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

Overview

Ocean Gene Atlas (OGA) is a web service to explore the biogeography of marine genes based on sequence similarity with environmental genomics datasets (Fig. 1). OGA is currently implemented with the *Tara* Ocean Microbiome - Reference Gene Catalog database (OM-RGC; Sunagawa et al., 2015) (OM-RGC v2; Salazar et al.,2019) 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 also includes 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) and metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean (Arctic MAG+G ; Royo-Llonch et al.,2021).
. We plan to update the website gradually as other marine gene catalogs are released .

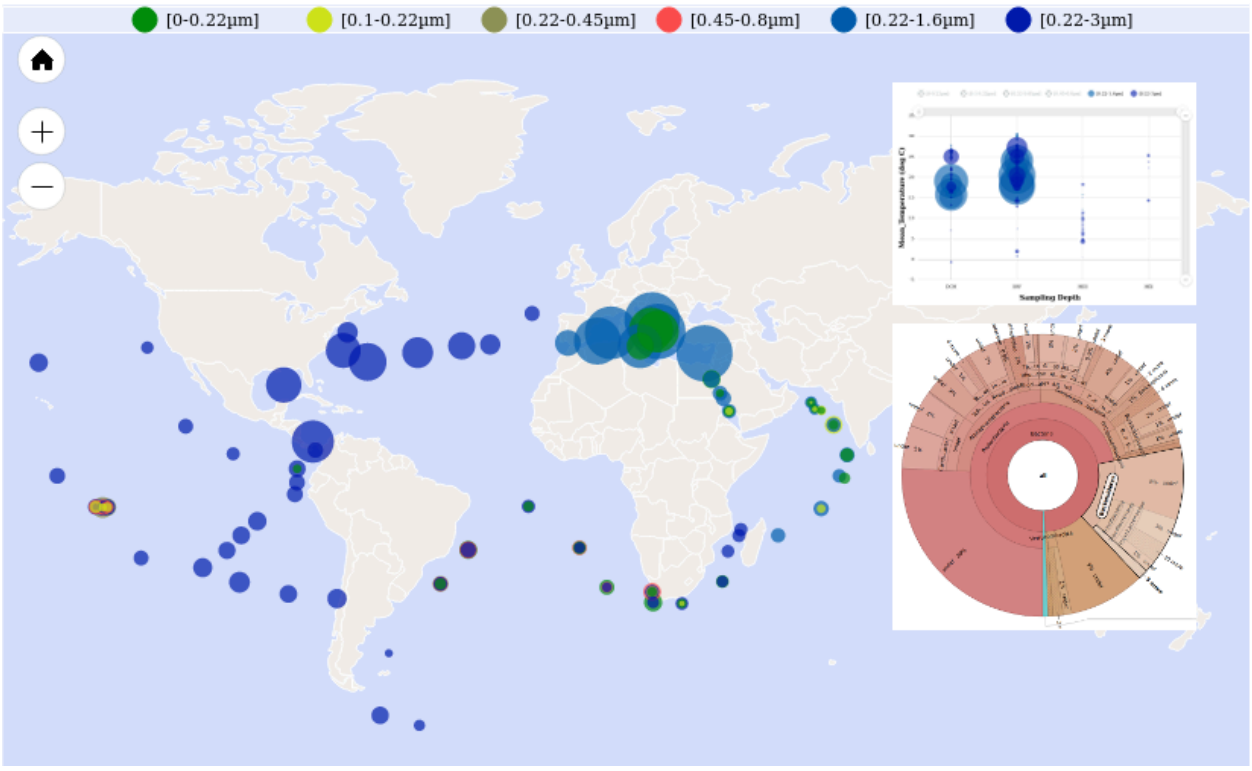

Figure 1: Interactive Ocean Gene Atlas results.

If you use this web service, please cite:

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.

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 web server (yellow) combines the following data published by distinct archives (pink): EBI ENA for sequencing reads, published articles companion websites for gene catalogs and taxonomic annotations, PANGAEA for contextual environmental data.

