User Guide

Last updated: 2020/12/01

- User Guide
  - Overview
  - I) Ocean Barcode Atlas workflow
  - II) Submission interfaces
    - II.1) Definition of the query
    - II.2) Analysis parameters
    - II.3) Accessory administrative parameters
  - III) Intermediate results
    - III.1) List of matching reference sequences
    - III.2) List of sequence similarity matches
  - IV) Results interface
    - IV.1) Job details
    - IV.2) Percent identity distribution
    - IV.3) Maps
    - IV.4) Bubble plots
    - IV.5) Taxonomic distribution
    - IV.6) Phylogeny
    - IV.7) Graphic representation of diversity
    - IV.8) One map per barcode
    - IV.9) Downloading publication grade figures
  - V) Application Programming Interface (API)
  - VI) References
    - VI.1 Environmental context files
    - VI.2 Barcode and OTU catalogs
    - VI.3 Literature cited

Overview

The Ocean Barcode Atlas (OBA) is a web service to explore the biogeography of marine diversity based on barcodes (or Operational Taxonomic Units: OTU) sequence similarity with environmental metabarcoding datasets (Fig. 1). OBA is currently implemented with the following tree datasets:

- i) *Tara* Oceans with arctic data : 18SV9 v2 (Ibarbalz et al. 2019) and PR^2 (Guillou et al. 2013)
- ii) *Tara* Oceans with arctic data 18S and 16S metagenomic : miTAGs (Salazar et al. 2019) SILVA release 128 (Quast et al. 2013)
- iii) metabarcoding of 16S-V4 of the Malaspina expedition (Salazar et al. 2016) with SILVA release 115

Barcode abundance is expressed by the counts of rDNA reads. We plan to update the website gradually with available public marine (meta) barcode datasets.

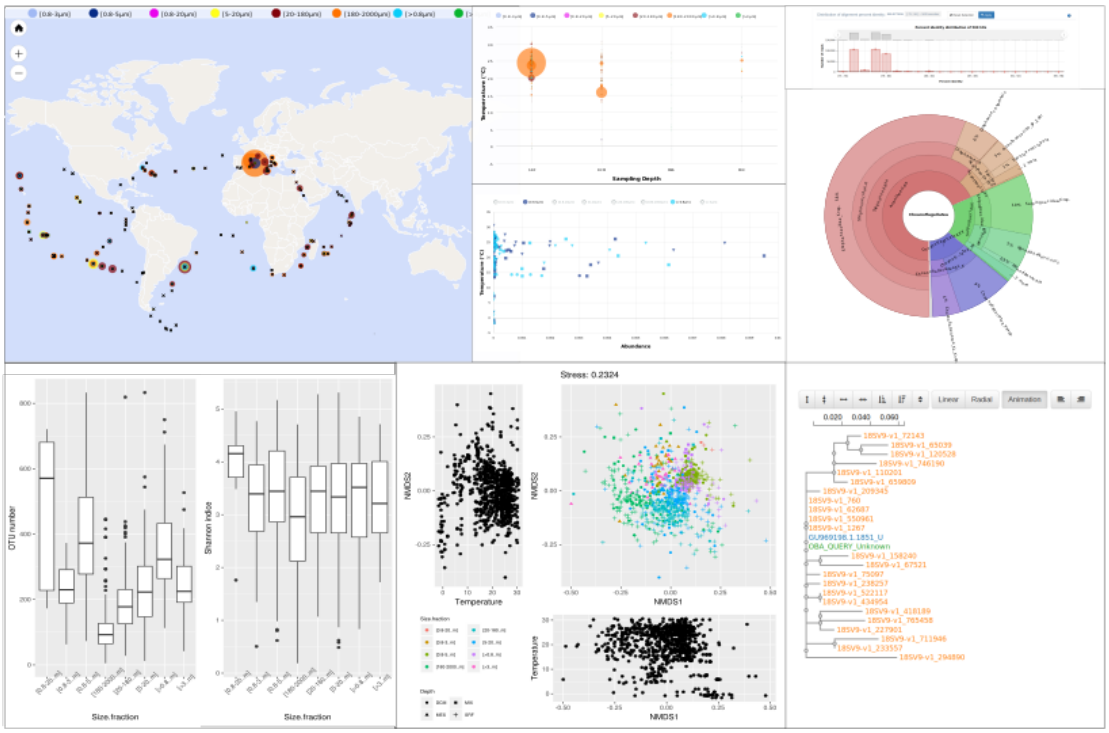


Figure 1: Interactive Ocean Barcode Atlas results.

This web service was designed and developed by: Caroline Vernette, Nicolas Henry, Julien Lecubin, Colomban de Vargas, Pascal Hingamp, Magali Lescot.

URL: http://oba.mio.osupytheas.fr/ocean-atlas/
Contact: oceanbarcodeatlas@mio.osupytheas.fr

The following browsers have been tested and are listed by decreasing order of compatibility with the interactive displays in the OBA 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 OBA.

