User Guide

If you use this web service, please cite:

1) The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. C. Vernette, J. Lecubin, P. Sanchez, Tara Oceans Coordinators, S. Sunagawa, T.O. Delmont, S.G. Acinas, E. Pelletier, P. Hingamp, M. Lescot. (2022) *Nucleic Acides Research*. gkac420, https://doi.org/10.1093/nar/gkac420

2) The Ocean Gene Atlas: exploring the biogeography of plankton genes online. E. Villar, T. Vannier, C. Vernette, M. Lescot, M. Cuenca, A. Alexandre, P. Bachelerie, T. Rosnet, E. Pelletier, S. Sunagawa, P. Hingamp. (2018). *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W289–W295, https://doi.org/10.1093/nar/gky376

Last updated: 2022/05/13

User Guide

Overview 2
I) Ocean Gene Atlas workflow 3
II) Submission interface 3
II.1) Definition of the query 3
II.2) Analysis parameters 4
II.3) Accessory administrative parameters 5
III) Results interface 6
III.1) Job details 6
III.2) E-values distribution 6
III.3) Maps 7
III.4) Bubble plots 8
III.5) Taxonomic distribution 8
III.6) Phylogenetic analysis 9
III.7) Downloading publication grade figures 11
IV) Interpretation of results 11
V) Application Programming Interface (API) 11
VI) References 12
VI.1 Environmental context files 12
VI.2 Gene catalogs and sequencing reads 13
VI.3 Literature cited 13

Overview

Ocean Gene Atlas 2.0 (OGA) is a web service to explore the biogeography of marine genes based on sequence similarity with environmental genomics datasets (Fig. 1). OGA was first implemented with the *Tara* Ocean Microbiome - Reference Gene Catalog database (OM-RGC; Sunagawa et al. 2015) and the Marine Atlas of *Tara* Ocean Unigenes (MATOU; Carradec et al. 2018). Gene abundance estimates are computed for DNA metagenomes from the smallest *Tara* Oceans size fractions (from 0 to 3 µm, OM-RGC), and for RNA metatranscriptomes from *Tara* Oceans larger size fractions (0.8 to 2000 µm, MATOU). OGA 2.0 was updated recently and includes the version 2 of OM-RGC (Salazar et al. 2019), curated Tara Oceans Eukaryotic Metagenome and Single-Cell Assembled Genomes (MAGs and SAGs ; Delmont et al. 2021), MetaGenomics-based Transcriptomes (MGTs ; (Vorobev et al. 2020), 530 metagenome-assembled bacterial and archaeal genomes from the *Tara* polar circle expedition (Arctic MAG+G ; Royo-Llonch et al. 2021), 1,888 Bacterial and Archaeal Genomes (BAC_ARC_MAGs **;** Delmont, Pierella Karlusich, et al. 2021) and Malaspina Deep Metagenome Assembled Genomes (Mdeep-MAGs; Acinas et al. 2021).

We plan to update the website gradually as other marine gene catalogs are released .

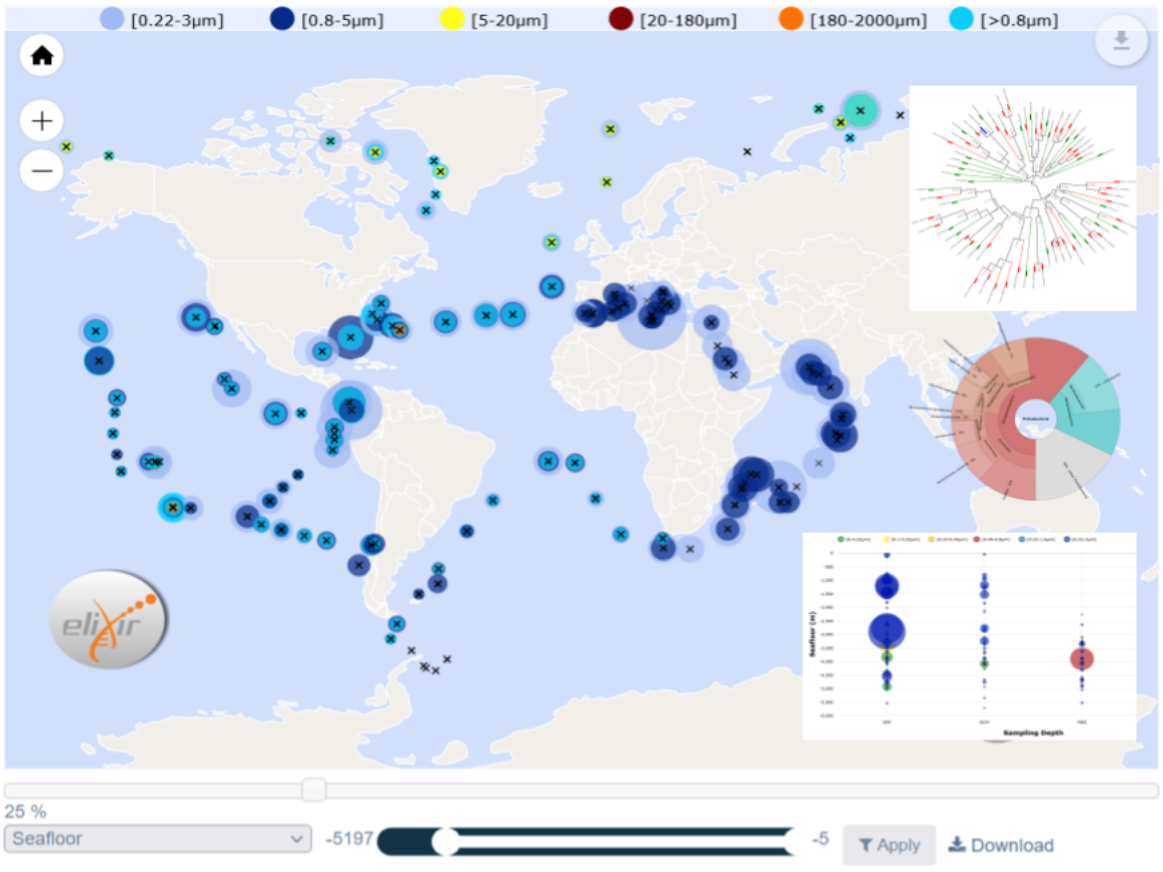

Figure 1: Interactive Ocean Gene Atlas results

If you use this web service, please cite:

1) The Ocean Gene Atlas: exploring the biogeography of plankton genes online. E. Villar, T. Vannier, C. Vernette, M. Lescot, A. Alexandre, P. Bachelerie, T. Rosnet, E. Pelletier, S. Sunagawa, P. Hingamp. *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W289–W295, https://doi.org/10.1093/nar/gky376

2) The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. (Vernette et al, submitted)

URL: http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/
Contact: oceangeneatlas@mio.osupytheas.fr

The following browsers have been tested and are listed by decreasing order of compatibility with the interactive displays in the OGA result panels:

1. Firefox (on Linux, Windows and Mac OS)
2. Chrome (on Linux, Windows and Mac OS)
3. Microsoft Edge (Windows)
4. Safari (Mac OS).
5. Microsoft Internet Explorer (Windows, Mac OS)

We recommend using Firefox to query OGA (100% operational).

