Title: Patterns and ecological drivers of ocean viral communities

Authors: Jennifer R. Brum[§,1], J. Cesar Ignacio-Espinoza[§,2], Simon Roux[§,1], Guilhem Doulcier[1,3], Silvia G. Acinas[4], Adriana Alberti[5], Samuel Chaffron[6,7,8], Corinne Cruaud[5], Colomban de Vargas[9,10], Josep M. Gasol[4], Gabriel Gorsky[11,12], Ann C. Gregory[13], Lionel Guidi[11,12], Pascal Hingamp[14], Daniele Iudicone[15], Fabrice Not[9,10], Hiroyuki Ogata[16], Stephane Pesant[17,18], Bonnie T. Poulos[1], Sarah M. Schwenck[1], Sabrina Speich[19,†], Celine Dimier[9,10,20], Stefanie Kandels-Lewis[21,22], Marc Picheral[11,12], Sarah Searson[11,12], *Tara* Oceans Coordinators[‡], Peer Bork[21,23], Chris Bowler[20], Shinichi Sunagawa[21], Patrick Wincker[5,24,25], Eric Karsenti[20,22,*], & Matthew B. Sullivan[1,2,13,*]

[§]co-first authors

Affiliations:

[1] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, 85721, USA

[2] Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, 85721, USA

[3] Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS, UMR8197, INSERM U1024, 75230 Paris, France

[4] Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona, E08003, Spain

[5] CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057, Evry, France

[6] Department of Microbiology and Immunology, Rega Institute KU Leuven, Herestraat 49, 3000 Leuven, Belgium

[7] Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium

[8] Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[9] CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[10] Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[11] CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France

[12] Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-mer, France

[13] Soil, Water, and Environmental Science, University of Arizona, Tucson, Arizona, 85721, USA

[14] Aix Marseille Université CNRS IGS UMR 7256 13288, Marseille, France

[15] Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy

[16] Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan

[17] PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany

[18] MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

[19] Laboratoire de Physique des Océan UBO-IUEM Palce Copernic 29820 Polouzané, France

[20] Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005, France

[21] Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

[22] Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg, Germany

[23] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany

[24] CNRS, UMR 8030, CP5706, Evry, France

[25] Université d'Evry, UMR 8030, CP5706, Evry, France

[†] Current address: Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris, Cedex 05, France

[‡] *Tara* Oceans coordinators and affiliations are listed at following the Acknowlegments.

*Correspondence to: mbsulli@gmail.com, karsenti@embl.de

Abstract: Viruses influence ecosystems by modulating microbial population size, diversity, metabolic outputs, and gene flow. Here we use quantitative double-stranded DNA (dsDNA) viral-fraction metagenomes (viromes) and whole viral community morphological datasets from 43 *Tara* Oceans expedition samples to assess viral community patterns and structure in the upper ocean. Protein cluster cataloging defined pelagic upper-ocean viral community pan and core gene sets and suggested this sequence space is well-sampled. Analyses of viral protein clusters, populations, and morphology revealed biogeographic patterns whereby viral communities were passively transported on oceanic currents and locally structured by environmental conditions that impact host community structure. Together these investigations establish a global ocean dsDNA viromic dataset with analyses supporting the seed-bank hypothesis to explain how oceanic viral communities maintain high local diversity.

One Sentence Summary: Global patterns that emerge from the *Tara* Oceans Virome dataset support a seed-bank structure underlying observed biogeography in ocean viral communities.

Main Text: Ocean microbes produce half of the oxygen we breathe (*1*) and drive much of the substrate and redox transformations that fuel Earth's ecosystems (*2*). However, they do so in a constantly evolving network of chemical, physical and biotic constraints – interactions which are only beginning to be explored. Marine viruses are presumably key players in these interactions (*3, 4*) as they affect microbial populations through lysis, reprogramming of host metabolism, and horizontal gene transfer. Here we strive to develop an overview of ocean viral community patterns and ecological drivers.

The *Tara* Oceans expedition provided a platform for sampling ocean biota from viruses to fish larvae with comprehensive environmental context (*5*). Prior virus-focused work from this expedition has helped optimize the dsDNA viromic sample-to-sequence workflow (*6*), evaluate ecological drivers of viral community structure as inferred from morphology (*7*), and map ecological patterns in the large dsDNA nucleo-cytoplasmic viruses using marker genes (*8*). Here we explore global patterns and structure of ocean viral communities using 43 samples from 26 stations in the *Tara* Oceans expedition (Supplementary File S1) to establish dsDNA viromes from viral-fraction (<0.22 μm) concentrates and quantitative whole viral community morphological datasets from unfiltered seawater. Viruses lack shared genes that can be used for investigation of community patterns. Therefore, we used three levels of information to study such patterns: (*i*) protein clusters (PCs, *9*) as a means to organize virome sequence space commonly dominated by unknown sequences (63–93%, *10*), (*ii*) populations, using established metrics for viral contig recruitment (*11*), and (*iii*) morphology, using quantitative transmission electron microscopy (qTEM, *7*).

The *Tara* Oceans Viromes (TOV) dataset

The 43 *Tara* Oceans Viromes (TOV) dataset is comprised of 2.16 billion ~101-bp paired-end Illumina reads (Supplementary File 1), largely representing epipelagic ocean viral communities (only 1 of 43 viromes are from mesopelagic waters, Environment Ontology feature ENVO:00000213) from the surface (ENVO:00002042) and deep chlorophyll maximum (DCM;

ENVO:01000326) throughout seven oceans and seas (Supplementary File S1). The TOV dataset offers deeper sampling of surface ocean viral communities, but under-represents the deep ocean relative to the Pacific Ocean Viromes dataset (POV, *10*) which includes 16 viromes from aphotic zone waters. In all viromes, sampling and processing affects what viruses are represented (*6*, *12-14*). We filtered TOV seawater samples through 0.22 µm pore-sized filters and then concentrated viruses in the filtrate using iron chloride flocculation (*15*). These steps would have removed most cells, but also excluded any viruses larger than 0.22 µm. We then purified the resulting TOV viral concentrates using DNase treatment, which is as effective as density gradients for purifying ocean viral concentrates (*14*). This DNAse-only step is unlikely to impact viral representation in the viromes, but reduces non-viral DNA contamination. Finally, we extracted DNA from the samples and prepared sequence libraries using linker amplification (*13*). These steps preserve quantitative representation of dsDNA viruses in the resulting viromes (*12, 13*), but the ligation step excludes RNA viruses, and is biased against single-stranded DNA (ssDNA) viruses (*12*).

We additionally applied qTEM (*7*) to paired whole seawater samples to evaluate patterns in whole viral communities. This method simultaneously considers ssDNA, dsDNA, and RNA viruses, though without knowledge of their relative abundances since particle morphology does not identify nucleic acid type. In the oceans, total virus abundance estimates based on TEM analyses, which include all viral particles, are similar to estimates based on fluorescent staining, which inefficiently stains ssDNA and RNA viruses (*16-24*). This suggests that most ocean viruses are dsDNA viruses. However, one study quantifying nucleic acids at a single marine location suggests RNA viruses may constitute as much as half of the viral community there (*16*). It remains unknown what the relative contribution of these viral types is to the whole viral community, but our analyses suggest small dsDNA viruses likely dominate as follows. The viromes capture the <0.22 µm dsDNA viruses of bacteria and archaea that are thought to dominate marine viral communities, whereas qTEM analysis includes all viruses regardless of size, nucleic acid type, or host (*7*). In these whole seawater samples used for qTEM, we found that viral capsid diameters ranged from 26 to 129 nm, with the per-sample average capsid diameter constrained at 46–66 nm (Fig. 1). We detected no viral particles larger than 0.22 µm among 100 randomly counted particles from each of 41 qTEM samples. These findings are similar to those from a subset of these *Tara* Oceans stations (14 of the 26 stations; *7*), and indicate that size fractionation using 0.22 µm filtration to prepare viromes did not substantially bias the TOV dataset.

TOV Protein Clusters for Comparison of Local and Global Genetic Richness and Diversity

Across the 43 viromes, a total of 1,075,763 PCs were observed, with samples beyond the 20th virome adding few PCs (Fig. 2A). When combining TOV with 16 photic-zone viromes from the POV dataset (*10*), the number of PCs increased to 1,323,921, but again approached a plateau (Fig. 2B). These results suggest that, while impossible to sample completely, the sequence space corresponding to dsDNA viruses from the epipelagic ocean is now relatively well sampled. This contrasts results from marine microbial metagenomic surveys using older sequencing

technologies (*9*), but is consistent with those from this expedition (*25*), as well as findings from viral sequence datasets which suggest a limited range of functional diversity derived from bacterial and archaeal viral isolates (*26*) and the POV dataset (*27*).

PCs were next used to establish the core genes shared across the TOV dataset (Fig. 2A). Broadly, there were 220, 710 and 424 core PCs shared across all surface and DCM viromes, surface viromes only, and DCM viromes only, respectively. The number of core PCs in the upper-ocean TOV samples (220 PCs) was thus less than the number of photic-zone core PCs in POV (565 PCs; *28*), likely because the POV dataset includes only the Pacific Ocean while TOV includes samples from seven oceans and seas. However, the number of core PCs in the upper-ocean TOV samples exceeded the total number of core PCs observed in POV (180 PCs; *28*), likely because of deep-ocean representation in POV (half of the samples in POV are from the aphotic zone). Consistent with the latter, the addition of the sole deep-ocean TOV sample, TARA_70_MESO, decreased the number of core PCs shared by all TOV samples from 220 to 65, which suggests that deep-ocean viral genetic repertoires are different from those in the upper oceans. Indeed, niche-differentiation has been observed in viromes sampled across these oceanic zones in the POV dataset (*28*), and similar findings were observed in the microbial metagenomic counterparts from the *Tara* Oceans Expedition (*25*). Thus viral communities from the deep ocean remain poorly explored and appear to hold different gene sets from those in the epipelagic oceans.

Beyond core and pan metagenomic analyses, PCs also provide a metric for viral community diversity comparisons (Fig. 3A; Supplementary File S1) from which three trends emerge in the TOV dataset. First, high-latitude viromes (82_DCM and 85_DCM) were least diverse (Shannon's H' of 8.93 and 9.22 nats), consistent with patterns in marine macroorganisms (*29*) and epipelagic ocean bacteria (*25, 30*). Second, the remaining viromes had similar diversity (Shannon's H' between 9.47 and 10.55 nats) and evenness (Pielou's *J* from 0.85 to 0.91) indicating low dominance of any particular PCs (*31*). Third, local diversity was relatively similar to global diversity (local:global ratios of H' from 0.73 to 0.87), suggesting high dispersal of viral genes (*32*) across the sampled ocean viral communities.

TOV Viral Populations for Assessing Global Viral Community Structure

We next estimated abundances of the 5,476 dominant viral populations in TOV, which represented up to 14.5% of aligned reads in a sample and were defined by applying empirically-derived recruitment cut-offs from naturally-occurring T4-like cyanophages (*11*) to high-confidence contigs from bacterial and archaeal viruses (see Methods). Assigning viral populations based on virome data remains challenging (*11, 33*), but here assembly of large contigs (up to 100 kb) aided our ability to accomplish not only analyses at the gene-level using PCs, but also the genome-level using viral populations. Viral populations were rarely endemic to one station (15%), and instead were commonly observed across >4 stations (47%), and up to 24 of the 26 stations (Fig. 4 and Fig. 5A). Exceptional samples include those from the Benguela upwelling region (TARA_67_SUR) and high-latitude samples from the Antarctic Circumpolar

and Falklands currents (TARA_82_DCM and TARA_85_DCM, respectively). These samples were also divergent when assessing microbial communities (TARA_82_DCM and TARA_85_DCM displayed lower microbial genetic richness; (*25*)) and eukaryotic communities (TARA_67_SUR had specific and unique eukaryotic communities in all size fractions; *34*). While many viral populations were broadly distributed, they were much more abundant at the original location (origin inferred from longest contig assembled; see Methods) compared to alternate stations (Fig. 5B). Thus most populations were relatively widespread, but with variable sample-to-sample abundances. As was observed with PCs, diversity and evenness estimates based on viral populations were similar across all samples except for high-latitude samples (TARA_82_DCM and TARA_85_DCM) and one sample in the Red Sea (TARA_32_DCM) that displayed lower diversity (Fig. 3B; Supplementary File S1). Finally, local diversity was relatively similar to global diversity (local:global ratios of H' from 0.23 to 0.86, average 0.74, Supplementary File S1), reflecting the high dispersal of viruses as highlighted by PC analysis.

Only 39 of the 5,476 populations we identified could be affiliated to cultured viruses, reflecting the dearth of reference viral genomes in databases. These cultured viruses include those infecting the abundant and widespread hosts SAR11, SAR116, *Roseobacter*, *Prochlorococcus* and *Synechococcus* (Fig. 6). The most abundant and widespread viral populations observed in TOV lack cultured representatives (Fig. 6), which suggests that most upper ocean viruses remain to be characterized even though viruses from known dominant microbial hosts (*35*-*39*) have been cultured. Methods independent of cultivation, including viral tagging (*11*) and mining of microbial genomic datasets (*40, 41*), show promise to expand the number of available viral reference genomes (*33*).

Drivers of Global Viral Community Composition and Distribution

We next leveraged this global dataset to evaluate ecological drivers (including environmental variables, sample location, and microbial abundances; Supplementary File S1) of viral community structure using all three data types – morphology, populations, and PCs. These metrics revealed increasing resolution, respectively, and showed that viral community structure was influenced by region and/or environmental conditions (Table 1). We conducted the analysis of ecological drivers using all samples in this study as well as a sample subset that omitted samples with exceptional environmental conditions and divergent viral communities observed using PC and population analyses (see above; TARA_67_SUR, TARA_82_DCM, TARA_85_DCM, and TARA_70_MESO). Within the sample subset, oceanic viral communities varied significantly with Longhurst province, biome, latitude, temperature, oxygen concentration, and microbial concentrations (including total bacteria, *Synechococcus*, and *Prochlorococcus*). Viral communities were not structured by depth (surface vs DCM) except when considering PCs, likely reflecting minimal variation between samples in the epipelagic zone compared to that of globally-sourced samples, and higher resolution provided by PCs. Nutrients influenced viral community structure when considering the whole dataset, but were much less explanatory when the few high-nutrient samples were removed, except for the

influence of phosphate concentration on viral populations. Thus nutrient concentrations may influence viral community structure, but testing this hypothesis would require analysis of samples across a more continuous nutrient gradient.

