OCEAN PLANKTON

Structure and function of the global ocean microbiome

Shinichi Sunagawa,[1*†] Luis Pedro Coelho,[1*] Samuel Chaffron,[2,3,4*] Jens Roat Kultima,[1] Karine Labadie,[5] Guillem Salazar,[6] Bardya Djahanschiri,[1] Georg Zeller,[1] Daniel R. Mende,[1] Adriana Alberti,[5] Francisco M. Cornejo-Castillo,[6] Paul I. Costea,[1] Corinne Cruaud,[5] Francesco d'Ovidio,[7] Stefan Engelen,[5] Isabel Ferrera,[6] Josep M. Gasol,[6] Lionel Guidi,[8,9] Falk Hildebrand,[1] Florian Kokoszka,[10,11] Cyrille Lepoivre,[12] Gipsi Lima-Mendez,[2,3,4] Julie Poulain,[5] Bonnie T. Poulos,[13] Marta Royo-Llonch,[6] Hugo Sarmento,[6,14] Sara Vieira-Silva,[2,3,4] Céline Dimier,[10,15,16] Marc Picheral,[8,9] Sarah Searson,[8,9] Stefanie Kandels-Lewis,[1,17] *Tara* Oceans coordinators‡ Chris Bowler,[10] Colomban de Vargas,[15,16] Gabriel Gorsky,[8,9] Nigel Grimsley,[18,19] Pascal Hingamp,[12] Daniele Iudicone,[20] Olivier Jaillon,[5,21,22] Fabrice Not,[15,16] Hiroyuki Ogata,[23] Stephane Pesant,[24,25] Sabrina Speich,[26,27] Lars Stemmann,[8,9] Matthew B. Sullivan,[13§] Jean Weissenbach,[5,21,22] Patrick Wincker,[5,21,22] Eric Karsenti,[10,17†] Jeroen Raes,[2,3,4†] Silvia G. Acinas,[6†] Peer Bork[1,28†]

Microbes are dominant drivers of biogeochemical processes, yet drawing a global picture of functional diversity, microbial community structure, and their ecological determinants remains a grand challenge. We analyzed 7.2 terabases of metagenomic data from 243 *Tara* Oceans samples from 68 locations in epipelagic and mesopelagic waters across the globe to generate an ocean microbial reference gene catalog with >40 million nonredundant, mostly novel sequences from viruses, prokaryotes, and picoeukaryotes. Using 139 prokaryote-enriched samples, containing >35,000 species, we show vertical stratification with epipelagic community composition mostly driven by temperature rather than other environmental factors or geography. We identify ocean microbial core functionality and reveal that >73% of its abundance is shared with the human gut microbiome despite the physicochemical differences between these two ecosystems.

Microorganisms are ubiquitous in the ocean environment, where they play key roles in biogeochemical processes, such as carbon and nutrient cycling (*1*). With an estimated 10^4 to 10^6 cells per milliliter, their biomass, combined with high turnover rates and environmental complexity, provides the grounds for immense genetic diversity (*2*). These microorganisms, and the communities they form, drive and respond to changes in the environment, including climate change–associated shifts in temperature, carbon chemistry, nutrient and oxygen content, and alterations in ocean stratification and currents (*3*).

With recent advances in community DNA shotgun sequencing (metagenomics) and computational analysis, it is now possible to access the taxonomic and genomic content (microbiome) of ocean microbial communities and, thus, to study their structural patterns, diversity, and functional potential (*4*, *5*). The *Sorcerer II* Global Ocean Sampling (GOS) expedition, for example, collected, sequenced, and analyzed 6.3 gigabases (Gb) of DNA from surface-water samples along a transect from the Northwest Atlantic to the Eastern Tropical Pacific (*6*, *7*) but also indicated that the vast majority of the global ocean microbiome still remained to be uncovered (*7*). Nevertheless, the GOS project facilitated the study of surface picoplanktonic communities from these regions by providing an ocean metagenomic data set to the scientific community. Several studies have demonstrated that such data could, in principle, identify relationships between gene functional compositions and environmental factors (*8*–*10*). However, an extended breadth of sampling (e.g., across depth layers, domains of life, organismal-size classes, and around the globe), combined with in situ measured environmental data, could provide a global context and minimize potential confounders.

To this end, *Tara* Oceans systematically collected ~35,000 samples for morphological, genetic, and environmental analyses using standardized protocols across multiple depths at global scale, aiming to facilitate a holistic study on how environmental factors and biogeochemical cycles affect oceanic life (*11*). Here we report the initial analysis of 243 ocean microbiome samples, collected at 68 locations representing all main oceanic regions (except for the Arctic) from three depth layers, which were subjected to metagenomic Illumina sequencing. By integrating these data with those from publicly available ocean metagenomes and reference genomes, we assembled and annotated a reference gene catalog, which we use in combination with phylogenetic marker genes (*12*, *13*) to derive global patterns of functional and taxonomic microbial community structures. The vast majority of genes uncovered in *Tara* Oceans samples had not previously been identified, with particularly high fractions of novel genes in the Southern Ocean and in the twilight, mesopelagic zone. By correlating genomic and environmental features, we infer that temperature, which we decoupled from dissolved oxygen, is the strongest environmental factor shaping microbiome composition in the sunlit, epipelagic ocean layer. Furthermore, we define a core set of gene families that are ubiquitous in the ocean and differentiate variable, adaptive functions from stable core functions; the latter are compared between ocean depth layers and to those in the human gut microbiome.

Ocean microbial reference gene catalog

To capture the genomic content of prevalent microbiota across major oceanic regions (Fig. 1A), *Tara* Oceans collected seawater samples within the epipelagic layer, both from the surface water and the deep chlorophyll maximum (DCM) layers, as well as the mesopelagic zone (*14*). From 68 selected locations, 243 size-fractionated samples targeting organisms up to 3 μm [virus-enriched fraction (<0.22 μm): n = 45; girus/prokaryote-enriched fractions (0.1 to 0.22 μm, 0.22 to 0.45 μm,

[1]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [3]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [4]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [5]CEA–Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [6]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [7]Sorbonne Universités, UPMC, Université Paris 06, CNRS-IRD-MNHN, LOCEAN Laboratory, 4 Place Jussieu, 75005 Paris, France. [8]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [9]Sorbonne Universités, UPMC Université Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [10]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [11]Laboratoire de Physique des Océans UBO-IUEM, Place Copernic 29820 Plouzané, France. [12]Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. [13]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [14]Department of Hydrobiology, Federal University of São Carlos (UFSCar), Rodovia Washington Luiz, 13565-905 São Carlos, São Paulo, Brazil. [15]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [16]Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [17]Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [19]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [20]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [21]CNRS, UMR 8030, CP5706, Evry, France. [22]Université d'Evry, UMR 8030, CP5706, Evry, France. [23]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan. [24]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [25]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [26]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [27]Laboratoire de Physique des Océans UBO-IUEM, Place Copernic, 29820 Plouzané, France. [28]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany.
*These authors contributed equally to this work **†Corresponding author. E-mail: sunagawa@embl.de (S.S.); karsenti@embl.de (E.K.); jeroen.raes@vib-kuleuven.be (J.R.); sacinas@icm.csic.es (S.G.A.); bork@embl.de (P.B.)** ‡*Tara* Oceans coordinators and affiliations are listed at the end of this paper. §Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

