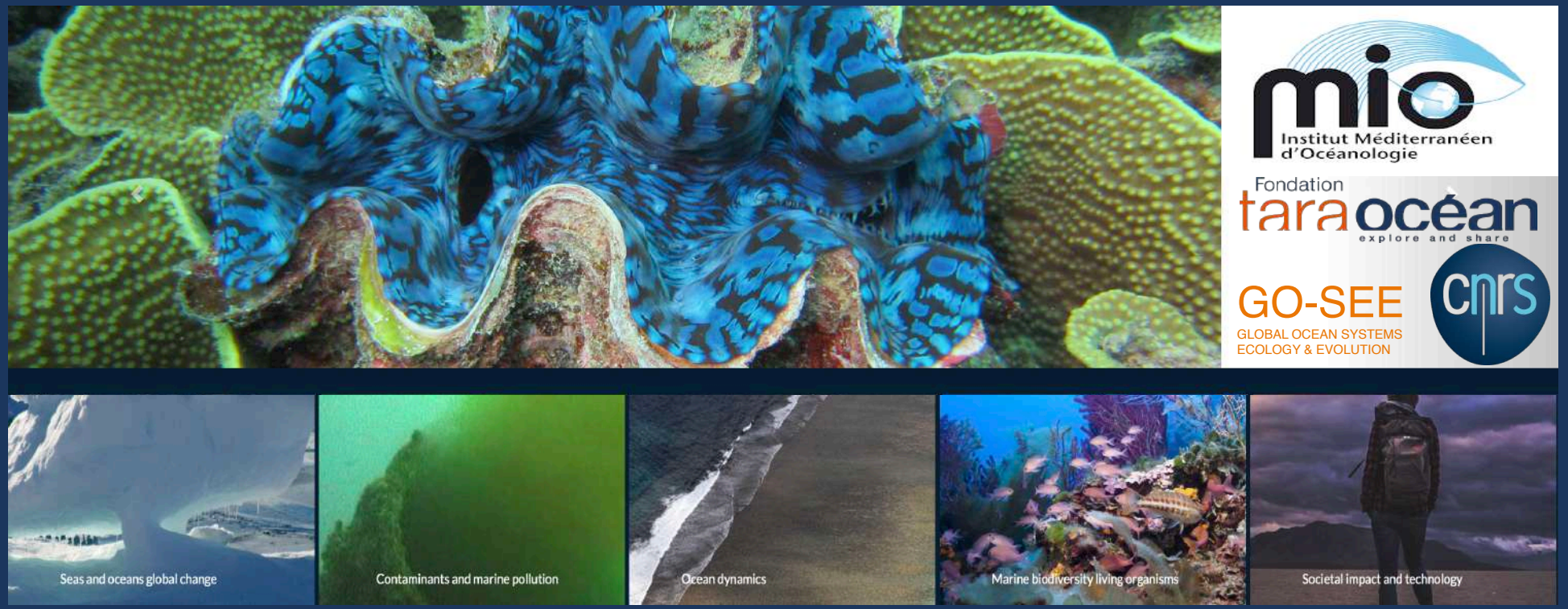
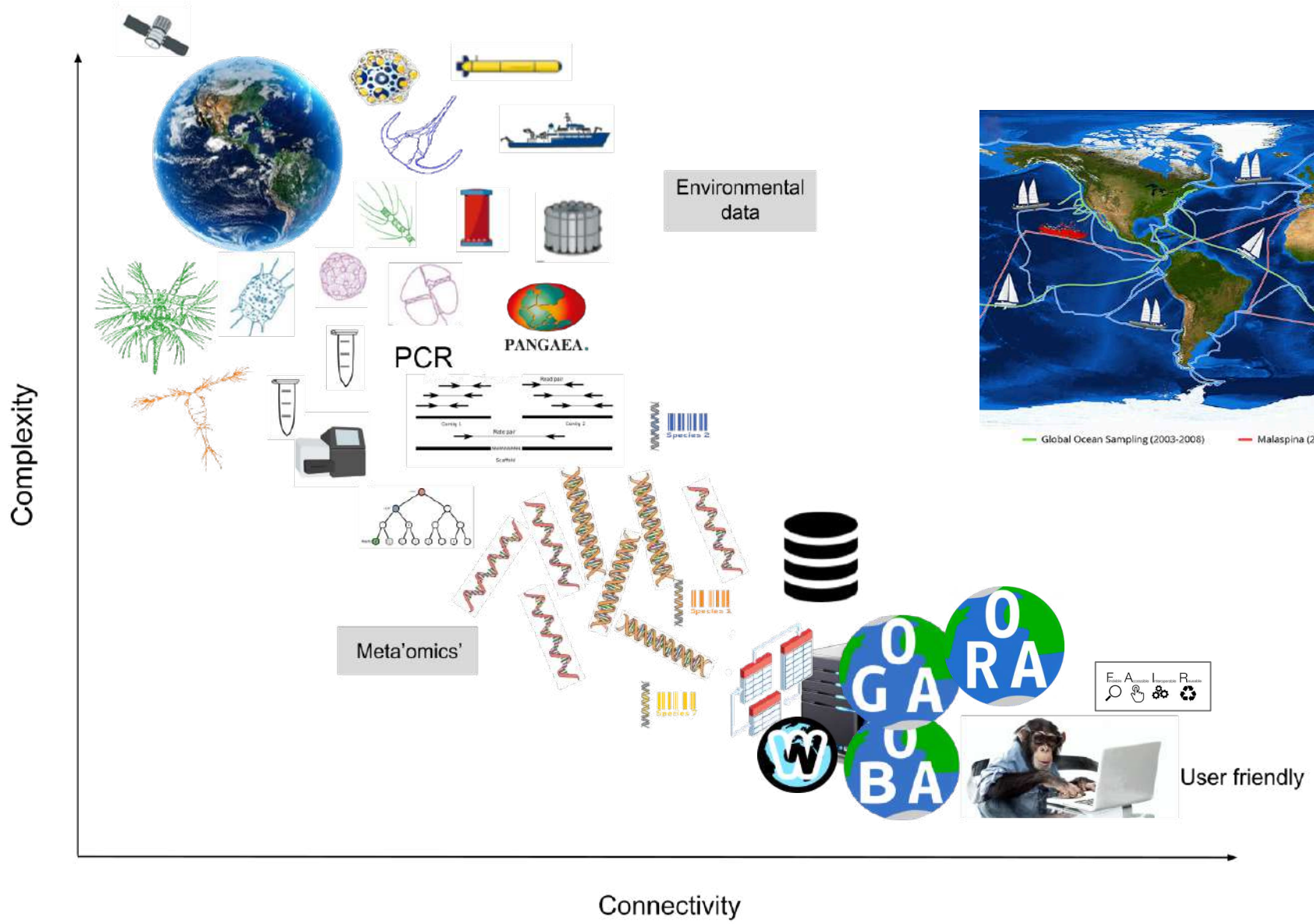
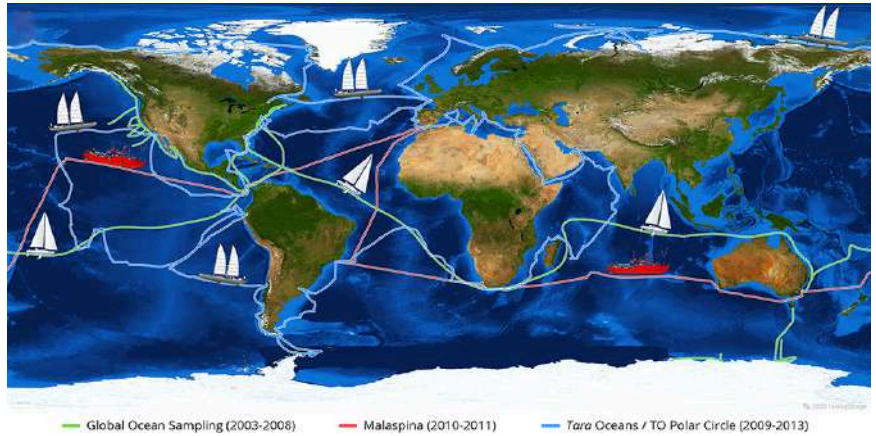
# Marine genomics web services




