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

Emilie Villar, Thomas Vannier, Caroline Vernette, Magali Lescot, Miguelangel Cuenca, Aurélien Alexandre, Paul Bachelerie, Thomas Rosnet, Eric Pelletier, Shinichi Sunagawa, et al.

*doi: 10.1093/nar/gky376*

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

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

[1]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin UMR7144, Station Biologique de Roscoff, Roscoff, France, [2]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France, [3]Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland and [4]Génomique Métabolique, Genoscope, Institut de Biologie François-Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91000 Evry, France

Received January 31, 2018; Revised April 16, 2018; Editorial Decision April 26, 2018; Accepted May 02, 2018

## ABSTRACT

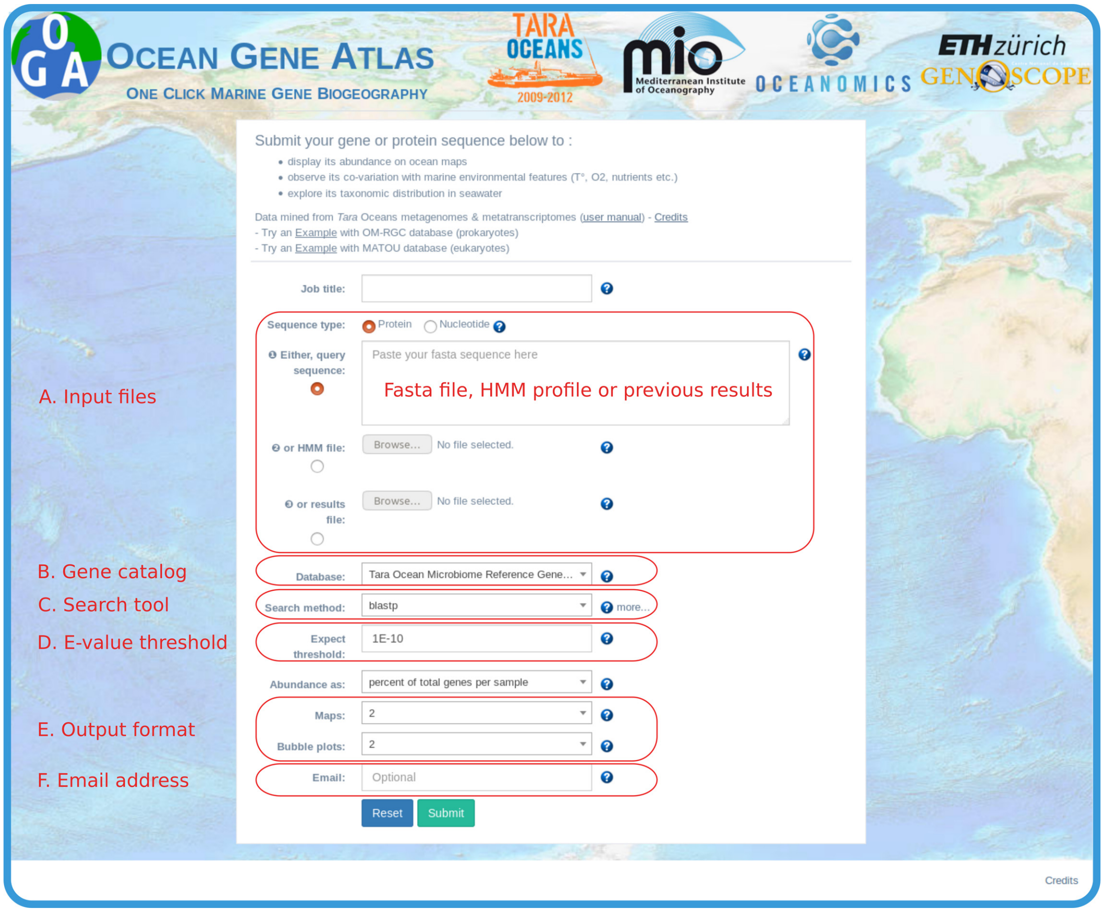
**The Ocean Gene Atlas is a web service to explore the biogeography of genes from marine planktonic organisms. It allows users to query protein or nucleotide sequences against global ocean reference gene catalogs. With just one click, the abundance and location of target sequences are visualized on world maps as well as their taxonomic distribution. Interactive results panels allow for adjusting cutoffs for alignment quality and displaying the abundances of genes in the context of environmental features (temperature, nutrients, etc.) measured at the time of sampling. The ease of use enables non-bioinformaticians to explore quantitative and contextualized information on genes of interest in the global ocean ecosystem. Currently the Ocean Gene Atlas is deployed with (i) the Ocean Microbial Reference Gene Catalog (OM-RGC) comprising 40 million non-redundant mostly prokaryotic gene sequences associated with both *Tara* Oceans and Global Ocean Sampling (GOS) gene abundances and (ii) the Marine Atlas of *Tara* Ocean Unigenes (MATOU) composed of >116 million eukaryote unigenes. Additional datasets will be added upon availability of further marine environmental datasets that provide the required complement of sequence assemblies, raw reads and contextual environmental parameters. Ocean Gene Atlas is a freely-available web service at: http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/.**

