Revised Manuscript: Confidential 29 January 2015

Colors used throughout the revision of pre-edited manuscript

Edits by Science editor in blue

Edits by the Authors in green

Title: Structure and Function of the Global Ocean Microbiome

Authors:

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[5,26,27], Fabrice Not[15,16], Hiroyuki Ogata[21], Stephane Pesant[22,23], Sabrina Speich[24,25], Lars Stemmann[8,9], Matthew B. Sullivan[13], Jean Weissenbach[5,26,27], Patrick Wincker[5,26,27], Eric Karsenti[10,17,*], Jeroen Raes[2,3,4,*], Silvia G. Acinas[6,*], Peer Bork[1,28*]

Affiliations:

[1] Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 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 Science (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain.
[7] Sorbonne Universités, UPMC, Univ Paris 06, CNRS-IRD-MNHN, LOCEAN Laboratory, 4 Place Jussieu, 75005, Paris, France.
[8] CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-mer, France.
[9] Sorbonne Universités, UPMC Univ 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, Paris, F-75005 France.
[11] Laboratoire de Physique des Océan UBO-IUEM Palce Copernic 29820 Polouzané, France.
[12] Aix Marseille Université CNRS IGS UMR 7256 13288 Marseille France.
[13] Department of Ecology and Evolutionary Biology, University of Arizona, 1007 E Lowell Street, Tucson, AZ, 85721, USA.
[14] Department of Hydrobiology, Federal University of São Carlos (UFSCar), Rodovia Washington Luiz, 13565-905 - São Carlos, SP – Brazil.
[15] CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
[16] Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
[17] Directors' Research, European Molecular Biology Laboratory, 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] Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan.
[22] PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.
[23] MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.
[24] Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris Cedex 05 France.
[25] Laboratoire de Physique des Océan UBO-IUEM Palce Copernic 29820 Polouzané, France.
[26] CNRS, UMR 8030, CP5706, Evry France.
[27] Université d'Evry, UMR 8030, CP5706, Evry France.
[28] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany.

[‡] *Tara* Oceans coordinators and affiliations are listed at the end of this manuscript.
[†] These authors contributed equally to this work
[*] Correspondence to: sunagawa@embl.de; karsenti@embl.de; jeroen.raes@vib-kuleuven.be; sacinas@icm.csic.es; bork@embl.de

Abstract: 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 non-redundant, 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, given the physicochemical differences, a surprisingly high fraction of its abundance (>73%) to be shared with the human gut microbiome.

One Sentence Summary: *Tara* Oceans provides a gene catalogue and analysis of ocean microbes in their environmental context across three depth layers at global scale.

Main Text: 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 - 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*), which 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 a large-scale 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 ca. 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 previously not 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, which 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 (<0.2 µm): *n*=45; girus/prokaryote-enriched fractions (0.1-0.2 µm, 0.2-0.45 µm, 0.45-0.8 µm): *n*=59; prokaryote-enriched fractions (0.2-1.6 µm, 0.2-3 µm): *n*=139) were paired-end shotgun Illumina sequenced to generate a total of more than 7.2 terabases (29.6 ± 12.7 gigabases (Gb) per sample) of metagenomic data (*14*), which are in the same order of magnitude as data from the US Human Microbiome Project (phase I) and the European Metagenomics of the Human Intestinal Tract project combined (*15-17*).

To generate an ocean microbial reference gene catalog (see also (*17, 18*)), we first reconstructed the genomic content of these new data by metagenomic assembly and gene prediction (*19*), and combined these results with publicly available ocean metagenomic data and reference genomes (*14*). Specifically, approximately 111.5 million (M) protein-coding nucleotide sequences were predicted from *Tara* Oceans metagenomes, which were clustered at 95% nucleotide sequence identity with 24.4 M sequences from other ocean metagenomes and 1.6 M sequences from ocean prokaryotic (*n*=433) and viral (*n*=121) reference genomes (*14*). This resulted in a global Ocean Microbial Reference Gene Catalog (OM-RGC), which comprises >40 M non-redundant representative genes from viruses, prokaryotes and picoeukaryotes (Fig. 1B).

Compared to a human gut microbial reference gene catalog (*18*), the OM-RGC comprises more than four times the number of genes, most of which (72.3% of the annotated fraction) appear prokaryotic (Fig. 1B). 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 one sample only amounted to 3.6%, which may originate from localized specialists.

To complement the work of *Tara* Oceans Consortium partners who analyzed viral and protist-enriched size fractions (*20, 21*) and integrated data across domains of life (*22, 23*), we focused our analyses on 139 prokaryote-enriched samples, which included: 63 surface water samples (5 m; s.d. 0 m), 46 epipelagic subsurface water samples mostly from the DCM (71 m; s.d. 41 m), and 30 mesopelagic samples (600 m; s.d. 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 it could also be due to seasonal and locational differences of sampling in relatively well-studied regions.

To put the degree of taxonomic novelty into context, we extracted a total of >14 M metagenomic 16S rRNA gene tags (16S $_{\mathrm{mi}}$tags; (*12*)) and mapped these to operational taxonomic units (OTUs) based on 97% sequence identity clustering of reference 16S sequences (*24*). 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 (*25*). The fraction of total 16S $_{\mathrm{mi}}$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 been already captured by 16S rRNA sequencing, the OM-RGC now provides a link to their genomic content.

Diversity and Stratification of the Ocean Microbiome

Given the global scale of *Tara* Oceans samples, we interrogated our data set for the composition and stratifying factors of ocean microbial communities. Taxonomic and phylogenetic diversity were highly (R^2=0.96) correlated (*14*) and 16S $_{mi}$tags identified in our metagenomic data set mapped to a total of 35,650 OTUs (2,937 OTUs; s.d. 585 OTUs). The total richness estimate of 37,470 is comparable to the numbers from a previous study, which detected about 44,500 OTUs based on PCR-amplified 16S tags from 356 globally distributed pelagic samples (*26*) that were collected in the context of the International Census of Marine Microbes (ICoMM) project (*27*). At phylum level, more than 93% of 16S $_{mi}$tags could be annotated. We found typical members of Proteobacteria, including the ubiquitous clades SAR11 (Alphaproteobacteria) and SAR86 (Gammaproteobacteria), to dominate the sampled areas of the ocean both in terms of relative abundance and taxonomic richness (*28, 29*). 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 (*30*).