**Magali Lescot** **MIO** UMR 7294 – CNRS AMU IRD - Marseille

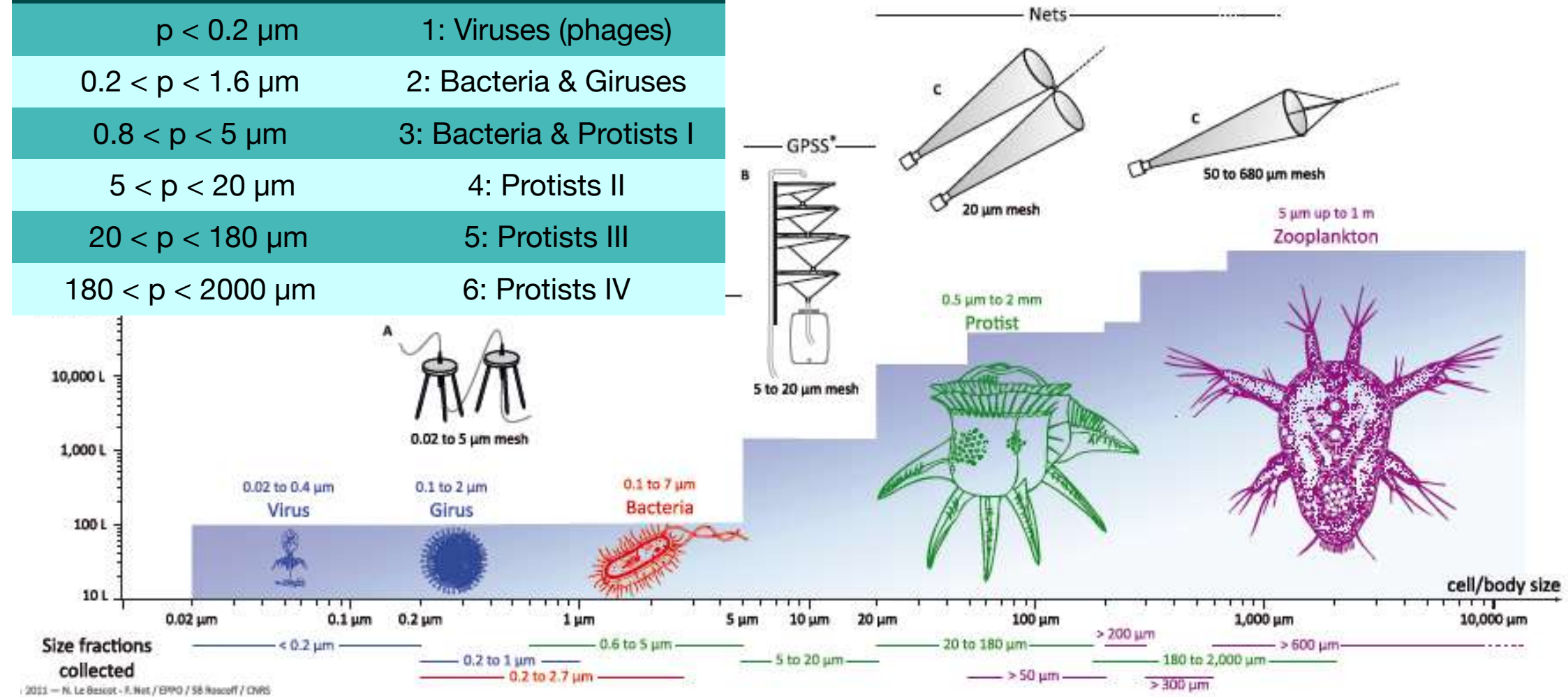
# Data integration



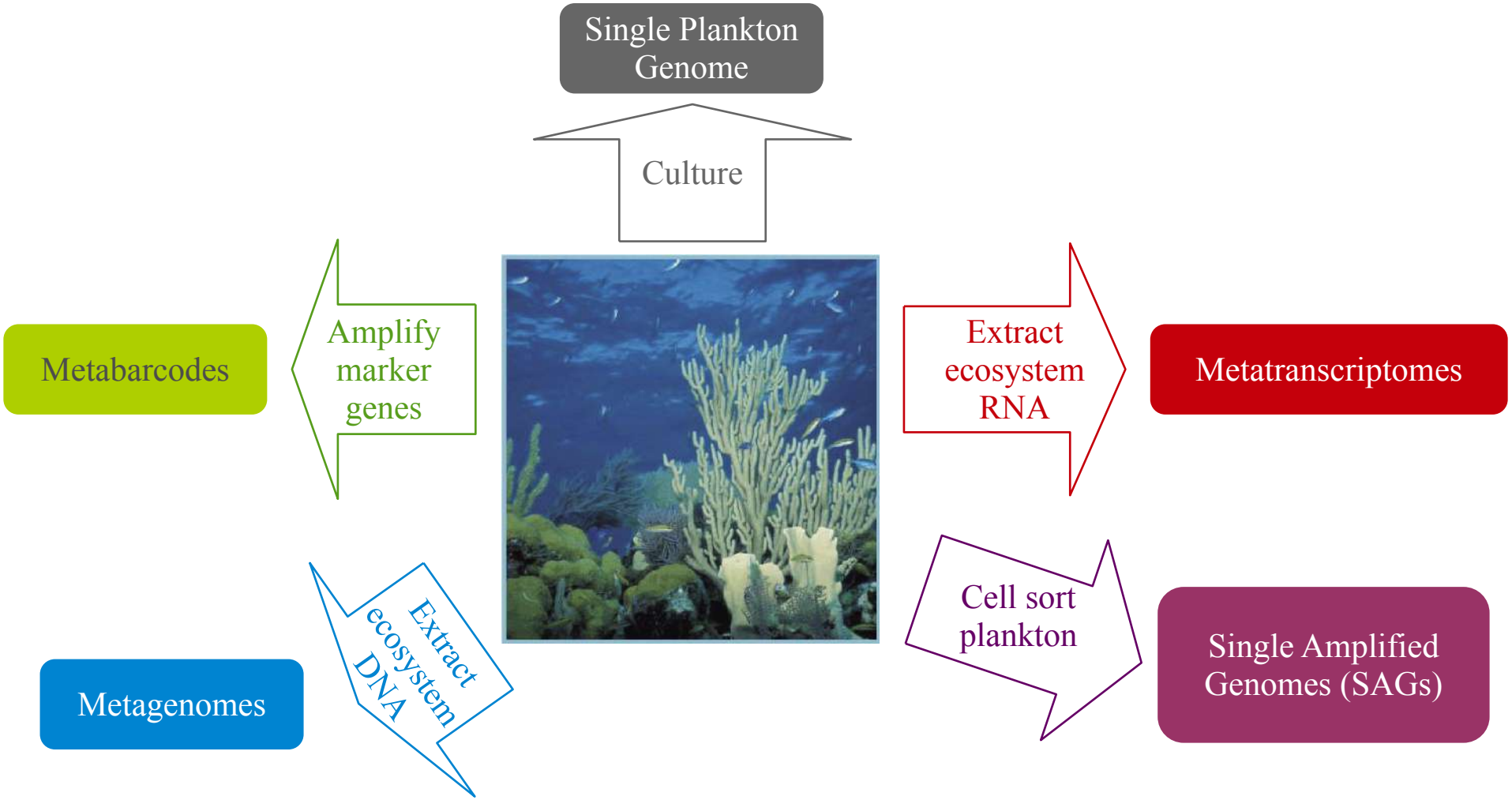



# Sampling strategy

| Filter pore sizes | Plankton fraction |
|---|---|
| p < 0.2 µm | 1: Viruses (phages) |
| 0.2 < p < 1.6 µm | 2: Bacteria & Giruses |
| 0.8 < p < 5 µm | 3: Bacteria & Protists I |
| 5 < p < 20 µm | 4: Protists II |
| 20 < p < 180 µm | 5: Protists III |
| 180 < p < 2000 µm | 6: Protists IV |



# Environmental genomics



# Environmental genomics

Single Plankton Genome

Culture

Amplify marker genes

Metabarcodes

→ which species are present ?

Extract ecosystem RNA

Metatranscriptomes

→ which functions are active now ?

Extract ecosystem DNA

Metagenomes

→ what biological functions are present ?

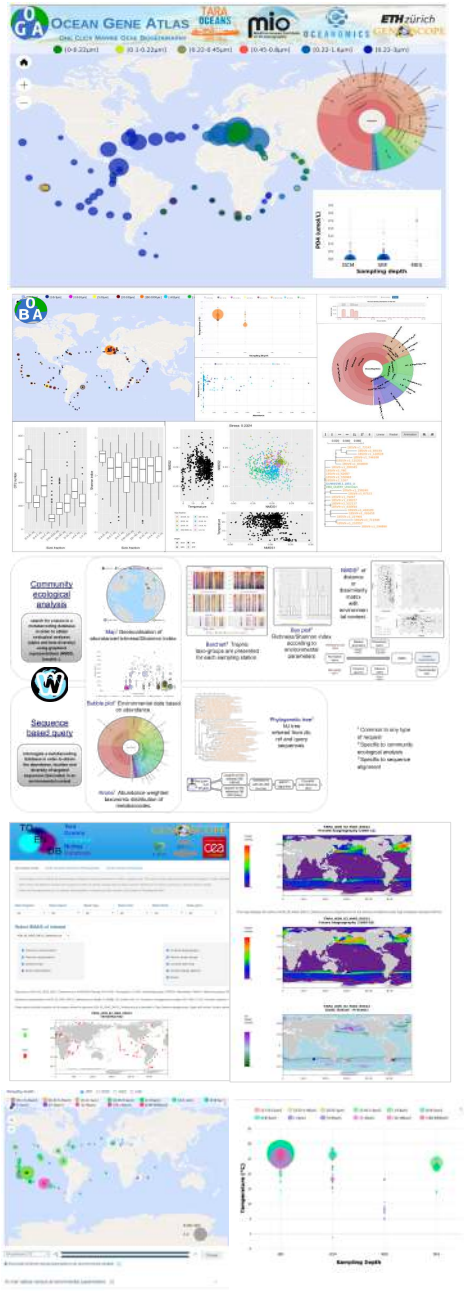
Cell sort plankton

Single Amplified Genomes (SAGs)

→ Full genome for representative uncultured species


tara-oceans.mio.osupytheas.fr
Ocean Atlas - Home page
Ticket internet
OCEAN ATLAS
ONE CLICK MARINE BIOGEOGRAPHY
TARA OCEANS
2009-2012
mio
Institut Méditerranéen d'Océanologie
OCEANOMICS
http://tara-oceans.mio.osupytheas.fr/
Explore the biogeography of a gene/protein sequence in a panoceanic collection of plankton metagenomes :
OGA
Ocean Gene Atlas
Explore the distribution of plankton taxa across a panoceanic collection of metabarcodes :
OBA
Ocean Barcode Atlas
Explore the biogegraphy of shared k-mer ratios for one or multiple DNA sequence(s) :
ORA
Ocean Read Atlas
Select below the SMAG of interest and visualize the effect of climate change :
TOENDB
Tara Oceans Ecological Niches Database

Web services




**Ocean Gene Atlas =>** **Ocean Gene Atlas V2.0**
- Marine metagenomic and metatranscriptomic datasets
  *Tara* Oceans Procaryotes and Eukaryotes genes catalogues

**8 trillion of read sequences for 228 million genes**

**Ocean Barcode Atlas**
- Marine metabarcoding datasets
  500 000 barcode sequences

**Ocean Read Atlas**

First indexing method capable of processing terabyte-sized genomic sequence in one dataset!

- index of 1,393 metagenome samples of raw sequences from virus to mesozooplankton from *Tara* Oceans
- 36.7TB raw data = one index (4.63TB)

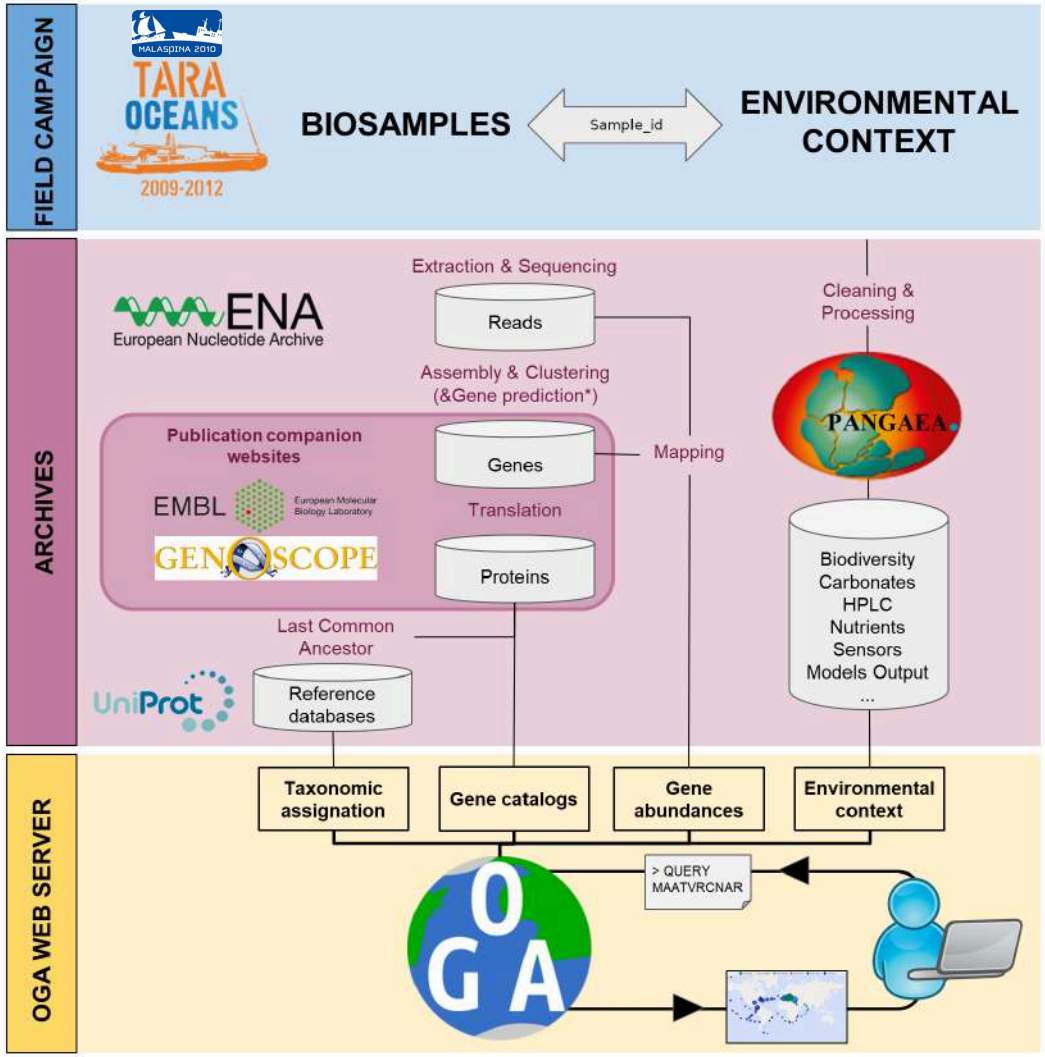
**TOENDB**
- *Tara* Oceans Ecological Niches

