Open science resources for the discovery and analysis of Tara Oceans data

Stéphane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Sabrina Speich, Romain Troublé, et al.

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Biodiversity

» Ecological genetics

» Ocean sciences

» Biooceanography

Received: 03 March 2015

Accepted: 27 April 2015

Published: 26 May 2015

Open science resources for the discovery and analysis of *Tara* Oceans data

Stéphane Pesant[1,2], Fabrice Not[3,4], Marc Picheral[5,6], Stefanie Kandels-Lewis[7,8], Noan Le Bescot[3], Gabriel Gorsky[5,6], Daniele Iudicone[9], Eric Karsenti[8,10], Sabrina Speich[11,12], Romain Troublé[13], Céline Dimier[3], Sarah Searson[4,14] & *Tara* Oceans Consortium Coordinators*

The *Tara* Oceans expedition (2009–2013) sampled contrasting ecosystems of the world oceans, collecting environmental data and plankton, from viruses to metazoans, for later analysis using modern sequencing and state-of-the-art imaging technologies. It surveyed 210 ecosystems in 20 biogeographic provinces, collecting over 35,000 samples of seawater and plankton. The interpretation of such an extensive collection of samples in their ecological context requires means to explore, assess and access raw and validated data sets. To address this challenge, the *Tara* Oceans Consortium offers open science resources, including the use of open access archives for nucleotides (ENA) and for environmental, biogeochemical, taxonomic and morphological data (PANGAEA), and the development of on line discovery tools and collaborative annotation tools for sequences and images. Here, we present an overview of *Tara* Oceans Data, and we provide detailed registries (data sets) of all campaigns (from port-to-port), stations and sampling events.

Design Type(s)	observation design • global survey
Measurement Type(s)	Registry
Technology Type(s)	Written Documentation
Factor Type(s)	
Sample Characteristic(s)	marine biome • Bay of Biscay • Strait of Gibraltar • Mediterranean Sea • Mediterranean Sea, Western Basin • Ligurian Sea • Tyrrhenian Sea • Ionian Sea • Adriatic Sea • Mediterranean Sea, Eastern Basin • Red Sea • Arabian Sea • Indian Ocean • Mozambique Channel • Southeast Atlantic Ocean • South Atlantic Ocean • Southwest Atlantic Ocean • Drake Passage • South Pacific Ocean • Equatorial Pacific Ocean • North East Pacific Ocean • Gulf of Mexico • Florida Straits • NW Atlantic Ocean • North Atlantic Ocean • NE Atlantic Ocean • Norwegian Sea • Kara Sea • Arctic Ocean • Beaufort Sea • Northwest Passages • Baffin Bay

[1]PANGAEA, Data Publisher for Earth and Environmental Science, 28359 Bremen, Germany. [2]MARUM, Center for Marine Environmental Sciences, Universität Bremen, 28359 Bremen, Germany. [3]CNRS, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France. [4]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France. [5]CNRS, UMR 7093, Observatoire Océanologique de Villefranche-sur-Mer (OOV), 06230 Villefranche/mer, France. [6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, Observatoire Océanologique de Villefranche-sur-Mer (OOV), 06230, Villefranche/mer, France. [7]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. [8]Directors' Research, European Molecular Biology Laboratory Meyerhofstrasse 1, 69117 Heidelberg, Germany. [9]Laboratory of Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, 80121 Naples, Italy. [10]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, CNRS, UMR 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, 75005 Paris, France. [11]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 75005, Paris, France. [12]Laboratoire de Physique des Océan, UBO-IUEM, 29280 Polouzané, France. [13]Tara Expéditions, Base Tara, 11 boulevard Bourdon, 75004 Paris, France. [14]Department of Oceanography, University of Hawaii at Manoa, Honolulu, HI 96822, USA. *see Author Contributions Section. Correspondence and requests for materials should be addressed to S.P. (email: spesant@marum.de).

Background & Summary

Over many centuries, global expeditions have led to major scientific breakthroughs, notably with the early voyages of the H.M.S. Beagle (1831–1836) and the H.M.S. Challenger (1872–1876). Ocean exploration now provides promising first steps towards understanding the role of the ocean in global biogeochemical cycles and the impact of global climate change on ocean processes and marine biodiversity. Recently, the *Sorcerer II* expeditions (2003–2010)[1] and the Malaspina expedition (2010–2011)[2] carried out global surveys of prokaryotic metagenomes from the ocean's surface and bathypelagic layer (>1,000 m), respectively. The *Tara* Oceans Expedition (2009–2013) complemented these surveys by collecting a wide variety of planktonic organisms (from viruses to fish larvae) from the ocean's surface (0–200 m) and mesopelagic zone (200–1,000 m) at a global scale. Overall, *Tara* Oceans surveyed 210 ecosystems in 20 biogeographic provinces, collecting over 35,000 samples of seawater and plankton. Organising such a knowledge base is essential to safeguard, discover and share *Tara* Oceans data. To address this challenge, *Tara* Oceans offers open science resources, including the use of open access data archives and the development of online tools for the collaborative annotation of sequences and images, and the discovery of *Tara* Oceans data.

Tara Oceans adopts the principle of open access and early release of raw and validated data sets. In the case of molecular data, raw short sequence reads are archived at the European Bioinformatics Institute *short read archive* (http://www.ebi.ac.uk/ena/) and made available immediately after manual curation of metadata. More advanced data (assemblies, annotations, etc.) will be released immediately after validation and before publication, and other versions will be released when available. In the case of environmental, biogeochemical, taxonomic and morphological measurements, data are published at PANGAEA, Data Publisher for Earth and Environmental Science (http://www.pangaea.de) and made available immediately after manual curation of metadata.

By combining modern sequencing and state-of-the-art imaging technologies, *Tara* Oceans is at the cutting edge of marine science[3]. The amount of data generated by these technologies is unprecedented in the field of plankton ecology and requires adapted storage infrastructures and collaborative platforms to carry out manual and automated annotation of sequences and high throughput images. These open science resources are currently being developed by *Tara* Oceans.

A first series of publications has demonstrated the potential of *Tara* Oceans data to study the ecology of plankton and the structural and functional diversity of viruses, prokaryotes and eukaryotes in the global ocean[4–11]. These publications are based on a fraction of the samples analysed so far and thus represent only the tip of the iceberg. The exploration of *Tara* Oceans data by the scientific community will undoubtedly lead to new hypotheses and emerging concepts in domains unforeseen by the *Tara* Oceans Consortium. The current discovery portal of *Tara* Oceans offers a simple map interface that links each sampling location to available environmental and molecular data (http://www.taraoceans-dataportal.org/). It will however evolve to offer advanced search functionalities based on geospatial, methodological, environmental, morphological, taxonomic, phylogenetic and ecological criteria.

Here, we present an overview of the sampling strategy and size-fractionation approach of the *Tara* Oceans Expedition (***Methods Section***) and we explain the rationale behind the choice of sampling devices (***Technical Validation Section***). Most importantly, we provide registries (data sets) describing all campaigns (from port-to-port), stations and sampling events (***Data Records Section***). These registries contain geospatial, temporal and methodological information that will be essential for researchers to explore and assess the quality of *Tara* Oceans data. Environmental data sets are already available openly, in whole or in part, and additional data sets will be progressively released to the community. We intend to submit additional publications describing specific data types (e.g., Data Citations 1–5) in more detail, further extending the value of this resource as the data becomes available.

Methods

As a research infrastructure, the *Tara* Oceans Expedition mobilised over 100 scientists to sample the world oceans on board a 36 m long schooner (SV *Tara*) refitted to operate state-of-the-art oceanographic equipment (Fig. 1). On board the schooner, the team was consistently composed of five sailors and six scientists, including one chief scientist, two oceanography engineers in charge of deck operations, instrument maintenance and data management, two biology engineers preparing and preserving samples for later morphological and genetic analyses, and one optics engineer in charge of imaging live samples on board. A winch equipped with 2,400 m of cable was installed to deploy sampling devices from the stern of the ship, and an industrial peristaltic pump was installed on starboard to sample large volumes of water from various depths down to 60 m. Peristaltic and vacuum filtration systems used to concentrate plankton on membranes of various pore sizes were setup in a laboratory container (wet lab) located outside on port side. Flow-through instruments connected to the continuous surface sampling system were installed in the fore peak and in a laboratory (dry lab) inside the schooner at the centre of the ship on port side.

The sampling strategy and methodology of the *Tara* Oceans Expedition is presented in six sub-sections. The first four describe why and how the environmental context was determined [**1**] at the mesoscale using remote sensing and meteorological data; [**2**] from sensors mounted on the continuous surface water sampling system; [**3**] from sensors mounted on the vertical profile sampling system; and [**4**] from discrete water samplers (Niskin bottles) mounted on the vertical profile sampling system. The last

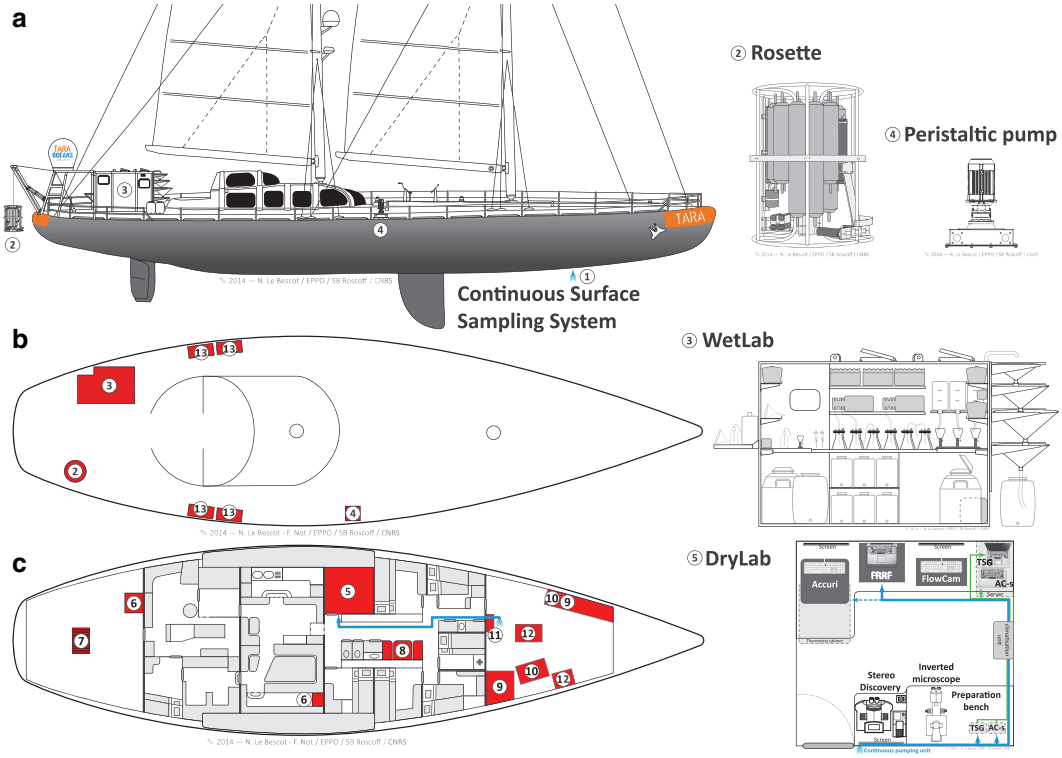

Figure 1. Sampling devices and working areas on-board SV *Tara*. Sampling devices and working areas on-board SV *Tara* are shown from the vessel's **[a]** side-view, **[b]** bird's-eye-view of the deck, and **[c]** inside-view. They consist of the **[1]** Continuous Surface Sampling System [CSSS]; **[2]** Rosette Vertical Sampling System [RVSS]; **[3]** wet lab and storage in liquid nitrogen; **[4]** High Volume Peristaltic pump [HVP-PUMP]; **[5]** dry lab; **[6]** oceanography engineers data acquisition and processing area; **[7]** winch; **[8]** video imaging area; **[9]** storage areas at room temperature; **[10]** storage areas at +4 °C and −20 °C; **[11]** MilliQ water system and AC-s system; **[12]** diving equipment, flowcytobot and ALPHA instruments; and **[13]** storage boxes. The flow of seawater from the continuous surface sampling system to the dry lab is shown in blue.