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 several studies that have reported a vertical stratification of microbial taxa and viruses according to changes in physico-chemical parameters, such as light, temperature and nutrients (*31, 32*). Given the 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 of both taxonomic and functional richness with depth while cell abundance, as measured by flow cytometry, and potential maximum growth rates (*33*) 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 (*34*), 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 micro-environments in the mesopelagic zone (*35*). In addition, slower growth, due 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 (*36*). Indeed, our metagenomic analysis provides molecular support for these models by identifying a significant 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 to which extent 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 for such a pattern, albeit being still controversial; while some authors found a negative correlation (*37*), others reported maxima in intermediate latitudinal ranges (*10, 38*). The latter is supported by our findings (Fig. 4A), as an increase in richness with temperature was found from 4 ºC to about 12 ºC, followed by a negative correlation for the remainder of the sampled temperature range (up to 30 ºC). This is also congruent with previous reports on oceanic groups of eukaryotes (*39*). A modeling study predicted season as a driver of biodiversity (*40*). For our data, however, the association of richness with temperature and latitude is robust to the confounding effect of seasonality (partial Mantel test, p-value < 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 (*41*). In addition to latitudinal biodiversity patterns, we found taxonomic community dissimilarity to increase up to about

5,000 km within an ocean region (Fig. 4B). Together, these findings support biogeographic patterns of microbial communities in line with a number of previous studies (*10, 37, 38*).

To further investigate the underlying mechanisms, we investigated 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. On the other hand, 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 (*42*). 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, exceeding the influence of temperature (*43*). In contrast to this, an analysis of functional trait composition in ocean environments suggested temperature and light to have stronger effects than nutrients or salinity (*10, 44*).

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 identify environmental drivers in our data set, we correlated geographic 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 epipelagic layer (Fig. 5B and below), 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 (*45*) to independently model associations of each of these two factors with taxonomic/functional composition within surface samples (Fig. 6A). We then tested the strength of these associations in DCM layers, where correlations between the two factors were much weaker (R^2=0.16), which allowed us to effectively decouple dissolved oxygen from temperature. The surface-fitted model of temperature continues to achieve high prediction accuracy when applied at the DCM layers. The oxygen model, on the other hand, cannot generalize 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 using data from the GOS project (R^2=0.66) despite a number of differences in sampling and sequencing procedures between these two studies (Fig. 6B).

Taken together, our data suggest geographic distance to play a subordinate role and reveal temperature to be the major environmental factor in shaping taxonomic and functional microbial community compositions in the photic open ocean. Thus, a global dispersal potential for microorganisms (*46*) 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, and characterizing these variations at strain-level resolution represents an important challenge for the future.

Core Functional Analysis Between Ecosystems

The generation of non-redundant 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, due to the lack of other large-scale, ecosystem-wide metagenomic data sets, it has been unknown how much 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, (*47*)), and selected prokaryotic genes to ensure comparability between the data sets. In total, we detected 39,246 OGs (19,524 OGs; s.d. 2,682 OGs). Of those, the number of shared OGs rapidly decreased with sample size reaching a minimum of 5,755 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 shared core functions between these two very different ecosystems and to identify differentially abundant core functional categories (*48*) to contrast their relative importance in each one of them (Fig. 7C). Despite large physicochemical differences between the two ecosystems, we found the majority of the prokaryotic gene abundance (73% of ocean total; 63% of gut total) to be attributable to a shared functional core. Although the respective gene abundances differed only by 10%, 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. 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 while 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 (*49*). 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 (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 (*16*). Thus, despite the presumably greater environmental complexity in the ocean, the congruent functional redundancy observed in both ecosystems may indeed be an ecosystem-independent property of microbial communities.

We differentiated ocean core from non-core OGs as the latter are more relevant for environment-specific adaptations. Within the ocean, 67% (s.d. 5%) of the total gene abundance was attributed to ocean core OGs. After removing these and the 29% (s.d. 5%) of gene abundance from genes that were not assigned to any OG, 4% (s.d. 1%) remained as the functionally characterized non-core fraction. The abundance distribution among these non-core OGs, of which the largest fraction are of 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 (*16*), functional variation similar in scale to that of the phylogenetic one can be detected when focusing on non-core, potentially adaptive gene families. As an example for such an environmental adaptation, we found an increase of lipid metabolism in oxygen minimum zones of the Eastern Pacific and Northern Indian Ocean (Fig. 8A).

In order 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 fact that the mesopelagic zone is a key re-mineralization 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 (*35*), 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 the role of microbial communities contributing to re-mineralization and carbon export into the ocean interior.

Conclusions

Tara Oceans has generated, in addition to global biodiversity resources for larger organismal size spectra (*21*), 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 (*20*), cross kingdom species interactions (*22*) and planktonic community connectivity across an ocean circulation chokepoint (*23*). Despite some limitations in the sampled organismal size range, oceanic depth layers and temporal resolution, our approach generated an ecosystem-wide dataset that will be useful for improving predictive models of the ocean. Finding temperature to drive microbial community variation has wide-ranging implications in relation to climate changes. The *Tara* Oceans dataset supports progress not only towards a holistic understanding of the ocean ecosystem, but also microbial communities in general by facilitating comparative analyses between ecosystems.

Materials and Methods

Sample and environmental data collection

From 2009-2013, morphological, genetic and environmental data were collected at >200 sampling stations across all major oceanic provinces in the context of *Tara* Oceans. The rationale for sampling, detailed methods and contextual data 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 details are available in table S1 and sample-associated environmental data (*14*) were inferred at the depth of sampling.