I) Ocean Barcode Atlas workflow

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

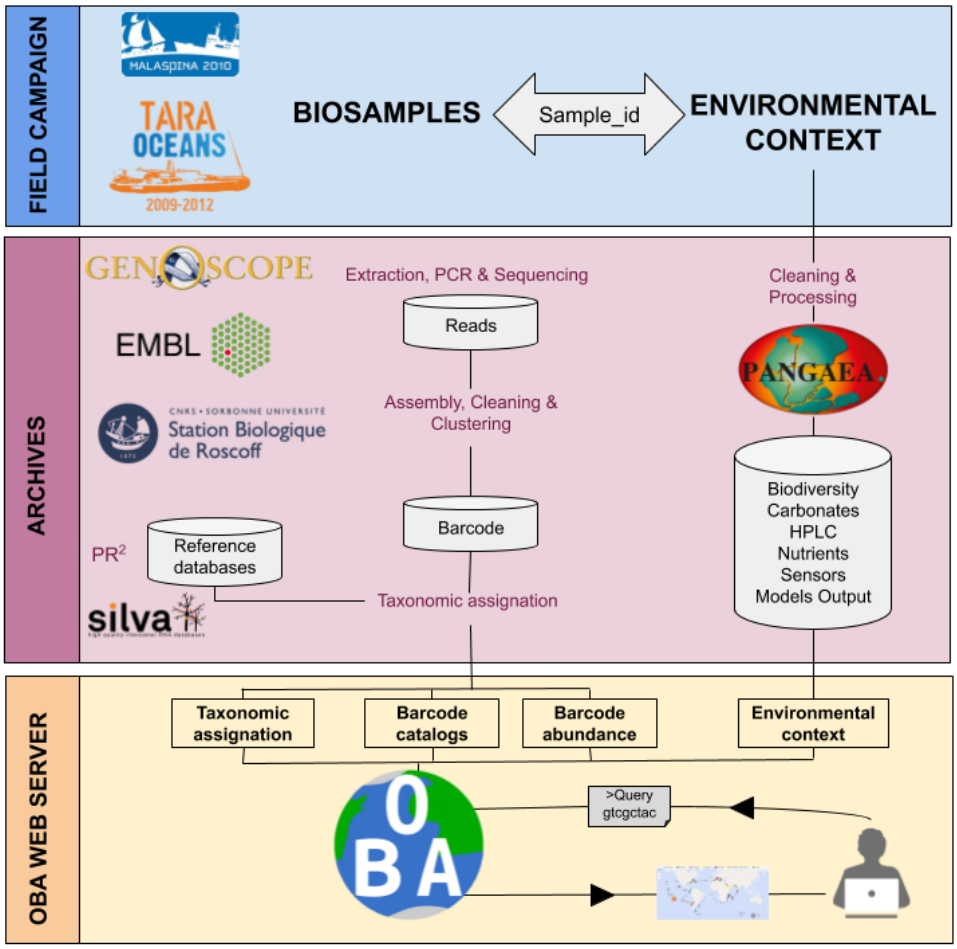


Figure 2: Data sources for the Ocean Barcode Atlas workflow.

II) Submission interfaces

On the OBA home page, two buttons allow you to choose the best interface for your needs (Fig. 3):

- "Community ecological analysis" allows one to search for a taxon by name in a metabarcoding database in order to obtain ecological analyses (alpha and beta diversity) using graphical representation (NMDS, ACP..).
- "Sequence based query" allows one to interrogate a metabarcoding database by sequence similarity in order to obtain the abundance, location and diversity of targeted sequences (barcodes) in an environmental context.

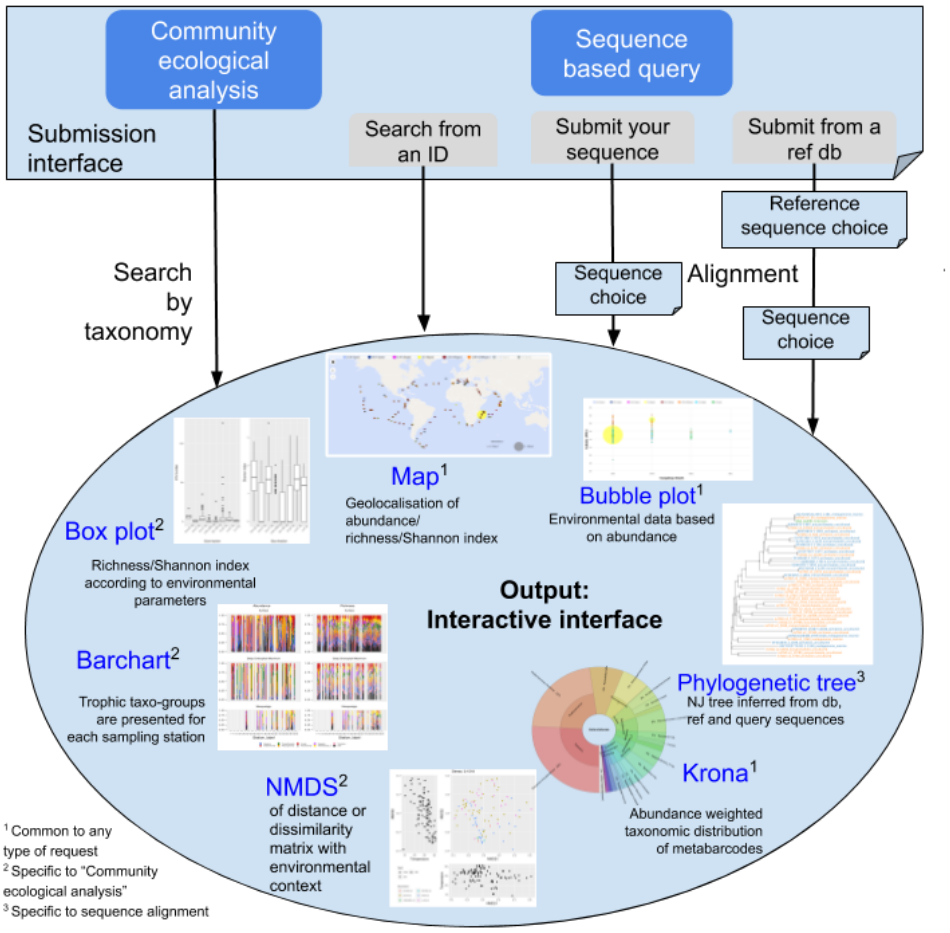


Figure 3: Submission interface

II.1) Definition of the query

II.1.1) Community ecological analysis

Search by taxon name. An interactive tree allows one to make a preselection. The auto-completion feature, when the cursor is in the “Taxonomy” text field, offers access to more complete taxonomy classifications. The beta diversity calculation option provides an NMDS (Non-metric multidimensional scaling) plot from a dissimilarity matrix if the "Bray & Curtis coefficient" is selected or from a distance matrix if the "Jaccard index" option is selected.

II.1.2) Sequence based query

According to which of the three available tabs is selected, the input query may be one of the following (Fig. 3):

1/ **Submit your sequence**. The FASTA-formatted sequence with a header line should be pasted in the text field. If you have a sequence that matches the exact barcode sub-region, click on the first option (eg “18SV9 region”). If not, select the second one (eg “18S complete”) which will extract the corresponding sub-region with cutadapt (version 2.1 (Martin 2011)).

2/ **Submit a ref db sequence**. It’s possible with the taxonomic search tools (tree and text

field) to use as a query a barcode from a reference database (PR² for Tara oceans OTU 18S V9 (Guillou et al. 2013) and Silva release 115 (Quast et al. 2013) for Malaspina OTU 16S V4). If several sequences match your query in the reference database, an intermediate results page will allow you to select which query sequences you wish to proceed with.

3/ **Search from an OTU ID**. Search from an identifier (e.g. OTU identifiers) or a list of IDs. You can upload a file containing your identifiers. The “one map per barcode” option is accessible if less than 5 barcodes are selected. It allows each barcode to be visualized on a separate map and a specific bubble plot (see IV.8).

With “Submit your sequence” and “Submit a ref db sequence” options an alignment is made between the query sequence and the selected database. It is also possible to select the phylogenetic tree option (see IV.6 Phylogeny).

II.2) Analysis parameters

Parameters that define the search method and output configuration :
- Database: Currently either *Tara* Oceans 18S V9, *Tara* Oceans 18S V9 version 2 with arctic data, *Tara* Oceans miTAGs and Malaspina 16S V4 rDNA OTU are available. More databases will be added as they become available.

- Maps: The number of maps can be defined to visualize the geographical distribution of homologous OTUs (each map can display one OTU homolog abundance in distinct size fractions and depths).

