

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 Sarmiento,^{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 Colomán 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 10^5 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 prin-

ciple, 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 de-

coupled 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 ,

¹Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. ²Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. ³Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. ⁴Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. ⁵CEA-Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. ⁶Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. ⁷Sorbonne Universités, UPMC, Université Paris 06, CNRS-IRD-MNHN, LOCEAN Laboratory, 4 Place Jussieu, 75005 Paris, France. ⁸CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. ⁹Sorbonne Universités, UPMC Université Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. ¹⁰Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm UI024, and CNRS UMR 8197, F-75005 Paris, France. ¹¹Laboratoire de Physique des Océans UBO-IUEM, Place Copernic 29820 Plouzané, France. ¹²Aix Marseille Université CNRS IGS UMR 7256, 13288 Marseille, France. ¹³Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. ¹⁴Department of Hydrobiology, Federal University of São Carlos (UFSCar), Rodovia Washington Luiz, 13565-905 São Carlos, São Paulo, Brazil. ¹⁵CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. ¹⁶Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. ¹⁷Directors' Research, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. ¹⁸CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. ¹⁹Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. ²⁰Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. ²¹CNRS, UMR 8030, CP5706, Evry, France. ²²Université d'Evry, UMR 8030, CP5706, Evry, France. ²³Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan. ²⁴PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. ²⁵MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. ²⁶Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. ²⁷Laboratoire de Physique des Océans UBO-IUEM, Place Copernic, 29820 Plouzané, France. ²⁸Max-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.