Global-scale analyses of oceanic macro- (*29*) and micro-organisms (*30*) have been conducted, including a concurrent *Tara* Oceans study showing that temperature and oxygen influence microbial community structure (*25*). Environmental conditions have also been shown to affect global viral community morphological traits (*7*). Our TOV study is consistent with these earlier findings in that viral communities are influenced by temperature and oxygen concentration, but not chlorophyll concentration (Table 1). Biogeographic structuring of TOV viral communities based on the significant influence of latitude and Longhurst provinces is also consistent with the conclusion that geographic region influences community structure in Pacific Ocean viruses (*42*). While only PC analysis showed depth-based divergence, this likely reflects poor (n=1) deep sample representation in the TOV dataset as discussed above. Prior POV viral investigation and concurrent *Tara* Oceans microbial analysis, both of which have better deep-water representation, show stronger depth patterns whereby photic and aphotic zone communities diverge (*25, 28, 42*). Thus our results suggest biogeography of upper-ocean viral communities is structured by environmental conditions.

Since viruses require host organisms to replicate, viral community structure follows from environmental conditions shaping the host community, as observed in paired *Tara* Oceans microbial samples (*25*), which would then indirectly affect viral community composition. However, global distribution of viruses can also be directly influenced by environmental conditions, such as salinity, that affect their ability to infect their hosts (*43*). Additionally, the variable decay rates of cultivated viruses and whole viral communities (*44*) could also influence their distribution as viruses with lower inherent decay rates will persist for longer in the environment, and environments with more favorable conditions (such as fewer extracellular enzymes) will also contribute to increased viral persistence. Until methods to link viruses to their host cells in natural communities mature to the point of investigating this issue at larger scales (emerging possible methods reviewed by *33, 45*), analyses such as ours remain the only means to assess ecological drivers of viral community structure.

To further investigate how ocean viral communities are distributed throughout the oceans, we compared population abundances between neighboring samples to assess the net direction and magnitude of population exchange (Fig. 7, see Methods). These genomic signals revealed that population exchange between dsDNA viral communities was largely directed along major oceanic current systems (*46*). For example, the Agulhas current and subsequent ring formation (*47*) connects viral communities between the Indian and Atlantic Oceans, as also observed in planktonic communities from the *Tara* Oceans expedition (*48*), while increased connection between the high-latitude stations (TARA_82 and TARA_85) reflects their common origin at the divergence of the Falklands and Antarctic Circumpolar currents. Further, current strength (*46*) was generally related to the magnitude of inter-sample population exchange, as higher and lower exchange was observed, respectively, in stronger currents such as the Agulhas current, and

within the open ocean gyres or between land-restricted basins such as the Mediterranean and Red Seas. These findings suggest that the intensity of water mass movement, in addition to environmental conditions, may explain the degree to which viral populations cluster globally (Fig. 4). Beyond such current-driven biogeographic evidence, vertical viral transport from surface to DCM samples was also observed (Fig. 4). This is consistent with POV observations wherein deep-sea viromes include a modest influx of genetic material derived from surface-ocean viruses that are presumably transported on sinking particles (*28*). Exceptions include areas such as the Arabian Sea upwelling region, where increased mixing and upwelling likely exceed sinking within the upper ocean.

Our TOV results enabled evaluation of a hypothesis describing the structure of viral communities in the environment. Gene-marker-based studies targeting subsets of ocean viruses previously found high local and low global diversity (*49*), a pattern also recently observed genome-wide in natural cyanophage populations (*11*). To explain this, a seed-bank viral community structure has been invoked whereby high local genetic diversity can exist by drawing variation from a common and relatively limited global gene pool (*49*). Our results support this hypothesis regarding viral community structure. Ecological driver analyses suggests that such local 'seed' communities are influenced by environmental conditions, which directly impact their microbial hosts and then indirectly restructure viral communities. These seed communities then form the 'bank' in neighboring samples, presumably when passively transported by ocean currents as shown here through the population-level analyses of net viral movement between samples. This systematically-sampled, global dataset suggests large- and small-scale processes play roles in structuring viral communities and offers empirical grounding for the seed-bank hypothesis with regards to viral community distribution and structure.

Conclusions

Our large-scale dataset provides a picture of global upper-ocean viral communities in which we assessed patterns using multiple parameters including morphology, populations and PCs. Our data provide advanced and complementary views on viral community structure including non-marker-gene-based diversity estimates and broad application of population-based viral ecology. We affirm the seed-bank model for viruses, hypothesized nearly a decade ago (*49*), which explains how high local viral diversity can be consistent with limited global diversity (*11, 27*). The mechanism underlying this seed-bank population structure appears to be a local production of viruses under small-scale environmental constraints and passive dispersal with oceanic currents. Improving sequencing, assembly and experimental methods are transforming the investigation of viruses in nature (*33, 45*), and pave the way towards assessment of viral community structure and analysis of virus-host co-occurrence networks (*50*) without requiring marker genes (*51, 52*). Such experimental and analytical progress, coupled to sampling opportunities from the *Tara* Oceans expedition, are advancing viral ecology towards the quantitative science needed to model the nano- (viruses) and micro- (microbes) scale entities driving Earth's ecosystems.

Materials and Methods

Sample Collection

Forty-three samples were collected between November 2, 2009, and May 13, 2011, at 26 locations throughout the world's oceans (Supplementary File S1) through the *Tara* Oceans Expedition (*5*). These included samples from a range of depths (surface, deep chlorophyll maximum, and one mesopelagic sample) located in 7 oceans and seas, 4 different biomes and 11 Longhurst oceanographic provinces (Supplementary File S1). Longhurst provinces and biomes are defined based on Longhurst (*53*) and environmental features are defined based on Environment Ontology (http://environmentontology.org/). Sampling strategy and methodology for the *Tara* Oceans Expedition is fully described by Pesant *et al.* (*54*).

Environmental Parameters

Temperature, salinity, and oxygen data were collected from each station using a CTD (Sea-Bird Electronics, Bellevue, WA, USA; SBE 911plus with Searam recorder) and dissolved oxygen sensor (Sea-Bird Electronics; SBE 43). Nutrient concentrations were determined using segmented flow analysis (*55*) and included nitrite, phosphate, nitrite plus nitrate, and silica. Nutrient concentrations below the detection limit (0.02 µmol kg^{-1}) are reported as 0.02 µmol kg^{-1}. Chlorophyll concentrations were measured using HPLC (*56, 57*). These environmental parameters are available in Pangaea (www.pangaea.de) using the accession numbers in Supplementary File S1.

Microbial Abundances

Flow-cytometry was used to determine the concentration of *Synechococcus*, *Prochlorococcus*, total bacteria, low-DNA bacteria, high-DNA bacteria, and the percent of bacteria with high DNA in each sample (*58*).

Morphological Analysis of Viral Communities

Quantitative transmission electron microscopy (qTEM) was used to evaluate the capsid diameter distributions of viral communities as previously described (*7*). Briefly, preserved unfiltered samples (EM-grade glutaraldehyde; Sigma-Aldrich, St. Louis, MO, USA; 2% final concentration) were flash-frozen and stored at -80ºC until analysis. Viruses were deposited onto TEM grids using an air-driven ultracentrifuge (Airfuge CLS, Beckman Coulter, Brea, CA, USA), followed by positive staining of the deposited material with 2% uranyl acetate (Ted Pella, Redding, CA, USA). Samples were then examined using a transmission electron microscope (Philips CM12, FEI, Hilsboro, OR, USA) with 100 kV accelerating voltage. Micrographs of 100 viruses were collected per sample using a Macrofire Monochrome CCD camera (Optronics, Goleta, CA, USA) and analyzed using ImageJ software (US National Institutes of Health, Bethesda, MD, USA; *59*) to measure the capsid diameter. A subset (21) of the 41 samples presented here had previously been analyzed in a different study (*7*).

Virome Construction

For each sample, 20 L of seawater were 0.22 µm-filtered and viruses were concentrated from the filtrate using iron chloride flocculation (*15*) followed by storage at 4ºC. After resuspension in ascorbic-EDTA buffer (0.1 M EDTA, 0.2 M Mg, 0.2 M ascorbic acid, pH 6.0), viral particles were concentrated using Amicon Ultra 100 kDa centrifugal devices (Millipore), treated with DNase I (100U/mL) followed by the addition of 0.1 M EDTA and 0.1 M EGTA to halt enzyme activity, and extracted as previously described (*14*). Briefly, viral particle suspensions were treated with Wizard PCR Preps DNA Purification Resin (Promega, WI, USA) at a ratio of 0.5 mL sample to 1 mL resin, and eluted with TE buffer (10 mM Tris, pH 7.5, 1 mM EDTA) using Wizard Minicolumns. Extracted DNA was Covaris-sheared and size selected to 160–180 bp, followed by amplification and ligation per the standard Illumina protocol. Sequencing was done on a HiSeq 2000 system at the Genoscope facilities (Paris, France).

Quality Control of Reads and Assembly

Individual reads of 43 metagenomes were quality controlled using a combination of trimming and filtering as previously described (*60*). Briefly, bases were trimmed at the 5' end if the number of base calls for any base (A, T, G, C) diverged by more than two standard deviations from the average across all cycles. Conversely, bases were trimmed at the 3' end of reads if the quality score was <20. Finally, reads that were shorter than 95 bp or reads with a median quality score <20 were removed from further analyses. Assembly of reads was done using SOAPdenovo (*61*) where insert and k-mer size are calculated at run-time and are specific to each virome as implemented in the MOCAT pipeline (*62*). On average, 34.2% of the virome reads were included in the assembled contigs (min: 21.08%, max: 48.52%). Virome reads were deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession numbers reported in Supplementary File S1.

Protein Clustering

Open Reading Frames (ORFs) were predicted from all quality-controlled contigs using Prodigal (*63*) with default settings. Predicted ORFs were clustered based on sequence similarity as described previously (*9, 10*). Briefly, ORFs were initially mapped to existing clusters (POV, GOS and phage genomes), using cd-hit-2d ('-g 1 -n 4 -d 0 -T 24 -M 45000'; 60% percent identity and 80% coverage). Then the remaining, unmapped ORFs were self-clustered, using cd-hit with the same options as above. Only protein clusters (PCs) with more than two ORFs were considered *bona fide* and were used for subsequent analyses. To develop read counts per PC for statistical analyses, reads were mapped back to predicted ORFs in the contigs dataset using Mosaik with the following settings: "-a all -m all -hs 15 –minp 0.95 –mmp 0.05 -mhp 100 -act 20" (version 1.1.0021; http://bioinformatics.bc.edu/marthlab/Mosaik). Read counts to PCs were normalized by sequencing depth of each virome. Shannon diversity (H') was calculated from PC read counts using only PCs with more than two predicted ORFs. Observed richness is reported as

the total number of reads in each PC. Pielou's evenness (*J*) was calculated as the ratio of H'/H_{max}, where $H_{max} = \ln N$, and N = total number of observed PCs in a sample.

Analysis of Viral Populations

Considering the size of the entire dataset (3,821,756 assembled contigs), we decided to focus the analysis of viral populations using contigs most likely originating from bacterial or archaeal viruses. For this, we mined only the 22,912 contigs with more than 10 predicted genes (corresponding to an average of 6.41% of the assembled reads per sample, min: 1.29%, max: 14.52%), as the origin of contigs with only a few predicted genes can be spurious. First, we removed 6,706 contigs suspected of having originated from cellular genomes (*64*), whether due to free genomic DNA contamination or viral-encapsidation of cellular DNA (for example, in gene transfer agents or generalized transducing phages). These suspect cellular contigs were those containing no typical viral genes (such as virion-related genes including major capsid proteins and large subunits of the terminase) and displaying as many 'characterized genes' (such as genes with a significant similarity to a PFAM domain through Hmmsearch, *65*) as a typical cellular genome, whereas phage genomes are typically enriched in 'uncharacterized genes' (*40*). We also removed all contigs posited to originate from eukaryotic viruses. These were contigs that contained at least three predicted proteins with best BLAST hits to a eukaryotic virus, and more than half of the affiliated proteins were not associated to bacteriophages or archaeal viruses. Not surprisingly, given that eukaryotes are outnumbered by bacteria and archaea in the marine environment, this step removed only 142 contigs associated with eukaryotic viruses. From the remaining 16,124 contigs most likely to have originated from bacterial or archaeal viruses, the population study only used those longer than 10kb in size – a total of 6,322 contigs, which corresponded to an average of 4.04% of the assembled reads per sample, min: 0.98%, max: 9.97%).

These 6,322 contigs were then clustered into populations if they shared more than 80% of their genes at >95% nucleotide identity; a threshold derived from naturally-occurring T4-like cyanophages (*11*). This resulted in 5,476 'populations' from the 6,322 contigs, where as many as 12 contigs (average 1.15 contigs) were included per population. For each population, the longest contig was chosen as the 'seed' sequence.

The relative abundance of each population was computed by mapping all quality-controlled reads to the set of 5,476 non-redundant populations (considering only mapping quality scores greater than 1) with Bowtie 2 (*66*). For each sample–sequence pair, if more than 75% of the reference sequence was covered by virome reads, the relative abundance was computed as the number of base pairs recruited to the contig normalized to the total number of base pairs available in the virome and the contig length. Shannon diversity index (H') and Pielou's evenness (*J*) were calculated as done for PCs using the relative abundance of viral populations.

The sample containing the seed sequence (the longest contig in a population) was also considered the best estimate of that population's origin. We reasoned this was because the longest contig in a population would derive most often from the sample with the highest

coverage (a metric for population abundance) and likely corresponded to the location with the greatest viral abundance for this population. This assumption was supported by the results showing that populations were most abundant in their original samples (Fig. 4, Fig. 5B). Even though some individual cases could diverge from this rule, we expected to correctly identify most of these original locations, hence getting an accurate global signal.