- Bubble plots: The number of plots can also be defined to visualize co-variation of barcode homologous abundances with different environmental parameters (each bubble plot can display co-variation in distinct size fractions and for distinct environmental features).

Specific to the “Community ecological analysis” interface:
- Taxonomy: Search from a taxon name. An interactive tree allows one to make a preselection. The auto-completion, when the cursor is in the text field, offers a more complete taxonomy.

- Calculate Beta diversity from: Provides an NMDS plot from a dissimilarity matrix if the "Bray & Curtis coefficient" is selected or from a distance matrix if the "Jaccard index" option is selected.

Specific to the “Sequence based query”:
- Extract sequence: If you have a sequence that matches the barcode sub-region, click on the first option. If not, select the second one which will extract the corresponding region with cutadapt.

- Search method: Choose an OTU sequence similarity search method from: either VSEARCH (Rognes et al. 2016) or BLAST (Altschul et al. 1997). VSEARCH assigns scores for the calculation of the percentage of identity and offerts a global alignment (recommended method).

- % identity search: Set the minimum percentage of identity threshold above which similar barcodes in the database are pre-selected in the intermediate page.

- Phylogenetic tree: The sequences of the selected database and of the reference database having a percentage of identity greater or equal to the threshold defined above with the query sequence are aligned and a phylogenetic tree is inferred.

-ID/ID file: provide identifiers to be searched for in the selected metabarcode databases.

-One map per barcode: If the number of OTUs is less than four, it is possible to select this option which produces one map and one bubble plot for each query OTU.

II.3) Accessory administrative parameters

Two parameters are optional:
- Job title: a free text will be used to name downloadable files.
- E-mail address: if provided, the output of the similarity search will be sent to the user via email. This results file can be used to generate the results page without recomputing the similarity search (i.e. saves server time). A hyperlink to the results page will be provided at the time of data submission and also included into the results file sent by email. The results will remain available online for 15 days.

III) Intermediate results

These intermediate pages are displayed only when the “Submit your sequence” and “Submit a ref db sequence” tabs were used (Fig.3).

III.1) List of matching reference sequences

This intermediate page allows one to select a sequence from the PR^2 or Silva reference databases. The taxonomies and nucleotide sequences of matching reference sequences are displayed and radio buttons allow the user to make his selection. The selected reference sequence will then be aligned with the selected metabarcode database, leading to a second intermediate page (below).

III.2) List of sequence similarity matches

This intermediate page lists the metabarcodes matching the user defined identity threshold (see II.2 Analysis parameters). The percent identity and number of matching reads are reported. Only the best matching 500 first barcodes are listed. Tick boxes allows the user to select which metabarcodes to submit to the final interactive plots (see below *III.2.2 Checkboxes*). A bar chart displays the barcodes distribution according to the number of reads and the percentage identity (see below *III.2.1 Percentage identity distribution*). The “one map per barcode” representation option is only available if less than 5 barcodes are selected. It allows each barcode to be visualized individually on separate maps and bubble plots (see IV.8).

III.2.1) Percentage identity distribution

The bar chart (Fig. 4) displays the distribution of the similarity hits and allows users to adjust inclusion thresholds. One can define the range of accepted identity percentage by selecting the range directly with the mouse on the histogram, and clicking on the "Apply" button to select or deselect the corresponding checkboxes. It is possible to know the number of hits for each barcode thanks to a tooltip that is displayed when you roll-over the graph bars with the mouse cursor.

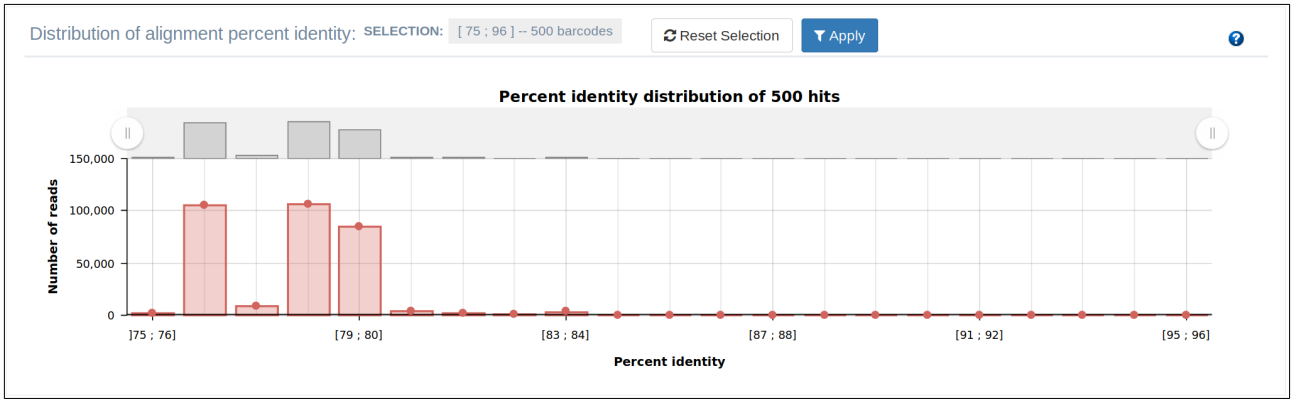


Figure 4: Dynamic percentage identity bar chart.

III.2.2) Checkboxes

Each barcode (Fig. 5) can be selected via a checkbox. The percentage of identity, the number of associated reads as well as the matching reference sequence is shown for each barcode. The reference sequences appear in blue with their associated taxonomy. It is possible to select or deselect all checkboxes with the “check all” and “uncheck” button in the top-right corner of the web page.

The barcodes selected in this intermediate page will then be carried over to be represented on the interactive results interface.



Figure 5: Checkoxes.

IV) Results interface

The results interface displays the computed results *via* maps, bubble plots, krona pie-charts and optional phylogenetic trees. The results are organized by sample (except for the overall krona pie-chart and trees), 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 15 days after job submission (the URL can be shared with collaborators).

IV.1) Job details

The top panel (Fig. 6) provides information about the submitted job (e.g. the shareable URL of this results page, parameters such as percentage of identity threshold etc.) and a summary of the similarity search results (number of “barcodes” hits, number of abundance measures).
If the “Submit a ref db sequence” tab was used to define the query, the query sequence is displayed. When the “Search from a taxonomy” tab was used, the selected taxonomy is displayed.
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 (barcode identifiers and % identity),
- the corresponding FASTA formatted sequences of the rDNA hits,
- the biosample abundance matrix and contextual environmental features.



Figure 6: Summary of results.

IV.2) Percent identity distribution

For queries defined by a sequence similarity search, a bar chart (Fig. 7) presents the percent identity distribution of the hits and allows the user to adjust the homolog inclusion threshold. One can change the range of percent identity by selecting with the cursor the chosen range directly on the histogram, and clicking on the “Apply” button to update all the maps, bubble plots and krona pie-charts.