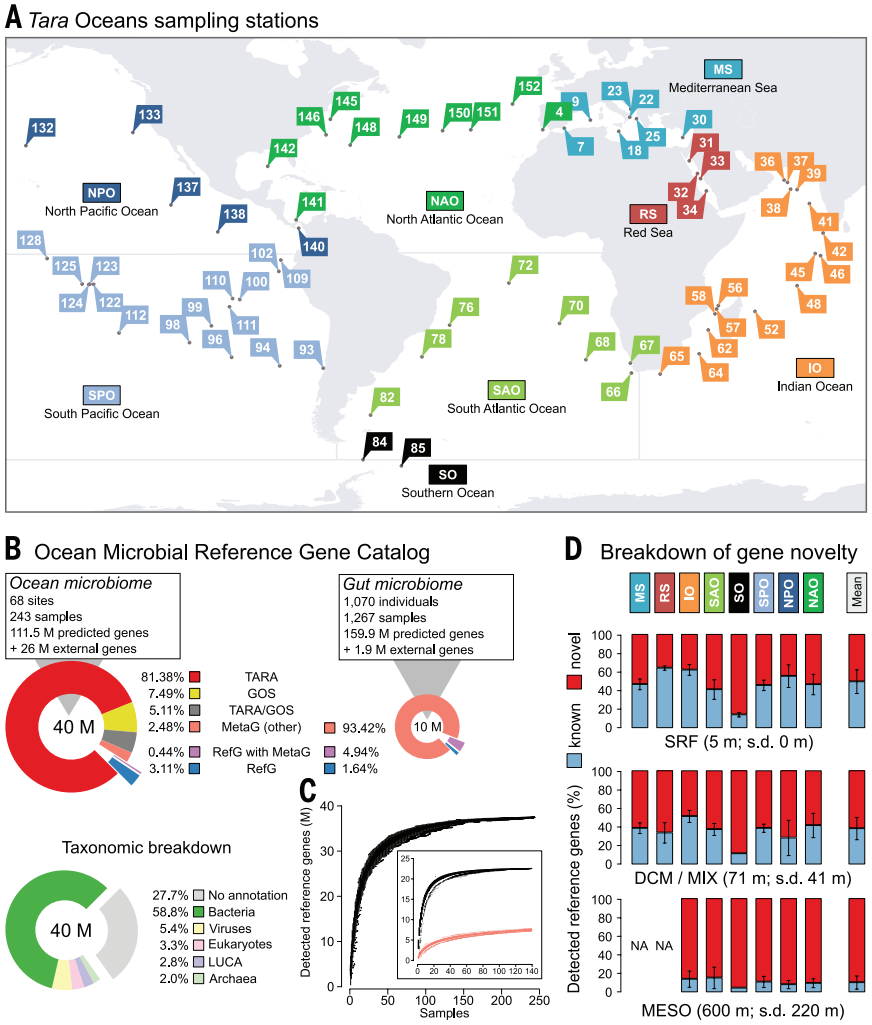

Fig. 1. *Tara* Oceans captures novel genetic diversity in the global ocean microbiome. (**A**) Geographic distribution of 68 (out of >200 in total) representative *Tara* Oceans sampling stations at which seawater samples and environmental data were collected from multiple depth layers. (**B**) Targeting viruses and microbial organisms up to 3 µm in size, deep Illumina shotgun sequencing of 243 samples, followed by metagenomic assembly and gene prediction, resulted in the identification of >111.5 M gene-coding sequences. The currently largest human gut microbial reference gene catalog (*16*) was built with similar amounts of data but from a substantially higher number of samples (n = 1,267). Genes identified in our study were clustered together with >26 M sequences from publicly available data [external genes; see (*14*)] to yield a set of >40 M reference genes (top left), which equals more than four times the number of genes in the human gut microbial reference gene catalog (top right). The combined clustering of genes identified in *Tara* Oceans samples with those obtained from public resources allowed us to annotate genes according to the composition of each cluster. For example, a gene was labeled as: "TARA/GOS" if its original cluster contained sequences from both *Tara* Oceans and GOS samples. More than 81% of the genes were found only in samples collected by *Tara* Oceans. A breakdown of taxonomic annotations (bottom left) shows that the reference gene catalog is mainly composed of bacterial genes (LUCA denotes genes that could not unambiguously be assigned to a domain of life). (**C**) Rarefaction curve of detected genes for 100-fold permuted sampling orders shows only a small increase in newly detected genes toward the end of sampling. The subplot compares sequencing depth-normalized rarefaction curves for 139 prokaryotic ocean samples (black) mapped to the prokaryotic subset of the OM-RGC (24.4 M genes) and the same number of random (100-fold permuted) human gut samples (pink) mapped to a human gut gene catalog (*16*). The lower asymptote for the human gut suggests that the ocean harbors a greater genetic diversity. (**D**) For the subset of 139 prokaryotic samples analyzed, the fraction of detected genes that had previously been available in public databases (blue) are compared to those that were newly identified in samples collected by *Tara* Oceans (red). The breakdown by ocean region and depths shows that the Southern Ocean and the mesopelagic zone had been vastly undersampled prior to *Tara* Oceans. NA, not available. Abbreviations: MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; GOS, *Sorcerer II* Global Ocean Sampling expedition; MetaG, genes of metagenomic origin; RefG, genes from reference genome sequences; LUCA, last universal common ancestor; SRF, surface water layer; DCM, deep chlorophyll maximum layer; MIX, subsurface epipelagic mixed layer; MESO, mesopelagic zone.

0.45 to 0.8 µm): n = 59; prokaryote-enriched fractions (0.22 to 1.6 µm, 0.22 to 3 µm): n = 139] were paired-end shotgun Illumina sequenced to generate a total of more than 7.2 terabases (Tb), 29.6 ± 12.7 Gb per sample (*14*), enabling comparative analyses with the human gut microbiome for which metagenomic data of the same order of magnitude have been published {U.S. Human Microbiome Project, phase I—stool [1.5 Tb; (*15*)]} and the European Metagenomics of the Human Intestinal Tract project [3.8 Tb; (*16*, *17*)].

To generate a reference gene catalog [see also (*16*, *17*)], we first reconstructed the genomic content of *Tara* Oceans samples by metagenomic assembly and gene prediction (*18*) and combined these data with those from publicly available ocean metagenomes and reference genomes (*14*). Specifically, ~111.5 million (M) protein-coding nucleotide sequences were predicted and clustered at 95% nucleotide sequence identity with 24.4 M sequences from other ocean metagenomes (*14*) and 1.6 M sequences from ocean prokaryotic (n = 433) and viral (n = 114) reference genomes (*14*). This resulted in a global Ocean Microbial Reference Gene Catalog (OM-RGC), which comprises >40 M nonredundant representative genes from viruses, prokaryotes, and picoeukaryotes (Fig. 1B). Compared to a human gut microbial reference gene catalog (*16*), the OM-RGC comprises more than four times the number of genes, most of which (59%) appear prokaryotic (Fig. 1B). Almost 28% of the genes could not be taxonomically annotated. A large fraction is, however, likely of viral origin, because in size fractions targeting organisms smaller than 0.22 µm, 37% (SD = 9%) of the profiled sequence data mapped to nonannotated genes [see also (*19*)], whereas in prokaryote-enriched samples, this fraction decreased to 9% (SD = 2%). As expected, eukaryotic genes (3.3%) include those from protists (unicellular eukaryotes) but also from multicellular, larger organisms whose gametes or fragmented cells may have been sampled (*14*).

In total, 81.4% of the genes were exclusive to *Tara* Oceans samples, with only 5.11 and 0.44% overlapping with GOS sequences and reference genomes, respectively (Fig. 1B), which highlights the extent of the unexplored genomic potential in our oceans. Rarefaction analysis showed that the rate of new gene detection decreased to 0.01% by the end of sampling (Fig. 1C), suggesting that the abundant microbial sequence space appears well represented, at least for the targeted size ranges, sampling locations, and depths. Genes found in only one sample amounted to 3.6% of the OM-RGC, which may originate from localized specialists.

To complement the work of *Tara* Oceans Consortium partners who analyzed viral and protist-enriched size fractions (*19*, *20*) and integrated data across domains of life (*21*, *22*), we focused our analyses on 139 prokaryote-enriched samples, which included 63 surface water samples (5 m; SD = 0 m), 46 epipelagic subsurface water samples mostly from the DCM (71 m; SD = 41 m), and 30 mesopelagic samples (600 m; SD = 220 m). Using this set, we revealed that gene novelty generally

increased from surface to DCM waters and remained relatively stable across ocean regions, with overall about half of the genes being novel. As exceptions to this pattern, we find in Southern Ocean (SO) and mesopelagic samples about 80 and 90% of novelty, respectively. In addition to higher novelty in hitherto uncharted regions, these patterns likely reflect the detection of rare organisms by deep sequencing, although seasonal and locational differences of sampling in relatively well-studied regions may be additional contributing factors.

To put the degree of taxonomic novelty into context, we extracted a total of >14 M metagenomic 16*S* ribosomal RNA gene (16*S*) tags [16*S* $_{mi}$tags; (*12*)] and mapped these to operational taxonomic units (OTUs) based on clustering of reference 16*S* sequences (*23*) at 97% sequence identity. This cutoff has been commonly used to group taxa at the species level, although it may rather represent clades somewhere between species and genus level (*24*). The fraction of total 16*S* $_{mi}$tags not matching any reference OTUs also increased with depth but was on average only 5.5% (*14*). Thus, although the vast majority of prokaryotic clades detected in *Tara* Oceans metagenomes had already been captured by 16*S* sequencing, the OM-RGC now provides a link to their genomic content.

