

# 1 Genomic evidence for global ocean plankton biogeography shaped 2 by large-scale current systems

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

13 \* and \$: equal contributions

14 \*\*: corresponding authors

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

21 **Data availability:** <http://doi.org/10.6084/m9.figshare.11303177>

22 Supplemental Tables 1-19 (including DDBJ/ENA/GenBank short read archive identifiers for *Tara*  
23 Oceans metagenomic & 18S V9 sequence reads, and distance matrices), Datasets 1-3 (18S V9  
24 metabarcode and OTU tables, and reference database).

25 1 Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, 29680 Roscoff, France

26 2 Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003  
27 Barcelona, Spain

28 3 Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

29 4 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110 , 13288, Marseille, France

30 5 Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA),  
31 CNRS, Université Evry, Université Paris-Saclay, Evry, France

32 6 Changing Ocean Research Unit, Institute for the Oceans and Fisheries, University of British Columbia. Aquatic Ecosystems  
33 Research Lab. 2202 Main Mall. Vancouver, BC V6T 1Z4. Canada.

34 7 Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA.

35 8 Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France.

36 9 INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes,  
37 France.

38 10 Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, 1350  
39 Copenhagen K, Denmark

40 11 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

41 12 Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

42 13 Dipartimento di Fisica e Astronomia 'G. Galilei' & CNISM, INFN, Università di Padova, Via Marzolo 8, 35131 Padova, Italy.

43 14 MalAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

44 15 Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France

45 16 Research Federation (FR2022) Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016 Paris, France

46 17 Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry,  
47 France

54 18 Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV),  
55 Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.  
56 19 Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.  
57 20 Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.  
58 21 Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS  
59 UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.  
60 22 Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg,  
61 Germany.  
62 23 Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.  
63 24 Sorbonne Universités, UPMC Univ Paris 06, UMR 7093 LOV, F-75005, Paris, France.  
64 25 CNRS, UMR 7093 LOV, F-75005, Paris, France.  
65 26 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.  
66 27 PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.  
67 28 Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, 70808, USA  
68 29 Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.  
69 30 Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.  
70 31 Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.  
71 32 School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.  
72 33 Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA.  
73 34 Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC, Barcelona, Spain.  
74 35 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome  
75 Campus, Hinxton, Cambridge CB10 1SD, United Kingdom  
76 36 Sorbonne Universités, CNRS, Laboratoire d'Oceanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France.  
77 37 CNRS, UMR 7232, BIOM, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.  
78 38 Sorbonne Universités Paris 06, OOB UPMC, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.  
79 39 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France  
80 40 Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan.  
81 41 Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.  
82 42 VIB Center for Microbiology, Herestraat 49, 3000 Leuven, Belgium.  
83 43 CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.  
84 44 National Science Foundation, Arlington, VA 22230, USA.  
85 45 Bigelow Laboratory for Ocean Sciences East Boothbay, ME, USA.  
86 46 Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.  
87 47 Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue  
88 Lhomond, 75231 Paris Cedex 05, France.  
89  
90