two Sub-Sections describe how **[5]** environmental features were selected and sampled; and how **[6]** plankton were collected for imaging and genetic analyses. These methods were also described briefly in Karsenti *et al.* (2011)[3].

[1] Atmospheric and oceanographic context at the mesoscale

The regular sampling programme was designed to study a variety of marine ecosystems and to target well-defined meso- to large-scale features such as gyres, eddies, currents, frontal zones, upwellings, hot spots of biodiversity, low pH or low oxygen concentrations. A total of 210 stations were characterised at the mesoscale to provide richer environmental context for the morphological and genomic study of plankton (Fig. 2). In order to identify these features before sampling but also to assess *a posteriori* if sampling events carried out during a station were taken within a relatively homogeneous environment, the atmospheric and oceanographic context were determined at the mesoscale, using climatologies, remote sensing products and arrays of Argo profiling floats. Meteorological forecast services, satellite observations (Chlorophyll *a*, sea surface temperature (SST) and altimetry) and real-time ocean model outputs (Mercator Ocean) were also used on a daily basis to revise sampling positions with respect to the selected oceanographic features.

Mapped altimetry from AVISO (Archiving Validation and Interpretation of Satellite Data in Oceanography), mapped operational SST (OSTIA), and satellite ocean colour (ACRI-ST GlobColour service) were used to describe the spatial and temporal variability of key environmental parameters at each sampling station. In addition, Temperature-Salinity profiles available around sampling stations were compiled from the Argo autonomous network array. Finally, a [BATOS] meteorological station mounted on-board *Tara* continuously measured wind speed and direction, and air temperature, pressure and humidity, which helped determine the variability of atmospheric conditions and vertical mixing of surface waters.

In addition to the regular sampling programme, topical experiments were designed to study ocean processes that operate at spatial and/or temporal scales larger or smaller than the mesoscale (Fig. 3).

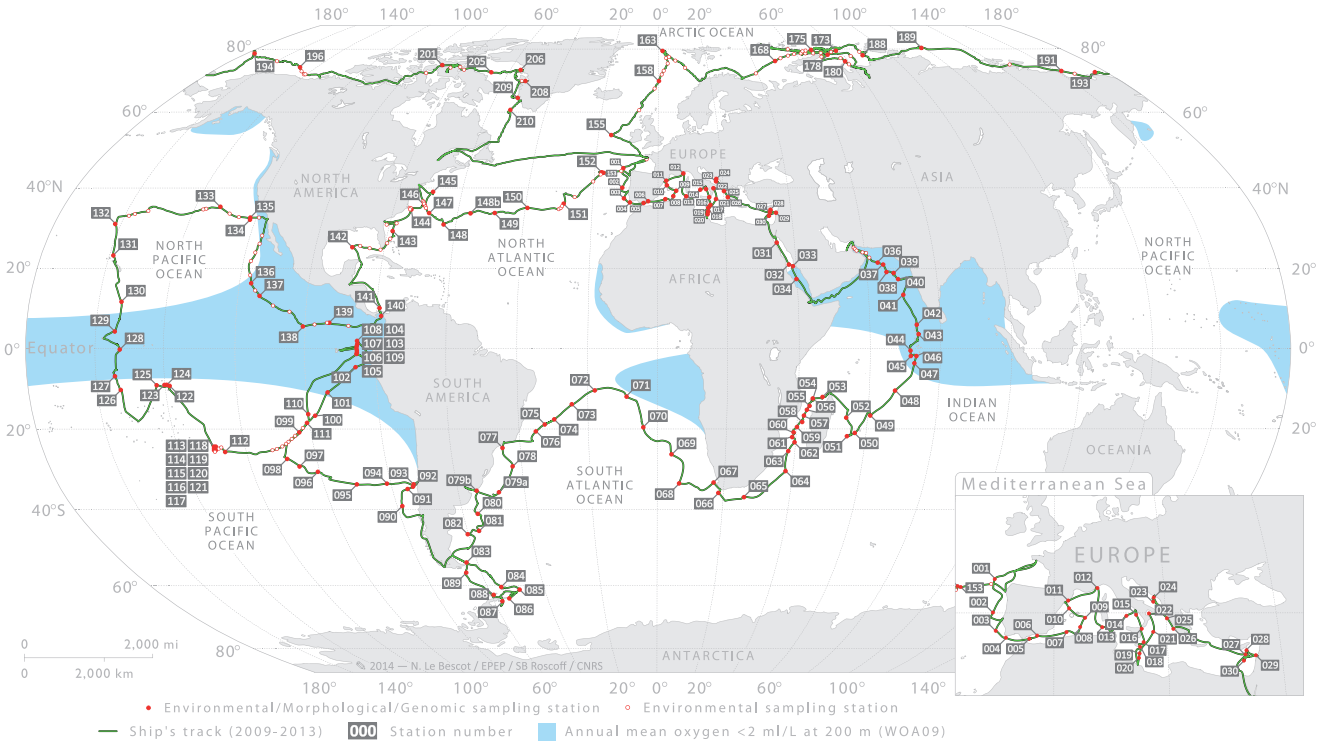

Figure 2. Sampling route and stations of the *Tara* Oceans Expedition. Sampling route of the *Tara* Oceans Expedition (green track), showing station labels and areas (blue shade) where the annual mean oxygen concentration is < 2 ml/l (WOA09), usually corresponding also to high CO_2 concentration and low pH. Detailed information about each station is given in (Data Citation 7).

Topics included, for example, diurnal processes, storm-induced perturbation of community structure and functions, latitudinal diversity gradients, oxygen minimum zone[6], island effects on iron fertilization, and longitudinal transport by Agulhas rings across the South Atlantic Ocean[11]. For topical experiments, oceanographic context was sometimes enriched by using automated underwater vehicles (e.g., gliders[12], ProvBio[13]), surface-tethered Argo drifters, lowered ADCP mounted on the rosette, basin scale eddy-field simulations and climatologies, and state-of-the-art physical models of global ocean circulation with biogeochemistry and genome-informed models of microbial processes[14]. The specific sampling strategy of each topical experiment is available in the respective campaign summary reports (see ***Data Records Section***). *Tara* Oceans data corresponding to methods described in this section are in part already open to the public at PANGAEA (Data Citation 1).

[2] Properties of seawater and particulate matter from physical, optical and imaging sensors mounted on the continuous surface water sampling system

Continuous measurements of surface water physical, chemical and biological properties serve the dual purpose of a) assessing the boundaries and the homogeneity/heterogeneity of an ecosystem studied during a station, and b) assessing the connectivity between stations. Underway measurements were often used to fine tune the location of sampling stations that were initially selected based on satellite images.

The in-line, Continuous Surface Sampling System [CSSS] installed on SV *Tara* (15,000 miles long track) comprised a SeaBird [TSG] temperature and conductivity sensor, a WETLabs [AC-S] spectrophotometer, a WETLabs chlorophyll [Fluorometer], and a Fast Repetition Rate Fluorometer [FRRF] to assess photosynthetic efficiency. All data were recorded simultaneously and archived daily in a single file, including navigation, date/time and GPS position. Water was pumped at the front of the vessel from ~2 m depth, then de-bubbled and circulated through the [AC-S], [TSG], [Fluorometer], and [FRRF]. An automated switching system provided periodic 0.2 µm filtered samples to the [AC-S], such that its particulate optical properties were not affected by instrument drift[15]. Systems maintenance (instrument cleaning, flushing) was done approximately once a week and in port between successive campaigns. In the Arctic Ocean and Arctic Seas (2013 campaigns), additional sensors for pH, PCO_2, optical backscattering (3 wavelengths), fluorescence emission [ALFA] and surface Photosynthetically Active Radiation [PAR] were added to the in-line system. A [FlowCytobot] also recorded images of microplankton every 20 min. Using daily discrete measurements of CDOM absorption with an [UltraPath] system, we calibrated the [AC-S] to also provide hourly CDOM absorption (besides particulate absorption and attenuation). Data were processed, quality-controlled, and are consistent with