713 Single Cell and Metagenomes Assembled Genomes

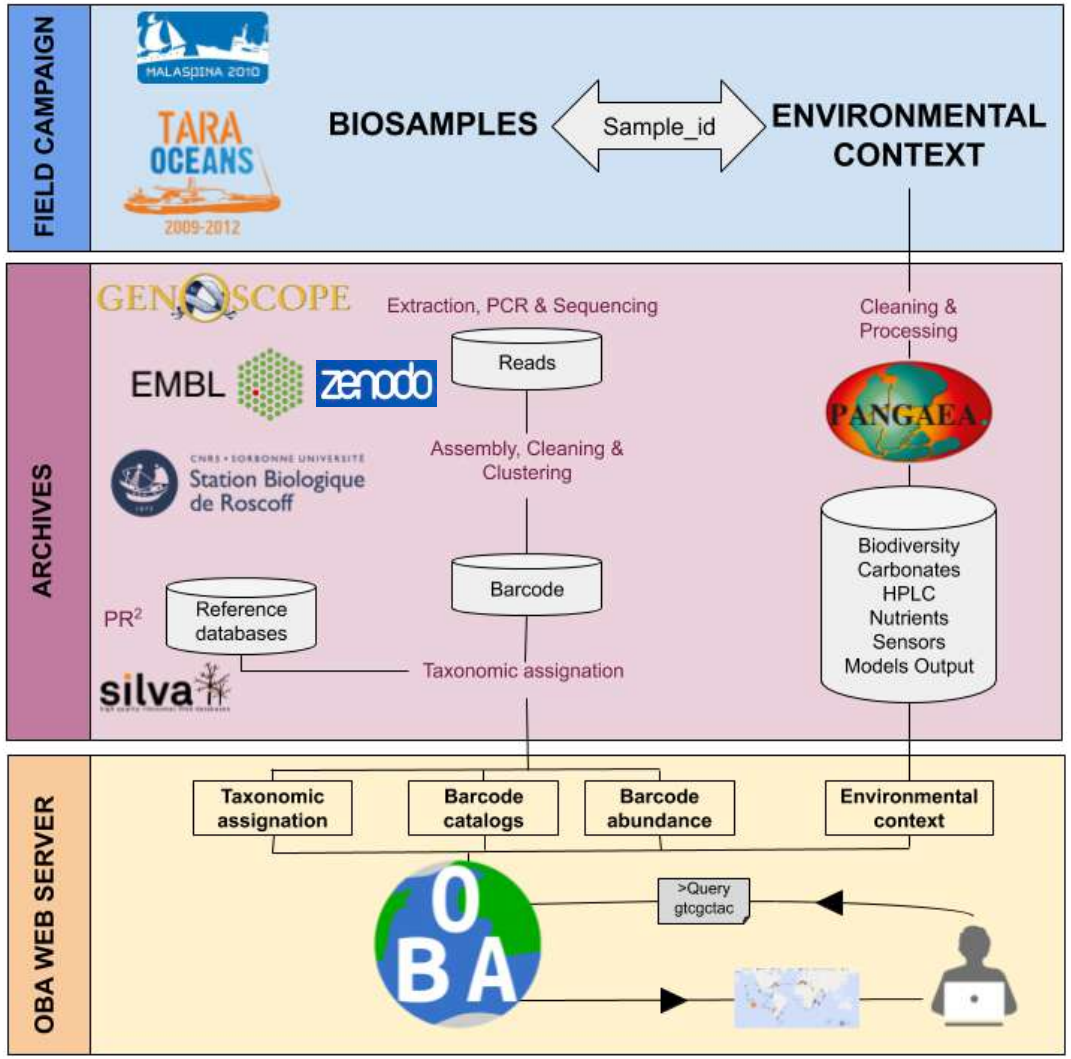
http://tara-oceans.mio.osupytheas.fr/

# Ocean Gene Atlas

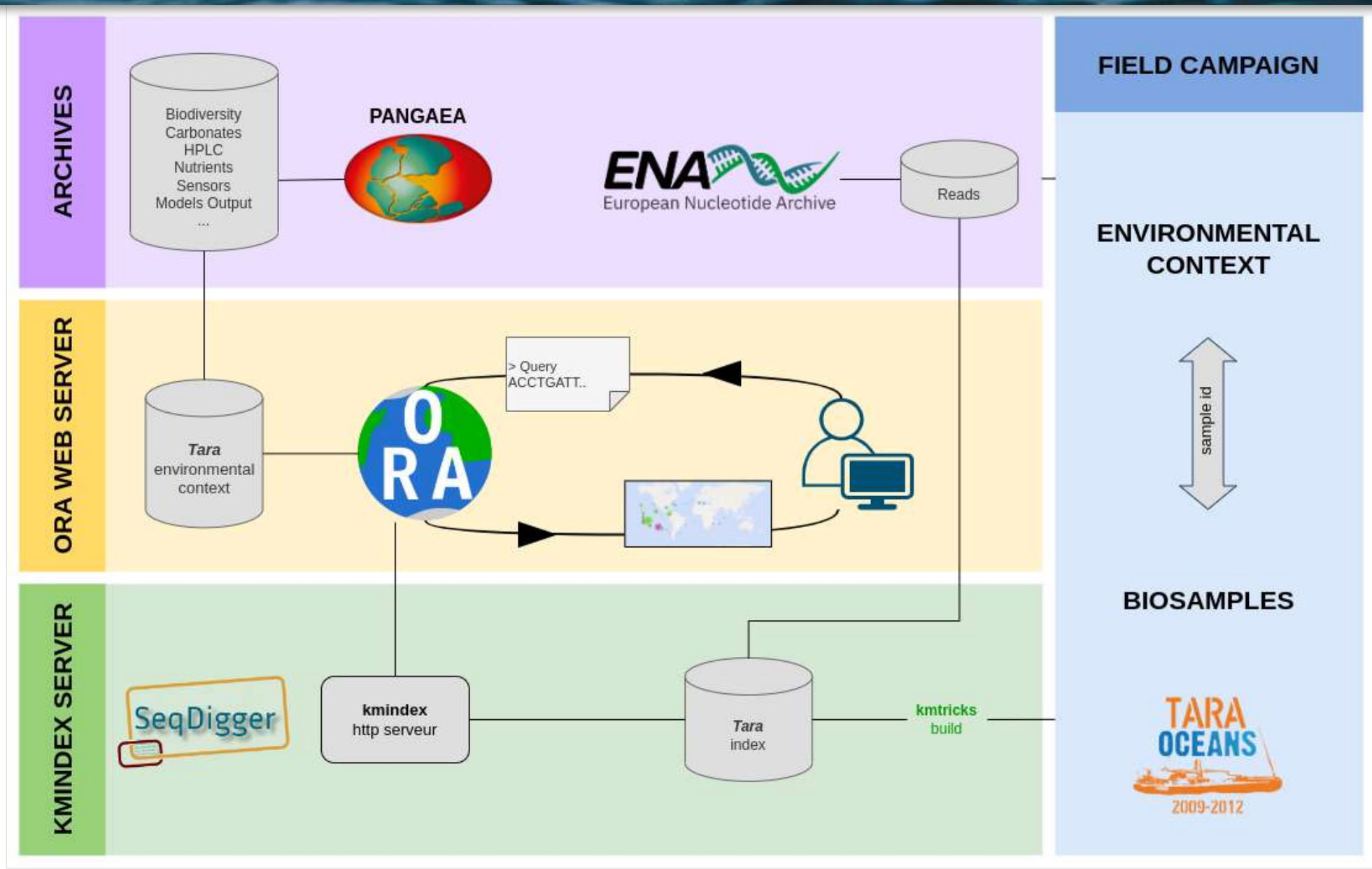
## Metagenomic & Metatranscriptomic datasets




