Metadata and MIxS* Checklists

*Minimal Information about any (x) Sequence

Jimena Linares MSc.
GFBio e.V.
jlinares@gfbio.org
@mia_lgo

Dr. Ivaylo Kostadinov
GFBio e.V.
ikostadi@gfbio.org
@tigroumaniac

www.gfbio.org
@GFBio_Project

State of art: Working with molecular data: Is the contextual data acknowledged?

Very Minimum expectations

The contextual data can and should be included in the databases!

- Contextual data is not integrated to the molecular data inside taxonomic databases (*e.g.* NCBI).
- Non-standardized collection of contextual data.
- Public databases (such as the ISNDC) depend on author-submitted information.
- Establishment of standardized procedures is required.

Is the contextual data acknowledged?

Comment | Open Access | Published: 19 June 2020

COVID-19 pandemic reveals the peril of ignoring metadata standards

Lynn M. Schriml ✉, Maria Chuvochina, Neil Davies, Emiley A. Eloe-Fadrosh, Robert D. Finn, Philip Hugenholtz, Christopher I. Hunter, Bonnie L. Hurwitz, Nikos C. Kyrpides, Folker Meyer, Ilene Karsch Mizrachi, Susanna-Assunta Sansone, Granger Sutton, Scott Tighe & Ramona Walls

Scientific Data **7**, Article number: 188 (2020) | Cite this article

7214 Accesses | **2** Citations | **216** Altmetric | Metrics

This article has been updated

Efficient response to the pandemic through the mobilization of the larger scientific community is challenged by the limited reusability of the available primary genomic data. Here, the Genomic Standards Consortium board highlights the essential need for contextual genomic data FAIRness, for empowering key data-driven biological questions.

A research program at the University of Oxford, "Our World in Data", maintains a global database on testing for COVID-19. Asked whether there are 'low-hanging fruit' to improve the response to the pandemic, Program Director Max Roser had a very simple answer: *"for all those who publish original data, provide a clear description of your data" (@MaxCRoser:* 1:39am · 12 Apr 2020 · Twitter Web App), highlighting the importance of maximizing the reusability of data. In the age of COVID-19, we are seeing where value really lies. Describing the WHO, WHAT, HOW, WHERE, and WHEN of genomic data enables comparative analysis, informs

Is the contextual data acknowledged?

From: COVID-19 pandemic reveals the peril of ignoring metadata standards

a

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: SAMN14751340; Sample name: WA-UW-6185; SRA: SRS6545124	
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus	
Package	Pathogen: clinical or host-associated; version 1.0	
Attributes	**isolate**	WA-UW-6185
	collected by	University of Washington Virology Lab
	collection date	missing
	geographic location	USA
	host	Homo sapiens
	host disease	COVID-19
	isolation source	missing
	latitude and longitude	missing
Submission	University of Washington, Pavitra Roychoudhury; 2020-04-27	

b

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: SAMN14656632; Sample name: hCoV-19/USA/WI-176/2020; SRA: SRS6514341	
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus	
Package	Pathogen: clinical or host-associated; version 1.0	
Attributes	**strain**	hCoV-19/USA/WI-176/2020
	isolate	Homo sapien
	collected by	Milwaukee Public Health Department
	collection date	2020-03-20
	geographic location	USA: Milwaukee, Wisconsin
	host	Homo sapiens
	host disease	COVID-19
	isolation source	nasal swab
	latitude and longitude	43.042180 N 87.908670 W
	ARTIC barcode identifiers	NB23

Lost opportunities for data reuse, SARS-CoV-2 (txid2697049[Organism:noexp]) BioSample records, where (**a**) **collection date** = "missing": 143; **latitude and longitude** = "missing": 1375; (**b**) SARS-CoV-2 BioSample record with complete metadata.

Is the contextual data acknowledged?

From: COVID-19 pandemic reveals the peril of ignoring metadata standards

a

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: SAMN14751340; Sample name: WA-UW-6185; SRA: SRS6545124
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus
Package	Pathogen: clinical or host-associated; version 1.0

Attributes

isolate	WA-UW-6185
collected by	University of Washington Virology Lab
collection date	missing
geographic location	USA
host	Homo sapiens
host disease	COVID-19
isolation source	missing
latitude and longitude	missing

Submission: University of Washington, Pavitra Roychoudhury; 2020-04-27

b

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: SAMN14656632; Sample name: hCoV-19/USA/WI-176/2020; SRA: SRS6514341
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus
Package	Pathogen: clinical or host-associated; version 1.0

Attributes

strain	hCoV-19/USA/WI-176/2020
isolate	Homo sapien
collected by	Milwaukee Public Health Department
collection date	2020-03-20
geographic location	USA: Milwaukee, Wisconsin
host	Homo sapiens
host disease	COVID-19
isolation source	nasal swab
latitude and longitude	43.042180 N 87.908670 W
ARTIC barcode identifiers	NB23

Lost opportunities for data reuse, SARS-CoV-2 (txid2697049[Organism:noexp]) BioSample records, where (**a**) **collection date** = "missing": 143; **latitude and longitude** = "missing": 1375; (**b**) SARS-CoV-2 BioSample record with complete metadata.