## INTRODUCTION

Marine plankton provide essential ecosystemic functions on the planet: at the basis of the ocean food web, they contribute about half of global primary production (1). Plankton are also key players of biogeochemical cycles in the ocean and drive the biological carbon pump (2). Due to their large distribution (oceans represents 71% of the Earth surface) and their substantial biogeochemical roles, plankton are considered to be main actors in the global climate regulation (3) and also constitute a potential source of innovations for blue biotechnology (4). Because of the difficulties associated with sampling plankton in the open and deep ocean, and because of the ultra high throughput required to sequence the genetic makeup of such complex communities, environmental genomics resources for these elusive organisms have only become available recently (reviewed by (5)). The first large scale sequencing of marine microbiomes led by the Global Ocean Sampling expedition (6) produced a 6.1 million gene catalog mostly from sunlit ocean prokaryotes.

More recently, the *Tara* Oceans pan-oceanic expedition deployed a holistic sampling of plankton ranging in size from viruses to fish larvae, coupled with comprehensive *in situ* biogeochemical measurements which provide the detailed environmental contexts necessary for ecological interpretation of marine microbiomes (7). Two complementary genesets have been released so far from the *Tara* Oceans sequencing effort: (i) the Ocean Microbial Reference Gene Catalog (OM-RGC) and (ii) the Marine Atlas of *Tara* Oceans Unigenes (MATOU). The OM-RGC is a comprehensive collection of 40 million genes from viruses, prokaryotes and picoeukaryotes with a size up to 3 µm (8) retrieved from public marine plankton metagenomes and reference genomes. The MATOU is a catalog of 116 million unigenes

[*]To whom correspondence should be addressed. Tel: +33 4 86 09 06 66; Fax: +33 4 86 09 06 41; Email: emilie.villar1@gmail.com
Correspondence may also be addressed to Pascal Hingamp. Email: pascal.hingamp@mio.osupytheas.fr



Submit your gene or protein sequence below to :

- display its abundance on ocean maps
- observe its co-variation with marine environmental features (T°, O2, nutrients etc.)
- explore its taxonomic distribution in seawater

Data mined from *Tara* Oceans metagenomes & metatranscriptomes (user manual) - Credits
- Try an Example with OM-RGC database (prokaryotes)
- Try an Example with MATOU database (eukaryotes)

Job title:

Sequence type: ● Protein ○ Nucleotide

❶ Either, query sequence: ● Paste your fasta sequence here

❷ or HMM file: ○ No file selected.

❸ or results file: ○ No file selected.

Database: Tara Ocean Microbiome Reference Gene...

Search method: blastp

Expect threshold: 1E-10

Abundance as: percent of total genes per sample

Maps: 2

Bubble plots: 2

Email: Optional

**Figure 1.** The Ocean Gene Atlas query submission interface. (**A**) The query can be either (i) a fasta format sequence, (ii) an uploaded HMM profile or (iii) an uploaded results file from a previous search. (**B**) Two gene catalogs are currently available: OM-RGC, a catalog of mostly prokaryotic genes from plankton metagenomes (with associated abundances from Tara oceans and GOS biosamples), and MATOU, a catalog of mostly eukaryotic transcripts from plankton metatranscriptomes. (**C**) The sequence similarity search algorithm is one of BLAST, DIAMOND or HMMER. (**D**) *E*-value threshold to filter the results. (**E**) Selection of the number of interactive panels in the results page. (**F**) Optionally notification of results availability can be sent by email.

obtained from poly-A+ cDNA sequencing of different filter size fractions ranging from 0.8 to 2000 μm (9). About half of the unigenes have a predicted taxonomic assignation representing genes from >8000 mostly eukaryotic organisms.

Such environmental genomics datasets are increasingly used in marine biology, ecology and evolutionary studies to understand the mechanisms by which genes influence phenotype in the wild. In parallel to the holistic systems biology approaches that tackle these datasets as a whole, biologists also frequently apply reductionist approaches that target specific marker genes known or hypothesized to play a role in processes such as metabolic pathways, biotic and abiotic interactions. Testing such candidate marker gene hypotheses requires detailed analyses of voluminous genomic data in their precise environmental context, a task which requires extensive expertise in interrogating heterogeneous data types (i.e. sequencing reads, assemblies, genes together with their taxonomic and functional annotations, environmental variables) to provide integrated interpretations. In order to allow biologists to easily mine such large and complex datasets without the requirement for either significant hardware or programming skills, we make freely available a web service to visualize the geolocalized abundances of taxonomically annotated plankton genes in the context of environmental features.