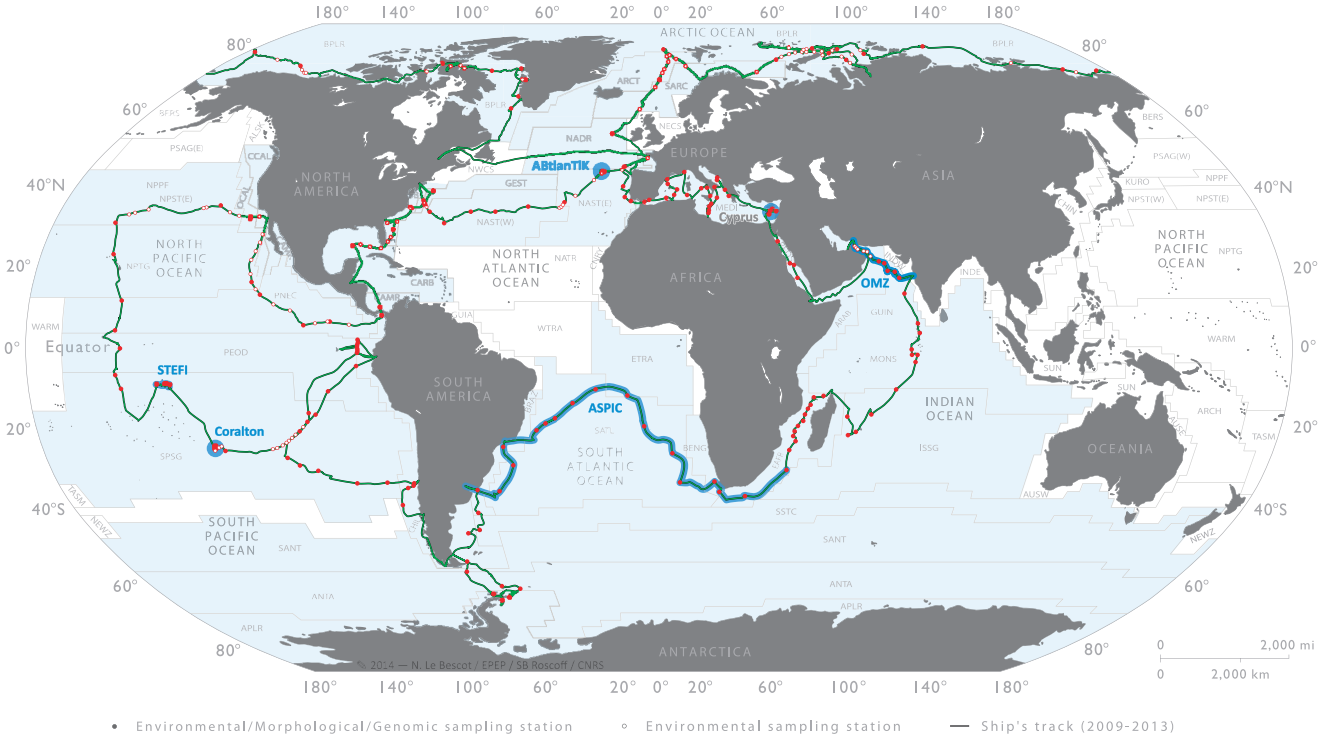

Figure 3. Sampling route, stations and topical experiments of the *Tara* Oceans Expedition. Sampling route of the *Tara* Oceans Expedition (green track), showing stations where plankton were sampled in their environmental context (full red dots) and where only environmental conditions were measured (open red dots). Topical experiments are identified along the sampling route (light blue). Longhurst biogeographical provinces[34] are shown in the background and those sampled during *Tara* Oceans Expedition are highlighted in blue.

remote sensing. A total of 60 tracks of continuous measurements corresponding to methods described in this section are in large part already open to the public at PANGAEA (Data Citation 2 and Data Citation 3).

[3] Properties of seawater and particulate & dissolved matter from physical, optical and imaging sensors mounted on the vertical profile sampling system

Repeated deployments of a Rosette Vertical Sampling System [RVSS] during day and night also served the dual purpose of a) assessing the boundaries and the homogeneity/heterogeneity of mesoscale features during a station, and b) assessing the connectivity between stations. These deployments were essential to locate features that have a vertical component and have a signature below the surface, such as eddies, upwellings, fronts, deep chlorophyll maxima, and oxygen minimum zones. The [RVSS] was specifically designed with various sensors, comprising 2 pairs of conductivity and temperature sensors (Sea-Bird), chrorophyll and CDOM fluorometers (WETLabs), a 25 cm transmissiometer for particles 0.5–20 µm (WETLabs), a one-wavelength backscatter meter for particles 0.5–20 µm (WETLabs), and a Underwater Vision Profiler[16] for particles >100 µm and zooplankton >600 µm (Hydroptic). A sbe43 oxygen sensor (Sea-Bird) and an In Situ Ultraviolet Spectrophotometer (ISUS) nitrate sensor (SATLANTIC) were also mounted on the Rosette. In the Arctic Ocean and Arctic Seas (2013 campaigns), a second sbe43 oxygen sensor (Sea-Bird) and a four frequency acoustic profiler (Aquascat) were added. Each component was powered on specific Li-Ion batteries and CTD data were self-recorded at 24 Hz. All sensors were calibrated in factory before, during and after the four year programme. Oxygen data were validated using climatologies. Nitrate and Fluorescence data were adjusted with discrete measurements from Niskin bottles mounted on the Rosette, and dark calibrations of the optical sensors were performed monthly on-board. A total of 837 vertical profiles were made during the Expedition. Additional stand-alone Sea-Bird components [sbe19] and [sbe9S] were exceptionally mounted directly on the oceanographic cable during harsh sea conditions, when the deployment of the rosette was not safe. In addition, apparent optical properties of sea water were measured using a surface tethered [RADIOMETRE-TSRB] in 2009–2012 and a profiling [RADIOMETRE-COPS] in 2013. Finally, a total of 101 deployments of the Secchi disk [SECCHI] provide a valuable, world-wide contribution to historical records of this very

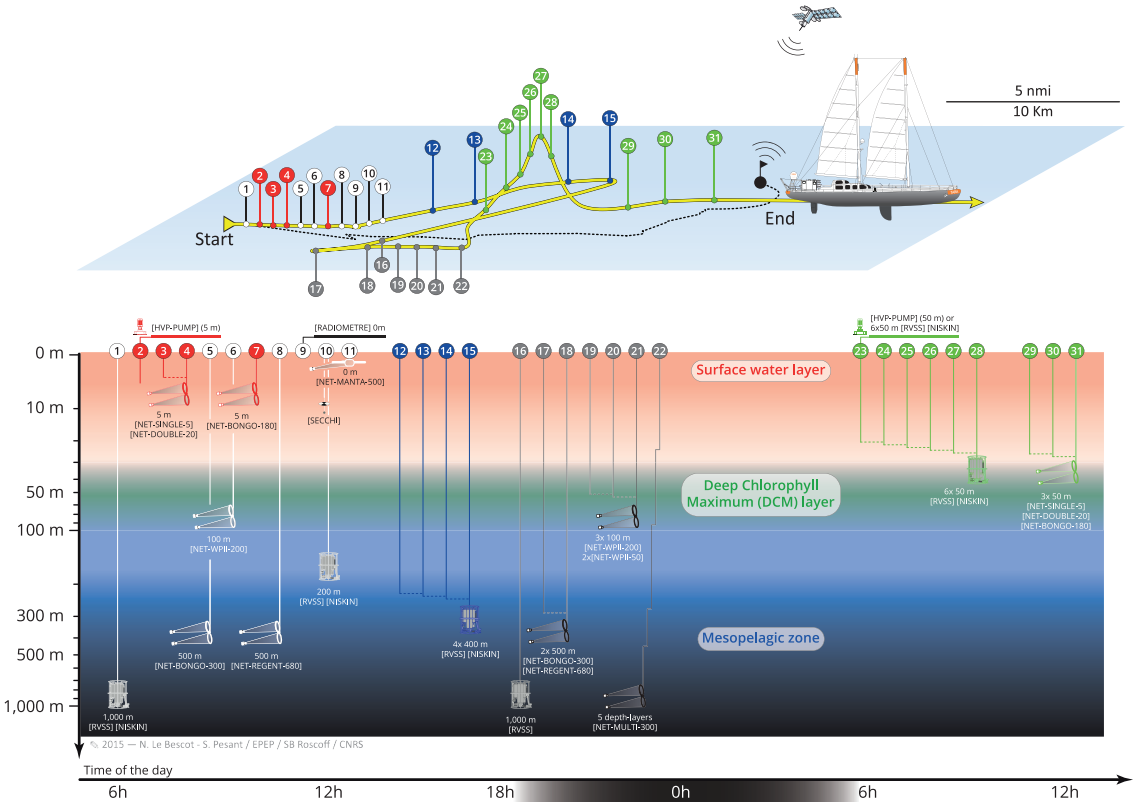

Figure 4. Spatial representation and chronology of sampling events during a 24-48 h station. Coloured markers along the route of *SV Tara* (yellow surface track) correspond to sampling events targeting the ***surface water layer*** (red,), ***deep chlorophyll maximum layer*** (green, here at 50 m), and the ***mesopelagic zone*** (blue, here at 400 m). At some stations, an Argo drifter (10-m floating anchor and satellite positioning) was used to follow the water mass during sampling (black surface track). White and grey markers correspond to day and night time deployments, respectively, of plankton nets [*TYPE-MESH*] and rosette [RVSS] casts that covered fixed depth layers of 0–100 m, 0–500 m or 0–1,000 m.

simple and yet fundamental sampling device. *Tara* Oceans data corresponding to methods described in this section are already open to the public at PANGAEA (Data Citation 4).

[4] Properties of seawater and particulate & dissolved matter from discrete water samples

In addition to sensors mounted on the Rosette Vertical Sampling System [RVSS], seawater was collected using Niskin bottles [NISKIN] (6 × 8-L Niskins and 4 × 12-L Niskins) in order to further characterise environmental conditions in the ecosystem under study. Measurements include pigment concentrations from HPLC analysis (10 depths per vertical profile; 25 pigments per depth), the carbonate system (Surface and 400 m; pHT, CO_2, pCO_2, fCO_2, HCO_3, CO_3, Total alkalinity, Total carbon, OmegaAragonite, OmegaCalcite, and quality Flag), nutrients (10 depths per vertical profile; NO_2, PO_4, $N0_2/NO_3$, SI, quality Flags), DOC, CDOM, and dissolved oxygen isotopes. More than 200 vertical profiles of these properties were made across the world ocean. DOC, CDOM and dissolved oxygen isotopes are available only for the Arctic Ocean and Arctic Seas (2013 campaigns). *Tara* Oceans data corresponding to methods described in this section are already open to the public at PANGAEA (Data Citation 5).

[5] Environmental features and sampling stations

During the *Tara* Oceans Expedition (2009–2013), plankton were sampled from 5–10-m thick layers in the water column, corresponding to specific environmental features that were characterised on-board from sensor measurements. Environmental features are defined by controlled vocabularies in the environmental ontology (EnvO; http://environmentontology.org/)[17].

The ***surface water layer*** (ENVO:00002042), sometimes labelled in the literature and databases as "surface", "SRF", "SUR", "SURF" or "S", was simply defined as a layer between 3 and 7 m below the sea surface. The ***deep chlorophyll maximum layer*** (ENVO:01000326), often labelled in the literature and databases as "DCM" or "D", was determined from the chlorophyll fluorometer (WETLabs optical sensors) mounted on the Rosette Vertical Sampling System [RVSS]. The presence of a DCM may indicate

a maximum in the abundance of plankton bearing chlorophyll pigments, or it may result from the higher chlorophyll content of plankton living in a darker environment[18]. This can be assessed *a posteriori* using water samples analysed for pigments by HPLC methods and from plankton counts. The ***mesopelagic zone*** (ENVO:00000213), also labelled in the literature and databases as "MESO" or "M", corresponds to the layer between 200 and 1000 m depths. The sampling depth within the ***mesopelagic zone*** was selected based on vertical profiles of temperature, salinity, fluorescence, nutrients, oxygen, and particulate matter. The selected depth varied from station to station, targeting for example a nutricline, a minimum concentration of oxygen, a maximum concentration of particulate matter, or a fixed depth of ca. 400 m when no particular feature could be identified. Other environmental features of special scientific interest include the ***oxygen minimum zone*** (ENVO:01000065), often labelled in the literature and data sets as "OMZ" or "O", and the ***epipelagic mixing layer*** (ENVO:01000061), also labelled in the literature and data sets as "ML", "MIX" or "X".

