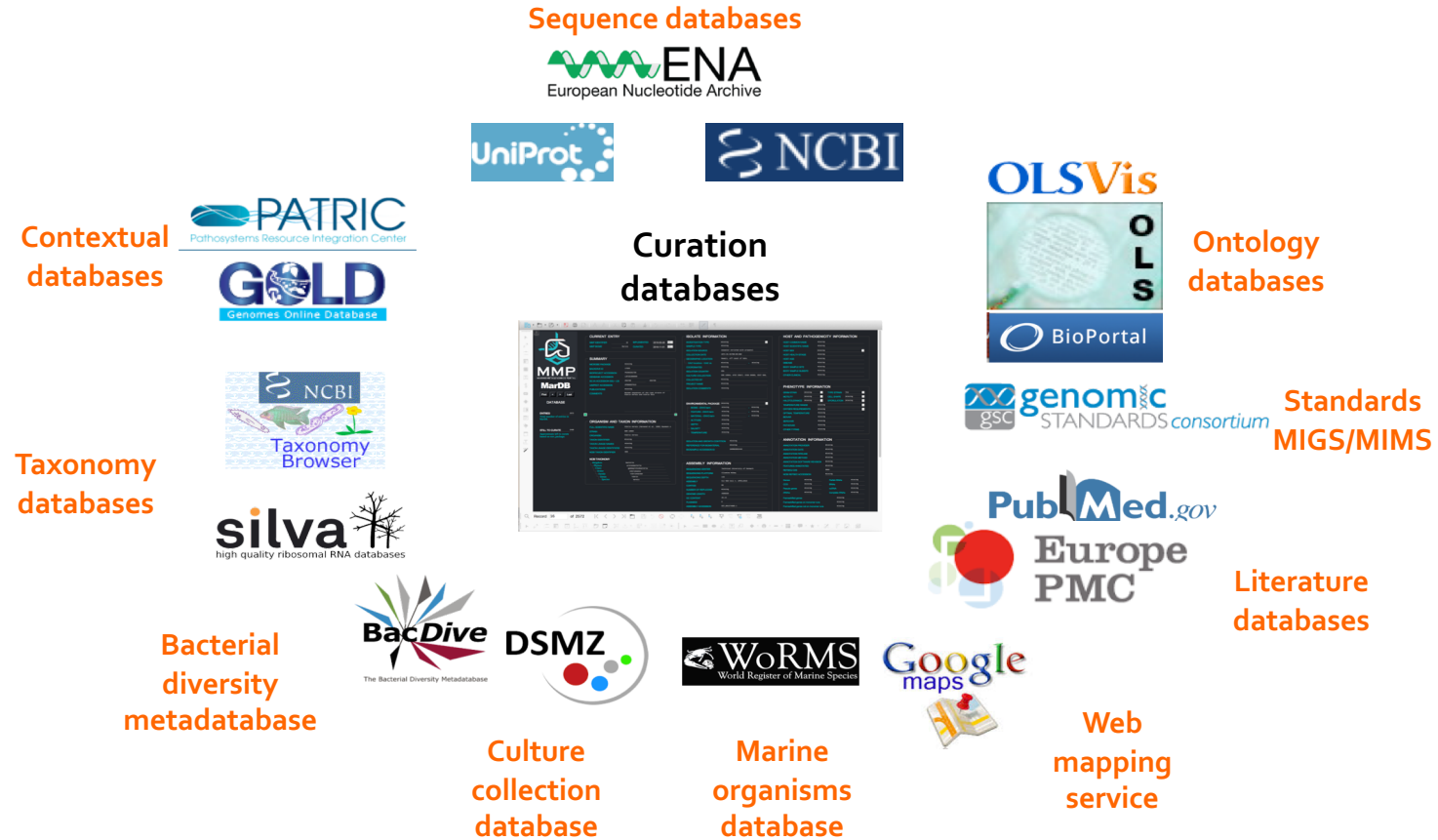
# MAR databases

## Implementation



# MAR databases

## Data resources used



# MAR databases

## Curation database

**MMP**  
MARINE METAGENOMICS PORTAL  
**MarREF**

First < > Last

**DATABASE**

ENTRIES 484  
Total number of entries in database.

STILL TO CURATE 18  
Approximate left to curate based on env\_package.

**CURRENT ENTRY**  
Thermaerobacter marianensis DSM 12885

MMP IDENTIFIER	MMP00713570	IMPLEMENTED	2016-05-20
MMP BIOME	Marine	CURATED	2016-05-01
BASE IDENTIFIER	12	UPDATED	2016-05-01

**SUMMARY**

MICROBE PACKAGE NA

GENOME STATUS Complete

BACKDIVE ID 18848

BIOPROJECT ACCESSION PRJNA38025

GENBANK ACCESSION CP002344

SILVA ACCESSION SSU - LSU 38025 38025

UNPROT ACCESSION UP000000915

PUBLICATIONS 21384738

COMMENTS Thermaerobacter marianensis DSM 12885. This strain will be used for comparative genome analysis.

**ISOLATE INFORMATION**

INVESTIGATION TYPE Bacteria

SAMPLE TYPE Cell culture

ISOLATION SOURCE Mud sample of the Challenger Deep in th

COLLECTION DATE 1996

GEOGRAPHIC LOCATION Pacific Ocean:Mariana Trench:Challenger

GAZ location - GAZ nr. Challenger Deep - GAZ:00007907

COORDINATES 11.350000, 142.410000

ISOLATION COUNTRY Pacific Ocean

CULTURE COLLECTIONS DSM 12885:ATCC 790841:JCM 10246

COLLECTED BY Japan MarineScience and TechnologyCente

PROJECT NAME Thermaerobacter marianensis DSM 12885 g

ISOLATION COMMENTS missing

**ENVIRONMENTAL PACKAGE** Sediment

BIOME - ENVO term	Marine benthic biome	ENVO:01000024
FEATURE - ENVO term	Ocean trench	ENVO:00000275
MATERIAL - ENVO term	Mud	ENVO:01000001
ALTITUDE	missing	m
DEPTH	10897	m
SALINITY	missing	psu
TEMPERATURE	missing	°C

**ISOLATION AND GROWTH CONTITION** 10319484

REFERENCE FOR BIOMATERIAL 10319484

BIOSAMPLE ACCESSION ID SAMN00713570

**ASSEMBLY INFORMATION**

SEQUENCING CENTER US DOE Joint Genome Institute (JGI-)

SEQUENCING PLATFORM Illumina GAII:454 GS FLX Titanium

SEQUENCING DEPTH 340.8:91.0 x

ASSEMBLY Newbler v. 2.1:Phrap:Velvet

CONTIGS 0

NUMBER OF REPLICONS 1

GENOME LENGTH 2844696 nt

GC CONTENT 72.48 %

PLASMIDS 0

ASSEMBLY ACCESSION GCA\_900184785.1

**HOST AND PATHOGENICITY INFORMATION**

HOST COMMON NAME NA

HOST SCIENTIFIC NAME NA

HOST SEX NA

HOST HEALTH STAGE NA

HOST AGE NA

PATHOGENICITY missing

DISEASE missing

BODY SAMPLE SITE NA

BODY SAMPLE SUBSITE NA

OTHER CLINICAL NA

**PHENOTYPE INFORMATION**

GRAM STAIN	Positive	TYPE STRAIN	Yes
MOTILITY	No	CELL SHAPE	Bacilli (Rod)
HALOTOLERANCE	missing	SPORULATION	Yes
TEMPERATURE RANGE	Hyperthermophilic		
OXYGEN REQUIREMENTS	Aerobic		
OPTIMAL TEMPERATURE	74-76 °C		
BIOVAR	missing		
SEROVAR	missing		
PATHOVAR	missing		
OTHER TYPING	missing		

**ANNOTATION INFORMATION (REFSEQ)**

ANNOTATION PROVIDER NCBI

ANNOTATION DATE 2015-08-10T01:34:20

ANNOTATION PIPELINE NCBI Prokaryotic Genome Annotatic

ANNOTATION METHOD Best-placed reference protein set

ANNOTATION SOFTWARE REVISION 3.0

FEATURES ANNOTATED Gene:CDS:rRNA:tRNA:ncRNA:repeat\_r

REFSEQ CDS 2235

NCBI REFSEQ ACCESSION NC\_014831

Genes	2370	Partial rRNAs	0, 0, 0
CDS	2235	tRNAs	50
Pseudo genes	77	ncRNA	2
rRNAs	2, 2, 2	Complete rRNAs	2, 2, 2

Frameshifted genes 13

Frameshifted genes on monomer runs 1

Frameshifted genes not on monomer runs 0

**ORGANISM AND TAXON INFORMATION**

FULL SCIENTIFIC NAME Thermaerobacter marianensis DSM 12885

STRAIN 7p75a:DSM 12885:ATCC 790841:JCM 10246

ORGANISM Thermaerobacter marianensis

NCBI TAXON IDENTIFIER 644066

TAXON LINEAGE NAMES cellular organisms:Bacteria:Terrabacter.

