nature REVIEWS MICROBIOLOGY

SUPPLEMENTARY INFORMATION

In format as provided by the authors

Tara Oceans: towards global ocean ecosystems biology

Shinichi Sunagawa , Silvia G. Acinas, Peer Bork, Chris Bowler, Tara Oceans Coordinators, Damien Eveillard, Gabriel Gorsky, Lionel Guidi, Daniele Iudicone, Eric Karsenti, Fabien Lombard, Hiroyuki Ogata, Stephane Pesant, Matthew B. Sullivan, Patrick Wincker and Colomban de Vargas

https://doi.org/10.1038/s41579-020-0364-5

Supplementary material

Tara Oceans: towards global ocean eco-systems biology

Shinichi Sunagawa[1,†], Silvia G. Acinas[2], Peer Bork[3,4,5], Chris Bowler[6,7], *Tara* Oceans Coordinators*, Damien Eveillard[7,8], Gabriel Gorsky[7,9], Lionel Guidi[7,9], Daniele Iudicone[10], Eric Karsenti[6,7,11], Fabien Lombard[7,9], Hiroyuki Ogata[12], Stephane Pesant[13,14], Matthew B. Sullivan[15,16,17], Patrick Wincker[7,18], Colomban de Vargas [7,19,†]

[1] Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich, Switzerland
[2] Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Barcelona, Spain
[3] Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany
[4] Max Delbrück Centre for Molecular Medicine, Berlin, Germany
[5] Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany
[6] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, Paris, France
[7] Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* GOSEE, Paris, France
[8] Université de Nantes, CNRS - UMR6004 - LS2N, Nantes, France
[9] Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, Villefranche-sur-Mer, France
[10] Stazione Zoologica Anton Dohrn, Naples, Italy
[11] Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany
[12] Institute for Chemical Research, Kyoto University, Kyoto, Japan
[13] PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany
[14] MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany
[15] Department of Microbiology, the Ohio State University, Columbus, OH, USA
[16] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA
[17] Center for RNA Biology, the Ohio State University, Columbus, OH, USA
[18] Génomique Métabolique, Genoscope, Institut de biologie Francois Jacob, Commissariat à l'Énergie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France
[19] Sorbonne Université & CNRS, UMR 7144 (AD2M), ECOMAP, Station Biologique de Roscoff, Roscoff, France
[†] Correspondence to S.S. ssunagawa@ethz.ch and C.d.V. vargas@sb-roscoff.fr

Supplementary BOX 1: Technologies and resources for ocean eco-systems biology

High Throughput Sequencing (HTS)

Tara Oceans sampled an extensive size range of plankton (10^{-2} - 10^{5} µm) at each sampling station (Figure 2) from several water layers to generate a variety of HTS data. In addition to the genomic potential, the goal was to assess the transcriptional activity of plankton and to contextualize the sequencing data by *in situ* environmental parameters. The study design included metaB sequencing to provide a baseline survey of the diversity and relative abundance of prokaryotic and eukaryotic taxa in their environmental context. In addition, metaG and metaT sequencing were performed (Figure 2) to uncover the genomic potential and expression of genes in viruses, prokaryotes, and eukaryotes[1,2], with the additional advantage of gaining insights into the population structure and evolution of, and selective pressure on, the most abundant planktonic organisms in the ocean.

To achieve these goals, molecular biology protocols were either newly developed or adapted from existing ones in order to limit technical biases and to ensure the comparability of the generated data. For example, new quantitative "low-input" methods for viral metaG analysis[3] and cDNA synthesis[4] were developed to extract both DNA and RNA with sufficiently high quantity and quality from several size-fractionated samples representing the whole diversity of plankton[5]. Furthermore, achieving high coverage of the eukaryotic genetic space by metaG sequencing is currently challenging, due to the relatively low coding density and presence of long, repetitive regions compared to prokaryotes. As an alternative, enrichment of poly-adenylated RNA allowed for efficient metaT sequencing of 441 eukaryote-enriched samples in order to reconstruct the content of global ocean eukaryotic genes[6]. The goal to "deeply" sequence each DNA and RNA sample was timely, as the costs of HTS and data storage/computing dropped dramatically over the course of the project. In total, the current sequencing effort has produced >60 terabases of data from more than 2,800 plankton samples (Supplementary Table 2).