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 (*20*), respectively. 30 to 50 ng of DNA was sonicated to a 100 - 800 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 using 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 qPCR quantification (MxPro, Agilent Technologies, USA), each library was sequenced using 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) (*19*), 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 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 co-linearity between assembled contigs and genes, and unassembled 454 sequencing reads using a subset of 11 overlapping samples (*58*). Based on this analysis, we estimate that 1.5% of contigs had breakpoints and thus may suffer from errors (*14*). This error rate is >6.5 times lower 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 latter were selected by the following procedure: an initial set of 3,496 reference genomes (all high-quality genomes available at as of 23 February 2012) was clustered into 1,753 species clusters (*25*), 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 >1x 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” at the IMG 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 (*25*) 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 using the same criteria as in (*18*), 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., non-redundant contiguous gene-coding nucleotide sequences operationally defined as ‘genes’; see also (*17, 18*)), 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 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 non-redundant 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) using SmashCommunity (version 1.6) (*47, 66, 67*). In total, 38% and 57% of the genes could be annotated by homology to a 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 16S rRNA gene tags and metagenomic operational taxonomic units 16S rRNA gene (16S) fragments directly identified in Illumina-sequenced metagenomes ($_{mi}$tags) were identified as described in (*12*). 16S $_{mi}$tags were mapped to cluster centroids of taxonomically annotated 16S reference sequences from the SILVA database (*24*) (release 115: SSU Ref NR 99) that had been clustered at 97% sequence identity using USEARCH v6.0.307 (*68*). 16S $_{mi}$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 16S OTUs as described in (*13, 25*). Relative abundances of mOTU linkage groups were quantified using MOCAT (version 1.3) (*19*).

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 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

16S $_{\mathrm{mi}}$tag counts were rarefied 100 times to the minimum number of total 16S $_{\mathrm{mi}}$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 16S $_{\mathrm{mi}}$tags was calculated from full-length 16S sequences, using parts of the LotuS 16S pipeline (*69*). This phylogenetic tree was midpoint rooted in R and used with the $_{\mathrm{mi}}$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 (*33*).

Distance correlations between genomic and environmental data

We computed pairwise distances between samples based on: (i) relative abundances of taxonomic (16S $_{\mathrm{mi}}$tags and mOTUs) and gene functional compositions (at KEGG module level) – the compositional data, (ii) in-situ measurements of physico-chemical data – the environmental data, and (iii) geographic location of sampling stations – the geographic data. Data from the three southern-most 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 are the original relative abundances and x_0 a small constant, where $x_0 < \min(x)$.

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 prior to 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 (*45*) using inner cross-validation to set the hyperparameters as implemented by the *scikit-learn* python package (*71*). For spatial auto correlation-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 R² (also known as Q²) (*72*), defined as $1-\sum\frac{(y_i-\hat{y}_i)^2}{(y_i-\bar{y})^2}$, where y_i is the value of the parameter for sample i, $\hat{y}_i$ is the prediction for that same sample (obtained by held-out cross validation), and $\bar{y}$ 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, in order to avoid spatial auto-correlation, 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. Due to 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 presence/absence of the taxa (which was modeled by using a binary input matrix to the elastic net fitting routines).

References and Notes

1. P. G. Falkowski, R. T. Barber, V. V. Smetacek, *Science* **281**, 200-207 (1998).
2. W. B. Whitman, D. C. Coleman, W. J. Wiebe, *Proc. Natl. Acad. Sci. USA* **95**, 6578-6583 (1998).
3. S. C. Doney *et al.*, *Ann. Rev. Mar. Sci.* **4**, 11-37 (2012).
4. J. A. Fuhrman, *Nature* **459**, 193-199 (2009).
5. J. B. H. Martiny *et al.*, *Nat. Rev. Microbiol.* **4**, 102-112 (2006).
6. D. B. Rusch *et al.*, *PLoS Biol.* **5**, e77 (2007).
7. S. Yooseph *et al.*, *PLoS Biol.* **5**, e16 (2007).
8. A. Barberán, A. Fernández-Guerra, B. J. M. Bohannan, E. O. Casamayor, *Mol. Ecol.* **21**, 1909-1917 (2012).
9. T. A. Gianoulis *et al.*, *Proc. Natl. Acad. Sci. USA* **106**, 1374-1379 (2009).
10. J. Raes, I. Letunic, T. Yamada, L. J. Jensen, P. Bork, *Mol. Syst. Biol.* **7**, (2011).
11. E. Karsenti *et al.*, *PLoS Biol.* **9**, e1001177 (2011).
12. R. Logares *et al.*, *Environ. Microbiol.* **16**, 2659-2671 (2014).
13. S. Sunagawa *et al.*, *Nat. Methods* **10**, 1196-1199 (2013).
14. Companion web site tables W1-W8, data and information available at: http://ocean-microbiome.embl.de/companion.html
15. H. M. P. Consortium, *Nature* **486**, 215-221 (2012).
16. H. M. P. Consortium, *Nature* **486**, 207-214 (2012).
17. J. Qin *et al.*, *Nature* **464**, 59-65 (2010).
18. J. Li *et al.*, *Nat. Biotechnol.* **32**, 834-841 (2014).
19. J. R. Kultima *et al.*, *PLoS One* **7**, e47656 (2012).
20. J. Brum, et. al., (submitted).
21. C. de Vargas, et. al., (submitted).
22. G. Lima-Mendez, et. al., (submitted).
23. E. Villar, et. al., (submitted).
24. C. Quast *et al.*, *Nucleic Acids Res.* **41**, D590-D596 (2013).
25. D. R. Mende, S. Sunagawa, G. Zeller, P. Bork, *Nat. Methods* **10**, 881-884 (2013).
26. L. Zinger *et al.*, *PLoS One* **6**, e24570 (2011).
27. L. Amaral-Zettler *et al.*, in *Life in the World's Oceans,* A. D. McIntyre, Ed. (Wiley-Blackwell, 2010), pp. 221-245.
28. C. L. Dupont *et al.*, *ISME J* **6**, 1186-1199 (2012).
29. R. M. Morris *et al.*, *Nature* **420**, 806-810 (2002).
30. K. Lochte, C. M. Turley, *Nature* **333**, 67-69 (1988).
31. S. J. Giovannoni, U. Stingl, *Nature* **437**, 343-348 (2005).
32. B. L. Hurwitz, J. R. Brum, M. B. Sullivan, *ISME J*, (2014).
33. S. Vieira-Silva, E. P. C. Rocha, *PLoS Genet.* **6**, e1000808 (2010).
34. T. Pommier *et al.*, *Aquat. Microb. Ecol.* **61**, 221-233 (2010).
35. R. Stocker, *Science* **338**, 628-633 (2012).
36. J. Pernthaler, *Nat. Rev. Microbiol.* **3**, 537-546 (2005).