Diversity and depth stratification of the ocean microbiome

Given the global scale of *Tara* Oceans samples, we assessed patterns of diversity and stratifying factors of ocean microbial community composition. 16*S* $_{mi}$tags identified in our metagenomic data set mapped to a total of 35,650 OTUs (2937 OTUs; SD = 585 OTUs), and taxonomic and phylogenetic diversity were highly (R^2 = 0.96) correlated (*14*). The total richness estimate of 37,470 is comparable to the numbers from a previous study, which detected about 44,500 OTUs based on polymerase chain reaction (PCR)–amplified 16*S* rRNA tags from 356 globally distributed pelagic samples (*25*) that were collected in the context of the International Census of Marine Microbes (ICoMM) project (*26*). More than 93% of 16*S* $_{mi}$tags could be annotated at the phylum level. We found that typical members of Proteobacteria, including the ubiquitous clades SAR11 (Alphaproteobacteria) and SAR86 (Gammaproteobacteria), dominate the sampled areas of the ocean both in terms of relative abundance and taxonomic richness (*27*, *28*). Cyanobacteria, Deferribacteres, and Thaumarchaeota were also abundant, although the taxonomic richness within these phyla was smaller (Fig. 2). Photosynthetic cyanobacterial taxa such as *Prochlorococcus* and *Synechococcus* were detected in all mesopelagic samples and contributed about 1% of the abundance (Fig. 2), which is in line with previous reports suggesting a role for cyanobacteria in sinking particle flux (*29*).

To explore the overall variability in community composition, we performed a principal coordinate analysis (PCoA), which revealed that depth explained 73% of the variance (PC1 in Fig. 3A). This is consistent with a vertical stratification of microbial taxa and viruses according to changes in physicochemical parameters, such as light, temperature, and nutrients (*30*, *31*). Given this vertical stratification, we further characterized taxonomic and functional richness, between-sample dissimilarity (β-diversity), total cell abundance, and potential growth rates across three depth layers. Our results revealed an increase in both taxonomic and functional richness with depth, whereas cell abundance, as measured by flow cytometry, and potential maximum growth rates (*32*) decreased with depth (Fig. 3B).

Although increasing species richness from the surface to the mesopelagic has been reported locally, e.g., in the Mediterranean Sea (*33*), our findings emphasize the global relevance of this pattern. The observed increase in taxonomic and functional richness may reflect diversified species adapted to a wider range of niches, such as particle-associated microenvironments in the mesopelagic zone (*34*). In addition, slower growth, due

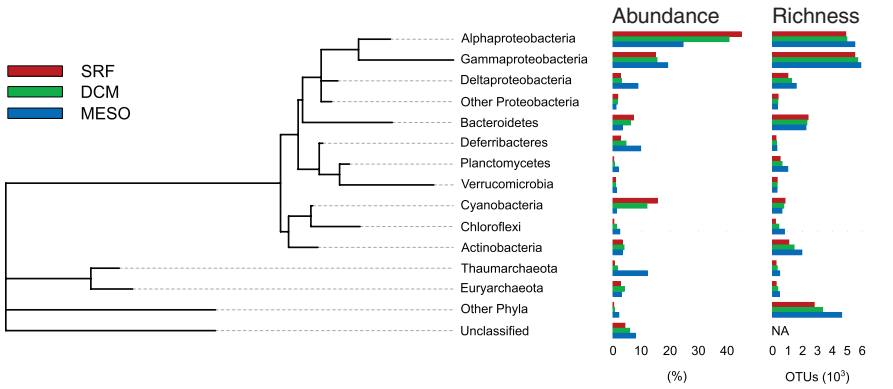

Fig. 2. Taxonomic breakdown of *Tara* Oceans samples. A phylum-level (class-level for Proteobacteria) breakdown of relative abundances is shown for all prokaryotic samples from three depth layers along with the number of detected taxa at the OTU level. SRF, surface water layer; DCM, deep chlorophyll maximum layer; MESO, mesopelagic zone.

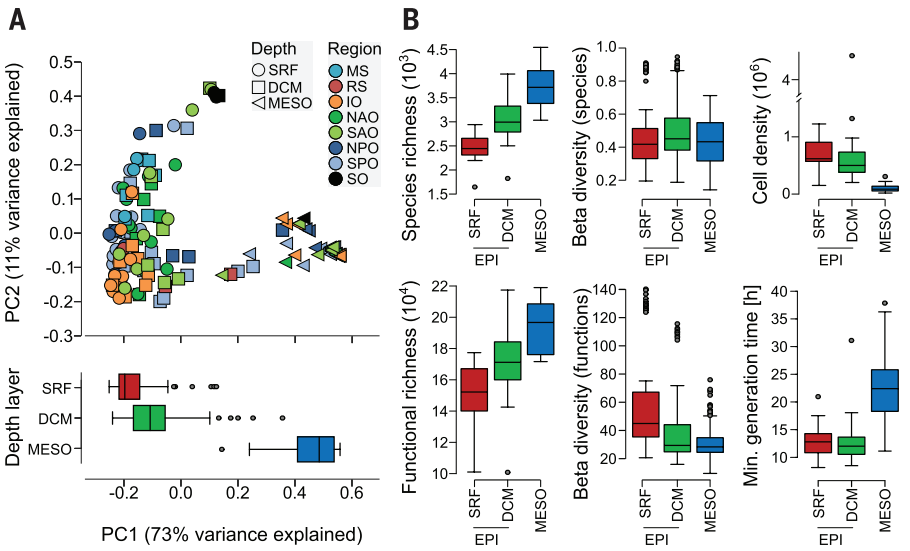

Fig. 3. Depth stratification of the ocean microbiome. (**A**) Principal coordinate (PC) analysis performed on community composition dissimilarities (Bray-Curtis) of 139 prokaryotic samples based on 16*S* $_{mi}$tag relative abundances shows that samples are significantly separated by their depth layer of origin, i.e., surface (SRF), deep chlorophyll maximum (DCM), or mesopelagic (MESO). Boxplots of the first PC illustrate differences between depth layers. Differences between samples from SRF and DCM were significant, but small compared to those with mesopelagic samples. Abbreviations for ocean regions are the same as in Fig. 1. (**B**) For a matched sample set from 20 stations where SRF, DCM, and MESO were sampled, calculations of within-sample species richness (top left) and between-sample diversities (top-center; Bray-Curtis) and cell densities per milliliter (top right) suggest an increase in species richness and a decrease in cell density with depth (pairwise Mann-Whitney U-test: $P < 0.001$), whereas no significant trend was found for between-sample dissimilarity. For gene functional groups (bottom left and center), richness increased with depth, whereas between-sample dissimilarity decreased. Minimum potential generation time of microbial communities (bottom right) is predicted to be higher in the mesopelagic compared to the epipelagic (EPI).

to more limited carbon sources in the mesopelagic zone, and higher motility have been suggested to reduce predation by flagellates and ciliates, as well as viral infection rates (*35*). Our metagenomic analysis now provides molecular support for these models by identifying a significant (P < 0.001) enrichment of chemotaxis and motility genes in the mesopelagic zone (see below).

Environmental drivers of community composition

A key question in ocean microbial ecology is the extent to which limited dispersal and historical contingency on the one hand, and global dispersion combined with selection by environmental factors on the other, are responsible for contemporary biogeographic patterns (*4*, *5*). The relationship between absolute latitude and biodiversity is an example of such a pattern, albeit one that is still controversial; while some authors found a negative correlation (*36*), others reported maxima in intermediate latitudinal ranges (*10*, *37*). The latter is supported by our findings (Fig. 4A), as an increase in richness with temperature was found from 4° to about 12°C, followed by a negative correlation for the remainder of the sampled temperature range (up to 30°C). This is also congruent with previous reports on oceanic groups of eukaryotes (*38*). A modeling study predicted season as a driver of biodiversity (*39*). For our data, however, the association of richness with temperature and latitude is robust to the confounding effect of seasonality (partial Mantel test, P < 0.01), although more data are needed for a rigorous statistical evaluation of such questions; for example, by periodically sampling the ocean across the globe on the same day (*40*). In addition to latitudinal biodiversity patterns, we found that taxonomic community dissimilarity increased up to about 5000 km within an ocean region (Fig. 4B). Together, our data support biogeographic patterns of microbial communities, in line with previous studies (*10*, *36*, *37*).

To further investigate the underlying mechanisms, we tested whether samples were more similar within than across ocean regions by focusing on surface samples only. If dispersal limitation rather than environmental selection dominated, we would expect a higher similarity within than across ocean regions. By contrast, if environmental selection explained biogeographic patterns, we would expect environmental factors to correlate with community similarity. Previous studies on selected ocean microbial taxa have shown a strong impact of light and temperature (*41*). For entire community assemblages, however, expectations are less clear. In a large-scale meta-analysis, salinity has been suggested as the major determinant across many (including ocean) ecosystems and to exceed the influence of temperature (*42*). In contrast, an analysis of functional trait composition in ocean environments suggested that temperature and light have stronger effects than nutrients or salinity (*10*, *43*).

