The Ocean Barcode Atlas: a web service to explore the biodiversity and biogeography of marine organisms.

Caroline Vernette[1], Nicolas Henry[2], Julien Lecubin[1], Colomban de Vargas[2], Pascal Hingamp[1], and Magali Lescot[1]

[1]CNRS
[2]Station Biologique de Roscoff

October 5, 2020

Abstract

The Ocean Barcode Atlas (OBA) is a user friendly web service designed for biologists who wish to explore the biodiversity and biogeography of marine organisms locked in otherwise difficult to mine planetary scale DNA metabarcode datasets. Using just a web browser, a comprehensive picture of the diversity of a taxon or a barcode sequence is visualized graphically on world maps and interactive charts. Interactive results panels allow dynamic threshold adjustments and the display of diversity results in their environmental context measured at the time of sampling (temperature, oxygen, latitude, etc.). Ecological analyses such as alpha and beta-diversity plots are produced via publication quality vector graphics representations. Currently, the Ocean Barcode Altas is deployed online with the i) Tara Oceans eukaryotic 18S-V9 rDNA metabarcodes, ii) Tara Oceans 16S/18S rRNA miTags, and iii) 16S-V4V5 metabarcodes collected during the Malaspina-2010 expedition. Additional prokaryotic or eukaryotic plankton barcode datasets will be added upon availability, given they provide the required complement of barcodes (including raw reads to compute barcode abundance) associated with their contextual environmental variables. Ocean Barcode Atlas is a freely-available web service at: http://oba.mio.osupytheas.fr/ocean-atlas/.

1. INTRODUCTION

Our planet is losing biodiversity at an unprecedented rate, and it is urgent today to map total biodiversity on Earth in order to assess how biodiversity is affected by global climate change. The ocean contains 97% of all water on our planet and is thus a fundamental biodiversity reservoir and driver of global ecology. Marine plankton form the base of ocean food webs and play a major role in the planet's global biogeochemistry balance by accounting for almost half of the net primary production (Falkowski et al., 2008; Field et al., 1998), and thus drive ocean oxygen production and the biological carbon pump (Guidi et al., 2016). However, global ocean physics and chemistry are changing rapidly and it is expected that plankton diversity and geographic distribution will be fundamentally altered in the coming decades (Ibarbalz et al., 2019).

Ever since the first large scale DNA sequencing survey of marine plankton undertaken by the Global Ocean Sampling expedition in 2007 (Rusch et al., 2007), other planetary-scale expeditions have deployed holistic sampling protocols to assess ocean ecosystems. Importantly, the latter have measured the *in situ* biogeochemical parameters that provide the environmental context necessary for ecological interpretation of plankton communities. One such international endeavour, *Tara* Oceans 2009-2013 (Karsenti et al., 2011) sampled viruses to zooplancton using a standardized pan-ecosystemic protocol at 210 globally distributed stations and three depths down to 1,000 m. The Malaspina-2010 (2010-2011) global circumnavigation expedition (Duarte, 2015) applied a similar approach with a particular emphasis in sampling the dark meso- and bathy-pelagic tropical and subtropical waters from surface down to 4,000 m depth.

During the same decade, rapid progress in high-throughput DNA sequencing technology (HTS) has led

to a thorough re-assessment of biodiversity in ecosystems and biomes. In particular, deep sequencing of environmental DNA or RNA amplicons can now reveal prokaryotic and eukaryotic biological diversity close to saturation in even the richest samples (Geisen et al., 2019). Such a metabarcode approach has provided comprehensive surveys of biological communities contained in plankton samples collected during the *Tara* Oceans and Malaspina expeditions. The resulting ocean metabarcodes have allowed a re-evaluation of eukaryotic diversity (de Vargas et al., 2015), a global description of plankton biogeography (Richter et al., 2019), and insights into key plankton players in carbon export (Guidi et al., 2016).

However, the Terabyte magnitude and complexity of these new datasets restrict their access to specialized bioinformatics teams, leaving a large majority of researchers interested in plankton diversity high and dry. Apart from the sheer volume of sequencing reads, their clustering and annotation as well as their connection to environmental data, contribute to rendering this precious data underexploited by biological oceanographers. The simple ergonomic tools to access and extract biological meaningful information that were developed for marine gene catalogs derived from metagenomes and metatranscriptomes (Villar et al., 2018) have so far been lacking for metabarcode datasets. The Ocean Barcode Atlas (OBA) has been developed to assist ocean researchers without specific bioinformatics expertise to easily explore metabarcodes (metaB) of interest across the global ocean ecosystem using nothing else than a web browser. Robust quantitative and contextualized analyses are carried out on the fly within minutes, compared to the several hours (more frequently days) of specialized bioinformatics computation on dedicated high-performance hardware that are required without such a web service. The OBA service (http://tara-oceans.mio.osupytheas.fr/) is independent but complementary to the previously described Ocean Gene Atlas (OGA, http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/; Villar et al., 2018). Indeed, the OBA reported here relies on metabarcode sequences, and as such allows users to explore plankton biodiversity from a taxonomic perspective, providing answers such as "how is a specific plankton taxon distributed across the oceans?". The previously published OGA, being based on metagenomic sequences, is designed to explore the biogeography of plankton gene functions, enabling users to answer questions such as "where in the marine biome are genes related to anaerobic ammonium oxidation to be found?".

The initial version of the OBA currently integrates three large metabarcode datasets: i) the *Tara* Oceans 18S-V9 rRNA metaB (de Vargas et al., 2015; Ibarbalz et al., 2019), ii) the *Tara* Oceans 16S/18S rRNA $_{\text{mi}}$Tags (Logares et al., 2014; Salazar et al., 2019) and iii) the Malaspina-2010 16S-V4V5 rRNA metaB (Salazar et al., 2015).