37. W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, M. L. Sogin, *Proc. Natl. Acad. Sci. USA* **110**, 2342-2347 (2013).
38. J. A. Fuhrman *et al.*, *Proc. Natl. Acad. Sci. USA* **105**, 7774-7778 (2008).
39. D. P. Tittensor *et al.*, *Nature* **466**, 1098-U1107 (2010).
40. J. Ladau *et al.*, *ISME J* **7**, 1669-1677 (2013).
41. Ocean Sampling Day: http://www.microb3.eu/osd
42. Z. I. Johnson *et al.*, *Science* **311**, 1737-1740 (2006).
43. C. A. Lozupone, R. Knight, *Proc. Natl. Acad. Sci. USA* **104**, 11436-11440 (2007).
44. D. P. Herlemann *et al.*, *ISME J* **5**, 1571-1579 (2011).
45. H. Zou, T. Hastie, *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **67**, 301–320 (2005).
46. B. J. Finlay, *Science* **296**, 1061-1063 (2002).
47. S. Powell *et al.*, *Nucleic Acids Res.* **40**, D284-D289 (2011).
48. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33-36 (2000).
49. T. Bell, J. A. Newman, B. W. Silverman, S. L. Turner, A. K. Lilley, *Nature* **436**, 1157-1160 (2005).
50. J. Arístegui, J. M. Gasol, C. M. Duarte, G. J. Herndl, *Limnol. Oceanogr.* **54**, 1501-1529 (2009).
51. E. F. DeLong *et al.*, *Science* **311**, 496-503 (2006).
52. C. Matz, K. Jurgens, *Appl. Environ. Microbiol.* **71**, 921-929 (2005).
53. Y. Yawata *et al.*, *Proc. Natl. Acad. Sci. USA* **111**, 5622-5627 (2014).
54. S. Pesant *et al.*, *Scientific Data*, (in review).
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. Chaffron, Samuel; D'Ovidio, Francesco; De Monte, Silvia; Guidi, Lionel; Iudicone, Daniele; Picheral, Marc; Speich, Sabrina; Pesant, Stephane; *Tara* Oceans Consortium, Coordinators; *Tara* Oceans Expedition, Participants (2014): Contextual environmental data of selected samples from the *Tara* Oceans Expedition (2009-2013): doi:10.1594/PANGAEA.840718
57. Chaffron, Samuel; D'Ovidio, Francesco; De Monte, Silvia; Pesant, Stephane; *Tara* Oceans Consortium, Coordinators; *Tara* Oceans Expedition, Participants (2014): Contextual biodiversity data of selected samples from the *Tara* Oceans Expedition (2009-2013): doi:10.1594/PANGAEA.840698
58. P. Hingamp *et al.*, *ISME J* **7**, 1678-1695 (2013).
59. C. Clerissi *et al.*, *Appl. Environ. Microbiol.* **80**, 3150-3160 (2014).
60. W. Zhu, A. Lomsadze, M. Borodovsky, *Nucleic Acids Res.* **38**, e132-e132 (2010).
61. D. R. Mende *et al.*, *PLoS One* **7**, e31386 (2012).
62. B. L. Hurwitz, L. Deng, B. T. Poulos, M. B. Sullivan, *Environ. Microbiol.* **15**, 1428-1440 (2013).
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, *Bioinformatics* **28**, 125-126 (2012).
66. M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, P. Bork, *Bioinformatics* **26**, 2977-2978 (2010).
67. M. Kanehisa *et al.*, *Nucleic Acids Res.* **36**, D480-D484 (2008).
68. R. C. Edgar, *Bioinformatics* **26**, 2460-2461 (2010).
69. F. Hildebrand, R. Tadeo, A. Y. Voigt, P. Bork, J. Raes, *Microbiome* **2**, 30 (2014).
70. D. P. Faith, *Biol. Conserv.* **61**, 1-10 (1992).
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, *Appl. Environ. Microbiol.* **73**, 5261-5267 (2007).

Figure legends

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 depths 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 (*18*) was built using 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 towards 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 (*18*). 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 mesopelagic zone had been vastly under-sampled prior to *Tara* Oceans. 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.

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 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 16S $_{mi}$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-right; Bray-Curtis) and cell densities per mL (top right) suggest an increase in species richness and a decrease in cell density with depth (pairwise MWU: *P*<0.001), while no significant trend was found for

between sample dissimilarity. For gene functional groups (bottom-right and center), richness increased while between sample dissimilarity decreased with depth. Minimum potential generation time of microbial communities is predicted to be higher in the mesopelagic compared to the epipelagic (EPI).

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 of richness up to about 15ºC followed by a decrease towards warmer waters. Richness is highest in mid-latitudinal ranges rather than towards 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 $_{\mathrm{mi}}$tag OTU abundances increases with distance between sampling stations up to about 5,000 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 show 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: $_{\mathrm{mi}}$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.

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 using 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 highly predict temperature and dissolved oxygen of samples used for validation (left). Models trained using subsets of taxonomic data from deep chlorophyll maximum (DCM) samples could highly predict temperature, but only moderately dissolved oxygen (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 $_{\mathrm{mi}}$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.

Fig. 7. Ocean vs. 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 were computed. With increasing sample size, the number of shared orthologous groups decreased first rapidly, then more gradually to a minimum of 5,755 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% - 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).

Fig. 8. Functional structuring of the ocean microbiome. (A) Phylum-level (class-level for Proteobacteria) taxonomic variability is higher (top, median relative s.d. 65%) relative to the functional composition (OG functional categories) of ocean microbial samples (center, median relative s.d. 7%). Removal of functions that are ubiquitous in the ocean environment reveals the variable (bottom, median relative s.d. 47%), non-core fraction, 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 non-core 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*<0.01, ****P*<0.001).

Acknowledgements

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.