## INTERFACE AND FUNCTIONALITY

The Ocean Gene Atlas (OGA) web service provides a submission interface to collect a nucleic or protein sequence query, it then executes data mining procedures on dedicated high performance hardware, and returns interactive result panels for data exploration (Figure 1). A user manual and two case study example sequences are provided online.

### The query submission interface

The user-provided query (Figure 1A) may be a nucleotide or amino acid sequence in FASTA format, or a hidden Markov model profile (HMM; 10). Users can search for sequences similar to their query in either one of two marine gene catalogues (Figure 1B): the OM-RGC (8) or the MATOU (9). Users can select different methods to identify sequence similarity (Figure 1C) depending on the type of query (nucleotide/amino acid sequence or HMM) and the user prefered trade-off between accuracy and speed of the

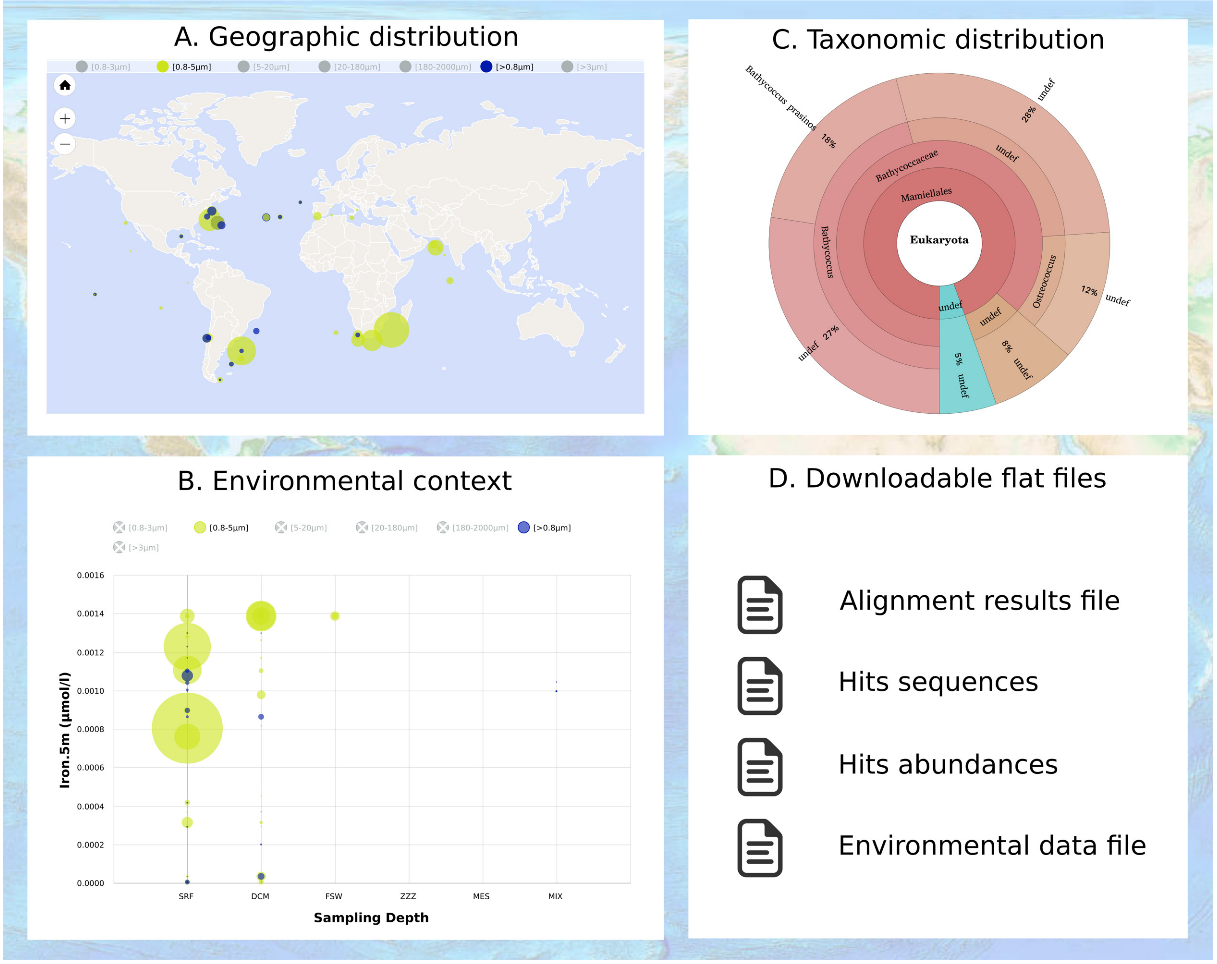


**Figure 2.** The Ocean Gene Atlas interactive results panels. (**A**) Hits abundances are represented by the diameter of filled circle for each sample at user selected sampling depths (e.g. subsurface or mesopelagic). Circle colors represent the filter size fractions (e.g. [0.2–3 µm]). (**B**) Co-variation of hits abundances with specific environmental variables are shown on bubble plots for each sampling depths: subsurface (SRF), deep chlorophyll maximum (DCM) and mesopelagic (MES). (**C**) Taxonomic distribution of the hits genes's predicted origins are represented on interactive Krona plots. (**D**) Result files can be downloaded as tab delimited flat files.

search (11). For protein sequence queries, similarity search tools are BLASTP (12) and Diamond (13). For nucleotide sequence queries, users can choose to carry out the similarity search against a nucleotide database (BLASTN), or to translate the nucleotide query to search against a protein database (BLASTX, DIAMOND BLASTX). For more sensitive sequence similarity searches, users can provide an HMM profile file instead of a FASTA sequence (either prebuilt, e.g. by Pfam (14), or custom-built from protein alignments using the hmmer package, http://hmmer.org). Prebuilt Pfam HMM profiles are also searchable from the OGA submission form by simply filling the corresponding field with the Pfam accession ID. For both sequence similarity and HMM searches, the e-value threshold may be customized (Figure 1D), as well as the number of interactive panels in the results page (Figure 1E). The email field (Figure 1F) is strictly optional since a bookmarkable URL for the results is immediately provided to the user upon submission. Results are usually returned to the user within 30 seconds, but for slower queries (e.g. HMM searches against the larger MATOU catalog which may take up to several minutes or longer in cases of high affluence), the user can choose to provide an email address in order to be notified when results are available (Figure 1F). Results will remain available for 15 days on the OGA web server.