Finally, studying genomic variation across populations benefits from data providing the highest possible phylogenetic/taxonomic resolution; however, despite targeted efforts[7], public resources of reference genome, transcriptome, or barcode sequences were insufficient or not available, in particular for eukaryotes, to analyse the vast amounts of sequencing data generated by *Tara* Oceans. Thus, additional efforts went into the sequencing of new draft genomes from single cells of protists and bacteria that were isolated by flow cytometry from cyropreserved seawater samples[8-10], as well as of individual zooplankton organisms[11]. Furthermore, new reference transcriptomes were generated from cultures of marine protists (Supplementary Table 2). Together with other efforts, such as the international Marine Microbial Eukaryote Transcriptome Sequencing Project[7], the sequencing data generated by *Tara* Oceans have led to a greatly expanded genomic representation of open ocean plankton.

It should be noted that *Tara* Oceans still maintains plankton samples for DNA/RNA sequencing at -80°C. This archive has been kept for work with future sequencing technologies, hopefully allowing deeper, longer and less-fractionated sequence reads from ultra-low-input nucleic acids extracts. Future work in *Tara* Oceans and *Tara*-related programmes will involve the use of long-read technologies to reconstruct more complete genome structures. While most analyses made at the gene level are not expected to change in their main conclusions, these improvements should provide better taxonomic assignments and insights into gene interaction networks at the organismal level, including for symbiotic interactions.

High Throughput Imaging (HTI)

Imaging data are highly complementary to HTS data, as they provide unique information on organismal concentrations, cellular biovolumes, and morphological attributes. However, they are rarely sampled together and analysed in conjunction. Imaging data can also be powerful for revealing how environmental and/or genomic variability impacts organismal or cellular phenotypes, and for verifying hypotheses based on 'omics' data about the nature of physical interactions between organisms.

In *Tara* Oceans, a series of automated imaging tools were adapted or newly developed to quantify the concentration, biomass, biovolume, taxonomic composition, and morphological features of plankton, as well as suspended and sinking particles across organismal size fractions. These tools included the Underwater Vision Profiler[12], ZooScan[13], FlowCAM[14], Imaging FlowCytobot[15], and e-HCFM, an environmental high-content fluorescence microscopy workflow developed during the project[16]. Together, these technologies allowed for automated imaging of organisms across different taxonomic and functional groups ranging in size from individual cells to large gelatinous zooplankton and marine snow particles (Figure 2). In addition, more classical imaging techniques, including light and confocal microscopy[16,17], as well as scanning[18] and transmission[19] electron microscopy, were used to generate smaller datasets at higher resolution. The set of imaging strategies was applied to >9,200 size-fractionated samples acquired from depths down to 1,000 m, with a current production of 6.8 Mio plankton images comprising >30 Tb of data (Supplementary Table 3).

HTI data have been instrumental for estimates of *in situ* cell numbers[17] and the biomass of giant protists[20], assessment of carbon fluxes from the surface down to 1,000 m during the entire expedition[21] and through the oxygen minimum zone of the Arabian Sea[22], demonstration of the dominance of non-tailed viruses in the global ocean[19], and estimates of prokaryotic cell abundances[23]. In conjunction with HTS data, the imaging data also provide new opportunities for integrative morpho-genetic analyses. This approach has been used to devise hypotheses and/or to validate organismal interactions[18,24-26], to associate carbon export with surface plankton interactomes[21], and to validate the feasibility of 3D automated fluorescence microscopy to generate quantitative data of taxonomic and morphological features, such as biovolumes, of microbial eukaryotes[16].

Bioinformatics

Tara Oceans has constantly faced the challenge of keeping pace with the rapidly evolving fields of HTS and HTI data production, processing, and analysis. Although state-of-the-art approaches and tools were used when available, other methods had to be improved or newly developed for the analysis of data from different organismal classes .