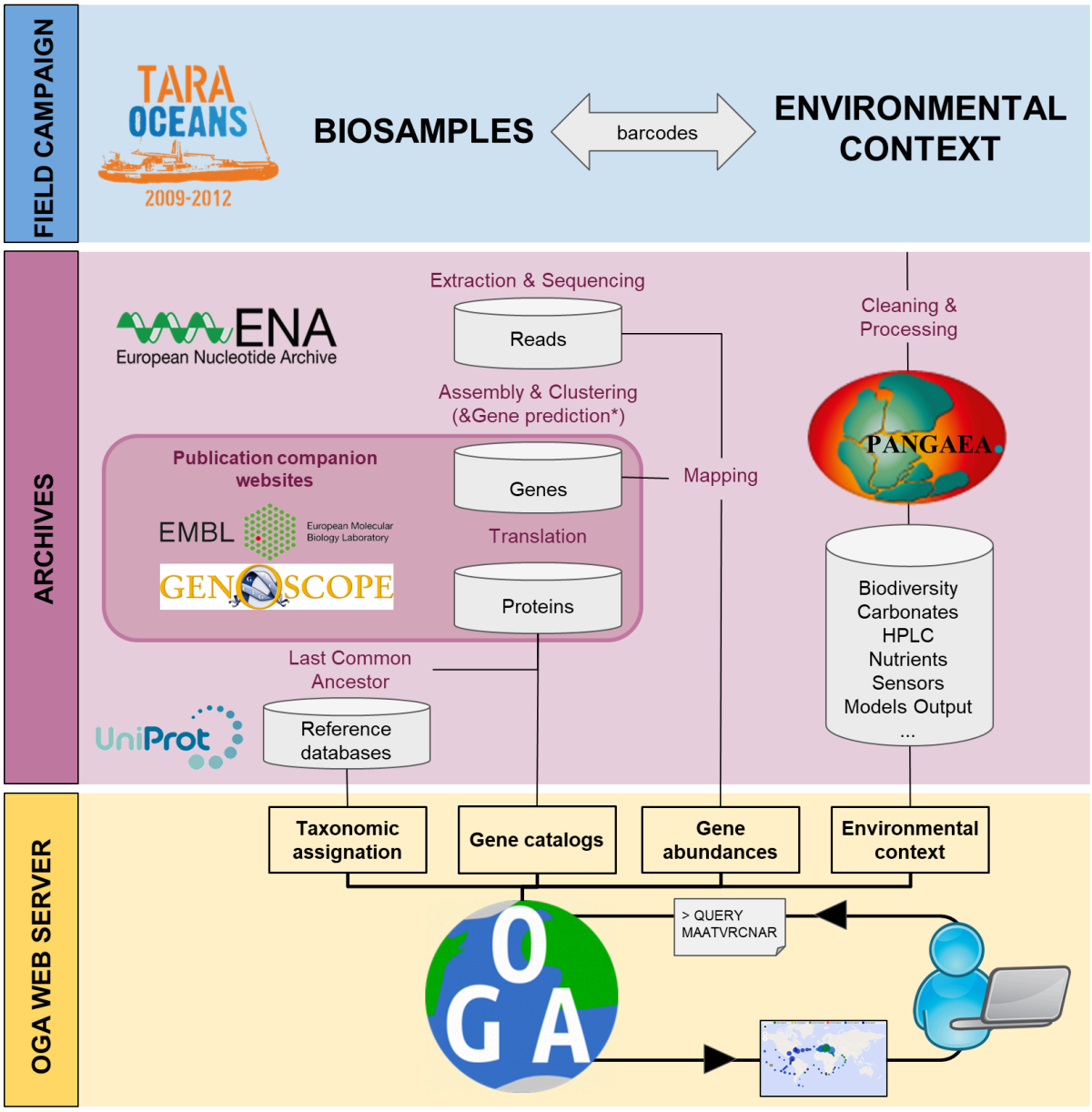

Figure 2: Data sources for the Ocean Gene Atlas workflow.