2. MATERIALS AND METHODS

2.1. IMPLEMENTATION

The Ocean Barcode Atlas web server is implemented through a classical Model-View-Controller pattern architecture using the Laravel 5.4 PHP framework. The application server communicates with the user through an Apache HTTP server using HTML5, CSS3, JS, BLADE and AJAX to retrieve user requests and display results. The PHP application server queries abundance and environmental data stored in a mySQL relational database. The amchart javascript library (https://www.amcharts.com/) was used to build maps and bubble plots, whilst phylotree (https://github.com/veg/phylotree.js/tree/master) and jstree (https://www.jstree.com/) were used to draw phylogenetic trees and choose a taxon respectively. The vegan R package was used for diversity analysis (https://cran.ism.ac.jp/web/packages/vegan/vegan.pdf whilst the ggplot R package (https://cran.r-project.org/src/contrib/Archive/ggplot/) was used for producing plots.

2.2. DATA SOURCES

2.2.1. Barcode/OTU datasets

The following three metabarcode datasets were released in OBA.

The *Tara* Oceans 18S-V9 rRNA metabarcode dataset consists of eight size-fractionated communities obtained from two depths in the photic zone (subsurface=SRF, Deep-Chlorophyll Maximum=DCM), one from the mesopelagic zone (MES) and one from the marine epipelagic mixed layer (MIX). Size fractionations

corresponded to filter collected pico- and nano-plankton (0.8-5 µm), and plankton net tows for the nano-, micro-, and meso- plankton (respectively, 5-20 µm, 20-180 µm and 180-2,000 µm) (http://taraoceans.sb-roscoff.fr/EukDiv/; de Vargas et al., 2015). This dataset was built by sequencing plankton metabarcodes and assembling 1,685,214,722 raw reads, from 1,046 samples including *Tara* Oceans Polar Circle expedition (https://figshare.com/s/cfbf869ca84310fda6bb; Ibarbalz et al., 2019). Metabarcodes were clustered into biologically meaningful 474,303 OTUs, using the '*Swarm'* approach (Mahé et al., 2014). For the taxonomic assignment of metabarcodes, the *P* rotist*R* ibosomal *R* eference -PR^2- database was used (Guillou et al., 2013).

The *Tara* Oceans 16S/18S rRNA $_{\text{mi}}$Tags dataset consists of two size-fractionated communities (0.22 to 1.6 µm and 0.22 to 3µm) that were obtained from two depths in the photic zone (subsurface=SRF, Deep-Chlorophyll Maximum=DCM), as well as one depth is the mesopelagic zone (MES) and one in the marine epipelagic mixed layer (MIX). The metagenomics reads corresponding to both size fractions (enriched in prokaryotes and giant viruses) described in (Salazar et al., 2019) are available at

https://www.ocean-microbiome.org and https://zenodo.org/record/3473199. For each prokaryote-enriched sample (N=180), merged 19,037,038 raw reads Illumina reads (miTags) that contained signatures of the 16S/18S rRNA gene were extracted (Logares et al., 2014). These fragments were mapped to a set of 16S/18S reference sequences that were downloaded from the SILVA database (Release 128: SSU Ref NR 99; https://www.arb-silva.de/fileadmin/silva_databases/release_128/Exports/SILVA_128_SSURef_-Nr99_tax_silva.fasta.gz). A total of 23,987 $_{\text{mi}}$Tags sequences were annotated. Abundance tables were built by counting the number of miTags assigned to each taxa in each sample and the number of unassigned$_{\text{mi}}$Tags (https://www.ebi.ac.uk/biostudies/files/S-BSST297/u/OM-RGC_v2_taxonomic_profiles.tar.gz).

The 16S-V4V5 metabarcode dataset from the Malaspina-2010 expedition was built from 60 samples of bathypelagic (BAT: 1000-4000 m) and abyssopelagic (ABY: 4000-6000 m) waters (Salazar et al., 2015) (https://github.com/GuillemSalazar/MolEcol_2015). This metabarcode dataset based on

1,789,427 raw reads contained 3,902 OTU sequences for two plankton size fractions (0.2 to 0.8 µm and 0.8 to 20 µm). The taxonomic assignment was performed using the SILVA database (release 115; https://www.arb-silva.de/fileadmin/silva_databases/release_115/Exports/SSURef_NR99_115_tax_silva.fasta.tgz). Abundance tables contained the number of reads for the OTUs of particle-attached (PA) and free-living (FL) prokaryotes detected in 30 globally distributed sampling stations (https://github.com/GuillemSalazar/MolEcol_-2015/blob/master/OTUtable_Salazar_etal_2015_Molecol_norarefac.txt).

2.2.2. Barcode/OTU abundances

The relative abundance of each metabarcode or OTU was estimated by dividing the number of barcode/OTU counts by the total number of metabarcodes in the corresponding sample.

2.2.3. Diversity computation

Diversity computations were carried out with the R *vegan* package (version 2.5-5) (Oksanen et al., 2019). Shannon diversity index was calculated with the *diversity* function. The richness was calculated from rarefied counts (*rarefy* function, random subsamples of the size of the sample with the lowest number of reads). For the computation of inter-sample diversity (beta-diversity), the matrices were generated from the rarefied abundances.

If the "*Bray-Curtis dissimilarity* " option is selected on the query page, the abundances are normalized using the *Hellinger*method from the *decostand* function and a dissimilarity matrix is produced with the *vegdist* function.

For the "*Jaccard index* " option, the abundance matrix is transformed into a presence/absence matrix (0 and 1). The distance matrix is calculated with the *vegdist* function using the Jaccard option.

Based on these matrices, a nMDS is performed with the *metaMDS*function. The stress value shown on the graph indicates how similar the distances between samples in the ordination space are to the original

distances. The closer the stress value is to 0, the more similar the distances are.

The "*With environmental vectors* " button plots an additional nMDS graphic with projections of variable vectors. The script uses the*envfit* function to fit environmental vectors or factors onto an ordination. The data projections onto vectors have a maximum correlation with corresponding environmental variables and the factors show the averages of factor levels. The variables with a percentage of missing data higher than 15 are removed.

