# Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems

Daniel J. Richter[1,2*], Romain Watteaux[3*], Thomas Vannier[4,5*], Jade Leconte[5], Paul Frémont[5], Gabriel Reygondeau[6,7], Nicolas Maillet[8], Nicolas Henry[1], Gaëtan Benoit[9], Antonio Fernàndez-Guerra[10,11,12], Samir Suweis[13], Romain Narci[14], Cédric Berney[1], Damien Eveillard[15,16], Frederick Gavory[17], Lionel Guidi[18,19], Karine Labadie[17], Eric Mahieu[17], Julie Poulain[5], Sarah Romac[1], Simon Roux[20], Céline Dimier[1,21], Stefanie Kandels[22,23], Marc Picheral[24,25], Sarah Searson[24,25], *Tara* Oceans Coordinators, Stéphane Pesant[26,27], Jean-Marc Aury[17], Jennifer R. Brum[20,28], Claire Lemaitre[9], Eric Pelletier[5], Peer Bork[22,29,30], Shinichi Sunagawa[22,31], Lee Karp-Boss[32], Chris Bowler[21], Matthew B. Sullivan[20,33], Eric Karsenti[21,23], Mahendra Mariadassou[14], Ian Probert[1], Pierre Peterlongo[9], Patrick Wincker[5], Colomban de Vargas[1**], Maurizio Ribera d'Alcalà[3**], Daniele Iudicone[3**$], Olivier Jaillon[5**$]

* and $: equal contributions
**: corresponding authors

***Tara* Oceans Coordinators**: Silvia G. Acinas[34], Peer Bork[22,29,30], Emmanuel Boss[32], Chris Bowler[21], Guy Cochrane[35], Colomban de Vargas[1], Gabriel Gorsky[36], Nigel Grimsley[37,38], Lionel Guidi[18,19], Pascal Hingamp[39], Daniele Iudicone[3], Olivier Jaillon[5], Stefanie Kandels[22,23], Lee Karp-Boss[32], Eric Karsenti[21,23], Fabrice Not[1], Hiroyuki Ogata[40], Stéphane Pesant[26,27], Jeroen Raes[41,42], Christian Sardet[18,43], Mike Sieracki[44,45], Sabrina Speich[46,47], Lars Stemmann[18], Matthew B. Sullivan[20,33], Shinichi Sunagawa[22,31], Patrick Wincker[5]

**Data availability**: http://doi.org/10.6084/m9.figshare.11303177
Supplemental Tables 1-19 (including DDBJ/ENA/GenBank short read archive identifiers for *Tara* Oceans metagenomic & 18S V9 sequence reads, and distance matrices), Datasets 1-3 (18S V9 metabarcode and OTU tables, and reference database).

1 Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, 29680 Roscoff, France
2 Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
3 Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.
4 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110 , 13288, Marseille, France
5 Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France
6 Changing Ocean Research Unit, Institute for the Oceans and Fisheries, University of British Columbia. Aquatic Ecosystems Research Lab. 2202 Main Mall. Vancouver, BC V6T 1Z4. Canada.
7 Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA.
8 Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France.
9 INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes, France.
10 Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark
11 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany
12 Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany
13 Dipartimento di Fisica e Astronomia 'G. Galilei' & CNISM, INFN, Università di Padova, Via Marzolo 8, 35131 Padova, Italy.
14 MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France
15 Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France
16 Research Federation (FR2022) Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016 Paris, France
17 Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

18 Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.
19 Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.
20 Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.
21 Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.
22 Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.
23 Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.
24 Sorbonne Universités, UPMC Univ Paris 06, UMR 7093 LOV, F-75005, Paris, France.
25 CNRS, UMR 7093 LOV, F-75005, Paris, France.
26 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.
27 PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.
28 Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, 70808, USA
29 Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.
30 Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.
31 Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
32 School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.
33 Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA.
34 Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC, Barcelona, Spain.
35 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
36 Sorbonne Universités, CNRS, Laboratoire d'Oceanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France.
37 CNRS, UMR 7232, BIOM, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.
38 Sorbonne Universités Paris 06, OOB UPMC, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.
39 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France
40 Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan.
41 Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.
42 VIB Center for Microbiology, Herestraat 49, 3000 Leuven, Belgium.
43 CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.
44 National Science Foundation, Arlington, VA 22230, USA.
45 Bigelow Laboratory for Ocean Sciences East Boothbay, ME, USA.
46 Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.
47 Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.

## Abstract

Biogeographical studies have traditionally focused on readily visible organisms, but recent technological advances are enabling analyses of the large-scale distribution of microscopic organisms, whose biogeographical patterns have long been debated[1,2]. The most prominent global biogeography of marine plankton was derived by Longhurst[3] based on parameters principally associated with photosynthetic plankton. Localized studies of selected plankton taxa or specific organismal sizes[1,4–7] have mapped community structure and begun to assess the roles of environment and ocean current transport in shaping these patterns[2,8]. Here we assess global plankton biogeography and its relation to the biological, chemical and physical context of the ocean (the 'seascape') by analyzing 24 terabases of metagenomic sequence data and 739 million metabarcodes from the *Tara* Oceans expedition in light of environmental data and simulated ocean current transport. In addition to significant local heterogeneity, viral, prokaryotic and eukaryotic plankton communities all display near steady-state, large-scale, size-dependent biogeographical patterns. Correlation analyses between plankton transport time and metagenomic or environmental dissimilarity reveal the existence of basin-scale biological and environmental continua emerging within the main current systems. Across oceans, there is a measurable, continuous change within communities and environmental factors up to an average of 1.5 years of travel time. Modulation of plankton communities during transport varies with organismal size, such that the distribution of smaller plankton best matches Longhurst biogeochemical provinces, whereas larger plankton group into larger provinces. Together these findings provide an

integrated framework to interpret plankton community organization in its physico-chemical context, paving the way to a better understanding of oceanic ecosystem functioning in a changing global environment.

**Main Text**

Plankton communities are constantly on the move, transported by ocean currents[9]. Transport involves both advection and mixing. While being advected by currents, plankton are influenced by multiple processes, both physico-chemical (fluxes of heat, light and nutrients[10]) and biological (species interactions, life cycles, behavior, acclimation/adaptation[11,12]), which act across various spatio-temporal scales. In turn, plankton impact seawater physico-chemistry while they are being advected[10]. The community composition and biogeochemical properties of a water mass are also partially dependent on its history of mixing with neighboring water masses during transport. These intertwined processes form the pelagic seascape[13] (Supplementary Fig. 1a). Previous studies on plankton distribution have tended to focus on individual factors, such as nutrient or light availability[3,14], or have investigated the role of transport for specific nutrients[15] or types of planktonic organisms[8,16]. Here, instead, we integrated uniformly collected metagenomic data across multiple size fractions with large-scale ocean circulation simulations in the context of the seascape.

We assessed global patterns of plankton biogeography in the context of the seascape using samples collected at 113 stations during the *Tara* Oceans expedition[17], including DNA sequence data from six organismal size fractions: one virus-enriched (0-0.22 µm)[5], one prokaryote-enriched (either 0.22-1.6 or 0.22-3 µm)[18], and four eukaryote-enriched (0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm[19]; Supplementary Fig. 1b). We analyzed 24.2 terabases of metagenomic sequence reads and 320 million new eukaryotic 18S V9 ribosomal DNA marker sequences (Supplementary Table 1), complementing previously described *Tara* Oceans data[5,18,19]. We used metagenomic data and Operational Taxonomic Units (OTUs, representing groups of genetically related organisms) as independent proxies to compute pairwise comparisons of plankton community dissimilarity (β-diversity). Metagenomic dissimilarity highlighted, at species and sub-species resolution, differences in the genomic identity of organisms between stations. Our metagenomic sampling resulted in pairwise metagenomic dissimilarities that likely represent an overestimate of true β-diversity (Supplementary Information 1). However, since we applied an identical procedure to compute dissimilarity between all pairs of samples, these values nevertheless provide an accurate picture of β-diversity variation among samples. The more deeply sampled OTU dissimilarity, in contrast, incorporated the numerous rare taxa within the plankton, but at genus or higher-level taxonomic resolution[19]. Metagenomic and OTU dissimilarities were correlated for all size fractions (Spearman's $\rho$ 0.53 to 0.97, $p \leq 10^{-4}$, Supplementary Fig. 2), indicating that both proxies, although characterized by different sampling depth and taxonomic resolution, provided coherent and complementary estimates of β-diversity (Supplementary Information 1). We performed subsequent analyses using both measures, which produced consistent results. We focus on analyses of metagenomic dissimilarity here, with accompanying results for OTU dissimilarity presented in Supplementary Figures.

Globally, we observed significant dissimilarities at both the metagenomic and OTU level between sampled stations (including adjacent sites) across all size fractions (Supplementary Fig. 3a, Supplementary Information 1). The resulting portrait is of a locally heterogeneous oceanic ecosystem dominated by a small number of abundant and cosmopolitan taxa, with a much larger number of less abundant taxa found at fewer sampling sites (Supplementary Fig. 3b-e), corroborating previous studies[19].

Underlying this local heterogeneity, we found robust evidence for the existence of large-scale biogeographical patterns within all plankton size classes using two complementary analyses of dissimilarity among samples (Fig. 1a, Supplementary Fig. 4a-f, Supplementary Fig. 5, Supplementary

Information 2). First, we grouped metagenomic samples within each size fraction into 'genomic provinces' via hierarchical clustering (Supplementary Fig. 6). Second, we derived colors for each sample based on a principal coordinates analysis (PCoA-RGB; see Methods) in order to visualize transitions in community composition within and between genomic provinces. Most genomic provinces were composed of large-scale geographically contiguous stations (consistent with previous studies documenting patterns in plankton biogeography[1,2,5,6]) with some independent distant samples (Fig. 1a, Supplementary Fig. 4a-f). Genomic provinces of smaller plankton (viruses, bacteria and eukaryotes <20 µm) tended to be limited to a single ocean basin and to approximately correspond to Longhurst biogeochemical provinces[3] (Supplementary Fig. 4a-d; Supplementary Information 3). In contrast, provinces of larger plankton (micro- and meso-plankton, >20 µm) spanned multiple basins (Supplementary Fig. 4e-f, Supplementary Information 4).

These large-scale biogeographical patterns derived from metagenomes were linked to environmental parameters including nutrients, temperature and trophic level. Seawater temperature was significantly different among genomic provinces for all plankton size classes (Kruskal-Wallis test, $p < 10^{-5}$), corroborating previous results for prokaryotes[18], whereas other environmental conditions were significantly different only with respect to specific size classes (Supplementary Fig. 7). The geography of combined nutrient and temperature variations resembled the biogeography of smaller plankton size classes (Fig. 1a-b, Supplementary Fig. 4a-d,g), whereas temperature alone more closely matched the distribution of larger plankton (Supplementary Fig. 4e,f,h), reflecting different potential ecological constraints. Many genomic provinces were spatially consistent with ocean basin-scale circulation patterns, such as western boundary currents or major subtropical gyres[20] (Fig. 1a, Supplementary Fig. 4a-f), suggesting a particular role for large-scale surface transport (a core component of the seascape) in the emergence of spatial patterns of plankton community composition, as previously proposed[21]. We therefore investigated community composition differences between sampled stations in light of the corresponding transit time. We inferred the time of mean transport between stations from trajectories computed with the physically well-constrained MITgcm ocean model (see Methods), which takes into account directionalities[9] and meso- to large-scale circulation, potential dispersal barriers and mixing effects[22,23]. We quantified transport using the minimum travel time[24] ($T_{min}$) between pairs of *Tara* stations. These trajectories corresponded to the dominant paths that transport the majority of water volume and its contents (e.g., heat, nutrients and plankton; Fig. 1c). For all plankton size classes, community composition differences between stations were correlated to travel time (Supplementary Fig. 8). Cumulative correlation values (correlations between metagenomic dissimilarity and $T_{min}$ computed for an increasing range of $T_{min}$) were maximal for pairs of stations separated by $T_{min}$ <~1.5 years for all size classes ($p \leq 10^{-4}$; Spearman's ρ 0.45 to 0.71 depending on size class, Fig. 2a, Supplementary Fig. 9a-e), hence revealing measurable plankton community dynamics on time scales far longer than typical plankton growth rates or life cycles. In contrast, no such unimodal pattern was found for correlations between metagenomic dissimilarity and geographic distance (without traversing land; Supplementary Fig. 9f). Over the timescale <~1.5 years, which corresponds well with the average time to travel across a basin or gyre, large-scale transport is therefore an appropriate framework for studying differences in plankton community composition (Fig. 2b). The fact that simulated transport times and metagenomic dissimilarity were correlated despite a 3 year pan-season sampling campaign highlights the overall stability of plankton dynamics along the main ocean currents.

Transit time also covaried (although less strongly) with differences in environmental conditions for pairs of stations for which $T_{min}$ <~1.5 years (Fig. 3). This indicates that along large-scale oceanic current systems, changes in environmental conditions and plankton community composition are concurrent. In our data, beyond ~1.5 years of transport, correlations of $T_{min}$ with metagenomic dissimilarity decreased (Fig. 2a, Fig. 3, Supplementary Fig. 9a-e), meaning the signature of transport in generating large-scale diversity changes weakened and travel time therefore becomes a less appropriate framework to study β-diversity. A similar trend was observed for the correlation between $T_{min}$ and

nutrient concentrations whereas temperature was better correlated when considering larger transit times (Fig. 3).

Together, these analyses suggest the existence in the seascape of stable biogeochemical continua induced by basin-scale currents with predictable, interlinked changes in environmental conditions and plankton community composition (Supplementary Information 5). It has previously been posited that transport could generate continuous transitions between niches[25], but it was not anticipated that this would occur on the scale of ocean basins. Beyond ~1.5 years, the correlation of metagenomic dissimilarity with differences in temperature increased while that with differences in nutrients decreased (Fig. 3, Supplementary Fig. 9a-e). However, both of these correlations with metagenomic dissimilarity remained strong on these time scales. This might be related to distant *Tara* Oceans stations experiencing similar oceanographic phenomena (notably temperature), for example upwelling zones, producing generally similar environmental conditions.

The existence of a size-class dependent (smaller or larger than 20 µm) plankton biogeography indicates that organisms contribute differently to the basin-scale biogeochemical continua present in the seascape. In the case of the North Atlantic current system (including the Mediterranean Sea), a simple exponential fit of metagenomic dissimilarity along $T_{min}$ for $T_{min}$ <~1.5 years (Fig. 2c) revealed that the smaller size classes (<20 µm) had a shorter metagenomic turnover time (ca. 1y) than larger plankton (ca. 2y) (Supplementary Fig. 10, Supplementary Information 6). At global geographical scales, the genomic provinces of small size classes, which are enriched in phytoplankton[18,19], corresponded with differences in environmental parameters such as nutrient levels (Fig. 1b, Supplementary Fig. 7) that are often constrained by regional oceanographic processes[26], as shown in our data. On the other hand, genomic provinces of larger plankton, dominated by heterotrophic and symbiotic organisms[19], often crossed biogeochemical boundaries and were more related to global scale gradients and circulation patterns, notably major latitudinal temperature zones or the separation between Atlantic and Indo-Pacific large-scale surface circulations (Supplementary Fig. 4e,f,h). These divergent effects were also evident in comparisons of metagenomic dissimilarity with variations in environmental conditions (Supplementary Fig. 9b). For smaller plankton, correlations with differences in nutrient concentrations were stronger for $T_{min}$ up to ~1.5 years, but for larger plankton, correlations were stronger with temperature variations for $T_{min}$ beyond ~1.5 years. These results indicate a significant size-based decoupling within planktonic food webs (see Supplementary Information 4).