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 (at the time of writing, Pfam 31.0 from March 2017 with 16712 entries). 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/GOS identifiers)

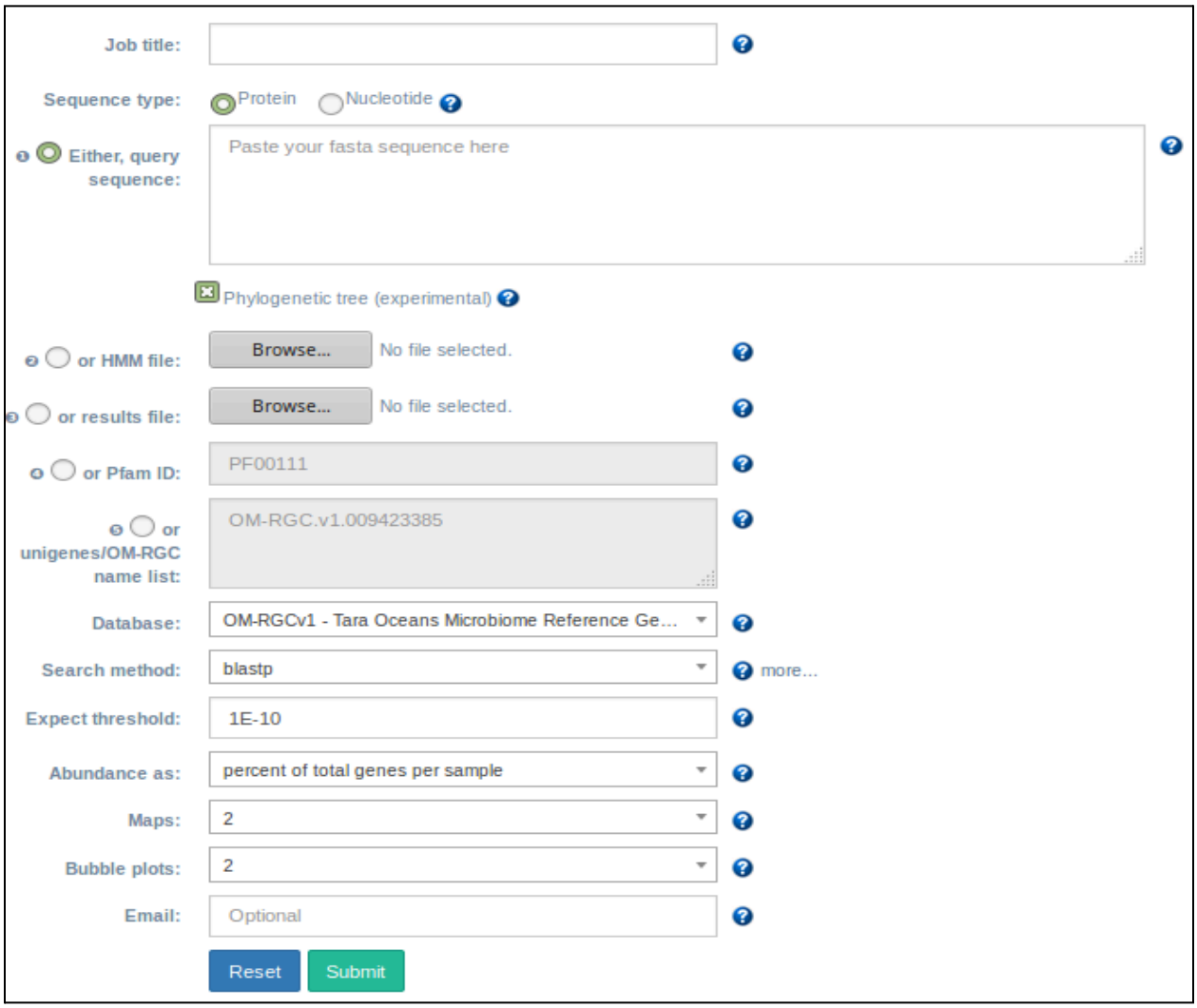

Figure 3: Submission form.

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 et al., 2014) and hmmer (Eddy et al., 2011) .

Each program has a specific computing speed and alignment sensitivity, see Steinegger & Soding (2017). 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).