A PCoA of taxonomic compositions of surface samples does not show a clear separation by regional origin, despite showing on average a higher similarity of communities within than across ocean regions (Fig. 5A). Instead, temperature was found to strongly correlate with PC1 (R^2 = 0.76). Thus, to verify the geographic independence of this pattern and to identify environmental drivers in our data set, we correlated distance-corrected dissimilarities of taxonomic and functional community composition with those of environmental factors (Fig. 5B). Overall, temperature and dissolved oxygen were the strongest correlates of both taxonomic and functional composition in the surface layer (Fig. 5B), while no significant correlation was found for salinity. Nutrients were only weakly correlated and, except for silicate, after the removal of a few extreme locations with very low temperatures, the correlations were not statistically significant.

Finally, we tackled the challenge of disentangling the high correlation between temperature and dissolved oxygen (R^2 = 0.87) in surface waters. To this end, we first used a machine learning–based approach (*44*) to independently model associations of each of these two factors with taxonomic and functional composition within surface samples (Fig. 6A). We then tested the strength of these associations in DCM layers, where the correlation between the two factors is much weaker (R^2 = 0.16), which allowed us to

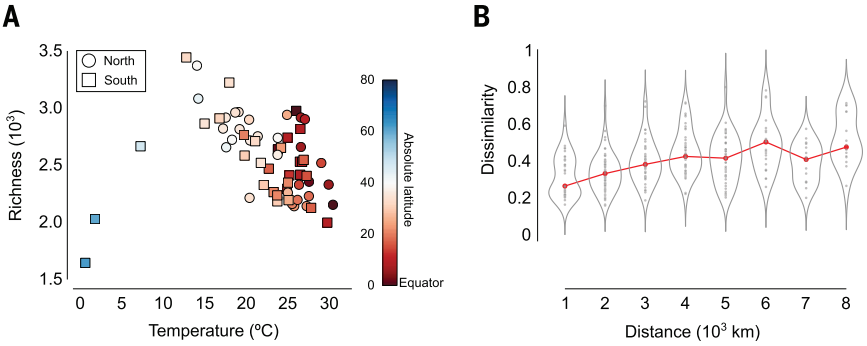

Fig. 4. Latitudinal diversity and distance decay of ocean microbial communities. (**A**) Plotting species richness against the temperature of sampling location shows an initial increase in richness up to about 15°C followed by a decrease toward warmer waters. Richness is highest in mid-latitudinal ranges rather than toward the equator. The color gradient denotes absolute latitudes (with increasing warmth of color from poles to equator). Shape of symbols denotes whether a sample originated from the Northern (circle) or Southern Hemisphere (square). (**B**) Pairwise microbial community dissimilarity (Bray-Curtis) based on relative $_{mi}$tag OTU abundances increases with distance between sampling stations up to about 5000 km. Pairwise distances were calculated only within ocean regions.

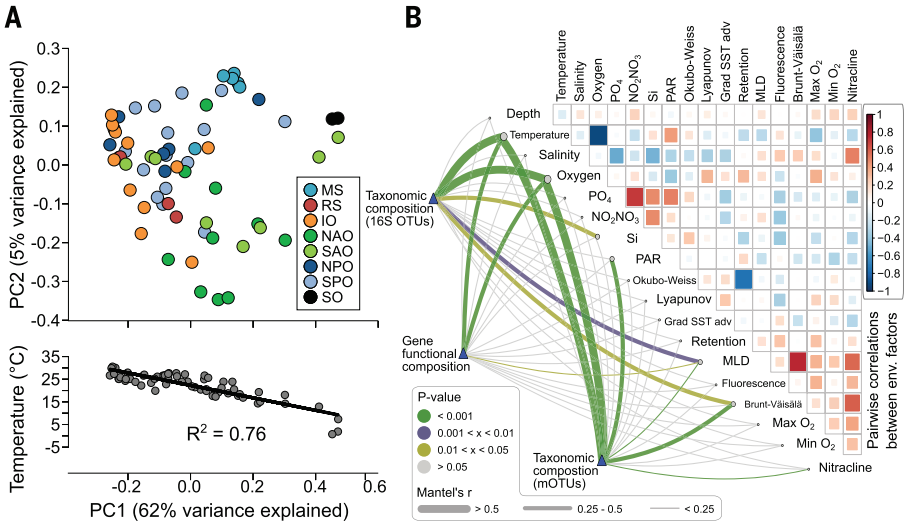

Fig. 5. Environmental drivers of surface microbial community composition. (**A**) Principal coordinate (PC) analysis of surface samples shows that samples are not clearly grouped by their regional origin (top), but rather separated by the local temperatures as shown by the strong correlation (R^2: 0.76) between the first PC and temperature (bottom). (**B**) Pairwise comparisons of environmental factors are shown, with a color gradient denoting Spearman's correlation coefficients. Taxonomic [based on two independent methods: $_{mi}$tags (*12*) and mOTUs (*13*)] and functional (based on biochemical KEGG modules) community composition was related to each environmental factor by partial (geographic distance–corrected) Mantel tests. Edge width corresponds to the Mantel's r statistic for the corresponding distance correlations, and edge color denotes the statistical significance based on 9,999 permutations.

effectively decouple dissolved oxygen from temperature. The surface-fitted model of temperature continued to achieve high prediction accuracy when applied at the DCM layers, whereas the oxygen model could not be generalized across depths. To illustrate the strength of these associations, we show that temperature could be predicted with an explained variance of 86%, using only species abundance as information (Fig. 6B). These results were validated with data from the GOS project (R^2 = 0.66) despite differences in sampling and sequencing procedures between the two studies (Fig. 6B).

Taken together, our data suggest that geographic distance plays a subordinate role and reveals temperature to be the major environmental factor shaping taxonomic and functional microbial community composition in the photic open ocean. Thus, a global dispersal potential for microorganisms (*45*) and subsequent environmental selection may, at least for some taxa, represent a mechanism for driving patterns of microbial biogeography. At the same time, localized adaptations by natural selection will lead to differences in spatially distant populations of phylogenetically similar organisms, so that characterizing these variations at strain-level resolution represents an important challenge for the future.

Core functional analysis between ecosystems

The generation of nonredundant gene abundance profiles from a large number (e.g., >100) of samples can be used to define a set of gene families, as a proxy for gene-encoded functions, which are ubiquitously found (core) in microbial communities. Such an analysis was performed for the human gut (*17*), which represents a fundamentally different microbial ecosystem (anoxic, host-associated, dominated by heterotrophs). However, owing to the lack of other large-scale, ecosystem-wide metagenomic data sets, it has been unknown how many of these core functions are shared with any other ecosystem. Thus, we first mapped the OM-RGC to known gene families, represented by clusters of orthologous groups [OGs, (*46*)] and selected prokaryotic genes to ensure comparability between the data sets. In total, we detected 39,246 OGs (19,524 OGs per sample; SD = 2682 OGs). Of those, the number of shared OGs rapidly decreased with sample size, reaching a minimum of 5755 ocean core OGs that were present in all (n = 139) prokaryote-enriched samples (Fig. 7A). Overall, we found that 40% of these ocean core OGs were of unknown function, compared to only 9% of the human gut core OGs (Fig. 7B).

We also sought to determine the overlap of core functions between the two ecosystems and to identify differentially abundant core functional categories (*47*), and contrast their relative importance in each of them (Fig. 7C). The ocean core contained almost twice as many OGs as the gut core, which may reflect the sampling of a greater number and higher complexity of niches in the ocean ecosystem than in the mostly anoxic, thermally stable human gut. However, despite large physicochemical differences between the two ecosystems, we found that most of the prokaryotic gene abundance (73% in the ocean; 63% in the gut) can be attributed to a shared functional core. Significant differential abundances between the two ecosystems were found across many functional categories. Most notably, those for defense mechanisms, signal transduction, and carbohydrate transport and metabolism were considerably more abundant in the gut, whereas those for transport mechanisms in general (coenzyme, lipid, nucleotide, amino acids, secondary metabolites) and energy production (including photosynthesis) were more abundant in the ocean (Fig. 7C).

Functional variability across ocean depths and regions

Functional redundancy across different taxa in microbial communities has been suggested to confer a buffering capacity for an ecosystem in scenarios of biodiversity loss (*48*). When contrasting taxonomic and functional variability in the ocean, we indeed found high taxonomic variability (even at phylum level) accompanied by relatively stable distributions of gene abundances summarized into functional categories (*47*) (Fig. 8A). This is also congruent with previous reports for the human gut, where gene abundances of metabolic pathways were found to be evenly distributed across samples, while taxonomic compositions varied markedly between subjects (*49*). Thus, despite the presumably greater environmental complexity in the ocean, the congruent functional redundancy observed in both ecosystems may be indicative of an ecosystem-independent property of microbial communities.