In this study, we provide genomic evidence for an organism-size-dependent global plankton biogeography shaped by currents at the scale of ocean basins. We measured, using metagenomes, the underlying plankton dynamics driven by seascape processes such as intrinsic biological dynamics, variation in environmental conditions, and/or long-range transport. Our analyses reveal that global plankton communities include components that are in a near steady-state that emerges from the integration of the seascape. This behavior resembles self-organizing systems within reaction-advection-diffusion contexts[27]. This work shows that studies of the dynamics of plankton communities must consider the critical influence of ocean currents in stretching and altering, on the scale of basins, the distribution of both planktonic organisms and the physico-chemical nature of the water mass in which they reside. In this context, our study confirms that the combination of ocean circulation modelling with the use of metagenomic DNA as a tracer of plankton communities is a key tool for unravelling the regulation of plankton dynamics. The planktonic ecosystem is fundamentally different in many ways from other major planetary ecosystems and this study provides a framework to understand and predict the structuring of the ocean ecosystem in a scenario of rapid environmental and current system changes[28,29].

## Methods

### Sampling, sequencing and environmental parameters

Sampling, size fractionation, measurement of environmental parameters and associated metadata, DNA extraction and metagenomic sequencing were conducted as described previously[30,31]. Samples were collected at 113 *Tara* Oceans stations for six size fractions (0-0.2, 0.22-1.6/3, 0.8-5, 5-20, 20-180, 180-2000 µm; Supplementary Fig. 1b; Supplementary Table 1) and two depths (subsurface and deep chlorophyll maximum (DCM)). The prokaryote-enriched size fraction was collected either a 0.22-1.6 µm or 0.22-3 µm filter[18,30].

We used physico-chemical data measured *in situ* during the *Tara* Oceans expedition (depth of sampling, temperature, chlorophyll, phosphate, nitrate and nitrite concentrations), supplemented with simulated values for iron and ammonium (using the MITgcm Darwin model described below in "Ocean circulation simulations"), day length, and 8-day averages calculated for photosynthetically active radiation (PAR) in surface waters (AMODIS, https://modis.gsfc.nasa.gov). In order to obtain PAR values at the deep chlorophyll maximum, we used the following formula[32]:

$$PAR(Z) = PAR(0)*\exp(-k*Z)$$

$$x=\log(Chl)$$

$$\log(Z)=1.524-0.426x-0.0145x^2+0.0186x^3$$

$$k=-\ln(0.01)/Z$$

in which k is the attenuation coefficient, and Z is the depth of the DCM (in meters). Other data, such as silicate and the nitrate/phosphate ratio, were extracted from the World Ocean Atlas 2013 (WOA13 version 2, https://www.nodc.noaa.gov/OC5/woa13/), by retrieving the annual mean values at the closest available geographical coordinates and depths to *Tara* sampling stations. For temperature and nitrate, we calculated seasonality indexes (SI) from monthly WOA13 data. For each sample, the index is the annual variation of the parameter (max - min) at this location divided by the highest variation value among all samples.

A list of samples, metagenomic and metabarcode sequencing information and associated environmental data is available in Supplementary Tables 1-2.

**Calculation of metagenomic community dissimilarity**

Metagenomic community distance between pairs of samples was estimated using whole shotgun metagenomes for all six size fractions. We used a metagenomic comparison method (Simka[33]) that computes standard ecological distances by replacing species counts by counts of DNA sequence k-mers (segments of length *k*). K-mers of 31 base pairs (bp) derived from the first 100 million reads sequenced in each sample (or the first 30 million reads for the 0-0.2 µm size fraction) were used to compute a similarity measure between all pairs of samples within each organismal size fraction. Based on a benchmark of Simka, we selected 100 million reads per sample (or 30 million for the 0-0.2 µm fraction) because increasing this number did not produce a qualitatively different set of results, and to ensure that the same number of reads were used in each pairwise comparison within a size fraction. Nearly all samples in our data set had at least 100 million reads (or at least 30 million for the 0-0.2 µm fraction; Supplementary Table 1).

We estimated β-diversity for metagenomic reads with the following equation within Simka:

$$\text{Metagenomic } \beta\text{-diversity} = (b + c) / (2a + b + c)$$

Where a is the number of distinct k-mers shared between two samples, and b and c are the number of distinct k-mers specific to each sample. We represented the distance between each pair of samples on a heatmap using the heatmap.2 function of the R-package[34] gplots_2.17.0[35]. The dissimilarity matrices we produced for each plankton size fraction (on a scale of 0 = identical to 100 = completely dissimilar) are available as Supplementary Tables 3-8.

**Calculation of OTU-based community dissimilarity**

Within the 0-0.2 µm size fraction, we used previously published viral populations (equivalent to OTUs)[36] and viral clusters (analogous to higher taxonomic levels)[5] based on clustering of protein content. For the 0.22-1.6/3 µm size fraction, we used previously derived miTAGs based on metagenomic matches to 16S ribosomal DNA loci and processed them as described[18]. For the four

eukaryotic size fractions, we added additional samples to a previously published *Tara* Oceans metabarcoding data set and processed them using the same methods[19] (also described at DOI: 10.5281/zenodo.15600).

We calculated OTU-based community dissimilarity for all size fractions as the Jaccard index based on presence/absence data using the vegdist function implemented in vegan 2.4-0[37] in the software package R. The dissimilarity matrices we produced for each plankton size fraction (on a scale of 0 = identical to 100 = completely dissimilar) are available as Supplementary Tables 9-14.

**Calculating distances of environmental parameters**

We calculated Euclidean distances[38] for physico-chemical parameters. Each were scaled individually to have a mean of 0 and a variance of 1 and thus to contribute equally to the distances. Then the Euclidean distance between two stations i and j for parameters P was computed as follows:

$$ED(i,j,P) = \sqrt{\sum_{p \in P} \left(x_{ip} - x_{jp}\right)^2}$$

**RGB encoding of environmental positions**

We color-coded the position of stations in environmental space for Fig. 1b and Supplementary Fig. 4g as follows. First, environmental variables were power-transformed using the Box-Cox transformation to have Gaussian-like distributions to mitigate the effect of outliers and scaled to have zero mean and unit variance. We then performed a principal component analysis (PCA) with the R command prcomp from the package stats 3.2.1[34] on the matrix of transformed environmental variables and kept only the first 3 principal components. Finally, we rescaled the scores in each component to have unit variance and decorrelated them using the Mahalanobis transformation. Each component was mapped to a color channel (red, green or blue) and the channels were combined to attribute a single composite color to each station. The components (x, y, z) were mapped to color channel values (r, g, b) between 0 and 255 as r = 128 * (1 + x / max(abs(x)), and similarly for g and b. This map ensures that the global dispersion is equally distributed across the three components and composite colors span the whole color space.

**Definition of genomic provinces**

We used a hierarchical clustering method on the metagenomic pairwise dissimilarities produced by Simka for all surface and DCM samples, and multiscale bootstrap resampling for assessing the uncertainty in hierarchical cluster analysis. We focused on metagenomic dissimilarity due to its higher resolution, and confirmed that the patterns found in metagenomic data were consistent when using OTU data (Supplementary Fig. 5). We used UPGMA (Unweighted Pair-Group Method using Arithmetic averages) clustering, as it has been shown to have the best performance to describe clustering of regions for organismal biogeography[39]. The R-package pvclust_1.3-2[40], with average linkage clustering and 1,000 bootstrap replications, was used to construct dendrograms with the approximately unbiased p-value for each cluster (Supplementary Fig. 6). Because the number of genomic provinces by size fraction was not known *a priori*, we applied a combination of visualization and statistical methods to compare and determine the consistency within clusters of samples. First, the silhouette method[41] was used to measure how similar a sample was within its own cluster compared to other clusters using the R package cluster_2.0.1[42]. The Silhouette Coefficient *s* for a single sample is given as:

$$s = (b - a) / \max(a,b)$$

Where *a* is the mean distance between *a* sample and all other points in the same class and *b* is the mean distance between *a* sample and all other points in the next nearest cluster. We used the value of *s*, in addition to bootstrap values, to partition each tree into genomic provinces (see Supplementary Information 2 for further details on statistical validation of genomic provinces). Additionally, we used the Radial Reingold-Tilford Tree representation from the JavaScript library D3.js (https://d3js.org/)[43]

to visualize sample partitions from the dendrogram. Single samples were not considered as genomic provinces.

In a complementary approach, we performed a principal coordinates analysis (PCoA) with the R command cmdscale (eig = TRUE, add = TRUE) from the package stats 3.2.1[34] on the matrices of pairwise metagenomic dissimilarities calculated by Simka (or OTU dissimilarity measured with the Jaccard index) within each size fraction and kept only the first 3 principal coordinates. We then converted those coordinates to a color using the RGB encoding described above, with one modification: scaling factors $\lambda_r$, $\lambda_g$ and $\lambda_b$ were calculated as the ratios of the second and third eigenvalues to the first (dominant) eigenvalue to ensure that the dispersion of stations along each color channel reproduced the dispersion of the stations along the corresponding principal component (the ratio for the color corresponding to the dominant eigenvalue is 1). The components (x, y, z) were then mapped to color channel values (r, g, b) between 0 and 255 as $r = 128 * (1 + \lambda_c x / \max(\text{abs}(x))$, where $\lambda_c$ is the ratio of the eigenvalue of color c to the dominant eigenvalue.

We represented number and PCoA-RGB color of genomic provinces for each sample on a world map (Fig. 1, Supplementary Fig. 4a-f) generated with the R packages maps_3.0.0.2[44], mapproj 1.2-4[45], gplots_2.17.0[35] and mapplots_1.5[46]. We also plotted phosphate and temperature (Supplementary Fig. 4a-f) obtained from the *Csiro Atlas of Regional Seas* (CARS2009, http://www.cmar.csiro.au/cars) using the phosphate_cars2009.nc and temprerature_cars2009a.nc files and the R package RNetCDF[47].

**Comparison of genomic provinces to previous ocean divisions**

To evaluate the spatial similarity between the clusters obtained in our study for each size fraction and previous biogeographic divisions, we performed an analysis of similarity (ANOSIM, Fathom toolbox, matlab®). First, we collected coordinates for three spatial divisions at a resolution of 0.5° x 0.5°: biomes, biogeochemical provinces (BGCPs)[3,48] and objective global ocean biogeographic provinces (OGOBPs)[49]. Second, we assigned *Tara* Oceans stations to biomes, BGCPs, and OGOBPs based on their GPS coordinates. Third, for each size fraction we performed an ANOSIM with the metagenomic dissimilarity matrix calculated by Simka, using biogeographic clusters (biome, BGCP, OGOBP) as group membership for each station. Each ANOSIM was bootstrapped 1,000 times to evaluate the interval of confidence around the strength of the relationships we detected (Supplementary Fig. 4a-f).

**Environmental differences among genomic provinces**

For each size fraction, we tested which environmental parameters significantly discriminated among genomic provinces (Supplementary Fig. 7). A total of 12 parameters characterizing each sample, grouped by genomic provinces, were evaluated with a Kruskal-Wallis test within each size fraction with a significance threshold of $p < 10^{-5}$. Selected parameters for each size fraction were then used to perform a principal components analysis of the samples using the R package vegan_1.17-11[37]. Samples were plotted with the same PCoA-RGB colors used in the genomic province maps above and each genomic province surrounded by a grey polygon. In analyses where Southern Ocean (including Antarctic) stations were considered independently from other stations, the following were considered Southern Ocean stations: 82, 83, 84, 85, 86, 87, 88, 89.

**Ocean circulation simulations**

We derived travel times from the MITgcm Darwin simulation[50] based on an optimized global ocean circulation model from the ECCO2 group[51]. The horizontal resolution of the model was approximately 18 km, with 1,103,735 total ocean cells. We ran the model for six continuous years in order to smooth anomalies that might occur during any single year. We used surface velocity simulation data to compute trajectories of floats originating in ocean cells containing all *Tara* Oceans stations, and applied the following stitching procedure to generate a large number of trajectories for each initial position. (The use of surface velocity data implies that Ekman transport also influences trajectories within the simulation.)

First, we precomputed a set of monthly trajectories: for each of the 72 months in the dataset, we released floats in every ocean cell of the model grid and simulated transport for one month. We used a fourth-order Runge-Kutta method with trilinearly interpolated velocities and a diffusion of 100 $m^2/s$. Second, following previous studies[4], we stitched together monthly trajectories to create 10,000 year trajectories: for each float released within a 200 km radius of a *Tara* station, we constructed 1,000 trajectories, each 10,000 years long. To avoid seasonal effects, we began by selecting a random starting month. We followed the trajectory of a float released within that month to the grid cell containing its end point at the end of the month. Next, we randomly selected a trajectory starting on the following month (e.g., February would follow January) from that grid cell, and repeated until reaching a 10,000 year trajectory.

We searched the resulting 50.8 million trajectories for those that connected pairs of *Tara* Oceans stations. To ensure robustness of our results, we only included pairs of stations that were connected by more than 1,000 trajectories. For each pair of stations, $T_{min}$ was defined as the minimum travel time of all trajectories (if any) connecting the two stations. The travel time matrix we produced (measured in years) is available as Supplementary Table 15. Standard minimum geographic distance without traversing land[52] is available as Supplementary Table 16.

**Correlations of β-diversity, $T_{min}$ and environmental parameters**

We excluded stations that were not from open ocean locations from correlation analyses to avoid sites impacted by coastal processes (those numbered 54, 61, 62, 79, 113, 114, 115, 116, 117, 118, 119, 120, and 121). In analyses where Southern Ocean (including Antarctic) stations were considered independently from other stations, the following were considered Southern Ocean (including Antarctic) stations: 82, 83, 84, 85, 86, 87, 88, 89. We calculated rank-based Spearman correlations between β-diversity, $T_{min}$ and environmental parameters (either differences in temperature or the Euclidean distance composed of differences in $NO_2NO_3$, $PO_4$ and Fe, see above) for surface samples with a Mantel test with 1,000 permutations and a nominal significance threshold of $p < 0.01$. For the correlations presented in Fig. 2a, Fig. 3 and Supplementary Fig. 9 correlation values were derived from pairs of stations connected by $T_{min}$ up to the value on the x-axis. We calculated partial correlations of metagenomic and OTU dissimilarity and $T_{min}$ by controlling for differences in temperature and for differences in nutrient concentrations, and partial correlations of dissimilarity with temperature or nutrient variation by controlling for $T_{min}$.

**Community turnover in the North Atlantic**

*Tara* Oceans stations numbered 72, 76, 142, 143, 144, and all stations from 146 to 151 were located along the main current system connecting South Atlantic and North Atlantic oceans and continuing to the strait of Gibraltar. In addition, we included stations 4, 7, 18, and 30 located on the main current system in the Mediterranean Sea (Supplementary Fig. 10). As the *Tara* Oceans samples within the subtropical gyre of the North Atlantic and in the Mediterranean Sea were all collected in winter, seasonal variations should not play a role in the variability in community composition that we observed (see Supplementary Table 2). We calculated genomic e-folding times (the time after which the detected genomic similarity between plankton communities changes by 63%) over scales from months to years based on an exponential fit of metagenomic dissimilarity to $T_{min}$ with the form $y = C_0 e^{-x/\tau}$ (where $C_0$ is a constant and $\tau$ the folding time). Exponential fits for size fractions 0-0.2 µm and 5-20 µm were not calculated due to an insufficient number of sampled stations in the North Atlantic (Supplementary Information 6).