I) Ocean Gene Atlas workflow

OGA imports heterogeneous datasets in order to present an integrated explorative display of the quantitative distribution of genes in the oceans (Fig. 2). Field campaigns (blue) have collected plankton biosamples and measured *in situ* environmental parameters. The OGA 2.0 web server (yellow) combines the following data published by distinct archives (pink): EBI ENA for sequencing reads, published articles and companion websites for gene catalogs and taxonomic annotations, PANGAEA for contextual environmental data.

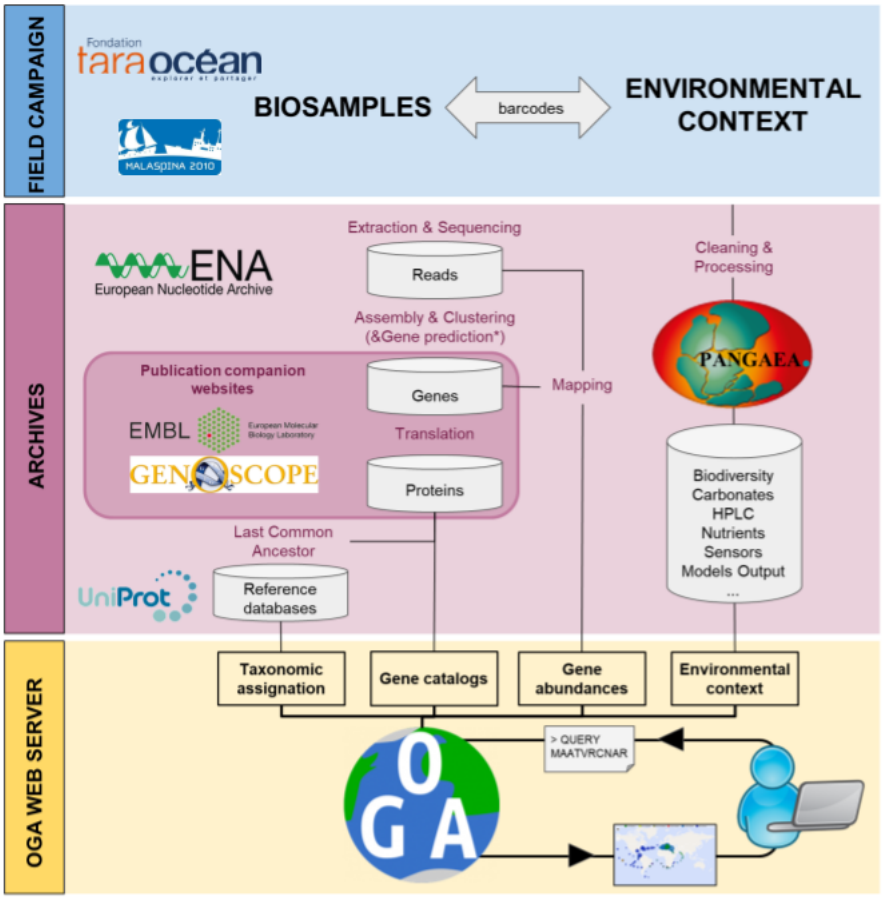

Figure 2: Data sources for the Ocean Gene Atlas workflow

II) Submission interface

II.1) Definition of the query

The input query may be one of the following (Fig. 3):
1/ a gene/protein sequence. The FASTA-formatted sequence with a header line should be pasted in the text field. Sequence type (nucleotide or protein) should be specified in the check box.
2/ a hidden Markov model profile (HMM) built from any protein alignment using hmmalign package (http://hmmer.org) with the default ASCII flat file format (either custom built or standard Pfam HMMs can be used).
3/ Previous result files (.tsv). Result files (sent by email if the email is provided in the submission form, or downloaded from the results page) can be uploaded in order to rebuild the results page in a few seconds (hence shunting the more lengthy similarity search step).
4/ a Pfam identifier (Pfam ID) following the format PFXXXXX, e.g. PF00111 for Ferredoxin. The OGA service downloads the Pfam HMM directly from the Pfam website. Pfam ID and entry annotation details can be found on https://pfam.xfam.org.
5/ a list of gene identifiers (e.g. unigenes/OM-RGC)

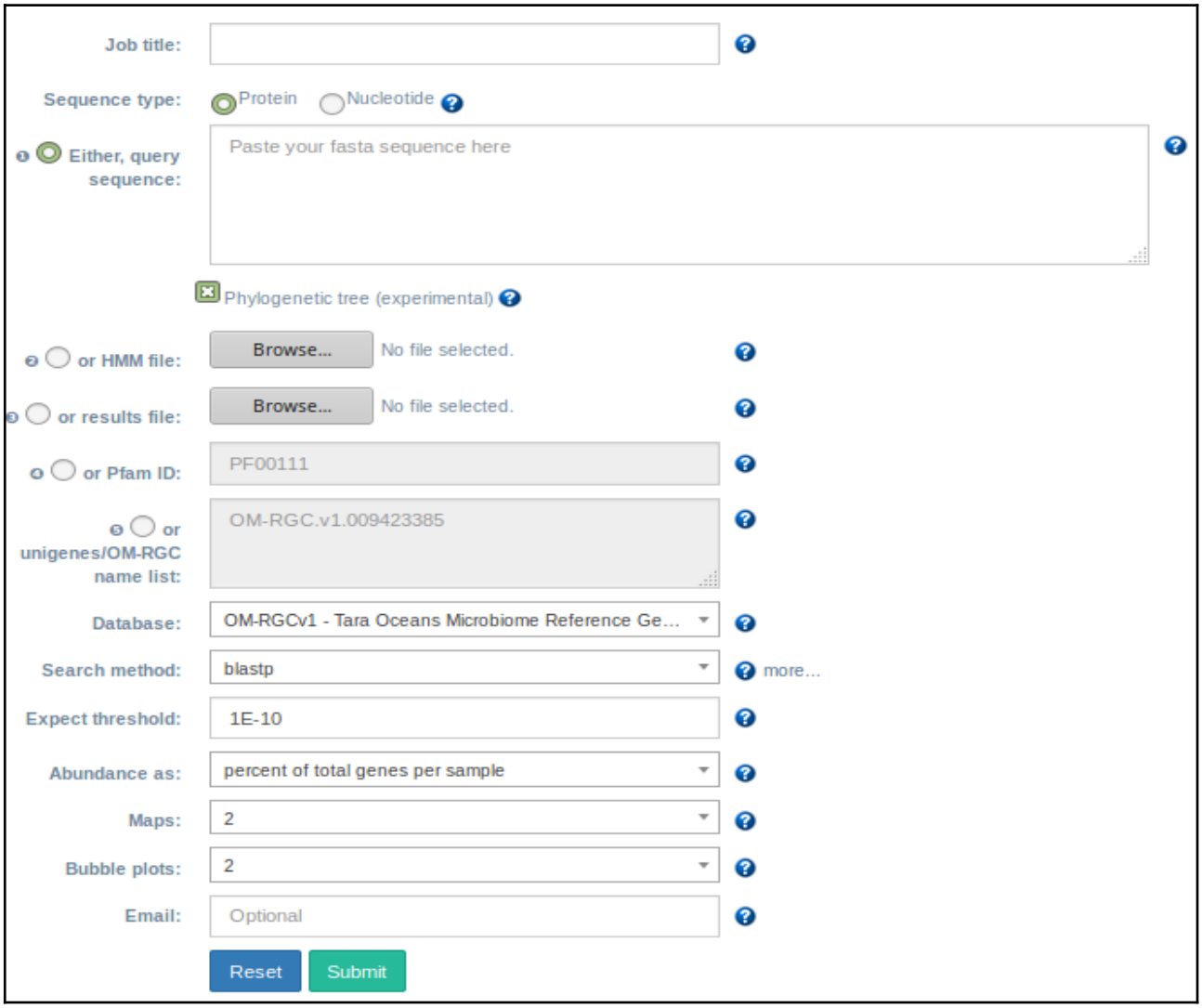

Figure 3: Submission form

II.2) Analysis parameters

The seven parameters that define the search method and output configuration are (Fig. 3):

- **Search method**: Choose a sequence similarity search method amongst: blast, (Altschul et al. 1997), Diamond (Buchfink, Xie, et Huson 2015) and hmmer (Eddy 2011). Each program has a specific computing speed and alignment sensitivity, see Steinegger et Söding (2017) for advice. For sequence-based queries, we recommend blastp (the default) which offers a very good sensitivity/speed tradeoff (usually returns results in less than 20 seconds for the OM-RGC catalog, and 120-300 seconds for the MATOU catalog).