We next differentiated ocean core from noncore OGs, as the latter are more relevant for environment-specific adaptations. Within the ocean, 67% (SD = 5%) of the total gene abundance was attributed to ocean core OGs. After removing these and the 29% (SD = 5%) of gene abundance from genes that were not assigned to any OG, 4% (SD = 1%) remained as the noncore fraction. The abundance distribution among these noncore OGs, of which the largest fraction encode unknown functions, displayed a much greater variability across samples even when summarized into functional categories (Fig. 8A). Thus, in addition to the stable abundance distribution of core functional processes, as reported here and for human body habitats (*49*), functional variation similar in scale to that of the phylogenetic one can be detected when focusing on noncore, potentially adaptive gene families. As an example for such an environmental adaptation, we found an increase in lipid metabolism in oxygen minimum zones of the Eastern Pacific and Northern Indian Ocean (Fig. 8A).

Finally, to globally investigate the functional basis for the large community structural differences between the epipelagic layer and mesopelagic zone (Fig. 3A), we defined depth-specific core OGs using the approach introduced above. Unexpectedly, we found that the epipelagic core is almost completely contained in the mesopelagic core (Fig. 8B). When testing between-depth functional differences (Fig. 8B), we observed an enrichment of aerobic respiration genes in the ventilated mesopelagic zone, which is coherent with the finding that the mesopelagic zone is a key

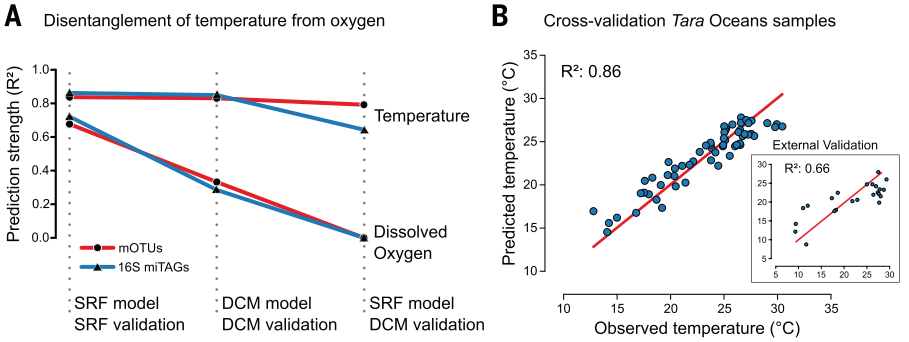

Fig. 6. Temperature as main environmental driver for microbial community composition in the epipelagic layer. (**A**) The strength of association between (meta)genomic and environmental data was tested by statistical models that were first generated with a subset of data for training and then validated on the remaining data. The prediction accuracy was used as a measure for the strength of association. Models that were trained on subsets of taxonomic data from surface water (SRF) samples could predict with high accuracy temperature and dissolved oxygen of samples used for validation (left). Models trained with subsets of taxonomic data from deep chlorophyll maximum (DCM) samples could predict temperature with high accuracy, but could predict dissolved oxygen with only moderate accuracy (middle). To demonstrate across-depth conservation of associations, we show that models trained on data from SRF samples could highly predict temperature, but failed to predict dissolved oxygen in DCM samples. (**B**) To illustrate prediction accuracy, and thus, strength of association between taxonomic composition (using 16S $_{mi}$tag abundances) and temperature, we show that in situ measured temperature could be predicted with 86% explained variance. The red diagonal shows the theoretical curve for perfect predictions. Sanger sequencing reads from the GOS project were used to calculate relative genus abundance tables. Using temperature prediction models trained at genus level using *Tara* Oceans data, we show (inset) that the results could be validated at relatively high accuracy given the large differences in sampling and sequencing methods between these two studies.

remineralization site of exported production (*50*). Flagellar assembly and chemotaxis were also enriched in mesopelagic samples, which is in contrast to previous findings (*51*) but congruent with the model that motility reduces grazing mortality in planktonic bacteria (*52*). In addition, these motility traits are potentially of great utility for bacteria in the dark ocean to colonize sinking particles or marine snow aggregates. Our taxonomic analysis (Fig. 2), combined with the detection of photosynthesis genes in the mesopelagic zone (Fig. 8B), indeed suggests microbial sedimentation from the epipelagic layer into the mesopelagic zone. Moving among aggregates to exploit nutrient patches and potentially new niches (*34*) may drive the diversification of mesopelagic zone–adapted microbial populations (*53*). In the future, matching *Tara* Oceans metatranscriptomic data should help in differentiating active from dead sinking biomass and give further insights into how microbial communities contribute to remineralization and carbon export into the ocean interior.

Conclusions

Tara Oceans has generated, in addition to global biodiversity resources for larger organismal size spectra (*20*), the OM-RGC, which makes ocean microbial genetic diversity accessible for various targeted analyses. Here we analyzed prokaryote-enriched size fractions, whereas related papers studied viral ecology (*19*), cross-kingdom species interactions (*21*), and planktonic community connectivity across an ocean circulation choke-point (*22*). Despite some limitations in the sampled organismal size range, oceanic depth layers, and temporal resolution, our approach generated an ecosystem-wide data set that will be useful for improving predictive models of the ocean. Finding that temperature drives microbial community variation and revealing the high functional redundancy in ocean microbial communities at global scale have wide-ranging implications for potential climate change–related effects. The *Tara* Oceans data set supports progress not only toward a holistic understanding of the ocean ecosystem but also of microbial communities in general, by facilitating comparative analyses between ecosystems.

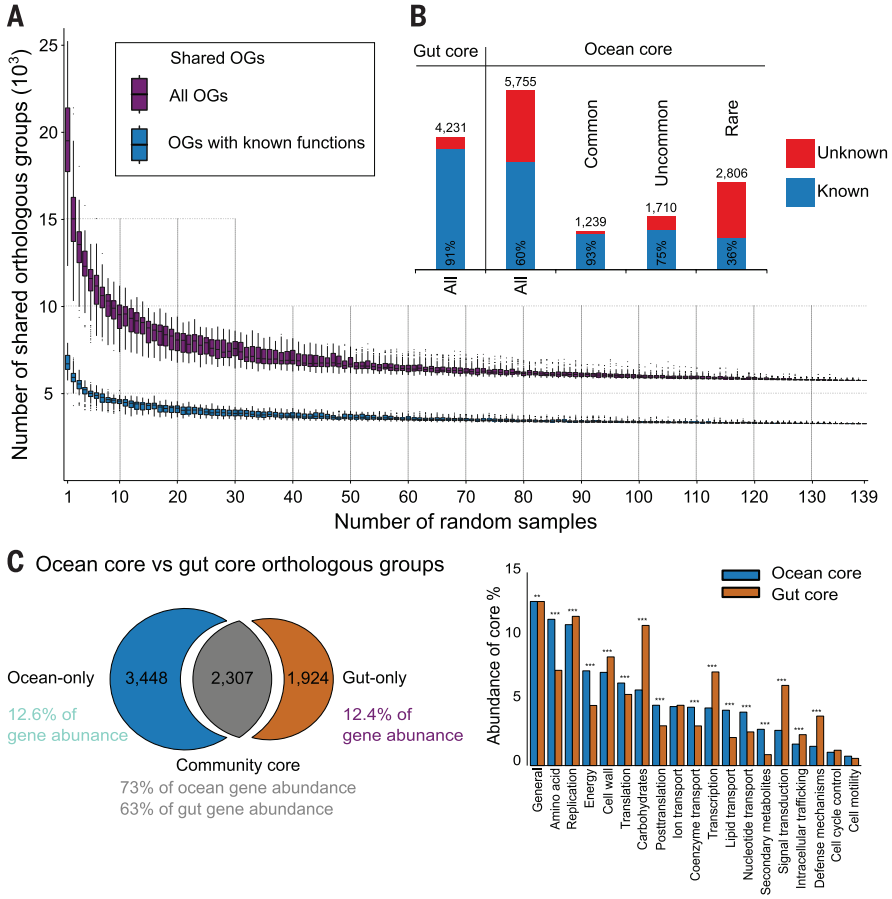

Fig. 7. Ocean versus human gut core orthologous groups. (**A**) The number of orthologous groups (OGs) that were shared among randomly selected sets of samples with sizes ranging from 1 to 139 was computed. With increasing sample size, the number of shared orthologous groups decreased first rapidly, then more gradually to a minimum of 5755 OGs at 139 samples, which was considered the set of ocean core OGs. Purple boxplots show the data for all OGs; blue boxplots show the data for OGs of known function. (**B**) Comparative statistics between ocean and human gut core OGs, showing that for a large fraction of ocean core OGs (40%), the functionality is unknown, which is in stark contrast to the human gut ecosystem (9%). Ocean core OGs are further subdivided into groups of OGs that are commonly (>50%), uncommonly (10% to 50%), or rarely (<10%) found in marine reference genomes. (**C**) A comparison of ocean and human gut core OGs (left) shows a large overlap of functions between these two fundamentally different ecosystems both qualitatively and quantitatively. The bar chart (right) displays a comparison of gene abundance summarized into OG functional categories to illustrate functional enrichments. Asterisks denote Mann-Whitney U-test results (**$P < 0.01$, ***$P < 0.001$).