### The interactive result panels

The quantitative distribution of environmental sequences presenting similarities to the user query are displayed in three interactive panels: geographic distribution (Figure 2A), co-variation with environmental features (Figure 2B), and taxonomic distribution (Figure 2C). Furthermore, the underlying data necessary to build the figures can be downloaded as tab delimited flat files for further analysis outside of OGA (Figure 2D): list of similarity search hits and corresponding FASTA formatted sequences, gene per sample abundance matrix, as well as contextual environmental features for each sample. The set of similarity search hits that are included in the three interactive panels can be interactively filtered by click-and-drag adjustment of the E-value threshold directly over the provided E-value distribution histogram. The world maps of Figure 2A display quantitative geographical distributions of the hits as filled circles with sizes proportional to their combined abundance at the user-selected sampling depth, whilst circle colors indicate the size fractionation applied to the sample (e.g. [0.2–3 μm] represents plankton collected on 0.2 μm pore membranes after a 3 μm prefiltration step). The meaning of the acronyms and references to source databases are provided in a user guide hyperlinked on the results page. The side-to-side display of multiple maps enables abundances comparisons between distinct size fractions and/or sampling depths. Co-variation of gene abundances and environmental features can be examined on bubble plots (Figure 2C) for user selectable combinations of sampling depth and size fractions. Finally, the taxonomic distribution of the target genes are displayed in multi-layered and interactive Krona pie-charts (15) either for each distinct sample (by clicking on the circles in the world maps) or for the full dataset (Figure 2C). The charts displayed on the Ocean Gene Atlas results page can be annotated online and downloaded as image files in vector graphics formats (SVG and PDF) suitable for publication.

## DATA SOURCES

Datasets suitable for inclusion in the Ocean Gene Atlas require three complementary data objects: gene sequence catalogs, gene abundances in samples, and sample environmental context (Figure 3).

### Gene catalogs

The building of the OM-RGC and MATOU gene catalogs included in the first release of OGA are detailed in their corresponding release articles (8 and 9 respectively). Briefly, to construct OM-RGC, 7.2 terabases of plankton metagenome shotgun sequencing reads were assembled for 243 *Tara* Oceans samples. Genes were predicted in the assemblies and were clustered at 80% sequence identity together with genes from publicly available marine genomic and metagenomic datasets to generate a non-redundant set of 40 million reference genes. These genes were translated and taxonomically annotated by retrieving the last common ancestor of homologs identified in reference protein sequence databases. The MATOU catalog was obtained from the assembly of 16.5 terabases of plankton metatranscriptome (cDNA sequences corresponding to polyA+ enriched RNA), representing 441 *Tara* Oceans samples. The subsequent contigs obtained for each assembled sample were then clustered at 95% sequence identity to construct a catalog of 116.8 million transcribed sequences. Due to the difficulty of accurate eukaryotic gene calling from low coverage metatranscriptomes, the proteic version of the MATOU catalog was obtained by six frame translation of the nucleic MATOU gene catalog using the sixpack package from the EMBOSS suite (16). About half of the unigenes were taxonomically annotated using a similar last common ancestor approach as described above for the OM-RGC. The vast majority of unigenes were assigned to eukaryota, but a minor proportion (<5%) was annotated as putative bacterial sequences.