TAXON LINEAGE IDENTIFIERS 131567:2:1783272:1299:1088001:1088002:530

**NCBI TAXONOMY**

- Kingdom Bacteria
- Phylum Firmicutes
- Class Clostridia
- Order Clostridiales
- Family Clostridiales Family XVII. Incer
- Genus Thermaerobacter
- Species Thermaerobacter marianensis

# 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:( <a href="http://www.environmentontology.org/Browse-EnvO">http://www.environmentontology.org/Browse-EnvO</a> )	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:( <a href="http://www.environmentontology.org/Browse-EnvO">http://www.environmentontology.org/Browse-EnvO</a> )	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).	<a href="https://www.ebi.ac.uk/metagenomics/projects/SRP000183/samples/SRS0000447">https://www.ebi.ac.uk/metagenomics/projects/SRP000183/samples/SRS0000447</a>	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:( <a href="http://www.environmentontology.org/Browse-EnvO">http://www.environmentontology.org/Browse-EnvO</a> )	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	<a href="http://purl.obolibrary.org/obo/ENVO_00000447">http://purl.obolibrary.org/obo/ENVO_00000447</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000045">http://purl.obolibrary.org/obo/ENVO_01000045</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000020">http://purl.obolibrary.org/obo/ENVO_01000020</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000046">http://purl.obolibrary.org/obo/ENVO_01000046</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000024">http://purl.obolibrary.org/obo/ENVO_01000024</a>
Marine mud	A liquid or semi-liquid mixture of water and some combination of soil, silt, and clay.	ENVO:00005795	<a href="http://purl.obolibrary.org/obo/ENVO_00005795">http://purl.obolibrary.org/obo/ENVO_00005795</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000023">http://purl.obolibrary.org/obo/ENVO_01000023</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000022">http://purl.obolibrary.org/obo/ENVO_01000022</a>
Marine upwelling biome	A marine biome which contains communities adapted to living in an environment determined by an upwelling process.	ENVO:01000898	<a href="http://purl.obolibrary.org/obo/ENVO_01000898">http://purl.obolibrary.org/obo/ENVO_01000898</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_00001999">http://purl.obolibrary.org/obo/ENVO_00001999</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000047">http://purl.obolibrary.org/obo/ENVO_01000047</a>
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	<a href="http://purl.obolibrary.org/obo/ENVO_01000048">http://purl.obolibrary.org/obo/ENVO_01000048</a>

*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

Refine MarDB\_full\_68083 csv Permalink

Facet / Filter Undo / Redo 68081 rows Show as: rows records Show: 5 10 25 50 rows

Depth	All	Genome ID	Genome Name	Organism Name	NCBI Taxon ID	Genome Status	Strain	Serovar	Biovar	Pathovar	MLST	Other Typing	Culture Collectio	Type Strain	Completion Date	Publicat
225 choices Sort by: name count																
24_24m 10																
5365 1																
58 1																
5800 m 1		68066	Vibrio parahaemolyticus strain S499-7		670	WGS	S499-7								2015-08-18	
5904 m 1		68067	Vibrio parahaemolyticus strain S439-9		670	WGS	S439-9								2015-08-18	
5m 3																
600 metres 1																
63 1																
63 m 1		68068	Vibrio parahaemolyticus strain S487-4		670	WGS	S487-4								2015-08-18	
642 M 1																
770m 1		68069	Vibrio parahaemolyticus strain S440-7		670	WGS	S440-7								2015-08-18	
79 meters 1																
8_12cm 25																
800m 3		68070	Vibrio parahaemolyticus strain S357-21		670										2015-09-21	
8219 3																
835 1																
840 1		68071	Vibrio parahaemolyticus strain S372-5		670										2015-09-09	
850 m 1																
851 m 1		68072	Vibrio vulnificus 99-796 DP-67		672										2014-12-02	
852 m 1																
853 m 1																
854 m 1																
855 m 1																
856 m 1																
857 m 1																
858 m 1																
859 m 1																
860 m 1																
861 m 1																
9 m 1		68073	Vibrio parahaemolyticus VIP4-0430		1408170										2013-11-29	
9.0 m 1																
90 cm below water surface 2																
9161 below sea level 1		68074	Vibrio sp. CY15		1440054										2014-08-20	25278542
95 m 16																
below surface 1																
missing 1																
N.A. 1		68075	Vibrio parahaemolyticus 49		1288779										2015-02-18	
not applicable 5																
NOT APPLICABLE 4																
not collect 1																
Not collected 1		68076	Vibrio parahaemolyticus 646		1288783										2015-02-18	
not collected 26																
sea level 1																
sediment 1																
soil surface 5																
Superficial water 1		68077	Vibrio parahaemolyticus 930		1288785	WGS		930							2015-02-18	
Surface 3																
surface 19																
surface sample 1																
surface sea water 1																
surface soil 2																
surface water 4		68078	Vibrio parahaemolyticus strain CT4287		670	WGS		CT4287							2015-06-09	25904905
The average depth is 25 metre. 1																
unclear 1																

Depth:  
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 e.g. 3.06 or 1.80-2.15

- 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)