- **Dataset**:
  1) *Tara* Oceans Microbiome Reference Gene Catalog (OM-RGC; Sunagawa et al. 2015)
  2) *Tara* Oceans Microbiome Reference Gene Catalog with arctic data (OM-RGCv2; Salazar et al. 2019)
  3) Marine Atlas of *Tara* Ocean Unigenes (MATOU; Carradec et al. 2018)
  4) *Tara* Oceans Single-Cell and Metagenome Assembled Genomes (EUK-SMAGs; (Delmont, Gaia, et al. 2021)
  5) Metagenome-assembled bacterial and archaeal genomes from *Tara* Polar Circle expedition (Arctic MAG; Royo-Llonch et al. 2021)
  6) Metagenomics-based transcriptomes from *Tara* Oceans expedition (MGTs; (Vorobev et al. 2020)
  7) *Tara* Oceans Bacterial and Archaeal Genomes (BAC_ARC_MAGs; Delmont, Pierella Karlusich, et al. 2021)
  8) Malaspina Deep Bacterial Genomes (MDeep-MAGs; Acinas et al. 2021)

More datasets will be added as they become available.

- **Expect threshold**: define the maximum allowed E-value above which similar sequences from the dataset are excluded from the analysis.

- **Abundance as**: select the per sample homologs abundance normalization method.
The abundance of each catalog gene (for OM-RGCv1 and MATOU) in specific biosamples was estimated by evaluating the coverage of raw sequencing reads mapped to the gene's nucleotide sequence. Briefly, depending on the database queried,, abundance estimates may be expressed in one of three available normalization schemes: (i) the gene's read coverage is divided by the sum of the total gene coverages for the sample ('percent of total coverage'), (ii) the gene's read coverage is divided by the total number of reads for the sample ('percent of total reads'), (iii) the gene's read coverage is divided by the median of the coverages of a set of 10 universal single copy marker genes ('average copies per cell'). The ten single marker genes used are: COG0012, COG0016, COG0018, COG0172, COG0215, COG0495, COG0525, COG0533, COG0541, COG0552.
In order to estimate the abundance and expression of each MGT unigene in each sample, cleaned reads (from metagenomes and metatranscriptomes) were mapped against the reference catalog as described in (Vorobev et al. 2020). Reads covering at least 80% of read length with at least 95% of identity were retained for further analysis. Unigene expression values and genomic occurrences were computed in RPKM (reads per kilobase covered per million of mapped reads). Gene abundance from MAG catalogs was computed using reads per genomic kilobase and metagenomic gigabase (RPKG).

For Euk_SMAGs, BAC_ARC_MAGs and Arctic MAGs gene abundance, we attributed to gene its MAGs abundance computed as described in (Delmont, Gaia, et al. 2021; 2020;

Royo-Llonch et al. 2021). For the MDeep-MAGs dataset, the abundance of each MAG was expressed by the number of mapped reads per genomic kilobase and sample gigabase as described in (Acinas et al. 2021). And each gene abundance is expressed as mean read coverage (best read map, with at least 95% identity over at least 90% of the read length).

- **Phylogenetic tree:** select this option to enable a phylogenetic analysis of your query sequence in context of its metagenomic and RefSeq homologs. This option is not available when submitting metagenome gene identifiers or nucleotide sequence queries.

- **Maps**: choose the number of maps used to visualize the geographical distribution of the homologs (each map can display homolog abundance in distinct size fractions and distinct depths).

- **Bubble plots:** choose the number of plots used to visualize co-variation of homologs abundances with different environmental parameters (each bubble plot can display co-variation in distinct size fractions and for distinct environmental features).

II.3) Accessory administrative parameters

Two parameters are accessory:
- Job title: a free text will be used to annotate and name downloadable files.
- Optional email address: if provided, the output of the similarity search will be attached to an email sent to the user. This results file can then be used to generate the results page without having to recompute the similarity search (i.e. saves user and server time when one wishes to access the interactive plots described below). A hyperlink to the results page will be provided at the time of data submission and also included into the optional email. The results will remain available online for 15 days. If you wish to visualize your results after this delay, you may download the results file and resubmit it using option N°3 (see II.1 above).

III) Results interface

⚠ Please enable pop-ups in your browser so you can access to the results interface after the query.

The results interface displays all the computed results via maps, bubble plots and Krona pie-chart. The results are organized by sample (except for the overall Krona pie-chart), the identity of which are available on mouse hover over the colored circles on the maps and bubble plots. The results will be available on the web page URL for 48 hours after job submission.

Attention the abundances have been rounded to 10-12 it is possible to obtain homolog without the associated abundance.

III.1) Job details

The top panel (Fig. 4) provides information about the submitted job (e. g. the shareable URL of results page, the E-value threshold etc.) and a summary of the similarity search results (number of genes hit and associated with abundance estimates). Three sets of text files that encapsulate the full dataset required to reproduce the figures are available for download:

- the list of similarity search hits (gene identifiers and E-values),
- the corresponding FASTA formatted sequences of the hits (DNA & proteins),
- the gene x biosample abundance matrix and contextual environmental features for each biosample.

- when the EUK-SMAGs have been selected on the request page, it is possible to download the first three EUK-SMAGs which contain the most abundant genes

Figure 4: Summary of results

III.2) E-values distribution

The bar chart (Fig. 5) displays the distribution of the hits E-values and allows the user to adjust the homolog inclusion threshold. One can change the range of E-values by selecting the chosen range directly in the histogram, and then clicking on the “Apply” button to update all the maps, bubble plots and Krona pie-charts.