- **Database**:
Tara Oceans Microbiome Reference Gene Catalog (OM-RGC; Sunagawa et al., 2015)
Tara Oceans Microbiome Reference Gene Catalog with arctic data (OM-RGCv2; Salazar et al.,2019)
Marine Atlas of *Tara* Ocean Unigenes (MATOU; Carradec et al., 2018)
Curated Tara Oceans Single-Cell and Metagenome Assembled Genomes (EUK-SMAGs; Delmont et al.,2021)
More databases will be added as they become available.

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

- **Abundance as**: select the per sample homologs abundance normalization method. The default normalization ("*percent of total genes per sample*") simply divides for each sample the sum of the abundances of the homologs by the sum of the total gene abundance for the sample (i.e. the abundance results represent the homologs fraction in the total gene set in each sample). For the OM-RGC abundance, an alternative normalization method ("*average copies per cell*") divides for each sample the sum of the abundances of the homologs by the observed median abundance of ten prokaryotic single marker genes which were previously benchmarked for their suitability for metagenomics data analysis (Sunagawa et al., 2013). Hence, using "*average copies per cell*" normalization, the abundance of the homologs are equal to one if the query homologs are equally abundant in the sample as genes known to be present as a single copy in each prokaryote genome. The ten single marker genes used for "*average copies per cell*" normalization are: COG0012, COG0016, COG0018, COG0172, COG0215, COG0495, COG0525, COG0533, COG0541, COG0552.
MATOU abundances are expressed as coverage divided for each gene in every sample, computed in RPKM (Reads Per Kilobase covered per Million of mapped reads).

- **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.

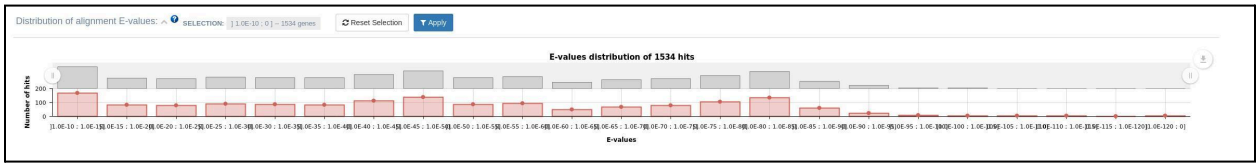

Figure 5: Dynamic E-values bar chart.

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, J.R. et al. 2012), normalised by one of the two methods selected in the submission form (normalisation 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).