For viruses, an automated genome annotation pipeline (ViPTree) was set up to support viral sequence classification[27]. In addition, as viruses lack universal marker genes, new approaches had to be established in what eventually became the 'iVirus' ecosystem of applications[28]. Specifically, viral taxonomy had to be defined to establish 'species' level viral populations using average nucleotide identity of shared genes[29] and 'genus' level assignments via clustering in gene-sharing networks[30]. Furthermore, datasets and insights from *Tara* Oceans were leveraged to establish benchmarks for quantitative viromic analyses[31,32]. The taxonomic composition analysis of prokaryote-enriched metagenomes benefited from a method based on a set of protein-coding, universal single-copy marker genes, which was originally developed for analysis of human gut microbiome data[33,34]. Complementary to this approach, a new method was developed for identifying and quantifying 16S rRNA gene fragments in metaG data as an alternative to PCR-based metaB profiling[35]. Furthermore, to enable community-level integration of metaT and metaG data to derive gene expression levels, that is, relative transcript per gene copy numbers, and to quantify the composition and differences between metatranscriptomes, constitutively expressed microbial 'housekeeping genes' were identified[33] and their abundances required for the normalization of metaG and metaT data[2]. For the analysis of eukaryotic diversity using 18S rDNA metabarcodes, a novel open source amplicon clustering program was developed that produces fine-scale operational taxonomic units and scales linearly with increasing amounts of data[36]. Furthermore, due to the large genome sizes of eukaryotes compared to prokaryotes and viruses, the analysis of metaG and metaT data from larger plankton size fractions required the development of new bioinformatics workflows. Rather than first assembling and annotating metaG data to analyze metaT data, as conventionally done for prokaryotes and viruses, the analysis of eukaryotic data required to assemble and annotate metatranscriptomic data first before eukaryotic metaG data could be analysed[6]. Finally, to integrate data from heterogeneous sources, *Tara* Oceans benefited not only from established methods such as sparse partial least square analysis and weighted gene correlation network analysis to detect significant associations in the highly dimensional data sets[21,37,38], but also from the development of new methods for network analysis to determine species co-occurrence[25,37,39], and to capture and classify high-content confocal microscopy images[16].

The products of bioinformatic analyses in *Tara* Oceans have generated or enriched numerous databases and resources. For example, automatically reconstructed metabolic diagrams for individual microbial metaG have been deposited in KEGG/MGENES (https://www.genome.jp/mgenes/) to facilitate the exploration of gene functions in microbial ecosystems. For ocean viruses, sequence data for nearly 200,000 viral populations have been made available in the last few years[1,29,40]. These data are needed for virus-host databases[41] and the iVirus 'ecosystem' for viral ecology[28]. Similarly, for prokaryotes and eukaryotes, reference collections of environmental genes[6,23], select taxonomic marker genes[42-45], and single amplified genomes[8-10,46] were made available (Supplementary BOX 6). Across these organisms, co-occurrence network analyses were conducted to provide predictions of plankton interactions[25]. Finally, to promote the accessibility of these data, a download service of *Tara* Oceans registries, a web platform for environmental image archiving, automated recognition, and expert annotation (EcoTaxa), and a web service enabling the exploration of the abundance and location of ocean plankton genes in the context of *in situ* environmental features (Ocean Gene Atlas)[47] have been developed (Supplementary BOX 5). Despite the number of bioinformatics tools that have emerged, there is still a pressing need to develop interfaces and analytics for scientists to fully exploit eco-systems biology data generated by planetary-scale projects such as *Tara* Oceans.

1 Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* (2019).

2 Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083 e1021 (2019).

3 Solonenko, S. A. *et al.* Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**, 320 (2013).

4 Alberti, A. *et al.* Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15**, 912 (2014).

5 Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Sci Data* **4**, 170093 (2017).

6 Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat Commun* **9**, 373 (2018).

7 Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12**, e1001889 (2014).

8 Royo-Llonch, M. *et al.* Exploring Microdiversity in Novel Kordia sp. (Bacteroidetes) with Proteorhodopsin from the Tropical Indian Ocean via Single Amplified Genomes. *Front Microbiol* **8**, 1317 (2017).

9 Seeleuthner, Y. *et al.* Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat Commun* **9**, 310 (2018).

10 Sieracki, M. E. *et al.* Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci Rep* **9**, 6025 (2019).

11 Madoui, M. A. *et al.* New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Mol Ecol* **26**, 4467-4482 (2017).

12 Picheral, M. *et al.* The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnol Oceanogr-Meth* **8**, 462-473 (2010).

13 Gorsky, G. *et al.* Digital zooplankton image analysis using the ZooScan integrated system. *J Plankton Res* **32**, 285-303 (2010).