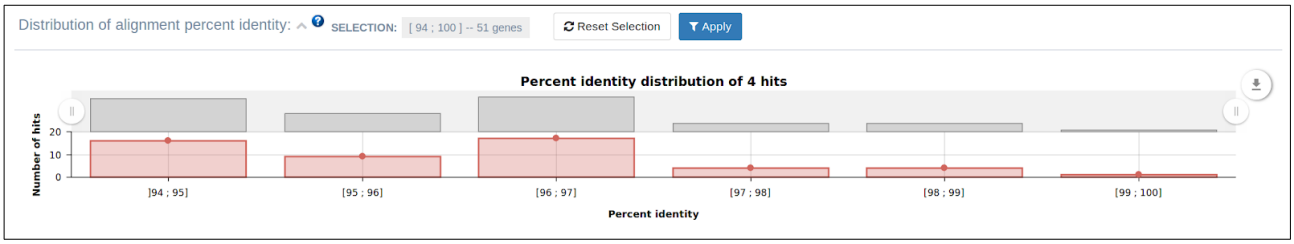


Figure 7: Dynamic percent identity bar chart.

IV.3) Maps

On the interactive geographical maps (Fig. 8), each circle represents the abundance of selected barcode homologous in one environmental sample. The abundance is the read number normalized by the total number of reads in the sample. 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 size of the circles may be scaled using the interactive slider below each map. A scale - entitled "abundance" - is displayed on the bottom of the map in order to compare several independent results. The two grey circles of the map scales represent the maximum and minimum abundance related to the barcode numerical values.
A click on a given circle will open a krona taxonomic distribution pie-chart specific to the selected sample. The different acronyms of sampling depth stand for: DCM: Deep Chlorophyll Maximum layer; SRF: upper layer zone; MES: MESopelagic zone (200-1000 m); MIX: marine epipelagic MIXed layer; BAT: BAThypelagic zone (1000-4000 m); ABY: ABYssopelagic zone (4000-6000 m). Using the down arrow in the top-right corner of each map, users may edit, print and/or download the maps in several formats.

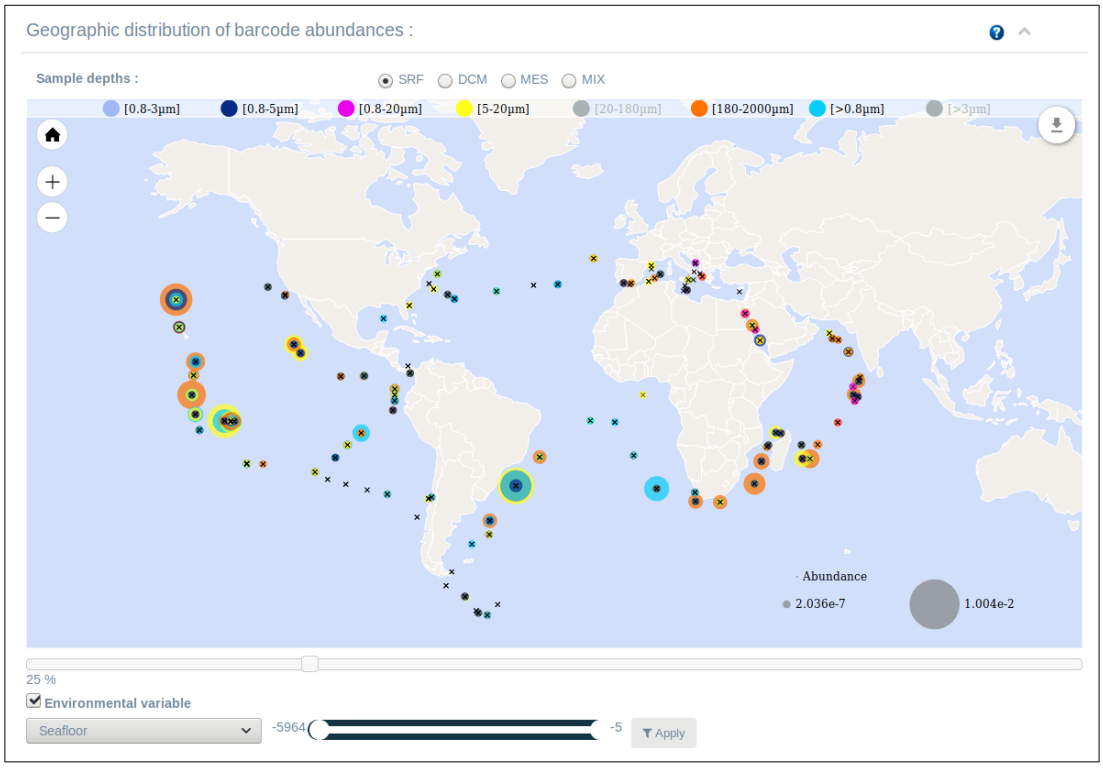


Figure 8: Interactive world map of the barcode abundance distributions.

With “sequence based queries”, users can select the "Environmental Variable" option to choose an environmental variable from a list which presents a slider spanning minimum to maximum values. When the button "Apply" is pressed, only the samples corresponding to the selected range are displayed on the map.

IV.3.1) Map with “community ecological analysis”

Three distinct metrics can be used to define the diameter of the sample circles on the maps: **Abundance**, **Shannon index** and **Richness** (number of OTUs). Shannon diversity index is calculated under R (version 3.5.3) with the vegan package (version 2.5-5 (Oksanen, Blanchet, and Friendly 2019)). The richness is calculated from rarefied abundances with the minimum number of sample reads.

IV.4) Bubble plots

The bubble plots associate environmental context to metabarcode abundance for each sampling depth (Fig. 9). A drop-down menu allows the user to change the displayed

environmental parameter. A logarithmic scale is available via the corresponding checkbox (log10).
The different acronyms of sampling depths 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 “V.1 Environmental context files”.
Similarly to the geographical maps above, the size of the sample circles are proportional to the abundance of the query homologs. The circles are colored according to the selected fractions (color codes are provided just above each bubble plot). The y-axis represents the environmental parameter value such as 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 or Temperature. Values estimated from oceanographic models are indicated by a star.