A complete sampling station consisted of collecting plankton from three distinct environmental features, typically the ***surface water layer***, ***deep chlorophyll maximum layer***, and ***mesopelagic zone*** (Fig. 4). Such a station lasted typically 24–48 h and special care was taken to reposition *SV Tara* in order to remain within a radius of 10 km and sample a homogeneous ecosystem as much as possible (see previous two sub-sections). The sequence of sampling deployments varied but generally followed the order illustrated in Fig. 4, i.e., ***surface water layer*** and ***mesopelagic zone*** during daytime on the first day, night sampling over fixed depths, and ***deep chlorophyll maximum layer*** during daytime on the second day. Sampling devices consisted essentially of a High Volume Peristaltic pump [HVP-PUMP], a Rosette Vertical Sampling System [RVSS] equipped with sensors and [NISKIN] bottles, instrumented plankton nets [NET-*TYPE-MESH*], a [SECCHI] disc, and a [RADIOMETRE] (Table 1 (available online only)).

Plankton were sampled from a total of 210 stations, of which: 51 stations did not target a specific environmental feature and conducted classical vertical profiles of physical and optical sensors and depth integrated net tows; 57 stations sampled only the ***surface water layer***; 62 stations sampled the ***surface water layer*** and a second depth-specific feature; and 40 stations sampled the ***surface water layer***, the ***deep chlorophyll maximum layer*** and a third depth-specific feature. *Tara* Oceans data corresponding to methods described in this section are already open to the public at PANGAEA (Data Citations 6–8) and are described in the ***Data Records Section*** of the present paper.

[6] Marine plankton

Plankton sampled during the *Tara* Oceans Expedition cover six orders of magnitude in size (10^{-2}–10^{5} µm) and correspond to viruses, giant viruses (giruses), prokaryotes (bacteria and archaea), unicellular eukaryotes (protists), and multicellular eukaryotes (metazoans). These five groups form the bulk of biomass throughout the oceans and drive the global biogeochemical cycles that regulate the Earth system[19–21]. Ocean viruses play an important role in plankton ecology by inducing mortality, horizontal gene transfer, and modulating microbial metabolism[22]. They are thought to target diverse prokaryotic and eukaryotic hosts including microalgae and heterotrophic protists, and to play a role in the evolution of their hosts[23]. Small viruses (< 0.2 µm) are known to be ubiquitous and the most abundant plankton in seawater, while the larger *giant viruses* or *giruses* (0.2–1 µm) were discovered more recently and are increasingly observed in marine samples[24,25]. Prokaryotes are believed to be responsible for 30% of primary production and 95% of community respiration in oceans[26] and are thus a fundamental component of marine food webs and biogeochemical processes. They are often divided and studied as two size fractions: the free-living prokaryotes range in size from 0.22–3 µm, and those that are attached to larger cells, particles, or aggregates are found in the 3–20 µm size-fraction[27]. In most cases, they are very difficult to culture. Unicellular eukaryotes, or protists, cover a broad range of cell size (0.8–2,000 µm). They are taxonomically very diverse with representatives in all of the 8 super-groups of the eukaryotic tree of life[28], whose roles in marine and Earth systems ecology are largely unexplored. Only the most abundant groups, such as diatoms and dinoflagellates, have been studied extensively in the field and cultured successfully[29]. Meso-zooplankton (metazoans; multicellular eukaryotes) range in size from 50 µm to tens of metres in colonial forms, and play a pivotal role in both the transfer of energy to higher trophic levels such as fish and other large predators, and in the vertical export of particulate matter produced at the surface of the ocean[30]. Their life history (e.g., metabolism, development, locomotion, reproduction, feeding) and their body-size are important properties affecting these two processes[31].

Various sampling methods were used to capture the diversity of both the dominant and less abundant organisms described above (see ***Technical Validation Section***). These methods effectively separated organisms into 10 size fractions: < 5 µm (or < 3 µm), 5–20 µm (or 3–20 µm), < 20 µm, 20–180 µm and 180–2,000 µm for planktonic viruses, prokaryotes and unicellular eukaryotes, and >50 µm, >200 µm, >300 µm, >500 µm and >680 µm for large planktonic unicellular eukaryotes and metazoans. Whenever possible, replicate sampling was performed to assess plankton natural variability and to ensure long-term storage of samples in view of future re-analysis using new technologies, notably in the fields of high throughput imaging and –omics which are evolving extremely rapidly.

Detailed protocols concerning the filtration, preservation and storage of plankton samples will be described in detail as a separate publication. Morphological data will be openly released at PANGAEA (http://www.pangaea.de) and nucleotides data will be openly released as they become available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/).

[6a] ***Sampling planktonic viruses, prokaryotes and unicellular eukaryotes***. Sampling devices used to collect small size organisms (< 20 µm size fractions) include Niskin bottles [NISKIN] mounted on the rosette [RVSS] or occasionally attached individually on the oceanographic cable, a High Volume Peristaltic Pump [HVP-PUMP], and exceptionally a 10-L plastic bucket [BUCKET] (Table 1 (available online only)). Waste waters from *Tara* were not released automatically at sea and were not purged during stations. The pump [HVP-PUMP] was fixed on deck and connected to a 40 mm diameter flexible tube, lowered from either starboard to sample surface water (3–7 m) or from the stern to sample other depths down to 60 m. At the beginning of each sampling activity, the tube was rinsed by letting seawater flow overboard for 10 min. A single deployment of the pump [HVP-PUMP] brought back up to 1,000 l, whereas 4–6 deployments of the rosette [RVSS] [NISKIN] were necessary to collect 400–600 l of seawater.

The choice of the sampling device was determined by weather conditions and the depth of the targeted environmental features. The ***surface water layer*** was systematically sampled using the pump [HVP-PUMP], and exceptionally (at 3 stations) using a 10-Litres plastic bucket [BUCKET]. The ***deep chlorophyll maximum layer*** was sampled preferentially with the pump [HVP-PUMP] or alternatively using multiple deployments of the rosette [RVSS] [NISKIN] when the sampling depth was >60 m. A WETLabs three-optical-sensor [ECOTriplet], measuring pressure, temperature, CDOM and chlorophyll fluorescence, was attached to the intake of the pump tube to monitor sampling depth, temperature and chlorophyll concentrations in real time during pumping. The ***mesopelagic zone*** was systematically sampled using multiple deployments of the rosette [RVSS] [NISKIN].

Whole seawater collected by these devices was then pre-filtered successively on nylon conical sieves with a mesh of 200 µm and 20 µm, and additionally 5 µm for protists (Table 1 (available online only)). The filtrate was collected in four to six 100-L polyethylene containers, which were thoroughly washed with 0.1% bleach, rinsed twice with fresh water and rinsed again twice with the filtrate. Depending on protocols, the < 5 µm and < 20 µm filtrates were further fractionated on-board using one or a combination of membranes with pore sizes 0.1 µm, 0.2 µm, 0.45 µm, 0.7 µm, 0.8 µm, 1.6 µm or 3 µm. The retention efficiency of meshes, pore-membranes and fibre-filters is a constant debate in plankton ecology. Organisms display various shapes, including high length-to-width ratios, some may easily "squeeze" through pores smaller than their "normal" size, and others may form colonies or tend to aggregate into particles much larger than their individual size. We do not intend to assess the efficiency of the various meshes and filters in retaining the different groups of organisms targeted during the *Tara* Oceans Expedition. We simply picked commonly used size-fractions and accept the fact that organisms or parts of organisms from the different groups may be present in several size-fractions.

The choice of size thresholds used to collect small eukaryotes, and prokaryotes associated with small particles or with eukaryotes varied during the *Tara* Oceans Expedition, between 3–20 µm and 5–20 µm. Plankton from that size fraction comprise organisms that are often not abundant enough in ***whole seawater*** and often too fragile to be collected with plankton nets that are themselves too delicate to be deployed in rough seas. The sampling method was therefore weather-dependent and often a combination of using either the pump [HVP-PUMP] or Niskin bottles [RVSS] [NISKIN] for ***whole seawater***, and using a plankton net with a mesh size of 5 µm [NET-SINGLE-5] when sea conditions allowed.

When the pump [HVP-PUMP] or Niskin bottles [RVSS] [NISKIN] were used to collect the 3–20 µm or 5–20 µm size-fraction, ***whole seawater*** was pre-filtered successively on 200 µm and 20 µm, and either collected on 3 µm membrane filters (1 to 5 replicates) or concentrated in a conical, 5 µm mesh nylon sieve [SIEVE-5]. The volume of water passing through each 3 µm pore size membrane was consistently 100 l, whereas the volume passing through the [SIEVE-5] was estimated by recording the pumping rate and the start and end time of pumping. Material collected in the [SIEVE-5] was washed-off into an 8-L polyethylene bottle using seawater pre-filtered on 0.1 µm, up to a final volume of 3 l. The 3–20 µm method was used throughout the *Tara* Oceans Expedition (2009–2013), whereas the 5–20 µm method was used only in 2009–2012, i.e., not in the Arctic.

When nets [NET-SINGLE-5] were used for the 5–20 µm size-fraction, they were lowered to the selected environmental feature and towed horizontally for 5–15 min at a speed of 0.3 m/s. Net samples from the cod-end were sieved through 20 µm [SIEVE-20] and poured into an 8-L polyethylene bottle, which was thoroughly pre-washed with 0.1% bleach, rinsed twice with fresh water and rinsed again twice with seawater pre-filtered on 0.1 µm. The volume of ***net sample*** was adjusted to 3 l with 0.1 µm pre-filtered seawater. After each use, nets, cod-ends, and sieves were rinsed with fresh water and checked for holes. That method was used only in 2009–2012, i.e., not in the Arctic.

Plankton from the 20–180 µm size-fraction were collected using a double plankton net with a 20 µm mesh size [NET-DOUBLE-20]. Nets were lowered to the selected environmental feature and towed horizontally for 5–15 min at a speed of 0.3 m/s. Net samples from the two cod-ends were sieved through 180 µm [SIEVE-180] and poured into an 8-L polyethylene bottle, which was thoroughly pre-washed with 0.1% bleach, rinsed twice with fresh water and rinsed again twice with seawater pre-filtered on 0.1 µm. The volume of ***net sample*** was adjusted to 3 l with 0.1 µm pre-filtered seawater. After each use, nets, cod-ends, and sieves were rinsed with fresh water and checked for holes.