14 Sieracki, C. K., Sieracki, M. E. & Yentsch, C. S. An imaging-in-flow system for automated analysis of marine microplankton. *Mar Ecol Prog Ser* **168**, 285-296 (1998).

15 Olson, R. J. & Sosik, H. M. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnol Oceanogr-Meth* **5**, 195-203 (2007).

16 Colin, S. *et al.* Quantitative 3D-imaging for cell biology and ecology of environmental microbial eukaryotes. *Elife* **6** (2017).

17 Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **179**, 1084-1097 e1021 (2019).

18 Vincent, F. J. *et al.* The epibiotic life of the cosmopolitan diatom Fragilariopsis doliolus on heterotrophic ciliates in the open ocean. *ISME J* **12**, 1094-1108 (2018).

19 Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J* **7**, 1738-1751 (2013).

20 Biard, T. *et al.* In situ imaging reveals the biomass of giant protists in the global ocean. *Nature* **532**, 504-507 (2016).

21 Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465-470 (2016).

22 Roullier, F. *et al.* Particle size distribution and estimated carbon flux across the Arabian Sea oxygen minimum zone. *Biogeosciences* **11**, 4541-4557 (2014).

23 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

24 Cornejo-Castillo, F. M. *et al.* Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat Commun* **7**, 11071 (2016).

25 Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).

26 Mordret, S. *et al.* The symbiotic life of Symbiodinium in the open ocean within a new species of calcifying ciliate (Tiarina sp.). *ISME J* **10**, 1424-1436 (2016).

27 Nishimura, Y. *et al.* ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379-2380 (2017).

28 Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L. & Sullivan, M. B. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J* **11**, 7-14 (2017).

29 Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).

30 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632-639 (2019).

31 Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).

32 Roux, S. *et al.* Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777 (2016).

33 Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**, 1014 (2019).

34 Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**, 1196-1199 (2013).

35 Logares, R. *et al.* 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 (2014).

36 Mahe, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, e1420 (2015).

37 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**, 538-550 (2012).

38 Le Cao, K. A., Rossouw, D., Robert-Granie, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* **7**, Article 35 (2008).

39 Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* **8** (2017).

40 Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693 (2016).

41 Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66 (2016).

42 de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).

43 Decelle, J. *et al.* PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol Ecol Resour* **15**, 1435-1445 (2015).

44 Farrant, G. K. *et al.* Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc Natl Acad Sci U S A* **113**, E3365-3374 (2016).

45 Grebert, T. *et al.* Light color acclimation is a key process in the global ocean distribution of Synechococcus cyanobacteria. *Proc Natl Acad Sci U S A* **115**, E2010-E2019 (2018).

46 Mangot, J. F. *et al.* Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep-Uk* **7** (2017).

47 Villar, E. *et al.* The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res* **46**, W289-W295 (2018).

Supplementary BOX 2. *Tara* Oceans' sample and data resources

Tara Oceans provides a foundational knowledge base[1,2] that draws its success from:

- a consistent sampling strategy across the world oceans, describing 210 contrasting ecosystems in 20 biogeographic provinces;
- a coherent suite of protocols that effectively collected plankton across the entire size and taxonomy spectra of life, from viruses to metazoans;
- a unique and persistent identification of about 35,000 samples and a systematic record of their provenance; and
- an extensive set of environmental contexts obtained from concurrent field measurements, satellite products, climatology and gazetteers.

The quality of samples collected during the *Tara* Oceans expedition has enabled state-of-the-art imaging and sequencing, the vast potential of which still remains to be unveiled. Data are progressively released in open access repositories and already represent the largest coherent set of environmental, imaging and sequencing data collected in the global ocean. Yet, only a fraction of the available samples has been analysed, while replicates and aliquots are preserved in trusted biobanks for future generations (see Supplementary BOX 3).

Tara Oceans fosters a stable infrastructure of open access archives for sequences (ENA), high throughput images (EuBI) and environmental data (PANGAEA), ensuring that its knowledge base remains FAIR[3] (Findable, Accessible, Interoperable, and Reusable). Sample provenance is key to connecting environmental, sequencing and imaging data, and *Tara* Oceans set the trend in marine science with the most comprehensive registries of sample provenance and context to date (see Supplementary BOX 4). Web services are in place to use these registries and navigate the infrastructure.