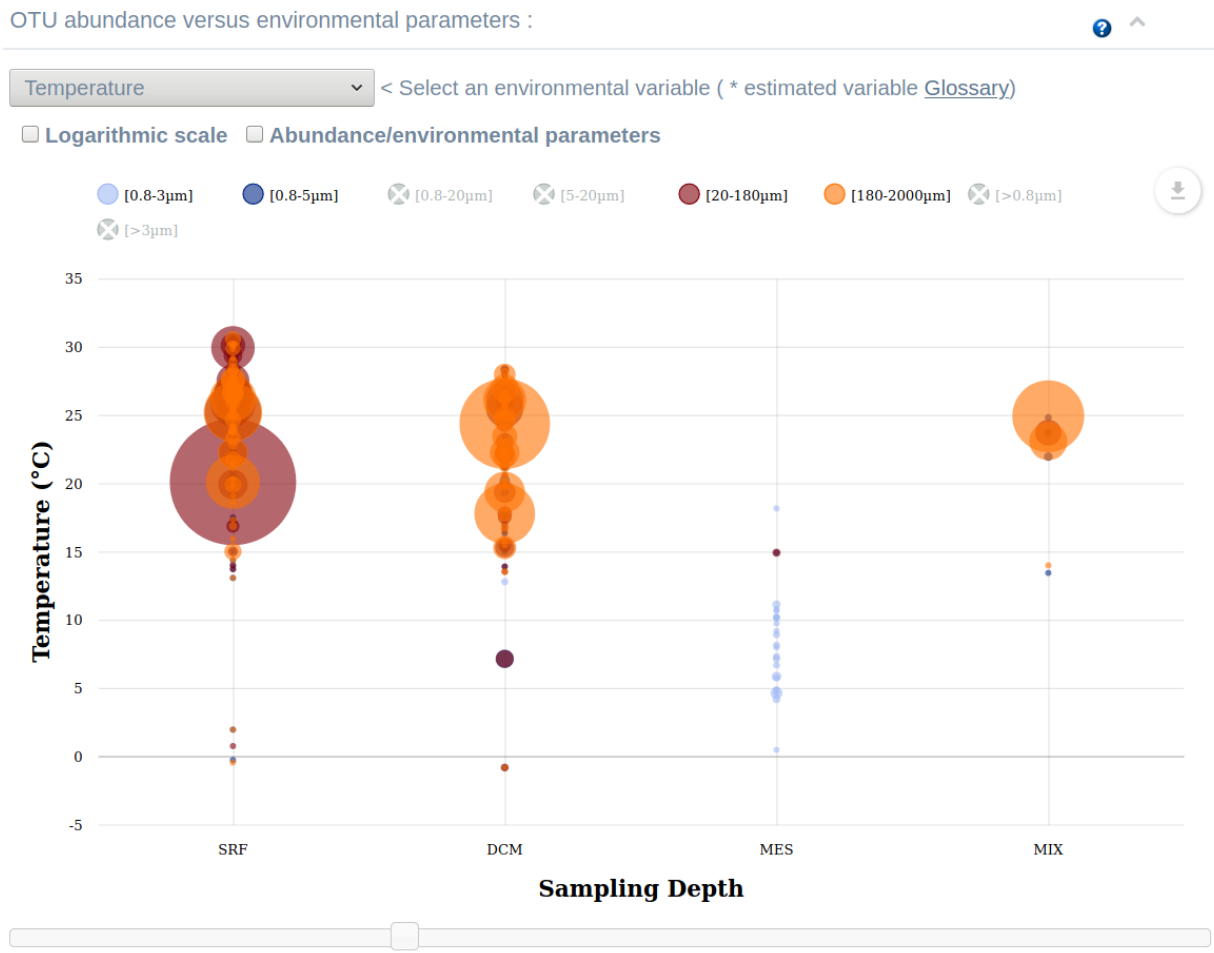


Figure 9: Bubble plots representing the co-variation of barcode abundances and an environmental feature (e.g. Fraction Lower) for each depth and size fraction combination.

By selecting the “abundance/environmental parameters” option, it is possible to visualize a dot plot of the abundance versus environmental parameter values (Fig. 10). Each symbol shape is associated with a depth.



Figure 10: Abundance/environmental parameters option.

IV.5) Taxonomic distribution

At the bottom of the results page, a Krona pie-chart (Ondov, Bergman, and Phillippy 2011) presents an overview of the abundance weighted taxonomic distribution of metabarcodes summed over all samples (Fig. 11). The diagram allows taxonomic data to be explored with a zoomable multi-layered pie-chart. Metabarcode taxonomic distribution for each distinct biosample can be explored by clicking on the corresponding circles in the geographic maps (see above “IV.3 Maps”).

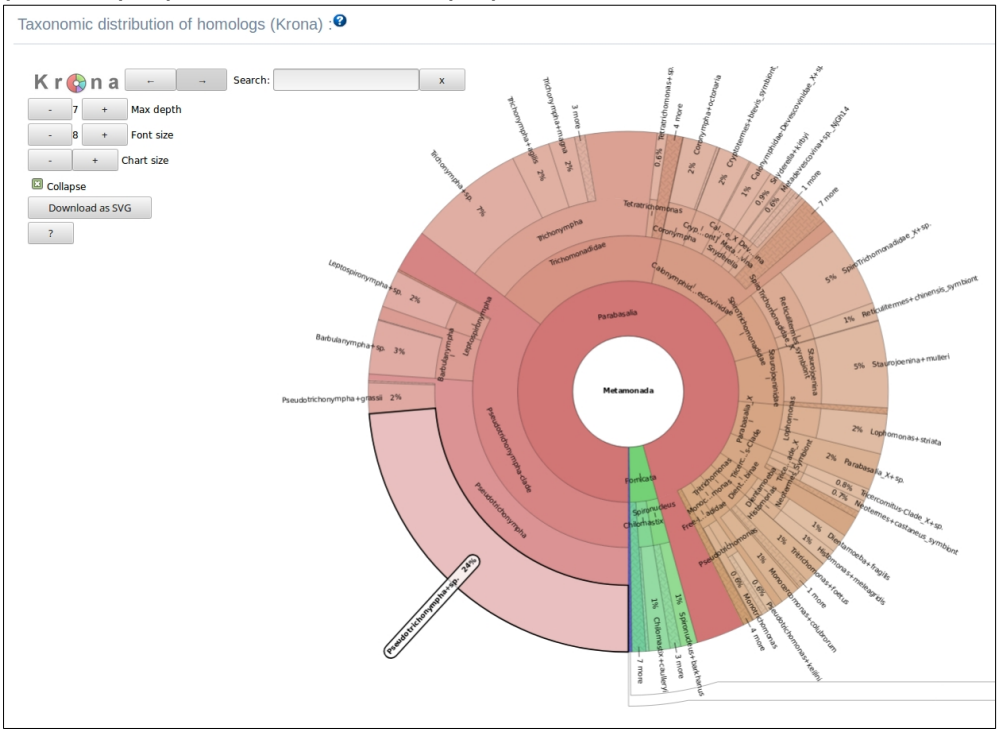


Figure 11: Krona pie-chart representing taxonomic distribution of homologous sequences in all samples.