We would like to thank the following people for their contribution to *Tara* Oceans: Ameer Abdulla, Chantal Abergel, Denis Allemand, Aldine Amiel, Leif Anderson, David Antoine, Detlev Arendt, Roberto Arrigoni , Defne Arslan, Francois Artiguenave, Stephane Audic, Jean-Marc Aury, Marcel Babin, Celine Bachelier, Xavier Bailly, Andrew Baker, Cecilia Balestra, Benedetto Barone, Daniela Basso, Daniel Bayley, Gregory Beaugrand, Laurent Beguery, Elia Benito-Guttierez, Francesca Benzoni, Eric Beraud, Lionel Bigot, Lucie Bittner, Martine Boccara, Roxane Boonstra, Peer Bork, Emmanuel Boss, Christophe Boutte, Chris Bowler, Annick Bricaud, Jennifer Brum, Jeremie Capoulade, Luigi Caputi, Annalisa Caragnano, Margaux Carmichael, Raffaela Casotti, Ivona Cetinic, Samuel Chaffron, Aurélie Chambouvet, Patrick Chang, Ali Chase, Claudia Chica, Hervé Claustre, Jean-Michel Claverie, Camille Clerissi, Sebastien Colin, Montse Coll-Lladó, Steeve Comeau, Christian Conrad, Laurent Coppola, Miguel Francisco Cornejo, Marcella Cornejo , Daniel Cossa, Maryam Cousin, Corinne Cruaud, Corrine

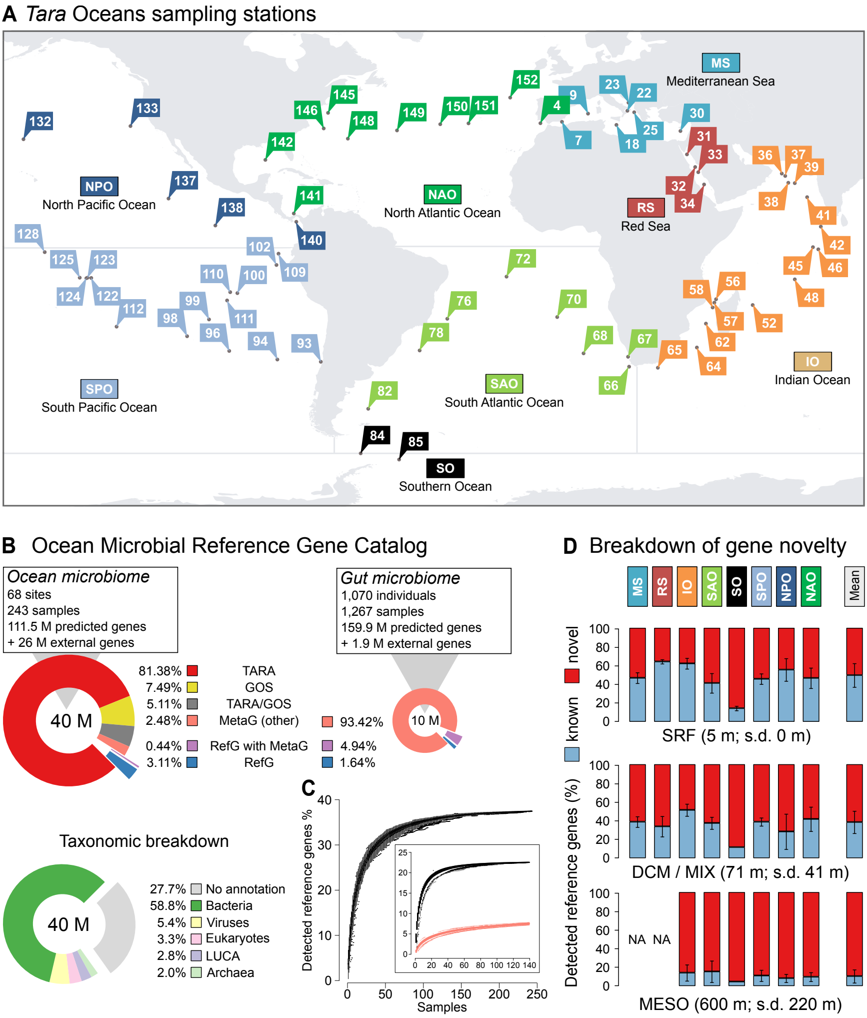
Cuck, Marcela D'Ottone, Corinne Da Silva, Denis Dausse, Denis de la Broise, Silvia De Monte, Colomban de Vargas, Johan Decelle, Alan Deidun, Javier del Campo, Evelyne Derelle, Yves Desdevises, Elodie Desgranges, Valerie Desplanches, Floriane Despres, Nicolas Desreumaux, Rosanna di Mauro, Celine Dimier, John Dolan, Fabrizio D'Ortenzio, Francesco d'Ovidio, Anne Doye, Melissa Duhaime, Emilie Duperche, Xavier Durrieu de Madron, Stephanie Dutkiewicz, Karoline Faust, Janine Felden, Beatriz Fernández, Isabel Ferrera, Regis Ferriere, Christine Ferrier-Pagès, Mick Follows, Rainer Friedrich, Françoise Gaill, Alexandre Ganachaud, Laurence Garczarek, Josep M Gasol, Stéphane Gasparini, Jean-Pierre Gattuso, Gabriella Gilkes, Jennifer Gillette, Silvia G. Acinas, Gabriel Gorsky, Brett Grant, Nigel Grimsley, Jean-Michel Grisoni, Michel Groc, Lionel Guidi, Cedric Guigand, Luis Gutierrez-Herredia, Roland Hellig, Pascal Hingamp, Danwei Huang, Julio Ignacio-Espinoza, Daniele Iudicone, Olivier Jaillon, Jean-Louis Jamet, Stefanie Kandels-Lewis, Lee Karp-Boss, Eric Karsenti, Michael Katinka, Yuko Kitano, Zbigniew Kolber, Philippe Koubbi, Uros Krzic, Hironobu Kukami , Karine Labadie, Pamela Labbe-Ibanez, Tomas Larsson, Alban Lazar, Herve Le Goff, Corinne Le Quere, Brian Leander, Philippe Lebaron, Noan LeBescot, Thomas Lefort, Louis Legendre, Cristophe Lejeusne, Cyrille Lepoivre, Magali Lescot, Mangan Lewis, Edouard Leymarie, Gipsi Lima-Mendez, Ramiro Logares, Frédéric Mahé, Cornelia Maier, Shruti Malviya, Catarina Marcolin, Claudie Marec, Sophie Marinesque, Ramon Massana, Lydiane Mattio, Maria Grazia Mazzochi, Raphaël Morard, Hervé Moreau, Pascal Morin, Simon Morisset, David Mountain, Paul Muir, Harry Nelson, Sophie Nicaud, Paul Nival, Benjamin Noel, Fabrice Not, Grigor Obolensky, David Obura, Hiroyuki Ogata, Philippe Pages, Claude Payri, Javier Paz Yepes, Carlos Pedros-Alio, Eric Pelletier, Rainer Pepperkok, Fabien Perault, Yvan Perez, Stephane Pesant, Marc Picheral, Michel Pichon, Gwenaël Piganeau, Ruby Pillay, Olivier Poirot, Julie Poulain, Nicole Poulton, Franck Prejger, Judit Prihoda, Ian Probert, Gabriele Procaccini, Jeroen Raes, Jeannine Rampal, Josephine Ras, Gilles Reverdin, Emmanuel G. Reynaud, Stephanie Reynaud, Francois Ribalet, Maurizio Ribera d'Alcala, Eric Roettinger, Sarah Romac, Jean-Baptiste Romagnan, Cecile Rottier, Francois Roullier, Christian Rouviere, Anne Royer, Marta Royo Llonch, Martina Sailerova, Guillem Salazar, Gaelle Samson, Sébastien Santini, Christian Sardet, Hugo Sarmento, Eleonora Scalco, Claude Scarpelli, Antoine Sciandra, Sarah Searson, Raffaele Siano, Mike Sieracki, Bianca Silva, Oleg Simakov, Sergei Solonenko, Sabrina Speich, Silvia Spezzaferri, Fabio Stalder, Fabrizio Stefani, Halldor Stefansson, Ernst Stelzer, Lars Stemmann, Lucie Subirana, Matt Sullivan, Shinichi Sunagawa, Jarred Swalwell, Vincent Taillandier, Eric Tambutté, Sylvie Tambutté, Atsuko Tanaka, Isabelle Taupier-Letage, Pierre Testor, Anne Thompson , Doris Thuillier, Virgine Tichanné-Seltzer, Leila Tirichine, Eve Toulza, Sasha Tozzi, Jean-Éric Tremblay, Aline Tribollet, Antoine Triller, Didier Velayoudon, Alaguraj Veluchamy, Emilie Villar, Flora Vincent, Carden Wallace, Markus Weinbauer, Jean Weissenbach, Maureen Williams, Patrick Wincker, Sheree Yau, Alexis Yelton, Adriana Zingone, Didier Zoccola.