The sheer quantity of imaging and sequencing data, and the richness of their environmental context require innovative bioinformatics tools, next generation global ocean modelling, and cloud computing. A number of bioinformatics and modelling tools have started to spur in the framework of *Tara* Oceans (see Supplementary BOX 5) and several comprehensive processed data products are beginning to emerge (see Supplementary BOX 6). However, there is a pressing need to develop more comprehensive tools and interfaces that enable scientists to explore and exploit the full potential of *Tara* Oceans' knowledge base, in particular to further our understanding of interactions, adaptation, and evolution of life.

[1] Pesant, S. *et al.* Open science resources for the discovery and analysis of *Tara* Oceans data. *Sci Data* **2**, 150023 doi: 10.1038/sdata.2015.23 (2015)

[2] Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Sci Data* **4**, 170093, doi:10.1038/sdata.2017.93 (2017)

[3] Wilkinson, M. D *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 doi: 10.1038/sdata.2016.18 (2016)

Supplementary BOX 3: Collection of *Tara* Oceans plankton samples for molecular (sequencing) (1 to 3), and imaging (4 to 6) analyses.

1. Samples of viruses, consisting in original filters and extracted nucleotides: The Sullivan lab, The Ohio State University, Columbus, OH, USA.
2. Samples of prokaryotes and eukaryotes, consisting in original filters and extracted nucleic acids: Genoscope, National Sequencing Centre, Paris, France.
3. Samples of prokaryotes and eukaryotes collected in 2009-2012 for Single-Amplified Genomes: *Cryopreserved Microbial Culture Collection*, Bigelow Single Cell Genomics Center, Bigelow Laboratory for Oceans Science, Maine, USA.
4. Samples of prokaryotes and eukaryotes collected in 2013 for Single-Amplified Genomes: Institute of Marine Sciences (ICM-CSIC), Barcelona, Spain.
5. Samples of eukaryotes (0.8 to 2,000 µm) collected at discrete depths: Station Biologique de Roscoff (SBR), Roscoff, France.
6. Samples of eukaryotes collected from vertical deployments of nets: Centre de Collection du Zooplancton J-C Pomerol, Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), France.

Supplementary BOX 4: *Tara* Oceans data and metadata

Sample provenance (meta)data

- Campaigns registry:
  https://dx.doi.org/10.1594/PANGAEA.842191
- Stations registry:
  https://dx.doi.org/10.1594/PANGAEA.842237
- Events registry:
  https://dx.doi.org/10.1594/PANGAEA.842227
- Samples registry:
  https://doi.pangaea.de/10.1594/PANGAEA.875582

Methodological context (meta)data

- Methodology used on board to prepare samples:
  https://doi.pangaea.de/10.1594/PANGAEA.875580
- Methodology used in the lab for sequencing analyses:
  https://doi.pangaea.de/10.1594/PANGAEA.875581
- Methodology used in the lab for imaging analyses: in prep

Environmental context (meta)data

- Meso-scale features:
  https://doi.pangaea.de/10.1594/PANGAEA.875577
- Water column features:
  https://dx.doi.org/10.1594/PANGAEA.858207
- Sensor data at the discrete depth of each sample:
  https://doi.pangaea.de/10.1594/PANGAEA.875576
- Carbonate chemistry at the discrete depth of each sample:
  https://doi.pangaea.de/10.1594/PANGAEA.875567
- Nutrients at the discrete depth of each sample:
  https://doi.pangaea.de/10.1594/PANGAEA.875575
- Pigments concentration at the discrete depth of each sample:
  https://doi.pangaea.de/10.1594/PANGAEA.875569

Nucleotides (sequencing) data

- Barcoding and shotgun sequencing:
  https://www.ebi.ac.uk/ena/data/view/PRJEB402

Supplementary BOX 5: *Tara* Oceans bioinformatics resources and tools

- The Ocean Gene Atlas[1]. Webserver to explore the abundance, biogeography, and environmental correlates of genes in the global ocean: http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas
- ViPTree[2]. Bioinformatics tool for enhancing viral proteomic tree analyses of users' viral genomes together with public viral genome data: http://www.genome.jp/viptree/
- R package for miTAGs[3]-based taxonomic profiling of metagenomes: in prep.
- Genome-scale ecosystem modelling[4]. Code for modelling multiple organism interactions via their metabolic networks: https://gitlab.univ-nantes.fr/mbudinich/MultiObjective-FBA-FVA
- EcoTaxa. Web application dedicated to the visual exploration and automated/expert taxonomic annotation of plankton images: http://ecotaxa.obs-vlfr.fr/
- MGnify[5]. Resource for browsing automated analyses of *Tara* Oceans metagenomic and metatranscriptomic data: https://www.ebi.ac.uk/metagenomics/search?query=TARA+OCEANS
- PANGAEA data download service. Service to query tabular data sets in *Tara* Oceans registries: https://ws.pangaea.de/dds-fdp