2.2.4. Environmental context

For the *Tara* Oceans samples (Pesant et al., 2015), geo-localization and biogeochemical characteristics of the sampled seawater were obtained from PANGAEA (https://doi.org/10.1594/PANGAEA.875582). The environmental variables provided by OBA (Figure 2) were either classical oceanographic measures obtained *in situ* (e.g. depth, salinity, temperature, oxygen, chlorophyll a, etc.) or mesoscales features estimated from oceanographic models and remote satellite observations (e.g. nutrient concentration at 5m depth or net primary production). Estimated values are indicated by a star in the drop-down menu of bubble plot panels (Figure 2B and C). Descriptions of the environmental variables available in OBA and corresponding PANGAEA hyperlinks are provided in the OBA user manual hyperlinked from the OBA results page.

For each Malaspina-2010 sample, environmental variables collected as described in (Fernández-Castro et al., 2014; Logares et al., 2020; Salazar et al., 2015) were extracted from the metadata file found at:

https://github.com/GuillemSalazar/MolEcol_2015/blob/master/Metadata_Salazar_etal_2015_Molecol.txt.

2.3. INTERFACE AND FUNCTIONALITY

An OBA online analysis session begins with a submission interface that collects a user defined barcode sequence query. The OBA server then executes data mining procedures on dedicated high-performance hardware, and returns interactive result panels for data exploration (Figure1). A user manual is available online as well as case study example sequences (http://oba.mio.osupytheas.fr/ocean-gene-atlas/build/pdf/Ocean-Barcode-Atlas_User_Manual.pdf).

Two distinct query submission interfaces are proposed (Figure 1). On the one hand, the "Community ecological analysis" allows the user to search for a taxonomically annotated OTU in a metabarcode dataset in order to obtain the biogeographic abundance of this taxon (world maps and bubble plots) as well as ecological analyses (alpha- and beta-diversity) in the form of graphical representations such as non-Metric multiDimensional Scaling (nMDS), scatter plots and boxplots. On the other hand, the "Sequence based query" allows the user to interrogate a metabarcode dataset by sequence similarity in order to obtain the abundance, location and diversity of targeted sequences (barcodes) in an environmental context.

2.3.1. Community ecological analysis

An interactive tree first allows the user to make a taxon preselection. Then, when the cursor is placed in the taxonomic classification text field, an auto-completion feature assists the user in defining a more specific taxonomic range. The resulting beta-diversity calculation option provides a nMDS ordination from either a Bray-Curtis dissimilarity matrix if the "Bray-Curtis dissimilarity" option is selected or a Jaccard distance matrix if the "Jaccard index" option is selected.

2.3.2. Sequence based query

Three origins for the sequence query are proposed in three distinct tabs of the submission interface:

Submit your sequence. If the user has a FASTA-formated barcode sequence that matches the barcode sub-region targeted by the dataset (eg. the V9 sub-region of the 18S), the first option (eg "18SV9 region") can be selected. If not, the user can select the second option (eg "18S complete") which will extract the corresponding sub-region from the sequence query with cutadapt version 2.1 software (Martin, 2011).

Search from a ref db sequence. A taxonomic search allows the user to identify a barcode from a reference database (PR^2 for *Tara* Oceans OTU 18S-V9 (Guillou et al., 2013) or SILVA release 115 (Quast et al., 2013) for *Tara* Oceans 16S rRNA$_{\text{mi}}$Tags and Malaspina-2010 OTU 16S-V4V5).

Search from an ID. This third tab caters for users with a list of OTU identifiers to use as queries (the OTU identifiers must correspond to those used in the original datasets). A "*one map per barcode* " option is available if less than 5 barcodes are selected, allowing each barcode to be presented on a separate map or bubble plot in the results panels.

If one of the two first tabs has been used to define the query ("*Submit your sequence* " or "*Search from a ref db sequence* "), an alignment is computed (using VSEARCH) (Rognes et al., 2016) between the selected barcode query and the OTU sequences of the selected metabarcode dataset. An optional phylogenetic tree can be built in order to compare the user barcode query sequence with its homologous target OTU sequences.

3. RESULTS

3.1. Interactive results panels

The results interface presents the user with graphical representations of computation results via maps, bubble plots, krona charts (Ondov et al., 2011), phylogeny trees, alpha and beta-diversity plots, nMDS ordinations and barcharts (Figure 2). The results are organized by sample on graphs (except for the overall krona chart), and sample contextual information is displayed on mouse hover over the coloured circles on the maps and bubble plots. The results are stored and remain available via the web page URL for 15 days after job submission.

3.1.1. *Community ecological analysis results*

The maps representing OTU geographic distribution presented as a result of a "*Community ecological analysis* " query are computed using three distinct metrics: i) relative abundance, ii) Shannon diversity index and iii) richness (number of OTUs) for each sample depths (Figure 2A). Abundance is calculated as barcode counts divided by the sum of counts for the corresponding sample. Richness is computed from rarefied barcode counts. Relative abundance, Shannon diversity index or richness comparisons are possible between distinct plankton size fractions and/or sampling depths using distinct world maps.

Co-variation of barcodes abundance and environmental variables can be examined on bubble plots (Figure 2B) for selected combinations of sampling depths and plankton size fractions. A scatter plot is drawn if the user selects the option "*Abundance/environmental variables* " (Figure 2C).

The taxonomic distributions of the target barcodes are displayed in multi-layered and interactive krona charts for each distinct sample (by clicking on the circles in the world map; Figure 2A) and for the full dataset (Figure 2E).

The beta-diversity of the selected taxon using either Bray-Curtis dissimilarity or Jaccard distance is represented by nMDS ordinations with optional projections of the 3 selected variable vectors using the R vegan package *envfit* function (Figure 2F). A glossary of available environmental variables - both measured *in situ* and estimated from modelling - is provided to help users select variables of interest. Dot colors and shapes on the graphs are related to sample depths and plankton size fractions.

Alpha-diversity is represented by two box- or scatter plots (Figure 2G). The first plot represents the richness (OTU number) according to a user selected environmental variable, whilst the second represents the Shannon diversity index.

Finally, relative abundance and richness of rDNA metabarcodes are represented on bar charts for the 9 most abundant taxonomy groups depending on sample depths and plankton size fractions (Figure 2H).