The seed sequence was also used to assess taxonomic affiliation of the viral population. Cases where >50% of the genes were affiliated to a specific reference genome from RefSeq (based on a BLASTp comparison with thresholds of 50 for bit score and 10^{-5} for e-value) with an identity percentage of at least 75% (at the protein sequence level) were considered as confident affiliations to the corresponding reference virus.

Finally, estimations of net viral population movement between samples were made based on the relative abundance of populations in one sample compared to that of its neighboring samples (Fig. 4). For each neighboring sample pair, the average relative abundance of populations originating from sample A in sample B was compared with the relative abundance of populations originating from sample B in sample A. The origin of each population was defined as the sample in which the longest contig of the population was assembled. The magnitude of these differences was carried through the analysis to estimate the level of transport between each pair of samples (depicted as line width in Fig. 7) and the difference between these values was used to estimate the directionality of the transfer. For example, if sample B contains many populations from sample A, but very few populations from sample B are detected in sample A, we calculate that the net movement is from sample A to sample B. Again, while the sampling of some populations may not be strong, the net movement was calculated as the average of all shared populations between neighboring sample pairs, which corresponded to 105 different populations on average (ranging from 2 to 412).

Statistical Ordination of Samples

Viral community composition based on capsid diameter distributions (from qTEM; using 7-nm histogram bin sizes), population abundances, and normalized PC read counts (using only protein clusters with more than 20 representatives) were compared using non-metric multidimensional scaling (NMDS) performed using the 'metaMDS' function (default parameters) of the vegan package (*67*) in R version 2.15.2 (*68*). The influence of metadata on sample ordination was evaluated using the functions 'envfit' for factor variables including depth category, Longhurst province, and biome, and 'ordisurf' for all linear variables, in the vegan package (*67*, *69*). Several samples had exceptional environmental conditions (TARA_67_SUR, TARA_70_MESO, TARA_82_DCM, and TARA_85_DCM), thus all statistical ordination analyses were conducted with and without these samples (referred to as the 'sample subset') to evaluate their influence.

References

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237-240 (1998).

2. P. G. Falkowski, T. Fenchel, E. F. Delong, The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034-1039 (2008).
3. M. Breitbart, Marine viruses: Truth or dare. *Ann. Rev. Mar. Sci.* **4**, 425-448 (2012).
4. C. A. Suttle, Marine viruses - major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801-812 (2007).
5. E. Karsenti, *et al.*, A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
6. S. Solonenko, *et al.*, Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**, 320 (2013).
7. J. R. Brum, R. O. Schenck, M. B. Sullivan, Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* **7**, 1738-1751 (2013).
8. P. Hingamp, *et al.*, Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678-1695 (2013).
9. S. Yooseph, *et al.*, The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5**, 0432-0466 (2007).
10. B. L. Hurwitz, M. B. Sullivan, The Pacific Ocean Virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE* **8**, e57355 (2013).
11. L. Deng, *et al.*, Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513**, 242 (2014).
12. M. B. Duhaime, M. B. Sullivan, Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434**, 181-186 (2012).
13. M. B. D. Duhaime, L. Deng, B. T. Poulos, M. B. Sullivan, Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* **14**, 2526-2537 (2012).
14. B. L. Hurwitz, L. Deng, B. T. Poulos, M. B. Sullivan, Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15**, 1428-1440 (2013).
15. S. G. John, *et al.*, A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195-202 (2011).
16. G. F. Steward, *et al.*, Are we missing half of the viruses in the ocean? *ISME J.* **7**, 672-679 (2013).
17. K. Holmfeldt, D. Odic, M. B. Sullivan, M. Middelboe, L. Riemann, Cultivated single stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA binding stains. *Appl. Environ. Microbiol.* **78**, 892-894 (2012).
18. Y. Tomaru, K. Nagasaki, Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae. *J. Oceanogr.* **63**, 215-221 (2007).
19. C. P. D. Brussaard, D. Marie, G. Bratbak, Flow cytometric detection of viruses. *J. Virol. Methods* **85**, 175-182 (2000).
20. Y. Bettarel, T. Sime-Ngando, C. Amblard, H. Laveran, A comparison of methods for counting viruses in aquatic systems. *Appl. Environ. Microbiol.* **66**, 2283-2289 (2000).
21. K. P. Hennes, C. A. Suttle, Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. *Limnol. Oceanogr.* **40**, 1050-1055 (1995).

22. M. G. Weinbauer, C. A. Suttle, Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. *Aquat. Microb. Ecol.* **13**, 225-232 (1997).
23. R. T. Noble, J. A. Fuhrman, Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb. Ecol.* **14**, 113-118 (1998).
24. D. Marie, C. P. D. Brussaard, R. Thyrhaug, G. Bratbak, D. Vaulot, Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl. Environ. Microbiol.* **65**, 45-52 (1999).
25. S. Sunagawa, Structure and function of the global ocean microbiome. (in review).
26. D. M. Kristensen, *et al.*, Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* **195**, 941-950 (2013).
27. C. J. Ignacio-Espinoza, S. A. Solonenko, M. B. Sullivan, The global virome: not as big as we thought? *Curr. Opin. Virol.* **3**, 566-571 (2013).
28. B. L. Hurwitz, J. R. Brum, M. B. Sullivan, Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* **9**, 472-484 (2015).
29. D. P. Tittensor, *et al.*, Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098-1101 (2010).
30. W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, M. L. Sogin, Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2342-2347 (2013).
31. D. I. Jarvis, *et al.*, A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5326-5331 (2008).
32. S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press, Princeton, NJ, 2001).
33. J. R. Brum, M. B. Sullivan, Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro3404 (2015).
34. C. de Vargas, *et al.*, Global oceans eukaryotic plankton diversity. (in review).
35. I. Kang, H.-M. Oh, D. Kang, J.-C. Cho, Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12343-12348 (2013).
36. S. J. Labrie, *et al.*, Genomes of marine cyanopodoviruses reveal multiple origins of diversity. *Environ. Microbiol.* **15**, 1356-1376 (2013).
37. M. B. Sullivan, *et al.*, Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**, 3035-3056 (2010).
38. Y. Zhao, *et al.*, Abundant SAR11 viruses in the ocean. *Nature* **494**, 357-360 (2013).
39. F. Rohwer, *et al.*, The complete genomic sequence of the marine phage Roseaphage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45**, 408-418 (2000).
40. S. Roux, *et al.*, Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics. *eLife* 10.7554/eLife.03125 (2014).
41. C. M. Mizuno, F. Rodriguez-Valera, N. E. Kimes, R. Ghai, Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
42. B. Hurwitz, A. Westvald, J. Brum, M. Sullivan, Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10714-10719 (2014).

43. P. Kukkaro, D. H. Bamford, Virus-host interactions in environments with a wide range of ionic strengths. *Environ. Microbiol. Rep.* **1**, 71-77 (2009).
44. K. E. Wommack, R. R. Colwell, Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69-114 (2000).
45. V. Dang, M. B. Sullivan, Emerging methods to study bacteriophage infection at the single-cell level. *Front. Microbiol.* (in press).
46. L. D. Talley, G. L. Pickard, W. J. Emery, J. H. Swift, *Descriptive Physical Oceanography: An Introduction (Sixth Edition)* (Elsevier, Boston, 2011).
47. D. B. Olson, R. H. Evans, Rings of the Agulhas current. *Deep Sea Res. A* **33**, 27-42 (1986).
48. E. Villar, Dispersal and remodeling of plankton communities by Agulhas rings. (in review).
49. M. Breitbart, F. Rohwer, Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278-284 (2005).
50. G. Lima-Mendez, *et al.*, Top-down determinants of ocean microbial community structure. (in review).
51. D. M. Needham, *et al.*, Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J.* **7**, 1274-1285 (2013).
52. C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, J. A. Fuhrman, Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8**, 816-829 (2014).
53. A. Longhurst, *Ecological Geography of the Sea* (Elsevier, Inc., London, 2007).
54. S. Pesant, *et al.*, *Tara* Oceans Data: A sampling strategy and methodology for the study of marine plankton in their environmental context. (in review).
55. A. Aminot, R. Kerouel, S. C. Coverly, in *Practical Guidelines for the Analysis of Seawater*, O. Wurl, Eds (CRC Press, Boca Raton, 2009), vol. pp. 143-178.
56. J. Ras, H. Claustre, J. Uitz, Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: comparison between *in situ* and predicted data. *Biogeosciences* **5**, 353-369 (2008).
57. L. Van Heukelem, C. S. Thomas, Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *J. Chromatogr. A* **910**, 31-49 (2001).
58. J. M. Gasol, P. A. del Giorgio, Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci. Mar.* **64**, 197-224 (2000).
59. M. D. Abramoff, P. J. Magalhaes, S. J. Ram, Image processing with ImageJ. *Biophotonics International* **11**, 36-42 (2004).
60. S. Schloissnig, *et al.*, Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50 (2013).
61. R. Luo, *et al.*, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).
62. J. R. Kultima, *et al.*, MOCAT: A metagenomics assembly and gene prediction toolkit. *PLoS ONE* **7**, e47656 (2012).
63. D. Hyatt, *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
64. S. Roux, M. Krupovic, D. Debroas, P. Forterre, F. Enault, Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).

65. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-W37 (2011).
66. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
67. J. Oksanen, *et al.*, vegan: Community Ecology Package, R package version 2.1-27/r2451 (2013).
68. R Core Team, R: A language and environment for statistical computing. v. 2.15.2 (2012).
69. S. N. Wood, Fast stable restricted maximum likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. A* **73**, 3-36 (2011).

Acknowledgements. We thank Jesse Czekanski-Moir for advice on statistics and Laurent Coppola for assistance with validating nutrient data. We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders, Rega Institute, KU Leuven, The French Ministry of Research, the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376), ERC Advanced Grant Award to CB (Diatomite: 294823), Gordon and Betty Moore Foundation grant (#3790) to MBS, Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to SGA, TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to SGA, JSPS KAKENHI Grant Number 26430184 to HO, and FWO, BIO5, Biosphere 2 to MBS. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge excellent assistance from the European Bioinformatics Institute (EBI), in particular Guy Cochrane and Petra ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular Rainer Pepperkok. The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein is available at EBI (project identifiers PRJEB402 and PRJEB7988) and Pangaea (see Supplementary Table S1), and the data release policy regarding future public release of *Tara* Oceans data is described in Pesant et al (*54*). We also acknowledge the EMBL Advanced Light Microscopy Facility (ALMF), and in particular Rainer Pepperkok. This research is funded in part by the Gordon and Betty Moore Foundation through grants GBMF2631 and GBMF3790 to MBS. We also acknowledge support from UA high-performance computing; the foundation for France-American Cultural Exchange, Partner University Fund program, awarded to Ecole

Normale Supérieure and the University of Arizona; and a grant to the University of Arizona Ecosystem Genomics Institute through the UA Technology and Research Initiative Fund and the Water, Environmental and Energy Solutions Initiative. All authors approved the final manuscript. This article is contribution number XXX of the *Tara* Oceans Expedition. Supplement contains additional data.

Tara Oceans Coordinators

Silvia G. Acinas[1], Peer Bork[2,3], Emmanuel Boss[4], Chris Bowler[5], Colomban de Vargas[6,7], Michael Follows[8], Gabriel Gorsky[9,31], Nigel Grimsley[10,11], Pascal Hingamp[12], Daniele Iudicone[13], Olivier Jaillon[14,15,16], Stefanie Kandels-Lewis[2,17], Lee Karp-Boss[18], Eric Karsenti[5,17], Uros Krzic[19], Fabrice Not[6,7], Hiroyuki Ogata[20], Stephane Pesant[21,22], Jeroen Raes[23,24,25], Emmanuel G. Reynaud[26], Christian Sardet[27,28], Mike Sieracki[29,†], Sabrina Speich[30,‡], Lars Stemmann[9,31], Matthew B. Sullivan[32], Shinichi Sunagawa[2], Didier Velayoudon[33], Jean Weissenbach[14,15,16], Patrick Wincker[14,15,16]

[1] Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona, E08003, Spain

[2] Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

[3] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany

[4] School of Marine Sciences, University of Maine, Orono, Maine, USA

[5] Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France

[6] CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[7] Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[8] Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[9] CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France

[10] CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France

[11] Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France

[12] Aix Marseille Université CNRS IGS UMR 7256 13288 Marseille, France

[13] Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy

[14] CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France

[15] CNRS, UMR 8030, CP5706, Evry, France

[16] Université d'Evry, UMR 8030, CP5706, Evry, France

[17] Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg, Germany

[18] School of Marine Sciences, University of Maine, Orono, USA

[19] Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

[20] Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan

[21] PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany

[22] MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

[23] Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium

[24] Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium

[25] Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[26] Earth Institute, University College Dublin, Dublin, Ireland

[27] CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France

[28] Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-mer, France

[29] Bigelow Laboratory for Ocean Science, East Boothbay, Maine, USA

[30] Laboratoire de Physique des Océan UBO-IUEM Palce Copernic 29820 Polouzané, France

[31] Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-mer, France

[32] Department of Ecology and Evolutionary Biology, Depts Molecular and Cellular Biology and Soil, Water and Environmental Science, University of Arizona, Tucson, Arizona, 85721, USA

[33] DVIP Consulting, 92310, Sèvres, France

† Current address: National Science Foundation, Arlington, Virginia, USA

‡ Current address: Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris, Cedex 05, France

Supplementary Information

Supplementary File S1. Description of samples and relevant virome data. Metadata is presented for each *Tara* Oceans sample in this study including the PANGAEA accession numbers, sample location and environmental conditions, and the abundances of selected microorganisms. Detailed information is also presented for the viromes in this study including ENA accession numbers, the total number of reads and PCs for each virome, and diversity and evenness data for each virome based on PCs and viral populations.