## Metabarcoding datasets

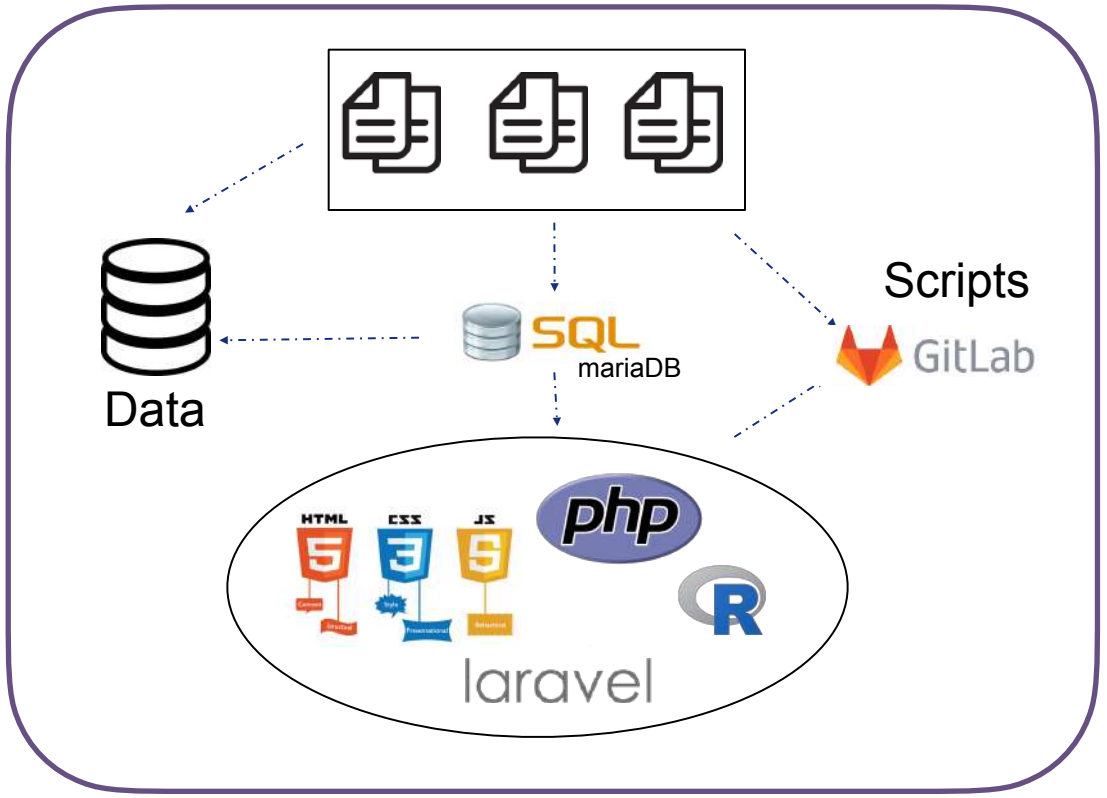


# The Ocean Read Atlas

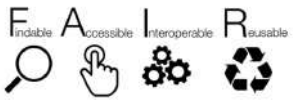


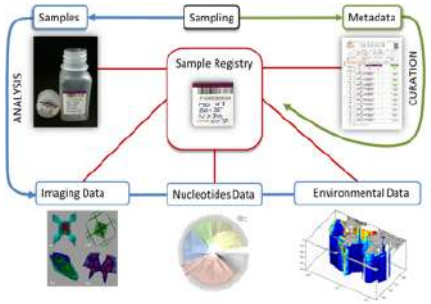
Nolan Lezzoche

Services web

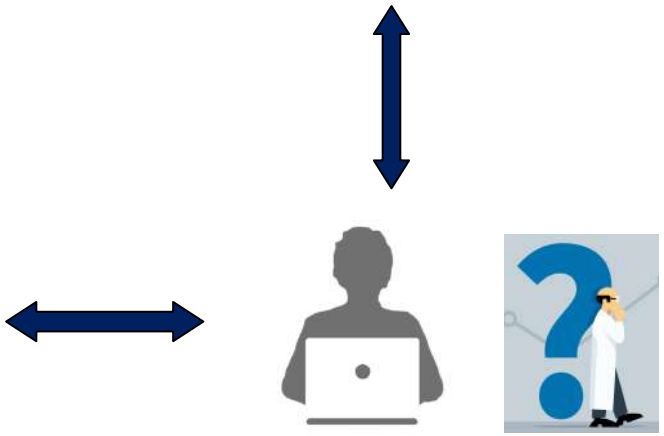




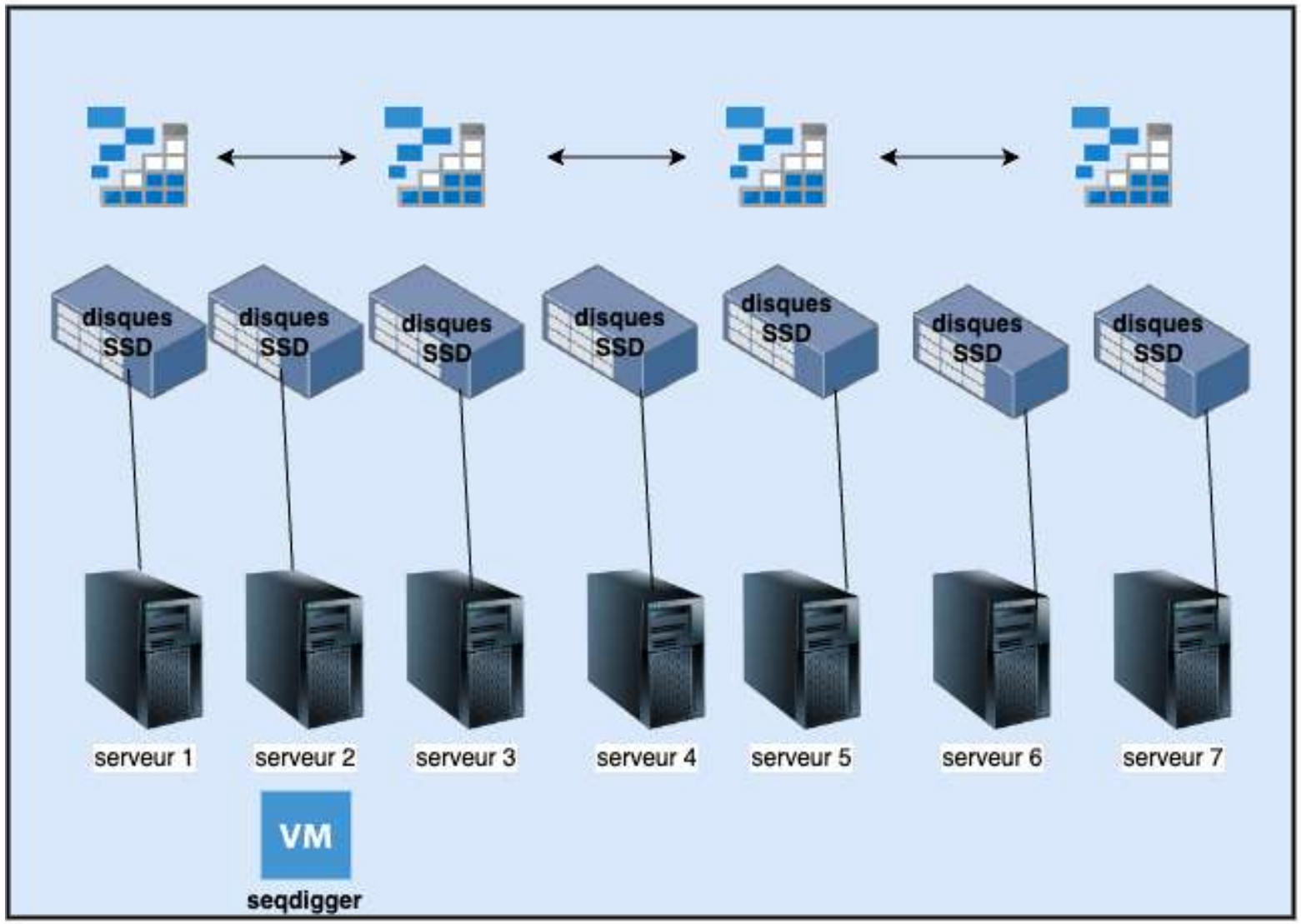



Metadata



# Infrastructure SeqDigger



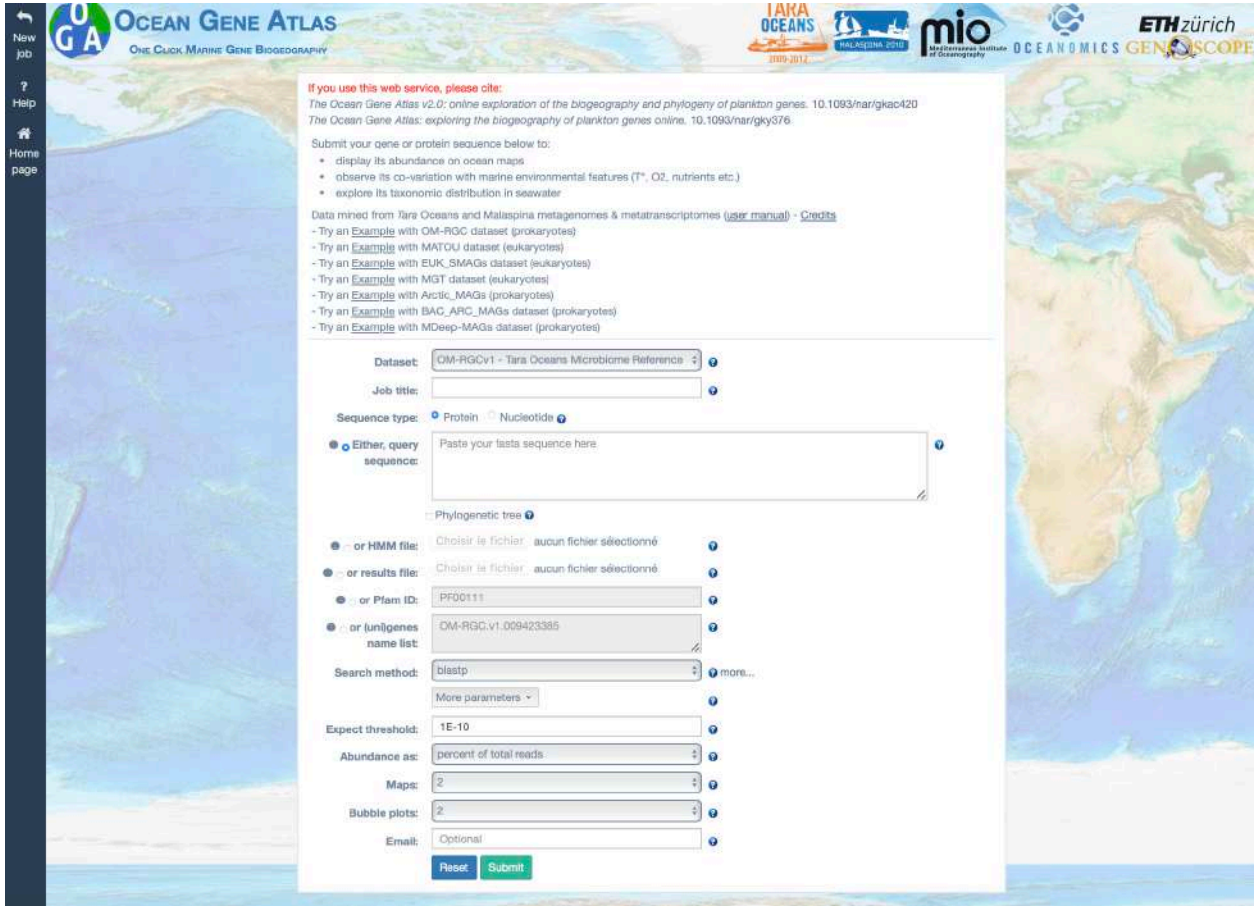

Julien Lecubin

Distributed storage system Ceph

virtualisation PROXMOX

# The Ocean Gene Atlas v2.0

Metagenomic (MetaG)
Metatranscriptomic (MetaT)




*Nucleic Acids Research, 2022* 1
https://doi.org/10.1093/nar/gkac420

**The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes**

Caroline Vernette[1,2], Julien Lecubin[3], Pablo Sánchez[4], Tara Oceans Coordinators, Shinichi Sunagawa[5], Tom O. Delmont[2,6], Silvia G. Acinas[4], Eric Pelletier[2,6], Pascal Hingamp[1] and Magali Lescot[1,2,*]

*Nucleic Acids Research, 2018* 1
doi: 10.1093/nar/gky376

**The Ocean Gene Atlas: exploring the biogeography of plankton genes online**

Emilie Villar[1,2,*], Thomas Vannier[2], Caroline Vernette[2], Magali Lescot[2], Miguelangel Cuenca[3], Aurélien Alexandre[2], Paul Bachelerie[2], Thomas Rosnet[2], Eric Pelletier[4], Shinichi Sunagawa[3] and Pascal Hingamp[2,*]