Materials and methods

Sample and environmental data collection

From 2009 to 2013, morphological, genetic, and environmental data were collected at >200 sampling stations across all major oceanic provinces during the *Tara* Oceans expedition. The sampling strategy and methodology are described in (*54*–*57*). Sampling and enumeration of heterotrophic prokaryotes, phototrophic picoplankton, and small eukaryotes by flow cytometry followed previously described procedures, which are summarized in (*58*). Sample provenance is described in table S1 and in (*55*). Sample-associated environmental data and sample-associated biodiversity indexes were inferred at the depth of sampling (*56*, *57*), and additional information is available at (*14*).

Extraction and sequencing of metagenomic DNA

Metagenomic DNA from prokaryote and girus-enriched size fraction filters, and from precipitated viruses, was extracted as described in (*12*), (*59*), and (*19*), respectively. DNA (30 to 50 ng) was sonicated to a 100– to 800–base pair (bp) size range. DNA fragments were subsequently end repaired and 3′-adenylated before Illumina adapters were added by using the NEBNext Sample Reagent Set (New England Biolabs). Ligation products were purified by Ampure XP (Beckmann Coulter), and DNA fragments (>200 bp) were PCR-amplified with Illumina adapter-specific primers and Platinum Pfx DNA polymerase (Invitrogen). Amplified library fragments were size selected (~300 bp) on a 3% agarose gel. After library profile analysis using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and quantitative PCR (MxPro, Agilent Technologies, USA), each library was sequenced with 101 base-length read chemistry in a paired-end flow cell on Illumina sequencing machines (Illumina, USA).

Metagenomic sequence assembly and gene predictions

Using MOCAT (version 1.2) (*18*), high-quality (HQ) reads were generated (option *read_trim_filter*; solexaqa with length cut-off 45 and quality cut-off 20) and reads matching Illumina sequencing adapters were removed (option *screen_fastafile* with e-value 0.00001). Screened HQ reads were assembled (option *assembly*; minimum length 500 bp), and gene-coding sequences [minimum length 100 nucleotides (nt)] were predicted on the assembled scaftigs [option *gene_prediction*;

MetaGeneMark (version 2.8) (*60*)], generating a total of 111.5 M gene-coding sequences (*14*). Assembly errors were estimated by testing for colinearity between assembled contigs and genes and unassembled 454 sequencing reads by using a subset of 11 overlapping samples (*58*). From this analysis, we estimate that 1.5% of contigs had breakpoints and thus may suffer from errors (*14*). This error rate is more than a factor of 6.5 less than previous estimates of contig chimericity in simulated metagenomic assemblies (9.8%) with similar N_{50} values (*61*).

Generation of the ocean microbial reference gene catalog

Predicted gene-coding sequences were combined with those identified in publicly available ocean metagenomic data and reference genomes: 22.6 M predicted genes from the GOS expedition (*6*, *7*), 1.78 M from Pacific Ocean Virome study (POV) (*62*), 14.8 thousand from viral genomes from the Marine Microbiology Initiative (MMI) at the Gordon & Betty Moore Foundation (*14*), and 1.59 M from 433 ocean microbial reference genomes (*14*). The reference genomes were selected by the following procedure: An initial set of 3496 reference genomes (all high-quality genomes available as of 23 February 2012) was clustered into 1753 species (*24*), from each of which we selected one representative genome. After mapping all HQ reads against these genomes, a genome was selected if the base coverage was >1× or if the fraction of genome coverage was >40% in at least one sample. In addition, we included prokaryotic genomes for which habitat entries matched the terms "Marine" or "Sea Water" in the Integrated Microbial Genomes database (*63*) or if a genome was listed under the Moore Marine Microbial Sequencing project (*64*) as of 29 July 2013. Finally, we applied previously established quality criteria (*24*), resulting in a final set of 433 ocean microbial reference genomes (*14*). For data from GOS, POV, and MMI, assemblies were downloaded from the CAMERA portal (*64*). A total of 137.5 M gene-coding nucleotide sequences were clustered by using the same criteria as in (*16*); i.e., 95% sequence identity and 90% alignment coverage of the shorter sequence. The longest sequence of each cluster was selected, and after removing sequences <100 nt, we obtained a set of 40,154,822 genes [i.e., nonredundant contiguous gene-coding nucleotide sequences operationally defined as "genes"; see also (*16*, *17*)] that we refer to as the Ocean Microbial Reference Gene Catalog (OM-RGC).

Taxonomic and functional annotation of the OM-RGC

We taxonomically annotated the OM-RGC using a modified dual BLAST-based last common ancestor (2bLCA) approach as described in (*58*). For modifications, we used RAPsearch2 (*65*) rather than BLAST to efficiently process the large data volume and a database of nonredundant protein sequences from UniProt (version: UniRef_2013_07) and eukaryotic transcriptome data not represented in UniRef. The OM-RGC was functionally annotated to orthologous groups in the eggNOG (version 3) and KEGG databases (version 62) with SmashCommunity (version 1.6) (*46*, *66*, *67*). In total, 38% and 57% of the genes could be annotated by homology to a KEGG ortholog group (KO) or an OG, respectively. Functional modules were defined by selecting previously described key marker genes for 15 selected ocean-related processes, such as photosynthesis, aerobic respiration, nitrogen metabolism, and methanogenesis (*14*).

Taxonomic profiling using 16*S* tags and metagenomic operational taxonomic units 16*S* fragments directly identified in Illumina-sequenced metagenomes ($_{mi}$tags) were identified as described in (*12*). 16*S* $_{mi}$tags were mapped to cluster centroids of taxonomically annotated 16*S* reference sequences from the SILVA database (*23*) (release 115: SSU Ref NR 99) that had been clustered at 97% sequence identity with USEARCH v6.0.307 (*68*). 16*S* $_{mi}$tag counts were normalized by the total sum for each sample. In addition, we identified protein-coding marker genes suitable for metagenomic species profiling using fetchMG (*13*) in all 137.5 M gene-coding sequences and clustered them into metagenomic operational taxonomic units (mOTUs) that group organisms into species-level clusters at higher accuracy than 16*S* OTUs as described in (*13*, *24*). Relative abundances of mOTU linkage groups were quantified with MOCAT (version 1.3) (*18*).

Functional profiling using the OM-RGC

Gene abundance profiles were generated by mapping HQ reads from each sample to the OM-RGC (MOCAT options *screen* and *filter* with length and identity cutoffs of 45 and 95%, respectively, and paired-end filtering set to *yes*). The abundance of each reference gene in each sample was calculated as gene length–normalized base and insert counts (MOCAT option *profile*). Functional abundances were calculated as the sum of the relative abundances of reference genes, or key