The synthetic map (Supplementary Fig. 10a) was generated with the R packages maps_3.0.0.2, mapproj 1.2.4, gplots_2.17.0 and mapplots_1.5. We derived dynamic sea surface height from the *Csiro Atlas of Regional Seas* (CARS2009, http://www.cmar.csiro.au/cars) using the hgt2000_cars2009a.nc file and plotted with the R package RNetCDF.

## References


1. Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. 4, 102–112 (2006).
2. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat. Rev. Microbiol. (2012). doi:10.1038/nrmicro2795
3. Longhurst, A. Ecological Geography of the Sea. (Academic Press, 2006).
4. Hellweger, F. L., van Sebille, E. & Fredrick, N. D. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science 345, 1346–1349 (2014).
5. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537, 689–693 (2016).
6. McGowan, J. A. & Walker, P. W. Structure in the Copepod Community of the North Pacific Central Gyre. Ecol. Monogr. 49, 195–226 (1979).
7. Reygondeau, G. & Dunn, D. Pelagic Biogeography. in Encyclopedia of Ocean Sciences 588–598 (Elsevier, 2019). doi:10.1016/B978-0-12-409548-9.11633-1
8. Villarino, E. et al. Large-scale ocean connectivity and planktonic body size. Nat. Commun. 9, 142 (2018).
9. Watson, J. R. et al. Realized and potential larval connectivity in the Southern California Bight. Mar. Ecol. Prog. Ser. 401, 31–48 (2010).
10. Moore, C. M. et al. Processes and patterns of oceanic nutrient limitation. Nat. Geosci. 6, 701–710 (2013).
11. Flynn, K. J. et al. Acclimation, adaptation, traits and trade-offs in plankton functional type models: reconciling terminology for biology and modelling. J. Plankton Res. 37, 683–691 (2015).
12. Armbrust, E. V. The life of diatoms in the world's oceans. Nature 459, 185–192 (2009).
13. Pittman, S.J. (ed.). Seascape Ecology. (Wiley-Blackwell, 2017).
14. Tagliabue, A. et al. The integral role of iron in ocean biogeochemistry. Nature 543, 51–59 (2017).
15. Letscher, R. T., Primeau, F. & Moore, J. K. Nutrient budgets in the subtropical ocean gyres dominated by lateral transport. Nat. Geosci. 9, 815–819 (2016).
16. Wilkins, D., van Sebille, E., Rintoul, S. R., Lauro, F. M. & Cavicchioli, R. Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. Nat. Commun. 4, 2457 (2013).
17. Karsenti, E. et al. A Holistic Approach to Marine Eco-Systems Biology. PLoS Biol. 9, e1001177 (2011).
18. Sunagawa, S. et al. Structure and function of the global ocean microbiome. Science 348, 1261359 (2015).
19. de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605–1261605 (2015).
20. Talley, L. D., Pickard, G. L., Emery, W. J. & Swift, J. H. Descriptive Physical Oceanography: An Introduction. (Elsevier, 2011).
21. Clayton, S., Dutkiewicz, S., Jahn, O. & Follows, M. J. Dispersal, eddies, and the diversity of marine phytoplankton. Limnol. Oceanogr. Fluids Environ. 3, 182–197 (2013).
22. Goetze, E. et al. Ecological dispersal barrier across the equatorial Atlantic in a migratory planktonic copepod. Prog. Oceanogr. (2016). doi:10.1016/j.pocean.2016.07.001
23. Mousing, E. A., Richardson, K., Bendtsen, J., Cetinić, I. & Perry, M. J. Evidence of small-scale spatial structuring of phytoplankton alpha- and beta-diversity in the open ocean. J. Ecol. 104, 1682–1695 (2016).
24. Jönsson, B. F. & Watson, J. R. The timescales of global surface-ocean connectivity. Nat. Commun. 7, 11239 (2016).
25. Lévy, M., Jahn, O., Dutkiewicz, S. & Follows, M. J. Phytoplankton diversity and community structure affected by oceanic dispersal and mesoscale turbulence. Limnol. Oceanogr. Fluids Environ. 4, 67–84 (2014).
26. Sarmiento, J. L. & Gruber, N. Ocean Biogeochemical Dynamics. (Princeton University Press, 2006).
27. Feudel, U. Pattern Formation in Marine Systems. in Complexity and Synergetics 179–196 (Springer International Publishing, 2018). doi:10.1007/978-3-319-64334-2_15
28. Beaugrand, G. Reorganization of North Atlantic Marine Copepod Biodiversity and Climate. Science 296, 1692–1694 (2002).
29. Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G. & Saba, V. Observed fingerprint of a weakening Atlantic Ocean overturning circulation. Nature 556, 191–196 (2018).
30. Pesant, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. Sci. Data 2, 150023 (2015).

31. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci. Data 4, 170093 (2017).
32. Morel, A. et al. Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. Remote Sens. Environ. 111, 69–88 (2007).
33. Benoit, G. et al. Multiple comparative metagenomics using multiset k -mer counting. PeerJ Comput. Sci. 2, e94 (2016).
34. R Core Team, T. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2017).
35. Warnes, G. R. et al. R package gplots: Various R Programming Tools for Plotting Data. (2015).
36. Brum, J. R. et al. Patterns and ecological drivers of ocean viral communities. Science 348, 1261498 (2015).
37. Oksanen, J. et al. R package vegan: Community Ecology Package. (2019).
38. Legendre, P. & Legendre, L. Numerical Ecology. (Elsevier, 2012).
39. Kreft, H. & Jetz, W. A framework for delineating biogeographical regions based on species distributions. J. Biogeogr. 37, 2029–2053 (2010).
40. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22, 1540–1542 (2006).
41. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
42. Maechler, M., Rousseeuw, P. J., Struyf, A., Hubert, M. & Hornik, K. R package cluster: Cluster Analysis Basics and Extensions. (2015).
43. Bostock, M., Ogievetsky, V. & Heer, J. D3 Data-Driven Documents. IEEE Trans. Vis. Comput. Graph. 17, 2301–2309 (2011).
44. Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P. & Deckmyn, A. R package maps: Draw Geographical Maps. (2018).
45. McIlroy, D., Brownrigg, R., Minka, T. P. & Bivand, R. R package mapproj: Map Projections. (2015).
46. Gerritsen, H. R package mapplots: Data Visualization on Maps. (2014).
47. Ridgway, K. R., Dunn, J. R. & Wilkin, J. L. Ocean Interpolation by Four-Dimensional Weighted Least Squares—Application to the Waters around Australasia. J. Atmospheric Ocean. Technol. 19, 1357–1375 (2002).
48. Reygondeau, G. et al. Dynamic biogeochemical provinces in the global ocean. Glob. Biogeochem. Cycles 27, 1046–1058 (2013).
49. Oliver, M. J. & Irwin, A. J. Objective global ocean biogeographic provinces. Geophys. Res. Lett. 35, L15601 (2008).
50. Clayton, S. et al. Biogeochemical versus ecological consequences of modeled ocean physics. Biogeosciences Discuss. 1–20 (2016). doi:10.5194/bg-2016-337
51. Menemenlis, D. et al. ECCO2: High resolution global ocean and sea ice data synthesis. Mercat. Ocean Q. Newsl. 31, 13–21 (2008).
52. Rattray, A. et al. Geographic distance, water circulation and environmental conditions shape the biodiversity of Mediterranean rocky coasts. Mar. Ecol. Prog. Ser. 553, 1–11 (2016).
53. Carradec, Q. et al. A global ocean atlas of eukaryotic genes. Nat. Commun. 9, 373 (2018).
54. Wu, S., Xiong, J. & Yu, Y. Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass Copepoda. PLoS ONE 10, e0131498 (2015).
55. Vannier, T. et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. Sci. Rep. 6, 37900 (2016).
56. Piganeau, G., Eyre-Walker, A., Grimsley, N. & Moreau, H. How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. PLoS ONE 6, e16342 (2011).
57. Worden, A. Z. et al. Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science 324, 268–272 (2009).
58. Seeleuthner, Y. et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nat. Commun. 9, 310 (2018).
59. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31, 3718–3720 (2015).
60. Sokal, R. R. & Rohlf, F. J. The Comparison of Dendrograms by Objective Methods. Taxon 11, 33–40 (1962).
61. Sneath, P. H. A. & Sokal, R. R. Numerical taxonomy. The principles and practice of numerical classification. (W.H. Freeman and Company, 1973).

569 62. Baker, F. B. & Hubert, L. J. Measuring the Power of Hierarchical Cluster Analysis. J. Am. Stat. Assoc. 70,
570 31–38 (1975).
571 63. Wei, T. & Simko, V. R package corrplot: Visualization of a Correlation Matrix. (2016).
572 64. Terada, Y. & von Luxburg, U. R package loe: Local Ordinal Embedding. (2016).
573 65. Speich, S., Blanke, B. & Cai, W. Atlantic meridional overturning circulation and the Southern
574 Hemisphere supergyre. Geophys. Res. Lett. 34, n/a–n/a (2007).
575 66. Madoui, M.-A. et al. New insights into global biogeography, population structure and natural selection
576 from the genome of the epipelagic copepod Oithona. Mol. Ecol. 26, 4467–4482 (2017).
577 67. Eppley, R. W. Temperature and phytoplankton growth in the sea. Fish Bull 70, 1063–1085 (1972).
578 68. Reygondeau, G. et al. Biogeography of tuna and billfish communities. J. Biogeogr. 39, 114–129 (2012).
579 69. Fofonoff, N. P. The Gulf Stream system. in Evolution of Physical Oceanography: Scientific Surveys in
580 Honor of Henry Stommel (eds. Warren, B. A. & Wunsch, C.) 112–139 (MIT Press, 1980).
581 70. Dornelas, M. et al. Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss.
582 Science 344, 296–299 (2014).
583 71. Franklin, B. A Letter from Dr. Benjamin Franklin, to Mr. Alphonsus le Roy, Member of Several
584 Academies, at Paris. Containing Sundry Maritime Observations. Trans. Am. Philos. Soc. 2, 294–329 (1786).
585

586 **Acknowledgements**
587

588 We acknowledge Oliver Jahn and M. J. Follows for providing numerical simulations of particle
589 trajectories from *Tara* Oceans stations. We thank the commitment of the following people and
590 sponsors who made this expedition possible: CNRS (in particular Groupement de Recherche
591 GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French
592 Government 'Investissement d'Avenir' programs OCEANOMICS (ANR-11-BTBR-0008) and FRANCE
593 GENOMIQUE (ANR-10-INBS-09), Fund for Scientific Research – Flanders, VIB, Stazione Zoologica
594 Anton Dohrn, UNIMIB, MEMO LIFE (ANR-10-LABX-54), Paris Sciences et Lettres (PSL) Research
595 University (ANR-11-IDEX-0001- 02),
596 ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, MAPPI/ANR-
597 2010-COSI-004, TARA-GIRUS/ANR-09-PCS-GENM-218, HYDROGEN/ANR-14-CE23-0001), EU FP7
598 MicroB3/No. 287589, US NSF grant DEB-1031049, FWO, BIO5, Biosphere 2, Agnès b., the Veolia
599 Environment Foundation, Région Bretagne, World Courier, Illumina, Cap L'Orient, the EDF
600 Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the
601 *Tara* schooner and its captain and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing
602 daily satellite data during the expedition. The bulk of genomic computations were performed using
603 the Airain HPC machine provided through GENCI- [TGCC/CINES/IDRIS] (grants t2011076389,
604 t2012076389, t2013036389, t2014036389, t2015036389 and t2016036389). We are also grateful to
605 the French Ministry of Foreign Affairs for supporting the expedition and to the countries who
606 granted us sampling permissions. *Tara* Oceans would not exist without continuous support from 23
607 institutes (http://oceans.taraexpeditions.org).
608 DJR was supported by postdoctoral fellowships from the Conseil Régional de Bretagne, the Beatriu
609 de Pinós programme of the Government of Catalonia's Secretariat for Universities and Research of
610 the Ministry of Economy and Knowledge, and a fellowship from "la Caixa" Foundation (ID
611 100010434) with the fellowship code LCF/BQ/PI19/11690008. RW, DI and MRd'A were supported by
612 the Italian Flagship Project RITMARE and Premiale MIUR NEMO. MBS was supported by US NSF
613 grants OCE-1536989 and OCE-1829831, grant #3709 from the Gordon and Betty Moore Foundation,
614 and HPC support from the Ohio Super Computer.
615 We also acknowledge Stéphane Audic for assistance with metabarcoding analyses, C. Scarpelli for
616 support in high-performance computing, Mathieu Raffinot and Dominique Lavenier for discussions
617 on sequence comparison algorithms, Samuel Chaffron for help with sample contextual data, Noan Le
618 Bescot (Ternog Design) for assistance in preparing figures, and Marion Gehlen. We thank all
619 members of the *Tara* Oceans consortium for maintaining a creative environment and for their
620 constructive criticism.
621

## Author Contributions

DI, OJ, CdV, and PW designed and directed the study. IP, DJR, RW, OJ, DI, MRd'A, TV and CdV wrote the manuscript .TV, GB, NM, PP, CL and OJ designed and computed pairwise metagenomic comparisons. TV, DJR, RW, JL and PF performed the analyses of genomic data with substantial input from MRd'A, DI, OJ and PW. RW, DI, TV, PF and DJR analyzed ocean circulation simulations. GR, NH, AF-G, S Suweis, RN, J-MA, MM and EP contributed additional analysis. S Sunagawa, LG, PB, CB, MBS and EK provided additional interpretation of results. KL, EM and JP coordinated the genomic sequencing with the informatics assistance of CD, FG and J-MA. S Roux, JRB and MBS contributed viral data, PB and S Sunagawa contributed bacterial data. CB, S Romac, NH, CdV and DJR analyzed eukaryotic metabarcoding data. CD, SK, MP, S Searson and JP coordinated collection and management of *Tara* Oceans samples. *Tara* Oceans Coordinators provided support and guidance throughout the study. All authors discussed the results and commented on the manuscript.

## Author Information

The authors 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 in whose waters the *Tara* Oceans expedition sampled. Metagenomic and metabarcoding sequencing reads have been deposited at the European Nucleotide Archive under accession numbers provided in Supplementary Table 1. Contextual metadata of *Tara* Oceans stations are available in Supplementary Table 2. Metagenomic dissimilarity, OTU community dissimilarity, simulated travel times and geographic distances are provided in Supplementary Tables 3-16. All Supplementary Tables, in addition to tables of 18S V9 barcodes and OTUs and the V9 reference database are available on FigShare at the following URL: http://doi.org/10.6084/m9.figshare.11303177
The authors declare no competing financial interests.
Correspondence and requests for materials should be addressed to Olivier Jaillon, Daniele Iudicone, Maurizio Ribero d'Alcalà, Colomban de Vargas.