Figure Legends

Fig. 1. Distribution of viral capsid diameters in each sample (n = 100 viruses per sample). Data are not available for samples TARA_18_DCM and TARA_70_MESO. Boxplots are constructed with the upper and lower lines corresponding to the 25th and 75th percentiles, while outliers are displayed as points. Longhurst provinces are indicated below samples (MEDI, Mediterranean Sea; REDS, Red Sea; ARAB, NW Arabian Upwelling; MONS, Indian Monsoon Gyres; ISSG, Indian S. Subtropical Gyre; EAFR, E. Africa Coastal; BENG, Benguela Current Coastal; SATL, S. Atlantic Gyre; FKLD, SW Atlantic Shelves; APLR, Austral Polar; PNEC, N. Pacific Equatorial Countercurrent).

Fig. 2. Protein cluster (PC) richness in core and pan viromes from the TOV and POV datasets. A) Accumulation curves of core and pan PCs in the TOV dataset. Vertical axis shows the number of shared (core virome) and total (pan virome) PCs when *n* viromes are compared (*n* = 1 to 43; from 3 to 41 only 1000 combinations are shown). Lines: i) total number of PCs (1,075,763 PCs), ii) core surface virome (710 PCs), iii) core DCM virome (424 PCs), iv) core surface and DCM virome (220 PCs), v) all samples (including the deep-ocean sample TARA_70_MESO; 65 PCs). B) Core and pan PCs in all TOV and photic-zone POV samples combined. Vertical axis shows the number of shared (core virome) and total (pan-virome) PCs when *n* viromes are compared (*n* = 1 to 57; from 3 to 57 only 1,000 combinations are shown). Overall, 1,323,921 PCs were identified in all viromes combined.

Fig. 3. Alpha diversity measurements in TOV dataset. A) Shannon's richness H' and Pielou's evenness J calculated from protein clusters counts for each sample and a pool of all samples, normalized to 5 million reads. B) Shannon's richness H' and Pielou's evenness J calculated from relative abundances of viral populations for each sample and a pool of all samples, with subsamples of 100,000 reads. Outliers corresponding to values outside of the average value plus or minus two standard deviations are colored in green and red, respectively. Values calculated

from the pool of all samples are colored in blue. Longhurst provinces are indicated below samples using the same abbreviations as in Fig. 1.

Fig. 4. Relative abundance of viral populations in TOV by sample. This heatmap displays the relative abundance of each population (sorted according to its original sample; y-axis) in each sample (x-axis). Relative abundance of one population in a sample is based on recruitment of reads to the population reference contig, and only considered if more than 75% of the reference contig is covered. Longhurst provinces are indicated below samples (using the same abbreviations as in Fig. 1) and outlined in black on the heatmap.

Fig. 5. Relative abundance of viral populations in TOV by station. A) Evaluation of viral population distribution showing the number of stations (y-axis) in which each population (sorted by their original station, x-axis) is distributed. Populations are grouped by station, merging surface and DCM samples from the same station. B) Relative abundance of populations at the original stations where the contigs were assembled compared to their abundance at other stations. Boxplots are constructed as in Fig. 1.

Fig. 6. Taxonomic affiliation of TOV viral populations sorted by distribution and average abundance. A population was considered as similar to a known virus when less than half of its reference contig genes were uncharacterized, and all characterized genes had taxonomic affiliations to the same reference genome. As in Fig. 4, the relative abundance (y-axis) is computed for each sample as the number of bp mapped to a contig per kb of contig per Mb of metagenome sequenced. Here, the relative abundance of a population is defined as the average abundance of its reference contig across all samples.

Fig. 7. Net movement of viral populations throughout the oceans. Calculations are based on reciprocal comparison of viral population abundances between neighboring samples (see Fig. 3 and Methods). For each sample pair, the average relative population abundances in one sample originating from a neighboring sample were calculated and compared (for example, relative abundance of populations from sample A found in sample B are compared with relative abundance of populations from sample B found in sample A). The sign of the relative abundance difference between neighboring samples was used to estimate the movement direction (arrowhead), and the absolute value of the difference was interpreted as reflecting the movement magnitude (line width). Stations are labeled with station number. ‘Down’ and ‘up’ refer to net vertical movement of viral populations between the surface and DCM samples at the same station.

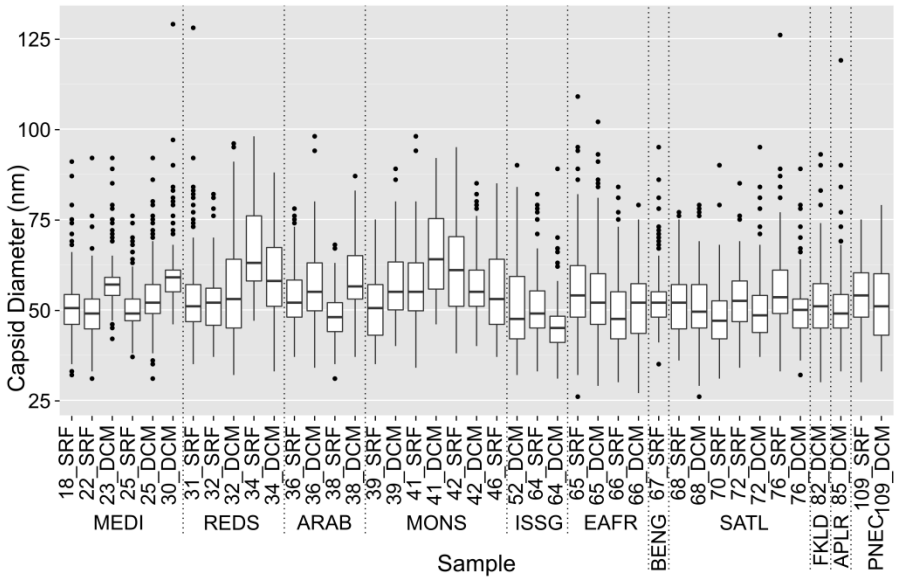

Fig. 1. Distribution of viral capsid diameters in each sample (n = 100 viruses per sample). Data are not available for samples TARA_18_DCM and TARA_70_MESO. Boxplots are constructed with the upper and lower lines corresponding to the 25th and 75th percentiles, while outliers are displayed as points. Longhurst provinces are indicated below samples (MEDI, Mediterranean Sea; REDS, Red Sea; ARAB, NW Arabian Upwelling; MONS, Indian Monsoon Gyres; ISSG, Indian S. Subtropical Gyre; EAFR, E. Africa Coastal; BENG, Benguela Current Coastal; SATL, S. Atlantic Gyre; FKLD, SW Atlantic Shelves; APLR, Austral Polar; PNEC, N. Pacific Equatorial Countercurrent).

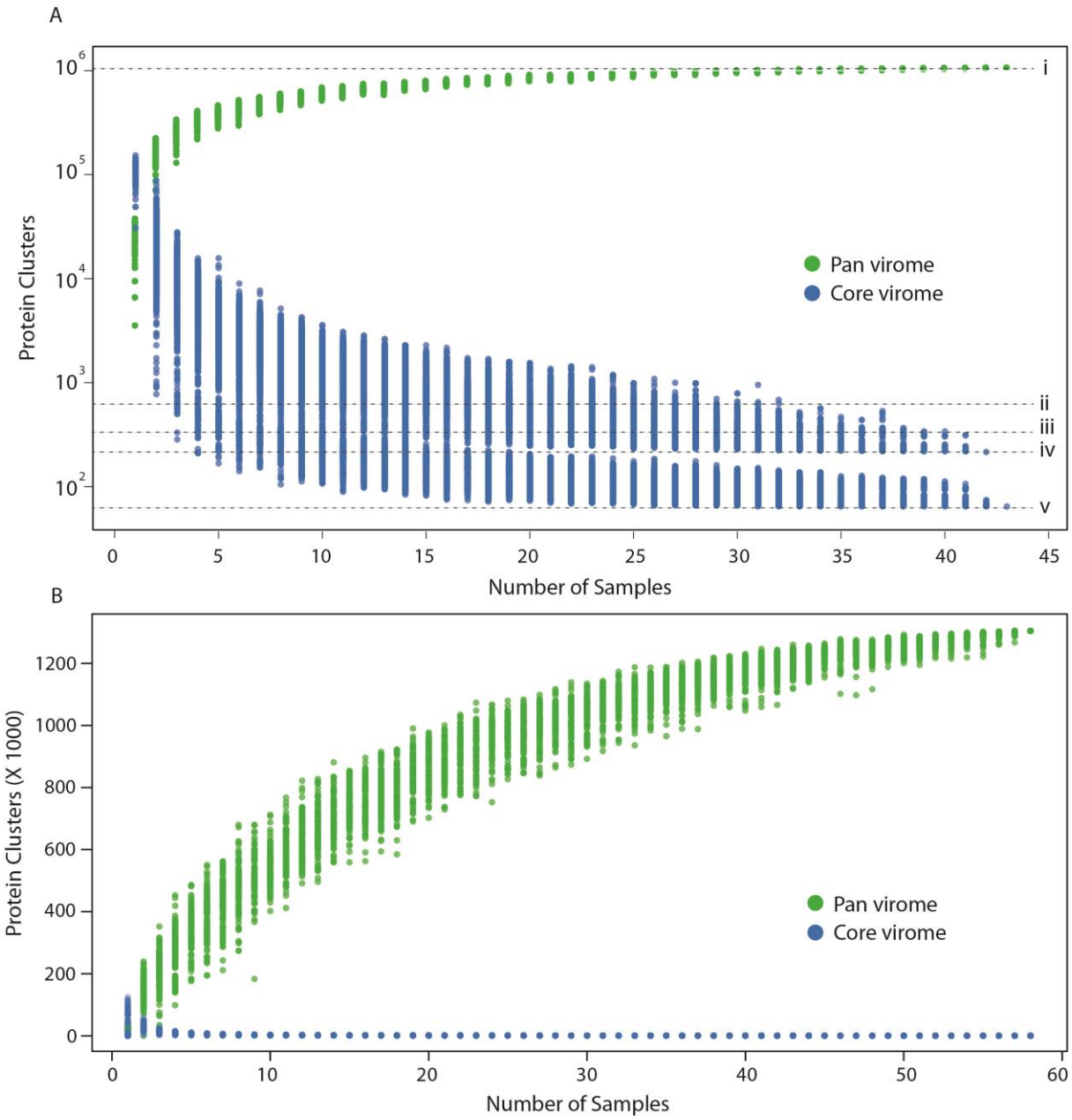

Fig. 2. Protein cluster (PC) richness in core and pan viromes from the TOV and POV datasets. A) Accumulation curves of core and pan PCs in the TOV dataset. Vertical axis shows the number of shared (core virome) and total (pan virome) PCs when *n* viromes are compared (*n* = 1 to 43; from 3 to 41 only 1000 combinations are shown). Lines: i) total number of PCs (1,075,763 PCs), ii) core surface virome (710 PCs), iii) core DCM virome (424 PCs), iv) core surface and DCM virome (220), v) all samples (including the deep-ocean sample TARA_70_MESO; 65 PCs). B) Core and pan PCs in all TOV and photic-zone POV samples combined. Vertical axis shows the number of shared (core virome) and total (pan-virome) PCs when *n* viromes are compared (*n* = 1 to 57; from 3 to 57 only 1,000 combinations are shown). Overall, 1,323,921 PCs were identified in all viromes combined.

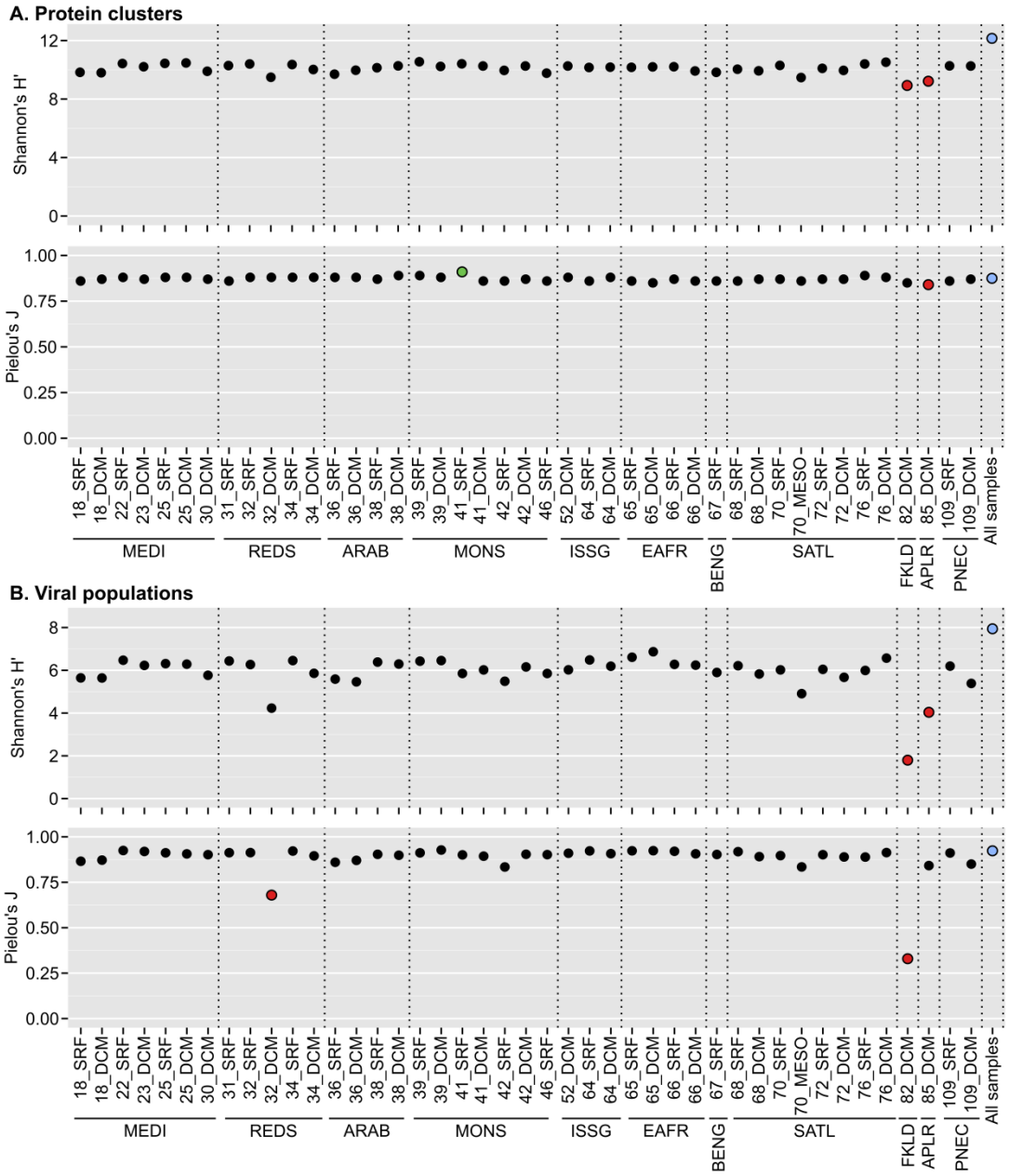

Fig. 3. Alpha diversity measurements in TOV dataset. A) Shannon's richness H' and Pielou's evenness J calculated from protein clusters counts for each sample and a pool of all samples, normalized to 5 million reads. B) Shannon's richness H' and Pielou's evenness J calculated from relative abundances of viral populations for each sample and a pool of all samples, with subsamples of 100,000 reads. Outliers corresponding to values outside of the average value plus or minus two standard deviations are colored in green and red, respectively. Values calculated from the pool of all samples are colored in blue. Longhurst provinces are indicated below samples using the same abbreviations as in Fig. 1.