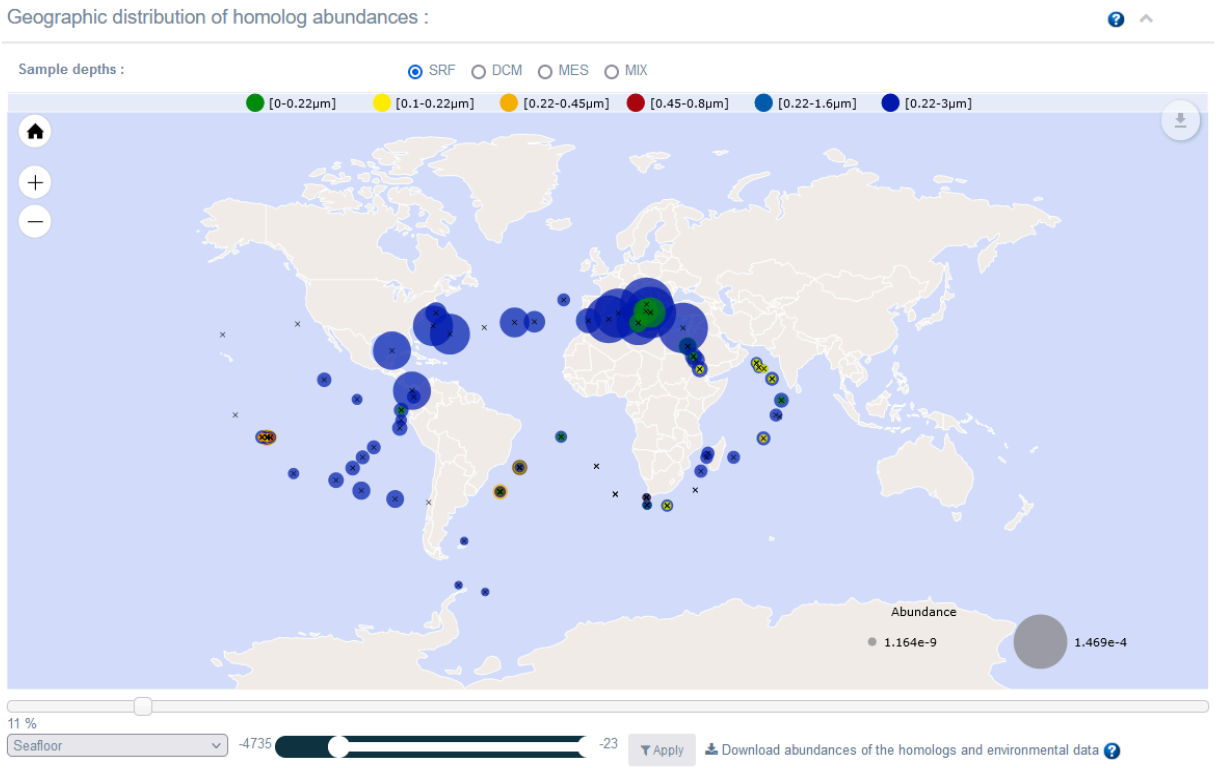

Figure 6: Interactive world maps of the quantitative geographical distributions.

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.

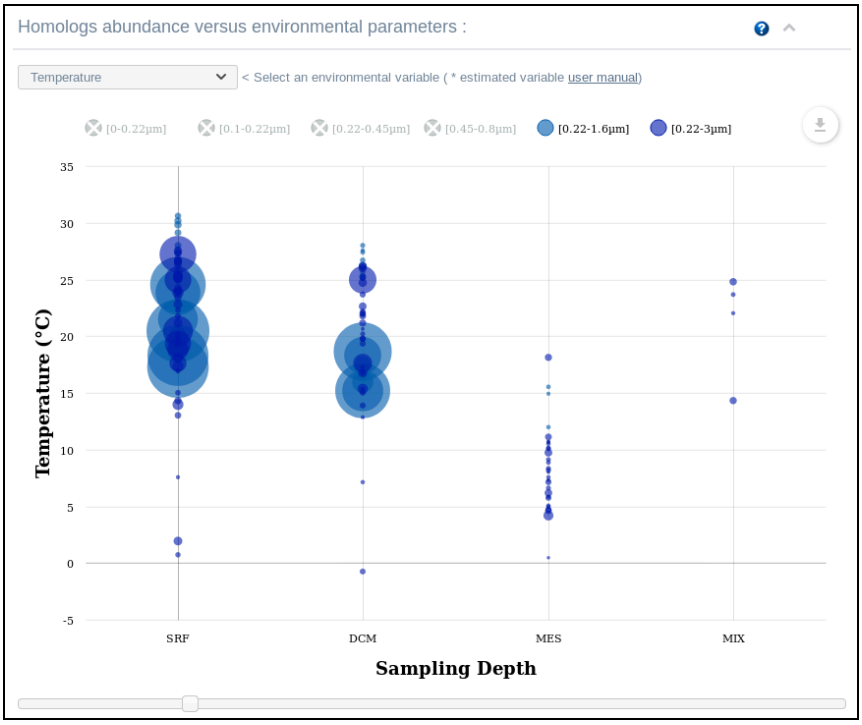

Figure 7: Bubble plots representing the co-variation of gene abundances and an environmental feature (e.g. Mean temperature) for each depth and size fraction combination.

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 et al., 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).