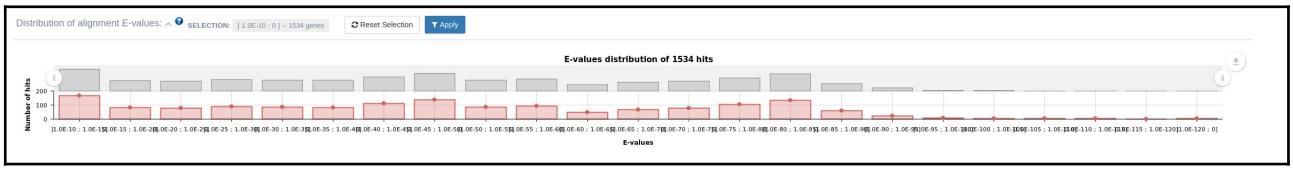

Figure 5: Dynamic E-values bar chart

III.3) Maps

In the interactive geographical maps (Fig. 6), each circle represents the abundance of selected homologs in one *Tara* Oceans sample. The abundance is estimated by the number of raw sequencing read nucleotides mapped to each gene from a gene catalog using MOCAT (Kultima et al. 2012), normalized by one of the two methods selected in the submission form (normalization method is recalled in the job details panel, see II.2 above for the description of the normalization schemes).
Different size fractions and sampling depths can be displayed on the maps by selecting the corresponding options above each map. Each size fraction is associated with a distinct color. The *Tara* Oceans sampling protocol for prokaryotes changed slightly during the cruise, shifting from 0.2-1.6µm to 0.2-3µm size fractions from the Indian Ocean onwards. OGA’s color codes are close shades of blue to remind the users that both fractions correspond to the major prokaryotes size fraction.
The size of the circles may be tuned using the interactive slider. A scale - entitled "abundance" - is displayed in the map in order to be able to compare several independent results with circles representing the maximum and minimum abundance as well as their numerical values.
A click on a given sample circle will open a Krona taxonomic distribution pie-chart specific for the selected sample (see III.5 below). The different acronyms of sampling depth stand for: DCM: deep chlorophyll maximum layer; SRF: upper layer zone; MES: mesopelagic zone; MIX: marine epipelagic mixed layer, FSW: filtered sea water, ZZZ: marine water layer. Using the top-right button, users may edit, print and/or download the map in several formats (see III.6 below).

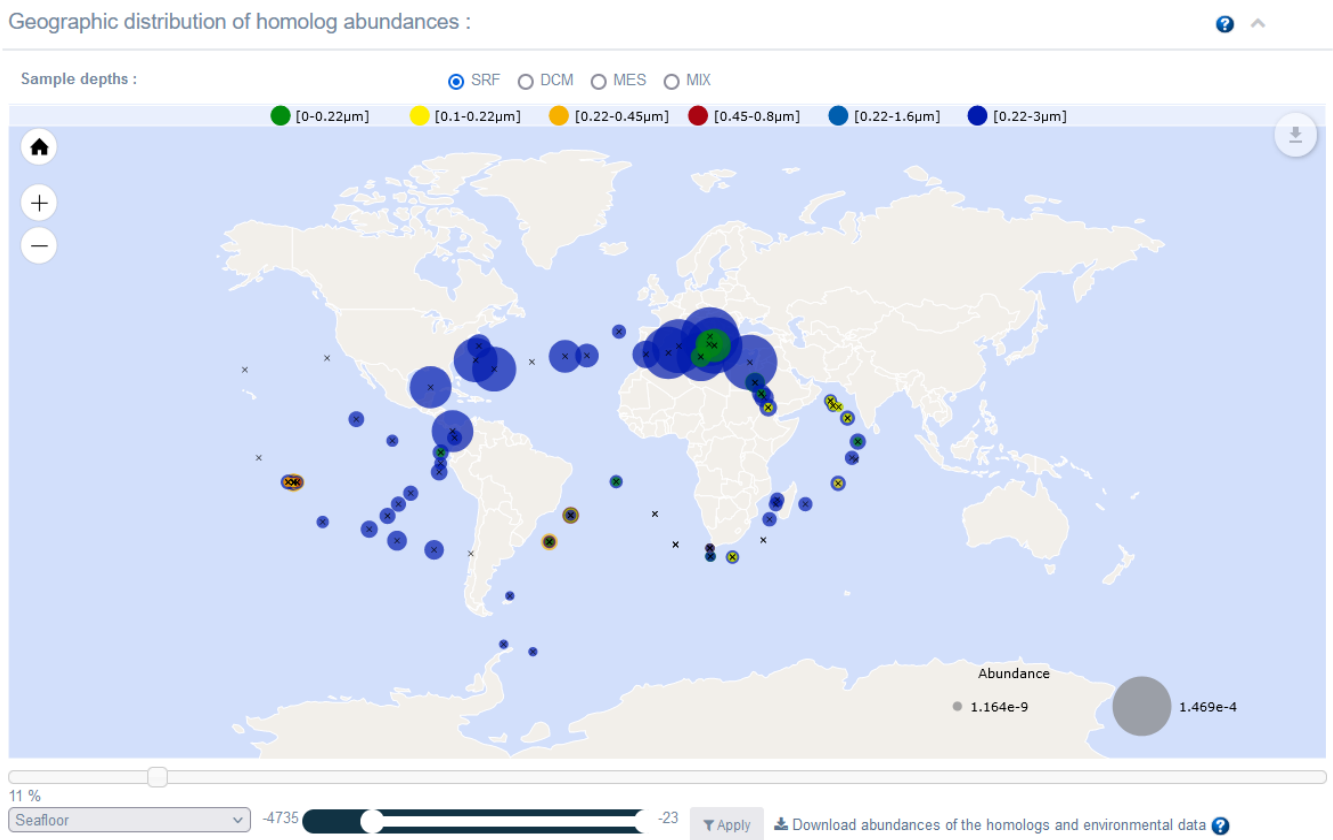

Figure 6: Interactive world maps

Users can choose an environment variable from a list. Define the maximum and minimum values using the slider. When the “Apply” button is clicked, only samples corresponding to the selected range are displayed on the map. It is possible to download the abundance files and environmental variables corresponding to the selection.

III.4) Bubble plots

The bubble plots associate environmental context with homologs abundance for each sampling depth (Fig. 7). A drop-down menu allows the user to change the displayed environmental parameter.

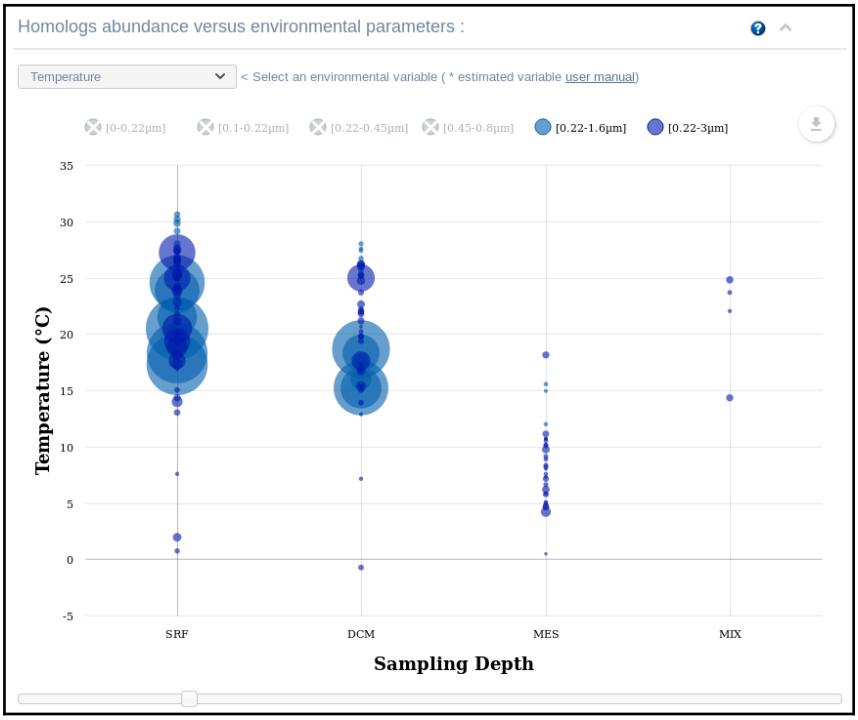

Figure 7: Bubble plots representing the co-variation of gene abundances and an environmental feature (e.g. Mean temperature) for each depth and size fraction combination

The different acronyms of sampling depth stand for: DCM: deep chlorophyll maximum layer; SRF: upper layer zone; MES: mesopelagic zone; MIX: marine epipelagic mixed layer, FSW: filtered sea water, ZZZ: marine water layer. Comprehensive detailed descriptions of the biosamples' environmental context can be found in the resources listed under IV.1 below.