### Gene abundances

The abundance of each catalog gene in specific biosamples was estimated by evaluating the coverage of raw sequencing reads mapped to the gene's nucleotide sequence. *Tara* Oceans sequencing reads from 243 metagenomes corresponding to the smallest size fractions (0–0.22, 0.1–0.22, 0.22–0.45, 0.45–0.8, 0.22–1.6 and 0.22–3 μm) and GOS reads from 38 metagenomes (from two size fractions, 0.1–0.8 and 0.8–3 μm, 6) were mapped onto OM-RGC, whilst the reads from 441 metatranscriptomes corresponding to the largest size fractions (0.8–5, 5–20, 20–180 and 180–2000 μm) were mapped onto MATOU. When OM-RGC is queried, abundance estimates may be expressed in one of two available normalization schemes: (i) the gene's read coverage is divided by the sum of the total gene coverages for the sample ('*percent of total genes per sample*'), or (ii) the gene's read coverage is divided by the median of the coverages of a set of 10 universal single copy marker genes ('*average copies per cell*') that were previously benchmarked for their suitability for metagenomics data analysis (17). MATOU gene abundance estimates are expressed as gene read coverage computed in RPKM (Reads Per Kilobase covered per Million of mapped reads) divided by the sum of the total gene coverages for the sample ('*percent of total genes per sample*').

### Environmental context

For *Tara* oceans biosamples, contextual environmental parameters are linked to the sequence datasets via barcodes assigned to each *Tara* Oceans sample (18). These metadata serve to geo-localize the samples and provide biogeochemical characteristics of the sampled seawater. Environmental parameters used by the OGA web service are obtained from PANGAEA (https://doi.org/10.1594/PANGAEA.875582), the open access library which archives and distributes georeferenced data from earth system research. For GOS biosamples, environmental data was extracted manually from Table 1 of Rusch *et al*. (6). The environmental parameters provided by OGA (Figure 3) are 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 2C). Descriptions of the environmental parameters available in OGA as well as corresponding PANGAEA hyperlinks are provided in the 'OGA user manual' hyperlinked from the OGA results page.

## IMPLEMENTATION

The Ocean Gene Atlas web server is implemented through a classical Model-View-Controller pattern architecture using the Laravel 5.4 PHP framework. Developed on the GNU/Linux, the application server communicates with the user through an Apache HTTP server using HTML5, CSS3, Javascript and AJAX to retrieve user requests and display results. The PHP application server queries abun-