Plankton from the 180–2,000 µm size-fraction were collected using a 180 µm Bongo Net [NET-BONGO-180]. Nets were lowered to the selected environmental feature and towed horizontally for 5–15 min at a speed of 0.3 m/s. Net samples from the two cod-ends were sieved through 2,000 µm [SIEVE-2000] and poured into an 8-L polyethylene bottle, which was thoroughly pre-washed with 0.1%

bleach, rinsed twice with freshwater and rinsed again twice with seawater pre-filtered on 0.1 µm. The volume of ***net sample*** was adjusted to 3 l with 0.1 µm pre-filtered seawater. After each use, nets, cod-ends, and sieves were rinsed with fresh water and checked for holes.

[6b] ***Sampling large planktonic unicellular eukaryotes and metazoans***. Sampling devices used to concentrate and collect the larger and less abundant organisms (>50 µm size fractions) consisted of plankton nets with mesh sizes ranging from 50 to 680 µm [NET-*TYPE-MESH*] and metal pan-shaped sieves [SIEVE-*MESH*] to remove large organisms as needed (Table 1 (available online only)). All nets were equipped with a flow meter and a temperature-depth recorder, and their depth was monitored and adjusted during deployments using an acoustic SCANMAR system. Upon recovery, all nets were rinsed from the outside with running seawater. Cod-ends and metal sieves used to size-fractionate samples were rinsed with running seawater pre-filtered successively on 25 µm and 0.1 µm, using Polygard-CR Cartridge Filters (CR2501006, CRK101006). After each use, nets, cod-ends, and sieves were rinsed with fresh water and checked for holes.

Organism-selectivity and capture-efficiency of plankton nets depend on the mesh size and deployment methods, i.e., depth, tow method (oblique/vertical/horizontal), tow speed and time of day[32]. During the *Tara* Oceans Expedition, plankton nets were deployed during day and night in order to capture the nycthemeral vertical migrations. Both [NET-WPII-50] and [NET-WPII-200] were towed vertically or obliquely from a depth of 100 m to the surface, during night and daytime, at a speed of 0.3–0.5 m/s depending on weather conditions. Both [NET-BONGO-300] and [NET-REGENT-680] were towed obliquely from a depth of 500 m to the surface, during night and daytime, at a speed of 0.5 m/s depending on weather conditions. Net samples were preserved on-board with buffered formaldehyde, ethanol or RNA-Later for later morphological and/or molecular analyses.

Where time and weather allowed, a multiple opening-closing net equipped with 5 nets of 300 µm mesh size [NET-MULTI-300] was deployed preferentially at night or during daytime to study the vertical distribution of zooplankton. Nets opened and closed at selected depths between 1,000 m and the surface, according to water column features identified from vertical profiles of temperature, salinity, fluorescence, nutrients, oxygen, and particulate matter. Samples were preserved on-board in buffered formaldehyde. The Underwater Vision Profiler (UVP) mounted on the rosette was also used to study the vertical distribution of zooplankton >600 µm during day and night.

In 2011–2013, a neuston net [NET-MANTA-500] was towed at the surface for about 1 h at a speed of 0.7 m/s in order to collect plastic particles and associated organisms. Samples were preserved in ethanol for later morphological and molecular analyses. Finally, a Continuous Plankton Recorder (CPR) was deployed between stations in 2013. Samples were preserved in formaldehyde and sent to the Sir Alister Hardy Foundation for Ocean Science (SAHFOS) for later morphological and molecular analyses.

Data Records

Tara Oceans developed best practices for the standardisation and interoperability of data generated across environmental, morphological and molecular analyses. This effort contributed to the publication of a set of standards for reporting and serving data in Marine Microbial Biodiversity, Bioinformatics and Biotechnology (M2B3)[33]. Here we describe three levels of the M2B3 reporting standard: ***campaigns***, ***stations*** and ***events***. For each level, we provide a registry of all campaigns/stations/events, pdf documents describing each campaign/station/event, and universal resource locator (URL) queries to access related nucleotides and environmental data.

Labels

Labels that identify each campaign, station or event are built using a consistent syntax that has the following format:

Campaign labels: TARA_date(yyyymmddZ), e.g.,TARA_20110401Z
Station labels: TARA_station#(001-210), e.g.,TARA_100
Event labels: TARA_datetime(yyyymmddThhmmZ)_station#(001-210)_EVENT_TYPE, e.g.,TARA_20110416T1306Z_100_EVENT_CAST

Registries

The *Tara* Oceans Expedition (2009–2013) comprised 60 campaigns (from port-to-port), 210 stations and over 3200 sampling events. Three registries listing the campaigns, stations and events are published at PANGAEA, Data Publisher for Earth and Environmental Science:

Campaigns registry (Data Citation 6): http://doi.pangaea.de/10.1594/PANGAEA.842191
Stations registry (Data Citation 7): http://doi.pangaea.de/10.1594/PANGAEA.842237
Events registry (Data Citation 8): http://doi.pangaea.de/10.1594/PANGAEA.842227

The ***campaigns*** registry provides details about the scientific interest of each campaign, a list of scientists on board, and URLs for the corresponding campaign summary report (pdf), environmental data sets and nucleotides data sets.

The ***stations*** registry provides details about the geographic context of each station, including mean and maximum bathymetric depth (extracted from the General Bathymetric Chart of the Oceans; GEBCO), minimum distance from the coast, the corresponding marine biomes and biogeographical provinces defined by Longhurst[34], and when applicable information about the corresponding exclusive economic zone and related legal aspects. Additionally for each station, we provide information about the environmental features that were sampled and their depth, and the number of deployments carried out with the different sampling devices listed in Table 1 (available online only). URLs provide access to the corresponding oceanographic context report (pdf), environmental data sets and nucleotides data sets.

The ***events*** registry provides details about the sampling date, time, location and methodology of each event. URLs provide access to the corresponding event log sheet (pdf), environmental data sets and nucleotides data sets. Sampling events occurring outside the context of a station were assigned the station label TARA_999. Such events include for example underway measurements of the on-board meteorological station [BATOS] and of the continuous surface sampling system [CSSS], and exceptional deployments of plankton nets [NET-*TYPE-MESH*] or of the rosette vertical sampling system [RVSS].

All available metadata information about a single campaign/station/event can be extracted from the respective registries using a web service (description: https://ws.pangaea.de/dds-fdp/). Here we provide an example for each registry:

Campaign-specific metadata query: https://ws.pangaea.de/dds-fdp/rest/panquery?datasetDOI=doi.pangaea.de/10.1594/PANGAEA.842191&filterParameter=Campaign&filterValue=TARA_20110401Z

Station-specific metadata query: https://ws.pangaea.de/dds-fdp/rest/panquery?datasetDOI=doi.pangaea.de/10.1594/PANGAEA.842237&filterParameter=Station&filterValue=TARA_100

Event-specific metadata query: https://ws.pangaea.de/dds-fdp/rest/panquery?datasetDOI=doi.pangaea.de/10.1594/PANGAEA.842227&filterParameter=Event&filterValue=TARA_20110416T1306Z_100_EVENT_CAST

Additionally, the same web service can be used to extract information about all stations from a campaign, all events from a campaign, or all events from a station as shown in these examples:

Stations-from-a-campaign metadata query: https://ws.pangaea.de/dds-fdp/rest/panquery?datasetDOI=doi.pangaea.de/10.1594/PANGAEA.842237&filterParameter=Campaign&filterValue=TARA_20110401Z

Events-from-a-campaign metadata query: https://ws.pangaea.de/dds-fdp/rest/panquery?datasetDOI=doi.pangaea.de/10.1594/PANGAEA.842227&filterParameter=Campaign&filterValue=TARA_20110401Z

Events-from-a-station metadata query: https://ws.pangaea.de/dds-fdp/rest/panquery?datasetDOI=doi.pangaea.de/10.1594/PANGAEA.842227&filterParameter=Station&filterValue=TARA_100

Reports and log sheets

Campaign reports were written by the chief scientist and scientific crew in order to document the objectives and main achievements of each campaign, as well as any deviations from the regular sampling programme (e.g.,TARA_20110401Z_report.pdf). The station reports were written by Flavian Kokoszka, Rémi Laxenaire and Sabrina Speich to provide background knowledge of the physical oceanography at each station (e.g.,TARA_100_oceanographic_context_report.pdf). Finally, event log sheets were filled on board each time a sampling device was deployed, recording the position, date, time, type of device, sampling depths, volume sampled, ID and filename of sensor outputs, operator's comments, and unique identifiers (barcodes) of samples collected during the event (e.g.,TARA_20110415T1312Z_100_EVENT_CAST.pdf). Most of the information found in reports and log sheets was extracted manually, quality checked using controlled vocabularies, and archived in the campaigns, stations and events registries. Nevertheless, these narrative documents remain a valuable and complementary source of information. The three registries contain universal resource locators (URLs) pointing to the reports and log sheets of each campaign/station/event. Reports and log sheets can also be browsed directly in the PANGAEA store:

Campaign reports: http://store.pangaea.de/Projects/TARA-OCEANS/Campaign_Reports/
Station reports: http://store.pangaea.de/Projects/TARA-OCEANS/Station_Reports/
Event log sheets: http://store.pangaea.de/Projects/TARA-OCEANS/Logsheets_Event/

Up-to-date lists of available nucleotides and environmental data sets

Tara Oceans data will progressively be released openly following their analysis and validation. Up-to-date lists of nucleotides and environmental data sets can be obtained from universal resource locator (URL) queries that are made specific to any campaign/station/event by using labels from the campaigns, stations and events registries.