Similarly to the geographical maps above, the sizes of the sample circles are proportional to the abundance of the query homologs. The circles are color coded according to the selected fractions. The y-axis represents the environmental parameter value: Alkalinity, Ammonium_5m*, Carbon Total, CDOM*, Chlorophyll_A, CO2, CO3, Density, Depth, Distance_coast, HCO3, Iron_5m*, Nitrate_5m*, Nitrite_5m*, NO2, NO3, NO3_NO2, NPP_C*, O2, PAR, pH, PIC*, PO4, POC*, Salinity, Si, Temperature. Values estimated from oceanographic models are indicated by a star. Comprehensive detailed descriptions of the biosamples' environmental context can be found in the resources listed under IV.1 below.

III.5) Taxonomic distribution

A general Krona pie-chart (Ondov, Bergman, et Phillippy 2011) at the bottom of the results page presents an overview of the abundance weighted taxonomic distribution of homologous sequences in all samples (Fig. 8). The diagram allows taxonomic data to be explored with a zoomable multi-layered pie-chart.
To explore homolog taxonomies for each distinct biosample, click on the corresponding circles in the geographic maps (see III.3 above).

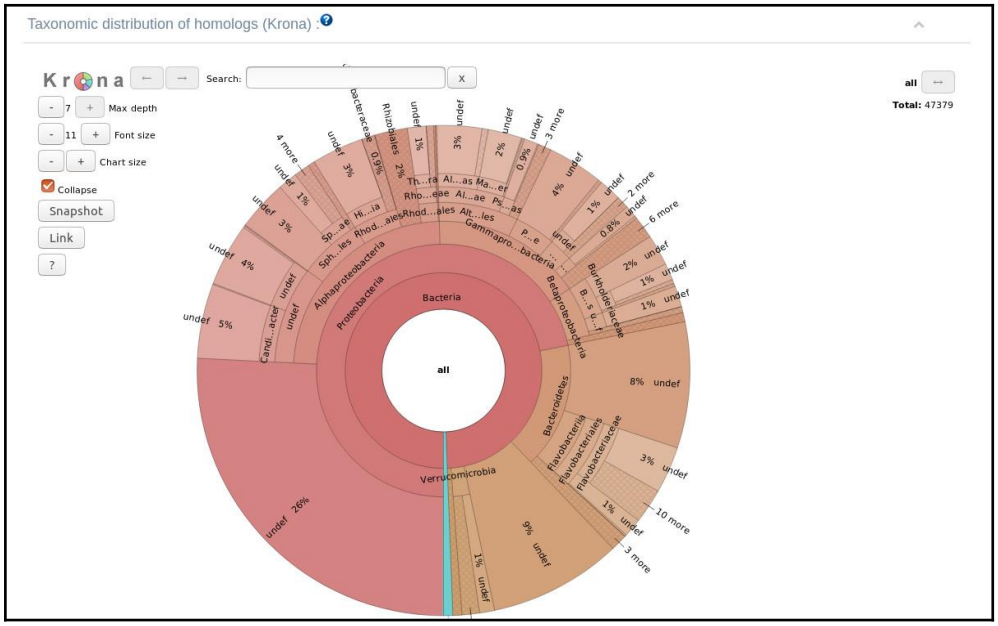

Figure 8: Krona pie-chart representing the taxonomic distribution of homologous sequences in all samples

The all genes option makes it possible to visualize the taxonomic distribution of genes with one or more abundances less than 10^{-12}.
For more information on Krona, see : https://github.com/marbl/Krona/wiki/Browsing%20Krona%20charts.

III.6) Phylogenetic analysis

By ticking the Phylogenetic tree option in the OGA submission form, an additional section in the results page will display a phylogenetic tree putting the user query in context of its metagenomic and reference homologs.
To this end, the user query is used to search homologs from the RefSeq reference database. If the number of RefSeq homologs is greater than the number of metagenomic homologs, then RefSeq homologs are progressively clustered with CD-HIT (Li et Durbin 2009) until their number is equal or less than that of metagenome homologs (in order to avoid the resulting tree to be too biased towards RefSeq homologs). This clustering is done iteratively by gradually decreasing the threshold of clustering from 100% to a minimum of 60%. The sequences in the full dataset - consisting of the user query, the metagenomic homologs, and the reference RefSeq homologs - are then aligned with MAFFT (v7.407) (Katoh et al. 2002). This alignment is finally cleaned with MaxAlign (v1.1) (Gouveia-Oliveira, Sackett, et Pedersen 2007) and TrimAl (v1.4.rev22) (Capella-Gutiérrez, Silla-Martinez, et Gabaldon 2009) before submission to FastTree (v2.1.10) (Price, Dehal, et Arkin 2010) for phylogenetic tree inference. The resulting tree is displayed (Fig.10) thanks to the javascript library plylotree.js (Shank, Weaver, et Kosakovsky Pond 2018). Several interactive display options are available to the user.

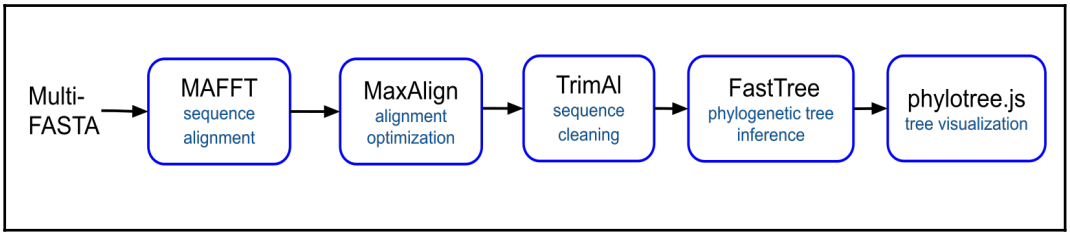

Figure 9: Phylogenetic pipeline

Once the phylogeny workflow has completed successfully, the resulting phylogenetic tree is rendered in the results interface (Fig. 10) together with associated phylogeny options (Fig. 11). In the tree, the user query sequence is represented in blue, the metagenome homologs appear in red, and the RefSeq reference homologs are labelled in green.
It is possible to download the tree in SVG format as well as all intermediate files used in the workflow (homologs multi-FASTA, multiple alignment, newick format tree). It is also possible to interact with the rendering of the tree (radial / linear), to change the substitution mode and gamma law correction, to root the tree (with the longest branch or branch specified by the user) and zoom in and out. The colored multiple sequence alignment with selected positions (as output by Trimal) can also be displayed ("View multiple alignment"). You can get a subtree by selecting branches (Click on a node, select "All descendant branches", and click on the recompute tree button).