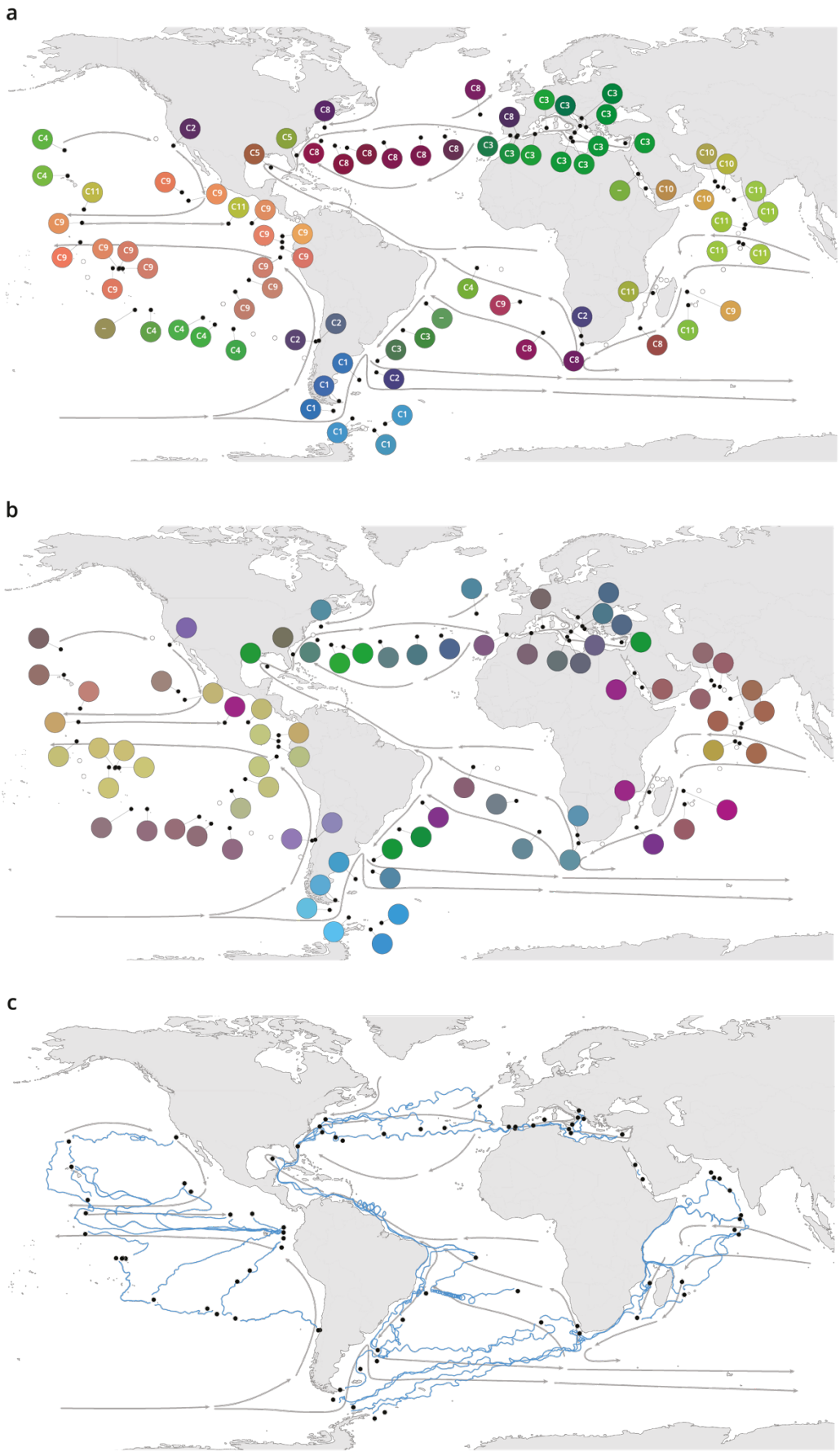



**Figure 1 | Plankton biogeography, environmental variation and ocean transport among *Tara* Oceans stations.** Major currents are represented by solid arrows. **a**, Genomic provinces of *Tara* Oceans surface

655 samples for the 0.8-5 µm size fraction, each labeled with a letter prefix ('C' represents the 0.8-5 µm size
656 fraction) and a number; samples not assigned to a genomic province are labeled with '-'. Maps of all six size
657 fractions and including DCM samples are available in Supplementary Fig. 4. Station colors are derived from an
658 ordination of metagenomic dissimilarities; more dissimilar colors indicate more dissimilar communities (see
659 Methods). **b**, Stations colored based on an ordination of temperature and the ratio of $NO_2NO_3$ to $PO_4$ (replaced
660 by $10^{-6}$ for 3 stations where the measurement of $PO_4$ was 0) and of $NO_2NO_3$ to Fe. Colors do not correspond
661 directly between maps; however, the geographical partitioning among stations is similar between the two
662 maps. **c**, Simulated trajectories corresponding to the minimum travel time ($T_{min}$) for pairs of stations (black
663 dots) connected by $T_{min} < 1.5$ years. Directionality of trajectories is not represented.

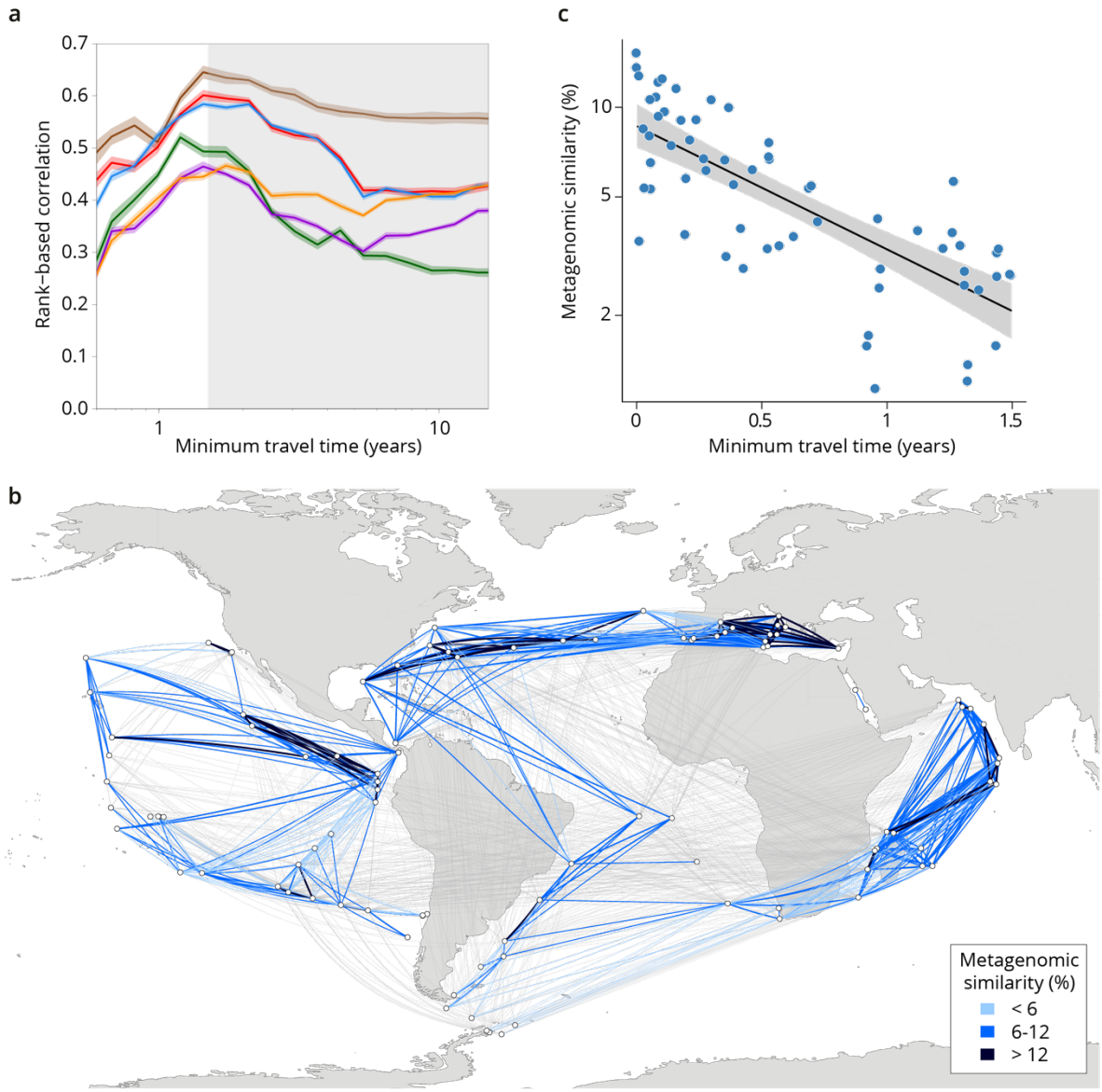



**Figure 2 | Metagenomic dissimilarity and travel time of plankton are maximally correlated up to ~1.5 years.** **a**, Spearman rank-based correlation by size fraction between metagenomic dissimilarity and minimum travel time along ocean currents ($T_{min}$) for pairs of *Tara* Oceans samples separated by a minimum travel time less than the value of $T_{min}$ on the x axis. Brown line: 0-0.2 µm size fraction, red: 0.22-1.6/3 µm, blue: 0.8-5 µm, green: 5-20 µm, purple: 20-180 µm, orange: 180-2000 µm. Shaded colored areas represent 95% confidence intervals. $T_{min}$ >1.5 years is shaded in grey. See plots for OTU dissimilarity in Supplementary Fig. 9. **b**, Pairs of *Tara* stations connected by $T_{min}$ <1.5 years in blue/black and >1.5 years in grey. Shading reflects metagenomic similarity from the 0.8-5 µm size fraction. **c**, The relationship of metagenomic similarity to $T_{min}$ with an exponential fit (black line, grey 95% CI), for pairs of surface samples in the 0.8-5 µm size fraction within the North Atlantic and Mediterranean current system (see map and plots for other size fractions and OTUs in Supplementary Fig. 10, and Supplementary Information 1 for a discussion of metagenomic similarity).

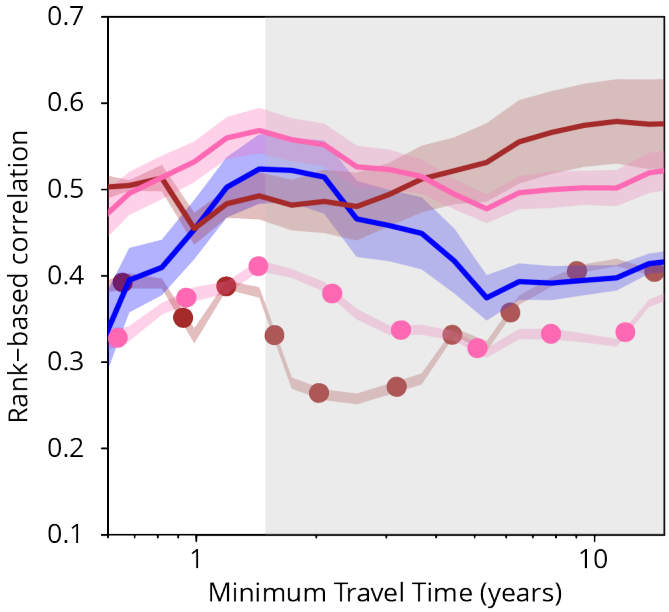
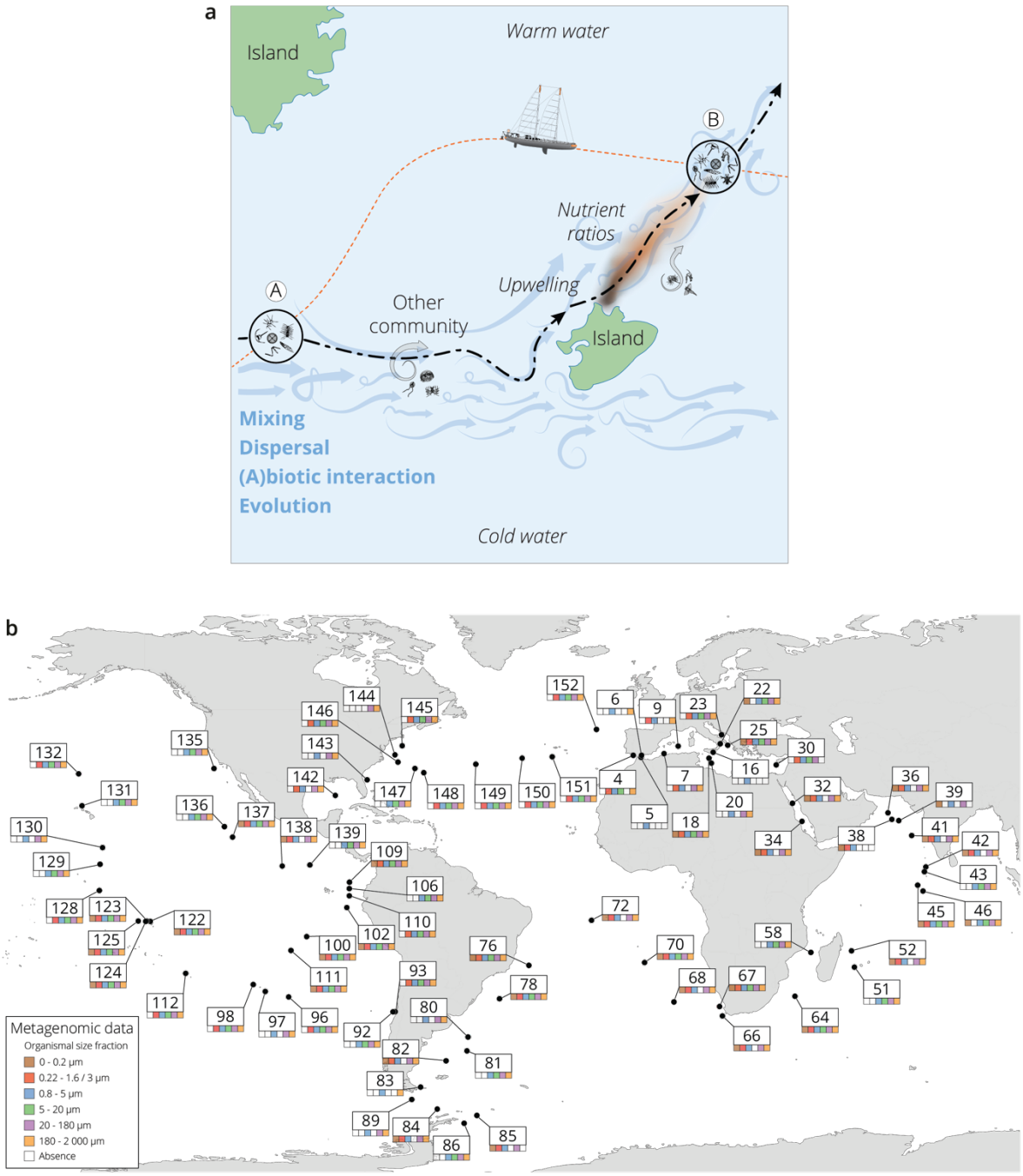
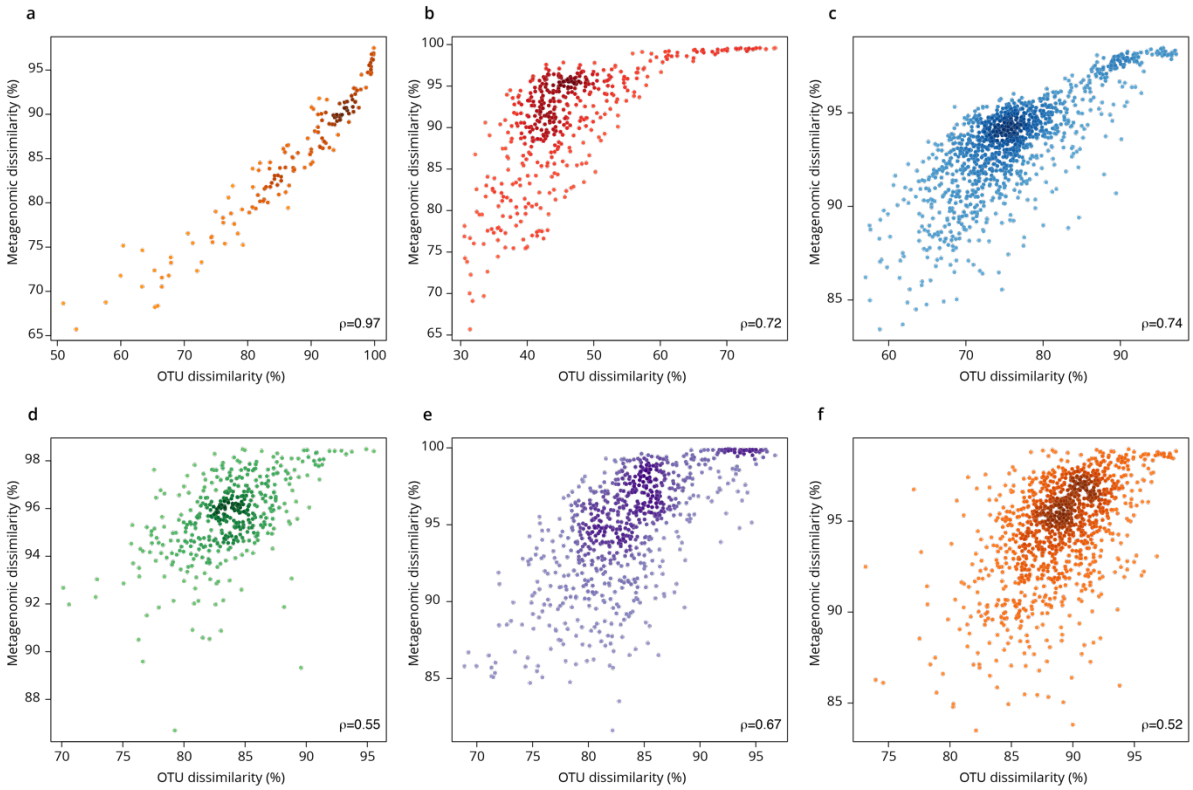