## 91 Abstract

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

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

113 **Main Text**

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

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

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

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

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

168 These large-scale biogeographical patterns derived from metagenomes were linked to environmental  
169 parameters including nutrients, temperature and trophic level. Seawater temperature was  
170 significantly different among genomic provinces for all plankton size classes (Kruskal-Wallis test,  $p <$   
171  $10^{-5}$ ), corroborating previous results for prokaryotes<sup>18</sup>, whereas other environmental conditions were  
172 significantly different only with respect to specific size classes (Supplementary Fig. 7). The geography  
173 of combined nutrient and temperature variations resembled the biogeography of smaller plankton  
174 size classes (Fig. 1a-b, Supplementary Fig. 4a-d,g), whereas temperature alone more closely matched  
175 the distribution of larger plankton (Supplementary Fig. 4e,f,h), reflecting different potential ecological  
176 constraints. Many genomic provinces were spatially consistent with ocean basin-scale circulation  
177 patterns, such as western boundary currents or major subtropical gyres<sup>20</sup> (Fig. 1a, Supplementary Fig.  
178 4a-f), suggesting a particular role for large-scale surface transport (a core component of the seascape)  
179 in the emergence of spatial patterns of plankton community composition, as previously proposed<sup>21</sup>.  
180 We therefore investigated community composition differences between sampled stations in light of  
181 the corresponding transit time. We inferred the time of mean transport between stations from  
182 trajectories computed with the physically well-constrained MITgcm ocean model (see Methods),  
183 which takes into account directionalities<sup>9</sup> and meso- to large-scale circulation, potential dispersal  
184 barriers and mixing effects<sup>22,23</sup>. We quantified transport using the minimum travel time<sup>24</sup> ( $T_{\min}$ )  
185 between pairs of *Tara* stations. These trajectories corresponded to the dominant paths that transport  
186 the majority of water volume and its contents (e.g., heat, nutrients and plankton; Fig. 1c). For all  
187 plankton size classes, community composition differences between stations were correlated to travel  
188 time (Supplementary Fig. 8). Cumulative correlation values (correlations between metagenomic  
189 dissimilarity and  $T_{\min}$  computed for an increasing range of  $T_{\min}$ ) were maximal for pairs of stations  
190 separated by  $T_{\min} < \sim 1.5$  years for all size classes ( $p \leq 10^{-4}$ ; Spearman’s  $\rho$  0.45 to 0.71 depending on size  
191 class, Fig. 2a, Supplementary Fig. 9a-e), hence revealing measurable plankton community dynamics  
192 on time scales far longer than typical plankton growth rates or life cycles. In contrast, no such unimodal  
193 pattern was found for correlations between metagenomic dissimilarity and geographic distance  
194 (without traversing land; Supplementary Fig. 9f). Over the timescale  $\sim 1.5$  years, which corresponds  
195 well with the average time to travel across a basin or gyre, large-scale transport is therefore an  
196 appropriate framework for studying differences in plankton community composition (Fig. 2b). The fact  
197 that simulated transport times and metagenomic dissimilarity were correlated despite a 3 year pan-  
198 season sampling campaign highlights the overall stability of plankton dynamics along the main ocean  
199 currents.

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

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

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

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

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

252

253

## 254 **Methods**

255

### 256 **Sampling, sequencing and environmental parameters**

257 Sampling, size fractionation, measurement of environmental parameters and associated metadata,  
258 DNA extraction and metagenomic sequencing were conducted as described previously<sup>30,31</sup>. Samples  
259 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,  
260 180-2000 µm; Supplementary Fig. 1b; Supplementary Table 1) and two depths (subsurface and deep  
261 chlorophyll maximum (DCM)). The prokaryote-enriched size fraction was collected either a 0.22-1.6  
262 µm or 0.22-3 µm filter<sup>18,30</sup>.

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

$$269 \text{PAR}(Z) = \text{PAR}(0) * \exp(-k*Z)$$

$$270 x = \log(\text{Chl})$$

$$271 \log(Z) = 1.524 - 0.426x - 0.0145x^2 + 0.0186x^3$$

$$272 k = -\ln(0.01)/Z$$

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

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

282

### 283 **Calculation of metagenomic community dissimilarity**

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

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

$$296 \text{Metagenomic } \beta\text{-diversity} = (b + c) / (2a + b + c)$$

297 Where a is the number of distinct k-mers shared between two samples, and b and c are the number  
298 of distinct k-mers specific to each sample. We represented the distance between each pair of samples  
299 on a heatmap using the heatmap.2 function of the R-package<sup>34</sup> gplots\_2.17.0<sup>35</sup>. The dissimilarity  
300 matrices we produced for each plankton size fraction (on a scale of 0 = identical to 100 = completely  
301 dissimilar) are available as Supplementary Tables 3-8.

302

### 303 **Calculation of OTU-based community dissimilarity**

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

308 eukaryotic size fractions, we added additional samples to a previously published *Tara* Oceans  
309 metabarcoding data set and processed them using the same methods<sup>19</sup> (also described at DOI:  
310 10.5281/zenodo.15600).

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

315

### 316 Calculating distances of environmental parameters

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

$$320 ED(i,j,P) = \sqrt{\sum_{p \in P} (x_{ip} - x_{jp})^2}$$

321

### 322 RGB encoding of environmental positions

323 We color-coded the position of stations in environmental space for Fig. 1b and Supplementary Fig. 4g  
324 as follows. First, environmental variables were power-transformed using the Box-Cox transformation  
325 to have Gaussian-like distributions to mitigate the effect of outliers and scaled to have zero mean and  
326 unit variance. We then performed a principal component analysis (PCA) with the R command prcomp  
327 from the package stats 3.2.1<sup>34</sup> on the matrix of transformed environmental variables and kept only  
328 the first 3 principal components. Finally, we rescaled the scores in each component to have unit  
329 variance and decorrelated them using the Mahalanobis transformation. Each component was mapped  
330 to a color channel (red, green or blue) and the channels were combined to attribute a single composite  
331 color to each station. The components (x, y, z) were mapped to color channel values (r, g, b) between  
332 0 and 255 as  $r = 128 * (1 + x / \max(\text{abs}(x)))$ , and similarly for g and b. This map ensures that the global  
333 dispersion is equally distributed across the three components and composite colors span the whole  
334 color space.