1. Villar, E. et al. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res* **46**, W289-W295 (2018).
2. Nishimura, Y. et al. Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere* **2** (2017).
3. Logares, R. et al. 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 (2014).
4. Budinich M et al. A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLoS ONE* **12**: e0171744 (2017).
5. Mitchell, A. L. et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res* **46**, D726-D735 (2018).

Supplementary BOX 6 - Processed data products generated in the framework of *Tara* Oceans

- Ocean Microbial Reference Gene Catalog (version 1 and 2). Genomic inventories of ocean microbial genes. Version 1 includes 40 million non-redundant sequences from 243 samples (Sunagawa et al., Science, 2015): https://www.ebi.ac.uk/ena/about/tara-oceans-assemblies, http://ocean-microbiome.embl.de/companion.html; version 2 includes 47 million non-redundant sequences form 370 samples (Salazar et al., Cell, 2019): https://www.ebi.ac.uk/biostudies/studies/S-BSST297
- Global ocean atlas of eukaryotic genes. Contains >115 million non- redundant eukaryotic genes obtained from a metaT sequencing effort on >400 samples (Carradec et al., Nat Comm, 2018): http://www.genoscope.cns.fr/tara/
- Global Ocean Virome datasets. Systematically collected and processed quantitative viral metaG datasets for the global oceans, predominantly composed of *Tara* Oceans data (Roux et al., Nature, 2016; Brum et al. Science 2015; Gregory et al., Cell, 2019): http://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/GOV

- Collection of annotated V9 rDNA metaB. Contains >100,000 taxonomically curated eukaryotic plankton OTUs from the sunlit ocean and associated analytical tools (de Vargas et al., Science, 2015): http://taraoceans.sb-roscoff.fr/EukDiv; organized at the OTU level: http://dx.doi.org/10.1594/PANGAEA.873275; organized at the metaB level: http://dx.doi.org/10.1594/PANGAEA.873277
- Collection of marker gene sequences to infer ecologically significant taxonomic units of picocyanobacterial diversity (Farrant et al., PNAS, 2016; Grébert et al, PNAS 2018).
- Global trans-kingdom interactome dataset. Data resource to study marine organismal interactions across all domains of life (Lima-Mendez et al., Science, 2015): http://www.raeslab.org/companion/ocean-interactome.html
- Virus-Host DB. A record of >8,000 viral genomes (Mihara et al., Viruses, 2016). Database linking viral genomes with taxonomic information of their hosts: http://www.genome.jp/virushostdb/
- KEGG/MGENES. A publicly accessible database of metaG/metaT genes including *Tara* Oceans data with pre-calculated taxonomic, functional, and metabolic pathway assignment: http://www.genome.jp/mgenes/

Supplementary Table 1. Selection of environmental parameters used widely in recent *Tara* Oceans publications. These include parameters measured *in situ* with sensors or from water samples, and additional parameters calculated from in situ measurements, climatologies and remote sensing products. The full suite of environmental parameters for the *Tara* Oceans expedition (2009-2013) is available at PANGAEA (https://doi.pangaea.de/10.1594/PANGAEA.875582).