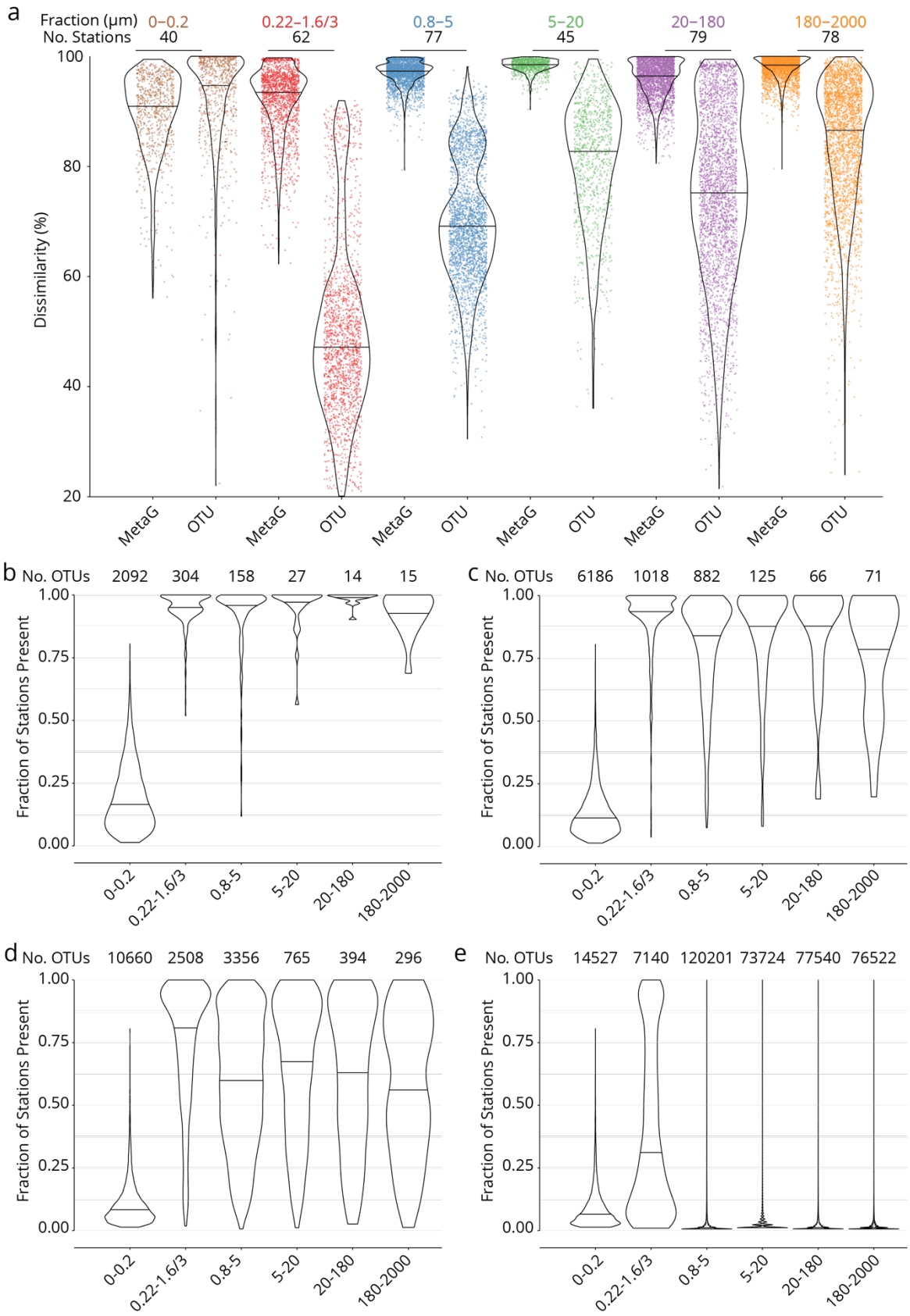

**Figure 3 | Plankton travel time, metagenomic dissimilarity and environmental differences show different temporal patterns of pairwise correlation.** Spearman rank-based correlations between metagenomic dissimilarity and minimum travel time ($T_{min}$, blue), metagenomic dissimilarity and differences in $NO_2NO_3$, $PO_4$ and Fe (pink), metagenomic dissimilarity and differences in temperature (red), $T_{min}$ and differences in $NO_2NO_3$, $PO_4$ and Fe (pink, dashed), and $T_{min}$ and differences in temperature (red, dashed) for pairs of *Tara* Oceans samples separated by a minimum travel time less than the value of $T_{min}$ on the x axis. Shaded regions represent standard error of the mean. Correlations represent averages across four of six size fractions represented in Fig. 2a; the 0-0.2 µm and 5-20 µm size fractions are excluded due to a lack of samples at the global level. Individual size fractions, partial correlations, and correlations with OTU data are in Supplementary Fig. 9.




**Supplementary Figure 1 | The seascape, plankton transport and community metagenomic samples of *Tara* Oceans stations. a**, A community sampled at a given location (A) changes over time as it travels along ocean currents (dashed bold line) to a second location (B). It is affected by numerous external processes, including mixing with water containing other communities and changes in local nutrient concentration, and by internal processes, such as biotic interactions. In this study, the *Tara* schooner followed a sampling route (orange dashed line) leading to an elapsed time between the 2 sampling sites A and B that was independent of plankton travel time. **b**, Location, station number, and sequenced surface metagenomic samples.

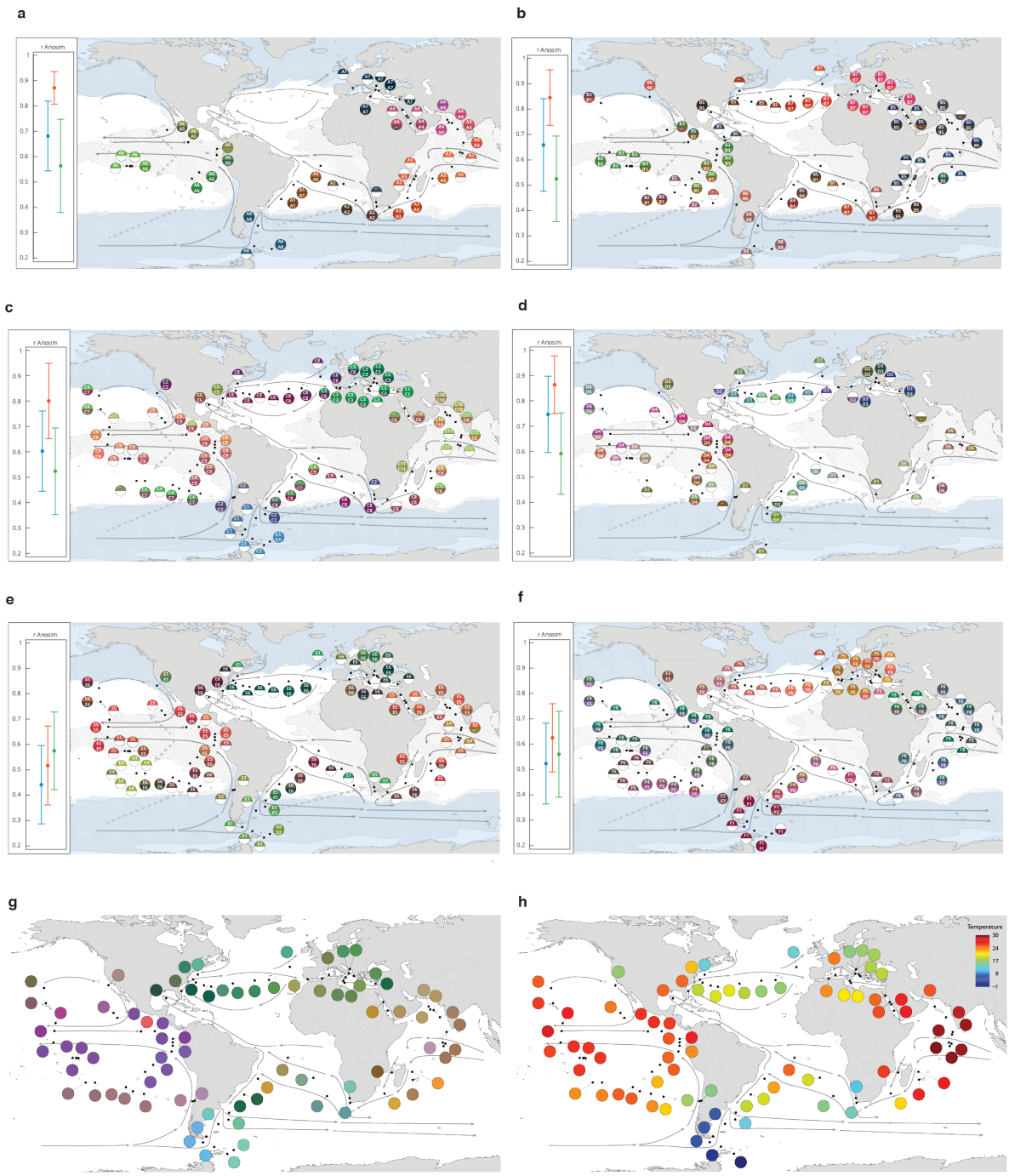



**Supplementary Figure 2 | β-diversity estimates from metagenomic and OTU-based dissimilarity are correlated.** Scatter plots of metagenomic dissimilarity versus OTU community dissimilarity for six organismal size fractions. Each point represents a pairwise comparison between two samples. **a**, 0-0.2 µm size fraction. **b**, 0.22-1.6/3 µm size fraction. **c**, 0.8-5 µm size fraction. **d**, 5-20 µm size fraction. **e**, 20-180 µm size fraction. **f**, 180-2000 µm size fraction. Global rank-based correlations (Spearman, $p \leq 10^{-4}$) are indicated in the bottom right of each plot.



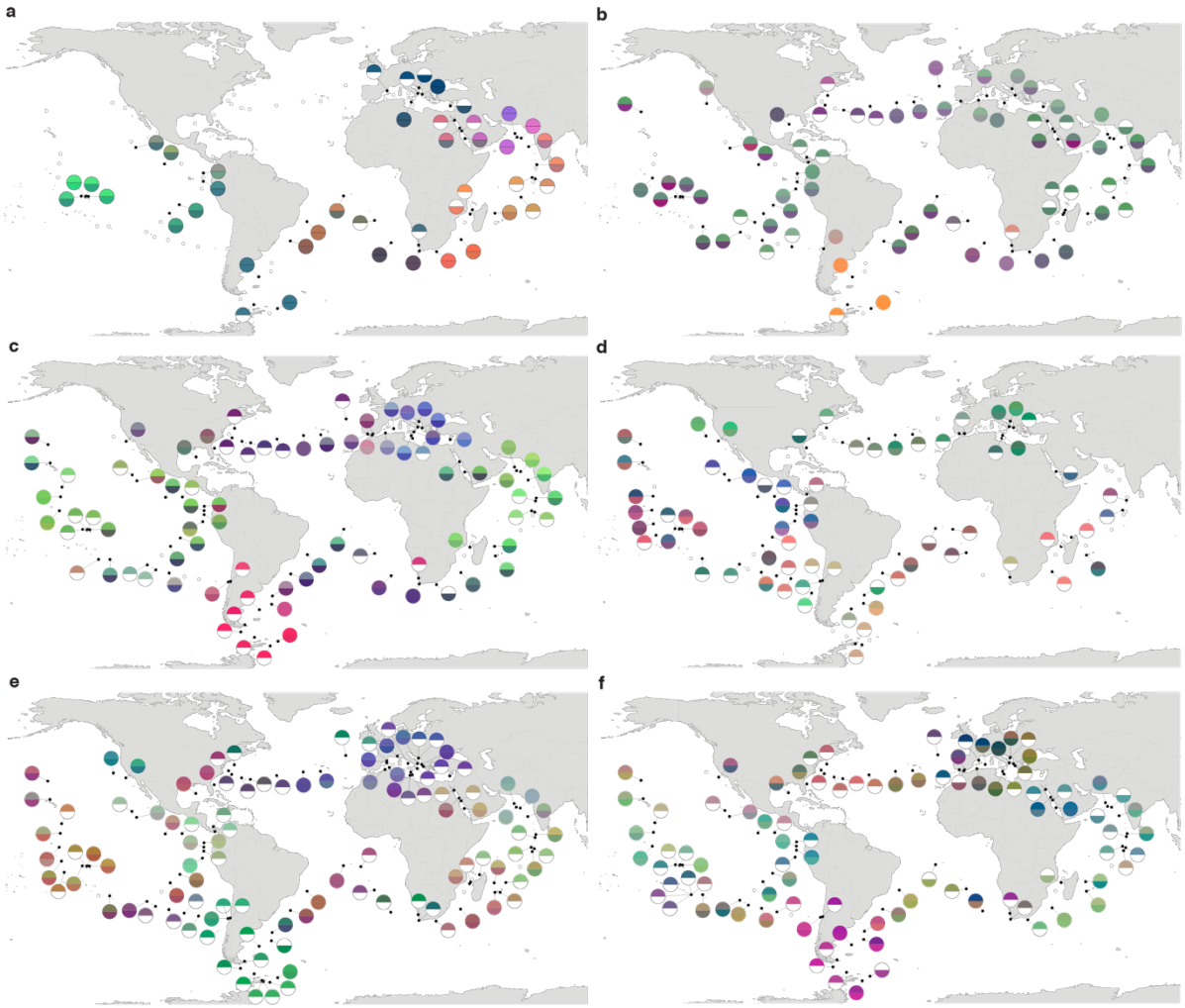

**Supplementary Figure 3** | **Global dissimilarity and OTU occupancy. a**, Distributions of dissimilarity for six organismal size fractions (measured either as metagenomic or OTU dissimilarity; see Supplementary Information 1). One colored point represents one pair of stations. Violin plots (horizontal line: median) summarize each distribution. The number of stations in common between the metagenomic/OTU data sets

within each size fraction is indicated above. **b-e**, **OTU occupancy for different proportions of total abundance.** Fraction of stations present (occupancy) for the minimum number of OTUs (indicated above) necessary to represent different proportions of the total abundance within each organismal size fraction. A relatively small number of abundant and cosmopolitan taxa represents the majority of the abundance within each size fraction; this effect is more pronounced with increasing organismal size. **b**, OTUs representing 50% of the total abundance within each size fraction. **c**, 80%. **d**, 95%. **e**, 100% (all OTUs).