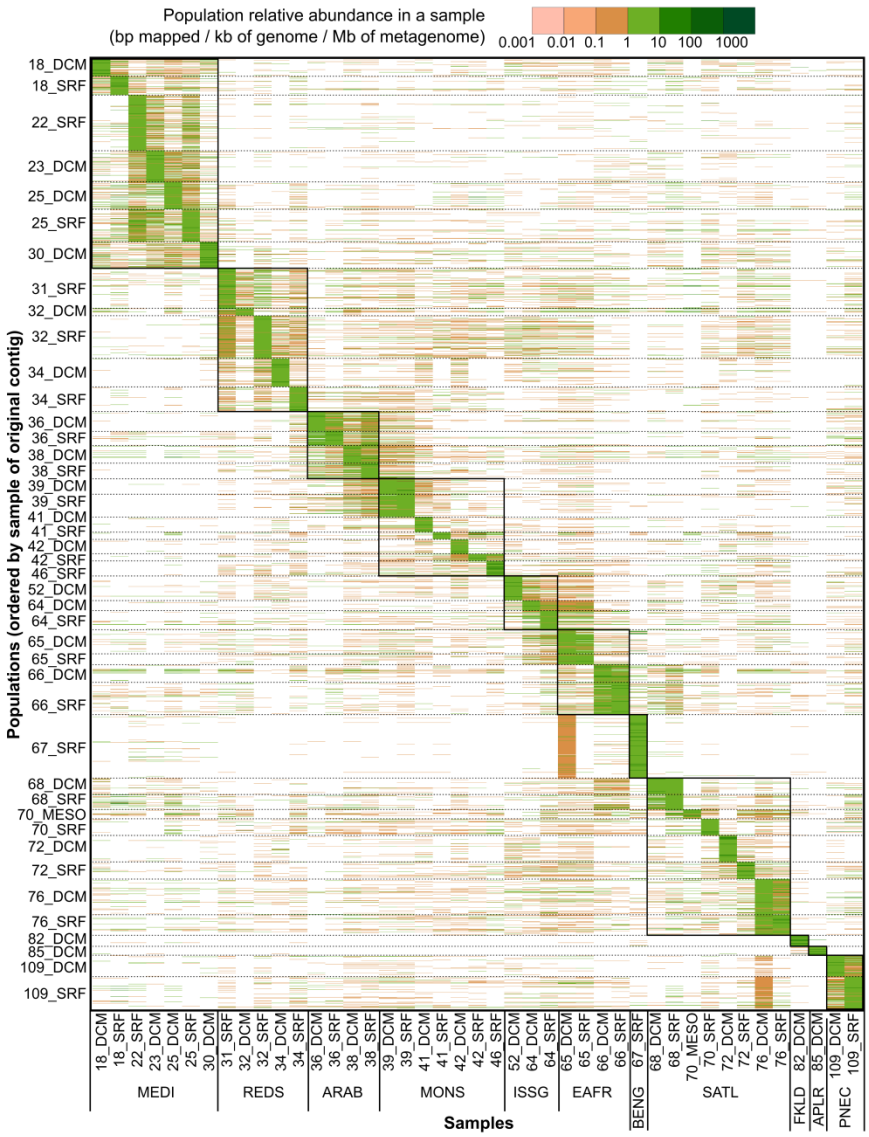

Fig. 4. Relative abundance of viral populations in TOV by sample. This heatmap displays the relative abundance of each population (sorted according to its original sample; y-axis) in each sample (x-axis). Relative abundance of one population in a sample is based on recruitment of reads to the population reference contig, and only considered if more than 75% of the reference contig is covered. Longhurst provinces are indicated below samples (using the same abbreviations as in Fig. 1) and outlined in black on the heatmap.

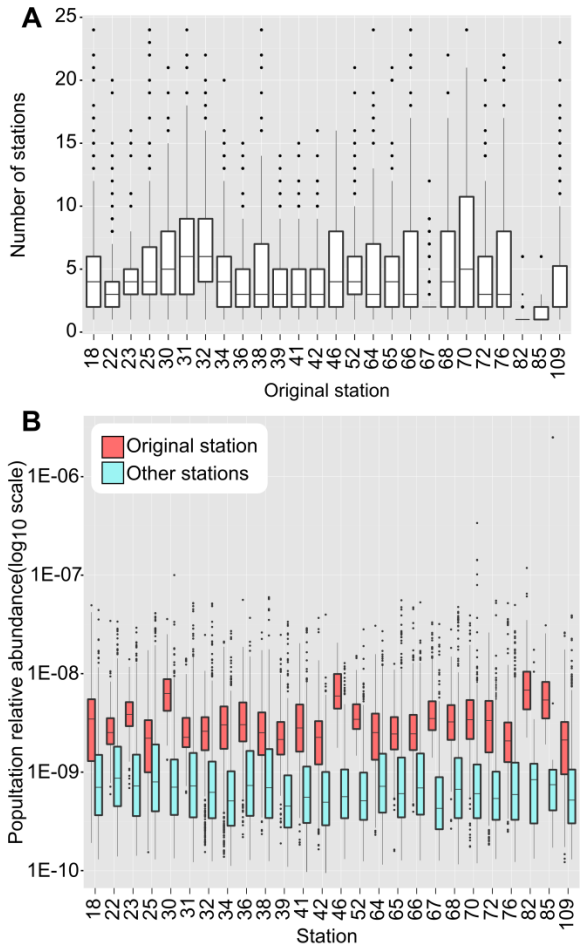

Fig. 5. Relative abundance of viral populations in TOV by station. A) Evaluation of viral population distribution showing the number of stations (y-axis) in which each population (sorted by their original station, x-axis) is distributed. Populations are grouped by station, merging surface and DCM samples from the same station. B) Relative abundance of populations at the original stations where the contigs were assembled compared to their abundance at other stations. Boxplots are constructed as in Fig. 1.

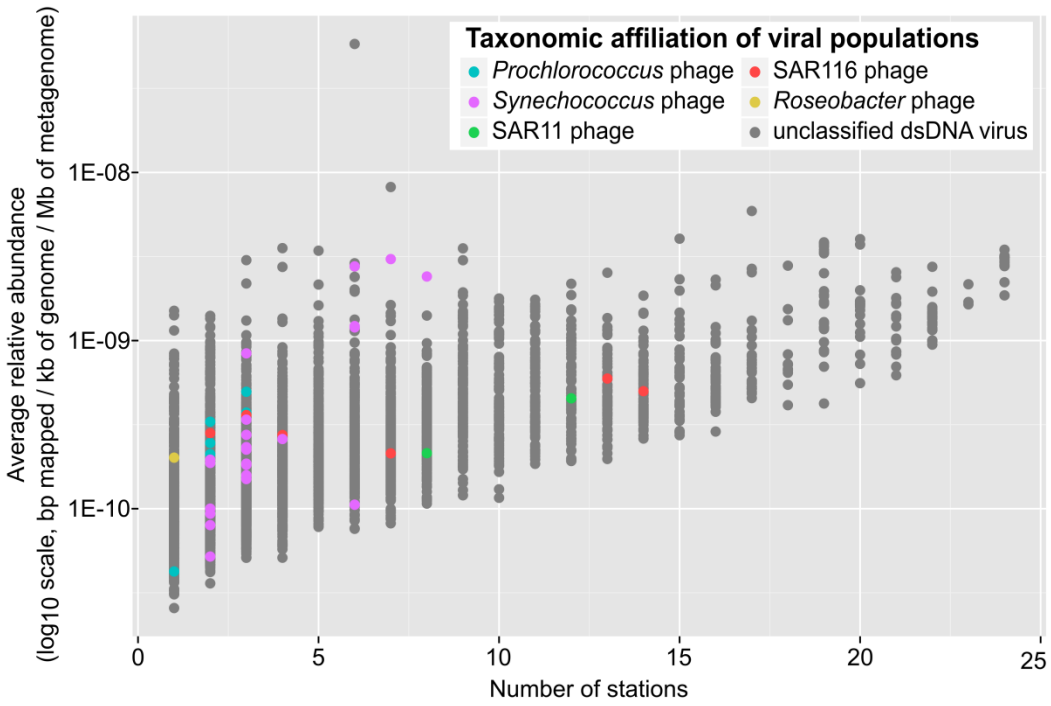

Fig. 6. Taxonomic affiliation of TOV viral populations sorted by distribution and average abundance. A population was considered as similar to a known virus when less than half of its reference contig genes were uncharacterized, and all characterized genes had taxonomic affiliations to the same reference genome. As in Fig. 4, the relative abundance (y-axis) is computed for each sample as the number of bp mapped to a contig per kb of contig per Mb of metagenome sequenced. Here, the relative abundance of a population is defined as the average abundance of its reference contig across all samples.

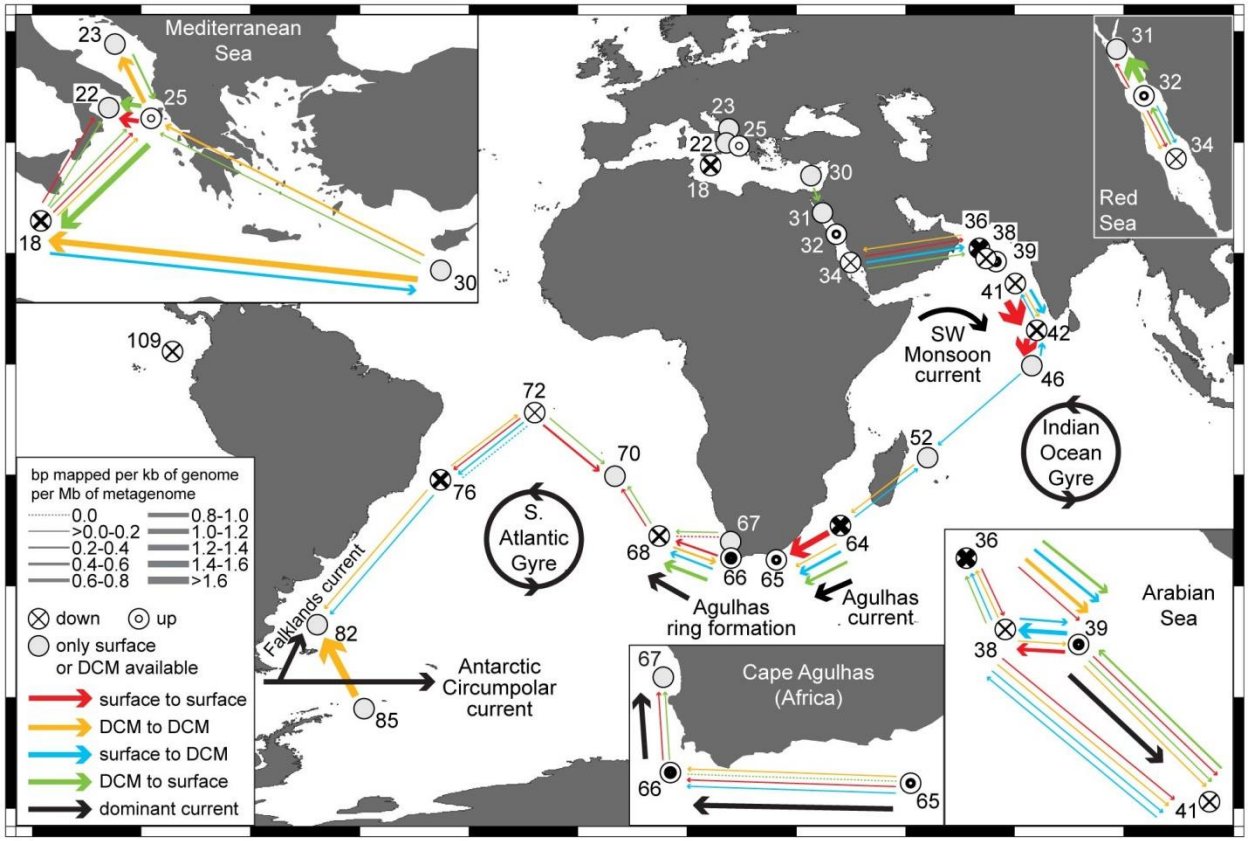

Fig. 7. Net movement of viral populations throughout the oceans. Calculations are based on reciprocal comparison of viral population abundances between neighboring samples (see Fig. 3 and Methods). For each sample pair, the average relative population abundances in one sample originating from a neighboring sample were calculated and compared (for example, relative abundance of populations from sample A found in sample B are compared with relative abundance of populations from sample B found in sample A). The sign of the relative abundance difference between neighboring samples was used to estimate the movement direction (arrowhead), and the absolute value of the difference was interpreted as reflecting the movement magnitude (line width). Stations are labeled with station number. 'Down' and 'up' refer to net vertical movement of viral populations between the surface and DCM samples at the same station.