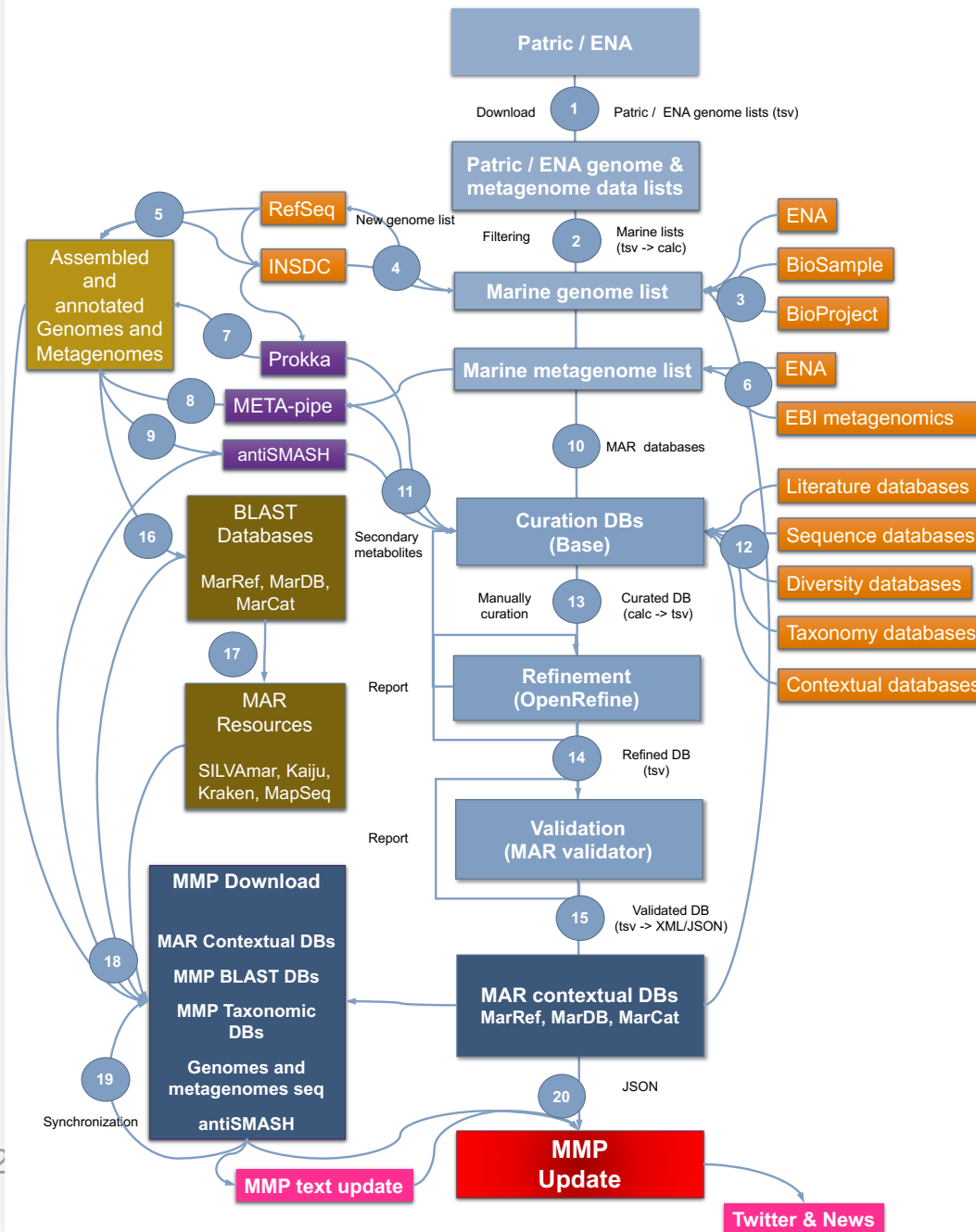
type	project_name	env_biome	env_biome_EVO	env_package	env_feature	env_feature_EVO	env_material
line	env_salinity	collected by	env_temp	host_sex	host_age	host_health_stage	body_material
her_typing	culture_collection			ref_biome	publication_publication	bioproject	bioproject
ingon	phylum	class	order	family	genus	species	taxon_lineage_ids
rtial_rnas	trnas	ncrna	framedshifted_genes	framedshifted_genes_on_monomer_runs	NA	framedshifted_genes_not	imp_labeled
issing	Marine	missing	missing	missing	missing	NA	missing
issing	Spleen	missing	Negative	missing	No	missing	Mesophilic
ipeoialis	Brucella pinnipedialis	8274	IMCC122090	missing	missing	missing	NCTC 12890
1395270	57.2	Bacteria	Proteobacteria	Alphaproteobacteria	Brucellales	Brucellaceae	
BI	2015-08-18T01:45:34Z	NCBI Prokaryotic Genome Annotation Pipeline	Best-placed reference protein set; GeneMarkES	NA	NA	NA	NA
issing	Marine	Marine benthic biome	ENVO:01000024	Sediment	Sea floor	ENVO:00000402	Sediment
00	University of California	missing	NA	NA	NA	Negative	NA
ty of which are on chromosome 1. There are 2 ATP synthase operons, one on each chromosome. Chromosome 1 contains more conserved genes							
issing	missing	ATCC BAA-1253	Yes	298386	15746425	15746425	PRJNA13128
oteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Photobacterium	Photobacterium	proferens	
BI	Prokaryotic Genome Annotation Pipeline	Best-placed reference protein set; GeneMarkES	3.3	GeneID:FRNA1FRNA1ncRN			
issing	Marine	Marginal sea biome	ENVO:01000045	Sea coast	ENVO:00000393	Sea water	ENVO:0000214
rden	Kongsfjorden	GAZ:00103000	79.100140	11.309200	missing	missing	Iha University
ip. IMC9063	Candidatus Pelagibacter	IMC9063	missing	missing	missing	IMC9063	missing
issing	1204727	292	Bacteria	Alphaproteobacteria	Pelagibacteriales	Pelagibacter	
ip. IMC9063	2016-05-20	2016-11-22	NA	missing	NCBI Prokaryotic Genome Annotation		
003	Complete						
issing	Marine	Marginal sea biome	ENVO:01000046	Water	Sea shore	ENVO:00000405	Sea water
issing	missing	missing	Korea Ocean Research & Development Institute	missing	NA	NA	NA
IMCC1322	Candidatus Puniceispirillum marinum	IMCC1322	missing	missing	missing	IMCC1322	missing
issing	275327	48.9	Bacteria	Proteobacteria	Alphaproteobacteria	missing	
tus	Puniceispirillum marinum	Candidatus Puniceispirillum marinum	IMCC1322	2016-05-20	2016-11-22	NA	missing
issing	missing	WWP020422	20081	Complete			
issing	Marine	Mediterranean sea biome	ENVO:01000047	Sediment	Sea coast	ENVO:00000393	Sediment
depo. 10 to 40 m) from Barcelona, Spain. This area is strongly polluted by industrial and urban wastewater. Missing							
killi (Rod)	Yes	No	Mesophilic	30	missing	Aerobic	Pseudomonas stutzeri
an 0 M. Evolutionary experiments to adapt strain AN16 to higher salicylate concentrations are now underway. Cell culture							
rative Genomics	454 GS FLX Titanium	illumina	hiseq	2000	241400	Celera Assembler v. 7.4	
Gammaproteobacteria	Pseudomonadales	Pseudomonadales	Pseudomonadaceae	Pseudomonas stutzeri group	Pseudomonas stutzeri	subgroup	Pseudomonas
DS	4218	62	4	4	0	12	48
issing	Marine	missing	missing	missing	missing	missing	NA
00	University of Delaware	missing	missing	missing	missing	Negative	Alvinella pac
nis of Alvinella pompejana.	Cell culture	Nautilia profunda	AMH	Nautilia profunda	AMH	DSM 18972	ATCC BAA-1016
lersa Assembler v.1	Missing	Missing	Missing	Missing	Missing	Missing	Missing
utilia profunda	Nautilia profunda	AMH	Missing	Missing	Missing	Missing	Missing
issing	3000208343	2016-05-20	2016-11-22	Pompeii worm	10430	NCBI	2015-07-30T01
issing	Marine	missing	missing	missing	missing	NA	NA
issing	missing	missing	missing	missing	missing	NA	Oncorhynchus
issing	missing	missing	missing	missing	missing	NA	NA
UPN000511	University of Chile	illumina	PaBio	05	190712100	MG-Assembler (Celera) v. 8.2	
Gammaproteobacteria	Thiotrichales	Psicirickettsiaceae	Psicirickettsia	Psicirickettsia	salmonis	2016-05-20	2016-11-22
44	missing	missing	missing	missing	missing	304700	Complete
issing	Marine	missing	missing	missing	missing	NA	Seriola quin
015930_1	UPN000821	missing	missing	missing	missing	NA	Complete
Bacillus	Streptococcaceae	Lactococcus	garvieae	Lactococcus garvieae	ATCC 4956	2016-05-20	2016-11-22
issing	missing	missing	missing	missing	missing	191896	Complete
22.569300	missing	missing	missing	missing	missing	NA	NA
530	missing	missing	missing	missing	missing	NA	Negative
phonadaceae	Maricoccus	Maricoccus	maris	1315672	122412821	1284453	69657
issing	Marine	Marine benthic biome	ENVO:01000024	Sediment	Ocean trench	ENVO:00000795	Mud
200002421	11.368300	142.438000	missing	18098	Japan Marine Science and Technology Center	missing	NA
ch. We report the genome sequence of this bacterium, which contribute to the understanding of denitrification and bioremediation in de							
hon University	454 GS FLX Titanium	31.5	Celera Assembler v. 5.3	1	missing	1	4061250
nadaceae	Pseudomonas sp. HT-1	2016-05-20	2016-11-22	NA	missing	NCBI	2016-05-05T01:43:59Z
issing	2890	2890	Complete				
issing	Marine	Marine benthic biome	ENVO:01000024	Sediment	Sea floor	ENVO:00000402	Sediment
Science and Technology Center	missing	NA	NA	NA	Negative	Bacilli	(Roc
4]ATCC BAA-071	JCM 11897	missing	missing	missing	DSM 12294	ATCC BAA-071	JCM 11897
0	missing	235308	35.24	2128	Bacteria	delta/Epsilon subdivisions	Epsilonproteobacteria
bacteraceae	Sulfurimonas	Sulfurimonas	autotrophica	Sulfurimonas autotrophica	DSM 12294	2016-05-20	2016-01-08
hemerobacter	marianensis	DSM 12285	genome sequencing project	Marine	Marine benthic biome	ENVO:01000024	Sediment
allinger Deep	GAZ:00000700	142.451000	missing	11.590000	Japan Marine Science and Technology Center	missing	NA
ll culture	Thermobacter	marianensis	DSM 12285	Thermobacter marianensis	7675	DSM 12285	ATCC 70884
GI-06F	illumina	GA454	FLX Titanium	340.010	Newbler v.1.3	Phrap	Consed
bacteria group	Firmicutes	Clostridiales	Clostridiales	Incertae sedis	Clostridiales	Family XVII. Incertae	Thermobac
repeat_region	2235	2370	2	2	0	58	15746425
issing	Marine	Marine benthic biome	ENVO:01000024	Water	Marine hydrothermal vent	ENVO:01000122	
12x 240 ml at a depth of 2650 m. Samples were immediately transferred into sterile filtered artificial seawater. 7							
al de la Recherche Scientifique et Universite de Bretagne Occidentale							
al de la most radioresistant hyperthermophilic archaea. Cell culture							
ic Analyzer	missing	Phred	Phrap	Consed	1	23.6	2015-08-17T22
cus	Thermococcus gammatolerans	Thermococcus gammatolerans	E13	2016-05-20	2016-11-22	NA	16885
issing	WWP020333	missing	Complete				
issing	Marine	Marine benthic biome	ENVO:01000024	Sediment	Marine hydrothermal vent	ENVO:01000122	
issing	missing	missing	missing	missing	missing	NA	NA
issing	missing	missing	missing	missing	missing	523849	10.1807
iprchaetota	Thermococcales	Thermococcales	Thermococcus	litoralis	Thermococcus litoralis	22493101	PRJNA81925
BI	Prokaryotic Genome Annotation Pipeline	Best-placed reference protein set; GeneMarkES	3	GeneID:FRNA1FRNA1ncRN			
issing	Marine	Marine benthic biome	ENVO:01000024	Marine hydrothermal vents	ENVO:01000122		
Juan de Fuca Ridge	Juan de Fuca Ridge	GAZ:00002438	47.950000	-120.100000	missing	2200	University of Washington
polyethene worm collected from hydrothermal vent on Endeavour Seamount							
ip. IMC9065.1	UPN001902	Kyova	Hic	University	454 GS FLX Titanium	illumina	HISEQ
eal	Euryarchaeota	Thermococci	Thermococcales	Thermococcus	parvilineae	Lae	2016-05-20
issing	Marine	Marine benthic biome	ENVO:01000024	Misc environment	Marine hydrothermal vent	ENVO:01000122	
the Central Indian Ocean. The chimney fragment was collected by the ROV Jason and was placed in an isolated container for trip 10							
field	GAZ:00002488	25.317000	70.033000	missing	428	missing	NA
Thermosulfator	Indius	DSM 15286	Thermosulfator	Indius	DSM 15286	JCM 11387	ATCC 72961
lumina Genome Analyzer II	454 GS FLX Titanium	30	Newbler v. 2.3	GC v. 3.4.6	Velvet v. 1.0.13	Phrap	Consed
951607814	cellular_organisms	Bacteria	Thermodesulfobacteria	Thermodesulfobacteriales	Thermodesulfobacterales	Thermodesulfobacterales	