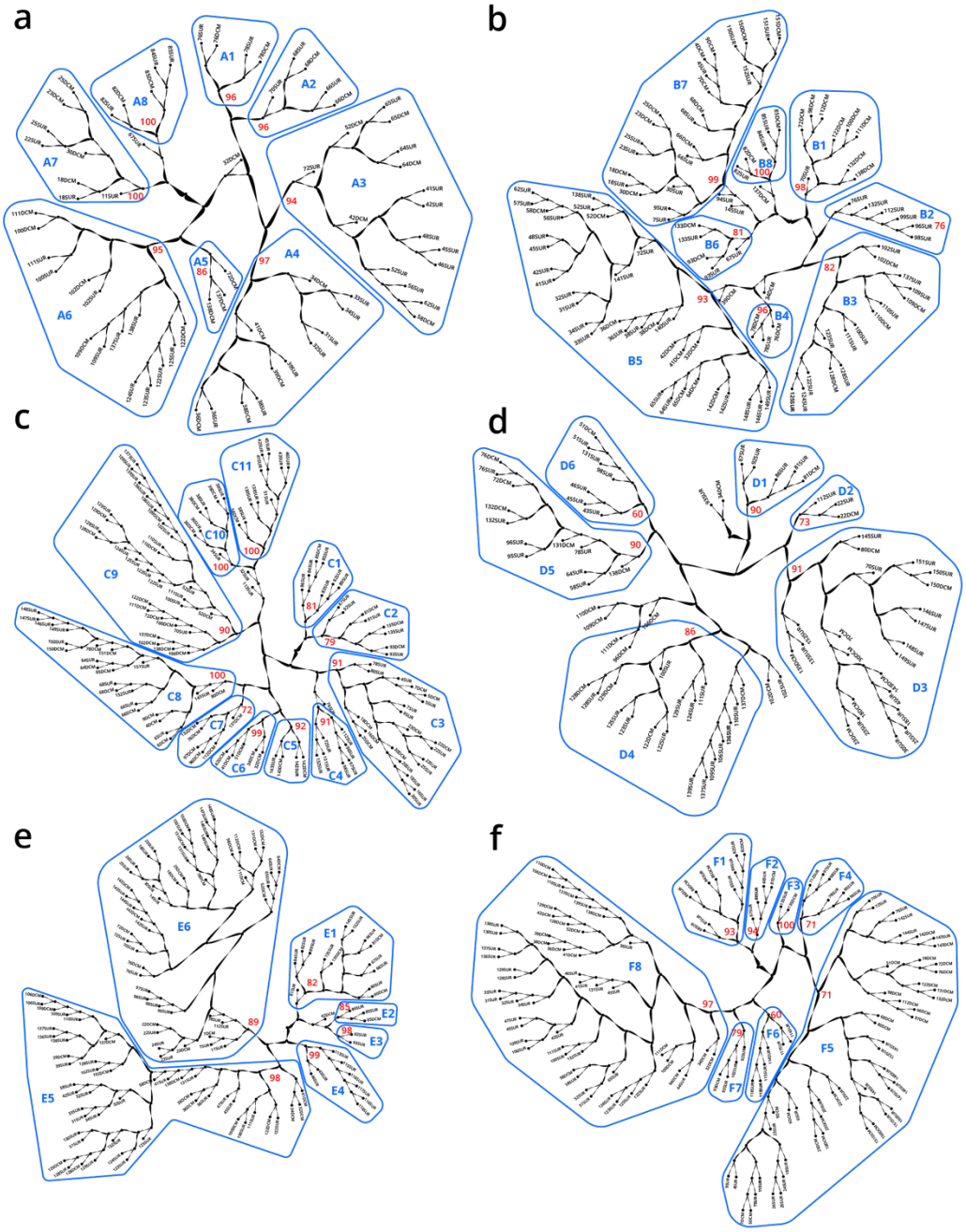



**Supplementary Figure 4 | Genomic provinces in comparison to previous ocean divisions, and ordination maps of environmental parameters. a-f**, Geographical maps of genomic provinces by organismal size fraction (see Supplementary Information 2). Circles denote stations with data available for the size fraction and contain the corresponding genomic province identifiers (one letter prefix per size fraction (A-F); stations not assigned to genomic provinces are shown as '-'). The top portion of each circle represents samples collected at the surface and the bottom portion represents the deep chlorophyll maximum (stations missing metagenomic data for one of the two depths are drawn as half circles). Colors are based on PCoA-RGB (Methods) and do not correspond among size fractions. Major currents are shown with solid black arrows, wind transport with dashed grey arrows. Blue zones indicate temperature < 14 °C. Hashed zones indicate phosphate concentration > 0.4 mmol. Hierarchical dendrograms that were used to build genomic provinces are shown in Supplementary Fig. 6. Maps with colors based on OTU dissimilarity are shown in Supplementary Fig. 5. **a**, 'A' prefix, 0-0.2 µm size fraction. **b**, 'B' prefix, 0.22-1.6/3 µm. **c**, 'C' prefix, 0.8-5 µm. **d**, 'D' prefix, 5-20 µm. **e**, 'E' prefix, 20-180 µm. **f**, 'F' prefix, 180-2000. **Insets**, Results of ANOSIM to determine, independently for each size fraction, the ability of three nested levels of ocean partitioning to explain metagenomic dissimilarities among stations (blue,

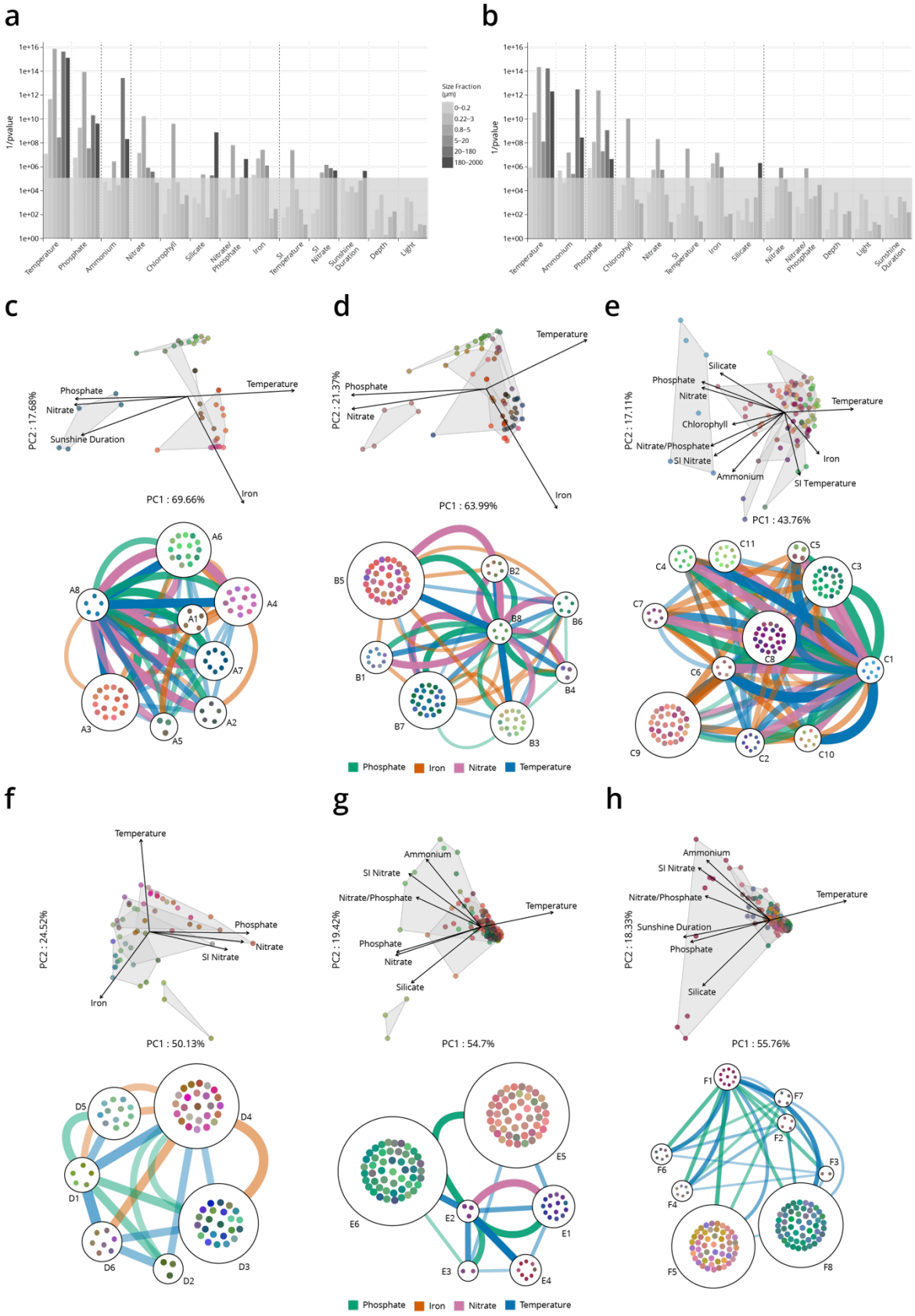
727 Longhurst biomes; red, Longhurst biogeochemical provinces; green, Oliver and Irwin objective provinces; see
728 Methods and Supplementary Information 3). **g**, The distribution of temperature and nutrient variations
729 matches the biogeography of small plankton (< 20 μm). Stations are colored based on an ordination of
730 Euclidean distances in temperature, $NO_2NO_3$, $PO_4$ and Fe. **h**, The distribution of temperature matches the
731 biogeography of large plankton (> 20 μm). Stations are colored following a Box-Cox transformation (Methods).




**Supplementary Figure 5 | Biogeography based on an ordination of OTU dissimilarity. a-f**, Principal coordinates analysis (PCoA)-RGB color maps for OTUs (see Methods). The top of each half circle represents samples collected at the surface and the bottom portion represents the deep chlorophyll maximum (stations missing OTU data for one of the two depths are drawn as half circles). Station colors do not correspond among size fractions. **a**, 0-0.2 µm size fraction. **b**, 0.22-1.6/3 µm. **c**, 0.8-5 µm. **d**, 5-20 µm. **e**, 20-180 µm. **f**, 180-2000 µm.

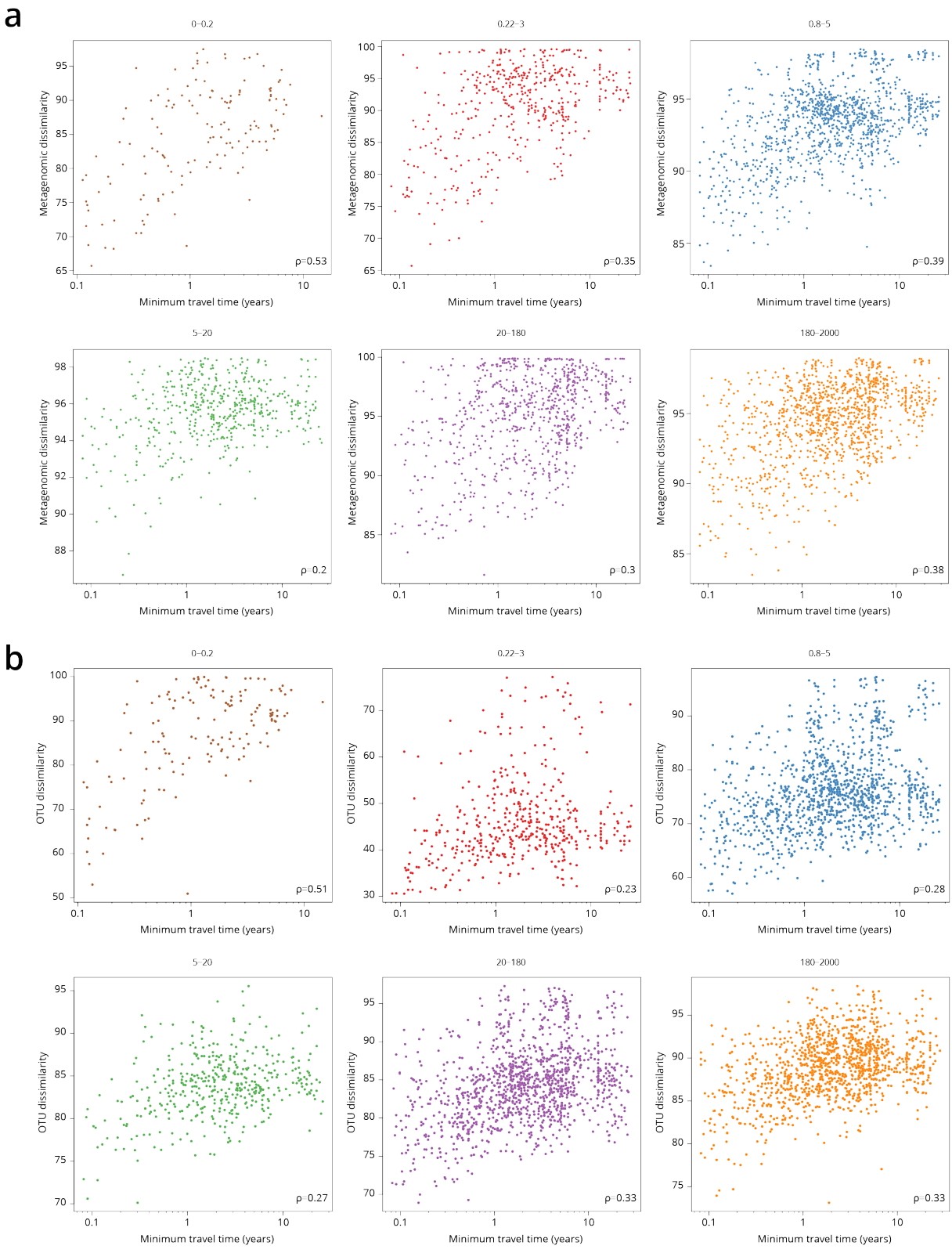



**Supplementary Figure 6 | Hierarchical trees illustrating how samples were partitioned into genomic provinces.** Dendrograms resulted from UPGMA clustering. Each sample (SUR: surface, DCM: deep chlorophyll maximum) is shown as a leaf. Genomic provinces are shown with their identifiers in blue polygons; identifiers are composed of one letter prefix per size fraction (A-F) and a number. Bootstrap values in red show the support at the key nodes that separate genomic provinces from one another. See also Supplementaryt Information 2 on the robustness of genomic provinces. **a**, 'A' prefix, 0-0.2 µm size fraction. **b**, 'B' prefix, 0.22-1.6/3 µm. **c**, 'C' prefix, 0.8-5 µm. **d**, 'D' prefix, 5-20 µm. **e**, 'E' prefix, 20-180 µm. **f**, 'F' prefix, 180-2000 µm.