## One Click Marine Gene Biogeography

| Dataset | Read number | MAG number | Gene number | Sample number |
|---|---|---|---|---|
| OM-RGCv1 | 7.20E+12 | - | 40,154,822 | 243 |
| OM-RGCv2_metaG | 1.13E+11 | 370 | 46,775,154 | 180 |
| OM-RGCv2_metaT | 5.00E+09 | 187 | | 187 |
| MATOU_metaG | 1.85E+11 | - | 116,849,350 | 445 |
| MATOU_metaT | 8.70E+10 | | | 440 |
| MGT | 5.80E+07 | 924 | 6,946,068 | 364 |
| EUK_SMAGs | 2.80E+11 | 713 | 10,207,435 | 939 |
| BAC_ARC_MAGs | 2.80E+11 | 1,888 | 4,567,982 | 922 |
| Arctic_MAGs_metaG | 1.40E+08 | 530 | 1,033,381 | 68 |
| Arctic_MAGs_metaT | 4.50E+07 | | | 53 |
| MDeep-MAGs | 6.49E+08 | 317 | 867,795 | 58 |
| **Total** | **8,15E+12** | 4,929 | 227,401,987 | |

mio Mediterranean Institute of Oceanography

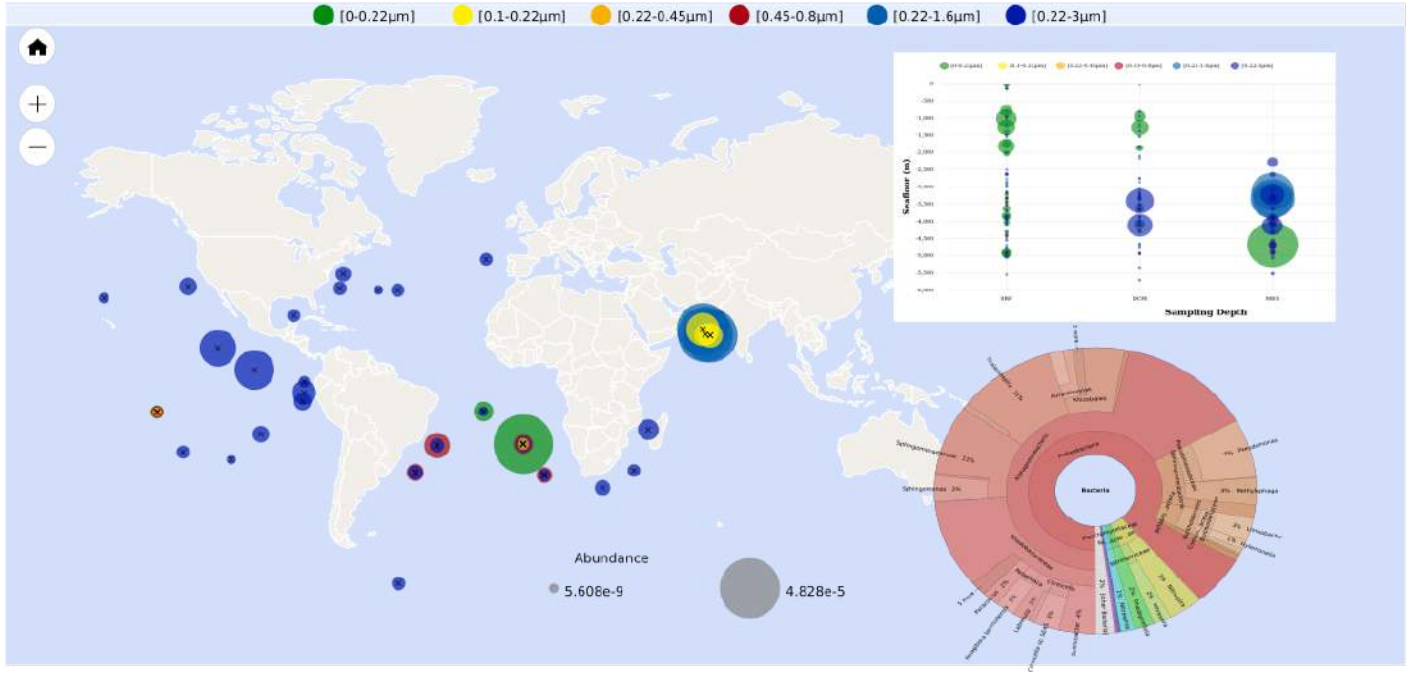
OCEANOMICS

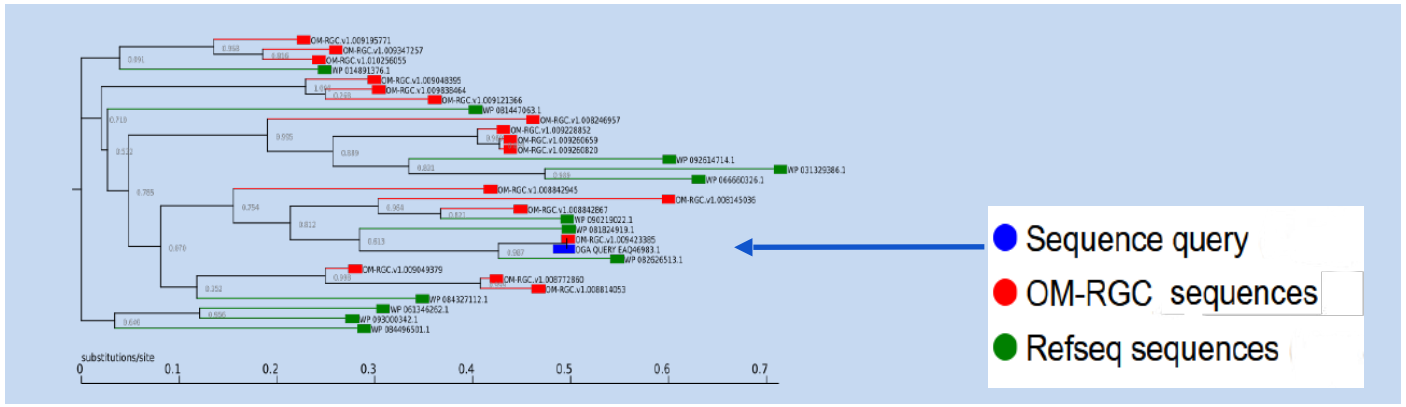
TARA OCEANS

elixir Data for Life

SeqDigger

# Biogeography of homologous sequences



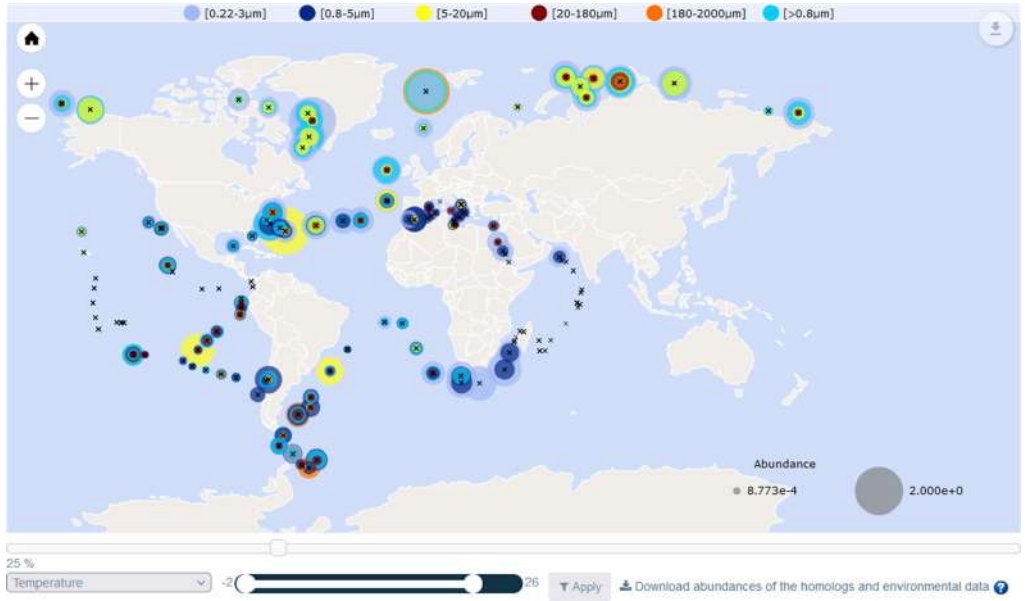

- Search for a gene or protein in the different samples (sequence alignment or hmm search)

- Computation of the homologous genes abundance in each sample (/ station and / filter)

- Correlation with environmental parameters

- Taxonomic assignment of identified sequences

- Phylogenetic tree of the homologous sequences




Phylogenetic tree : marine environmental genes in context of reference sequences :

View multiple alignment

- Radial / Linear tree
- Root on the longest branch
- Remove tree root
- Add branch length values
- Increase / Decrease leaf size, Hide / Show leaf labels
- Grow / Shrink tree size
- Reset (cancel all changes)

Download tree in svg format
Download sequences in fasta format
Download full fasta alignment
Download the intermediate column cleaned alignment (Trimal) in fasta format
Download the final cleaned alignment (MaxAlign) in fasta format
Download output from second alignment curation step (MaxAlign)
Download tree in newick format

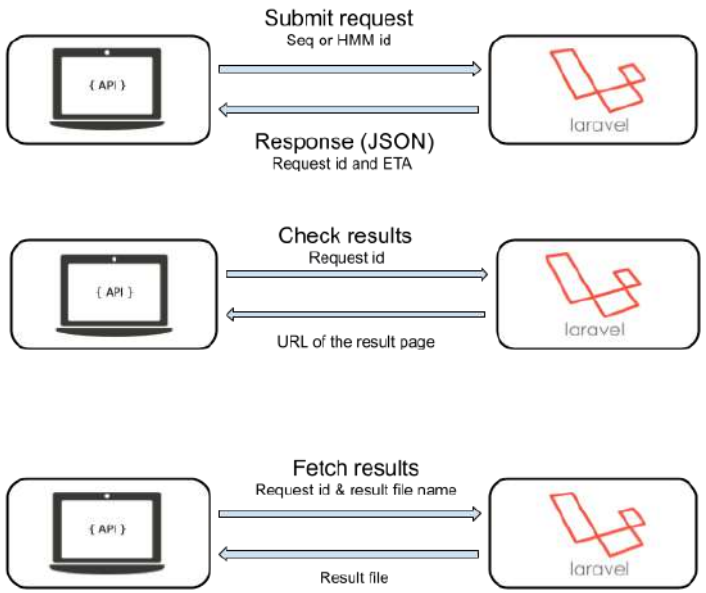
- Root the tree on the following leaf: Root
- FastTree option's :