Table 1. Relationships between viral community structure (based on viral morphology, populations, and PCs) and metadata using NMDS analysis of all samples and the sample subset (all samples except for TARA_67_SRF, TARA_70_MESO, TARA_82_DCM, and TARA_85_DCM due to exceptional environmental conditions at these locations). Significant relationships are italicized and in bold.

		Viral Morphology (qTEM)	Populations (contigs)	Protein Clusters (PCs)
Depth Category	all samples	p = 0.354 (n = 41)	p = 0.362 (n = 43)	***p = 0.033 (n = 43)***
	sample subset	p = 0.228 (n = 38)	p = 0.105 (n = 39)	***p = 0.011 (n = 39)***
Province	all samples	p = 0.098 (n = 41)	***p < 0.001 (n = 43)***	***p = 0.014 (n = 43)***
	sample subset	***p = 0.029 (n = 38)***	***p < 0.001 (n = 39)***	***p = 0.008 (n = 39)***
Biome	all samples	p = 0.099 (n = 41)	***p < 0.001 (n = 43)***	p = 0.097 (n = 43)
	sample subset	p = 0.120 (n = 38)	***p < 0.001 (n = 39)***	p = 0.543 (n = 39)
Latitude	all samples	***p = 0.003 (n = 41)***	***p < 0.001 (n = 43)***	***p = 0.002 (n = 43)***
	sample subset	***p = 0.014 (n = 38)***	***p < 0.001 (n = 39)***	***p = 0.010 (n = 39)***
Temperature	all samples	***p = 0.001 (n = 41)***	***p < 0.001 (n = 43)***	***p < 0.001 (n = 43)***
	sample subset	***p = 0.001 (n = 38)***	***p < 0.001 (n = 39)***	***p = 0.015 (n = 39)***
Salinity	all samples	p = 0.118 (n = 39)	***p = 0.035 (n = 41)***	***p = 0.029 (n = 41)***
	sample subset	p = 0.138 (n = 36)	p = 0.075 (n = 37)	***p = 0.001 (n = 37)***
Oxygen	all samples	***p = 0.001 (n = 41)***	***p < 0.001 (n = 43)***	***p < 0.001 (n = 43)***
	sample subset	***p = 0.005 (n = 38)***	***p < 0.001 (n = 39)***	***p < 0.001 (n = 39)***
Chlorophyll	all samples	p = 0.711 (n = 41)	***p < 0.001 (n = 43)***	***p = 0.001 (n = 39)***
	sample subset	p = 0.738 (n = 38)	p = 0.412 (n = 39)	p = 0.059 (n = 39)
Nitrite	all samples	p = 0.951 (n = 39)	p = 0.648 (n = 41)	p = 0.828 (n = 41)
	sample subset	p = 0.851 (n = 36)	p = 0.509 (n = 37)	p = 0.999 (n = 37)
Phosphate	all samples	p = 0.275 (n = 39)	***p < 0.001 (n = 41)***	***p < 0.001 (n = 41)***
	sample subset	p = 0.411 (n = 36)	***p < 0.001 (n = 37)***	p = 0.583 (n = 37)
Nitrite+Nitrate	all samples	***p = 0.046 (n = 39)***	***p < 0.001 (n = 41)***	***p < 0.001 (n = 41)***
	sample subset	p = 0.290 (n = 36)	p = 0.052 (n = 37)	p = 0.643 (n = 37)
Silica	all samples	***p = 0.008 (n = 39)***	***p = 0.002 (n = 41)***	***p = 0.008 (n = 41)***
	sample subset	p = 0.255 (n = 36)	p = 0.285 (n = 37)	p = 0.191 (n = 37)
Bacteria	all samples	p = 0.579 (n = 39)	***p < 0.001 (n = 40)***	p = 0.119 (n = 40)
	sample subset	p = 0.329 (n = 36)	***p = 0.003 (n = 36)***	***p = 0.007 (n = 36)***
Low DNA bacteria	all samples	p = 0.227 (n = 39)	p = 0.090 (n = 40)	p = 0.123 (n = 40)
	sample subset	p = 0.468 (n = 36)	***p = 0.018 (n = 36)***	***p = 0.005 (n = 36)***
High DNA bacteria	all samples	p = 0.967 (p = 39)	***p < 0.001 (n = 40)***	p = 0.273 (n = 40)
	sample subset	p = 0.174 (n = 36)	***p = 0.027 (n = 36)***	***p = 0.024 (n = 36)***
% high DNA bacteria	all samples	***p = 0.007 (n = 39)***	p = 0.078 (n = 40)	***p = 0.009 (n = 40)***
	sample subset	***p = 0.017 (n = 36)***	p = 0.059 (n = 36)	***p < 0.001 (n = 36)***
Synechococcus	all samples	p = 0.143 (n = 39)	p = 0.094 (n = 40)	***p = 0.041 (n = 40)***
	sample subset	p = 0.142 (n = 36)	***p = 0.023 (n = 36)***	***p = 0.013 (n = 36)***
Prochlorococcus	all samples	p = 0.118 (n = 39)	p = 0.076 (n = 40)	p = 0.123 (n = 40)
	sample subset	p = 0.249 (n = 37)	p = 0.161 (n = 37)	p = 0.140 (n = 37)

Station identifier (TARA_station#)	Sample identifier (TARA_station#_environmental-feature)	Environmental Feature	PANGAEA sample identifier	Corresponding contextual data published at PANGAEA	Date/Time (yyyy-mm-ddThh:mm)	Latitude (degrees North)	Longitude (degrees East)	Sampling depth (m)	Marine pelagic biomes (Longhurst 2007)
TARA_018	TARA_018_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_A100000172	http://www.pangaea.de/search?ALL&q=TARA_A100000172	2009-11-02T14:07	35.7528	14.2765	60	Westerlies Biome
TARA_018	TARA_018_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_A100000171	http://www.pangaea.de/search?ALL&q=TARA_A100000171	2009-11-02T08:13	35.759	14.2574	5	Westerlies Biome
TARA_022	TARA_022_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_S200002703	http://www.pangaea.de/search?ALL&q=TARA_S200002703	2009-11-18T08:18	39.8394	17.4185	5	Westerlies Biome
TARA_023	TARA_023_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_X000001382	http://www.pangaea.de/search?ALL&q=TARA_X000001382	2009-11-18T13:50	42.1735	17.7252	65	Westerlies Biome
TARA_025	TARA_025_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_E500000305	http://www.pangaea.de/search?ALL&q=TARA_E500000305	2009-11-23T13:52	39.3991	19.3997	60	Westerlies Biome
TARA_025	TARA_025_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_E500000318	http://www.pangaea.de/search?ALL&q=TARA_E500000318	2009-11-23T08:12	39.3888	19.3906	5	Westerlies Biome
TARA_030	TARA_030_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_X000001348	http://www.pangaea.de/search?ALL&q=TARA_X000001348	2009-12-15T16:23:37	33.9236	32.8118	70	Westerlies Biome
TARA_031	TARA_031_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_A100001391	http://www.pangaea.de/search?ALL&q=TARA_A100001391	2010-01-09T07:15	27.16	34.835	5	Coastal Biome
TARA_032	TARA_032_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_A100001510	http://www.pangaea.de/search?ALL&q=TARA_A100001510	2010-01-11T14:17	23.4193	37.245	80	Coastal Biome
TARA_032	TARA_032_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_A100001515	http://www.pangaea.de/search?ALL&q=TARA_A100001515	2010-01-11T07:21	23.38	37.2183	5	Coastal Biome
TARA_034	TARA_034_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000008	http://www.pangaea.de/search?ALL&q=TARA_R100000008	2010-01-20T11:50	18.4417	39.8587	60	Coastal Biome
TARA_034	TARA_034_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000005	http://www.pangaea.de/search?ALL&q=TARA_R100000005	2010-01-20T04:27	18.3907	39.875	5	Coastal Biome
TARA_036	TARA_036_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000030	http://www.pangaea.de/search?ALL&q=TARA_R100000030	2010-03-12T12:23	20.8122	63.6133	17	Coastal Biome
TARA_036	TARA_036_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000027	http://www.pangaea.de/search?ALL&q=TARA_R100000027	2010-03-12T04:08	20.8183	63.6047	5	Coastal Biome
TARA_038	TARA_038_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000084	http://www.pangaea.de/search?ALL&q=TARA_R100000084	2010-03-15T11:18	19.0284	64.6128	25	Trades Biome
TARA_038	TARA_038_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000081	http://www.pangaea.de/search?ALL&q=TARA_R100000081	2010-03-15T03:36	19.0393	64.4913	5	Trades Biome
TARA_039	TARA_039_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000482	http://www.pangaea.de/search?ALL&q=TARA_R100000482	2010-03-18T11:23	18.5639	66.4727	25	Trades Biome
TARA_039	TARA_039_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000479	http://www.pangaea.de/search?ALL&q=TARA_R100000479	2010-03-18T04:27	18.5916	66.622	5	Trades Biome
TARA_041	TARA_041_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000458	http://www.pangaea.de/search?ALL&q=TARA_R100000458	2010-03-30T10:58	14.9534	70.0128	80	Trades Biome
TARA_041	TARA_041_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000455	http://www.pangaea.de/search?ALL&q=TARA_R100000455	2010-03-30T02:47	14.8059	69.9776	5	Trades Biome
TARA_042	TARA_042_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000152	http://www.pangaea.de/search?ALL&q=TARA_R100000152	2010-04-04T09:50	5.9998	73.9047	80	Trades Biome
TARA_042	TARA_042_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000149	http://www.pangaea.de/search?ALL&q=TARA_R100000149	2010-04-04T02:47	6.0001	73.8955	5	Trades Biome
TARA_048	TARA_048_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000408	http://www.pangaea.de/search?ALL&q=TARA_R100000408	2010-04-15T02:01	-0.8825	73.181	5	Trades Biome
TARA_052	TARA_052_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000234	http://www.pangaea.de/search?ALL&q=TARA_R100000234	2010-05-17T11:42	-16.9534	53.9801	75	Trades Biome
TARA_064	TARA_064_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000315	http://www.pangaea.de/search?ALL&q=TARA_R100000315	2010-07-08T08:21	-29.5333	37.9117	85	Coastal Biome
TARA_064	TARA_064_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000322	http://www.pangaea.de/search?ALL&q=TARA_R100000322	2010-07-07T04:48	-29.5019	37.9883	5	Coastal Biome
TARA_065	TARA_065_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100001188	http://www.pangaea.de/search?ALL&q=TARA_R100001188	2010-07-12T11:03:22	-35.2421	26.3048	30	Coastal Biome
TARA_065	TARA_065_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100001230	http://www.pangaea.de/search?ALL&q=TARA_R100001230	2010-07-12T05:59	-35.1728	26.2888	5	Coastal Biome
TARA_066	TARA_066_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000908	http://www.pangaea.de/search?ALL&q=TARA_R100000908	2010-07-15T15:38:10	-34.6901	18.0458	30	Coastal Biome
TARA_066	TARA_066_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000900	http://www.pangaea.de/search?ALL&q=TARA_R100000900	2010-07-15T12:22	-34.9449	17.9183	5	Coastal Biome
TARA_067	TARA_067_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000951	http://www.pangaea.de/search?ALL&q=TARA_R100000951	2010-08-07T08:19	-32.2401	17.7103	5	Coastal Biome
TARA_068	TARA_068_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000985	http://www.pangaea.de/search?ALL&q=TARA_R100000985	2010-08-14T13:20	-31.027	4.6802	50	Trades Biome
TARA_068	TARA_068_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100000968	http://www.pangaea.de/search?ALL&q=TARA_R100000968	2010-08-14T08:55	-31.0288	4.665	5	Trades Biome
TARA_070	TARA_070_MESO	(MES) mesopelagic zone (ENVO:00000213)	TARA_R100001039	http://www.pangaea.de/search?ALL&q=TARA_R100001039	2010-08-21T11:57:54	-20.4075	-3.1841	800	Trades Biome
TARA_070	TARA_070_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100001015	http://www.pangaea.de/search?ALL&q=TARA_R100001015	2010-08-21T08:55	-20.4081	-3.1758	5	Trades Biome
TARA_072	TARA_072_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100001062	http://www.pangaea.de/search?ALL&q=TARA_R100001062	2010-10-05T15:35	-8.7288	-17.8804	100	Trades Biome
TARA_072	TARA_072_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100001079	http://www.pangaea.de/search?ALL&q=TARA_R100001079	2010-10-05T08:00	-8.7768	-17.9088	5	Trades Biome
TARA_078	TARA_078_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100001129	http://www.pangaea.de/search?ALL&q=TARA_R100001129	2010-10-18T10:59:37	-21.0182	-35.3488	150	Trades Biome
TARA_078	TARA_078_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100001128	http://www.pangaea.de/search?ALL&q=TARA_R100001128	2010-10-18T09:55	-20.9354	-35.1803	5	Trades Biome
TARA_082	TARA_082_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100000544	http://www.pangaea.de/search?ALL&q=TARA_R100000544	2010-12-08T18:58	-47.2007	-57.9448	40	Coastal Biome
TARA_085	TARA_085_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100001377	http://www.pangaea.de/search?ALL&q=TARA_R100001377	2011-01-08T19:34:21	-62.2231	-49.2139	90	Polar Biome
TARA_109	TARA_109_DCM	(DCM) deep chlorophyll maximum layer (ENVO:01000326)	TARA_R100001510	http://www.pangaea.de/search?ALL&q=TARA_R100001510	2011-05-12T22:15	2.0133	-84.5548	30	Coastal Biome
TARA_109	TARA_109_SRF	(SRF) surface water layer (ENVO:00002042)	TARA_R100001509	http://www.pangaea.de/search?ALL&q=TARA_R100001509	2011-05-12T14:00	1.9928	-84.5708	5	Coastal Biome