Machine-readable (XML/JSON)



# MAR databases

## Curation Workflow for the MAR DBs



# MAR databases

## Marine reference databases

The screenshot shows the homepage of the Marine Reference Databases. At the top, there is a navigation bar with links for SERVICES, DOCUMENTATION, COMMUNITY, HELP, and CONTACT. Below the navigation bar, the main heading is "MARINE REFERENCE DATABASES". A central text block describes the databases as a collection of richly annotated and manually curated contextual (metadata) and sequence databases. Below this, there are three main sections for MARREF, MARDB, and MARCAT. Each section includes a brief description, a "Browse" button, and "BLAST" and "Download" buttons. The footer contains logos for eLife NORWAY, Center for Bioinformatics - UIT, Terms and conditions, Contact: mmp@uit.no, eAccelerate, and the European Union flag.

- 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

Accessible from the Marine Metagenomics Portal (MMP)

<https://mmp.sfb.uit.no/>

Marine Metagenomics Workshop, Nov 26-30, 2018, Tromsø, Norway



# MAR databases

## Browse and filtering

MARINE METAGENOMICS PORTAL

SERVICES - DOCUMENTATION - COMMUNITY - HELP - CONTACT - HELPDESK

**MARREF**

MarRef is a manually curated marine microbial reference genome database that contains completely sequenced genomes. Each entry contains 106 metadata fields including information about sampling environment or host, organism and taxonomy, phenotype, pathogenicity, assembly and annotation information. The current version contain 618 genomes. [Help](#)

[Overview](#) [Browse](#)

Database overview	
Bacteria	530
Archaea	88
Total number of records	618

MarRef

Kartdata ©2018 Vilkår, 1,000 km

MARINE METAGENOMICS PORTAL

SERVICES - DOCUMENTATION - COMMUNITY - HELP - CONTACT - HELPDESK

**MARREF**

MarRef is a manually curated marine microbial reference genome database that contains completely sequenced genomes. Each entry contains 106 metadata fields including information about sampling environment or host, organism and taxonomy, phenotype, pathogenicity, assembly and annotation information. The current version contain 618 genomes. [Help](#)

[Overview](#) [Browse](#)

Database overview	
Bacteria	530
Archaea	88
Total number of records	618

← Thermosipho sp. 1063

nbvni  
Thermosipho sp. 1063  
beskrivelse  
MMP02644973  
(<https://mmp.sfb.uit.no/databases/marref/#/records/MMP02644973>) Genome sequencing of Thermotogales isolates from hydrothermal vents

MarRef

Kartdata ©2018 Vilkår, 1,000 km

# MAR databases

## Browse and filtering

The screenshot shows the MARREF website interface. At the top, there is a navigation bar with 'SERVICES', 'DOCUMENTATION', 'COMMUNITY', and 'CONTACT'. Below the navigation bar is the MARREF logo, which consists of a stylized blue and green icon resembling a hand or a microscope, followed by the text 'MARREF'. Underneath the logo are the links 'Overview' and 'Browse'. The main heading is 'Rhodothermus marinus DSM 4252'. Below the heading is a 'Summary' section with a table of metadata.

MMP ID	MMP00002700
Full Scientific Name	Rhodothermus marinus DSM 4252
Strain	R-50, DSM 4252, ATCC 43812
Type Strain	Yes
Geographic location	Iceland, Reykjanes, Hafjardardjup Bay
Collection Date	1988-01-01
BioSample Accession	SAMN00002700
BacDive ID	17791
Culture Collection(s)	DSM 4252, ATCC 43812
Isolation Country	Iceland
Environmental Package	Misc environment
Isolation Source	Submarine hot spring at Reykjanes, NW Iceland
Host Scientific Name	not applicable
Curation Date	2016-11-22
Updated Date	missing
Implementation Date	2016-05-20
Microbe Package	missing
Experiment/Investigation Type	Bacteria
Bioproject Accession	PRJNA149481
Genbank Accession	CP000807, CP000808
NCBI Taxon Identifier	518166
Reference for Biomaterial	10.1093/00222720/134-2-299
Silva Accession SSU	23481
Silva Accession LSU	23481
Uniprot Accession	UP000002231
Publication(s)	2194469
Comments	Rhodothermus marinus DSM 4252. Rhodothermus marinus DSM 4252 was isolated from a shallow marine hot spring off the coast of Iceland. This organism produces a number of thermostable enzymes, such as cellulase and xylanase, which may be important in industrial processes. Rhodothermus marinus DSM 4252 is the type strain and this genome will provide information on the production and regulation of thermostable enzymes. env_temp: 75-95

The screenshot shows the MARREF website interface for the 'Rhodothermus marinus DSM 4252' record. The page is titled 'Rhodothermus marinus DSM 4252'. Below the title is a navigation bar with 'SERVICES', 'DOCUMENTATION', 'COMMUNITY', and 'CONTACT'. The main content area is divided into several sections: 'Summary', 'Organism and taxon info', 'Isolate info', 'Host and pathogenicity info', 'Assembly info', and 'Annotation info'. The 'Assembly info' section contains a table of sequencing and assembly details.

Sequencing Centers	US DOE Joint Genome Institute (JGI-PG)DOE Joint Genome Institute(JGI-PGI)
Sequencing Platform	ABI 3730 Genetic Analyzer, 454 GS FLX
Sequencing Depth (x)	8.8, 23.9
Assembly Accession	GCA_000004845.1
GC Content (%)	64.3
Assembly Tool	Newbler v. 1.1.02.15, Phrap
Number of Replicons	1
Number of Contigs	missing
Genome Length (bp)	3386029
Plasmids	1

The 'Annotation info' section contains a table of annotation details.

Annotation Provider	NCBI
Annotation Date	2015-07-30 17:37:47
Annotation Pipeline	NCBI Prokaryotic Genome Annotation Pipeline
Annotation Method	Best-placed reference protein sets, GeneMarkS+
Annotation Software Revision	3
Features Annotated	Gene, CDS, rRNA, tRNA, ncRNA, repeat_region
Refseq CDS	2842
NCBI Refseq Accession	NC_043391.1, NC_043392.1
Genes	2931
CDS	2842
Pseudo Genes	19
tRNAs (S, 16S, 23S)	111
Complete rRNAs (S, 16S, 23S)	111
Partial rRNAs (S, 16S, 23S)	0100
tRNAs	45
ncRNAs	2
Frameshifted Genes	6
Frameshifted Genes On Monomer Runs	0
Frameshifted Genes Not On Monomer Runs	0



# MAR databases

## Browse and filtering

MarRef is a manually curated marine microbial reference genome database that contains completely sequenced genomes. Each entry contains 106 metadata fields including information about sampling environment or host, organism and taxonomy, phenotype, pathogenicity, assembly and annotation information. The current version contain 618 genomes.