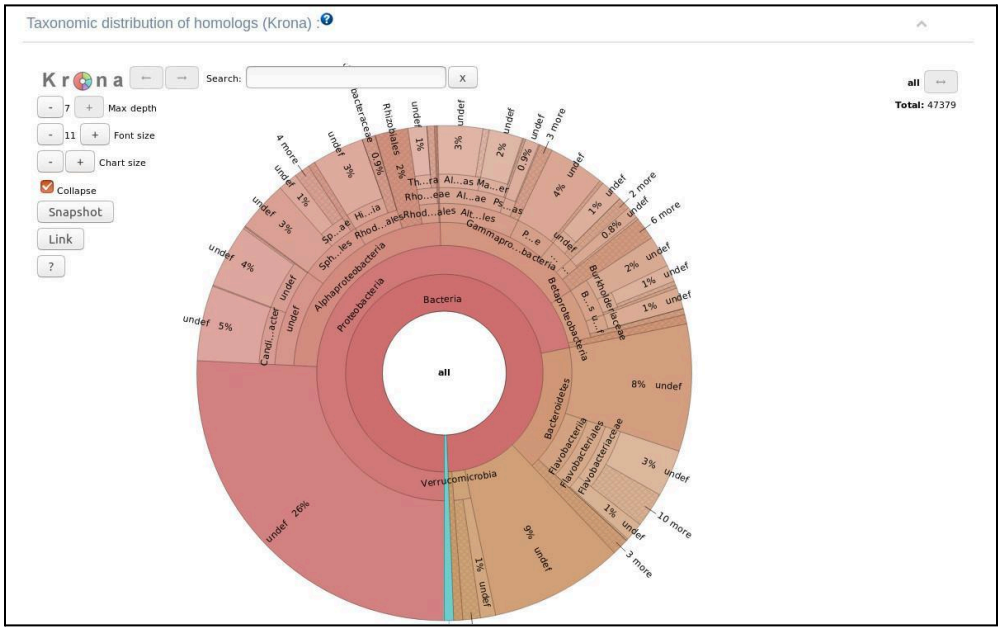

Figure 8: Krona pie-chart representing the taxonomic distribution of homologous sequences in all samples.

The all genes option makes it possible to visualize the taxonomic distribution of genes with one or more abundances less than 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 al, 2001) 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 (Katoh et al, 2002). This alignment is finally cleaned with TrimAl (Capella-Gutiérrez et al, 2009) and MaxAlign (Gouvaia et al, 2007) before submission to FastTree (Price et al,2010) for phylogenetic tree inference. The resulting tree is displayed (Fig.10) thanks to the javascript library plylotree.js (Shank, Weaver, and Kosakovsky Pond 2018). Several interactive display options are available to the user.

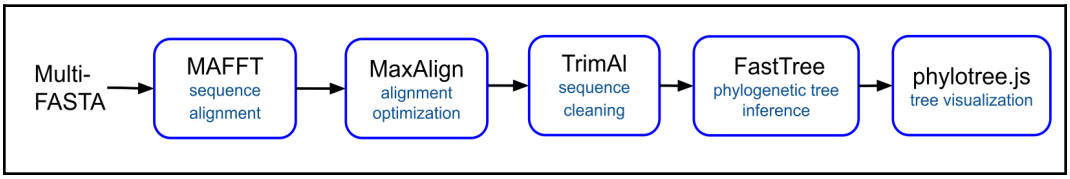

Figure 9: Phylogenetic pipeline.

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 orange, the metagenome homologs appear in green, and the RefSeq reference homologs are labeled in blue.

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 (Click on a leaf then “Reroot in this node”) 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).

Tree legend:

- Sequence query (OGA_QUERY_EAQ46983.1 Metallo-phosphoesterase [Roseobacter sp. MED193])
- OM-RGC_v1 sequences
- Refseq sequences (clustered at 60% identity, see table above)