For more information on Krona, see : https://github.com/marbl/Krona/wiki/Browsing%20Krona%20charts.

IV.6) Phylogeny

When the option was ticked in the initial submission interface (unchecked by default), a phylogenetic tree is inferred according to the method shown in Figure 12. The resulting tree is displayed (Fig.13) thanks to the javascript library plylotree.js (Shank, Weaver, and Kosakovsky Pond 2018). Several interactive display options are available to the user.

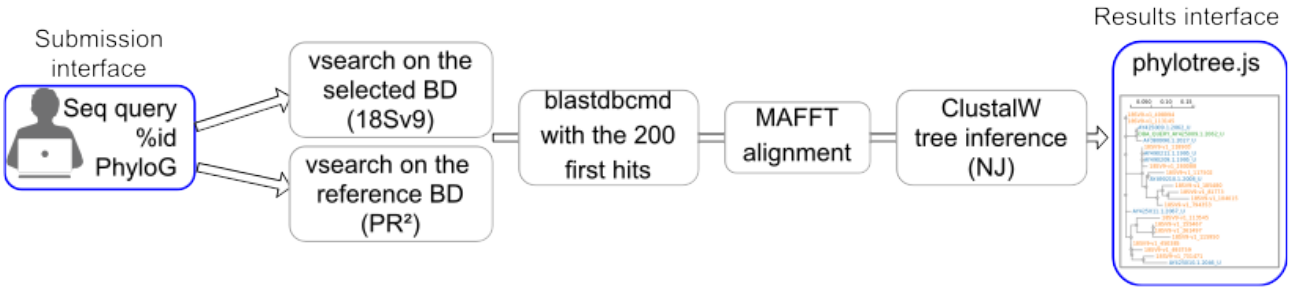


Figure 12: MAFFT version 7.407 (Katoh et al. 2002); ClustalW version 2.1 (Larkin et al. 2007)

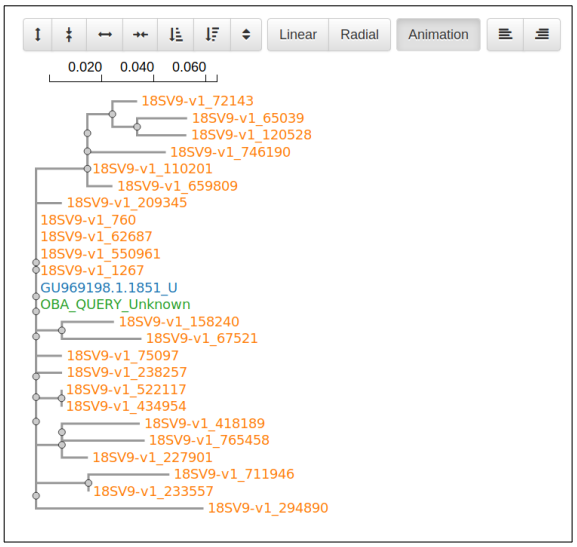


Figure 13: Phylogenetic tree rendering.

IV.7) Graphic representation of diversity

For jobs submitted via the “community ecological analysis” option, three additional panels illustrate plankton diversity.

IV.7.1) Abundance & richness barcharts

Both richness and abundance barcharts of rDNA metabarcodes assigned to various trophic taxo-groups are presented for each sampling station across:

- plankton organismal size fractions
- sampling depths

The selected taxonomic rank is located just after the taxonomic rank of the query. The 9

most abundant taxonomies are displayed, whilst the others are grouped together and classified as “others” (Fig 14).

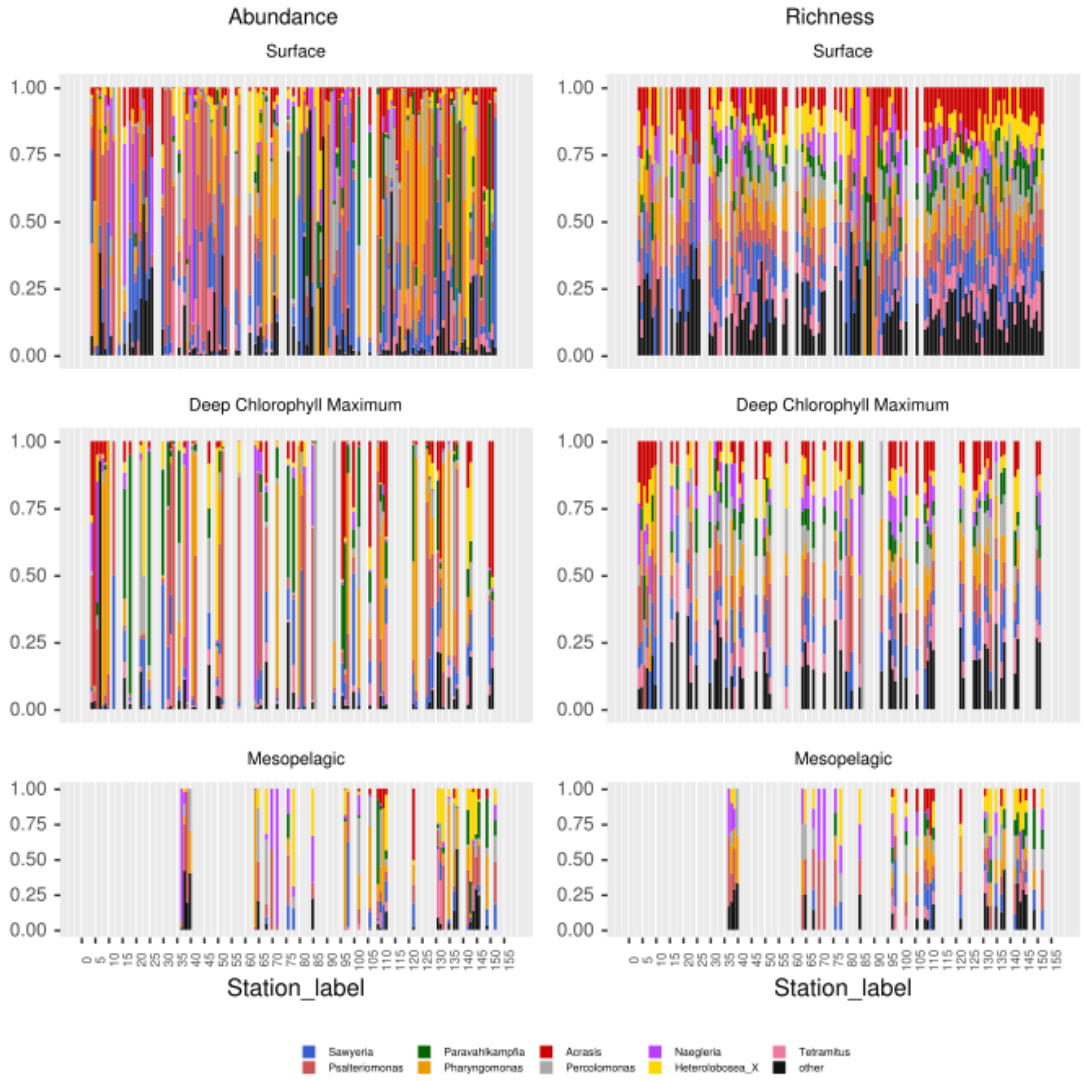


Figure 14: Barchart with taxo-group by station.