Sample identifier	Ocean and sea regions (IHO General Sea Areas 1953) [MRGID registered at www.marineregions.com]	Marine pelagic biomes (Longhurst 2007) [MRGID registered at www.marineregions.com]
TARA_018_DCM	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_018_SRF	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_022_SRF	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_023_DCM	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_025_DCM	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_025_SRF	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_030_DCM	(MS) Mediterranean Sea [MRGID:1905]	(MEDI) Mediterranean Sea, Black Sea Province [MRGID:21405]
TARA_031_SRF	(RS) Red Sea [MRGID:4264]	(REDS) Red Sea, Persian Gulf Province [MRGID:21474]
TARA_032_DCM	(RS) Red Sea [MRGID:4264]	(REDS) Red Sea, Persian Gulf Province [MRGID:21474]
TARA_032_SRF	(RS) Red Sea [MRGID:4264]	(REDS) Red Sea, Persian Gulf Province [MRGID:21474]
TARA_034_DCM	(RS) Red Sea [MRGID:4264]	(REDS) Red Sea, Persian Gulf Province [MRGID:21474]
TARA_034_SRF	(RS) Red Sea [MRGID:4264]	(REDS) Red Sea, Persian Gulf Province [MRGID:21474]
TARA_036_DCM	(IO) Indian Ocean [MRGID:1904]	(ARAB) Northwest Arabian Sea Upwelling Province [MRGID:21475]
TARA_036_SRF	(IO) Indian Ocean [MRGID:1904]	(ARAB) Northwest Arabian Sea Upwelling Province [MRGID:21475]
TARA_038_DCM	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_038_SRF	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_039_DCM	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_039_SRF	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_041_DCM	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_041_SRF	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_042_DCM	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_042_SRF	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_048_SRF	(IO) Indian Ocean [MRGID:1904]	(MONS) Indian Monsoon Gyres Province [MRGID:21471]
TARA_052_DCM	(IO) Indian Ocean [MRGID:1904]	(ISSG) Indian South Subtropical Gyre Province [MRGID:21472]
TARA_064_DCM	(IO) Indian Ocean [MRGID:1904]	(EAFR) Eastern Africa Coastal Province [MRGID:21473]
TARA_064_SRF	(IO) Indian Ocean [MRGID:1904]	(EAFR) Eastern Africa Coastal Province [MRGID:21473]
TARA_065_DCM	(IO) Indian Ocean [MRGID:1904]	(EAFR) Eastern Africa Coastal Province [MRGID:21473]
TARA_065_SRF	(IO) Indian Ocean [MRGID:1904]	(EAFR) Eastern Africa Coastal Province [MRGID:21473]
TARA_066_DCM	(SAO) South Atlantic Ocean [MRGID:1914]	(BENG) Benguela Current Coastal Province [MRGID:21470]
TARA_066_SRF	(SAO) South Atlantic Ocean [MRGID:1914]	(BENG) Benguela Current Coastal Province [MRGID:21470]
TARA_067_SRF	(SAO) South Atlantic Ocean [MRGID:1914]	(BENG) Benguela Current Coastal Province [MRGID:21470]
TARA_068_DCM	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_068_SRF	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_070_MESO	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_070_SRF	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_072_DCM	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_072_SRF	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_078_DCM	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_078_SRF	(SAO) South Atlantic Ocean [MRGID:1914]	(SATL) South Atlantic Gyral Province [MRGID:21459]
TARA_082_DCM	(SAO) South Atlantic Ocean [MRGID:1914]	(FKLD) Southwest Atlantic Shelves Province [MRGID:21469]
TARA_085_DCM	(SO) Southern Ocean [MRGID:1907]	(ANTA) Antarctic Province [MRGID:21502]
TARA_109_DCM	(NPO) North Pacific Ocean [MRGID:1908]	(CHIL) Chile-Peru Current Coastal Province [MRGID:21495]
TARA_109_SRF	(NPO) North Pacific Ocean [MRGID:1908]	(CHIL) Chile-Peru Current Coastal Province [MRGID:21495]

Sample identifier	Temperature (°C)	Salinity (psu)	Oxygen (µmol kg^{-1})	Chlorophyll (mg m^{-3})	Nitrite* (µmol kg^{-1})	Phosphate* (µmol kg^{-1})	Nitrite+Nitrate* (µmol kg^{-1})	Silica (µmol kg^{-1})	*Synechococcus* (x 10^4 ml^{-1})	*Prochlorococcus* (x 10^4 ml^{-1})	Bacteria (x 10^5 ml^{-1})	Low DNA Bacteria (x 10^5 ml^{-1})	High DNA Bacteria (x 10^5 ml^{-1})	High DNA Bacteria (% of total bacteria)
TARA_018_DCM	18.4	37.9	230	0.13	0.09	0.02	0.02	0.62	NA	NA	NA	NA	NA	NA
TARA_018_SRF	21.4	37.9	200	0.05	0.02	0.03	0.12	0.56	1.79	2.13	6.34	4.32	2.02	31.8
TARA_022_SRF	17	37.8	221	0.18	0.07	0.02	0.21	2.03	6.19	0.64	5.34	3.09	2.25	42.1
TARA_023_DCM	15.7	38.4	224	0.24	0.02	0.02	0.05	1.05	1.04	1.94	5.95	3.35	2.6	43.8
TARA_025_DCM	15.2	38.5	230	0.2	0.07	0.07	0.15	1.82	NA	NA	NA	NA	NA	NA
TARA_025_SRF	18.3	38.2	218	0.12	0.02	0.02	0.03	1.05	2.56	0.84	2.64	1.69	0.93	35.9
TARA_030_DCM	19	39.3	218	0.14	0.02	0.02	0.02	0.67	1.02	3.78	5.99	3.94	2.05	34.2
TARA_031_SRF	25.1	40	189	0.05	0.02	0.02	0.03	0.82	0.63	20.3	6.47	4.22	2.25	34.8
TARA_032_DCM	26.1	40.2	180	0.22	0.02	0.03	0.02	0.85	0.11	3.08	3.63	1.72	1.91	52.6
TARA_032_SRF	25.8	39.7	188	0	0.02	0.02	0.02	0.87	0.67	19.4	5.95	3.86	2.09	35.1
TARA_034_DCM	27.6	38.9	175	0.18	0.27	0.09	0.64	1.35	0.01	1.82	NA	NA	NA	NA
TARA_034_SRF	27.6	38.6	184	0.19	0.02	0.19	0.03	3.91	6.95	3.9	5.87	3.93	1.94	33
TARA_036_DCM	25.4	36.5	212	0.35	0.51	0.51	2.08	1.79	13.6	NA	11.7	7.31	4.42	37.7
TARA_036_SRF	25.6	36.5	211	0.13	0.05	0.37	0.13	1.2	18.7	1.91	18.4	12.1	6.26	34
TARA_038_DCM	25.5	NA	155	0.6	0.7	0.42	1.86	1.63	13.7	17.8	16.1	10	6.09	37.8
TARA_038_SRF	26.2	36.6	200	0.16	0.02	0.32	0.06	1.18	8.87	29.8	10.3	5.39	4.99	47.8
TARA_039_DCM	26.8	36.3	193	0.18	0.02	0.26	0.09	1.28	7.58	22.4	9.07	5.77	3.3	36.4
TARA_039_SRF	26.8	36.3	193	0.1	0.02	0.23	0.03	1.44	14.7	12.5	8.95	4.63	4.32	48.3
TARA_041_DCM	27.1	36.5	148	0.48	0.07	0.34	0.49	1.38	0.38	11.9	6.05	3.81	2.23	36.9
TARA_041_SRF	29.1	36	187	0.02	0.02	0.14	0.09	1.47	1.37	24.6	7.09	4.33	2.75	38.9
TARA_042_DCM	27.7	35.1	133	0.39	0.15	0.34	1.39	3.2	NA	3.29	3.22	1.59	1.63	50.7
TARA_042_SRF	30	34.6	189	0.01	0.02	0.08	0.03	2.21	1.37	24.6	7.09	4.33	2.75	38.9
TARA_048_SRF	30.1	35.1	186	0.12	0.02	0.1	0.05	1.94	17.8	15.8	8.43	5.82	2.61	31
TARA_052_DCM	24.9	34.9	193	0.39	0.03	0.18	0.47	3.22	0.58	14.6	5.2	3.29	1.91	36.8
TARA_064_DCM	22.3	35.3	207	0.21	0.02	0.08	0.02	1.75	3.95	13.6	5.6	3.76	3.71	66.3
TARA_064_SRF	22.2	35.3	210	0.16	0.02	0.08	0.02	1.77	5.66	28.8	6.58	4.38	4.38	66.6
TARA_065_DCM	21.8	NA	206	0.27	NA	NA	NA	NA	3.22	2.01	5.88	3.44	4.15	70.6
TARA_065_SRF	21.8	35.4	207	0.22	NA	NA	NA	NA	3.44	2.28	5.79	3.53	4.02	69.5
TARA_066_DCM	15	35.3	240	0.42	0.26	0.37	3.23	2.47	1.54	1.42	7.05	4.53	4.78	67.8
TARA_066_SRF	15	35.3	239	0.25	0.3	0.34	3.34	2.74	1.35	1.19	7.22	4.18	5.12	71
TARA_067_SRF	12.8	34.8	249	1.55	0.17	1.02	7.09	13.88	10.3	NA	35.1	9.68	25.4	72.4
TARA_068_DCM	16.8	35.7	232	0.42	0.29	0.23	1.08	2.44	3.75	13.3	5.85	3.07	2.78	47.5
TARA_068_SRF	16.8	35.7	232	0.2	0.25	0.23	1.3	2.6	3.54	11.7	4.89	2.06	3.45	70.6
TARA_070_MESO	4.2	34.4	162	0	0.02	2.52	36.73	44.86	0	0	0.51	0.31	0.34	67.3
TARA_070_SRF	19.8	36.4	216	0.32	0.05	0.36	0.99	1.36	1.49	8.42	6.86	3.93	4.89	71.3
TARA_072_DCM	24.1	36.6	194	0.26	0.02	0.1	0.02	0.96	0.07	11.7	4.01	2.49	1.52	38
TARA_072_SRF	25	36.4	199	0.05	0.02	0.1	0.02	0.87	0.26	19.4	7.89	2.67	5.22	66.2
TARA_078_DCM	21.6	36.7	204	0.15	0.02	0.04	0.02	0.71	0.08	9.42	3.83	2.01	2.42	63.1
TARA_078_SRF	23.3	37.2	206	0.03	0.02	0.06	0.02	0.81	0.42	18.6	5.25	2.34	4.07	77.4
TARA_082_DCM	7	34.1	306	1.02	0.14	1.42	19.07	3.08	0.01	0.35	5.15	2.02	3.12	60.1
TARA_085_DCM	0.8	34.3	325	0.54	0.06	2.31	31	81.82	0	0	1.94	0.68	1.26	65
TARA_109_DCM	26.5	34.3	293	0.74	0.11	0.5	4.64	3.23	2.5	15.54	8.24	5.06	3.18	38.6
TARA_109_SRF	27.6	33.4	199	0.28	0.04	0.27	0.89	1.75	1.39	NA	10.5	3.91	6.62	62.8

* Values below detection limit are reported as the detection limit of 0.02 µmol kg^{-1}