3.1.2. *Sequence based query results*

Intermediate query selection pages. When the query was submitted via the "*Submit your sequence* " and "*Submit a ref db sequence* " submission interface (Figure 1), intermediate pages allow users to select which homologous metabarcode sequences to use as queries for the analyses. In the case of submission via the "*Submit a ref db sequence* " option, a first intermediate page lists available reference sequences from the PR^2 or SILVA databases, including their taxonomic classification and nucleotide sequences, together with radio buttons to select the appropriate query sequence. The query sequence is then aligned (using VSEARCH by default) with OTU sequences from the selected metabarcode dataset. The following intermediate page then lists matching metabarcodes by order of decreasing sequence similarity (limited to the first 500 hits) (Figure 2D). Above the list, a bar chart showing the distribution of alignment percent identities can be used to select which homologs to carry over for analysis. The "*one map per barcode* " option is only applicable if less than 5 barcodes are selected, allowing each barcode to be visualized on separate maps and bubble plots.

Result interfaces. The biogeography of the metabarcode sequences homologous to the user query is displayed in the following four interactive panels: geographic distribution (Figure2A), co-variation with environmental variables (Figure 2B and C), taxonomic distribution (Figure 2E) and phylogenetic tree (Figure 2I) if the "*Phylogenetic tree* " option was ticked in the initial submission interface (unchecked by default). If the "*Environmental variable filtering* " option is selected on the world map, the user can filter out samples using the provided slider (Figure 3).

Regardless of submission route, the complete set of data necessary to independently re-build each figure is downloadable as flat files from the final results page (namely the list of rDNA hits, the corresponding FASTA formatted sequences, the biosample abundance matrix and the contextual environmental variables).

3.2. THREE CASE STUDIES

3.2.1. Case study 1

Decelle et al. (2018) have demonstrated, using the *Tara* Oceans 18S-V9 metabarcode dataset, that the dinoflagellate genus*Symbiodinium* well known for sustaining coral reefs photosymbiosis, is ecologically and economically important in the open ocean. The relative abundance of Symbiodiniaceae metabarcodes in this study was high in tropical and sub-tropical waters, whilst reaching undetectable levels at higher latitudes (Figure 1B in (Decelle et al., 2018)), indicating an ecological preference for warm and oligotrophic oceanic waters. We searched for Symbiodiniaceae as a taxon query in the OBA "*Community ecological analysis* " interface targeting the*Tara* Oceans 18S-V9 metabarcode dataset. The resulting world map (Figure 4A) and relative abundance versus latitude scatterplot (Figure 4B) confirm the strong bounding of Symbiodiniaceae under 40 degrees of latitude as in (Decelle et al., 2018).

3.2.2. Case study 2

The oceanic distribution of Chlorophyta was reported by (Lopes dos Santos et al., 2017) based on the *Tara* Oceans 18S-V9 rRNA metabarcode dataset. In this study, prasinophytes clade VII and mamiellophyceae appeared to be the most abundant group of Chlorophyta in the oceans. Prasinophytes clade VII were abundant in open ocean waters whereas mamielophyceae were abundant in a larger spectrum of environments, including high latitudes as well as the Benguela current (Figure 4 in (Lopes dos Santos et al., 2017)). Using the OBA "*Community ecological analysis* " interface applied to the same metabarcode dataset, we obtained similar oceanic distributions of these two Chlorophyta groups ("prasino clade VII" and "mamiellophyceae") in the plankton size fraction ranging from 0.8 to 5 µm. Indeed, figures 6A and 6B confirm respectively that prasinophytes clade VII were more abundant in open ocean waters, whereas Mamiellophyceae were also abundant in coastal waters. Moreover, the scatter-plots of their relative abundance (% OTUs) versus sampling latitude illustrates the clear bounding of prasinophytes clade VII within 40 degrees of latitude (Figure 5C), whereas Mamiellophyceae extend into high latitudes both in the northern and southern hemispheres (Figure 5D).

3.2.3. Case study 3

The study of marine diplonemids diversity (Flegontova et al., 2016) revealed that they were stratified ac-

cording to depth, with a large fraction of OTUs (35.6%) concentrated in the mesopelagic zone (Figure 4A, (Flegontova et al., 2016)). In Figure 4B of (Flegontova et al., 2016), a nMDS based on Bray-Curtis dissimilarities between samples supported the difference in diplonemid community composition and showed that the mesopelagic samples stood apart from the epipelagic samples (surface and DCM). Using the OBA "*Community ecological analysis* " and querying the *Tara* Oceans 18S-V9 metabarcode dataset with the taxonomy 'Diplonemida", we were able to obtain similar results (Figure 6). A nMDS ordination with depth as the environmental variable (Figure 6A) illustrated how mesopelagic samples (green square) clustered mostly separately from the surface and DCM samples (blue boxes and red circles). The boxplot in Figure 6B identified mesopelagic samples as containing the highest Diplonemids diversity, both in terms of richness (number of OTUs; Figure 6B left panel) and Shannon index (Figure 6B right panel).

4. COMPARISON WITH OTHER AVAILABLE SOFTWARE PROGRAMS

Alternative web servers propose subsets of the Ocean Barcode Atlas functionalities, but to the best of our knowledge none offer such tight taxon (or sequence) diversity integration with environmental context across multiple metabarcode datasets.

Authors of the MicrobiomeAnalyst tool (https://www.microbiomeanalyst.ca; Dhariwal et al., 2017) carried out a comparison between several web-based tools such as METAGEN-assist, EBI Metagenomics, MG-RAST and VAMPS. However in the extensive set of features proposed by MicrobiomeAnalyst, several modules were developed specifically for human and mouse models and are thus not suitable for the analysis of marine barcodes. However VAMPS allows to give access to marine datasets but a prior registration is needed and the data is a bite complex to query for a non-expert user.

In contrast, GLOSSary (http://bioinfo.szn.it/glossary/; Tangherlini et al., 2018) is specifically tailored for marine data, but only allows users to query the *Tara* 16S miTAGs dataset with interactive geographic exploration of prokaryotic sequences or taxon, but abundance, diversity indexes, and environmental parameters are lacking.