MIxS

Minimum **I**nformation about any (**x**) **S**equence

developed by:

genomic STANDARDS consortium

supported by:

INSDC International Nucleotide Sequence Database Collaboration

Available in: http://gensc.org/mixs/

nature biotechnology

Explore our content | Journal information | Publish with us

nature › nature biotechnology › perspectives › article

Published: 06 May 2011

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

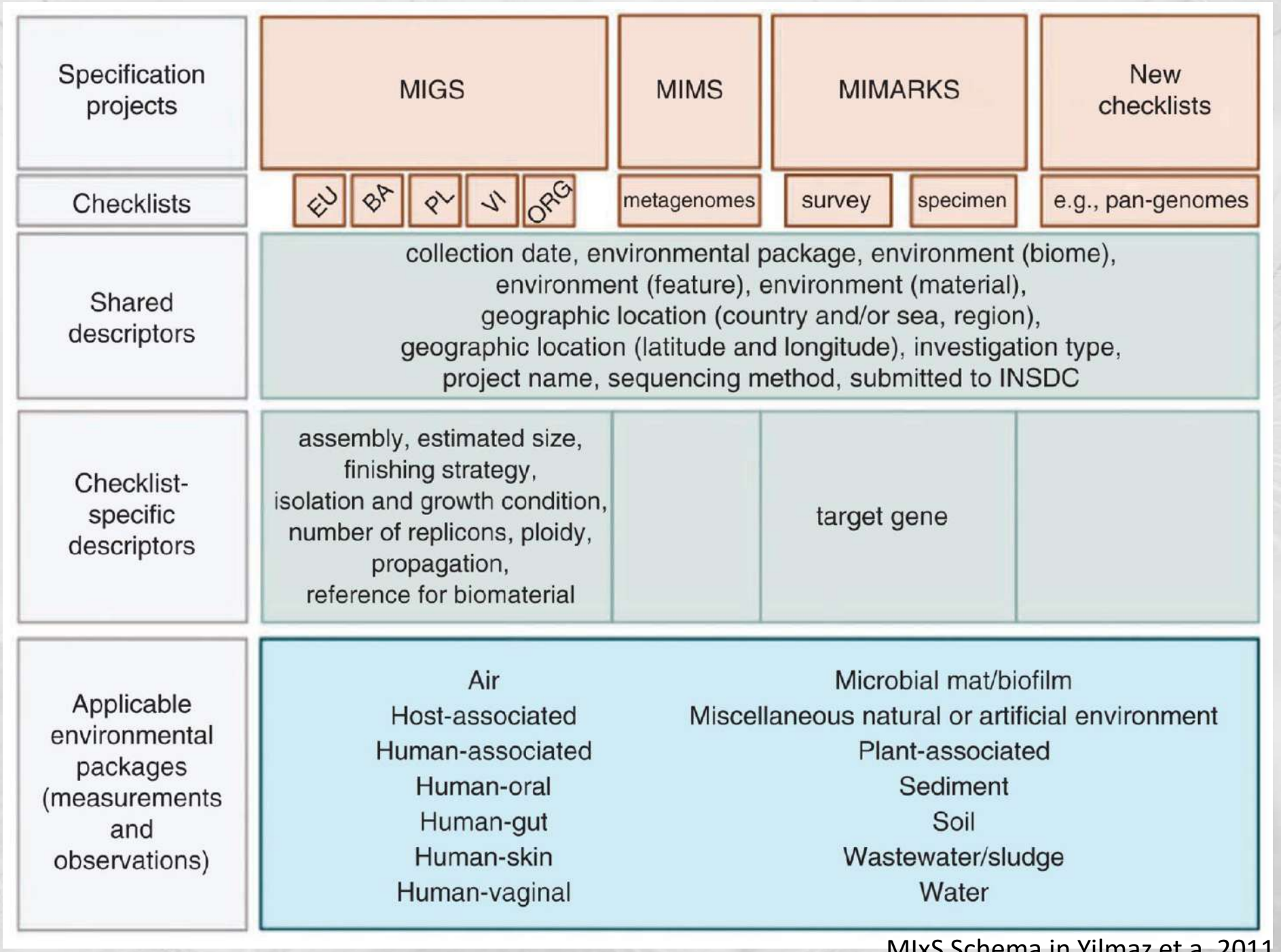
Pelin Yilmaz, Renzo Kottmann, [...] Frank Oliver Glöckner

Nature Biotechnology **29**, 415–420(2011) | Cite this article

2021 Accesses | **359** Citations | **35** Altmetric | Metrics

Abstract

Here we present a standard developed by the Genomic Standards Consortium (GSC) for reporting marker gene sequences—the minimum information about a marker gene sequence (MIMARKS). We also introduce a system for describing the environment from which a biological sample originates. The 'environmental packages' apply to any genome sequence of known origin and can be used in combination with MIMARKS and other GSC checklists. Finally, to establish a unified standard for describing sequence data and to provide a single point of entry for the scientific community to access and learn about GSC checklists, we present the minimum information about any (x) sequence (MIxS). Adoption of MIxS will enhance our ability to analyze natural genetic diversity documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

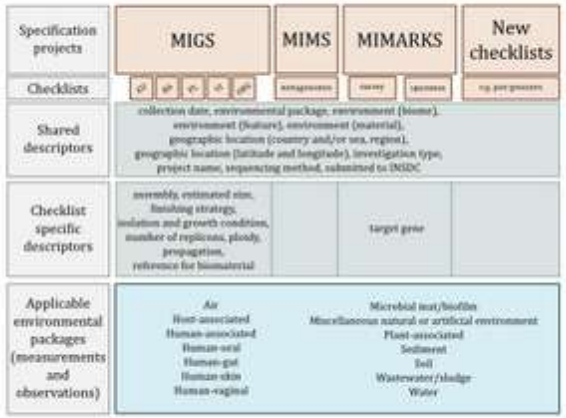


MIxS Schema in Yilmaz et a. 2011

Download MIxS Checklist v5.0: mixs_v5

or from the MIxS GitHub here

Download MIxS Checklist v4.0 here



The core MIxS team developed the following environmental packages These packages are available as separate spreadsheets:

1. Air (download MIxS-air only)
2. Built-environment (download MIxS-built environment only)
3. Host-associated (download MIxS-host associated only)
4. Human-associated (download MIxS-human associated only)
5. Human-gut (download MIxS-human gut only)
6. Human-oral (download MIxS-human oral only)
7. Human-skin (download MIxS-human skin only)
8. Human-vaginal (download MIxS-human vaginal only)
9. Microbial mat/biofilm (download MIxS-microbial mat/biofilm only)
10. Miscellaneous natural or artificial environment (download MIxS-misc. natural or artificial environment only)
11. Plant-associated (download MIxS-plant associated only)
12. Sediment (download MIxS-sediment only)
13. Soil (download MIxS-soil only)
14. Wastewater/sludge (download MIxS-wastewater/sludge only)
15. Water (download MIxS-water only)
16. Hydrocarbon resources-cores (download from GitHub)
17. Hydrocarbon resources-fluids/swabs (download from GitHub)

https://press3.mcs.anl.gov/gensc/mixs/

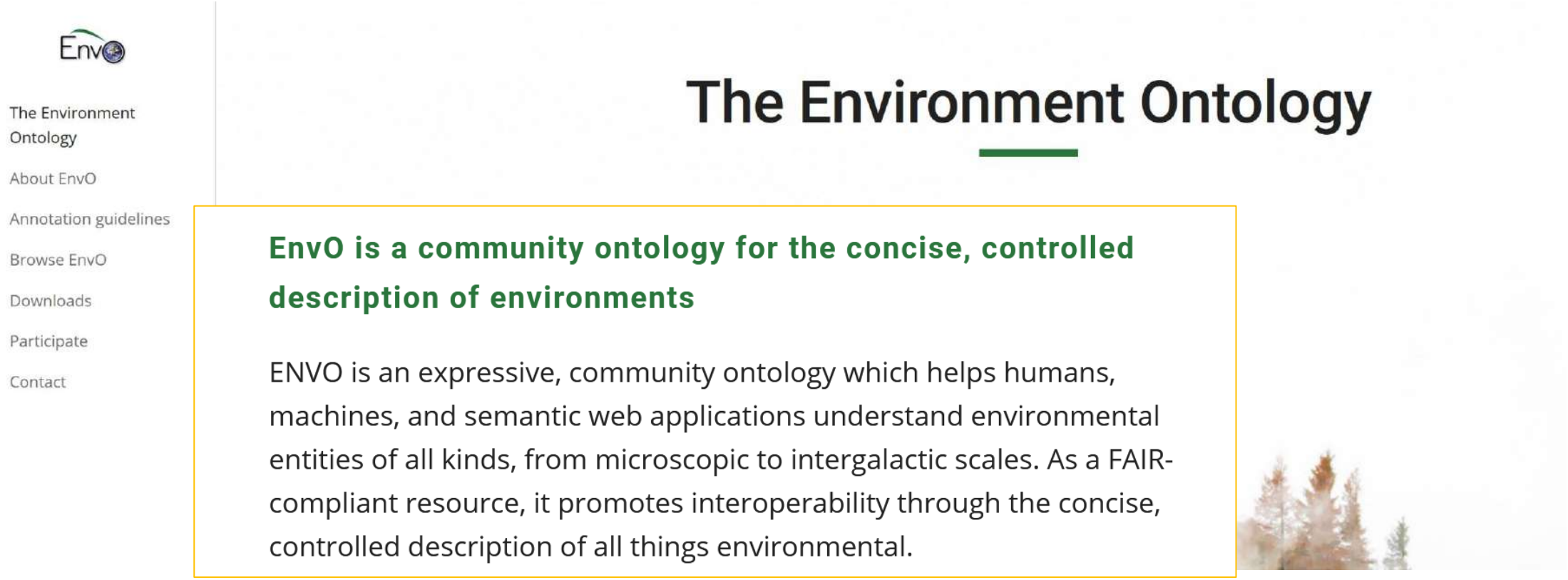
MIxS - Water Checklist

1	Structured comment name	Item	Definition	Example	Expected value	Section	eu	ba	pl	vi	org	me	MIMARKS Survey	MIMARKS Specimen	Value syntax
4	project_name	project name	Name of the project within which the sequencing was organized			investigation	M	M	M	M	M	M	M	M	{text}
5	experimental_factor	experimental factor	Experimental factors are essentially the variable aspects of an experiment design which can be used to describe an experiment, or set of experiments, in an increasingly detailed manner. This field accepts ontology terms from Experimental Factor Ontology (EFO) and/or Ontology for Biomedical Investigations (OBI). For a browser of EFO (v 2.43) terms, please see http://purl.bioontology.org/ontology/EFO; for a browser of OBI (v 2013-10-25) terms please see http://purl.bioontology.org/ontology/OBI		text or EFO and/or OBI	investigation	X	X	X	X	X	C	C	X	{term|text}
6	lat_lon	geographic location (latitude and longitude)	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system		decimal degrees	environment	M	M	M	M	M	M	M	M	{float} {float}
7	depth	geographic location (depth)	Please refer to the definitions of depth in the environmental packages		-	environment	E	E	E	E	E	E	E	E	-
8	alt_elev	geographic location (altitude/elevation)	Please refer to the definitions of either altitude or elevation in the environmental packages		-	environment	E	E	E	E	E	E	E	E	-
9	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 (http://insdc.org/country.html), or the GAZ ontology (v 1.512) (http://purl.bioontology.org/ontology/GAZ)	Germany:Sylt:Haustrand	country or sea name (INSDC or GAZ):region(GAZ):specific location name	environment	M	M	M	M	M	M	M	M	{term}:{term}:{text}
10	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 i.e. all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; Except: 2008-01; 2008 all are ISO8601 compliant		date and time	environment	M	M	M	M	M	M	M	M	{timestamp}
11	env_biome	environment (biome)	Biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Biome should be treated as the descriptor of the broad ecological context of a sample. Examples include: desert, taiga, deciduous woodland, or coral reef. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO		EnvO	environment	M	M	M	M	M	M	M	M	{term}
12	env_feature	environment (feature)	Environmental feature level includes geographic environmental features. Compared to biome, feature is a descriptor of the more local environment. Examples include: harbor, cliff, or lake. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO		EnvO	environment	M	M	M	M	M	M	M	M	{term}
13	env_material	environment (material)	The environmental material level refers to the material that was displaced by the sample, or material in which a sample was embedded, prior to the sampling event. Environmental material terms are generally mass nouns. Examples include: air, soil, or water. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO		EnvO	environment	M	M	M	M	M	M	M	M	{term}
			MIGS/MIMS/MIMARKS 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												associated|human-skin|human-oral|human-gut|human-vaginal|microbial mat/biofilm|misenvironment|plant-