Overview Browse

MMP ID	Full Scientific Name	Strain	Type Strain	Geographic location	Collection Date	Biosample Accession	Bacvide ID
<input type="checkbox"/> MMP02603341	<i>Brucella pinnipedialis</i> B2/94	B2/94	Yes	Scotland	missing	SAMN02603341	missing
<input type="checkbox"/> MMP13138356	<i>Photobacterium profundum</i> SS9	SS9	Yes	Philippines	missing	SAMEA13138356	missing
<input type="checkbox"/> MMP02603337	<i>Candidatus Pelagibacter</i> sp. IMCC9063	IMCC9063	missing	Norway	missing	SAMN02603337	missing
<input type="checkbox"/> MMP02603472	<i>Candidatus Puncicepirillum marinum</i> IMCC1322	IMCC1322	missing	South Korea	missing	SAMN02603472	missing
<input type="checkbox"/> MMP02603608	<i>Pseudomonas stutzeri</i> CCUG 29243	AN10	missing	Spain	missing	SAMN02603608	missing
<input type="checkbox"/> MMP02603431	<i>Nautilia profundicola</i> AmH	AmH	Yes	East Pacific Rise	1999	SAMN02603431	10430
<input type="checkbox"/> MMP04309069	<i>Piscirickettsia salmonis</i> strain PSCGR01	PSCGR01	missing	Chile	2010-10-20	SAMN04309069	missing
<input type="checkbox"/> MMP02596969	<i>Lactococcus garvieae</i> ATCC 49156	YT-3	No	Japan	1974	SAMN02596969	14683
<input type="checkbox"/> MMP02598982	<i>Maricaulis maris</i> MCS10	MCS10	missing	USA	missing	SAMN02598982	missing
<input type="checkbox"/> MMP00017418	<i>Pseudomonas</i> sp. MT-1	MT-1	missing	Pacific Ocean	missing	SAM00017418	missing
<input type="checkbox"/> MMP00713560	<i>Sulfurimonas autotrophica</i> DSM 16294	OK10	Yes	Japan	2001-06	SAMN00713560	6114
<input type="checkbox"/> MMP00713570	<i>Thermaerobacter marianensis</i> DSM 12885	79754	Yes	Pacific Ocean	1996	SAMN00713570	18040
<input type="checkbox"/> MMP02603333	<i>Thermococcus gammatolerans</i> E13	E13	Yes	USA	1991	SAMN02603333	16885
<input type="checkbox"/> MMP02603679	<i>Thermococcus litoralis</i> DSM 5473	NS-C	Yes	Italy	missing	SAMN02603679	16862
<input type="checkbox"/> MMP03085513	<i>Thermococcus</i> sp. ES1	ES1	Yes	Pacific Ocean	missing	SAMN03085513	24653
<input type="checkbox"/> MMP02232057	<i>Thermodesulfator indicus</i> DSM 15286	CIR2982	Yes	Central Indian	2001-04	SAMN02232057	16889
<input type="checkbox"/> MMP021919351	<i>Thermotoga maritima</i> MSB8	MSB8	Yes	Italy	1982	SAMN021919351	17060
<input type="checkbox"/> MMP02603251	<i>Thermotoga neapolitana</i> DSM 4359	NS-E	Yes	Italy	missing	SAMN02603251	17061
<input type="checkbox"/> MMP00000279	<i>Thermotoga</i> sp. RQ2	RQ2	missing	Portugal	missing	SAMN00000279	missing
<input type="checkbox"/> MMP00713575	<i>Hemovibrio ammonificans</i> HB-1	HB-1	Yes	Pacific Ocean	2000-04	SAMN00713575	17646
<input type="checkbox"/> MMP02232063	<i>Thioflavococcus mobilis</i> 8321	8321	Yes	USA	1986	SAMN02232063	missing
<input type="checkbox"/> MMP02603567	<i>Verrucosipora maris</i> AB-18-032	AB-18-032	Yes	Japan	1991-08	SAMN02603567	8032
<input type="checkbox"/> MMP03232923	<i>Vibrio nigrificans</i> str. SFn1	SFn1	missing	France	2000-03	SAMN03232923	missing
<input type="checkbox"/> MMP02604302	<i>Vibrio parahaemolyticus</i> BB22OP	BB22OP	missing	Bangladesh	missing	SAMN02604302	missing
<input type="checkbox"/> MMP03085521	<i>Vibrio parahaemolyticus</i> UCM-V493	UCM-V493	missing	Spain	2002	SAMN03085521	missing

Download selected records

Unselect Download

tsv fasta xml Genbank

Number of selected records: 0

Search

Filter

Reset all filters

Phylum  
Select Phylum

Order  
Select Order

Genus  
Select Genus

Environment Biome  
Select Environment Biome

Environment Feature  
Select Environment Feature

Environment Material  
Select Environment Material

Environmental Package  
Select Environmental Package

Isolation Country  
Select Isolation Country

Temperature Range  
Select Temperature Range

Halotolerance  
Select Halotolerance

Oxygen Requirement  
Select Oxygen Requirement

Host Common Name  
Select Host Common Name

Collection Date  
min. max.

Depth  
min. max.

# 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

Filter

Reset all filters

Project Name  
Select Project Name

Phylum  
Select Phylum

Order  
Select Order

Genus  
Select Genus

Environment Biome  
Select Environment Biome

Environment Feature  
Select Environment Feature

Environment Material  
Select Environment Material

Environmental Package  
Select Environmental Package

Isolation Country  
Select Isolation Country

Temperature Range  
Select Temperature Range

Halotolerance  
Select Halotolerance

Oxygen Requirement  
Select Oxygen Requirement

Host Common Name  
Select Host Common Name

Collection Date  
min. max.

Depth  
min. max.

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

Filtering based on region or country

Filtering based on host (pathogen or host associate)