A list of environmental data sets published at PANGAEA can be obtained by combining the following base URL: http://www.pangaea.de/search?q= with a search term. The URL query is made specific to any *Tara* Oceans campaign, station or event by adding the corresponding label as the search term, see

Group of organism	Environment		Cell/body size Dimension: L^1 Unit: micrometre		Abundance Dimension: L^{-3} Unit: individuals per Litre		Species richness Dimension: L^{-3} Unit: species per Litre		Reference
	Coastal	Oceanic	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	
Viruses	X	X	2×10^{-2}	2×10^{-1}	1×10^{9}	2×10^{11}			37
Viruses	X	X	2×10^{-2}	2×10^{-1}	1×10^{10}				38
Viruses		X	2×10^{-2}	2×10^{-1}	6×10^{7}	$2{,}6\times10^{8}$			39
Viruses		X	2×10^{-2}	2×10^{-1}	$1{,}2\times10^{7}$	$3{,}5\times10^{8}$			39
Viruses	X	X					1×10^{5}	1×10^{6}	40
Prokaryotes	X				3×10^{9}				41
Prokaryotes	X	X	<	1×10^{0}	4×10^{9}				42
Prokaryotes	X		<	2×10^{0}	1×10^{9}	5×10^{9}			43
Prokaryotes		X	<	2×10^{0}	5×10^{8}				43
Prokaryotes		X	<	2×10^{0}	2×10^{8}	9×10^{8}			44
Prokaryotes		X	<	2×10^{0}	2×10^{8}	6×10^{8}			44
Prokaryotes	X		<	2×10^{0}	5×10^{8}	7.5×10^{9}			45
Prokaryotes		X	<	2×10^{0}	$2{,}25\times10^{8}$	3×10^{8}			46
Prokaryotes		X	<	2×10^{0}	2×10^{8}	8×10^{8}			46
Prokaryotes					1×10^{9}				47
Prokaryotes							1×10^{7}		48
Prokaryotes							2×10^{4}		49
Prokaryotes							2×10^{4}		50
Protists	X				3×10^{4}				41
Protists		X	8×10^{-1}	2×10^{0}	$1{,}83\times10^{5}$				51
Protists		X	6×10^{-1}	1×10^{1}	1×10^{5}	4×10^{6}			52
Protists		X	2×10^{-1}	2×10^{0}	2×10^{7}	1×10^{9}			53
Protists	X		8×10^{-1}	2×10^{0}	5×10^{6}	3×10^{9}			54
Protists	X	X	2×10^{-1}	3×10^{0}	1×10^{6}	2×10^{7}			55
Protists		X	8×10^{0}	1×10^{1}	1×10^{7}	3×10^{8}			56
Protists	X		<	2×10^{1}	$1{,}08\times10^{5}$				57
Protists	X		<	2×10^{1}	$2{,}6\times10^{5}$				57
Protists	X		1×10^{1}	1×10^{2}	1×10^{7}				58
Protists		X	5×10^{0}	2×10^{3}	1×10^{10}				58
Protists	X		$1{,}8\times10^{1}$	5×10^{2}	2×10^{4}	5×10^{5}			59
Protists							3×10^{3}		50
Protists							1.4×10^{6}	$1{,}6\times10^{7}$	60
Protists			1×10^{0}	1×10^{4}					61
Protists							1×10^{6}	>	62
Metazoa	X				2×10^{-2}	8×10^{1}			41
Metazoa, copepods	X		1×10^{2}	>	few	$2{,}5\times10^{2}$			63
Metazoa, copepods			2×10^{2}	>			3×10^{0}	4×10^{1}	64
Metazoa, copepods		X	5×10^{2}	>	few	8×10^{1}			65
Metazoa, copepods		X	1×10^{2}	>	few	$1{,}5\times10^{2}$			65
Metazoa	X		1×10^{4}	>	5×10^{-5}	7×10^{-4}			66
Metazoa	X	X	2×10^{3}	>	6×10^{-4}	5×10^{0}			67
Metazoa							7×10^{3}		68

Table 2. Cell/body size, abundance and species richness reported in the literature for the four groups of plankton in coastal and oceanic environments.

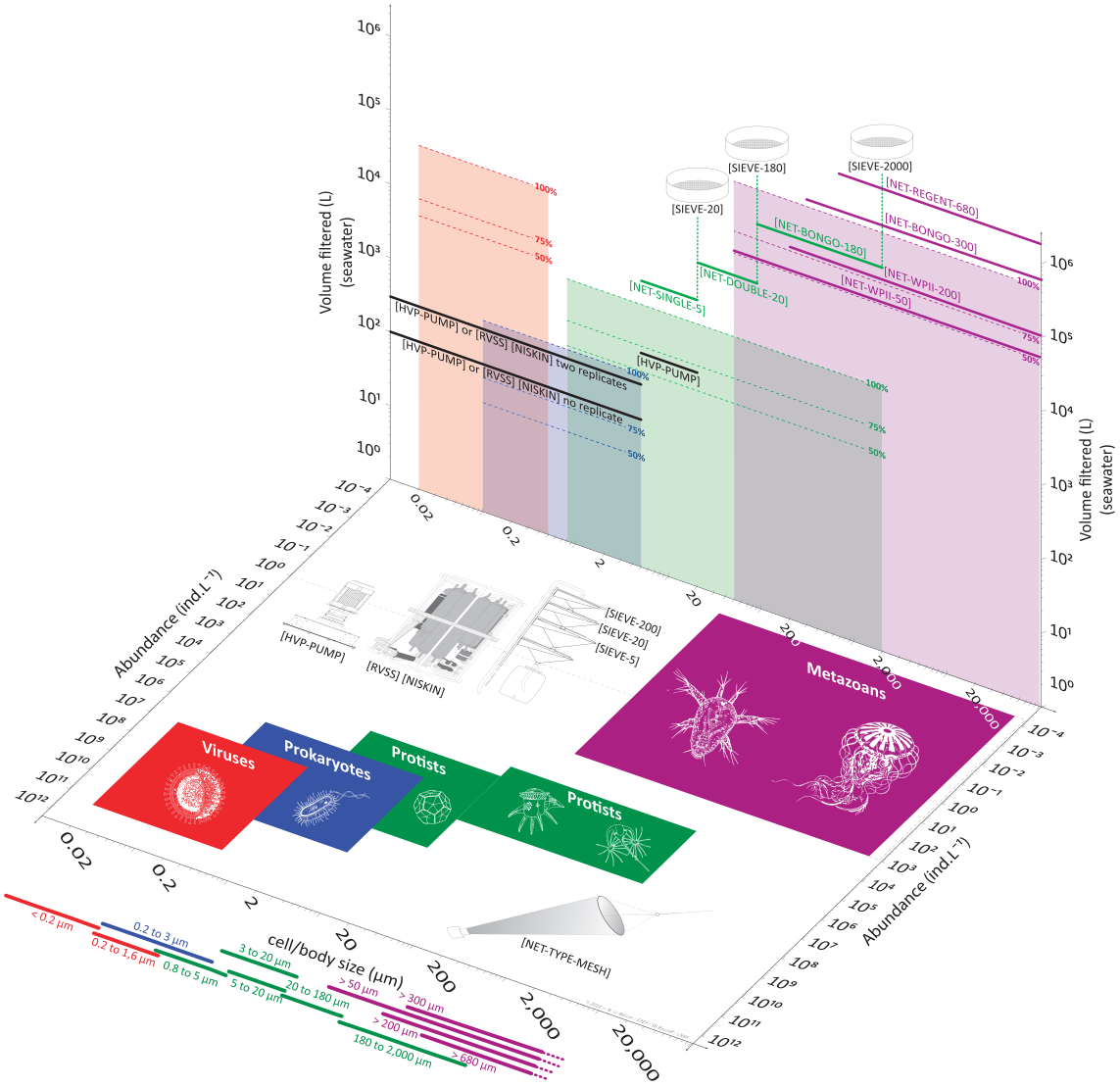

Figure 5. Empirical basis for the size-fractionation approach and the choice of sampling devices. The horizontal plane shows the range of body/cell size and natural abundances reported in the literature (Table 2) for viruses (including giant viruses), prokaryotes, protists and metazoans (coloured boxes). The sampling devices used to collect plankton < 5 µm in size (i.e., high volume peristaltic pump and rosette with Niskin bottles) and >5 µm in size (i.e., plankton nets) are illustrated as well on the horizontal plane. The vertical plane shows the volume of seawater required to capture 100, 75 and 50% of species richness reported in the literature (Table 2) for viruses (including giant viruses), prokaryotes, protists and metazoans (shaded boxes). The typical volume of seawater collected by sampling devices are shown in comparison (horizontal thick lines). Also illustrated on the vertical plane: Sieves were used to remove large organisms from protists net samples.

examples below. Already-built URL queries for each campaign, station or event are provided in the three respective registries (Data Citations 6–8).

Campaign-specific environmental data query: http://www.pangaea.de/search?q=TARA_20110401Z
Station-specific environmental data query: http://www.pangaea.de/search?q=TARA_100
Event-specific environmental data query: http://www.pangaea.de/search?q=TARA_20110416T1306Z_100_EVENT_CAST

A list of nucleotides data published at ENA can be obtained by combining the following base URL: http://www.ebi.ac.uk/ena/data/search?query= with a search term. The URL query is made specific to any *Tara* Oceans campaign, station or event by adding the corresponding label as the search term. Already-built URL queries for each campaign, station or event are provided in the three respective registries (Data Citations 6–8).

Technical Validation

Here we provide a first order validation of the *Tara* Oceans sampling methodology by compiling published values of plankton cell/body size, natural abundance and richness (Table 2). These are compared to the sampling volume and mesh size of the different sampling methods (Fig. 5).

Life history traits such as cell/body size and the natural range of abundance determine the general structure and dynamics of food webs and other ecological networks, across multiple scales of organisation[31,35]. Here we characterise the five groups of plankton by their size and abundance in seawater, using values from the literature (Table 2). The range of these characteristics are summarised for each plankton group using coloured areas on the horizontal plane of Fig. 5. As already described for a wide range of organisms[36], the literature shows an inverse relationship between plankton size and abundance in the natural environment, so that small viruses (10^{-2}–10^{0} µm) generally form the most abundant group (10^{7}–10^{11} ind. L^{-1}), whereas the larger metazoans (10^{1}–10^{5} µm) are generally the least abundant group (10^{-4}–10^{3} ind. L^{-1}).

Species richness and evenness are used to estimate species diversity, and should therefore be considered when designing sampling strategies and methodologies for biodiversity studies. Here we characterise the five groups of plankton by their species richness in seawater, using values from the literature (Table 2, refs. 37–68). From these values we made "back of the envelope" calculations of the volume of seawater required to capture 100, 75 and 50% of species within each group of plankton (Fig. 5; coloured areas on the vertical plane). The effectiveness of our sampling strategy can be assessed by comparing these coloured areas with the sampling volume of the various sampling devices used during the *Tara* Oceans Expedition (horizontal full lines on the vertical plane).

Based on this assessment, it appears that our sampling strategy would capture $< 50\%$ of total richness for viruses and small size protists (0.8–5 µm). Accordingly, one would need to filter thousands of litres of seawater in order to capture 75% of total richness for these groups. This is both impractical for most field campaigns and dependent on how one defines the currency of richness for these groups, i.e., the concept of species. In all other groups and size fractions, our sampling strategy appears to capture >75% of species richness, and 100% in the case of large size protists and metazoans. It is important to note that data about plankton richness is very scarce in the literature, so that this assessment is only a first approximation. *Tara* Oceans data will undoubtedly contribute to fill this knowledge gap and improve the sampling design of future ocean biodiversity surveys.