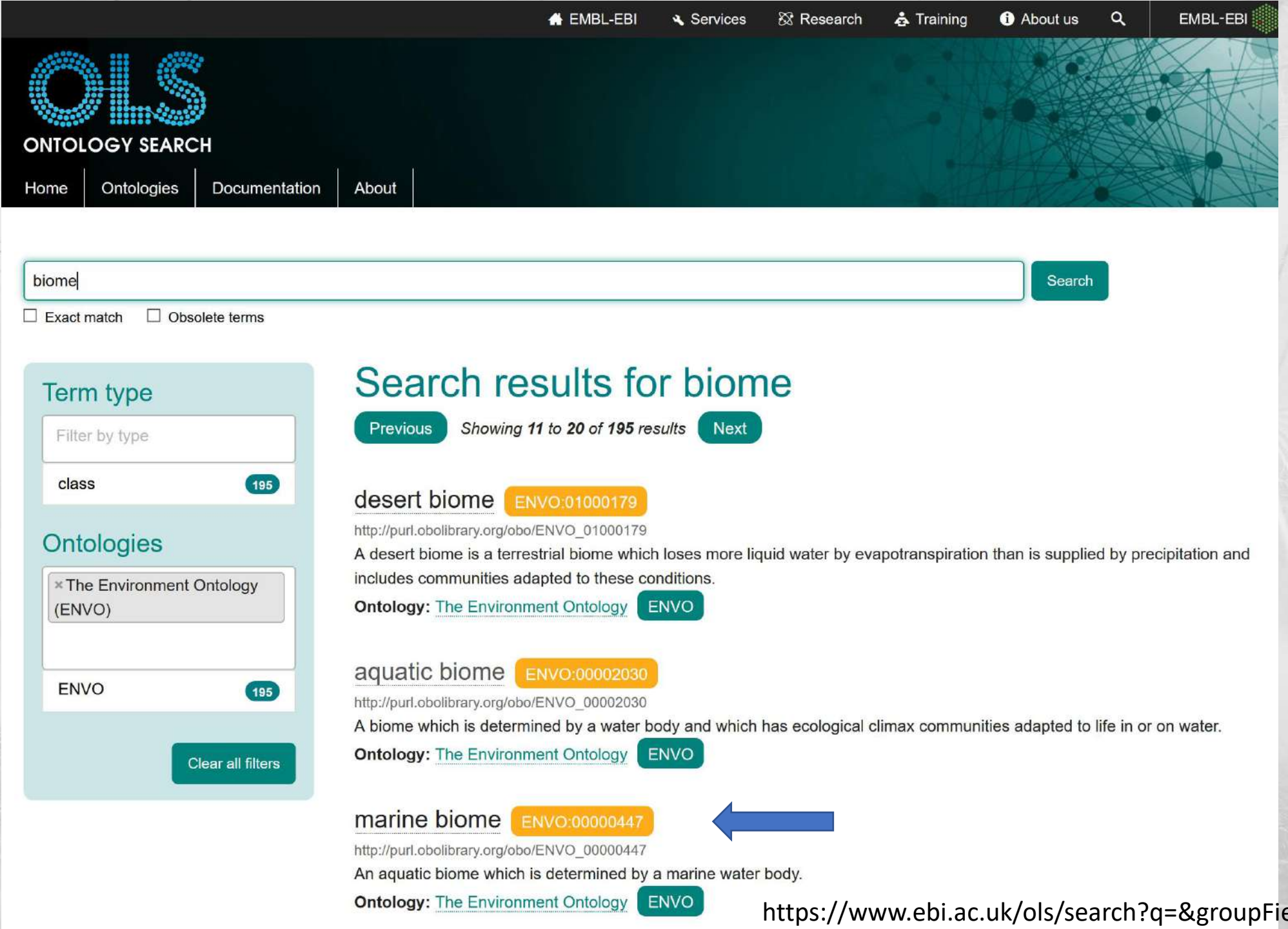
README | MIxS-water

https://press3.mcs.anl.gov/gensc/mixs/

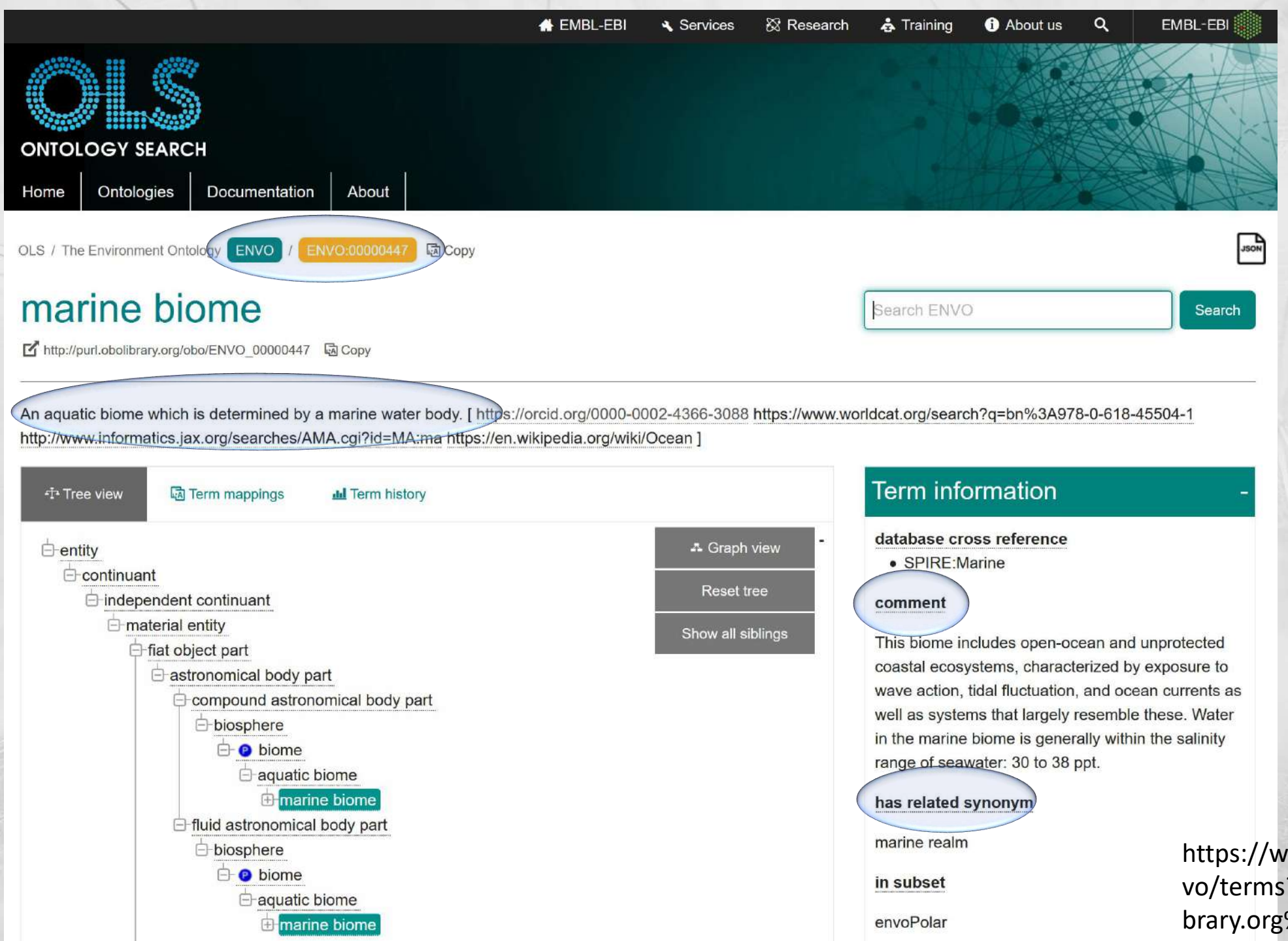
EnvO – The Environment Ontology



https://sites.google.com/site/environmentontology/



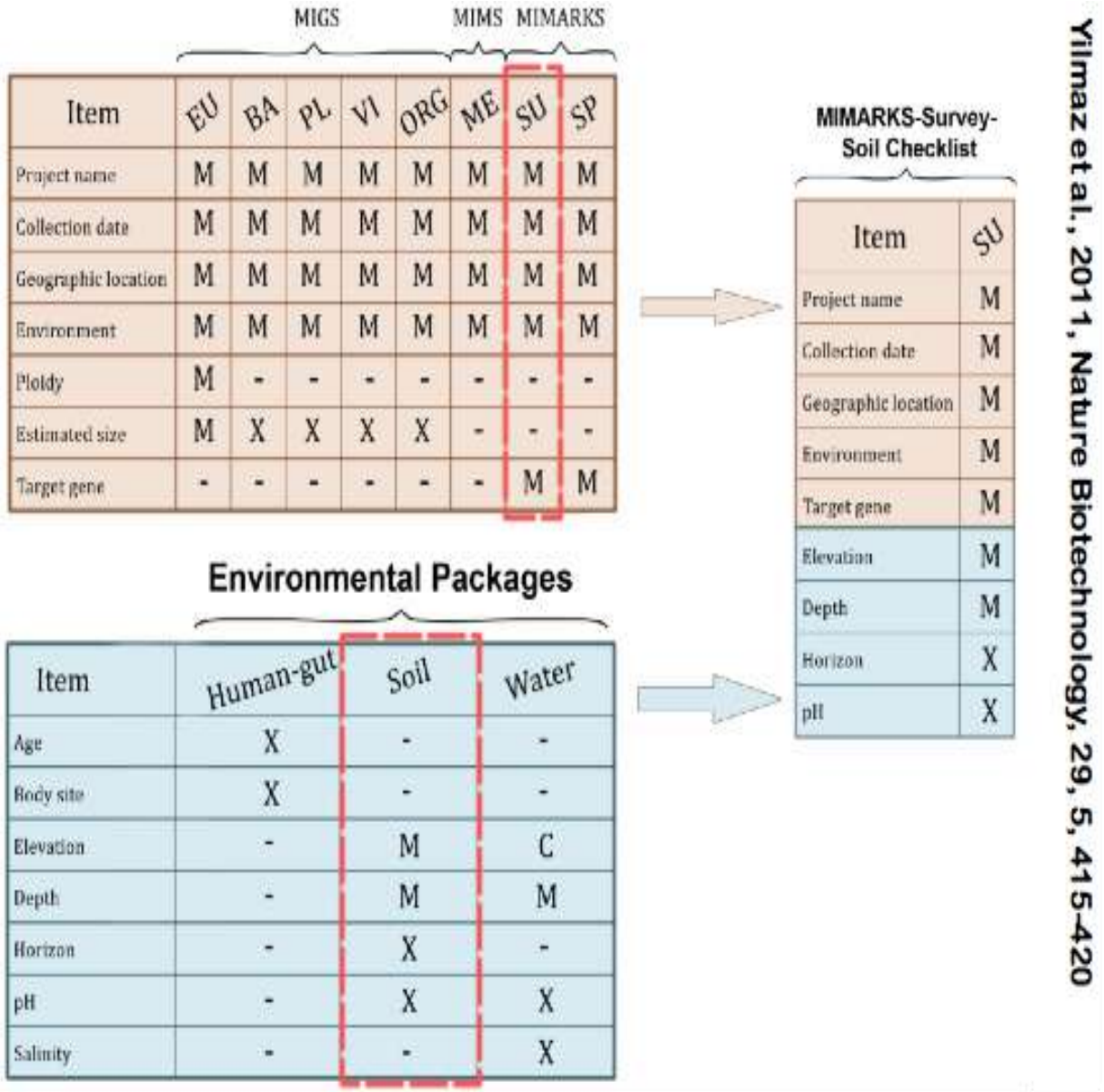
https://www.ebi.ac.uk/ols/search?q=&groupField=iri&start=0&ontology=envo



https://www.ebi.ac.uk/ols/ontologies/envo/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FENVO_00000447

EnvO – The Environment Ontology