IV.7.2) Beta diversity

The beta diversity panel represents beta diversity calculated using the R (version 3.5.3) vegan package (version 2.5-5 (Oksanen, Blanchet, and Friendly 2019)). Figure 15 below describes the analysis pipeline.

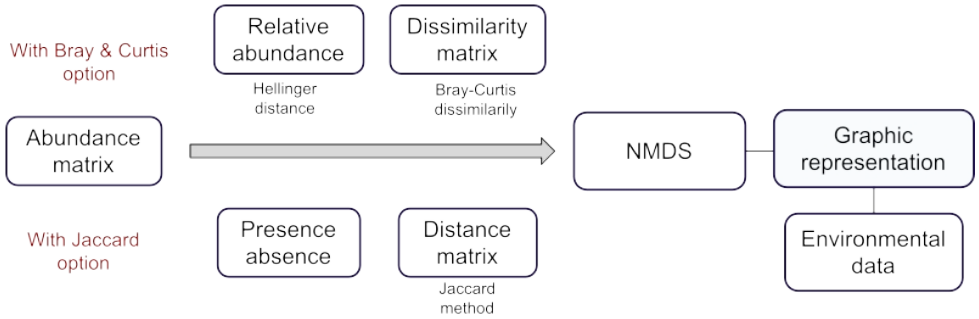


Figure 15: Pipeline for the graphical representation of beta diversity.

Three drop-down lists are proposed (Fig 16) with a choice of environmental variables (a glossary is accessible with their descriptions). The first drop-down list allows users to change the variable associated with the colors, the second changes the variable associated with the symbols of the NMDS points. The last drop-down list changes which variable is plotted to the left and below the NMDS representation. Remember to click on the “Apply” button to update the plots after changing the variables.
The stress value shown on the graph indicates how similar the distances between samples of the NMDS are to the distances analyzed. The closer the stress value is to 0, the more similar the distances.

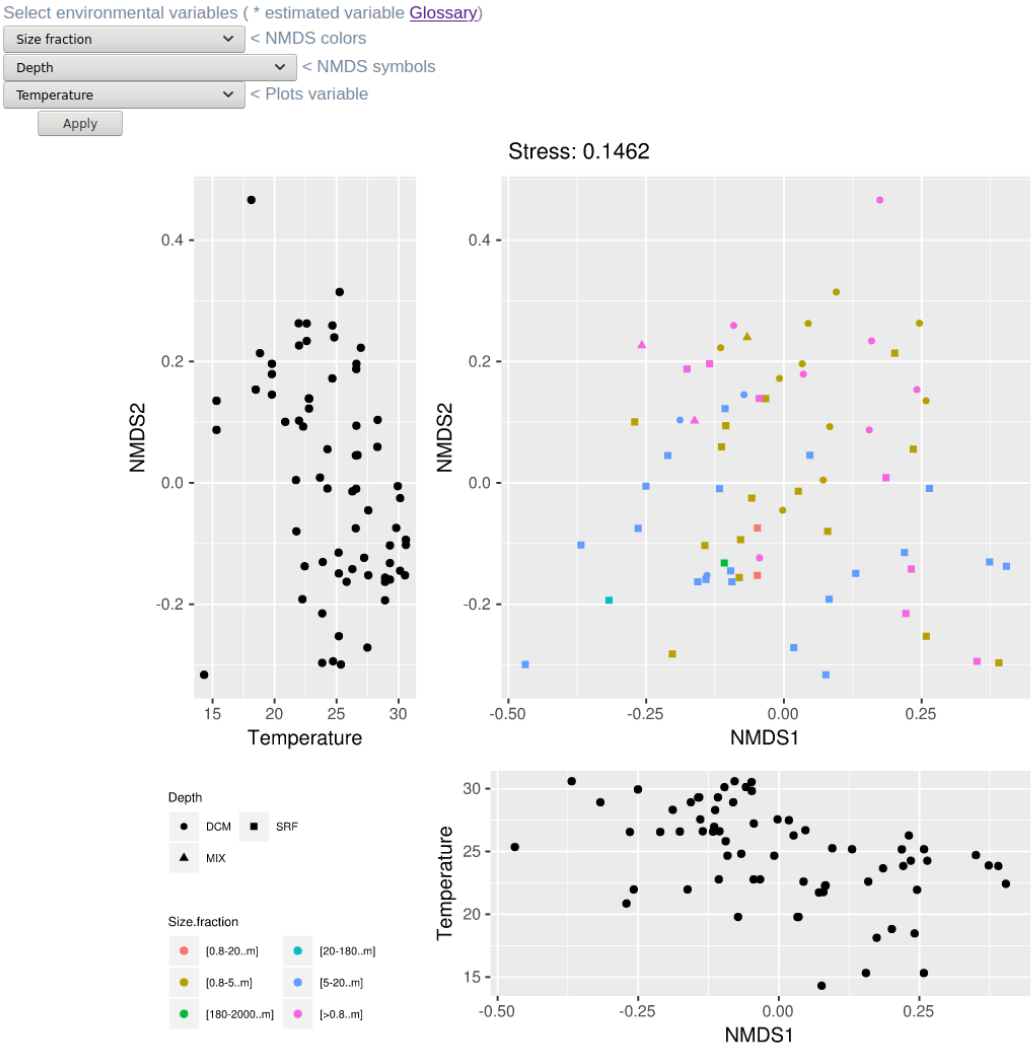


Figure 16: Graphical representation of beta diversity

The “With environmental vectors” button plots an additional NMDS graphic with projections of variable vectors. The script uses the envfit function (vegan package) to fit environmental vectors or factors onto the ordination. The projections of points onto vectors have maximum correlation with corresponding environmental variables and the factors show the averages of factor levels. The variables with pourcentage of missing data better than 15 are removed.

IV.7.3) Alpha diversity

The “Alpha diversity” panel contains two graphs (Fig.17). The first graph represents richness according to environmental parameters. The second graph represents the Shannon index as a function of environmental parameters selected by the user in the drop-down list above the graphs.