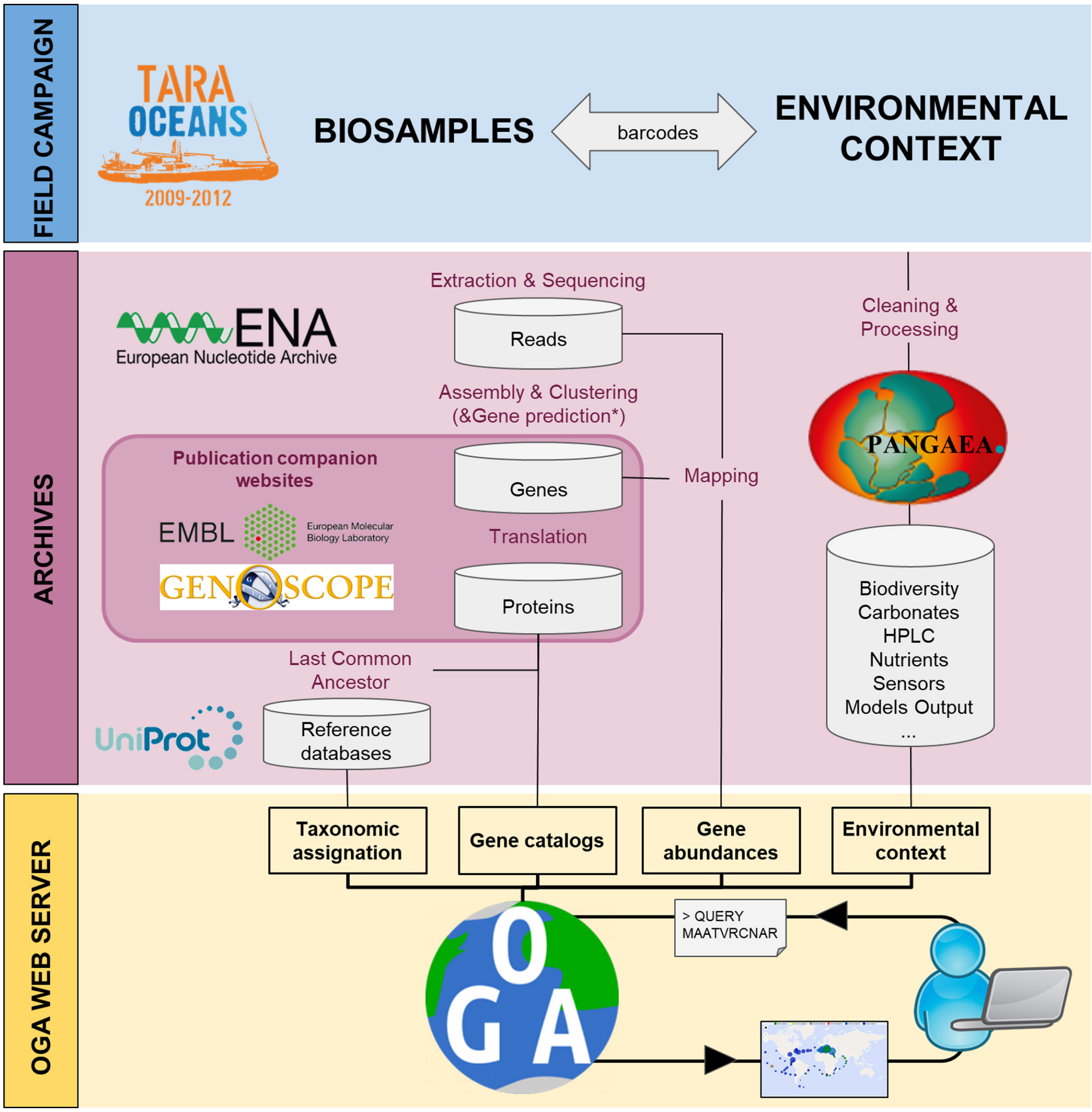


**Figure 3.** *Tara* oceans data sources for the the Ocean Gene Atlas workflow. Field campaigns (blue) have collected plankton biosamples and measured *in situ* environmental parameters. The OGA web server (yellow) combines heterogeneous data published by distinct archives (pink): EBI ENA for sequencing reads, published articles companion websites for gene catalogs and taxonomic annotations, PANGAEA for contextual environmental data. For GOS, metadata was manually extracted from table 1 of Rusch *et al.* (6).

dance and environmental data stored in a MySQL relational database.

## CASE STUDY

Upon phosphorus deficiency, bacterioplankton have established a widespread strategy of replacing membrane phospholipids with alternative non-phosphorus lipids. Sebastián *et al.* (19) 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 GOS and *Tara* Oceans have confirmed that PlcP is abundant in low phosphate concentrations areas. We reproduced the analysis of Sebastián *et al.* (19) using OGA to verify that the web service conforms to the published results. We used the same phospholipase C (EAQ46983) as a BLASTP query sequence with the author's e-value threshold of 1e$^{-40}$ to search for similar sequences in the OM-RGC catalog (the same metagenome dataset used by Sebastián *et al.*). The 952 PlcP hits identified showed higher abundances in Mediterranean subsurface samples (Figure 4A) related to low phosphorus concentration (Figure 4B) and mostly originated from Proteobacteria and Bacteroidetes (Figure 4C), 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.