- biome - e.g. ***concentration basin mediterranean sea biome***
- environmental feature - e.g. ***circalitoral zone***
- environmental material - e.g. ***coastal sea water***



MIGS MIMS MIMARKS

Item	EU	BA	PL	VI	ORG	ME	SU	SP
Project name	M	M	M	M	M	M	M	M
Collection date	M	M	M	M	M	M	M	M
Geographic location	M	M	M	M	M	M	M	M
Environment	M	M	M	M	M	M	M	M
Ploidy	M	-	-	-	-	-	-	-
Estimated size	M	X	X	X	X	-	-	-
Target gene	-	-	-	-	-	-	M	M

MIMARKS-Survey-Soil Checklist

Item	SU
Project name	M
Collection date	M
Geographic location	M
Environment	M
Target gene	M
Elevation	M
Depth	M
Horizon	X
pH	X

Environmental Packages

Item	Human-gut	Soil	Water
Age	X	-	-
Body site	X	-	-
Elevation	-	M	C
Depth	-	M	M
Horizon	-	X	-
pH	-	X	X
Salinity	-	-	X

Yilmaz et al., 2011, Nature Biotechnology, 29, 5, 415-420

EMBL-EBI

BioSamples

Start your search

Examples: leukaemia, virid

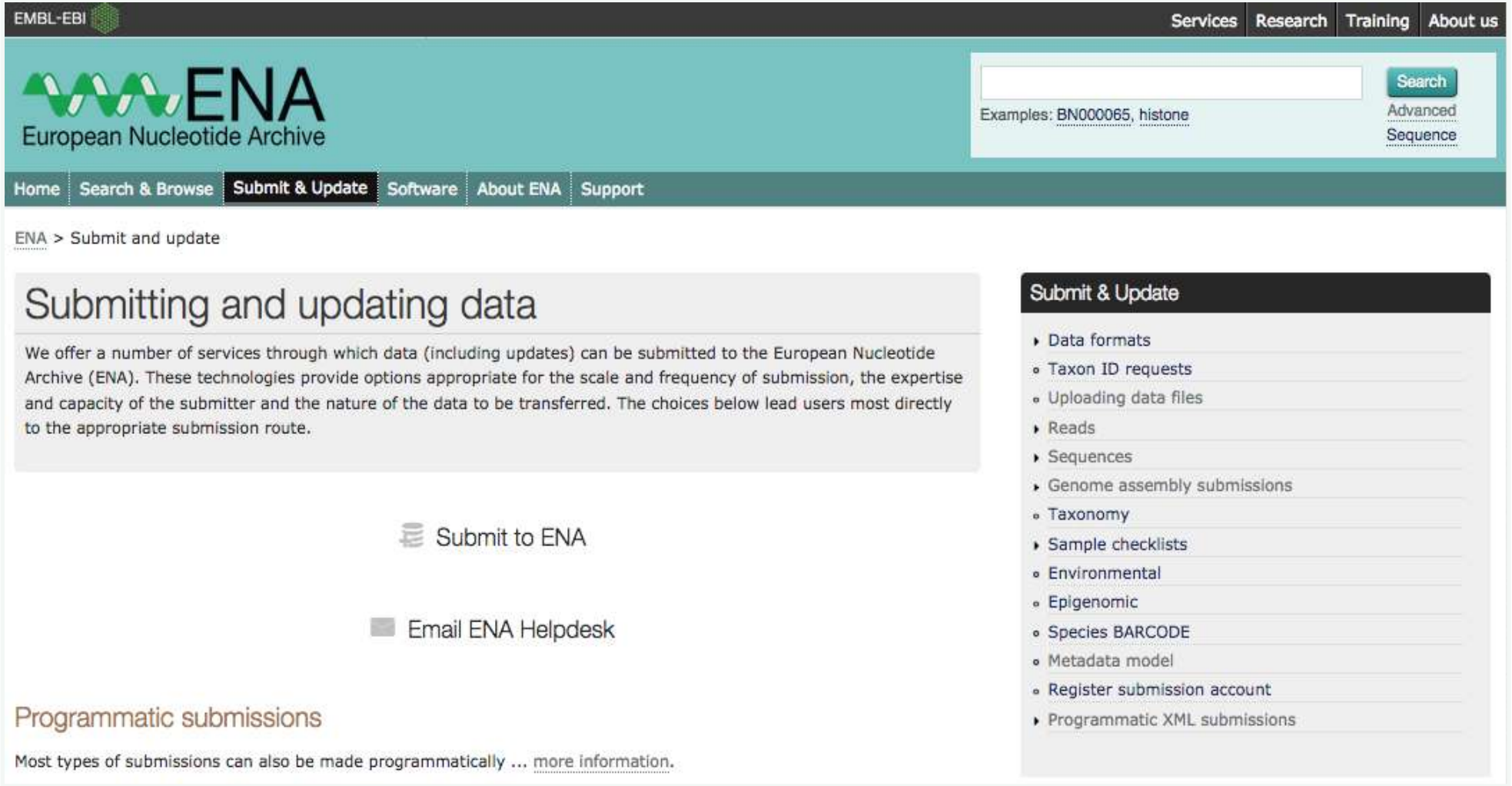
Home | Search | Submit | Help | About BioSamples