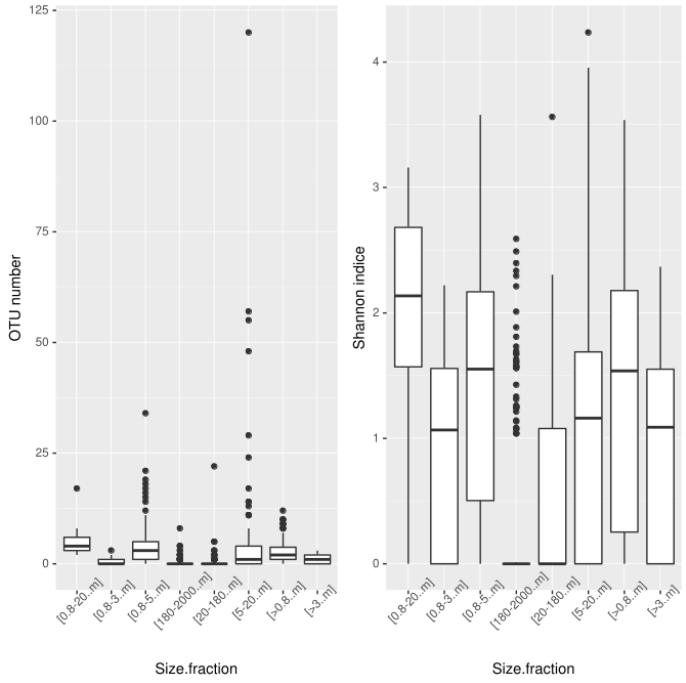


Figure 17: Alpha diversity graphics.

IV.8) One map per barcode

A specific interface is presented when the “one map per barcode” option was ticked in the submission interface. This interface displays a separate map and bubble plot for each barcode (up to a maximum number of 5 barcodes). A panel indicates the identifier of the barcode, the number of abundance measures and the associated taxonomy.

IV.9) Downloading publication grade figures

On the maps and bubble plot display panels, a download arrow in the top right corner (Fig. 18) allows users to download the figure in several formats such as Scalable Vector Graphics (SVG) format suitable for high resolution post-treatment and publication. For the Krona chart, barchart, beta and alpha panel the “Download as SVG” button allows users to save the figure as an “.svg” file.

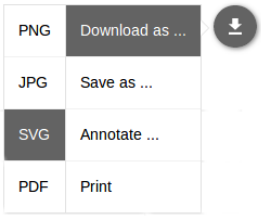


Figure 18: Download figure as Scalable Vector Graphics.

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, seq or taxon. 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.

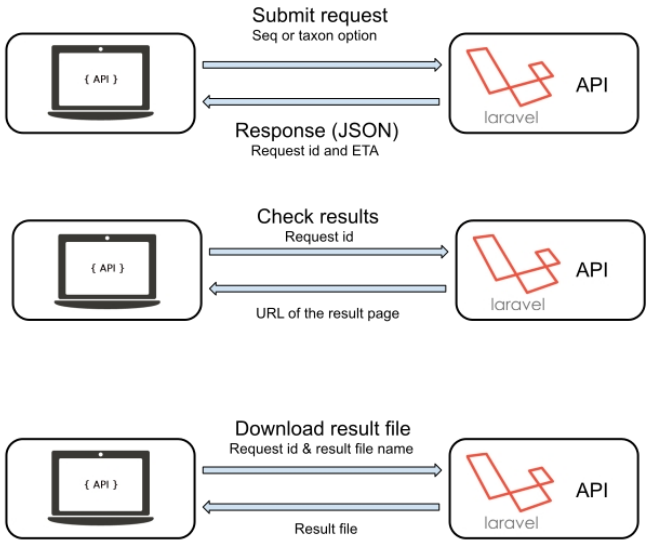


Figure 13: The three types of API requests

A script (oba_APIpipeline.sh) is available at the following address:

http://oba.mio.osupytheas.fr/ocean-atlas/build/script/oba_APIpipeline.sh

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 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 OBA upon request at oceanbarcodeatlas@mio.osupytheas.fr

VI.2 Barcode and OTU catalogs

Tara Oceans 18S-V9 catalog: http://taraoceans.sb-roscoff.fr/EukDiv/
Tara Oceans 18S-V9 v2 catalog: https://www.omicsdi.org/PRJEB9737
Malaspina 16S-V4 catalog: https://github.com/GuillemSalazar/MolEcol_2015
PR² database: https://github.com/pr2database/pr2database
Silva release 115: SSURef_NR99_115_tax_silva.fasta.tgz
Silva release 128: SSURef_NR99_128_tax_silva.fasta.tgz

VI.3 Literature cited

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402.

Guillou, Laure, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, et al. 2013. “The Protist Ribosomal Reference Database (PR2): A Catalog of Unicellular Eukaryote Small Sub-Unit RRNA Sequences with Curated Taxonomy.” *Nucleic Acids Research* 41 (Database issue): D597–604. https://doi.org/10.1093/nar/gks1160.

Ibarbalz, Federico M., Nicolas Henry, Manoela C. Brandão, Séverine Martini, Greta Busseni, Hannah Byrne, Luis Pedro Coelho, et al. 2019. “Global Trends in Marine Plankton Diversity across Kingdoms of Life.” *Cell* 179 (5): 1084-1097.e21. https://doi.org/10.1016/j.cell.2019.10.008.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30 (14): 3059–66.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, et al. 2007. “Clustal W and Clustal X Version 2.0.” *Bioinformatics* 23 (21): 2947–48. https://doi.org/10.1093/bioinformatics/btm404.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Oksanen, Jari, F. Guillaume Blanchet, and Michael Friendly. 2019. “Vegan.Pdf.” Package “Vegan.” May 8, 2019. https://cran.ism.ac.jp/web/packages/vegan/vegan.pdf.

Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. “Interactive Metagenomic Visualization in a Web Browser.” *BMC Bioinformatics* 12 (1): 385. https://doi.org/10.1186/1471-2105-12-385.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools.” *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. “VSEARCH: A Versatile Open Source Tool for Metagenomics.” *PeerJ* 4: e2584. https://doi.org/10.7717/peerj.2584.

Salazar, Guillem, Francisco M Cornejo-Castillo, Verónica Benítez-Barrios, Eugenio Fraile-Nuez, X Antón Álvarez-Salgado, Carlos M Duarte, Josep M Gasol, and Silvia G Acinas. 2016. “Global Diversity and Biogeography of Deep-Sea Pelagic Prokaryotes.” *The ISME Journal* 10 (3): 596–608. https://doi.org/10.1038/ismej.2015.137.

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.

Shank, Stephen D., Steven Weaver, and 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.

Vargas, Colomban de, Stéphane Audic, Nicolas Henry, Johan Decelle, Frédéric Mahé, Ramiro Logares, Enrique Lara, et al. 2015. “Ocean Plankton. Eukaryotic Plankton Diversity in the Sunlit Ocean.” *Science (New York, N.Y.)* 348 (6237): 1261605. https://doi.org/10.1126/science.1261605.