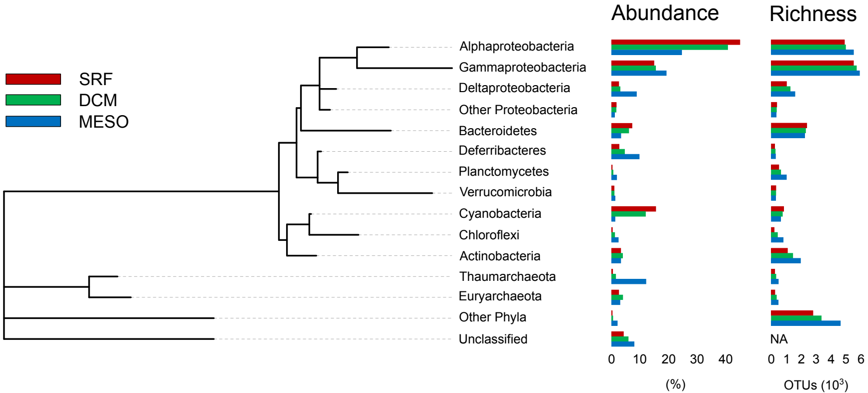
The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein is available at http://ocean-microbiome.embl.de/companion.html, at the EBI under the project identifiers PRJEB402 and PRJEB7988, and at Pangaea (for details see table S1). The data release policy regarding future public release of *Tara* Oceans data is described in Pesant et al (*54*). All authors approved the final manuscript. This article is contribution number ZZZ of *Tara* Oceans. Supplement contains additional data.