BioSamples > SAMEA104172529

Sample SAMEA104172529

Name	ERS1831547
Release date	2017-08-31
Last updated	2017-08-31
Submission title	121 marine metagenome samples from ENA SRA
Submission identifier	GEN-ERA979722
Organism	marine metagenome
Amount or size of sample collected	0.005-0.01 (L)
Colection date	2014-04-26T10:00:00/2014-05-01T15:30:00
Ena first public	2017-08-08
Ena last update	2017-08-30
Environment biome	concentration basin mediterranean sea biome [ENVO:01000004]
Environment feature	circalittoral zone [ENVO:01000412]|coastal water body [ENVO:02000049]
Environment material	coastal sea water [ENVO:00002150]|oligotrophic water [ENVO:00002223]
Geographic location country and or sea	France
Geographic location depth	10-12 (m)
Investigation type	mimarks-survey
Latitude	43.402301 (DD)
Longitude	7.195801 (DD)
Observed biotic relationship	free living
Pcr primers	28F: 5'-GAGTTTGATCNTGGCTCAG-3', 519R: 5'-GTNTTACNGCGGCKGCTG-3'
Project name	IN-EX
Sample collection device or method	Insitu Incubation
Sample material processing	filtering of seawater, DNA extraction
Sequencing method	Illumina MiSeq
Submitted to insdc	true
Synonym	1131:6d930886-7aaa-4eef-a260-6c6b426e4d50 130 Tf
Target gene	16S rRNA
Water environmental package	water
External references	ERS1831547