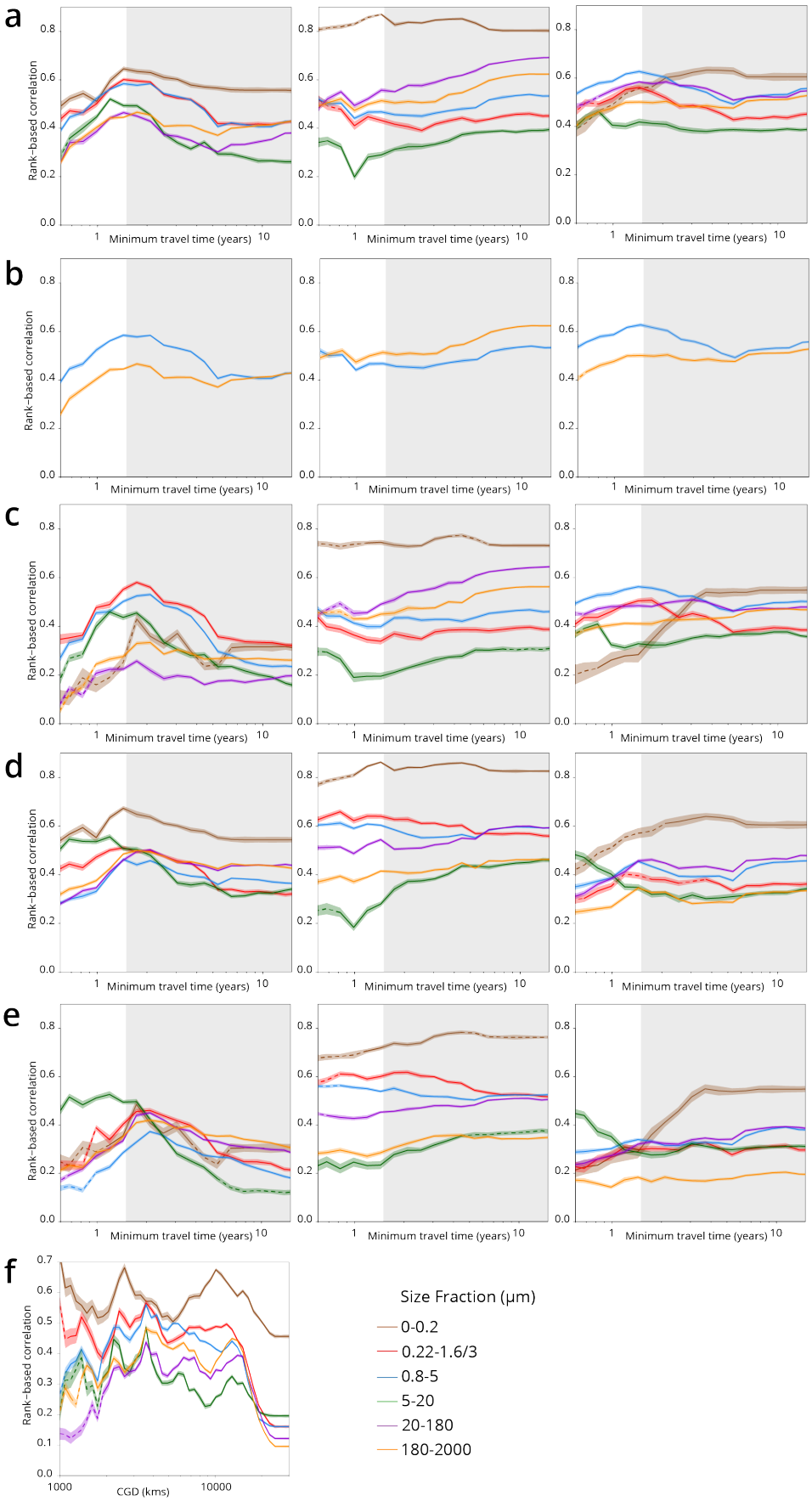



**Supplementary Figure 7 | Environmental parameters that distinguish genomic provinces. a-b**, Environmental parameters that significantly differentiate among genomic provinces (Kruskal-Wallis test, grey box indicates p values > $10^{-5}$). SI = Seasonality Index. **a**, all stations. **b**, Antarctic stations removed (see Methods). Eliminating Antarctic stations does not result in a large change in the parameters that significantly differentiate among

provinces. **c-h**, Two types of visualizations of the relationships between genomic provinces and environmental parameters. Sample colors are those from Supplementary Fig. 4. **Top plots within panels c-h**: principal components analysis-based visualization. Samples, and environmental parameters differing significantly ($p \leq 10^{-5}$) among genomic provinces, are projected onto the first two axes of variation. Grey polygons enclose different genomic provinces. **Bottom plots within panels c-h**: network-based visualization. Each genomic province is represented as a node, with the individual samples composing the province within the node. Edges between nodes represent differences in temperature, nitrate, phosphate and iron that significantly differentiate ($p \leq 10^{-5}$) among genomic provinces, that are statistically significantly different between individual pairs of genomic provinces (*post hoc* Tukey test, $p < 0.01$) and whose difference in median parameter values is ≥ 1 standard deviation (calculated from the parameter values of all samples in the size fraction). Thicker edges represent larger differences. **c**, 0-0.2 µm size fraction. **d**, 0.22-1.6/3 µm. **e**, 0.8-5 µm. **f**, 5-20 µm. **g**, 20-180 µm. **h**, 180-2000 µm.

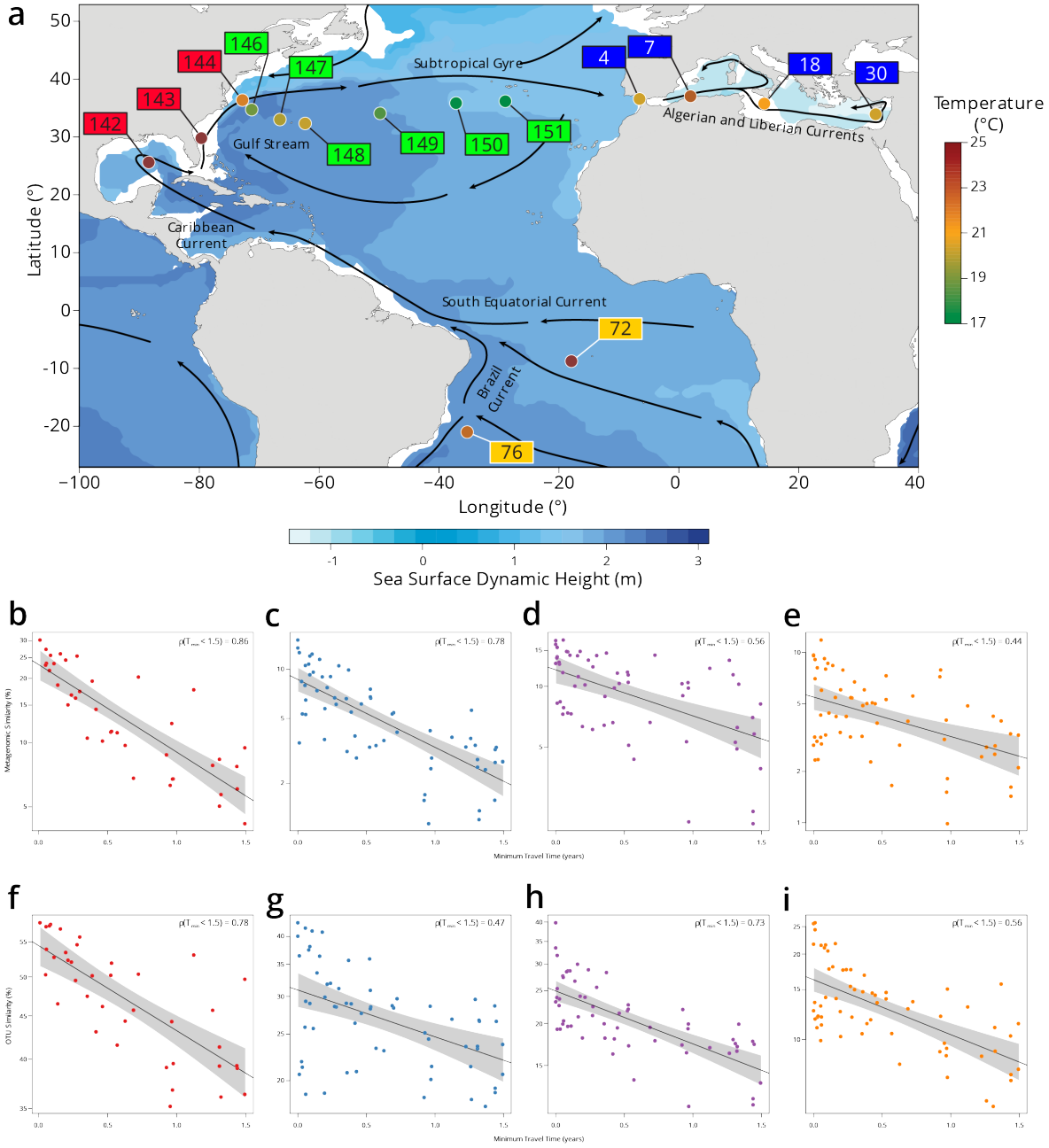



**Supplementary Figure 8 | Global correlations of dissimilarity with minimum travel time ($T_{min}$).** Scatter plots of dissimilarity versus $T_{min}$. One point represents a pair of samples. **a**, metagenomic dissimilarity. **b**, OTU dissimilarity. Global Spearman correlation values are indicated within each panel.




**Supplementary Figure 9 | Plankton travel time, dissimilarity, environmental distance and geographic distance show different temporal patterns of pairwise correlation.** Spearman correlation values are shown

separately by organismal size fraction. Non-significant correlations ($p > 0.01$) are shown with dashed lines. **a-e**, Correlations for pairs of *Tara* Oceans samples separated by a minimum travel time less than the value of $T_{min}$ on the x axis. $T_{min}$ >1.5 years is shaded in grey. Left panels: correlation of dissimilarity with $T_{min}$; middle panels, dissimilarity with temperature; right panels: dissimilarity with differences in $NO_2NO_3$, $PO_4$ and Fe. **a-c**, metagenomic dissimilarity. **d-e**, OTU dissimilarity. There is a maximum correlation of dissimilarity with $T_{min}$ (and, for most size fractions, of dissimilarity with nutrients) for $T_{min}$ <~1.5 years, but the correlation between dissimilarity and temperature does not display a similar maximum. **b** displays only the 0.8-5 µm (blue) and 180-2000 µm (orange) size fractions from **a**, to highlight that for smaller plankton, correlations with differences in nutrient concentrations were stronger for $T_{min}$ up to ~1.5 years, but for larger plankton, correlations were stronger with temperature variations for $T_{min}$ beyond ~1.5 years. **c** and **e**, Partial correlations to estimate the independent effects of $T_{min}$ and environmental distances on β-diversity. Left panels: controlling for differences in temperature and for differences in $NO_2NO_3$, $PO_4$ and Fe; middle and right panels: controlling for $T_{min}$. Partial correlations do not affect the maximum correlation of dissimilarity with $T_{min}$ for $T_{min}$ <~1.5 years. **f**, Correlation of geographic distance (without traversing land) with metagenomic dissimilarity for pairs of *Tara* Oceans samples separated by a geographic distance less than the value on the x axis.




**Supplementary Figure 10 | Plankton community composition turnover through the North Atlantic. a**, Map of *Tara* Oceans stations, currents (solid lines), temperature by station (colored circles) and sea surface climatological dynamic height from CARS2009 (http://www.cmar.csiro.au/cars). Each station label has a color corresponding to a sub-region: South Atlantic in orange, Gulf Stream in red, Recirculation/Gyre in green and Mediterranean Sea in blue. **b-e**, Scatter plots of metagenomic similarity versus minimum travel time ($T_{min}$) for these stations in the **b**, 0.22-3 µm; **c**, 0.8-5 µm; **d**, 20-180 µm; and **e**, 180-2000 µm size fractions. **f-i**, Scatter plots of OTU community similarity for the **f**, 0.22-3 µm; **g**, 0.8-5 µm; **h**, 20-180 µm; and **i**, 180-2000 µm size fractions**.** The black line represents an exponential fit, with a light grey shaded 95% confidence interval. The resulting turnover times using metagenomic similarity are τ = 0.91 y for 0.22-3 µm, τ = 0.91 y for 0.8-5 µm, τ = 2.22 y for 20-180 µm and τ = 1.99 y for 180-2000 µm. Turnover times using the OTU community similarity are τ = 4.23 y for 0.22-3 µm, τ = 4.08 y for 0.8-5 µm, τ = 2.6 y for 20-180 µm and τ = 2.1 y for 180-2000 µm. The viral-enriched 0-0.2 µm and the nanoplanktonic 5-20 µm size fractions are not shown due to insufficient sampling of these stations.

# Supplementary Information

## *Supplementary Information 1. Comparison of metagenomes and OTUs*

Metagenomic comparisons reflect fine-scale differences in genome content at the community level as a function of diversity, genome size and organismal abundance, and also depend on the rate of evolution of each specific lineage. With exhaustive sampling, metagenomic dissimilarity could theoretically distinguish among genomes in a sample separated by a single mutation. However, our metagenomic sequencing depth was likely not able to reach saturation due to the number of genomes per sample and their putative large size (metatranscriptomes, which contain fewer sequences per species than do metagenomes, did not reach saturation within *Tara* Oceans samples[53]). For example, if for a pair of samples we sequence 50% of the total amount of the unique genomic DNA present, we expect the maximum similarity of the two samples to be roughly 25% (0.5 x 0.5). Therefore, the pairwise metagenomic dissimilarities we calculated between samples probably reflected a combination of genomic differences weighted towards more abundant organisms. In contrast, OTUs, obtained by sequencing single marker genes, approach biodiversity saturation[5,18,19]. However, OTU resolution depends on the choice of the marker to be used, the threshold of similarity for the marker, and its lineage-specific substitution rate, and may therefore confound evolutionarily and/or ecologically distant organisms[54–58]. We observed a significant agreement between the two proxies (Supplementary Fig. 2), although dissimilarities based on OTUs were generally lower than those computed from metagenomic data (Supplementary Fig. 3a).