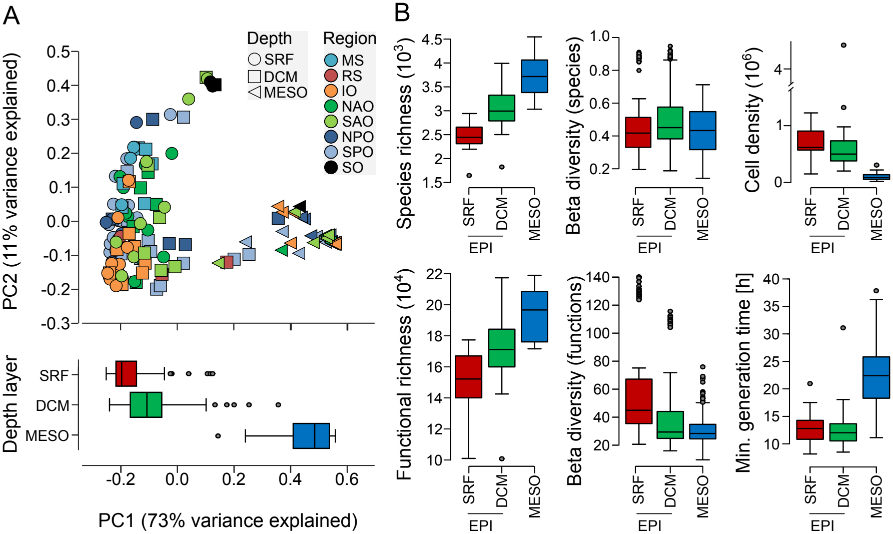
***Tara* Oceans Coordinators**

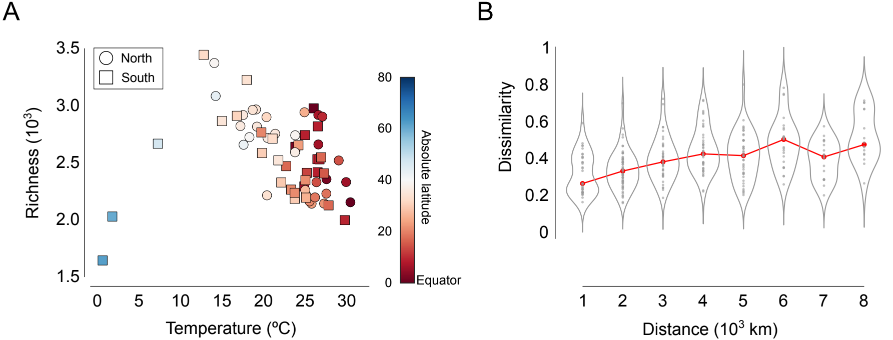
Silvia G. Acinas[1], Peer Bork[2], Emmanuel Boss[3], Chris Bowler[4], Colomban de Vargas[5,6], Michael Follows[7], Gabriel Gorsky[8,9], Nigel Grimsley[10,11], Pascal Hingamp[12], Daniele Iudicone[13], Olivier Jaillon[14,15,16], Stefanie Kandels-Lewis[2,17], Lee Karp-Boss[3], Eric Karsenti[4,17], Uros 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]

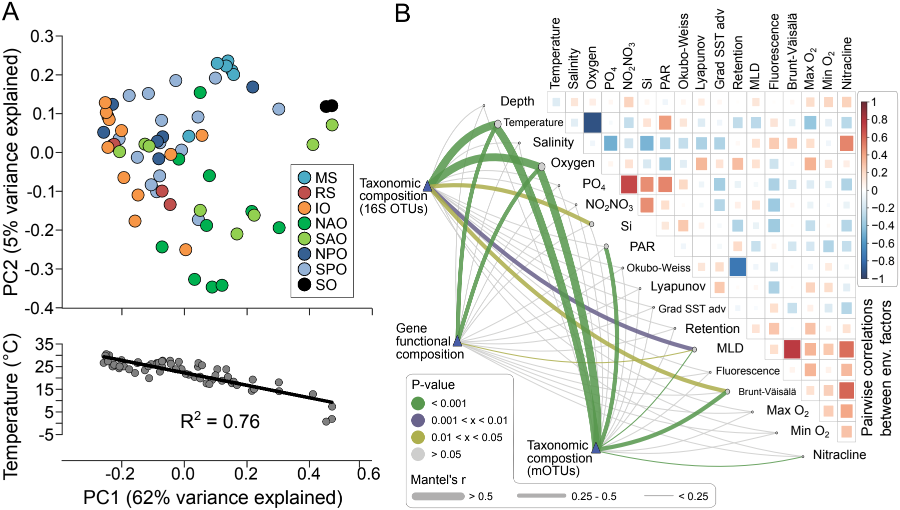
[1]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain.
[2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.
[3]School of Marine Sciences, University of Maine, Orono, Maine, USA.
[4]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France.
[5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
[6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
7Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA.
[8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-mer, France.
[9]Sorbonne Universités, UPMC Univ 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, Heidelberg, Germany.
[18]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstr. 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 Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-mer, France.
[28]Bigelow Laboratory for Ocean Sciences, East Boothbay, 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 Océan UBO-IUEM Palce Copernic 29820 Polouzané, France.
[31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 E Lowell Street, Tucson, AZ, 85721, USA.
[32]DVIP Consulting, Sèvres, France.

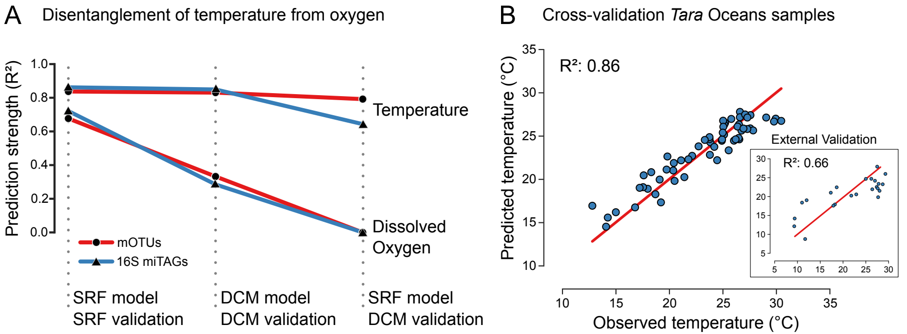

A *Tara* Oceans sampling stations
MS Mediterranean Sea
NPO North Pacific Ocean
NAO North Atlantic Ocean
RS Red Sea
SAO South Atlantic Ocean
IO Indian Ocean
SPO South Pacific Ocean
SO Southern Ocean
132 133 137 138 140 141 142 145 146 148 149 150 151 152 4 9 7 23 22 25 18 30 31 33 32 34 36 37 39 38 41 42 45 46 48 56 58 57 52 62 64 65 67 66 68 70 72 76 78 82 84 85 93 94 96 98 99 100 102 109 110 111 112 122 123 124 125 128
B Ocean Microbial Reference Gene Catalog
Ocean microbiome
68 sites
243 samples
111.5 M predicted genes
+ 26 M external genes
Gut microbiome
1,070 individuals
1,267 samples
159.9 M predicted genes
+ 1.9 M external genes
40 M
10 M
81.38% TARA
7.49% GOS
5.11% TARA/GOS
2.48% MetaG (other) 93.42%
0.44% RefG with MetaG 4.94%
3.11% RefG 1.64%
Taxonomic breakdown
40 M
27.7% No annotation
58.8% Bacteria
5.4% Viruses
3.3% Eukaryotes
2.8% LUCA
2.0% Archaea
C
Detected reference genes %
Samples
0 50 100 150 200 250
0 10 20 30 40
0 20 40 60 80 100 120 140
0 5 10 15 20 25
D Breakdown of gene novelty
MS RS IO SAO SO SPO NPO NAO Mean
novel
known
Detected reference genes (%)
0 20 40 60 80 100
SRF (5 m; s.d. 0 m)
DCM / MIX (71 m; s.d. 41 m)
NA NA
MESO (600 m; s.d. 220 m)