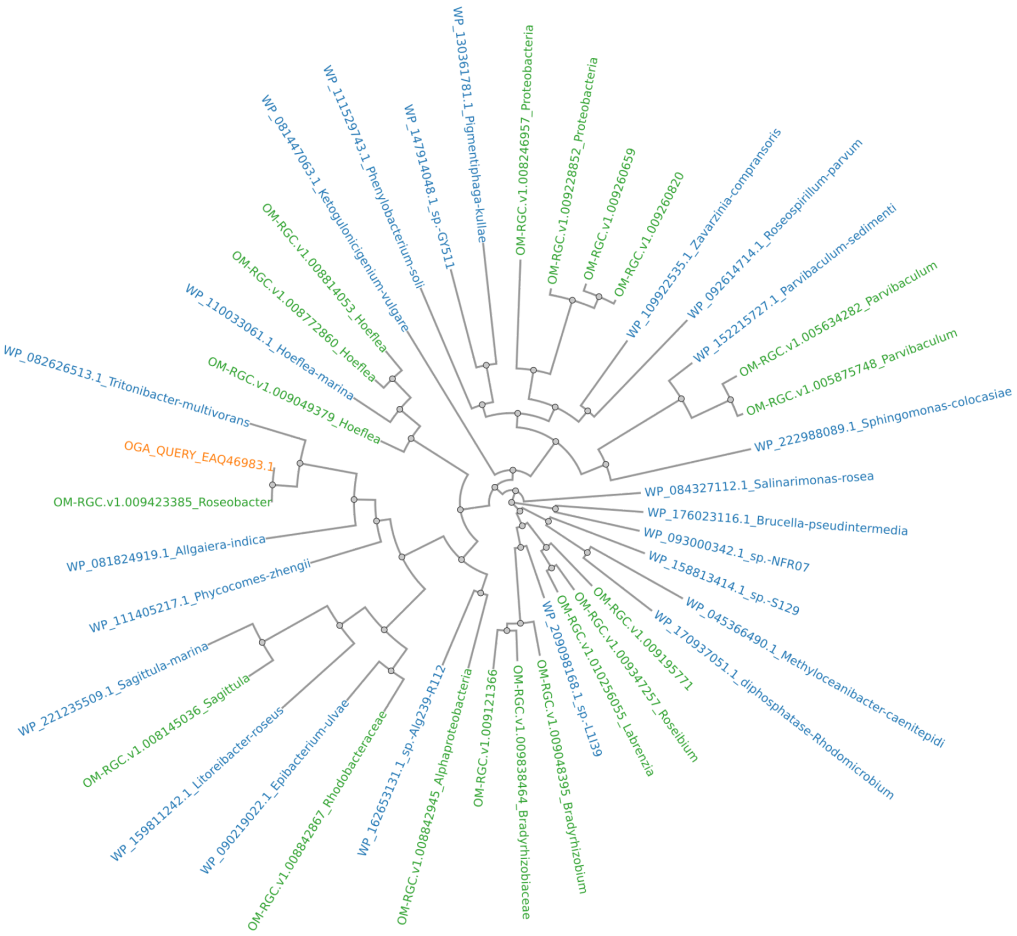

Figure 10: Phylogenetic tree rendering.

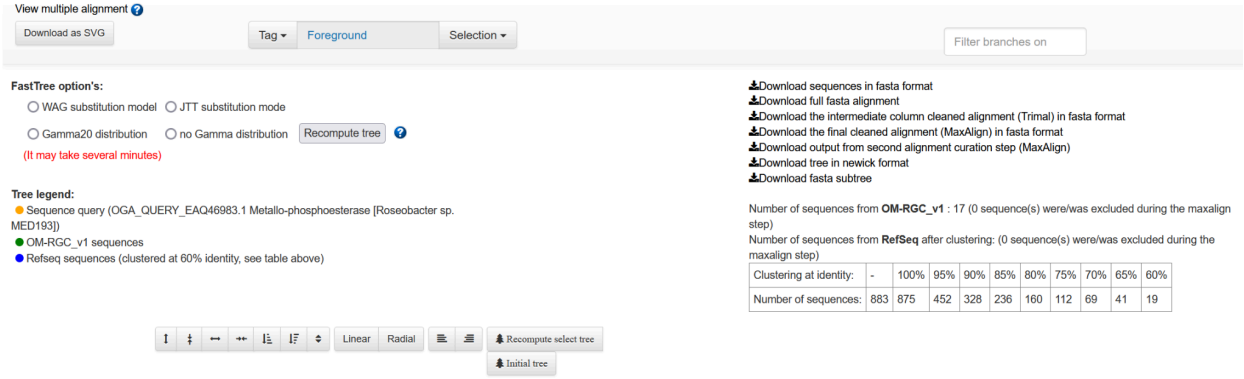

Figure 11:: Phylogenetic analysis options.