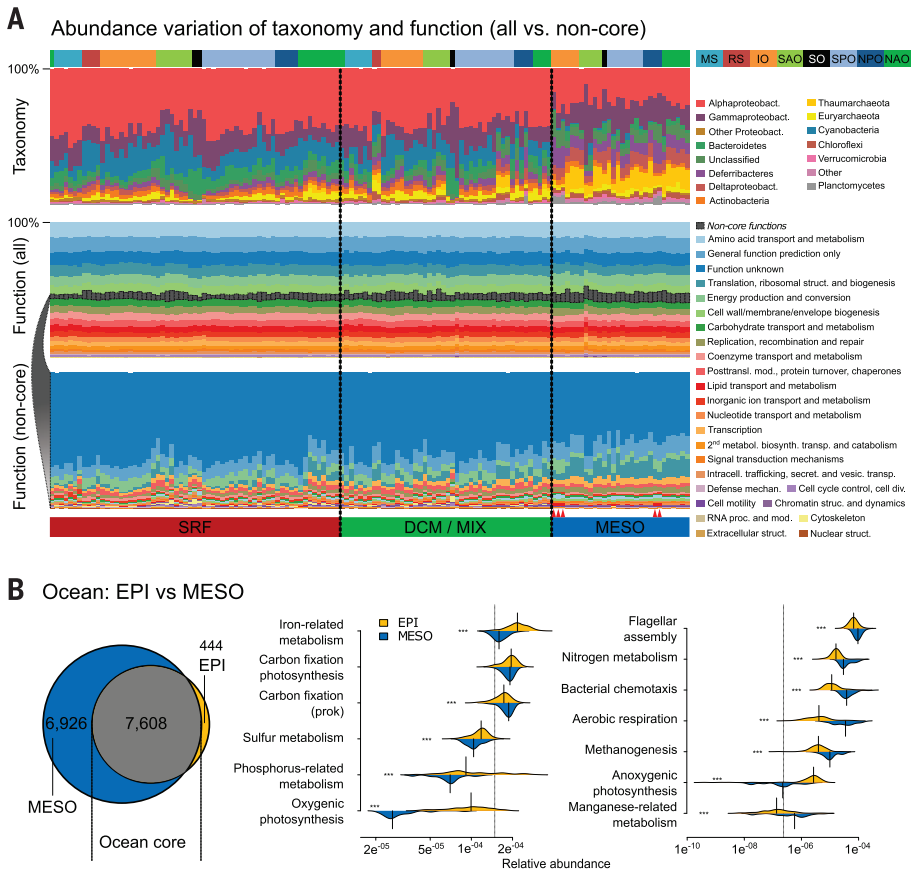

Fig. 8. Functional structuring of the ocean microbiome. (**A**) Phylum-level (class-level for Proteobacteria) taxonomic variability is higher (top, median relative SD = 65%) relative to the functional composition (OG functional categories) of ocean microbial samples (center, median relative SD = 7%). Removal of functions that are ubiquitous in the ocean environment reveals the variable, noncore fraction (bottom, median relative SD = 47%), which amounts on average to 4% of the total gene abundance. Red triangles on *x* axis highlight mesopelagic samples collected in oxygen minimum zones of the Indian Ocean and Eastern Pacific, which show increased levels of lipid metabolism in noncore functions. (**B**) Venn diagram (left) showing that core OGs in the epipelagic layer of the ocean are almost completely contained in mesopelagic core OGs (left). The bean charts (right) display differential abundances of marker genes (based on KO annotations) for selected functional processes in the ocean. Asterisks denote Mann-Whitney U test results (***$P < 0.001$).

marker genes (*14*), annotated to different functional groups (OGs, KOs, and KEGG modules). For each functional module, the abundance was calculated as the sum of relative abundances of marker KOs normalized by the number of KOs. For comparative analyses with the human gut ecosystem, we used the subset of the OM-RGC that was annotated to Bacteria or Archaea (24.4 M genes). Using a rarefied (to 33 M inserts) gene count table, an OG was considered to be part of the ocean microbial core if at least one insert from each sample was mapped to a gene annotated to that OG. Samples from the human gut ecosystem were processed similarly, and a list of all OGs that were defined in either the ocean or the gut as core is provided in (*14*).

Microbial community structural analyses and prediction of minimum generation times

16*S* $_{mi}$tag counts were rarefied 100 times to the minimum number of total 16*S* $_{mi}$tags per sample (39,410), and OTU richness and Chao1 richness estimators were calculated as the mean of all rarefactions (*14*). A phylogenetic tree of 16*S* $_{mi}$tags was calculated from full-length 16*S* sequences, by using parts of the LotuS 16*S* pipeline (*69*). This phylogenetic tree was midpoint rooted in R and used with the $_{mi}$tag abundance matrix rarefied to 39,000 reads per sample to calculate Faith's phylogenetic diversity (*70*) as the mean value of five repetitions (*14*). Similarly, OG richness was computed as the average of 10 rarefactions (*14*). Community growth potential from genomic traits was estimated as the average minimum generation time of the organisms present in the sample, weighted by their abundance, as previously described (*32*).

Distance correlations between genomic and environmental data

We computed pairwise distances between samples on the basis of (i) relative abundances of taxonomic (16*S* $_{mi}$tags and mOTUs) and gene functional compositions (at KEGG module level)—the compositional data; (ii) in situ measurements of physicochemical data—the environmental data; and (iii) geographic location of sampling stations—the geographic data. Data from the three southernmost stations were removed from the analysis, as these stations are outside the range of the rest of the data in parameters such as temperature, oxygen, and nutrients. For compositional data, we applied a logarithmic transformation to relative abundances using the function $\log_{10}(x + x_0)$, where x is the original relative abundance and x_0 is a small constant, and $x_0 < \min(x)$.

We applied an additional low-abundance filter, which removed features whose relative abundance did not exceed 0.0001 in any sample. Environmental data were transformed to z-scores before calculating distances. We used Euclidean distances for compositional and environmental data and Haversine distances for geographic data. Given these distance matrices, we computed partial Mantel correlations between compositional and environmental data given geographic distance (9,999 permutations) using the *vegan* R software package. Partial Mantel tests were also performed between species richness and both temperature and latitude, while controlling for season.

Statistical modeling and correlation analysis

Compositional data (see above) were normalized to ranks across samples and then used to learn a regression model to predict environmental measures. In particular, we fitted an elastic net model (*44*) using inner cross-validation to set the hyperparameters as implemented by the *scikit-learn* Python package (*71*). For spatial autocorrelation-corrected cross-validation, samples from each ocean basin were iteratively held out for testing on a model learned from the rest of the samples.

As a measure of association between the environmental parameter and the compositional data, we computed the cross-validated R^2 (also known as Q^2) (*72*), defined as $1 - \sum \frac{(y_i - \widehat{y}_i)^2}{(y_i - \overline{y}_i)^2}$, where y_i is the value of the parameter for sample i, $\widehat{y}_i$ is the prediction for that same sample (obtained by held-out cross-validation), and $\overline{y}$ is the overall mean (the summation runs over all the samples). To disentangle effects of temperature and oxygen, we trained models on surface samples, which were then evaluated in DCM samples. Again, to avoid spatial autocorrelation, cross-validation by ocean basin was used. An external cross-validation was performed by classifying GOS reads using the RDP database (*73*). Only genera detected in both studies were considered. Because of the lower and varying sequencing depth of the GOS data, for each GOS sample, we downsampled *Tara* Oceans data to match the corresponding sequencing depth and learned a model based on this downsampled data set. This model was based on the presence or absence of the taxa (which was modeled by passing a binary input matrix to the elastic net fitting routines).

REFERENCES AND NOTES

1. P. G. Falkowski, R. T. Barber, V. Smetacek, Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**, 200–206 (1998). doi: 10.1126/science.281.5374.200; pmid: 9660741
2. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578–6583 (1998). doi: 10.1073/pnas.95.12.6578; pmid: 9618454
3. S. C. Doney *et al.*, Climate Change Impacts on Marine Ecosystems. *Annu. Rev. Mar. Sci.* **4**, 11–37 (2012). doi: 10.1146/annurev-marine-041911-111611
4. J. A. Fuhrman, Microbial community structure and its functional implications. *Nature* **459**, 193–199 (2009). doi: 10.1038/nature08058; pmid: 19444205
5. J. B. H. Martiny *et al.*, Microbial biogeography: Putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006). doi: 10.1038/nrmicro1341; pmid: 16415926
6. D. B. Rusch *et al.*, The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLOS Biol.* **5**, e77 (2007). doi: 10.1371/journal.pbio.0050077; pmid: 17355176
7. S. Yooseph *et al.*, The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLOS Biol.* **5**, e16 (2007). doi: 10.1371/journal.pbio.0050016; pmid: 17355171
8. A. Barberán, A. Fernández-Guerra, B. J. M. Bohannan, E. O. Casamayor, Exploration of community traits as ecological markers in microbial metagenomes. *Mol. Ecol.* **21**, 1909–1917 (2012). doi: 10.1111/j.1365-294X.2011.05383.x; pmid: 22121910
9. T. A. Gianoulis *et al.*, Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1374–1379 (2009). doi: 10.1073/pnas.0808022106; pmid: 19164758
10. J. Raes, I. Letunic, T. Yamada, L. J. Jensen, P. Bork, Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**, 473 (2011). doi: 10.1038/msb.2011.6; pmid: 21407210
11. E. Karsenti *et al.*, A holistic approach to marine eco-systems biology. *PLOS Biol.* **9**, e1001177 (2011). doi: 10.1371/journal.pbio.1001177; pmid: 22028628
12. R. Logares *et al.*, Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2014). doi: 10.1111/1462-2920.12250; pmid: 24102695
13. S. Sunagawa *et al.*, Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013). doi: 10.1038/nmeth.2693; pmid: 24141494
14. Companion website tables W1 to W8, data, and information are available at http://ocean-microbiome.embl.de/companion.html
15. Human Microbiome Project Consortium, *Nature* **486**, 215–221 (2012). doi: 10.1038/nature11209; pmid: 22699610
16. J. Li *et al.*, An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014). doi: 10.1038/nbt.2942; pmid: 24997786
17. J. Qin *et al.*, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010). doi: 10.1038/nature08821; pmid: 20203603
18. J. R. Kultima *et al.*, MOCAT: A metagenomics assembly and gene prediction toolkit. *PLOS ONE* **7**, e47656 (2012). doi: 10.1371/journal.pone.0047656; pmid: 23082188
19. J. R. Brum *et al.*, Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
20. C. de Vargas *et al.*, Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
21. G. Lima-Mendez *et al.*, Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
22. E. Villar *et al.*, Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* **348**, 1261447 (2015).
23. C. Quast *et al.*, The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013). doi: 10.1093/nar/gks1219; pmid: 23193283
24. D. R. Mende, S. Sunagawa, G. Zeller, P. Bork, Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013). doi: 10.1038/nmeth.2575; pmid: 23892899
25. L. Zinger *et al.*, Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLOS ONE* **6**, e24570 (2011). doi: 10.1371/journal.pone.0024570; pmid: 21931760
26. L. Amaral-Zettler *et al.*, in *Life in the World's Oceans*, A. D. McIntyre, Ed. (Wiley-Blackwell, Oxford, 2010), pp. 221–245.
27. C. L. Dupont *et al.*, Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012). doi: 10.1038/ismej.2011.189; pmid: 22170421
28. R. M. Morris *et al.*, SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002). doi: 10.1038/nature01240; pmid: 12490947
29. K. Lochte, C. M. Turley, Bacteria and cyanobacteria associated with phytodetritus in the deep sea. *Nature* **333**, 67–69 (1988). doi: 10.1038/333067a0
30. S. J. Giovannoni, U. Stingl, Molecular diversity and ecology of microbial plankton. *Nature* **437**, 343–348 (2005). doi: 10.1038/nature04158; pmid: 16163344
31. 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). doi: 10.1038/ismej.2014.143; pmid: 25093636
32. S. Vieira-Silva, E. P. C. Rocha, The systemic imprint of growth and its uses in ecological (meta)genomics. *PLOS Genet.* **6**, e1000808 (2010). doi: 10.1371/journal.pgen.1000808; pmid: 20090831
33. T. Pommier *et al.*, Spatial patterns of bacterial richness and evenness in the NW Mediterranean Sea explored by pyrosequencing of the 16S rRNA. *Aquat. Microb. Ecol.* **61**, 221–233 (2010). doi: 10.3354/ame01484
34. R. Stocker, Marine microbes see a sea of gradients. *Science* **338**, 628–633 (2012). doi: 10.1126/science.1208929; pmid: 23118182