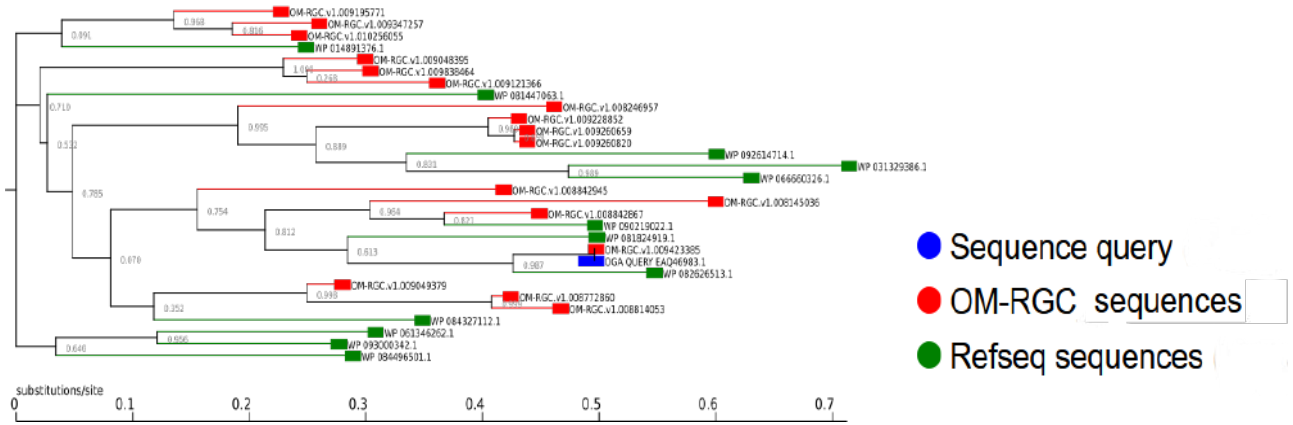
WAG substitution model  JTT substitution mode
Gamma20 distribution  no Gamma distribution  Recompute tree
(It may take several minutes)

Number of sequences from **SMAGs** : 77 (4 sequence(s) were/was excluded during the maxalign step)
Number of sequences from **RefSeq** after clustering: (8 sequence(s) were/was excluded during the maxalign step)

| Clustering at identity: | - | 100% | 95% | 90% | 85% | 80% | 75% | 70% | 65% | 60% |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of sequences: | 1254 | 1234 | 891 | 685 | 545 | 399 | 280 | 189 | 112 | 64 |

# API: Application and Programming Interface

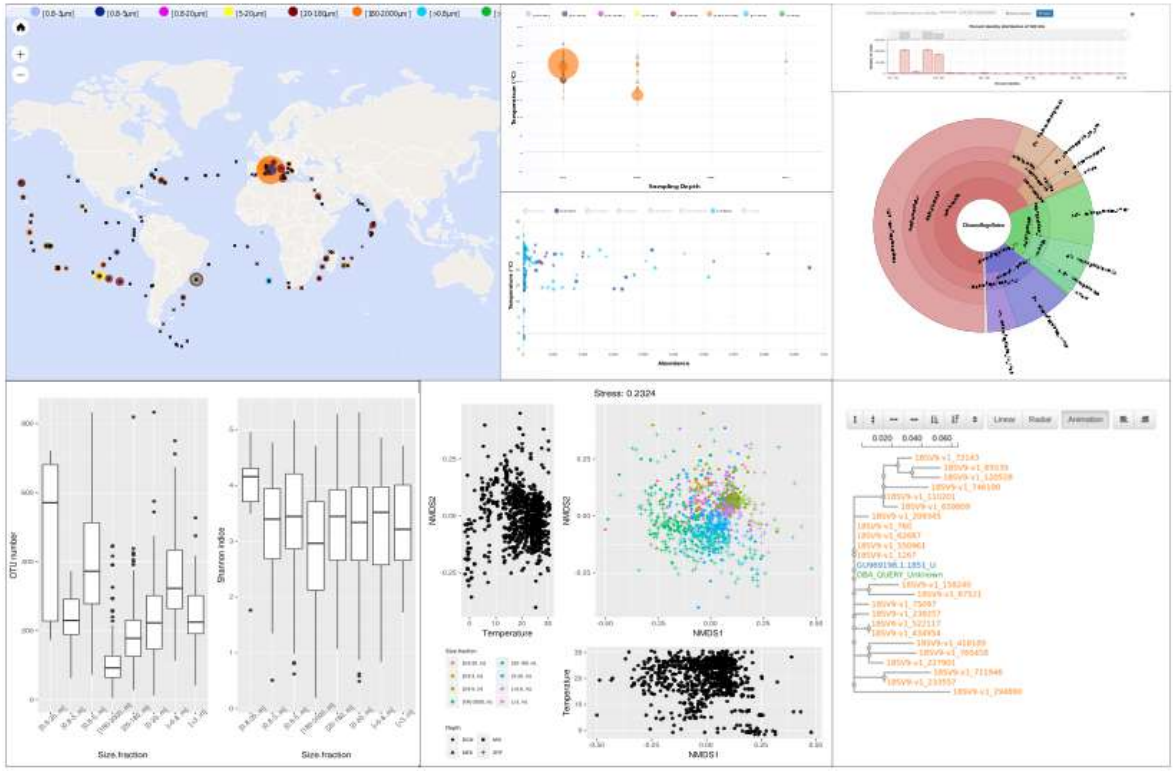



Vernette et al, 2022. doi: 10.1093/nar/gkac420

# The Ocean Barcode Atlas

## Metabarcoding datasets



http://oba.mio.osupytheas.fr/ocean-atlas/

*The Ocean Barcode Atlas: a web service to explore the biodiversity and biogeography of marine organisms.*
C. Vernette, N. Henry, J. Lecubin, C. de Vargas, P. Hingamp, M. Lescot (2020)
*Molecular Ecology Resources.* Link: https://doi.org/10.22541/au.160193452.23998228/v1