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 database (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.

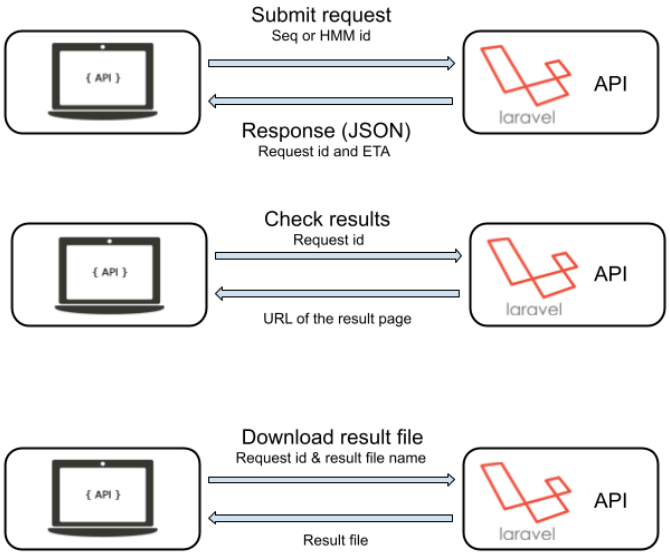

Figure 13: The three types of API requests

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

MATOU catalog: http://www.genoscope.cns.fr/tara/

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

EUK-SMAGs catalog: https://www.genoscope.cns.fr/tara/

VI.3 Literature cited

1. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” *Nucleic Acids Res* 25, 3389–3402.
2. Buchfink, B., Xie, C., and Huson, D.H. (2015). “Fast and sensitive protein alignment using DIAMOND.” *Nat Meth* 12, 59–60.
3. Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M. , Searson, S., Tara Oceans Coordinators, Jaillon, O., Aury, JM., Karsenti, E., Sullivan, M. B. , Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler, C., Wincker, P. (2018). “A global ocean atlas of eukaryotic genes.” *Nat Commun* 9(1):373. doi: 10.1038/s41467-017-02342-1.
4. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Computational Biology*, **7**, e1002195.
5. Kultima,J.R. *et al.* (2012) “MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit.” *PLOS ONE*, **7**, e47656.
6. Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). “Interactive metagenomic visualization in a Web browser.” *BMC Bioinformatics* 12, 385.6.

7. Steinegger M, Söding J. (2017). “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.” *Nat Biotechnol* 35(11):1026-1028. doi: 10.1038/nbt.3988. PMID:29035372
8. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). “Structure and function of the global ocean microbiome.” *Science* 348, 1261359.
9. Li H & Durbin R (2009) “Fast and accurate short read alignment with Burrows-Wheeler transform.” Bioinformatics 25: 1754–1760
10. Katoh K, Misawa K, Kuma K & Miyata T (2002) “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.” Nucleic Acids Res 30: 3059–3066
11. Capella-Gutiérrez S, Silla-Martínez JM & Gabaldón T (2009) “trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.” Bioinformatics 25: 1972–1973
12. Gouveia-Oliveira, Rodrigo, Peter W. Sackett, and Anders G. Pedersen. “MaxAlign: Maximizing Usable Data in an Alignment.” *BMC Bioinformatics* 8, no. 1 (August 28, 2007): 312.

13. Price MN, Dehal PS & Arkin AP (2010) FastTree 2 – “Approximately Maximum-Likelihood Trees for Large Alignments.” PLOS ONE 5: e9490
14. Junier T & Zdobnov EM (2010) “The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell.” Bioinformatics 26: 1669–1670
15. Sebastian,M., Smith,A.F., Gonzalez,J.M., Fredricks,H.F., Van Mooy,B., Koblizek,M., Brandsma,J., Koster,G., Mestre,M., Mostajir,B., et al. (2016)” Lipid remodelling is a widespread strategy in marine heterotrophic bacteria upon phosphorus deficiency.” ISME J., 10, 968–978. doi.org/10.1038/ismej.2015.172