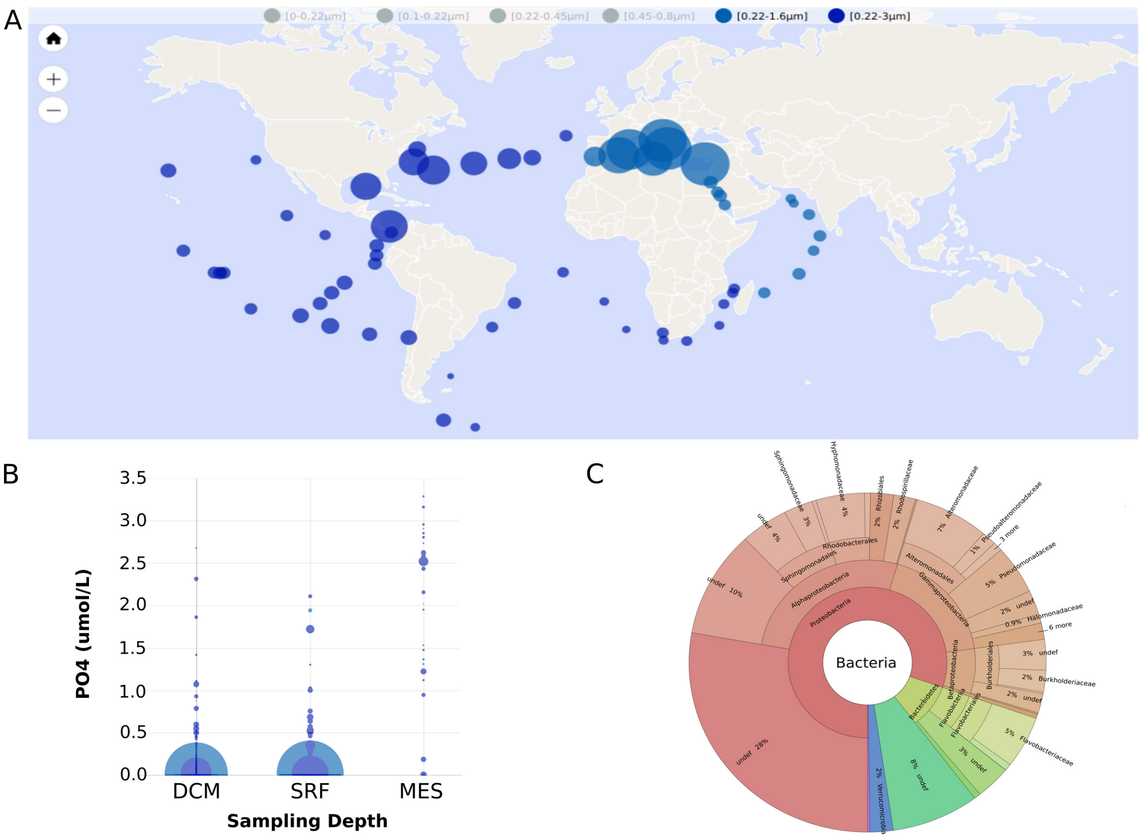


**Figure 4.** Phospholipase C (PlcP) biogeography produced by the Ocean Gene Atlas web service. (**A**) Abundance of PlcP hits in the OM-RGC subsurface samples. (**B**) Bubble plot of the PlcP abundance in relation to $PO_4$ concentrations; DCM: Deep Chlorophyll Maximum layer, SRF: subsurface and MES: mesopelagic zone. (**C**) Krona plot of the taxonomic distribution of the PlcP hits.

## CONCLUSION/PERSPECTIVES

By hiding the complex and time consuming integration of heterogeneous data sources behind a user-friendly minimalist web form, the Ocean Gene Atlas web server has the potential to broaden the access to the rapidly accumulating environmental marine genomics datasets. Enabling marine biologists to mine such data—without specific high performance hardware or programming skills—is one of the keys to extract knowledge and understanding from these valuable but underexploited resources. With the current first release of OGA, users can search by sequence similarity genes and proteins in two of the largest marine gene catalogs representing all three eukaryotic, prokaryotic and viral lineages.

These two first catalogs will be periodically updated as further marine environmental genomics databases are publically released. The prerequisites for inclusion in OGA are the availability of the three core resources: gene sequence catalogs, gene abundance estimates in biosamples, and geolocalized environmental context of biosamples. Leveraging the total 2100 *Tara* Oceans biosamples sequenced so far (20), our short term roadmap is to (i) extend the *Tara* Oceans datasets by including further sampling sites from the *Tara* Polar Circle expedition, (ii) complement the eukaryotic MATOU metatranscriptome catalog with corresponding metagenomic abundances and (iii) complement the OM-RGC with prokaryotic metatranscriptomes. In addition, we envisage adding the increasingly available pangenomes relevant to the marine biome, such as the large scale *Prochlorococcus* metapangenome (21). Such target OGA datasets would allow users with the powerful option to focus queries to their pangenome of interest, in order to track their species specific distribution across the world oceans.

We also plan to complement the current MATOU metatranscriptomes and OM-RGC metagenomes Ocean Gene Atlas results panels with corresponding barcoding based taxonomic profiles (e.g. Krona plots) available for 18S eukaryotic (22) and 16S prokaryotic rRNA genes (8,23). One limitation of the current OGA analyses is that pairwise sequence comparisons with uncultured fragmented microbial community genomes naturally impede fine grained taxonomic interpretations. Two alternatives may be considered to mitigate this shortcoming: (i) phylogenetic tree inference using the subset of full length environmental sequence hits aligned with known reference sequences and (ii) phylogenetic mapping on reference trees of all hits, including partial sequences. Both approaches are being actively pursued but compared to the well established and computationally efficient pairwise alignments, these still present conceptual and algorithmic challenges before being applicable to (semi) automatic online on-the-fly computation.