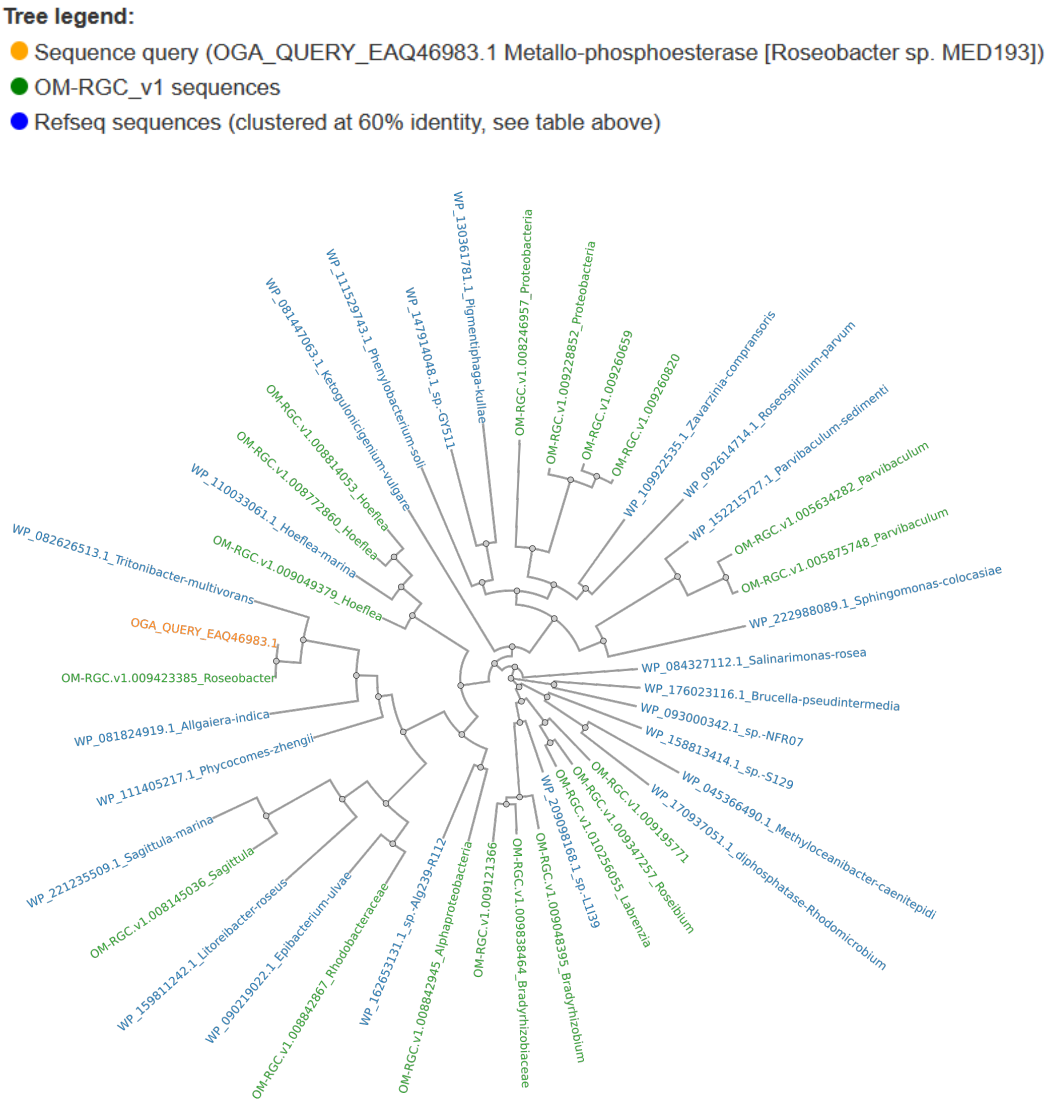

Figure 10: Phylogenetic tree

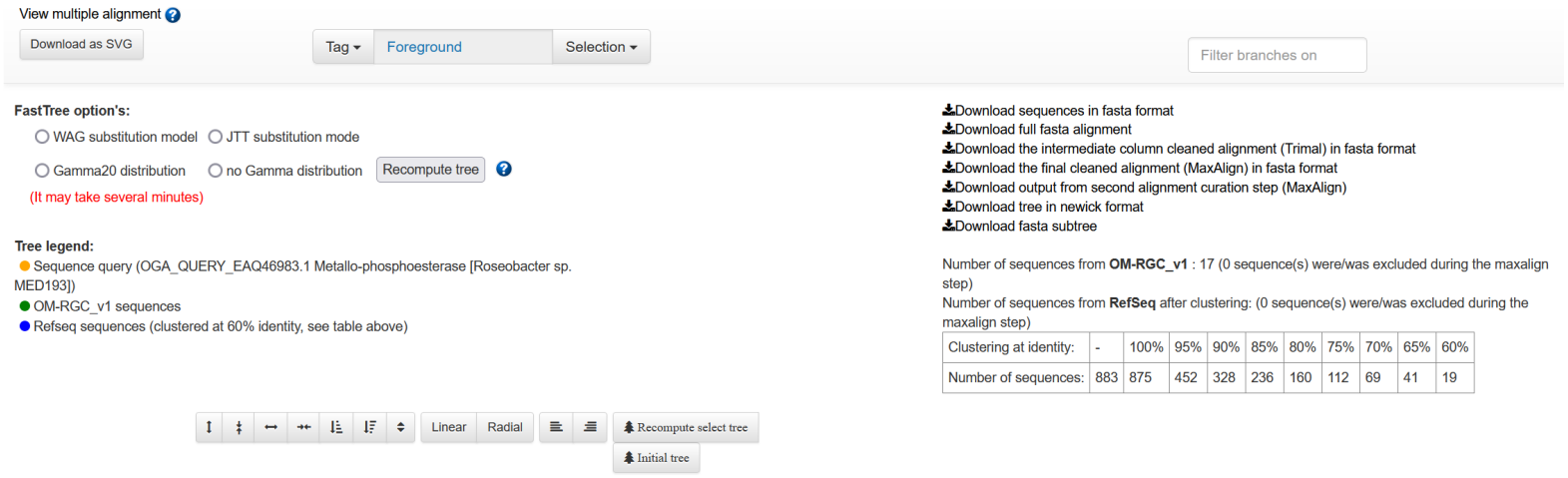

Figure 11: Phylogenetic analysis options

Command line details :

MAFFT options :

- if the number of sequences is less than 2000 : default options
- 2000 < nbseq <10,000 --retree 1 -maxiterate 0
- nbseq > 10,000 --retree 1 -maxiterate 0 --nofft --parttree

MaxAlign : maxalign.pl -v=1 -w -f=$input $output

trimAl : trimal -in $input -out $output -htmlout $Html -strict

III.7) Downloading publication grade figures

Click on the download arrow at the top right of each display panel (Fig. 12) to download in Scalable Vector Graphics format suitable for high resolution post-treatment and publication. For the Krona charts, click on the “Snapshot” button which will open the pie chart in a separate window, then save as .svg file.

Figure 12: Download figures as Scalable Vector Graphics

IV) Interpretation of results

To help with the interpretation of the results, the following case study reproduces the study carried out by (Sebastián et al. 2016).

Upon phosphorus deficiency, bacterioplankton have established a widespread strategy of

replacing membrane phospholipids with alternative non-phosphorus lipids. Sebastián et al. have shown that this response is conserved among diverse marine heterotrophic bacteria. Several experiments of mutagenesis and complementation have then confirmed the roles of the phospholipase C (PlcP) and a glycosyltransferase in lipid modelling. Analyses of metagenome datasets such as the Global Ocean Sampling (GOS) and Tara Oceans have confirmed that PlcP is abundant in low phosphate concentrations areas.