Parameter name	**Units**	**Method**
depth	m	measured[a]
conductivity	mS/cm	measured[a]
salinity	psu	measured[a]
temperature	°C	measured[a]
density, sigma-theta	kg/m^3	measured[a]
oxygen	$\mu mol/m^3$	measured[a,b]
nitrite	$\mu mol/m^3$	measured[c,d]
nitrate and nitrite	$\mu mol/m^3$	measured[d]
phosphate	$\mu mol/m^3$	measured[d]
silicate	$\mu mol/m^3$	measured[d]
chlorophyll *a*	mg/m^3	measured[a,c]
concentration of 23 pigments (HPLC analysis)	mg/m^3	measured[d]
pH	Total scale	measured[d]
total alkalinity	µmol/kg	measured[d]
dissolved inorganic carbon	µmol/kg	calculated[s]
carbon dioxide	µmol/kg	calculated[s]
carbonate ion	µmol/kg	calculated[s]
bicarbonate ion	µmol/kg	calculated[s]
partial pressure of carbon dioxide	µatm	calculated[s]
fugacity of carbon dioxide	µatm	calculated[s]
aragonite saturation state		calculated[s]
calcite saturation state		calculated[s]
fluorescence (fCDOM)	ppb	measured[a,o]
optical beam attenuation coefficient, 660 nm	1/m	measured[a]
beam attenuation coefficient of particles	1/m	measured[a]
radiation, photosynthetically active per day	$E/m^2/day$	measured[a,e,f]
diffuse coefficient of attenuation of PAR	1/m	calculated[g,h]
depth of Secchi Disk	m	measured[i]
depth of euphotic zone	m	calculated[j,k,l]
depth of the mixed layer	m	calculated[m]
depth of chlorophyll maximum	m	calculated[m]
depth of maximum Brunt Väisälä frequency	m	calculated[m]
depth of minimum oxygen concentration	m	calculated[m]
depth of nitracline	m	calculated[m]
chlorophyll *a*, areal concentration	mg $Chla/m^2$	calculated[n]
radiation, photosynthetically active per day	$E/m^2/day$	calculated[o,p,q]
net primary production of carbon per area	mg $C/m^2/day$	calculated[r]

[a] using in situ sensor data calibrated with factory settings.

[b] using in situ sensor data calibrated with the WOA09 climatology.

[c] using in situ sensor data calibrated with measurements on water sample.

[d] using measurements on water samples.

[e] and calculated from Kd(PAR)1 using eq. 9 in Morel et al. (2007; http://dx.doi.org/10.1016/j.rse.2007.03.012) and 1-day avg daily surface PAR (AMODIS).

[f] and calculated from Kd(PAR)1 using eq. 9 in Morel et al. (2007; http://dx.doi.org/10.1016/j.rse.2007.03.012) and 8-day avg daily surface PAR (AMODIS).

[g] Kd(PAR)1 is calculated for a layer, which thickness is equal to [Kd(490)]−1, using Kd490 (AMODIS; 4 km resolution; 30-day average) and eq. 9 in Morel et al. (2007; http://dx.doi.org/10.1016/j.rse.2007.03.012).

[h] Kd(PAR)1 is calculated for a layer, which thickness is equal to 2 [Kd(490)]−1, using Kd490 (AMODIS; 4 km resolution ; 30-day average) and eq. 9prime in Morel et al. (2007; http://dx.doi.org/10.1016/j.rse.2007.03.012).

[i] measured only in 2009-2012 using a 20 cm diameter white Secchi disk

[j] zeu(0.415) is the depth of the 0.415 mol quanta m^-2 day^-1 light level. Zeu(1%) was calculated from Secchi depth (equation 18 in Morel et al. 2007; http://dx.doi.org/10.1016/j.rse.2007.03.012) and converted to zeu(0.415) using daily surface PAR (AMODIS; 4 km resolution; 30-day average) and equations in Boss and Berhenfeld (2010; http://dx.doi.org/10.1029/2010GL044174).

[k] zeu(0.415) is the depth of the 0.415 mol quanta m^-2 day^-1 light level. Zeu(1%) was calculated from Kd(PAR)1 (equation 9 in Morel et al. 2007; http://dx.doi.org/10.1016/j.rse.2007.03.012) and converted to zeu(0.415) using daily surface PAR (AMODIS; 4 km resolution; 30-day average) and equations in Boss and Berhenfeld (2010; http://dx.doi.org/10.1029/2010GL044174).

[l] zeu(0.415) is the depth of the 0.415 mol quanta m^-2 day^-1 light level. Zeu(1%) was calculated from Kd(PAR)2 (equation 9prime in Morel et al. 2007; http://dx.doi.org/10.1016/j.rse.2007.03.012) and converted to zeu(0.415) using daily surface PAR (AMODIS; 4 km resolution; 30-day average) and equations in Boss and Berhenfeld (2010; http://dx.doi.org/10.1029/2010GL044174).