## Diversity analysis of planktonic communities

### Public

**18S V9**
474,303 OTUs
1,046 samples

**18S V4**
156,648 OTUs
1,191 samples

**16S V4**
3,902 OTUs
60 samples

**ASV 18S V4**
1,011 samples

**ASV 18S V9**
1,069 samples

**miTAGs**
23,987 sequences
180 samples

### Private

**ASV 16S V4V5**
1,361,502 ASVs
1,134 samples

**MOOSE 18S V4**
15,743 OTUs
271 samples

Mediterranean Institute of Oceanography

1872

Fondation tara océan explorer et partager

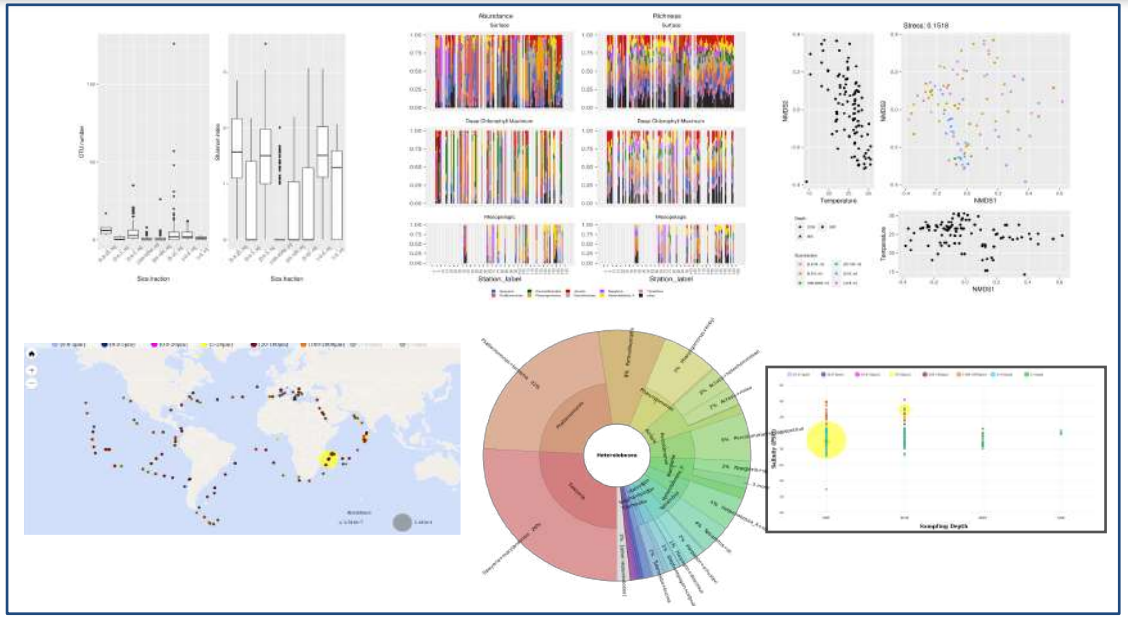
MALASPINA 2010

MOOSE Mediterranean Ocean Observing System for the Environment

# Request types

Community ecological analysis

Search by taxonomy



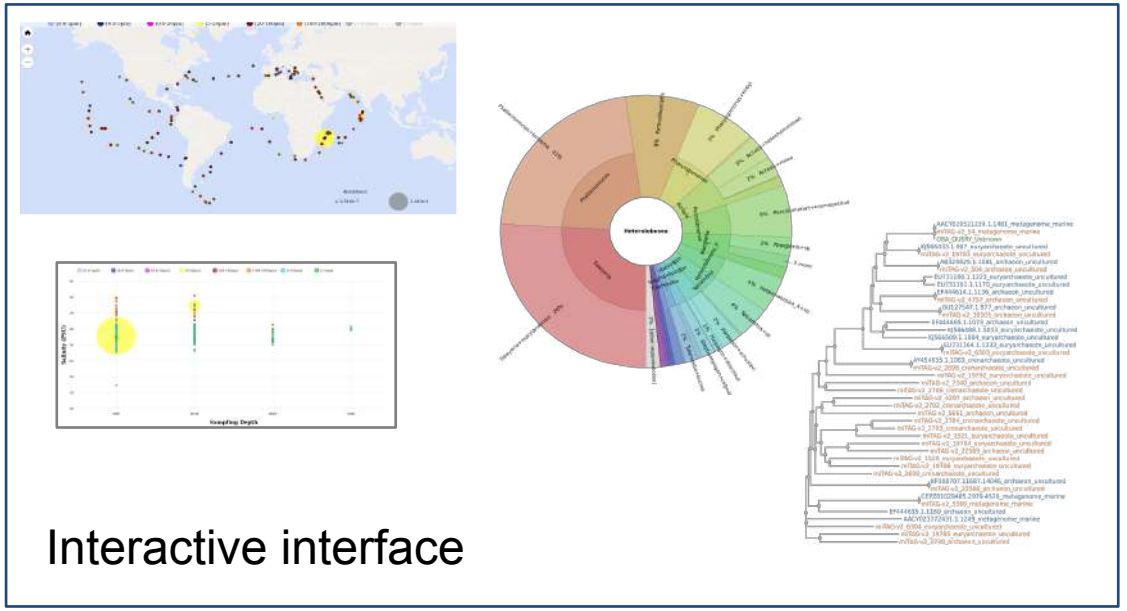
Sequence based query

Submit your sequence

Alignment

Submit from a ref db

Search from an ID



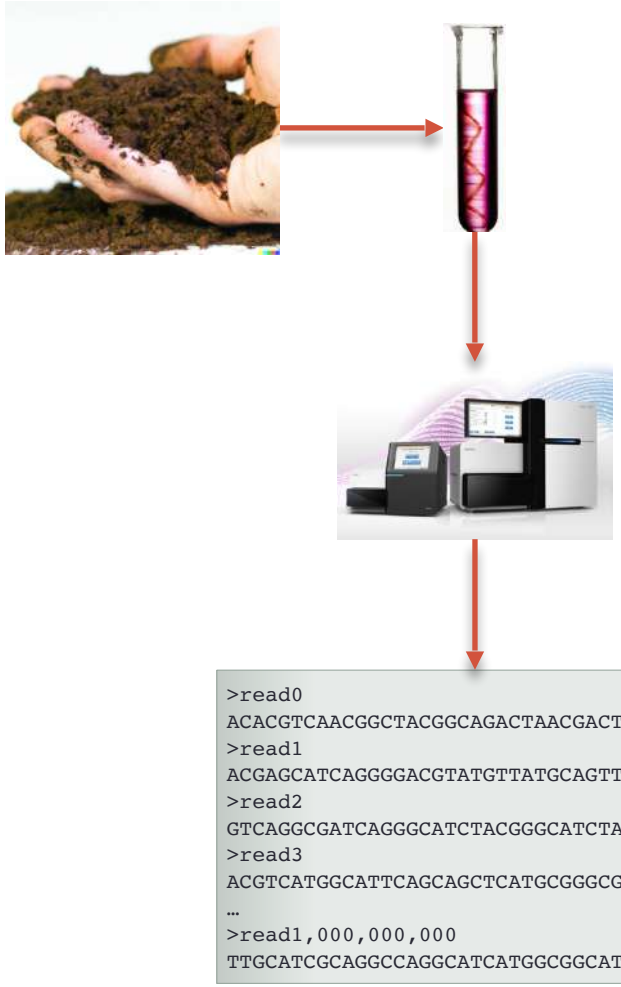
Interactive interface

Submission interface

User friendly web service

# Data

**Metagenome sequencing** (soil, seawater, waste water, air, gut, …)




## Raw sequences

- **Redundant** & fragmented data
- Error-prone (0.1% to 10% error rate)
- Important **background noise**
- Heterogenous
  - Quality and quantity
- **Volumes**:
  - hundreds millions fragments / experiment
  - Millions of experiments
- **Archived**

ENA European Nucleotide Archive
53.5 petabytes
doubles: 32 months

*P. Perterlongo*

**ENA** (European Nucleotide Archive) contains almost 50 Petabytes of raw data (raw sequences)

=> equivalent to 100,000 desktop computer

Query this data with a sequence « ATGAGAAAAGTAGCAATTTACGGAAAAGGC » is IMPOSSIBLE!

The cumulative total size of the compressed files representing these genomes is 125 Gigabytes.

That's around 500 thousand times smaller than the 50 Petabytes of raw sequencing data (also compressed).

=> querying these genomes would take several hundred days.

We know how to query assembled genomes.

But **we don't know how to query the raw data** 2 by 2.

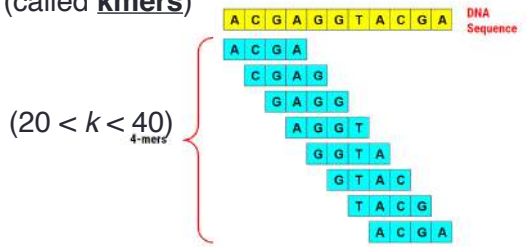
=> 2 weeks at TGCC on *Tara* Oceans datasets

=> so we can extend to 2000 weeks = 38 years of queries on "The Sequence Read Archive"

# kmers

## Words

- **No word in DNA**
- Split to subsequences of fixed length *k* (called **kmers**)




- Thousand billions distinct kmers
  - *(google indexes millions)*

## Compare sequences

- Sequence similarity ~ shared kmers count

```
ACGAGGTACGA        ACGAGTTACGA
ACGA               ACGA
 CGAG               CGAG
  GAGG               GAGT
   AGGT               AGTT
    GGTA               GTTA
     GTAC               TTAC
      TACG               TACG
       ACGA               ACGA
```

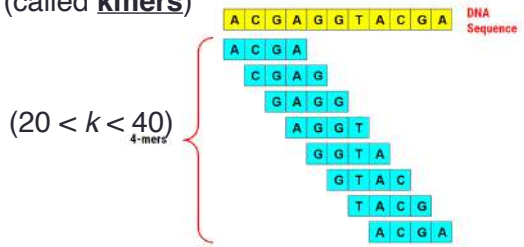
- 4 over 8 kmers shared

*P. Perterlongo*

# kmers

## Words

- **No word in DNA**
- Split to subsequences of fixed length *k* (called **kmers**)

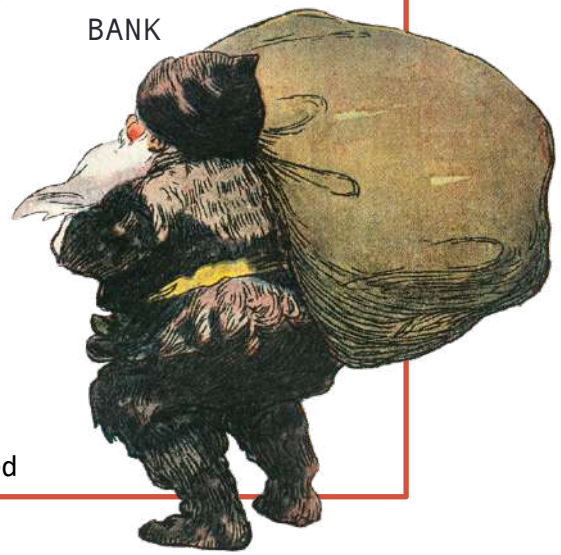



- Thousand billions distinct kmers
  - *(google indexes millions)*

## Query vs Bank

- Sequence similarity ~ shared kmers count

```
ACGAGGTACGA        BANK
ACGA
 CGAG
  GAGG
   AGGT
    GGTA
     GTAC
      TACG
       ACGA
```



- 6 over 8 kmers shared

*P. Perterlongo*

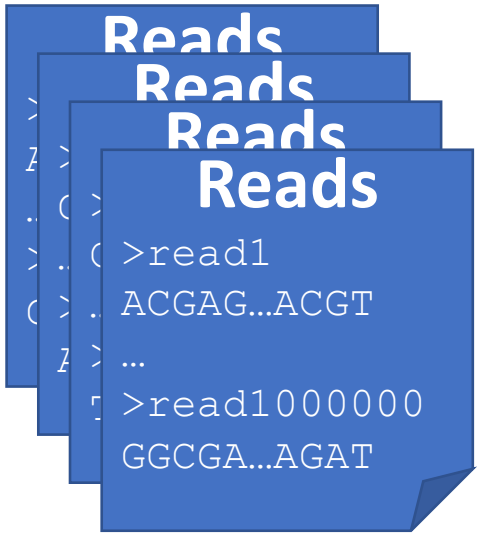
# Indexing: conceptual view

**One read set**:

- Extract & count kmers
- Filter kmers
- Generate a [counting] bloom filter with kmtricks

**N read sets**:

- Create N [counting] bloom filters
- This is the index



Probabilistic data structure



*P. Perterlongo*

# Kmindex & Ocean Read Atlas

**kmindex** and **ORA**: indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets

Téo Lemane[1,*], Nolan Lezzoche[2], Julien Lecubin[3], Eric Pelletier[4,5], Magali Lescot[2,5], Rayan Chikhi[6], and Pierre Peterlongo[1,*]

**Input:** *Tara* Oceans Metagenomic

Compressed fastq.gz files:
- 36.7 TB, 1,393 samples

Each sample:
- Position
- Species fraction sizes
- Physico-chemical env:
  - Ph, salinity, T°, …

Tara Schooner - Creative Commons Attribution 3.0

*P. Perterlongo*

# Kmindex & Ocean Read Atlas

**kmindex** and **ORA**: indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets

Téo Lemane[1,*], Nolan Lezzoche[2], Julien Lecubin[3], Eric Pelletier[4,5], Magali Lescot[2,5], Rayan Chikhi[6], and Pierre Peterlongo[1,*]

nature computational science

Explore content · About the journal · Publish with us · Subscribe

nature › nature computational science › brief communications › article

Brief Communication | Published: 26 February 2024

**Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA**

Téo Lemane, Nolan Lezzoche, Julien Lecubin, Eric Pelletier, Magali Lescot, Rayan Chikhi & Pierre Peterlongo

*Nature Computational Science* 4, 104–109 (2024) | Cite this article

**Input:** *Tara* Oceans Metagenomic

Compressed fastq.gz files:
- 36.7 TB, 1,393 samples

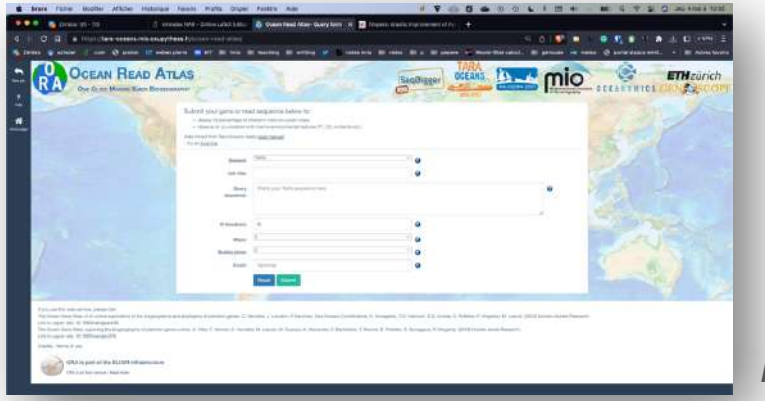
Each sample:
- Position
- Species fraction sizes
- Physico-chemical env:
  - Ph, salinity, T°, …

3 days
(TGCC CEA)