The results described below are obtained by clicking on the hyperlink of the OGA submission form entitled "*Try an Example with OM-RGC dataset (prokaryotes)*", entering "1e-40" in the "E-value threshold field, followed by "Submit". This uses the same phospholipase C (EAQ46983) as a BLASTp query sequence with the author's e-value threshold of $1e^{-40}$ to search for homologous sequences in the OM-RGC catalog (the same metagenome dataset used by (Sebastián et al. 2016).

The 922 PlcP homologs identified shows higher abundances in Mediterranean subsurface samples (geographical maps panel, after selection of "SRF" depth and [0.2-1.6 µm] & [0.2-3 µm] size fractions) related to low phosphorus concentration (environmental bubble plots panel, after selection of "PO4" in the dropdown menu) and mostly originated from Proteobacteria and Bacteroidetes (Krona taxonomy panel), which agrees with the previously published interpretations of Sebastián et al. that marine heterotrophic bacteria display reduced phosphorus requirements upon phosphorus deficiency by PlcP-mediated replacement of membrane phospholipids by alternative non-phosphorus lipids.

V) Application Programming Interface (API)

Three types of queries are accessible:
- the submit request packaged in a JSON file with the search parameters. Two options are available, sequence or pfam id. The server sends a response in JSON format with the identifier of the analysis and an estimation of the calculation time.
- the checkResults request with the identifier of the analysis. The server returns the URL of the result web page.
- the fetchResults request with the name of the result file and identifier file of the analysis.
Three files are possible: alignment result, homolog sequences or abundances & environmental data.

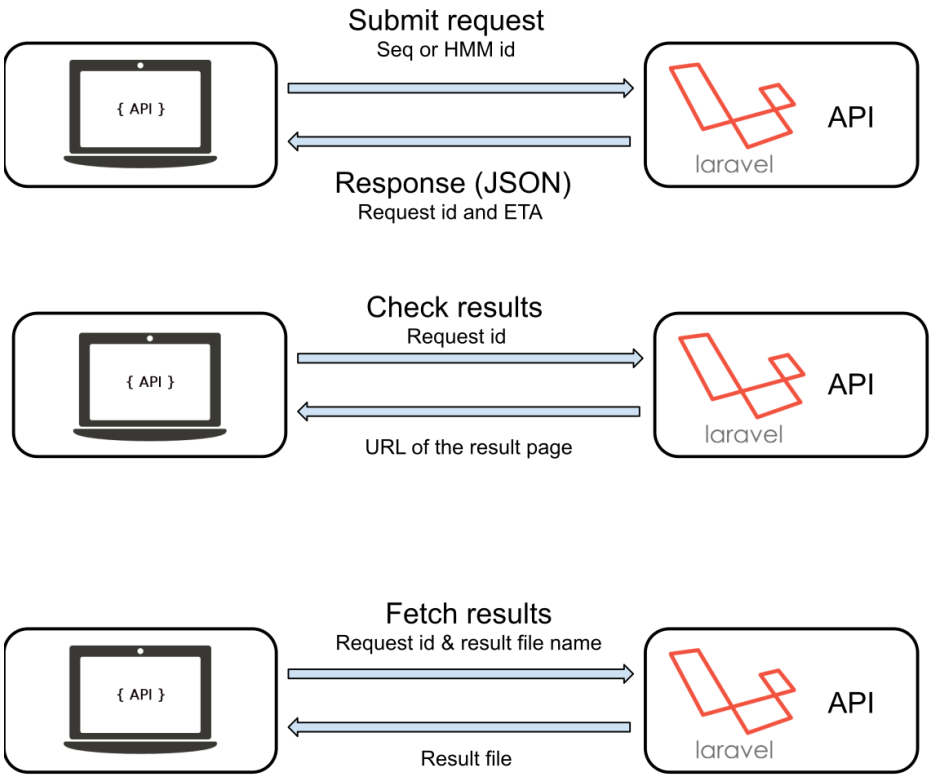

Figure 13: The three types of API requests

A tutorial with examples is available at the following address :
https://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/build/script/API_tutorial.pdf

A limit is set at 200 jobs per 24 hours and queries launched on the web interface have priority.

VI) References

VI.1 Environmental context files

Registry of all the samples from the *Tara* Oceans Expedition (2009-2013) have been deposited at *PANGAEA* : https://doi.org/10.1594/PANGAEA.875582. The environmental variables listed in Table 1 were retrieved from the following databases:
BIODIV: https://doi.org/10.1594/PANGAEA.853809
CARB: https://doi.org/10.1594/PANGAEA.875567
HPLC: https://doi.org/10.1594/PANGAEA.875569
MESOSCALE: https://doi.org/10.1594/PANGAEA.875577
NUT: https://doi.org/10.1594/PANGAEA.875575
SENSORS: https://doi.org/10.1594/PANGAEA.875576
SEQUENCING: https://doi.org/10.1594/PANGAEA.875581
WATERCOLUMN: https://doi.org/10.1594/PANGAEA.875579

Any additional variable deposited in PANGAEA database can be added to OGA upon request at oceangeneatlas@mio.osupytheas.fr

VI.2 Gene catalogs and sequencing reads

OM-RGC catalog: http://ocean-microbiome.embl.de/companion.html

OM-RGC reads: https://www.ebi.ac.uk/ena/data/view/PRJEB7988

OM-RGCv2 catalog: https://www.ebi.ac.uk/biostudies/studies/S-BSST297

BAC_ARC_MAGs, MATOU, EUK-SMAGs and MGTs catalogs:
https://www.genoscope.cns.fr/tara/

MATOU reads: https://www.ebi.ac.uk/ena/data/view/PRJEB6609

MGTs read: https://www.ebi.ac.uk/ena/browser/view/PRJEB4352

Arctic MAGs: https://www.ebi.ac.uk/biostudies/studies/S-BSST451 and
https://www.ebi.ac.uk/biostudies/studies/S-BSST451

Malaspina Deep MAGs:
https://www.ebi.ac.uk/ena/data/view/PRJEB40454
https://www.ebi.ac.uk/ena/data/view/PRJEB44456
https://www.ebi.ac.uk/biostudies/studies/S-BSST45

VI.3 Literature cited

Acinas, Silvia G., Pablo Sánchez, Guillem Salazar, Francisco M. Cornejo-Castillo, Marta Sebastián, Ramiro Logares, Marta Royo-Llonch, et al. 2021. « Deep Ocean Metagenomes Provide Insight into the Metabolic Architecture of Bathypelagic Microbial Communities ». *Communications Biology* 4 (1): 1-15. https://doi.org/10.1038/s42003-021-02112-2.

Altschul, S F, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, et D J Lipman. 1997. « Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. » *Nucleic Acids Research* 25 (17): 3389-3402.

Buchfink, Benjamin, Chao Xie, et Daniel H. Huson. 2015. « Fast and Sensitive Protein Alignment Using DIAMOND ». *Nature Methods* 12 (1): 59-60. https://doi.org/10.1038/nmeth.3176.

Capella-Gutiérrez, Salvador, José M. Silla-Martinez, et Toni Gabaldon. 2009. « trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses - PubMed ». 2009. https://pubmed-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/19505945/.