5. DISCUSSION

The Ocean Barcode Atlas is an interactive DNA metabarcode web service that supports exploratory analysis across the phylogenetic and ecological spaces of a given barcode or taxon. It allows users to concentrate on gaining biological meaning from large plankton barcode datasets, without being hampered by the inherent heterogeneity of the underlying data and the requirement for high performance computing resources. No user account or email address is required to run OBA analyses, which are fast enough to be rendered on the fly, stored for convenience on the server for 15 days and reachable via the URL alone.

Currently provided with five metabarcode datasets, we plan to complement the OBA service with further marine datasets as they become available. For instance, we plan to include the forthcoming *Tara* Oceans 18S-V4 rRNA metabarcode dataset which will usefully complement the current V9 based metabarcodes.

An additional ambition for OBA development, is to represent time series datasets, such as those produced during the Ocean Sampling Day (Kopf et al., 2015). Furthermore, we plan to create Docker containers running OBA in order to allow users to locally analyse and privately share geolocalized barcode datasets. An application programming interface (API) is also being developed allowing programmatic access to OBA resources similar to the OGA API that has proven to be a popular mode of access also within the bioinformatics specialists community.

ACKNOWLEDGEMENT

This article is contribution number XXX of *Tara* Oceans.

The authors would like to thank Guillem Salazar for providing us with the Malaspina-2010 16S-V4V5 OTU tables. We are grateful for the contribution of the ECOMAP team at Roscoff Marine Station, and in particular Miguel Méndez Sandin and Fabrice Not, for testing the web server and suggesting functionalities. Thanks also to Noan Le Bescot for the plankton drawing used on the interfaces. The web server is hosted

by the OSU Pythéas cluster with the help of Cyrille Blanpain and the SIP members. Adrien Malgoyre from the SIP is thanked for the development of the OSU Pythéas gitlab. We are grateful to the Institut Français de Bioinformatique for providing help and computing resources. *Tara*Oceans (which includes both the *Tara* Oceans and *Tara*Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (http://oceans.taraexpeditions.org). We further thank the commitment of the following sponsors: CNRS (in particular Groupement de Recherche GDR3280 and the Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* Oceans-GOSEE), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, The French Ministry of Research, and the French Government 'Investissements d'Avenir' programmes, FRANCE GENOMIQUE, MEMO LIFE and PSL* Research University. We also thank the support and commitment of agnès b. and Etienne Bourgois, the Prince Albert II de Monaco Foundation, the Veolia Foundation, Region Bretagne, Lorient Agglomeration, Serge Ferrari, Worldcourier, and KAUST. The global sampling effort was enabled by countless scientists and crew who sampled aboard the *Tara* from 2009-2013, and we thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expeditions. We are also grateful to the countries who graciously granted sampling permissions. The authors declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the analyses, publications, and ownership of data are free from legal entanglement or restriction by the various nations whose waters the*Tara* Oceans expeditions.

This work was supported by the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL* Research University (ANR-11-IDEX-0001-02).

AUTHOR CONTRIBUTIONS

C.V. designed the database schema, the associated query system, and implemented on the server. C.V. and N.H. conceived the overall analysis pipeline strategy. N.H. designed some analyses and tested the platform. J.L. implemented the server system architecture (virtual machine, storage and backups). C.d.V. contributed to the coordination of the scientific project and supervised developments at the SBR. P.H. tested the platform and contributed to the writing of the manuscript. M.L. contributed to the coordination of the scientific project and supervised developments at the MIO, tested the platform and wrote the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

Ocean Barcode Atlas is a free available web service at http:/oba.mio.osupytheas.fr/ocean-atlas/. Information about OBA, datasets and R scripts are available at: https://gitlab.osupytheas.fr/ocean_atlas/oba.

ORCID

Caroline Vernette: https://orcid.org/0000-0002-4784-9605

Nicolas Henry: https://orcid.org/0000-0002-7702-1382

Julien Lecubin: https://orcid.org/0000-0002-0090-4232

Colomban de Vargas: https://orcid.org/0000-0002-6476-6019

Pascal Hingamp: https://orcid.org/0000-0002-7463-2444

Magali Lescot: https://orcid.org/0000-0001-8189-0823

REFERENCES

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., ... Karsenti, E. (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

Decelle, J., Carradec, Q., Pochon, X., Henry, N., Romac, S., Mahé, F., Dunthorn, M., Kourlaiev, A., Voolstra, C. R., Wincker, P., & de Vargas, C. (2018). Worldwide Occurrence and Activity of the Reef-Building Coral Symbiont Symbiodinium in the Open Ocean. *Current Biology: CB* ,*28* (22), 3625-3633.e3. https://doi.org/10.1016/j.cub.2018.09.024

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., & Xia, J. (2017). MicrobiomeAnalyst : A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research* , *45* (W1), W180-W188. https://doi.org/10.1093/nar/gkx295

Duarte, C. M. (2015). Seafaring in the 21St Century : The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin* , *24* (1), 11-14. https://doi.org/10.1002/lob.10008

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science (New York, N.Y.)* , *320* (5879), 1034-1039. https://doi.org/10.1126/science.1153213

Fernandez-Castro, B., Mourino-Carballido, B., Benitez-Barrios, V. M., Choucino, P., Fraile-Nuez, E., Grana, R., Piedeleu, M., & Rodriguez-Santana, A. (2014). Microstructure turbulence and diffusivity parameterization in the tropical and subtropical Atlantic, Pacific and Indian Oceans during the Malaspina 2010 expedition. *Deep Sea Research Part I: Oceanographic Research Papers* , *94* , 15-30. https://doi.org/10.1016/j.dsr.2014.08.006

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary Production of the Biosphere : Integrating Terrestrial and Oceanic Components. *Science* , *281* (5374), 237-240. https://doi.org/10.1126/science.281.5374.237

Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., de Vargas, C., Bowler, C., Lukeš, J., & Horák, A. (2016). Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Current Biology* ,*26* (22), 3060-3065. https://doi.org/10.1016/j.cub.2016.09.031

Geisen, S., Vaulot, D., Mahe, F., Lara, E., Vargas, C. de, & Bass, D. (2019). A user guide to environmental protistology : Primers, metabarcoding, sequencing, and analyses. *BioRxiv* , 850610. https://doi.org/10.1101/850610

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J., Coelho, L. P., Espinoza, J. C. I., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., … Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* , *532* (7600), 465-470. https://doi.org/10.1038/nature16942

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (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-D604. https://doi.org/10.1093/nar/gks1160

Ibarbalz, F. M., Henry, N., Brandao, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahe, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Saez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., … Zinger, L. (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

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., … Tara Oceans Consortium. (2011). A holistic approach to marine eco-systems biology. *PLoS Biology* ,*9* (10), e1001177. https://doi.org/10.1371/journal.pbio.1001177

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., Wichels, A., Gerdts, G., Polymenakou, P., Kotoulas, G., Siam, R.,

Abdallah, R. Z., Sonnenschein, E. C., Cariou, T., O'Gara, F., ... Glockner, F. O. (2015). The ocean sampling day consortium.*GigaScience* , *4* , 27. https://doi.org/10.1186/s13742-015-0066-5

Logares, R., Deutschmann, I. M., Giner, C. R., Krabberod, A. K., Schmidt, T. S. B., Rubinat-Ripoll, L., Mestre, M., Salazar, G., Ruiz-Gonzalez, C., Sebastian, M., Vargas, C. de, Acinas, S. G., Duarte, C. M., Gasol, J. M., & Massana, R. (2018). Different processes shape prokaryotic and picoeukaryotic assemblages in the sunlit ocean microbiome. *BioRxiv* , 374298. https://doi.org/10.1101/374298

Logares, R., Deutschmann, I. M., Junger, P. C., Giner, C. R., Krabberod, A. K., Schmidt, T. S. B., Rubinat-Ripoll, L., Mestre, M., Salazar, G., Ruiz-Gonzalez, C., Sebastian, M., de Vargas, C., Acinas, S. G., Duarte, C. M., Gasol, J. M., & Massana, R. (2020). Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* , *8* (1), 55. https://doi.org/10.1186/s40168-020-00827-8

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., Raes, J., Poulain, J., Jaillon, O., Wincker, P., Kandels-Lewis, S., Karsenti, E., Bork, P., & Acinas, S. G. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology* , *16* (9), 2659-2671. https://doi.org/10.1111/1462-2920.12250

Lopes dos Santos, A., Gourvil, P., Tragin, M., Noel, M.-H., Decelle, J., Romac, S., & Vaulot, D. (2017). Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *The ISME Journal* , *11* (2), 512-528. https://doi.org/10.1038/ismej.2016.120

Mahe, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm : Robust and fast clustering method for amplicon-based studies.*PeerJ* , *2* , e593. https://doi.org/10.7717/peerj.593

Martin, M. (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, J., Blanchet, F. G., & Friendly, M. (2019, mai 8).*Vegan.pdf* . Package ¡¡ vegan ¿¿. https://cran.ism.ac.jp/web/packages/vegan/vegan.pdf

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* ,*12* (1), 385. https://doi.org/10.1186/1471-2105-12-385

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Trouble, R., Dimier, C., & Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* , *2* . https://doi.org/10.1038/sdata.2015.23

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project : Improved data processing and web-based tools.*Nucleic Acids Research* , *41* (D1), D590-D596. https://doi.org/10.1093/nar/gks1219

Richter, D. J., Watteaux, R., Vannier, T., Leconte, J., Fremont, P., Reygondeau, G., Maillet, N., Henry, N., Benoit, G., Fernandez-Guerra, A., Suweis, S., Narci, R., Berney, C., Eveillard, D., Gavory, F., Guidi, L., Labadie, K., Mahieu, E., Poulain, J., ... Coordinators, T. O. (2019). Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *BioRxiv* , 867739. https://doi.org/10.1101/867739

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). VSEARCH : A versatile open source tool for metagenomics. *PeerJ* ,*4* , e2584. https://doi.org/10.7717/peerj.2584

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., ... Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition : Northwest Atlantic through eastern tropical Pacific. *PLoS Biology* , *5* (3), e77. https://doi.org/10.1371/journal.pbio.0050077

Salazar, G., Cornejo-Castillo, F. M., Borrull, E., Diez-Vives, C., Lara, E., Vaque, D., Arrieta, J. M., Duarte, C. M., Gasol, J. M., & Acinas, S. G. (2015). Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. *Molecular Ecology* ,*24* (22), 5692-5706. https://doi.org/10.1111/mec.13419

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., Field, C. M., Coelho, L. P., Cruaud, C., Engelen, S., Gregory, A. C., Labadie, K., Marec, C., Pelletier, E., Royo-Llonch, M., Roux, S., Sanchez, P., Uehara, H., Zayed, A. A., ... Sunagawa, S. (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

Tangherlini, M., Miralto, M., Colantuono, C., Sangiovanni, M., Dell' Anno, A., Corinaldesi, C., Danovaro, R., & Chiusano, M. L. (2018). GLOSSary : The GLobal Ocean 16S subunit web accessible resource.*BMC Bioinformatics* , *19* (Suppl 15), 443. https://doi.org/10.1186/s12859-018-2423-8

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

TABLE AND FIGURES LEGENDS

Figure 1. The Ocean Barcode Atlas query schema.

Figure 2. Graphical results produced by the Ocean Barcode Atlas.

Figure 3. The Interactive world map of the barcodes abundance distribution using a range for an environmental variable.

Figure 4. Case study 1. Biogeography of the Symbiodiniaceae V9 rDNA metabarcodes (A) and scatter plot of their relative abundance (B) in the open ocean among the total reads of the eukaryotic community across different latitudes on surface samples and plankton size fractions (piconano: 0.8–5 µm; micro: 20–180 µm; meso: 180–2,000 µm).