**Output:** Index & web server
Index size: 4.63TB

Instant Queries
- Similarity metric
- Correlations with env. Metrics

A web server
https://ocean-read-atlas.mio.osupytheas.fr/




*P. Perterlongo*

# The Ocean Read Atlas - Like OGA for metagenomic reads

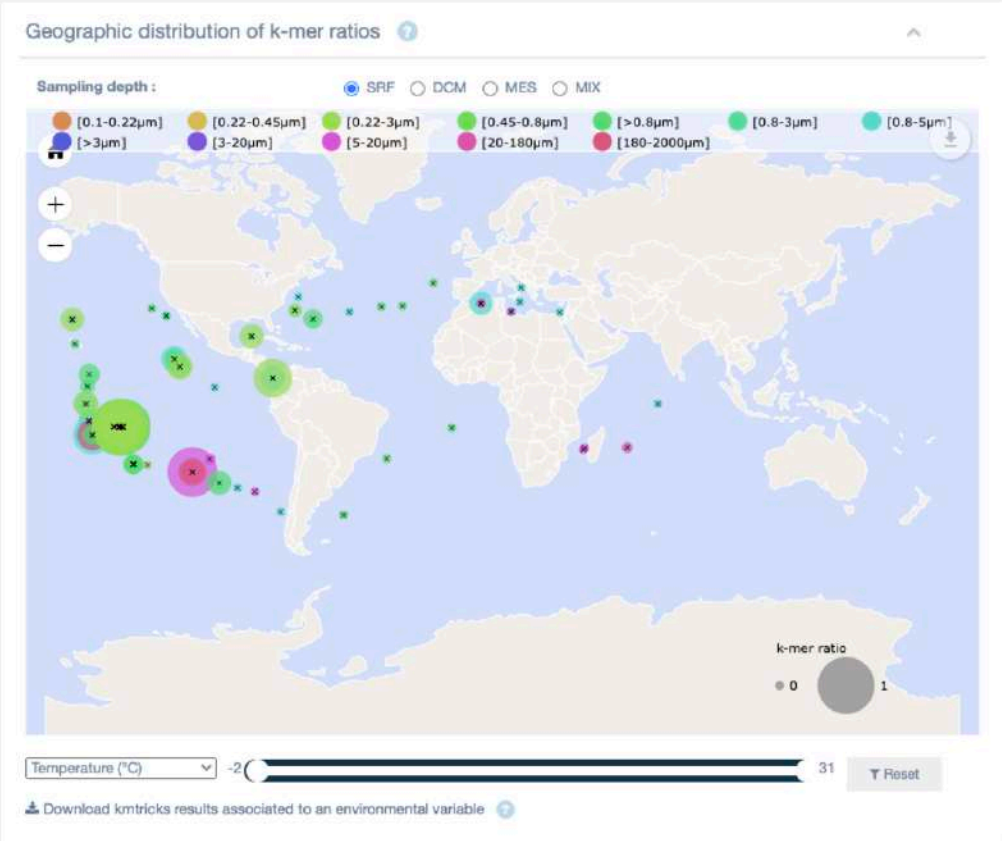
| | Build index | | | | Query time | | FP rate (%) | |
|---|---|---|---|---|---|---|---|---|
| | Time | RAM GB | Disk GB | Index size GB | Nb. queried reads 1 | Nb. queried reads 10 million | Average | Max |
| MetaProFi | 30h15 | 278 | 5,684 | 226 | 12s72 | 1h29 | 11.18 | 21.55 |
| COBS | 26h30 | 278 | 5,684 | 184 | 1s51 | 15h56 | 13.29 | 24.60 |
| kmindex | **2h56** | **107** | **878** | **164** | **0s06** | **4m21s** | **0.006** | **0.18** |

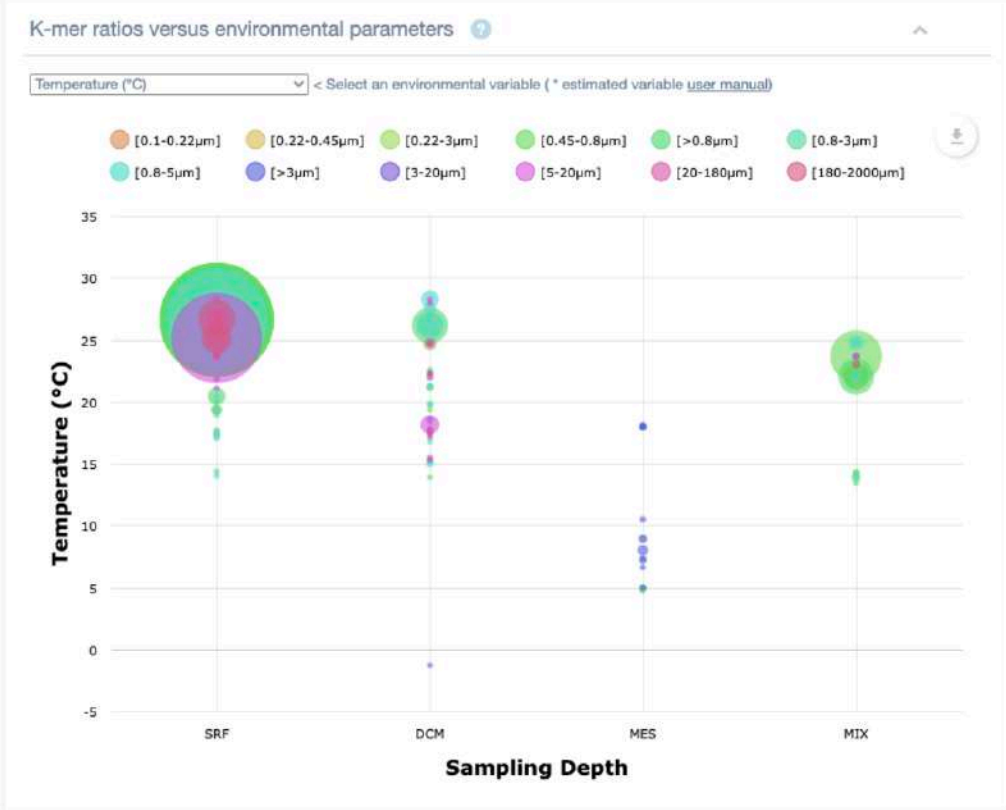
Index construction and read query performance of kmindex on 50 *Tara* Oceans samples.

*T. Lemane*

# The Ocean Read Atlas - Like OGA for metagenomic reads

- Exploring the biogeography of all sequences from *Tara* Oceans metagenomes (1393 samples)
- Limited to presence/absence query



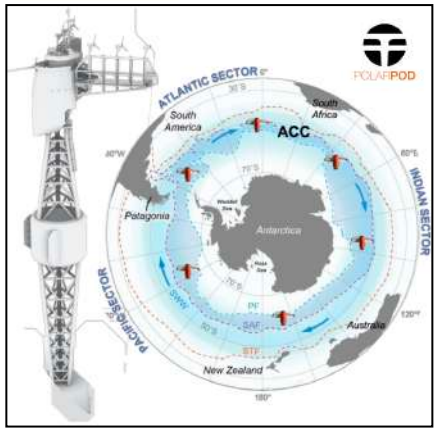

*T. Lemane*

TARA OCEANS

APERO

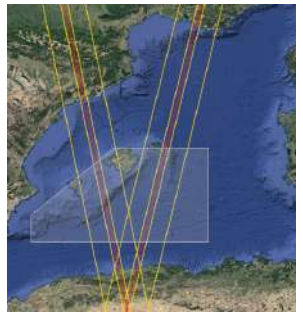
MOOSE
Mediterranean Ocean Observing System for the Environment

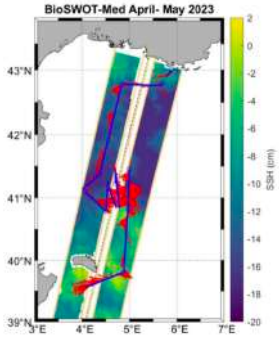
BioSWOT Med

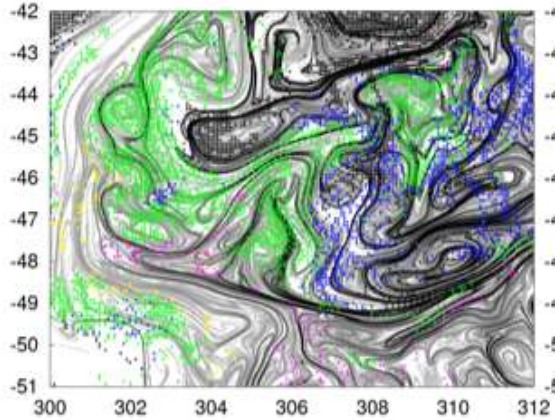



tara POLAR STATION

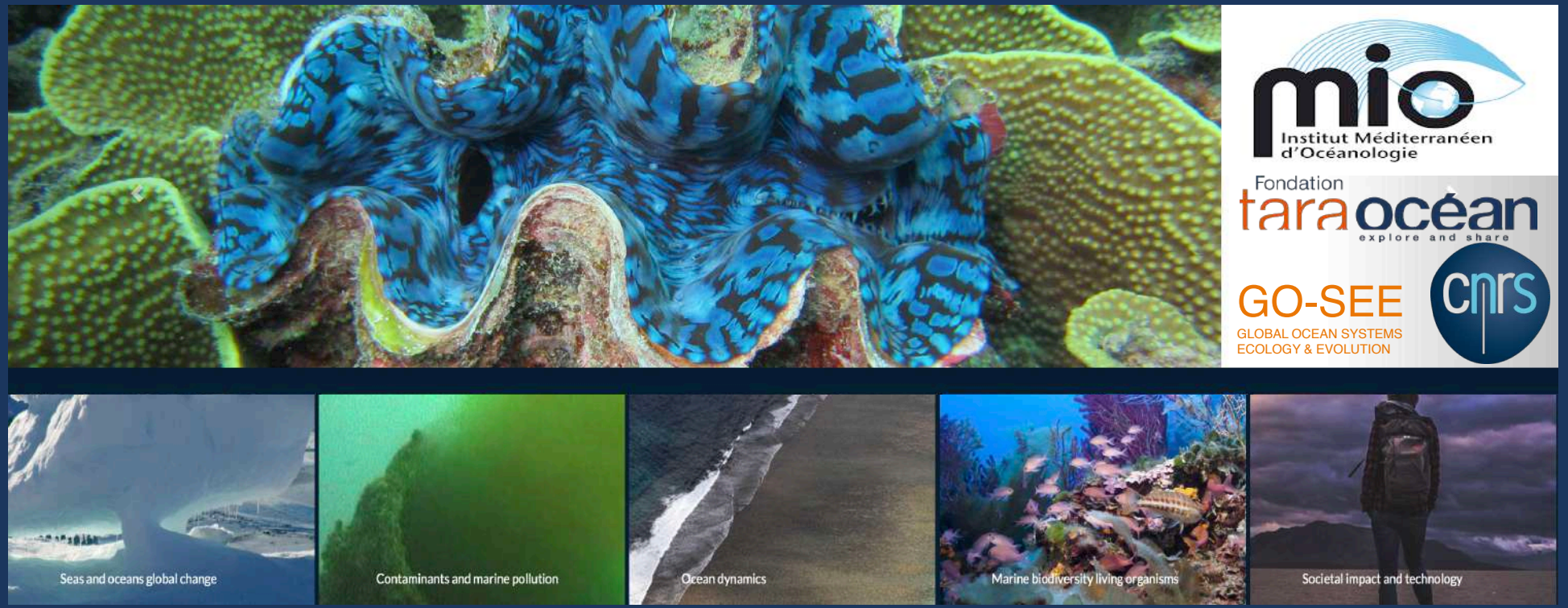
ORA






- Sequence abundance included soon
- Scale (from TB to PB)
- Reduce (environmental impact)
- Smart (queries and answers)
- Deploy (distribute indexes)
- Apply to all sequencing projects (environment, health, agronomy)




http://tara-oceans.mio.osupytheas.fr