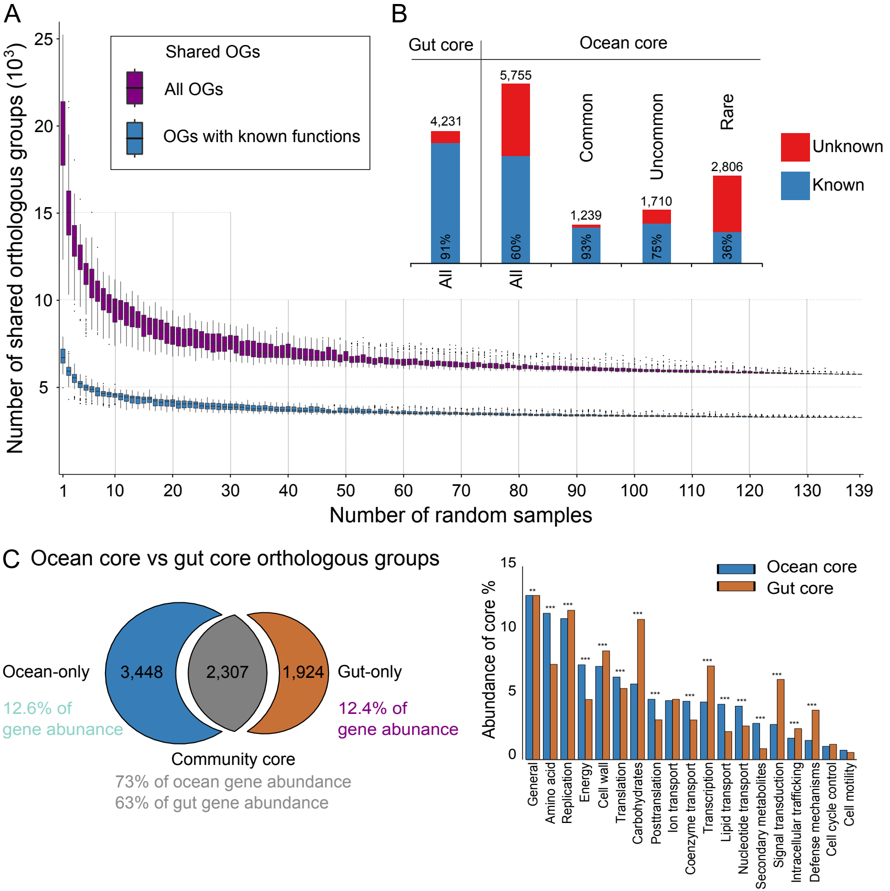

SRF
DCM
MESO
Abundance
Richness
Alphaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
Other Proteobacteria
Bacteroidetes
Deferribacteres
Planctomycetes
Verrucomicrobia
Cyanobacteria
Chloroflexi
Actinobacteria
Thaumarchaeota
Euryarchaeota
Other Phyla
Unclassified
NA
0 10 20 30 40
(%)
0 1 2 3 4 5 6
OTUs (10^3)


A
B
Depth
○ SRF
□ DCM
◁ MESO
Region
MS
RS
IO
NAO
SAO
NPO
SPO
SO
PC2 (11% variance explained)
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
Depth layer
SRF
DCM
MESO
-0.2
0.0
0.2
0.4
0.6
PC1 (73% variance explained)
Species richness (10^3)
4.5
4
3.5
3
2.5
2
1.5
Beta diversity (species)
1.0
0.8
0.6
0.4
0.2
Cell density (10^6)
4
1
0
Functional richness (10^4)
22
20
18
16
14
12
10
Beta diversity (functions)
140
120
100
80
60
40
20
Min. generation time [h]
40
35
30
25
20
15
10
SRF
DCM
MESO
EPI


A
North
South
3.5
3.0
2.5
2.0
1.5
Richness (10^3)
0
5
10
15
20
25
30
Temperature (ºC)
80
60
40
20
0
Absolute latitude
Equator
B
1
0.8
0.6
0.4
0.2
0
Dissimilarity
1
2
3
4
5
6
7
8
Distance (10^3 km)


A
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
PC2 (5% variance explained)
MS
RS
IO
NAO
SAO
NPO
SPO
SO
35
25
15
5
-5
Temperature (°C)
$R^2 = 0.76$
-0.2
0.0
0.2
0.4
0.6
PC1 (62% variance explained)
B
Temperature
Salinity
Oxygen
PO_4
NO_2NO_3
Si
PAR
Okubo-Weiss
Lyapunov
Grad SST adv
Retention
MLD
Fluorescence
Brunt-Väisälä
Max O_2
Min O_2
Nitracline
Depth
Taxonomic composition (16S OTUs)
Gene functional composition
Taxonomic compostion (mOTUs)
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
Pairwise correlations between env. factors
P-value
< 0.001
0.001 < x < 0.01
0.01 < x < 0.05
> 0.05
Mantel's r
> 0.5
0.25 - 0.5
< 0.25


A
Disentanglement of temperature from oxygen
Prediction strength (R²)
1.0
0.8
0.6
0.4
0.2
0.0
Temperature
Dissolved Oxygen
● mOTUs
▲ 16S miTAGs
SRF model SRF validation
DCM model DCM validation
SRF model DCM validation
B
Cross-validation Tara Oceans samples
R²: 0.86
Predicted temperature (°C)
35
30
25
20
15
10
Observed temperature (°C)
10
15
20
25
30
35
External Validation
R²: 0.66
30
25
20
15
10
5
5
10
15
20
25
30


A
Shared OGs
All OGs
OGs with known functions
Number of shared orthologous groups (10^3)
25
20
15
10
5
1
10
20
30
40
50
60
70
80
90
100
110
120
130
139
Number of random samples
B
Gut core
Ocean core
4,231
5,755
1,239
1,710
2,806
Common
Uncommon
Rare
91%
60%
93%
75%
36%
All
All
Unknown
Known
C
Ocean core vs gut core orthologous groups
Ocean-only
3,448
2,307
1,924
Gut-only
12.6% of
gene abundance
12.4% of
gene abunance
Community core
73% of ocean gene abundance
63% of gut gene abundance
15
10
5
0
Abundance of core %
Ocean core
Gut core
**
***
General
Amino acid
Replication
Energy
Cell wall
Translation
Carbohydrates
Posttranslation
Ion transport
Coenzyme transport
Transcription
Lipid transport
Nucleotide transport
Secondary metabolites
Signal transduction
Intracellular trafficking
Defense mechanisms
Cell cycle control
Cell motility