Usage Notes

The *Tara* Oceans data policy follows the open science principle of open access and early release of raw and validated data sets. All data presented here (Data Citations 6–8) are published under the Creative Commons Attribution 3.0 Unported (CC-by 3.0) and must therefore be cited when used in scientific papers, posters or presentations. As with most scholarly publications, data citations have authors, a title, a year of publication and a digital object identifier (see data citations in the Reference Section). Furthermore, we kindly ask to include the ***Tara* Oceans Consortium** in the acknowledgements. When referring to the *Tara* Oceans Data or to the sampling strategy and methodology of the *Tara* Oceans Expedition, please cite the present paper.

References

1. Gross, L. Untapped bounty: sampling the seas to survey microbial biodiversity. *PLoS Biol.* **5,** e85 (2007).
2. Laursen, L. Spain's ship comes. in. *Nature* **475,** 16–17 (2011).
3. Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* **9,** e1001177 (2011).
4. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in *Tara* Oceans microbial metagenomes. *The ISME Journal* **7,** 1678–1695 (2013).
5. Boss, E. *et al.* The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean; Contribution of the *Tara* Oceans expedition. *Methods in Oceanography* **7,** 52–62 (2013).
6. Roullier, F. *et al.* Particle size distribution and estimated carbon flux across the Arabian Sea oxygen minimum zone. *Biogeosciences* **11,** 4541–4557 (2014).
7. Brum, J. R. *et al.* Global patterns and ecological drivers of ocean viral communities. *Science* doi:10.1126/science.1261498 (2015).
8. de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit global ocean. *Science* doi:10.1126/science.1261605 (2015).
9. Lima-Mendez, G. *et al.* Top-down determinants of community structure in the global plankton interactome. *Science* doi:10.1126/science.1262073 (2015).
10. Sunagawa, S. *et al.* Structure and Function of the Global Ocean Microbiome. *Science* doi:10.1126/science.1261359 (2015).
11. Villar, E. *et al.* Environmental characteristics of Agulhas rings affect inter-ocean plankton transport. *Science* doi:10.1126/science.1261447 (2015).
12. Testor, P. *et al.* Gliders as a component of future observing systemsin *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society* Vol. 2. (ESA Publication WPP-306, 2010).
13. Xing, X. *et al.* Combined processing and mutual interpretation of radiometry and fluorimetry from autonomous profiling Bio-Argo floats: Chlorophyll a retrieval. *J. Geophys. Res.* **116,** C06020 (2011).
14. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. *Science* **315,** 1843–1846 (2007).
15. Slade, W. H. *et al.* Underway and moored methods for improving accuray in measurement of spectral particulate absorption and attenuation. *J. Atmos. Oceanic Technol* **27,** 1733–1746 (2010).
16. Falkowski, P. G. Physiological responses of phytoplankton to natural light regimes. *J. Plankton Res.* **6,** 295–307 (1984).
17. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semant* **4,** 43 (2013).
18. Picheral, M. *et al.* The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods* **8,** 462–473 (2010).

19. Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* **437,** 349–355 (2005).
20. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–1039 (2008).
21. Karl, D. M. Microbial oceanography: paradigms, processes and promise. *Nat. Rev. Microbiol.* **5,** 759–769 (2007).
22. Breitbart, M. Marine viruses: truth or dare. *Ann. Rev. Mar. Sci* **4,** 425–448 (2012).
23. Thomas, R. *et al.* Acquisition and maintenance of resistance to viruses in eukaryotic phytoplankton populations. *Environmental Microbiology* **13,** 1412–1420 (2011).
24. Monier, A., Claverie, J. M. & Ogata, H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biology* **9,** R106 (2008).
25. Claverie, J.-M. & Ogata, H. Ten good reasons not to exclude giruses from the evolutionary picture. *Nat. Rev. Microbiol.* **7,** 615 (2009).
26. Del Giorgio, P. A. & Duarte, C. M. Respiration in the open ocean. *Nature* **420,** 379–384 (2002).
27. Acinas, S. G., Anton, J. & Rodriguez-Valera, F. Diversity of free-living and attached bacteria in offshore Western Mediterranean waters as depicted by analysis of genes encoding 16S rRNA. *Appl. Environ. Microbiol.* **65**(2): 514–522 (1999).
28. Baldauf, S. L. The Deep Roots of Eukaryotes. *Science* **300,** 1703–1706 (2003).
29. Vaulot, D., Le Gall, F., Marie, D., Guillou, L. & Partensky, F. The Roscoff Culture Collection (RCC): a collection dedicated to marine picoplankton. *Nova Hedwigia* **79,** 49–70 (2004).
30. Banse, K. Reflections about chance in my career, and on the top-down regulated world. *Ann. Rev. Mar. Sci* **5,** 1–19 (2013).
31. Falkowski, P. Ocean Science: The power of plankton. *Nature* **483,** 17–20 (2012).
32. Wiebe, P. H. & Benfield, M. C. From the Hensen net toward four-dimensional biological oceanography. *Progress in Oceanography* **56,** 7–136 (2003).
33. Ten Hoopen, P. *et al.* Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. *Standards in Genomic Sciences* **10,** 1–10, doi:10.1186/s40793-015-0001-5 (2015).
34. Longhurst, A. *Ecological Geograhy of the Sea*. (London, 2007).
35. Woodward, G. *et al.* Body size in ecological networks. *Trends Ecol. Evol.* **20,** 402–409 (2005).
36. Peters, R. H.. *The Ecological Implications of Body Size*. (Cambridge University Press, 1983).
37. Brussaard, C. P., Payet, J. P., Winter, C. & Weinbauer, M. G. Quantification of aquatic viruses by flow cytometry. *Manual of Aquatic Viral Ecology* **11,** 102–107 (2010).
38. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
39. Weinbauer, M. G., Rowe, J. M. & Wilhelm, S. W. Determining rates of virus production in aquatic systems by the virus reduction approach. *Manual of Aquatic Viral Ecology* **1,** 1–8 (2010).
40. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biology* **4,** e368 (2006).
41. Gilabert, J. Short-term variability of the planktonic size structure in a Mediterranean coastal lagoon. *J. Plankton Res.* **23,** 219–226 (2001).
42. Buitenhuis, E. T. *et al.* Bacterial biomass distribution in the global ocean. *Earth System Science Data* **4,** 37–46 (2012).
43. Ducklow, H. W. Bacterial production and biomass in the oceans. *Microbial Ecology of the Oceans* **1,** 85–120 (2000).
44. Hyun, J.-H. & Kim, K.-H. Bacterial abundance and production during the unique spring phytoplankton bloom in the central Yellow Sea. *Mar. Ecol. Prog. Ser.* **252,** 77–88 (2003).
45. Li, W. K. W. Annual average abundance of heterotrophic bacteria and synechococcus in surface ocean waters. *Limnol. Oceanogr.* **43,** 1746–1753 (2007).
46. Church, M. J. Resource control of bacterial dynamics in the sea. *Microbial Ecology of the Oceans* **1,** 335–382 (2008).
47. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73,** 249–299 (2009).
48. Pedrós-Alió, C. Marine microbial diversity: can it be determined? *Trends in Microbiology* **14,** 257–263 (2006).
49. Pedrós-Alió, C. The Rare Bacterial Biosphere. *Annu. Rev. Marine. Sci* **4,** 449–466 (2012).
50. Amaral-Zettler, L. A. *et al.* Microbial community structure across the tree of life in the extreme Río Tinto. *The ISME Journal* **5,** 42–50 (2011).
51. Raghukumar, S. Ecology of the marine protists, the Labyrinthulomycetes (Thraustochytrids and Labyrinthulids). *European Journal of Protistology* **38,** 127–145 (2002).
52. Signorini, S. R. & McClain, C. R. Environmental factors controlling the Barents Sea spring-summer phytoplankton blooms. *Geophys. Res. Lett.* **36,** L10604 (2009).
53. Evans, C., Archer, S. D., Jacquet, S. & Wilson, W. H. Direct estimates of the contribution of viral lysis and microzooplankton grazing to the decline of a Micromonas spp. population. *Aquat. Microb. Ecol.* **30,** 207–219 (2003).
54. Countway, P. D. & Caron, D. A. Abundance and distribution of Ostreococcus sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Applied and Environmental Microbiology* **72,** 2496 (2006).
55. Worden, A. Z, Not, F.Ecology and Diversity of Picoeukaryotesin *Microbial Ecology of the Oceans*, (ed. Kirchman D. L.) 159–205 (John Wiley & Sons, Inc., 2008).
56. Holligan, P. M. *et al.* A biogeochemical study of the coccolithophore, Emiliania huxleyi, in the North Atlantic. *Geophys. Res. Lett.* **7,** 879 (1993).
57. Ya-Hui, G. *et al.* Marine Nanoplanktonic diatoms from the coastal waters of Hong Kong. In *Proceedings of an International Workshop Reunion Conference* **21-26,** 103–104 (2001).
58. Taylor, F. J. R. & Pahlinger, U.The Biology of Dinoflagellates *Botanical Monographs* 21, 399–529 (Wiley-Blackwell, 1987).
59. Smalley, G. W. & Coats, D. W. Ecology of the Red-Tide Dinoflagellate Ceratium furca: distribution, mixotrophy, and grazing impact on ciliate populations of Chesapeake Bay. *J. Eukaryot. Microbiol.* **49,** 63–73 (2002).
60. Adl, S. M. *et al.* Diversity, nomenclature, and taxonomy of protists. *Systematic Biology* **56,** 684 (2007).
61. Behnke, A. *et al.* Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental Microbiology* **13,** 340–349 (2010).
62. Edgcomb, V. *et al.* Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *The ISME Journal* **6,** 1–13 (2011).
63. Turner, J. T. The importance of small planktonic copepods and their roles in pelagic marine food webs. *Zool. Stud.* **43,** 255–226 (2004).
64. Rombouts, I. *et al.* Global latitudinal variations in marine copepod diversity and environmental factors. *Proceedings of the Royal Society B-Biological Sciences* **276,** 3053–3062 (2009).
65. Gallienne, C.P. & Robbins, B. D. Is Oithona the most important copepod in the world's oceans? *J. Plankton Res.* **23,** 1421–1432 (2001).
66. Stemmann, L. *et al.* Global zoogeography of fragile macrozooplankton in the upper 100–1000 m inferred from the underwater video profiler. *ICES Journal of Marine Science: Journal du Conseil* **65,** 433–442 (2008).
67. Moriarty, R., Buitenhuis, E. T., Le Quéré, C. & Gosselin, M.-P. Distribution of known macrozooplankton abundance and biomass in the global ocean. *Earth System Science Data* **5,** 241–257 (2013).
68. Bucklin, A., Steinke, D. & Blanco-Bercial, L. DNA Barcoding of Marine Metazoa. *Annu. Rev. Marine. Sci* **3,** 471–508 (2011).