Carradec, Q, Eric Pelletier, C Da Silva, Y Seeleuthner, et P Wincker. 2018. « A global ocean atlas of eukaryotic genes. - PubMed - NCBI ». 2018. https://www-ncbi-nlm-nih-gov.inee.bib.cnrs.fr/pubmed/29371626.

Delmont, Tom O., Morgan Gaia, Damien D. Hinsinger, Paul Fremont, Chiara Vanni, Antonio Fernandez Guerra, A. Murat Eren, et al. 2021. « Functional Repertoire Convergence of Distantly Related Eukaryotic Plankton Lineages Revealed by Genome-Resolved Metagenomics ». https://doi.org/10.1101/2020.10.15.341214.

Delmont, Tom O., Juan José Pierella Karlusich, Iva Veseli, Jessika Fuessel, A. Murat Eren, Rachel A. Foster, Chris Bowler, Patrick Wincker, et Eric Pelletier. 2021. « Heterotrophic Bacterial Diazotrophs Are More Abundant than Their Cyanobacterial Counterparts in Metagenomes Covering Most of the Sunlit Ocean ». *The ISME Journal*, octobre, 1-10. https://doi.org/10.1038/s41396-021-01135-1.

Eddy, Sean R. 2011. « Accelerated Profile HMM Searches ». *PLOS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Gouveia-Oliveira, Rodrigo, Peter W. Sackett, et Anders G. Pedersen. 2007. « MaxAlign:

maximizing usable data in an alignment ». *BMC Bioinformatics* 8 (1): 312. https://doi.org/10.1186/1471-2105-8-312.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, et Takashi Miyata. 2002. « MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform ». *Nucleic Acids Research* 30 (14): 3059-66. https://doi.org/10.1093/nar/gkf436.

Kultima, Jens Roat, Shinichi Sunagawa, Junhua Li, Weineng Chen, Hua Chen, Daniel R. Mende, Manimozhiyan Arumugam, et al. 2012. « MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit ». *PLOS ONE* 7 (10): e47656. https://doi.org/10.1371/journal.pone.0047656.

Li, Heng, et Richard Durbin. 2009. « Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform ». *Bioinformatics (Oxford, England)* 25 (14): 1754-60. https://doi.org/10.1093/bioinformatics/btp324.

Ondov, Brian D., Nicholas H. Bergman, et Adam M. Phillippy. 2011. « Interactive Metagenomic Visualization in a Web Browser ». *BMC Bioinformatics* 12 (septembre): 385. https://doi.org/10.1186/1471-2105-12-385.

Price, Morgan N., Paramvir S. Dehal, et Adam M. Arkin. 2010. « FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments ». 2010. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490.

Royo-Llonch, Marta, Pablo Sánchez, Clara Ruiz-González, Guillem Salazar, Carlos Pedrós-Alió, Marta Sebastián, Karine Labadie, et al. 2021. « Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean ». *Nature Microbiology* 6 (décembre): 1-14. https://doi.org/10.1038/s41564-021-00979-9.

Salazar, Guillem, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Christopher M. Field, et al. 2019. « Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome ». *Cell* 179 (5): 1068-1083.e21. https://doi.org/10.1016/j.cell.2019.10.014.

Sebastián, Marta, Alastair F. Smith, José M. González, Helen F. Fredricks, Benjamin Van Mooy, Michal Koblížek, Joost Brandsma, et al. 2016. « Lipid Remodelling Is a Widespread Strategy in Marine Heterotrophic Bacteria upon Phosphorus Deficiency ». *The ISME Journal* 10 (4): 968-78. https://doi.org/10.1038/ismej.2015.172.

Shank, Stephen D., Steven Weaver, et Sergei L. Kosakovsky Pond. 2018. « phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics ». *BMC Bioinformatics* 19 (1): 276. https://doi.org/10.1186/s12859-018-2283-2.

Steinegger, Martin, et Johannes Söding. 2017. « MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets ». *Nature Biotechnology* 35 (octobre). https://doi.org/10.1038/nbt.3988.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. « Structure and function of the global ocean microbiome ». *Science* 348 (6237): 1261359. https://doi.org/10.1126/science.1261359.

Vorobev, Alexey, Marion Dupouy, Quentin Carradec, Tom O. Delmont, Anita Annamalé, Patrick Wincker, et Eric Pelletier. 2020. « Transcriptome Reconstruction and Functional Analysis of Eukaryotic Marine Plankton Communities via High-Throughput Metagenomics and Metatranscriptomics ». *Genome Research* 30 (4): 647-59. https://doi.org/10.1101/gr.253070.119.

Dataset	Read number	Nucleic file (Go)	Protein file (Go)	DB table (Go)	MAG number	Gene number	Sample number	DOI	ENA ID	BioStudies ID	Companion website	Metadata
OM-RGCv1	7.2 trillion	26	14	35.7	-	40 154 822	243	10.1126/ science.1261359	PRJEB7988	-	http://ocean-microbiome.embl.de/ companion.html	https:// doi.org/ 10.1594/ PANGAEA.875582
OM-RGCv2_metaG OM-RGCv2_metaT	113 billion 5 billion	42	38	7.8 18.9	370 187	46 775 154	180 187	10.1016/ j.cell.2019.10.014	S-BSST297	-	https://www.ocean-microbiome.org/	https:// doi.org/ 10.1594/ PANGAEA.875582
MATOUv1_metaG MATOUv1_metaT	185 billion 87 bilion	44	196	73.7 130.8	-	116 849 350	445 440	10.1038/s41467-017-02342-1	PRJEB6609	-	http://www.genoscope.cns.fr/ tara/	https:// doi.org/ 10.1594/ PANGAEA.875582
MGT	58 million	1.8	25	0.7	924	6 946 068	364	10.1101/ gr.253070.119	PRJEB4352 PRJEB6609	ERZ480625	http://www.genoscope.cns.fr/ tara/	https:// doi.org/ 10.1594/ PANGAEA.875582
EUK_SMAGs	280 billion	15	8.4	1.2	713	10 207 435	939	10.1101/2020.10.15.341214	PRJEB402	-	http://www.genoscope.cns.fr/ tara/	https:// doi.org/ 10.1594/ PANGAEA.875582
BAC_ARC_MAGs	280 billion	5.9	3.5	0.6	1 888	4 567 982	922	10.1038/s41396-021-01135-1			http://www.genoscope.cns.fr/ tara/ https://figshare.com/articles/dataset/Marine_diazotrophs/14248283	https:// doi.org/ 10.1594/ PANGAEA.875582
Arctic_MAGs_metaG Arctic_MAGs_metaT	140 million 45 million	1.3	0.4	0.1 0.1	530	1 033 381	68 53	10.1038/s41564-021-00979-9	PRJEB41575	S-BSST451	-	https:// doi.org/ 10.1594/ PANGAEA.875582
MDeep-MAGs	649 million	1.1	0.67	0.9	317	867 795	58	10.1038/s42003-021-02112-2	PRJEB40454 PRJEB44456	S-BSST457	https://malaspina-public.gitlab.io/malaspina-deep-ocean-microbiome/	
Total		137	273	271	4 929	227 401 987	3 899					

Table 1: Dataset information

Datasets	Percent of total coverage RPKM or RPKG	Percent of total reads	Average copies per cell
OM-RGCv1	x	x	x
OM-RGVv2	x		
MATOU	x	x	
MGT	x		
EUK_SMAGs	x		
BAC_ARC_MAGs	x		
Arctic_MAGs	x		
MDeep-MAGs	x		

Table 2 : Dataset abundance normalization methods