Analyses of plankton biogeography produced consistent results based on metagenomic and OTU data (Supplementary Fig. 4, Supplementary Fig. 5, Supplementary Fig. 8, Supplementary Fig. 9). For simplicity, in the main text, we chose to highlight results based on metagenomes rather than on OTUs for three reasons. First, the metagenomic sequencing protocol and subsequent measurement of dissimilarity was uniform across size fractions, whereas OTUs were defined differently for the viral-enriched, bacterial-enriched and eukaryote-enriched size fractions (Methods). Second, the biogeographical patterns we obtained (see below) may be more evident in comparisons among metagenomic sequences (our data source in identifying genomic provinces), as genomes, accumulate single-base changes and other variants more quickly than a single ribosomal gene marker. Third, β-diversity estimated by metagenomic dissimilarity generally displayed higher correlation values with minimum travel time ($T_{min}$; Supplementary Fig. 8).

## *Supplementary Information 2. Robustness of genomic provinces*

We assessed the robustness of genomic provinces in five separate ways. First, we tested 5 different hierarchical clustering algorithms from R-package pvclust_1.3-2[40] (UPGMA - Unweighted Pair Group Method with Arithmetic mean; McQuitty's method; Complete linkage; Ward's method; Single linkage) on the metagenomic pairwise dissimilarities produced by Simka separately for the six organismal size fractions, followed by multiscale bootstrap resampling. We used the cophenetic correlation coefficient from the R-package dendextend_1.5.2[59] to measure how accurately the dendrograms produced by each method preserved the pairwise distances within the input dissimilarity matrices[60,61]. The ranking of the cophenetic correlation coefficient for different clustering methods within each size fraction was consistent with a published large-scale methodological comparison of clustering methods for biogeography (Supplementary Table 17), which considered UPGMA agglomerative hierarchical clustering to have consistently the best performance[39]. Second, we compared clustering results among all size fractions using Baker's Gamma Index[62] from the R-package corrplot_0.77[63], which is a measure of association (similarity) between two trees based on hierarchical clustering (dendrograms). The Baker's Gamma Index is defined as the rank correlation between the stages at which pairs of objects combine in each of the two trees. For each type of correlation, the UPGMA was consistently the most correlated with other clustering methods (Supplementary Table 18). This allowed us to conclude, in

agreement with previous results[39], that the UPGMA method is likely more robust than the other methods we tested.

Third, we compared the genomic provinces found by our UPGMA hierarchical clustering approach to those found by two different non-hierarchical methods: K-means on the positions found by multidimensional scaling and spectral clustering on the nearest-neighbor graph. Both methods rely on (i) a dissimilarity matrix and (ii) a tuning parameter (dimension of the projection space for K-means, and number of neighbors for spectral clustering). K-means uses the numeric values of the dissimilarities, whereas spectral relies only on their ordering (e.g., community A is closer to B than to C). We compared the genomic provinces to clusters found by K-means and spectral clustering for all values of the tuning parameter using the Rand Index (RI; from the GARI function of the loe R package version 1.1[64]), a score of agreement between partitions. Results are reported as mean +/- s.d. of the RI: 1 means perfect agreement and 0 complete disagreement. Fourth, in order to assess the significance of the genomic provinces, we performed a multivariate ANOVA to partition metagenomic dissimilarity across regions, using the adonis function of the vegan R package version 2.5-4[37]. Note, however, that since the same data were used both to construct the genomic provinces and to assess their significance, the p-values estimated by ADONIS might be anti-conservative. The results of the third and fourth analyses are presented in Supplementary Table 19.

Fifth, we found that clustering of samples in genomic provinces was consistent with a complementary visualization based on the same data: RGB colors derived from the first three axes of a principal coordinates analysis (PCoA-RGB) of β-diversity, in which similar colors represent similar communities (Supplementary Fig. 4; see Methods). Samples within the same genomic province generally shared the same range of PCoA-RGB colors. Because the clustering approach was hierarchical, samples sharing some similarity could have been assigned to different genomic provinces due to binary decisions during the clustering process. This was also reflected in the PCoA-RGB colors, where the boundaries of genomic provinces did not indicate a complete change of communities among genomic provinces (and, conversely, belonging to the same genomic province did not imply identical community). Nonetheless, samples with similar PCoA-RGB colors were generally situated in closely-related branches in the UPGMA tree (Supplementary Fig. 6). An illustrative example is genomic province F5 (of the 180-2000 µm size fraction; Supplementary Fig. 4f), which encompassed stations in the Atlantic, Mediterranean Sea and some subtropical stations in the Indo-Pacific. In this wide region, the PCoA-RGB colors indicate the variation in community composition within the genomic province, and also reflect the relatedness of F5 to its adjacent samples, in particular those in the subtropical Atlantic/Pacific region F4, its neighbor in the UPGMA tree (Supplementary Fig. 6f).

### *Supplementary Information 3. Comparison of genomic provinces to previous biogeographical divisions*

Current approaches in biogeographic theory divide the ocean into regions based either on expert knowledge applied to satellite data, as in the hierarchical nesting by Longhurst[3] into biomes (macro-scale, essentially representing a division of the world's oceans into cold and warm waters, and coastal upwelling zones) and biogeochemical provinces (BGCPs, areas within biomes defined by observable boundaries and predicted ecological characteristics), or, alternatively, into the objective provinces of Oliver and Irwin[49], which are based solely on statistical analyses. Longhurst BGCPs are based upon, primarily, monthly variations of chlorophyll a, the geography of the seasonal cycle of physical factors (such as the depth of the upper ocean mixed layer) and surface temperatures. In turn, these ocean properties are strongly modulated by oceanic currents (for example, moderate to large mixed layer depths are observed generally on the poleward side of the subtropical gyres). In contrast, the objective global ocean biogeographic provinces proposed by Oliver and Irwin[49] were based upon clustering temporal variability of chlorophyll concentration and surface temperatures, both measured from satellite data. They combined a proxy for the intensity of primary productivity with water temperature, therefore emphasizing regions similar in their temporal variability for both properties

(which essentially corresponds to the seasonal cycle). None of these ocean partitionings directly considered organismal community composition.

We tested whether genomic provinces were comparable with these partitionings by performing an analysis of similarity (ANOSIM; Supplementary Fig. 4, insets; Methods). The four small size classes, 0-0.2 µm, 0.22-1.6/3 µm, 0.8-5 µm, and 5-20 µm (Supplementary Fig. 4a-d) were more consistent with Longhurst BGCPs. In contrast, for the two larger size fractions 20-180 µm and 180-2000 µm, the three biogeographical divisions were not strongly different within the ANOSIM (Supplementary Fig. 4e-f).

From an oceanographic point of view, plankton should be quasi-neutrally redistributed (i.e., homogenized) by currents and their biogeography should follow the structure of the main recirculations, within a range of physiologically compatible temperatures. In this point of view, our results are consistent with the large-scale geographic distributions found by Hellweger *et al*.[4] using a neutral model.

### *Supplementary Information 4. Differences in genomic province sizes among organismal size fractions*

Globally, we obtained more numerous, smaller genomic provinces in the smaller size fractions and fewer, larger genomic provinces in the larger size fractions (Supplementary Fig. 4, Supplementary Fig. 7). We observed a similar pattern using OTU data (Supplementary Fig. 5). Whereas smaller size fractions generally lacked geographically widespread genomic provinces containing numerous *Tara* Oceans samples, the two largest size fractions were both characterized by two very widespread genomic provinces: F5 and F8 for the 180-2000 µm size fraction, and E5 and E6 for the 20-180 µm size fraction. These large genomic provinces were latitudinally limited by the boundary between the subtropics and subpolar regions, and spanned different oceanic basins. Notably, in the Southern Hemisphere the subtropical gyres actually form a single supergyre[65] and there are almost no metabolic (mainly temperature) barriers between the northern and southern subtropical gyres (see Supplementary Fig. 4), potentially explaining genomic provinces in the 20-180 µm and 180-2000 µm size fraction that contain samples from the North and South Atlantic. For example, in the 180-2000 µm size fraction, F5 mostly covered the North and South Atlantic Oceans and adjacent systems, and F8 covered the Indo-Pacific low- and mid-latitudes. No clear correspondence existed with biogeochemical patterns (e.g., nutrient ratios), except for the clusters coinciding with upwelling systems (F3 for the California upwelling, F7 for the Chile-Peru upwelling and F2 for the Benguela upwelling system) and for the samples collected at the deep chlorophyll maximum (DCM) in the Pacific subtropical gyres (F5); this is consistent with the comparison of genomic provinces to previous biographical divisions, in which the genomic provinces of smaller size fractions were more consistent with Longhurst BGCPs, but those of larger size fractions were not (Supplementary Information 3). A bimodal zooplankton species distribution (split into subtropical and subpolar communities, with ubiquitous warm water species) was also detected by a recent study on copepod population dynamics that used alternative approaches to analyze the same metagenomic dataset[66] (see their Fig. 2). More locally, within the North Atlantic (see also Supplementary Information 6), along the northern boundary of the subtropical gyre, cold and warm copepod species overlapped because of cross-current dispersal. Nonetheless, although both cold and warm species appeared to be able to travel long distances, mixing among them was not sufficient to create a local genomic province in our data.

We interpret the difference in genomic province sizes between smaller and larger size fractions as the result of various factors. Plankton smaller than 20 µm (femto-, pico- and nanoplankton), which represent most of the prokaryotic and eukaryotic phototrophs[18,19], are sensitive to a suite of environmental factors (i.e., temperature[67], nutrients and trace elements[10]; see also Supplementary Fig. 7) and generally have a shorter life cycle, together leading to faster fluctuations in their relative abundance in the communities we sampled. In contrast, larger plankton have longer life cycles and, if they are predators that are not strongly selective in their feeding, or are photosymbiotic hosts capable of partnering with multiple different symbionts, may cope with local fluctuations in environmental

conditions. Therefore, they should be affected primarily by large scale, mostly latitudinal, variations in the environment, leading to larger genomic provinces, whereas smaller plankton are grouped into smaller provinces more influenced by local environmental conditions. Overall, this difference in biogeography suggests a size-based decoupling between smaller and larger plankton (which may also extend to nekton such as tuna and billfish[68]), with implications for the structure and function of oceanic food webs and other types of biotic interactions.

***Supplementary Information 5. Genomic provinces as stable ecological continua***

As plankton communities are transported by ocean currents, they change over time due to the various processes that occur in the context of the seascape: variations in temperature, light and nutrients (where changes in the latter may also be induced by plankton communities), intra- and inter-individual and species biological interactions, and mixing with neighboring water masses. Thus, a continuum of composition among nearby samples is expected as a natural consequence of community turnover within the seascape over time. We observed the effects of continuous turnover in our biogeographical analyses (Fig. 1a, Supplementary Fig. 4, Supplementary Fig. 5, Supplementary Information 2) in which nearby samples often reflected gradual, but not complete changes in community composition.

We measured the time window of transport by currents separating two samples during which the changes in their community composition were maximally correlated with travel time, resulting in a global average of $T_{min}$ < roughly 1.5 years. This represents the travel time during which predictable continuous turnover occurs in our dataset. Notably, $T_{min}$ does not necessarily define the turnover rate itself which depends on how strongly different seascape processes affect communities with differing biological characteristics (see Supplementary Information 6).

The global ocean current system is composed of a series of large-scale main currents and associated recirculations (which are also referred to as gyres). Therefore, we present the following hypothesis as a potential explanation of our results: the average global timescale of 1.5 years is comparable to the crossing time of an ocean gyre (i.e., the amount of time it takes a water parcel to travel from one side of a gyre to the other), e.g., to cross the North Atlantic basin while riding the Gulf Stream system. This time scale of 1.5 years is probably an underestimate, since our sparse sampling did not cover all current systems. Within different systems, the transport by main currents leads to stable, continuous patterns of changes in community structure and nutrient concentrations, and also explains how temporally stable genomic provinces can exist in the face of ocean circulation. Within each system we have thus to expect that a community turnover is long enough to allow for this long range predictability due to smooth, continuous changes. Significant heterogeneity in environmental conditions among different circulation patterns means that moving from system to another (and therefore, in our case here, beyond the 1.5 year timescale; Supplementary Fig. 9c-f) disrupts the interlinked relationship among local seascape processes, leading to a global delimitation into separate ecological continua among different gyre-scale current systems.

***Supplementary Information 6. Community turnover in the North Atlantic***

In order to characterize the impact of physical and biological processes on changes in metagenomic composition during travel along currents, we focused on the well-known current systems crossing the North Atlantic into the Mediterranean Sea (the Gulf Stream and other currents around the subtropical gyre[20,69–71]; Supplementary Fig. 10a). Across this region, the piconanoplankton (0.8-5 µm) were split into three genomic provinces, C5, C8 and C3, each less than 5,000 km wide (~11 months of travel time; Supplementary Fig. 4c). In contrast, mesoplankton (180-2000 µm) biogeography corresponded to a single province, F5, spanning from the Caribbean to Cyprus (> 9,700 km or ~18 months of travel time; Supplementary Fig. 4f; see also Supplementary Information 4). Metagenomic dissimilarity and $T_{min}$ were strongly correlated within the region (Spearman's ρ between 0.44 and 0.86 depending on size

fraction, Supplementary Fig. 10b-e), which allowed us to explore the relationship of genomic province size, ocean transport and plankton community turnover over scales from months to years. We calculated metagenomic turnover times as e-folding times based on an exponential fit of metagenomic dissimilarity to $T_{min}$ (ranging from a few months to a few years, Methods). The metagenomic turnover time of smaller plankton (< 20 µm) was approximately one year. In contrast, for the larger size fractions, the metagenomic turnover time was approximately two years, suggesting that a lower turnover rate for larger plankton may explain their geographically larger genomic provinces.

We note that our results on metagenomic turnover time appear different from a recently published study that also calculated turnover rates for plankton, which found faster rates for larger organisms[8]. This may be explained by two significant differences between our approach and theirs: first, their measurements of β-diversity were based on presence/absence (Jaccard) comparisons among either morphological species or OTUs, whereas our calculations of turnover time above were based on metagenomic sequences. As described above (Supplementary Information 1), there are significant differences in resolution between OTU-based and metagenomic data, and we would expect similar differences in resolution between organismal observation data and metagenomic sequences. In fact, due to these differences in resolution, our estimates of metagenomic time based on OTU rather than metagenomic data show a similar trend to those of Villarino *et al*.[8] (Supplementary Fig. 10f-i). Second, their turnover rates were calculated separately for individual plankton groups (the 9 main groups were prokaryotes, coccolithophores, dinoflagellates, diatoms, all microbial eukaryotes, gelatinous zooplankton, mesozooplankton, macrozooplankton and myctophids), whereas our metagenomic data represent samples of the full plankton community within each size fraction. Among these, several groups (e.g., dinoflagellates or mesozooplankton) would be expected to be found across multiple *Tara* Oceans size fractions, blurring potential comparisons. Thus, our study and Villarino *et al*. calculated rates of change using broadly similar approaches, but based on very different underlying biological substrates.