Figure 5. Case study 2. Biogeography of prasinophytes Clade VII (A) and mamiellophyceae (B) V9 rDNA metabarcodes.Scatter plots show their relative abundance across different latitudes in surface samples for pico-nanoplankton size fraction (0.8–5 µm) corresponding to prasinophytes Clade VII (C) and mamiellophyceae (D).

Figure 6. Case study 3. Ecological analyses of Diplonemids V9 rDNA metabarcodes. Beta-diversity analysis (A) showing a non-metric multidimensional scaling analysis based on pairwise Bray-Curtis distances among samples (using depth as the environmental variable). Alpha-diversity analysis (B) showing boxplots of OTU number versus sampling depth and Shannon index versus sampling depth.

Figure 1

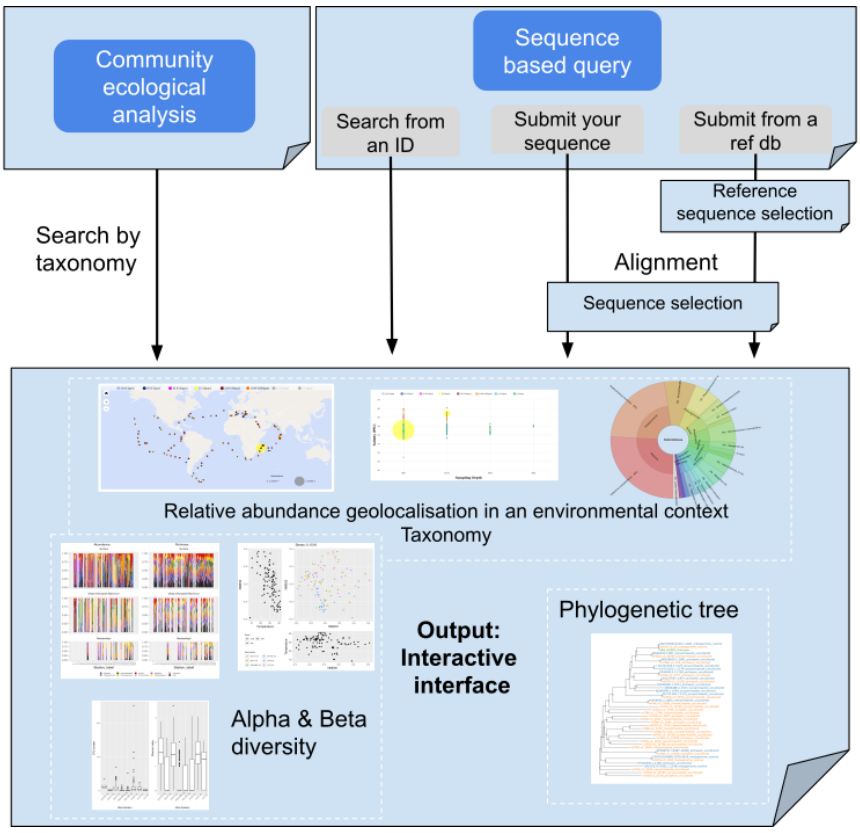

Figure 2

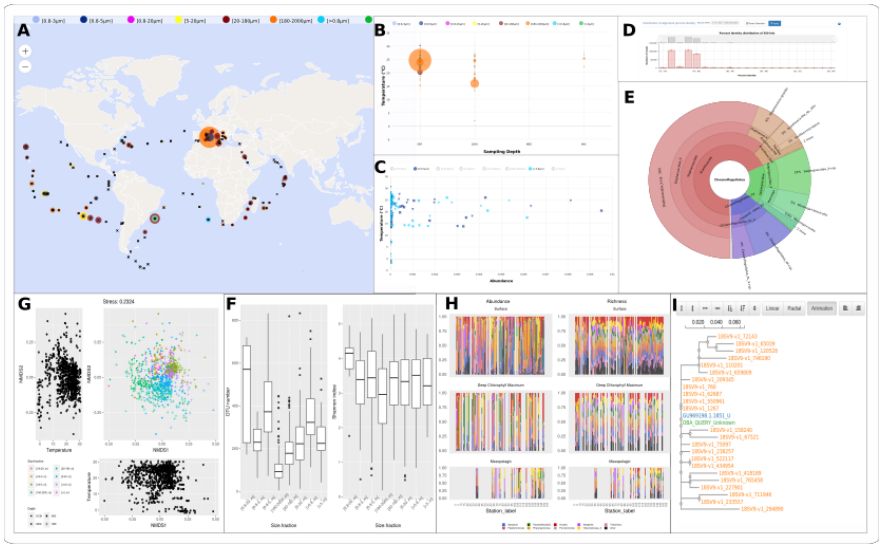

Figure 3

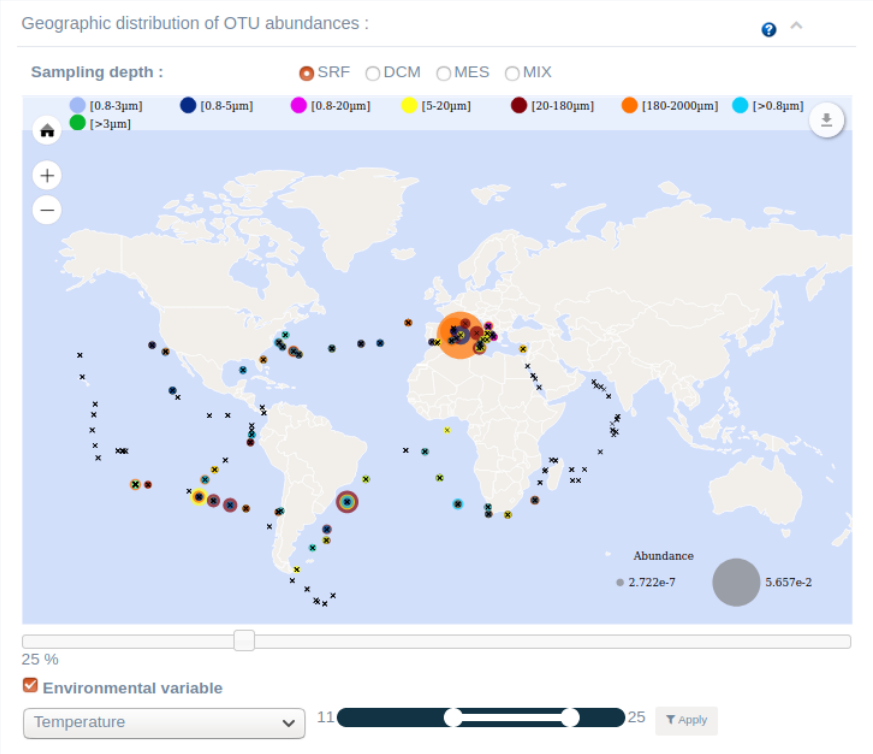

Figure 4

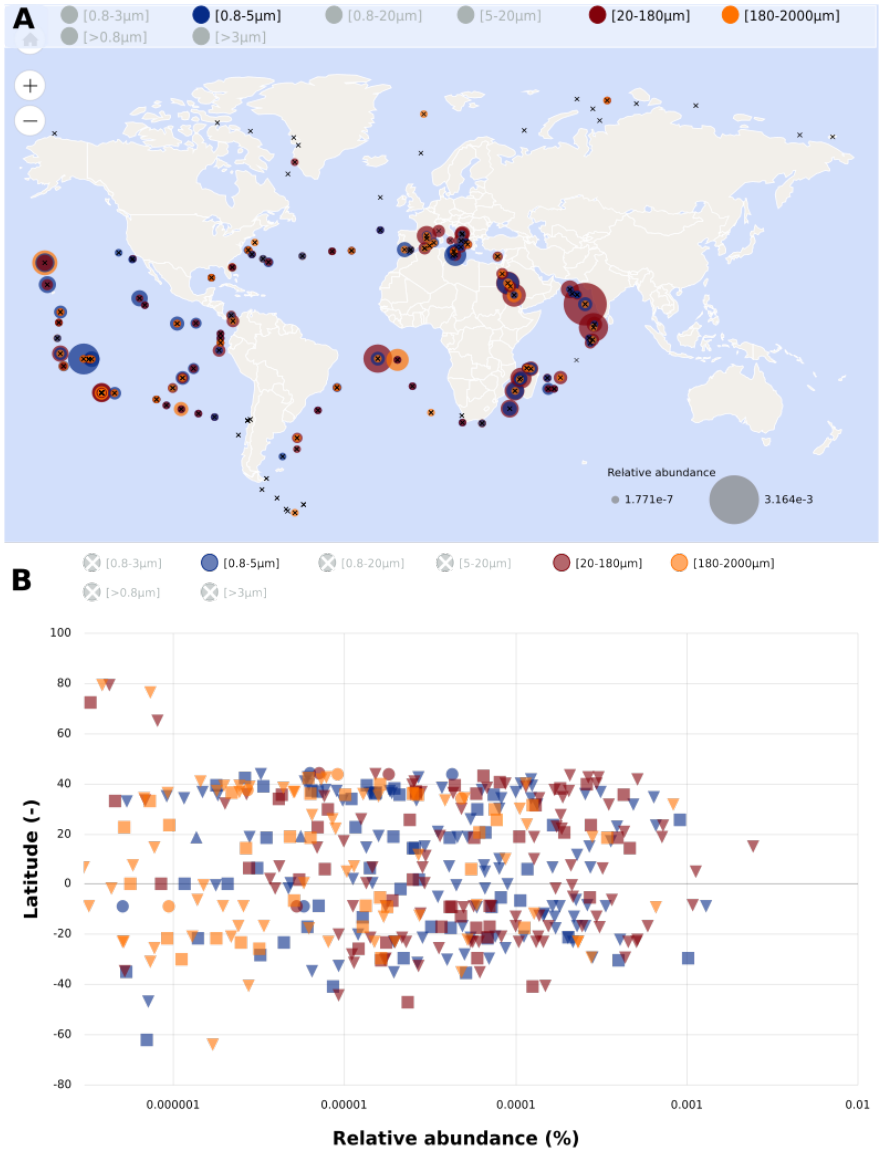

Figure 5

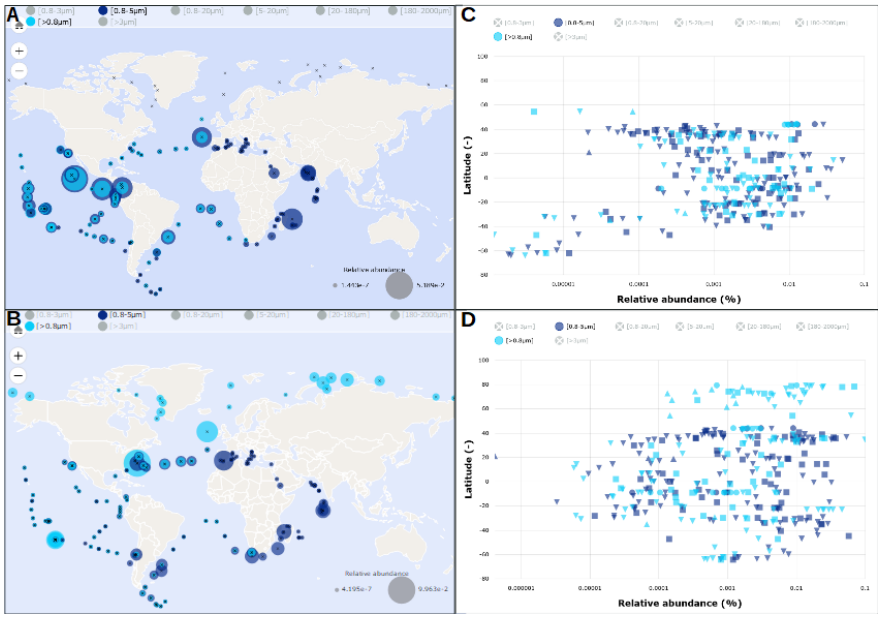

Figure 6