NA = not available

Station identifier [TARA_station#]	Sample identifier [TARA_station#_environmental-feature]	Sample label [TARA_station#_environmental-feature_size-fraction]	Size fraction lower threshold [micrometre]	Size fraction upper threshold [micrometre]	INSDC run accession number(s)	Corresponding nucleotides data published at ENA
TARA_018	TARA_018_DCM	TARA_018_DCM_<-0.22	<	0.22	ERR594352	http://www.ebi.ac.uk/ena/data/view/ERR594352
TARA_018	TARA_018_SRF	TARA_018_SRF_<-0.22	<	0.22	ERR594358	http://www.ebi.ac.uk/ena/data/view/ERR594358
TARA_022	TARA_022_SRF	TARA_022_SRF_<-0.22	<	0.22	ERR594378|ERR594406	http://www.ebi.ac.uk/ena/data/view/ERR594378,ERR594406
TARA_023	TARA_023_DCM	TARA_023_DCM_<-0.22	<	0.22	ERR594408	http://www.ebi.ac.uk/ena/data/view/ERR594408
TARA_025	TARA_025_DCM	TARA_025_DCM_<-0.22	<	0.22	ERR594375	http://www.ebi.ac.uk/ena/data/view/ERR594375
TARA_025	TARA_025_SRF	TARA_025_SRF_<-0.22	<	0.22	ERR594398	http://www.ebi.ac.uk/ena/data/view/ERR594398
TARA_030	TARA_030_DCM	TARA_030_DCM_<-0.22	<	0.22	ERR594405	http://www.ebi.ac.uk/ena/data/view/ERR594405
TARA_031	TARA_031_SRF	TARA_031_SRF_<-0.22	<	0.22	ERR594401|ERR594410	http://www.ebi.ac.uk/ena/data/view/ERR594401,ERR594410
TARA_032	TARA_032_DCM	TARA_032_DCM_<-0.22	<	0.22	ERR594390	http://www.ebi.ac.uk/ena/data/view/ERR594390
TARA_032	TARA_032_SRF	TARA_032_SRF_<-0.22	<	0.22	ERR594393	http://www.ebi.ac.uk/ena/data/view/ERR594393
TARA_034	TARA_034_DCM	TARA_034_DCM_<-0.22	<	0.22	ERR594380	http://www.ebi.ac.uk/ena/data/view/ERR594380
TARA_034	TARA_034_SRF	TARA_034_SRF_<-0.22	<	0.22	ERR594368|ERR594370	http://www.ebi.ac.uk/ena/data/view/ERR594368,ERR594370
TARA_036	TARA_036_DCM	TARA_036_DCM_<-0.22	<	0.22	ERR594402	http://www.ebi.ac.uk/ena/data/view/ERR594402
TARA_036	TARA_036_SRF	TARA_036_SRF_<-0.22	<	0.22	ERR594369	http://www.ebi.ac.uk/ena/data/view/ERR594369
TARA_038	TARA_038_DCM	TARA_038_DCM_<-0.22	<	0.22	ERR594388|ERR594399	http://www.ebi.ac.uk/ena/data/view/ERR594388,ERR594399
TARA_038	TARA_038_SRF	TARA_038_SRF_<-0.22	<	0.22	ERR594374|ERR594400	http://www.ebi.ac.uk/ena/data/view/ERR594374,ERR594400
TARA_039	TARA_039_DCM	TARA_039_DCM_<-0.22	<	0.22	ERR594363|ERR594397	http://www.ebi.ac.uk/ena/data/view/ERR594363,ERR594397
TARA_039	TARA_039_SRF	TARA_039_SRF_<-0.22	<	0.22	ERR594356|ERR594365	http://www.ebi.ac.uk/ena/data/view/ERR594356,ERR594365
TARA_041	TARA_041_DCM	TARA_041_DCM_<-0.22	<	0.22	ERR594368|ERR594367|ERR594371|ERR594372|ERR594373	http://www.ebi.ac.uk/ena/data/view/ERR594368,ERR594367,ERR594371,ERR594372,ERR594373
TARA_041	TARA_041_SRF	TARA_041_SRF_<-0.22	<	0.22	ERR594384	http://www.ebi.ac.uk/ena/data/view/ERR594384
TARA_042	TARA_042_DCM	TARA_042_DCM_<-0.22	<	0.22	ERR594413	http://www.ebi.ac.uk/ena/data/view/ERR594413
TARA_042	TARA_042_SRF	TARA_042_SRF_<-0.22	<	0.22	ERR594398|ERR594403	http://www.ebi.ac.uk/ena/data/view/ERR594398,ERR594403
TARA_046	TARA_046_SRF	TARA_046_SRF_<-0.22	<	0.22	ERR594376	http://www.ebi.ac.uk/ena/data/view/ERR594376
TARA_052	TARA_052_DCM	TARA_052_DCM_<-0.22	<	0.22	ERR594394	http://www.ebi.ac.uk/ena/data/view/ERR594394
TARA_064	TARA_064_DCM	TARA_064_DCM_<-0.22	<	0.22	ERR594385	http://www.ebi.ac.uk/ena/data/view/ERR594385
TARA_064	TARA_064_SRF	TARA_064_SRF_<-0.22	<	0.22	ERR594392	http://www.ebi.ac.uk/ena/data/view/ERR594392
TARA_065	TARA_065_DCM	TARA_065_DCM_<-0.22	<	0.22	ERR594362|ERR594414	http://www.ebi.ac.uk/ena/data/view/ERR594362,ERR594414
TARA_065	TARA_065_SRF	TARA_065_SRF_<-0.22	<	0.22	ERR594359|ERR594381	http://www.ebi.ac.uk/ena/data/view/ERR594359,ERR594381
TARA_066	TARA_066_DCM	TARA_066_DCM_<-0.22	<	0.22	ERR594389	http://www.ebi.ac.uk/ena/data/view/ERR594389
TARA_066	TARA_066_SRF	TARA_066_SRF_<-0.22	<	0.22	ERR594382	http://www.ebi.ac.uk/ena/data/view/ERR594382
TARA_067	TARA_067_SRF	TARA_067_SRF_<-0.22	<	0.22	ERR594395|ERR594404	http://www.ebi.ac.uk/ena/data/view/ERR594395,ERR594404
TARA_068	TARA_068_DCM	TARA_068_DCM_<-0.22	<	0.22	ERR594415	http://www.ebi.ac.uk/ena/data/view/ERR594415
TARA_068	TARA_068_SRF	TARA_068_SRF_<-0.22	<	0.22	ERR594391	http://www.ebi.ac.uk/ena/data/view/ERR594391
TARA_070	TARA_070_MESO	TARA_070_MES_<-0.22	<	0.22	ERR594407	http://www.ebi.ac.uk/ena/data/view/ERR594407
TARA_070	TARA_070_SRF	TARA_070_SRF_<-0.22	<	0.22	ERR594353	http://www.ebi.ac.uk/ena/data/view/ERR594353
TARA_072	TARA_072_DCM	TARA_072_DCM_<-0.22	<	0.22	ERR594378	http://www.ebi.ac.uk/ena/data/view/ERR594378
TARA_072	TARA_072_SRF	TARA_072_SRF_<-0.22	<	0.22	ERR594364	http://www.ebi.ac.uk/ena/data/view/ERR594364
TARA_076	TARA_076_DCM	TARA_076_DCM_<-0.22	<	0.22	ERR594355	http://www.ebi.ac.uk/ena/data/view/ERR594355
TARA_076	TARA_076_SRF	TARA_076_SRF_<-0.22	<	0.22	ERR594354	http://www.ebi.ac.uk/ena/data/view/ERR594354
TARA_082	TARA_082_DCM	TARA_082_DCM_<-0.22	<	0.22	ERR594409	http://www.ebi.ac.uk/ena/data/view/ERR594409
TARA_085	TARA_085_DCM	TARA_085_DCM_<-0.22	<	0.22	ERR594377	http://www.ebi.ac.uk/ena/data/view/ERR594377
TARA_109	TARA_109_DCM	TARA_109_DCM_<-0.22	<	0.22	ERR594357|ERR594360|ERR594361|ERR594383|ERR594387	http://www.ebi.ac.uk/ena/data/view/ERR594357,ERR594360,ERR594361,ERR594383,ERR594387
TARA_109	TARA_109_SRF	TARA_109_SRF_<-0.22	<	0.22	ERR594412	http://www.ebi.ac.uk/ena/data/view/ERR594412

Sample label [TARA_station#_environmental-feature_size-fraction]	Paired reads	Individual reads	Aligned reads	Reads in PCs	Reads in viral contigs	Reads in viral populations	Total PCs	Unique PCs	Shannon's H' for PCs (normalized to 5M reads)	Pielou's Evenness (J) for PCs (normalized to 5M reads)	Shannon's H' for PCs in virus contigs (normalized to 500k reads)	Local:Global diversity ratio for PCs (normalized to 5M reads)	Shannon's H' for viral populations (normalized to 100k reads)	Pielou's Evenness (J) for viral populations (normalized to 100k reads)	Local:Global diversity ratio for viral populations (normalized to 100k reads)
TARA_018_DCM_<-0.22	41756356	83512712	25174800	23357071	1793334	1026291	78974	49026	9.8	0.87	7.54	0.81	5.64	0.87	0.71
TARA_018_SRF_<-0.22	47508906	95017812	46106031	44220183	2047711	2179284	94270	56767	9.83	0.86	7.39	0.81	5.65	0.87	0.71
TARA_022_SRF_<-0.22	57573241	115146482	50619261	45860623	4783830	2623367	149546	78991	10.43	0.88	8.62	0.86	6.48	0.93	0.81
TARA_023_DCM_<-0.22	41387025	82774050	30778047	28011014	2870031	1464834	120126	66984	10.21	0.87	8.16	0.84	6.23	0.92	0.78
TARA_025_DCM_<-0.22	52556395	105112790	46691361	43479717	2180587	1382904	149410	92250	10.47	0.88	7.89	0.86	6.28	0.90	0.79
TARA_025_SRF_<-0.22	52114733	104229466	46939361	41788880	3609305	2285684	134834	64259	10.44	0.88	8.21	0.86	6.31	0.91	0.79
TARA_030_DCM_<-0.22	39094357	78188714	30119573	28438464	2910948	1686322	85995	53433	9.9	0.87	7.84	0.81	5.76	0.90	0.73
TARA_031_SRF_<-0.22	66963246	133926492	64232168	58034883	3746026	2644581	161635	100986	10.29	0.86	8.14	0.85	6.44	0.91	0.81
TARA_032_DCM_<-0.22	64643745	129287490	45773512	44846740	588548	448429	50432	15564	9.49	0.88	6.33	0.78	4.22	0.68	0.53
TARA_032_SRF_<-0.22	50756575	101513150	39347181	33568647	3043755	1771453	136282	74633	10.4	0.88	8.33	0.86	6.28	0.91	0.79
TARA_034_DCM_<-0.22	40848755	81697510	27781799	25512131	2475336	1441846	91705	56060	10.02	0.88	7.93	0.82	5.86	0.89	0.74
TARA_034_SRF_<-0.22	61395078	122790156	38897904	35425635	1731130	1126594	136362	90615	10.36	0.88	7.66	0.85	6.45	0.92	0.81
TARA_036_DCM_<-0.22	38625900	77251800	28199576	24009195	1792668	858377	80868	38517	9.97	0.88	7.71	0.82	5.46	0.87	0.69
TARA_036_SRF_<-0.22	37151587	74303174	26670863	23298945	1340632	795266	61808	31090	9.7	0.88	7.25	0.80	5.59	0.86	0.70
TARA_038_DCM_<-0.22	47121439	94242878	26769138	24918425	2343749	1628668	108957	50428	10.27	0.89	7.75	0.85	6.29	0.90	0.79
TARA_038_SRF_<-0.22	57024430	114048860	34124947	31088143	1873066	890075	120496	61967	10.14	0.87	7.55	0.83	6.39	0.90	0.80
TARA_039_DCM_<-0.22	45416113	90832226	21953764	20508973	1306449	655684	111796	38162	10.23	0.88	7.4	0.84	6.45	0.93	0.81
TARA_039_SRF_<-0.22	57246863	114493726	33219196	30600380	2093000	1249016	142047	68413	10.55	0.89	8.09	0.87	6.43	0.91	0.81
TARA_041_DCM_<-0.22	100024473	200048946	64282190	57894183	2755513	1169127	143206	97446	10.26	0.86	7.73	0.84	6.01	0.89	0.76
TARA_041_SRF_<-0.22	42658889	85317778	17983366	16842944	770079	558610	90032	50819	10.41	0.91	6.64	0.86	5.86	0.90	0.74
TARA_042_DCM_<-0.22	60004302	120008604	29940507	27764917	1142362	526364	128384	82624	10.26	0.87	7.45	0.84	6.16	0.90	0.77
TARA_042_SRF_<-0.22	49556238	99112476	30237421	28833837	1428407	1146194	100198	60985	9.96	0.86	6.74	0.82	5.48	0.83	0.69
TARA_046_SRF_<-0.22	38717821	77435642	25564107	23934444	1490412	923822	83427	52923	9.77	0.86	7.26	0.80	5.85	0.90	0.74
TARA_052_DCM_<-0.22	46496880	92993760	29085216	25854768	1901209	1134072	115331	74838	10.26	0.88	7.86	0.84	6.02	0.91	0.76
TARA_064_DCM_<-0.22	42826969	85653938	20073137	18787557	757872	406380	103357	60393	10.18	0.88	6.89	0.84	6.19	0.91	0.78
TARA_064_SRF_<-0.22	48271080	96542160	32186492	30179459	1688065	1296512	127492	69764	10.16	0.86	7.49	0.84	6.48	0.92	0.82
TARA_065_DCM_<-0.22	56703457	113406914	33968214	32008400	2161169	1441554	159111	71854	10.2	0.85	7.09	0.84	6.86	0.92	0.86
TARA_065_SRF_<-0.22	59658652	119317304	25209774	27843929	1053629	611695	143053	55796	10.17	0.86	6.94	0.84	6.61	0.92	0.83
TARA_066_DCM_<-0.22	46900600	93801200	34256927	33246333	3421381	3027385	108741	44925	9.92	0.86	7.57	0.82	6.24	0.91	0.79
TARA_066_SRF_<-0.22	43948626	87897252	36667746	34827911	2914606	2054397	120444	56628	10.21	0.87	8.06	0.84	6.28	0.92	0.79
TARA_067_SRF_<-0.22	43972034	87944068	37210358	35001146	5403501	3708818	98666	70819	9.83	0.86	8.63	0.81	5.90	0.90	0.74
TARA_068_DCM_<-0.22	40938660	81877320	22931296	21166806	1577597	879745	91497	44117	9.93	0.87	7.4	0.82	5.82	0.89	0.73
TARA_068_SRF_<-0.22	44816720	89633440	33145982	31695053	1418283	1192673	122997	75617	10.04	0.86	7.48	0.83	6.22	0.92	0.78
TARA_070_MES_<-0.22	41257033	82514066	37673669	35685550	1165029	1044424	62992	35984	9.47	0.86	6.69	0.78	4.90	0.83	0.62
TARA_070_SRF_<-0.22	47220407	94440814	29984244	28300746	1597071	1019761	133412	88128	10.3	0.87	7.55	0.85	6.02	0.90	0.76
TARA_072_DCM_<-0.22	48812251	97624502	32787182	29700584	2804641	1563306	98478	62170	9.96	0.87	7.92	0.82	5.68	0.89	0.71
TARA_072_SRF_<-0.22	39987130	79974260	24649114	22486780	1480960	843843	104462	59399	10.1	0.87	7.61	0.83	6.04	0.90	0.76
TARA_076_DCM_<-0.22	58870148	117740296	45747168	41445203	3091548	2180949	157970	80248	10.52	0.88	8.11	0.87	6.57	0.91	0.83
TARA_076_SRF_<-0.22	50148148	100296296	41354569	38608348	2263533	1755953	121463	43741	10.4	0.89	7.69	0.86	5.99	0.89	0.75
TARA_082_DCM_<-0.22	41877228	83754456	19062414	17969483	2092234	1285126	35761	24067	8.93	0.85	6.88	0.73	1.80	0.33	0.23
TARA_085_DCM_<-0.22	44370368	88740736	25609713	23868405	1246929	690679	55412	41835	9.22	0.84	6.93	0.76	4.04	0.84	0.51
TARA_109_DCM_<-0.22	82153717	164307434	55257622	50850295	3054367	1641370	132031	71720	10.16	0.87	7.87	0.84	5.38	0.85	0.68
TARA_109_SRF_<-0.22	45818926	91637852	31556400	28756552	2186577	1338874	144438	84127	10.27	0.86	8.1	0.85	6.19	0.91	0.78