## DATA AVAILABILITY

Ocean Gene Atlas is a freely-available web service at: http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/ Source code is available at: https://github.com/hingamp/oga *S*hotgun sequences are available at the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena) under accession number PRJEB402 (OM-RGC), PRJEB6609 (MATOU) and PRJNA13694 (GOS). The predicted genes from the OM-RGC are available at ENA under the accession numbers ERZ096909 to ERZ097151, and the protein sequences are available at: ftp://ftp.genome.jp/pub/db/mgenes/Environmental/Tara.pep.gz. All MATOU resources are available at http://www.genoscope.cns.fr/tara/. Registry of all the samples from the *Tara* Oceans Expedition (2009–2013) are available at *PANGAEA*: https://doi.org/10.1594/PANGAEA.875582.

## ACKNOWLEDGEMENTS

The authors thank N. Terrapon and L. Tichit for the master students co-supervision. This article is contribution number 75 of the *Tara* Oceans expedition.

## FUNDING

French government 'Investissements d'Avenir' project OCEANOMICS [ANR-11-BTBR-0008 to P.H., T.V., C.V., M.L., E.P.]; 'Agence Nationale de la Recherche' project IMPEKAB [ANR-15-CE02-0011 to E.V.]; ETH Zurich and Helmut Horten Foundation [to S.S.]; European Regional Development Fund project FEDER [1166-39417]. Funding for open access charge: French government 'Investissements d'Avenir' project OCEANOMICS [ANR-11-BTBR-0008].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Field,Behrenfeld, Randerson and Falkowski (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.
2. Guidi,L., Chaffron,S., Bittner,L., Eveillard,D., Larhlimi,A., Roux,S., Darzi,Y., Audic,S., Berline,L., Brum,J. *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, **532**, 465–470.
3. Falkowski,P. (2012) Ocean Science: The power of plankton. *Nature*, **483**, S17–S20.
4. Kennedy,J., Marchesi,J.R. and Dobson,A.D. (2008) Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb. Cell Factories*, **7**, 27.
5. Mineta,K. and Gojobori,T. (2016) Databases of the marine metagenomics. *Gene*, **576**, 724–728.
6. 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. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.
7. Karsenti,E., Acinas,S.G., Bork,P., Bowler,C., De Vargas,C., Raes,J., Sullivan,M., Arendt,D., Benzoni,F., Claverie,J.-M. *et al.* (2011) A holistic approach to marine eco-systems biology. *PLoS Biol.*, **9**, e1001177.
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) Ocean plankton. Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.
9. Carradec,Q., Pelletier,E., Da Silva,C., Alberti,A., Seeleuthner,Y., Blanc-Mathieu,R., Lima-Mendez,G., Rocha,F., Tirichine,L., Labadie,K. *et al.* (2018) A global ocean atlas of eukaryotic genes. *Nat. Commun.*, **9**, 373.
10. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
11. Steinegger,M. and Soding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
12. Altschul,S.F., Madden,T.L., Schaffer,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.
13. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
14. Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.
15. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
16. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet. TIG*, **16**, 276–277.
17. Sunagawa,S., Mende,D.R., Zeller,G., Izquierdo-Carrasco,F., Berger,S.A., Kultima,J.R., Coelho,L.P., Arumugam,M., Tap,J., Nielsen,H.B. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods*, **10**, 1196–1199.
18. Pesant,S., Not,F., Picheral,M., Kandels-Lewis,S., Le Bescot,N., Gorsky,G., Iudicone,D., Karsenti,E., Speich,S., Trouble,R. *et al.* (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data*, **2**, 150023.
19. 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.
20. Alberti,A., Poulain,J., Engelen,S., Labadie,K., Romac,S., Ferrera,I., Albini,G., Aury,J.-M., Belser,C., Bertrand,A. *et al.* (2017) Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data*, **4**, 170093.
21. Delmont,T.O. and Eren,A.M. (2018) Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*, **6**, e4320.
22. de Vargas,C., Audic,S., Henry,N., Decelle,J., Mahé,F., Logares,R., Lara,E., Berney,C., Le Bescot,N., Probert,I. *et al.* (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.
23. Logares,R., Sunagawa,S., Salazar,G., Cornejo-Castillo,F.M., Ferrera,I., Sarmento,H., Hingamp,P., Ogata,H., de Vargas,C., Lima-Mendez,G. *et al.* (2014) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.*, **16**, 2659–2671.