35. J. Pernthaler, Predation on prokaryotes in the water column and its ecological implications. *Nat. Rev. Microbiol.* **3**, 537–546 (2005). doi: 10.1038/nrmicro1180; pmid: 15953930
36. 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). doi: 10.1073/pnas.1212424110; pmid: 23324742
37. J. A. Fuhrman *et al.*, A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 7774–7778 (2008). doi: 10.1073/pnas.0803070105; pmid: 18509059
38. D. P. Tittensor *et al.*, Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010). doi: 10.1038/nature09329; pmid: 20668450
39. J. Ladau *et al.*, Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* **7**, 1669–1677 (2013). doi: 10.1038/ismej.2013.37; pmid: 23514781
40. Ocean Sampling Day, www.oceansamplingday.org.
41. Z. I. Johnson *et al.*, Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006). doi: 10.1126/science.1118052; pmid: 16556835
42. C. A. Lozupone, R. Knight, Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 11436–11440 (2007). doi: 10.1073/pnas.0611525104; pmid: 17592124
43. D. P. Herlemann *et al.*, Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* **5**, 1571–1579 (2011). doi: 10.1038/ismej.2011.41; pmid: 21472016
44. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005). doi: 10.1111/j.1467-9868.2005.00503.x
45. B. J. Finlay, Global dispersal of free-living microbial eukaryote species. *Science* **296**, 1061–1063 (2002). doi: 10.1126/science.1070710; pmid: 12004115
46. S. Powell *et al.*, eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–D289 (2012). doi: 10.1093/nar/gkr1060; pmid: 22096231
47. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000). doi: 10.1093/nar/28.1.33; pmid: 10592175
48. T. Bell, J. A. Newman, B. W. Silverman, S. L. Turner, A. K. Lilley, The contribution of species richness and composition to bacterial services. *Nature* **436**, 1157–1160 (2005). doi: 10.1038/nature03891; pmid: 16121181
49. Human Microbiome Project Consortium, *Nature* **486**, 207–214 (2012). doi: 10.1038/nature11234; pmid: 22699609
50. J. Arístegui, J. M. Gasol, C. M. Duarte, G. J. Herndl, Microbial oceanography of the dark ocean's pelagic realm. *Limnol. Oceanogr.* **54**, 1501–1529 (2009). doi: 10.4319/lo.2009.54.5.1501
51. E. F. DeLong *et al.*, Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006). doi: 10.1126/science.1120250; pmid: 16439655
52. C. Matz, K. Jürgens, High motility reduces grazing mortality of planktonic bacteria. *Appl. Environ. Microbiol.* **71**, 921–929 (2005). doi: 10.1128/AEM.71.2.921-929.2005; pmid: 15691949
53. Y. Yawata *et al.*, Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5622–5627 (2014). doi: 10.1073/pnas.1318943111; pmid: 24706766
54. S. Pesant *et al.*, Open science resources for the discovery and analysis of *Tara* Oceans Data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
55. *Tara* Oceans Consortium, Coordinators; *Tara* Oceans Expedition, Participants (2014): Registry of selected samples from the *Tara* Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840721
56. S. Chaffron *et al.*, (2014): Contextual environmental data of selected samples from the *Tara* Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840718
57. S. Chaffron *et al.*, (2014): Contextual biodiversity data of selected samples from the *Tara* Oceans Expedition (2009-2013). doi: 10.1594/PANGAEA.840698
58. P. Hingamp *et al.*, Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013). doi: 10.1038/ismej.2013.59; pmid: 23575371
59. C. Clerissi *et al.*, Unveiling of the diversity of Prasinoviruses (*Phycodnaviridae*) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. *Appl. Environ. Microbiol.* **80**, 3150–3160 (2014). doi: 10.1128/AEM.00123-14; pmid: 24632251
60. W. Zhu, A. Lomsadze, M. Borodovsky, Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132–e132 (2010). doi: 10.1093/nar/gkq275; pmid: 20403810
61. D. R. Mende *et al.*, Assessment of metagenomic assembly using simulated next generation sequencing data. *PLOS ONE* **7**, e31386 (2012). doi: 10.1371/journal.pone.0031386; pmid: 22384016
62. 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). doi: 10.1111/j.1462-2920.2012.02836.x; pmid: 22845467
63. Integrated Microbial Genomes database: https://img.jgi.doe.gov/cgi-bin/w/main.cgi.
64. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, http://camera.calit2.net/microgenome.
65. Y. Zhao, H. Tang, Y. Ye, RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**, 125–126 (2012). doi: 10.1093/bioinformatics/btr595; pmid: 22039206
66. M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, P. Bork, SmashCommunity: A metagenomic annotation and analysis tool. *Bioinformatics* **26**, 2977–2978 (2010). doi: 10.1093/bioinformatics/btq536; pmid: 20959381
67. M. Kanehisa *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36** (Database), D480–D484 (2008). doi: 10.1093/nar/gkm882; pmid: 18077471
68. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). doi: 10.1093/bioinformatics/btq461; pmid: 20709691
69. F. Hildebrand, R. Tadeo, A. Y. Voigt, P. Bork, J. Raes, LotuS: An efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014). doi: 10.1186/2049-2618-2-30
70. D. P. Faith, Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992). doi: 10.1016/0006-3207(92)91201-3
71. F. Pedregosa *et al.*, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
72. H. Wold, in *Systems Under Indirect Observation: Causality, Structure, Prediction,* K. G. Jöreskog, Ed. (North-Holland, 1982), vol. 2, pp. 1-54.
73. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007). doi: 10.1128/AEM.00062-07; pmid: 17586664

ACKNOWLEDGMENTS

We thank the following individuals and sponsors for their support: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, Università degli Studi di Milano-Bicocca, 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), Agence Nationale de la Recherche (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), European Research Council Advanced Grant Award to C.B. (Diatomite: 294823), Gordon and Betty Moore Foundation grant (no. 3790) to M.B.S., Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to SGA, Japan Society for the Promotion of Science KAKENHI grant no. 26430184 to H.O., and FWO, BIO5, Biosphere 2 to M.B.S. We also thank the following for their support: 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, and the *Tara* schooner and its captain 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 that 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 G. Cochrane and P. ten Hoopen, as well as the EMBL Advanced Light Microscopy Facility (ALMF), in particular R. 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. Data described herein are available at http://ocean-microbiome.embl.de/companion.html, at the EBI under the project identifiers PRJEB402 and PRJEB7988, and at PANGAEA (*55*–*57*). The data release policy regarding future public release of *Tara* Oceans data is described in (*54*). All authors approved the final manuscript. This article is contribution number 22 of *Tara* Oceans. Additional data are in the supplementary materials.

TARA OCEANS COORDINATORS

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

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)–CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, ME, USA. [4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, F-75005 Paris, France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [9]Sorbonne Universités, UPMC Université Paris 06, 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, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [18]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [27]Sorbonne Universités, UPMC Université Paris 06, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA. [29]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [30]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261359/suppl/DC1
Table S1

16 September 2014; accepted 24 February 2015
10.1126/science.1261359