[m] calculated from in situ sensor data calibrated with factory settings.

[n] calculated by vertical integration of fluorescence profiles from surface to 200 m or bottom, using in situ sensor data calibrated with factory settings.

[o] calculated from AMODIS products for the sampling location (4 km resolution) and exact date.

[p] calculated from AMODIS products for the sampling location (4 km resolution) and averaged for a period of 8 days around the sampling date.

[q] calculated from AMODIS products for the sampling location (4 km resolution) and averaged for a period of 30 days around the sampling date.

[r] calculated from VGPM products for the sampling location (4 km resolution) and exact date.

[s] calculated from pH and total alkalinity measurements, using seacarb (Nisumaa et al. 2010)

Supplementary Table 2. Summary of *Tara* Oceans' current HTS datasets, which were generated from size-fractionated plankton samples, and cryopreserved samples for single-cell sequencing, collected across the global ocean (Figure 2). Data are available at the European Nucleotide Archive under the project identifier PRJEB402. MetaB: metabarcoding; metaG: metagenomics; metaT: metatranscriptomics; SAG: single amplified genomes; giruses: giant DNA viruses.

Data type	Target group	Samples	Reads* (G)	Bases (Gb)	ENA project ID(s)	Main references
metaB	Protists and metazoa	1029	2.84	788	PRJEB6610; PRJEB9737	De Vargas 2015, Ibarbalz et al., 2019
metaG	Viruses	133	19.9	3912	PRJEB4419; PRJEB9742	Brum 2015, Roux 2016, Sunagawa 2015, Gregory 2019
metaG	Prokaryotes and giruses	252	41.1	8161	PRJEB1787; PRJEB9740	Sunagawa 2015
metaG	Protists and metazoa	674	130.20	25569	PRJEB4352	Carradec 2018
metaT	Protists and metazoa	457	87.06	16412	PRJEB6609	Carradec 2018
metaT	Prokaryotes	196	26.5	5264	PRJEB6608; PRJEB9741	Salazar et al., 2019
SAG	Protists	55	1.69	310	PRJEB4422; ERA768231; PRJEB6603	Seeleuthner et al., 2018; Vannier et al., 2016; Sieracki et al., 2019
SAG	Bacteria	1	0.11	0.06	PRJNA524487	Royo-LLonch, submitted
Reference transcriptomes	Protists	5	2.1		PRJEB21821	Carradec 2018
Reference genome	Zooplankton *Oithona nana*	1		0.085	PRJEB18938	Madoui et al., Mol Ecol 2017

*mainly paired-end Illumina reads

Supplementary Table 3. Summary of *Tara* Oceans' current imaging datasets, including ~6.8 million images extracted from >9,200 worldwide size-fractionated plankton communities (samples). TO: *Tara* Oceans; TOPC: *Tara* Oceans Polar Circle; IFCB: Imaging Flow CytoBot; e-HCFM: environmental High Content Fluorescence Microscopy; CLSM: Confocal Laser Scanning Microscopy; SEM: Scanning Electron Microscopy; TEM: Transmission Electron Microscopy. All datasets have been released in EcoTaxa (https://ecotaxa.obs-vlfr.fr): an online platform for automated and experts-based environmental images classification.

Imaging dataset **(Instrument & plankton sampling gear & size mesh)**	**Plankton size range**	**Plankton samples analysed**	**Number of objects**
UVP	> 0.7 mm	776	776497
ZooScan Regent net	> 0.68 mm	212	126389
ZooScan WP2 net	> 0.2 mm	203	394965
ZooScan Multinet net	> 0.3 mm	285	393382
ZooScan Bongo net	> 0.3 mm	92	283596
ZooScan Bongo net	> 300 µm	19	42365
FlowCam Bongo net & Niskin bottles	200 > 20 µm	317	744409
IFCB inline	160 > 5 µm	6982	2461263
SEM	200 > 0.1 µm	31	2469
e-HCFM H5 5-20µm	20 > 5 µm	76	336655
e-HCFM H20 20-180µm	180 > 20 µm	108	1235640
e-HCFM H0.2>0.2	> 0.2 µm	14	-
HiRes 3D-CLSM	2000 > 1 µm	65	3551
TEM virus	< 0.1 µm	43	4300
TOTAL		**9,223**	**6.8 million**