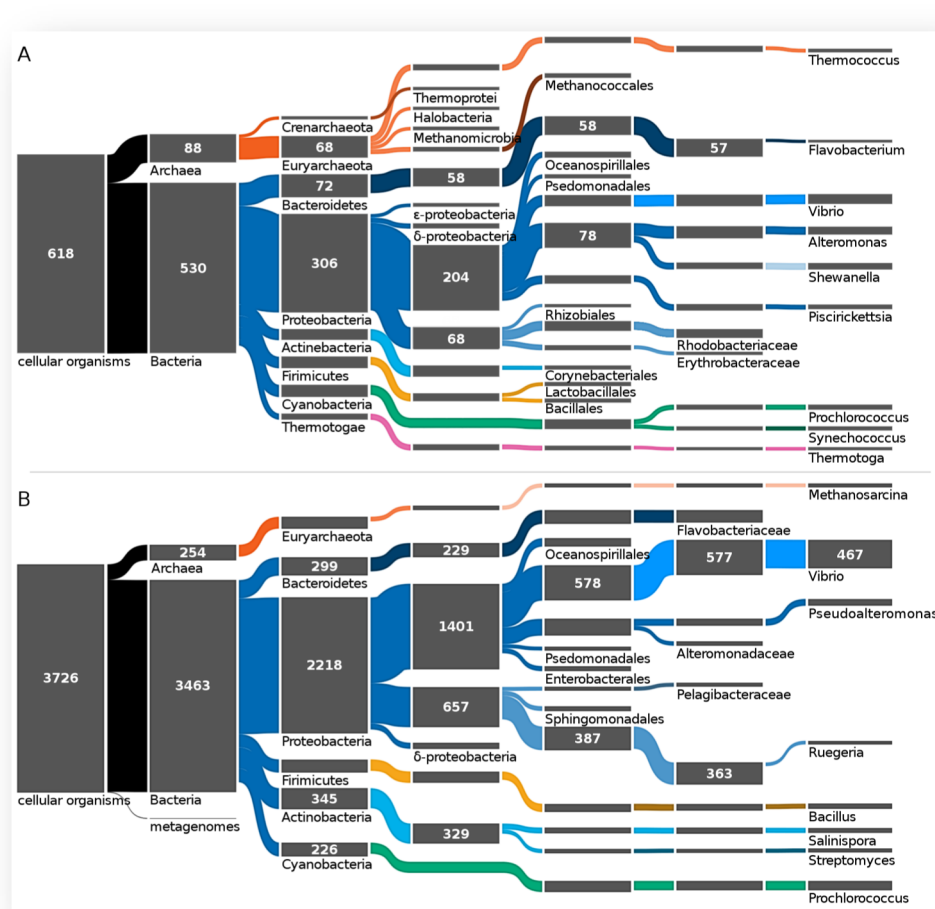
# MAR databases

## Taxonomic distribution

V1 - MarRef: 618, MarDB: 3726

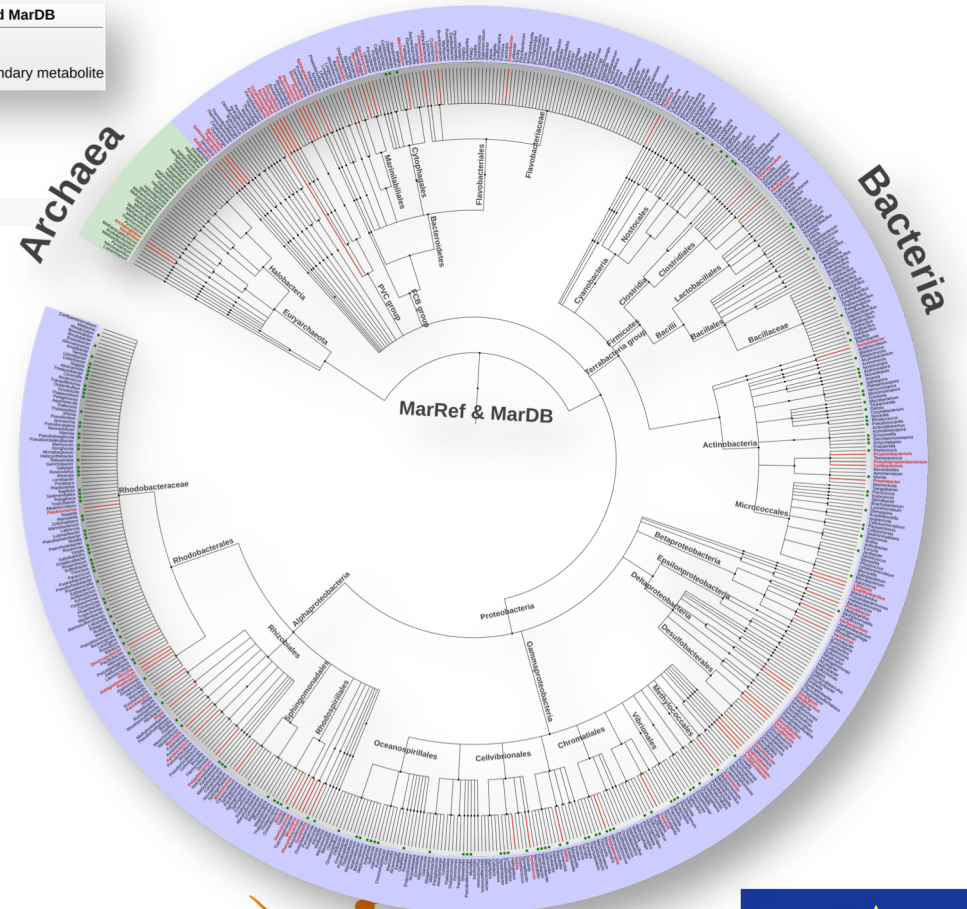
V2 - MarRef:771, MarDB: 8649

V3 - MarRef:852, MarDB: 10896



Genera of MarRef and MarDB

- MAG genome
- Synthesise secondary metabolite



# MAR databases

## Taxonomic distribution

JGI Gold database: 1608

JGI GOLD GENOMES ONLINE DATABASE

Home Search Distribution Graphs Biogeographical Metadata Statistics GOLD Usage Policy Team Help News

Studies	32,301
Biosamples	46,228
Sequencing Projects	197,643
Analysis Projects	155,882
Organisms	299,899

Your current search results are:

Studies	Biosamples	Organisms	Sequencing Projects	Analysis Projects
59	1	1,512	1,515	1,536

Current Filters:  
Organism.Domain → BACTERIAL  
Study.Ecosystem Type → Marine  
Analysis Project.Is Public → Yes

Refine Search Filters

Clear All Filters New Search

Secure https://gold.jgi.doe.gov/analysis\_projects?page=1&Study.Ecosystem+Type=Marine&entityFieldSubmit=Submit&Study.Ecosystem+Type=M

JGI GOLD GENOMES ONLINE DATABASE

Home Search Distribution Graphs Biogeographical Metadata Statistics GOLD Usage Policy Team Help News

Studies	32,301
Biosamples	46,228
Sequencing Projects	197,643
Analysis Projects	155,882
Organisms	299,899

Your current search results are:

Studies	Biosamples	Organisms	Sequencing Projects	Analysis Projects
10	0	96	97	98

Current Filters:  
Organism.Domain → ARCHAEL  
Study.Ecosystem Type → Marine  
Analysis Project.Is Public → Yes

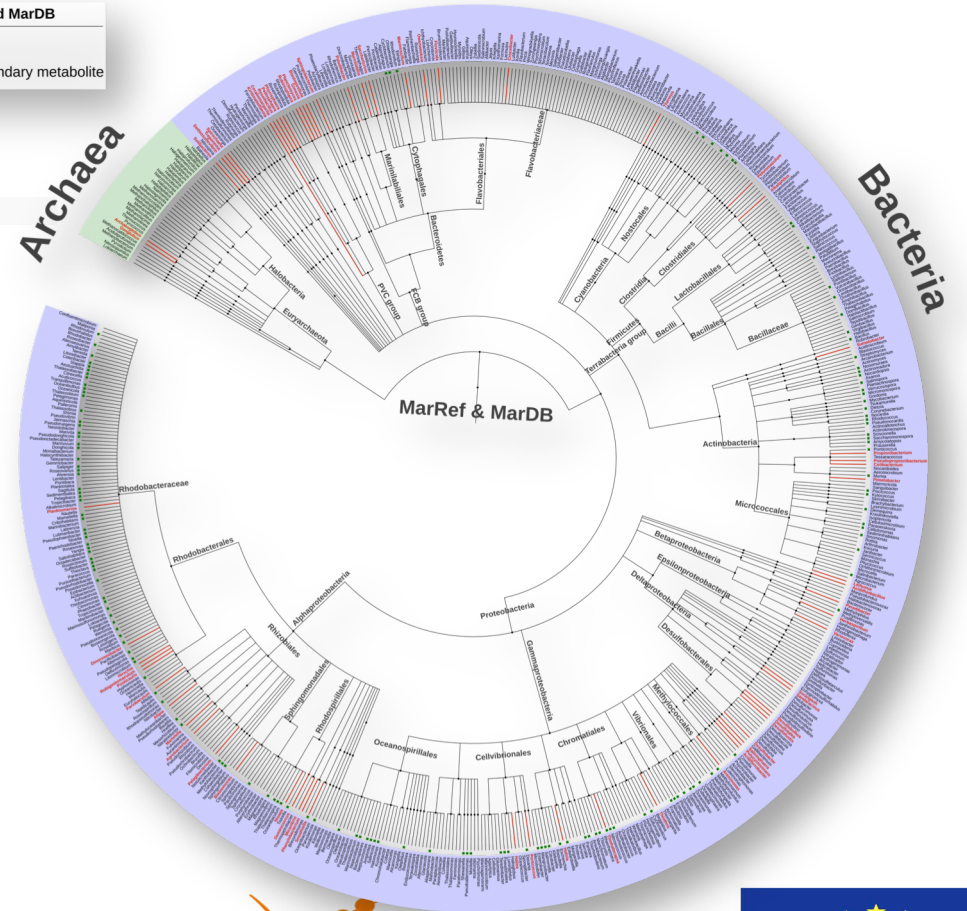
Refine Search Filters

Clear All Filters New Search

Version 2 - 9457

Genra of MarRef and MarDB

- MAG genome
- Synthesise secondary metabolite



# 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

eLife NORWAY Copyright © 2018 Center for Bioinformatics - UIT. [Terms and conditions.](#) eXcelerate