Data Citations

1. Météo France, Tara Oceans Consortium, C., Tara Oceans Expedition, P. *PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.836312 (2014).
2. Boss, E., Tara Oceans Consortium, C. & Tara Oceans Expedition, P. *PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.836318 (2014).
3. Reverdin, G. & Le Goff, H., Tara Oceans Consortium, C. & Tara Oceans Expedition, P. *PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.836320 (2014).
4. Picheral, M. *et al. PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.836321 (2014).
5. Picheral, M. *et al. PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.836319 (2014).
6. Tara Oceans Consortium, C. & Tara Oceans Expedition, P. *PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.842191 (2015).
7. Tara Oceans Consortium, C. & Tara Oceans Expedition, P. *PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.842237 (2015).
8. Tara Oceans Consortium, C. & Tara Oceans Expedition, P. *PANGAEA* http://doi.pangaea.de/10.1594/PANGAEA.842227 (2015).

Acknowledgements

We thank the commitment of the following people and sponsors who made this singular expedition possible: CNRS (in particular the Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, Fund for Scientific Research—Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects POSEIDON/ANR-09-BLAN-0348, BIOMARKS/ANR-08-BDVA- 003, PROMETHEUS/ANR-09-GENM-031, PROMETHEUS/ANR-09-PCS-GENM-217, TARAGIRUS/ANR-09-PCS-GENM-218, OCEANOMICS/ANR-11-BTBR-0008, FRANCE GENOMIQUE/ANR-10-INBS-09-08), EU FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, MetaCardis/HEALTH-F4-2012-305312), ERC Advanced Grant Awards to CB (Diatomite: 294823) and PB (CancerBiome: 268985), Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to SGA, JSPS KAKENHI Grant Number 26430184 to HO, FWO, BIO5, Biosphere 2, agne`s b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the EDF Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner and its captain and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). Special thanks to Cornelia Behrens, Janine Felden, Florentina Münzner, Lucy Schlicht, Adrian Tanara, Sany Tchanra and Marie-Jeanne Pesant for the manual curation of logsheets and archiving data at PANGAEA. We also acknowledge the work of Andree Behnken who developed the dds-fdp web service. All authors approved the final manuscript. This article is contribution number 26 of the *Tara* Oceans Consortium.

The collection of *Tara* Oceans data was made possible by those who contributed to sampling and to logistics during the *Tara* Oceans Expedition: Alain Giese, Alan Deidun, Alban Lazar, Aldine Amiel, Ali Chase, Aline Tribollet, Ameer Abdullah, Amélie Betus, André Abreu, Andres Peyrot, Andrew Baker, Anna Deniaud, Anne Doye, Anne Ghuysen Watrin, Anne Royer, Anne Thompson, Annie McGrother, Antoine Sciandra, Antoine Triller, Aurélie Chambouvet, Baptiste Bernard, Baptiste Regnier, Beatriz Fernandez, Benedetto Barone, Bertrand Manzano, Bianca Silva, Brett Grant, Brigitte Sabard, Bruno Dunckel, Camille Clérissi, Catarina Marcolin, Cédric Guigand, Céline Bachelier, Céline Blanchard, Céline Dimier-Hugueney, Céline Rottier, Chris Bowler, Christian Rouvière, Christian Sardet, Christophe Boutte, Christophe Castagne, Claudie Marec, Claudie Marec, Claudio Stalder, Colomban De Vargas, Cornelia Maier, Cyril Tricot, Dana Sardet, Daniel Bayley, Daniel Cron, Daniele Iudicone, David Mountain, David Obura, David Sauveur, Defne Arslan, Denis Dausse, Denis de La Broise, Diana Ruiz Pino, Didier Zoccola, Édouard Leymarie, Éloïse Fontaine, Émilie Sauvage, Emilie Villar, Emmanuel Boss, Emmanuel G. Reynaud, Éric Béraud, Eric Karsenti, Eric Pelletier, Éric Roettinger, Erica Goetz, Fabien Perault, Fabiola Canard, Fabrice Not, Fabrizio D'Ortenzio, Fabrizio Limena, Floriane Desprez, Franck Prejger, François Aurat, François Noël, Franscisco Cornejo, Gabriel Gorsky, Gabriele Procaccini, Gabriella Gilkes, Gipsi Lima-Mendez, Grigor Obolensky, Guillaume Bracq, Guillem Salazar, Halldor Stefansson, Hélène Santener, Hervé Bourmaud, Hervé Le Goff, Hiroyuki Ogata, Hubert Gautier, Hugo Sarmento, Ian Probert, Isabel Ferrera, Isabelle Taupier-Letage, Jan Wengers, Jarred Swalwell, Javier del Campo, Jean-Baptiste Romagnan, Jean-Claude Gascard, Jean-Jacques Kerdraon, Jean-Louis Jamet, Jean-Michel Grisoni, Jennifer Gillette, Jérémie Capoulade, Jérôme Bastion, Jérôme Teigné, Joannie Ferland, Johan Decelle, Judith Prihoda, Julie Poulain, Julien Daniel, Julien Girardot, Juliette Chatelin, Lars Stemmann, Laurence Garczarek, Laurent Beguery, Lee Karp-Boss, Leila Tirichine, Linda Mollestan, Lionel Bigot, Loïc Vallette, Lucie Bittner, Lucie Subirana, Luis Gutiérrez, Lydiane Mattio, Magali Puiseux, Marc Domingos, Marc Picheral, Marc Wessner, Marcela Cornejo, Margaux Carmichael, Marion Lauters, Martin Hertau, Martina Sailerova, Mathilde Ménard, Matthieu Labaste, Matthieu Oriot, Matthieu Bretaud, Mattias Ormestad, Maya Dolan, Melissa Duhaime, Michael Pitiot, Mike Lunn, Mike Sieracki, Montse Coll, Myriam Thomas, Nadine Lebois, Nicole Poulton, Nigel Grimsley, Noan Le Bescot, Oleg Simakov, Olivier Broutin, Olivier Desprez, Olivier Jaillon, Olivier Marien, Olivier Poirot, Olivier Quesnel, Pamela Labbe-Ibanez, Pascal Hingamp, Pascal Morin, Pascale Joannot, Patrick Chang, Patrick Wincker, Paul Muir, Philippe Clais, Philippe Koubbi, Pierre Testor, Rachel Moreau, Raphaël Morard, Roland Heilig, Romain Troublé, Roxana Di Mauro, Roxanne Boonstra, Ruby Pillay, Sabrina Speich, Sacha Bollet, Samuel Audrain, Sandra Da Costa, Sarah Searson, Sasha Tozzi, Sébastien Colin, Sergey Pisarev, Shirley Falcone, Sibylle Le Barrois d'Orgeval, Silvia G. Acinas, Simon Morisset, Sophie Marinesque, Sophie Nicaud, Stefanie Kandels-Lewis, Stéphane Audic, Stephane Pesant, Stéphanie Reynaud, Thierry Mansir, Thomas

Lefort, Uros Krzic, Valérian Morzadec, Vincent Hilaire, Vincent Le Pennec, Vincent Taillandier, Xavier Bailly, Xavier Bougeard, Xavier Durrieu de Madron, Yann Chavance, Yann Depays, Yohann Mucherie.

Author Contributions

Contributed to writing this paper: Stéphane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis and Noan Le Bescot. Contributed to sampling planning during the Tara Oceans Expedition (2009–2013): Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Stefanie Kandels-Lewis, Fabrice Not, Stéphane Pesant, Sabrina Speich, Romain Troublé. Céline Dimier, Marc Picheral, and Sarah Searson contributed extensively to sampling during the *Tara* Oceans Expedition and refined the sampling methods on board. Stefanie Kandels-Lewis contributed as coordinator of scientific operations and logistics. Stéphane Pesant contributed as coordinator of data management. Eric Karsenti contributed as scientific director of the *Tara* Oceans Consortium. *Tara* Oceans Coordinators contributed intellectually to this work.

***Tara* Oceans Consortium Coordinators**

Silvia G. Acinas[15], Peer Bork[7], Emmanuel Boss[16], Chris Bowler[10], Colomban De Vargas[3,4], Michael Follows[17], Gabriel Gorsky[5,6], Nigel Grimsley[18,19], Pascal Hingamp[20], Daniele Iudicone[9], Olivier Jaillon[21,22,23], Stefanie Kandels-Lewis[7,8], Lee Karp-Boss[16], Eric Karsenti[8,10], Uros Krzic[24], Fabrice Not[3,4], Hiroyuki Ogata[25], Stéphane Pesant[1,2], Jeroen Raes[26,27,28], Emmanuel G. Reynaud[29], Christian Sardet[30], Mike Sieracki[31], Sabrina Speich[11,12], Lars Stemmann[5], Matthew B. Sullivan[32], Shinichi Sunagawa[7], Didier Velayoudon[33], Jean Weissenbach[21,22,23], Patrick Wincker[21,22,23].

[15]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, 08003, Barcelona, Spain.
[16]School of Marine Sciences, University of Maine, Orono, ME 04469, USA.
[17]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[18]CNRS UMR 7232, Biologie Intégrative des Organismes Marins (BIOM), 66650, Banyuls-sur-Mer, France.
[19]Sorbonne Universités, Observatoire Océanologique de Banyuls-sur-Mer (OOB), UPMC Paris 06, 66650, Banyuls-sur-Mer, France.
[20]Aix Marseille Université, CNRS, IGS UMR 7256, 13288 Marseille Cedex 09, France.
[21]CEA, Genoscope, 91000, Evry France.
[22]CNRS, UMR 8030, 91000, Evry, France.
[23]Université d'Evry, UMR 8030, 91000, Evry, France.
[24]Cell Biology and Biophysics, European Molecular Biology Laboratory, 69117, Heidelberg, Germany.
[25]Institute for Chemical Research, Kyoto University, 611-0011, Kyoto, Japan.
[26]Department of Microbiology and Immunology, Rega Institute KU Leuven, 3000, Leuven, Belgium.
[27]VIB Center for the Biology of Disease, VIB, 3000, Leuven, Belgium.
[28]Laboratory of Microbiology, Vrije Universiteit Brussel, 1050, Brussels, Belgium.
[29]School of Biology and Environmental Science, University College Dublin, Dublin 4, Ireland.
[30]CNRS, UMR 7009, BioDev, Observatoire Océanologique de Villefranche-sur-Mer (OOV), 06230, Villefranche/mer, France.
[31]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA.
[32]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85719, USA.
[33]DVIP Consulting, 92310, Sèvres, France.

Additional Information

Table 1 is only available in the online version of this paper.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: 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).