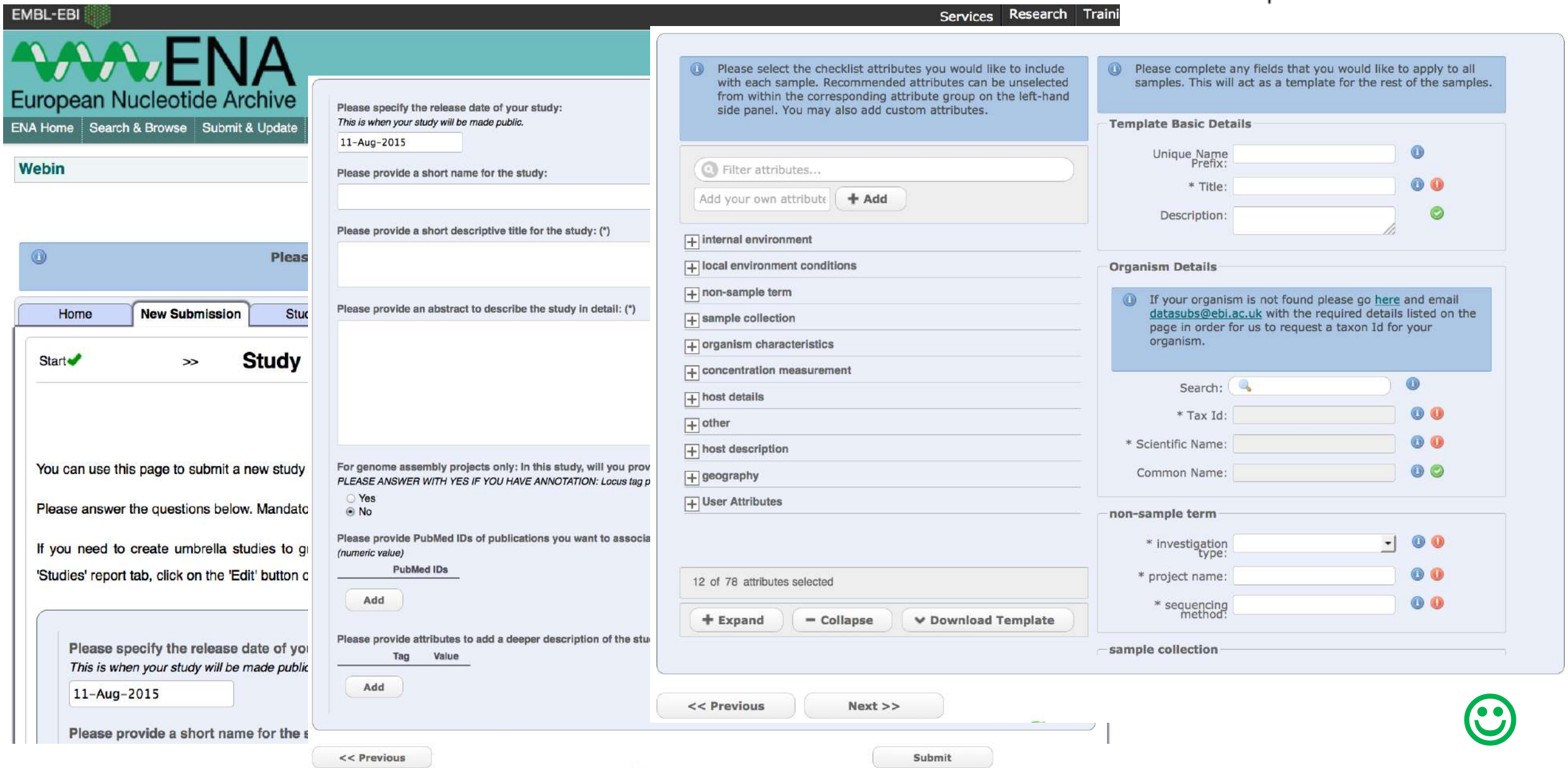
MIXS standard in EBI-ENA



https://www.ebi.ac.uk/ena/submit

EBI-ENA

ENA European Nucleotide Archive



https://www.ebi.ac.uk/ena/submit/sra/#home

Demo

- **Submission template for sequence data**

  Here: https://gitlab.gwdg.de/gfbio/molecular-submission-templates/-/blob/master/Template-Description.md

- **Available fields, defaults fields and suggested for FAIRness**

- **Walk through example**

Template Example

sample_title	taxon_id	sample_description	sequencing_plat	library_stra	library_sou	library_sel	library_lay	no	fc	fo	re	re	checksum_	investigation type	environmental	collection da	geographic	geographic	geogra	geographic	environment (biome)
01_VP1c_Kstria	1300146	Kappaphycus_striat	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.333	5	Vietnam	marine biome [ENVO:00000447]
03_VP1a_ Kstri	1300146	Kappaphycus_striat	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.333	5	Vietnam	marine biome [ENVO:00000447]
04_VP1b_Kstria	1300146	Kappaphycus_striat	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.333	5	Vietnam	marine biome [ENVO:00000447]
07_VP2a_Eden	1300146	Euchema_denticula	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.658	109.351	5	Vietnam	marine biome [ENVO:00000447]
09_VP5a_Eden	1300146	Euchema_denticula	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.657	109.333	16	Vietnam	marine biome [ENVO:00000447]
12_VP4a_Eden	1300146	Euchema_denticula	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.34	3	Vietnam	marine biome [ENVO:00000447]
14_VP4b_Eden	1300146	Euchema_denticula	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.34	3	Vietnam	marine biome [ENVO:00000447]
15_CR4a_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.957	109.201	0.5	Vietnam	marine biome [ENVO:00000447]
16_CR4b_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.957	109.201	0.5	Vietnam	marine biome [ENVO:00000447]
17_CR5a_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.948	109.194	0.5	Vietnam	marine biome [ENVO:00000447]
18_CR5b_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.948	109.194	0.5	Vietnam	marine biome [ENVO:00000447]
19_CR1a_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.981	109.2	0.5	Vietnam	marine biome [ENVO:00000447]
20_CR1b_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.981	109.2	0.5	Vietnam	marine biome [ENVO:00000447]
21_CR2a_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.965	109.204	0.5	Vietnam	marine biome [ENVO:00000447]
22_CR2b_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.965	109.204	0.5	Vietnam	marine biome [ENVO:00000447]
23_CR3a_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.969	109.203	0.5	Vietnam	marine biome [ENVO:00000447]
24_CR3b_Kalva	1300146	Kappaphycus_alvar	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	11.969	109.203	0.5	Vietnam	marine biome [ENVO:00000447]
43_VP1a_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.333	0.3	Vietnam	marine biome [ENVO:00000447]
44_VP1b_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.333	0.3	Vietnam	marine biome [ENVO:00000447]
45_VP2a_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.658	109.351	0.3	Vietnam	marine biome [ENVO:00000447]
46_VP2b_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.658	109.351	0.3	Vietnam	marine biome [ENVO:00000447]
47_VP3a_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.655	109.346	0.3	Vietnam	marine biome [ENVO:00000447]
48_VP3b_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.655	109.346	0.3	Vietnam	marine biome [ENVO:00000447]
49_VP4a_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.34	0.3	Vietnam	marine biome [ENVO:00000447]
50_VP4b_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.652	109.34	0.3	Vietnam	marine biome [ENVO:00000447]
51_VP5a_SPM	408172	Suspended_particul	Illumina MiSeq	AMPLICON	GENOMIC	PCR	paired	#	341F	341F	MD5	mimarks-survey	water	27/05/2017	12.657	109.333	0.3	Vietnam	marine biome [ENVO:00000447]

Example available in the GFBio Helpdesk (submission ID: DSUB-348)

DFG Deutsche Forschungsgemeinschaft

www.gfbio.org

@GFBio_Project