# MAR databases

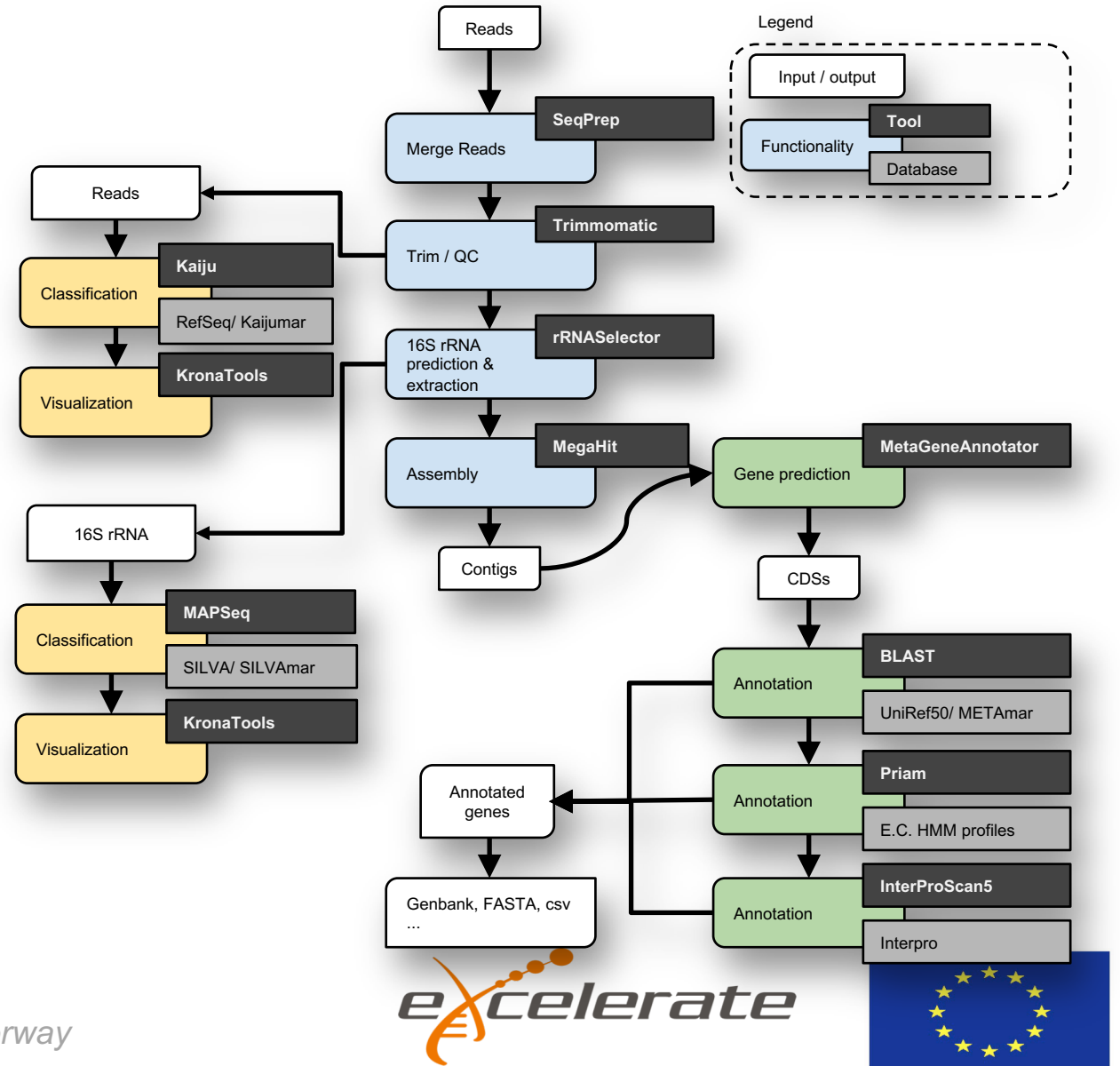
## Sequence databases

- **MarRef and MarDB**

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

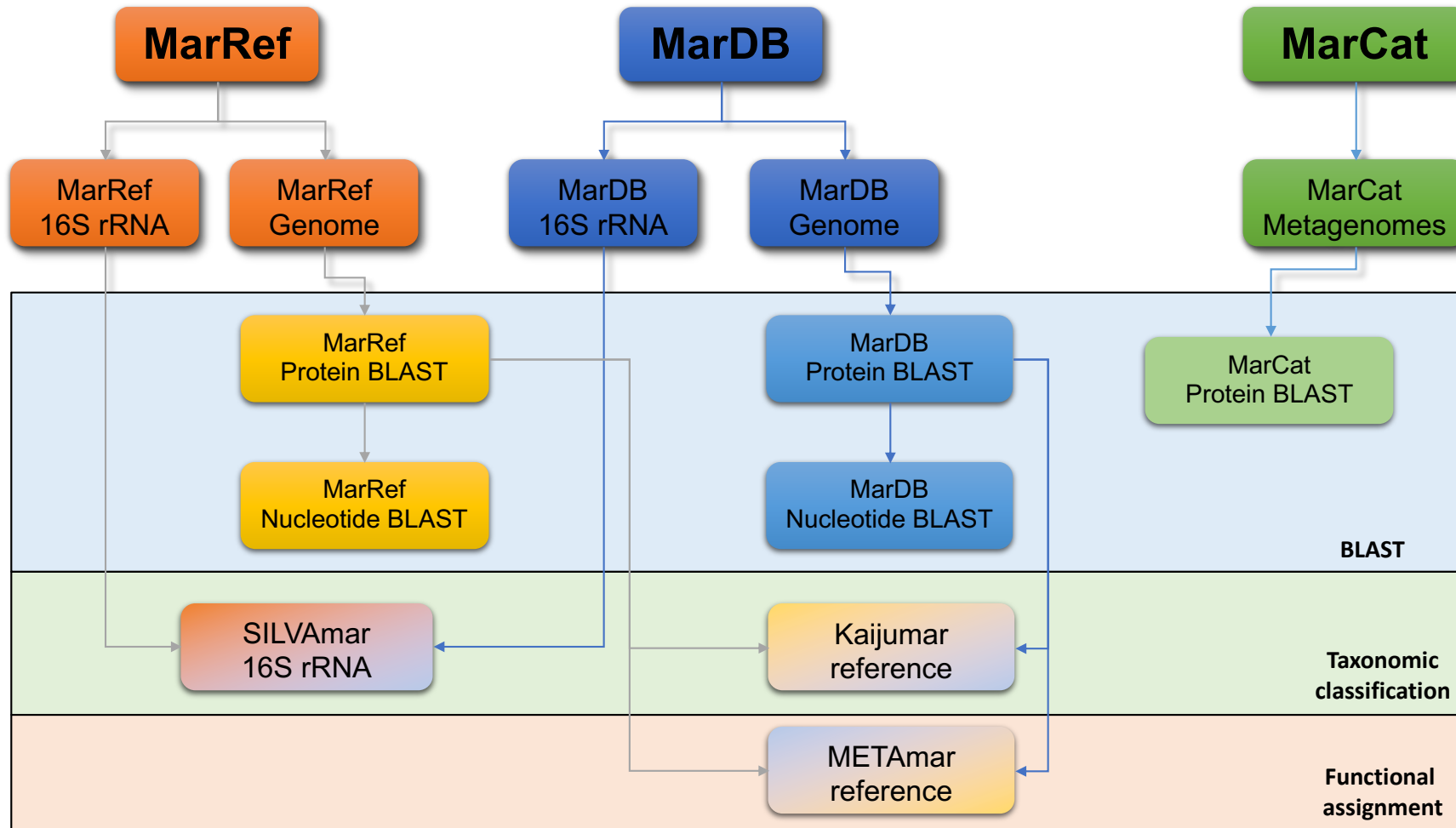
- **MarCat**

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



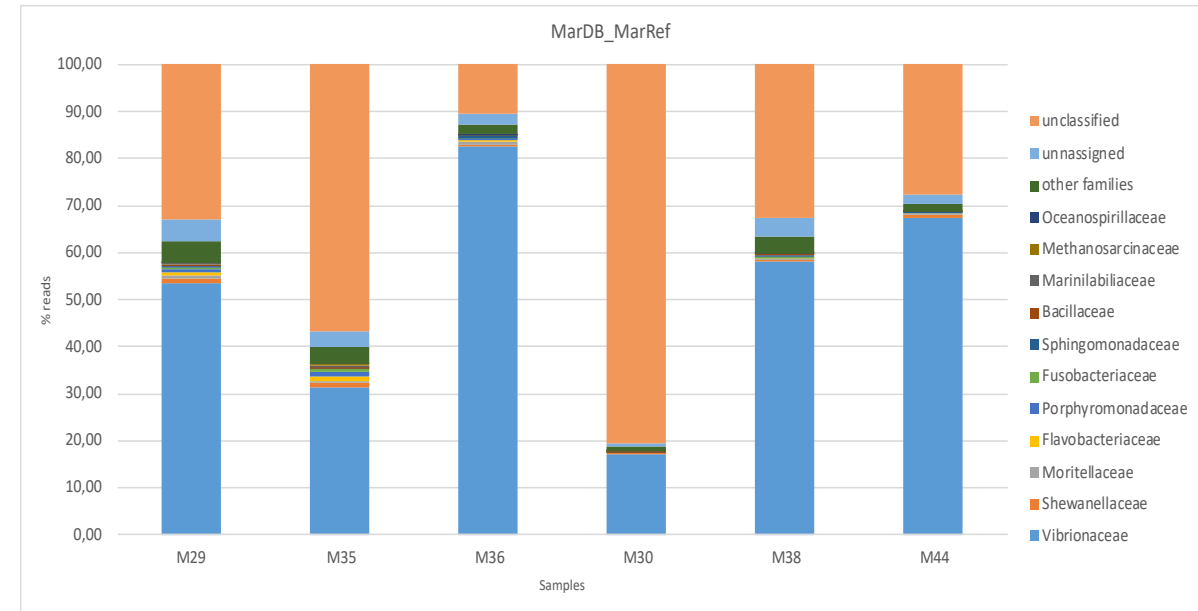
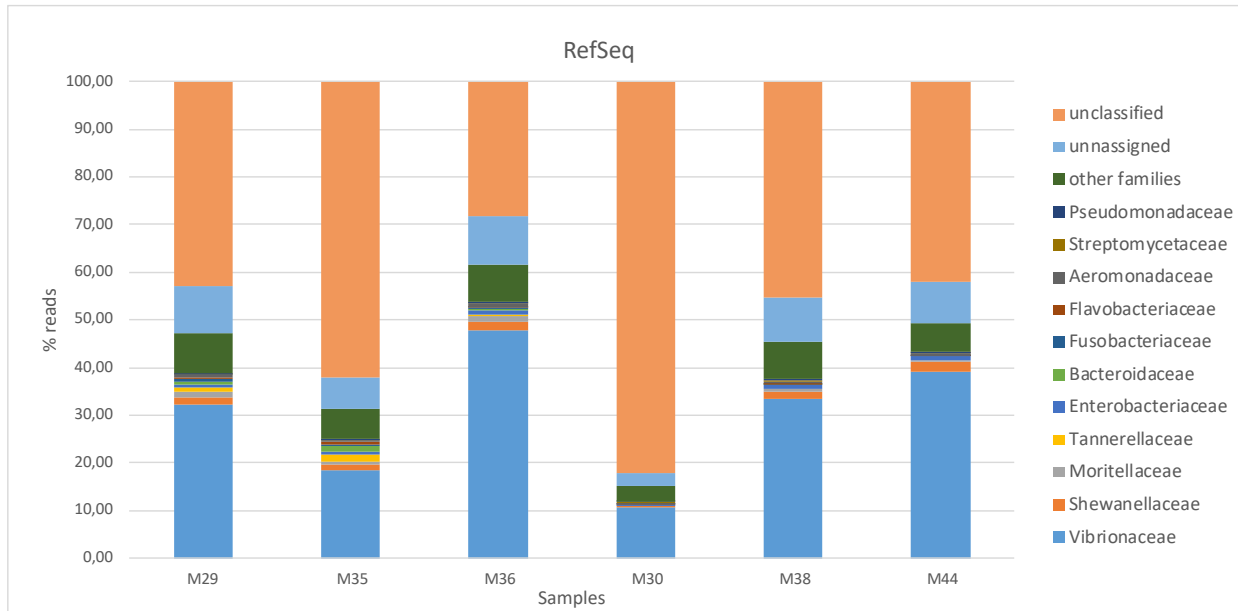
# MAR databases

## Sequence databases



# MAR databases

## Sequence databases



RefSeq: 52073 complete genomes

MarRef/DB: 9602 marine genomes incl. 4314 MAGs

Marine Metagenomics Workshop, Nov 26-30, 2018, Tromsø, Norway



# MAR databases

## BLAST

SequenceServer 1.0.9

Help & Support

```
>WP_012551601.1 catalase [Allivibrio salmonicida]
MSKLLTLAGCPFAHMQVQVAGKNGPQLQDVPFLRLAHPREVIPERRMIAKGGAGYQTFVTHDIT
KYTKAKIFSDIGKKTDMFAFSTVNGERGAADAERDINGSLKFTTECGNWDLAGNTPVFFLRDPLKFP
DLRHAVKRDPRTMRSANNNDFWTLSPALHQVTIVMSDRGIPATYRBMHGPGSHTFSEINSDNERVYV
KFHFVYQQIKNLSDAEAGELVGNDRSHQRDLLDSIDNQDFFKWTLVQIMPEADAATVPPYFDLTKV
WPHKDYPLIEVGEFELNRFQNYFAEVEQAFNPNVPGISFSDKMLQGRLEFAYDAQRYRLGVNHOH
IPVNAFPCPVHSHRDGAMRVDFNGSTLGYEPNDQQQWAEQDFSEPLNLDGAAAHDRHREDEDFSQ
PGDLFGLMTAEKQAILFDNFARNLNGVFKELQLRHVTHCYKADPAYGEGIGKLLGFDISEYNS]
```

Nucleotide databases

- MarDB CDS Nucleotides
- MarRef CDS Nucleotides

Protein databases

- MarCat CDS Proteins
- MarDB CDS Proteins
- MarRef CDS Proteins

Advanced Parameters:  ?

BLAST

BLAST against a custom, local database with SequenceServer. [Tweet](#)

Please cite: Priyam et al. (2015) Sequenceserver: a modern graphical user interface for custom BLAST databases & relevant data sources.

Query= WP\_012551601.1 catalase [Allivibrio salmonicida] 1 / 1

BLASTP: 1 query, 1 database

Query= WP\_012551601.1

Download FASTA, XML, TSV

FASTA of all hits

FASTA of selected hit(s)

Standard tabular report

Full tabular report

Full XML report

View More

Number	Sequences producing significant alignments	Total score	E value	Length
1.	MMP491463_23010	961.06	0.00	483
2.	MMP491463_108689	583.56	0.00	481
3.	MMP490065_67230	577.01	0.00	497
4.	MMP494431_93293	557.37	0.00	491
5.	MMP490065_161240	553.90	0.00	491
6.	MMP494431_35969	551.59	0.00	496
7.	MMP490065_142530	528.48	0.00	500
8.	MMP494431_182909	528.48	0.00	500
9.	MMP492357_229215	528.48	0.00	500

# MAR databases

## BLAST



The BLAST (Basic Local Alignment Search Tool) finds local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Help

Click here to open the BLAST page in a new tab

▼ MMP491463\_23010 [UniRef50=UniRef50\_A0A1J0LRC4;Cluster: Catalase] [Interpro=IPR002226;Catalase haem-binding site] 1 / 51  
[Pfam=1.111.6] [mmp\_id=MMP491463] [mmp\_db=marcat]

Hit length: 483  Select |  Sequence |  FASTA |  MMP |  UniRef50 |  Interpro |  Brenda

1. Score	E value	Identities	Gaps	Positives
961.06 (2483)	0.00	448/483 (92.75)	0/483 (0.00)	472/483 (97.72)

Query 1 MSKLLTAAGCPVAHNGVQTAGKRGFQLLDVWFLEKLAHFDREIVERRMHAAGSGAY 60  
Subject 1 MSKLLTAAGCPVAHNGVQTAGKRGFQLLDVWFLEKLAHFDREIVERRMHAAGSGAY 60

Query 61 GFTVTHDITKTKAKIFSDIGKKTDFARFSTVAGERGAADAERDIRGFLKPYTEEGN 120  
Subject 61 GFTVTHDITKTKAKIFSDIGKKTDFARFSTVAGERGAADAERDIRGFLKPYTEEGN 120

Query 121 WDLAGNPPVFFLRDLKFPDLNHAVKRDPRTMRSAKNNWDFWTSLEALHQ+IIVMSD 180  
Subject 121 WDMVGNPPVFFLRDLKFPDLNHAVKRDPRTMRSAKNNWDFWTSLEALHQ+IIVMSD 180

Query 181 RGIPTATYRHMGFGSHTFSFINSDNERVWVHFHFSQQGKLNLSDAEAGELVGDRESHQ 240  
Subject 181 RGIPTATYRHMGFGSHTFSFINSDNERVWVHFHFSQQGKLNLSDAEAAQVIGDRESHQ 240

Query 241 RDLDSIDNQDFPKWTLKQIPEADAATVYPNFFDLKVPWPKDYPLLEVGEFELNRP 300  
Subject 241 RDLDSIDNQDFPKWTLKQIPEADAATVYPNFFDLKVPWPKDYPLLEVGEFELNRP 300

Query 301 QNYFAEVEQAAFNPNVPGISFSDPKMLQGRIFAYGDAQRYRLGVNHHIIPVNAAPCPV 360  
Subject 301 QNYFAEVEQAAFNPNVPGISFSDPKMLQGRIFAYGDAQRYRLGVNHHIIPVNAAPCPV 360

Query 361 HSYHRDGMRVGDFGSLGYEFPNDCQWAEQDFPSPFPFLNLDGAAAHWDHREDDYFSQ 420  
Subject 361 HSYHRDGMRVGDFGSLGYEFPNDCQWAEQDFPSPFPFLNLDGAAAHWDHREDDYFSQ 420

Query 421 PGDLFGLMTAEQALFDNTARNLGVKPEIQLRHVTHCYKADPAVGEIGKLLGPDISE 480  
Subject 421 PGDLFRLMTEPKQALFDNTARNLGVKPEIQLRHLRHCYKADPAVGEIGKLLGPDISE 480

Query 481 YNS 483  
Subject 481 FNS 483

▼ MMP491463\_108689 [UniRef50=UniRef50\_P42321;Cluster: Catalase] [Interpro=IPR002226;Catalase haem-binding site] 2 / 51  
[Pfam=1.111.6] [mmp\_id=MMP491463] [mmp\_db=marcat]



# MAR databases

## Marine reference databases

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

The screenshot shows the homepage of the Marine Reference Databases. At the top is a navigation bar with links for SERVICES, DOCUMENTATION, COMMUNITY, HELP, and CONTACT. Below this is a main header with the MMP logo and the text: "The marine reference databases are a collection of richly annotated and manually curated contextual (metadata) and sequence databases. The contextual data can be accessed by browsing, searching or filtering, while the sequence data through BLAST. All data can be downloaded." A "Help" button is also present. The main content area features three database cards: MARREF, MARDB, and MARCAT. Each card includes a brief description, a "Browse" button, and "BLAST" and "Download" buttons. The footer contains logos for eLife, Accelerate, and the European Union, along with copyright information and contact details.

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

Terje Klemetsen<sup>1</sup>, Inge A. Raknes<sup>1</sup>, Juan Fu<sup>1</sup>, Alexander Agafonov<sup>1</sup>, Sudhagar V. Balasundaram<sup>1</sup>, Giacomo Tartari<sup>1,2</sup>, Espen Robertsen<sup>1</sup> and Nils P. Willassen<sup>1\*</sup>

<sup>1</sup>Centre for Bioinformatics, Faculty of science and technology, UIT The Arctic University of Norway, PO Box 6050 Langnes, TromsøN-9037, Norway and <sup>2</sup>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<sup>4</sup> to 10<sup>6</sup> cells per milliliter seawater and totally over 10<sup>29</sup> 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

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Accessible from the Marine Metagenomics Portal (MMP)

<https://mmp.sfb.uit.no/>

Marine Metagenomics Workshop, Nov 26-30, 2018, Tromsø, Norway





**MMP- team:**

Terje Klementsén

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](mailto:mmp@uit.no)



[www.elixir-europe.org](http://www.elixir-europe.org)