335

### 336 Definition of genomic provinces

337 We used a hierarchical clustering method on the metagenomic pairwise dissimilarities produced by  
338 Simka for all surface and DCM samples, and multiscale bootstrap resampling for assessing the  
339 uncertainty in hierarchical cluster analysis. We focused on metagenomic dissimilarity due to its higher  
340 resolution, and confirmed that the patterns found in metagenomic data were consistent when using  
341 OTU data (Supplementary Fig. 5). We used UPGMA (Unweighted Pair-Group Method using Arithmetic  
342 averages) clustering, as it has been shown to have the best performance to describe clustering of  
343 regions for organismal biogeography<sup>39</sup>. The R-package pvclust\_1.3-2<sup>40</sup>, with average linkage clustering  
344 and 1,000 bootstrap replications, was used to construct dendograms with the approximately  
345 unbiased p-value for each cluster (Supplementary Fig. 6). Because the number of genomic provinces  
346 by size fraction was not known *a priori*, we applied a combination of visualization and statistical  
347 methods to compare and determine the consistency within clusters of samples. First, the silhouette  
348 method<sup>41</sup> was used to measure how similar a sample was within its own cluster compared to other  
349 clusters using the R package cluster\_2.0.1<sup>42</sup>. The Silhouette Coefficient s for a single sample is given  
350 as:

$$351 s = (b - a) / \max(a, b)$$

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

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

359 In a complementary approach, we performed a principal coordinates analysis (PCoA) with the R  
360 command cmdscale (eig = TRUE, add = TRUE) from the package stats 3.2.1<sup>34</sup> on the matrices of  
361 pairwise metagenomic dissimilarities calculated by Simka (or OTU dissimilarity measured with the  
362 Jaccard index) within each size fraction and kept only the first 3 principal coordinates. We then  
363 converted those coordinates to a color using the RGB encoding described above, with one  
364 modification: scaling factors  $\lambda_r$ ,  $\lambda_g$  and  $\lambda_b$  were calculated as the ratios of the second and third  
365 eigenvalues to the first (dominant) eigenvalue to ensure that the dispersion of stations along each  
366 color channel reproduced the dispersion of the stations along the corresponding principal component  
367 (the ratio for the color corresponding to the dominant eigenvalue is 1). The components (x, y, z) were  
368 then mapped to color channel values (r, g, b) between 0 and 255 as  $r = 128 * (1 + \lambda_c x / \max(\text{abs}(x)))$ ,  
369 where  $\lambda_c$  is the ratio of the eigenvalue of color c to the dominant eigenvalue.  
370 We represented number and PCoA-RGB color of genomic provinces for each sample on a world map  
371 (Fig. 1, Supplementary Fig. 4a-f) generated with the R packages maps\_3.0.0.2<sup>44</sup>, mapproj 1.2-4<sup>45</sup>,  
372 gplots\_2.17.0<sup>35</sup> and mapplots\_1.5<sup>46</sup>. We also plotted phosphate and temperature (Supplementary Fig.  
373 4a-f) obtained from the *Csiro Atlas of Regional Seas* (CARS2009, <http://www.cmar.csiro.au/cars>) using  
374 the phosphate\_cars2009.nc and temprerature\_cars2009a.nc files and the R package RNetCDF<sup>47</sup>.  
375

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

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

386  
387 **Environmental differences among genomic provinces**  
388 For each size fraction, we tested which environmental parameters significantly discriminated among  
389 genomic provinces (Supplementary Fig. 7). A total of 12 parameters characterizing each sample,  
390 grouped by genomic provinces, were evaluated with a Kruskal-Wallis test within each size fraction  
391 with a significance threshold of  $p < 10^{-5}$ . Selected parameters for each size fraction were then used to  
392 perform a principal components analysis of the samples using the R package vegan\_1.17-11<sup>37</sup>. Samples  
393 were plotted with the same PCoA-RGB colors used in the genomic province maps above and each  
394 genomic province surrounded by a grey polygon. In analyses where Southern Ocean (including  
395 Antarctic) stations were considered independently from other stations, the following were considered  
396 Southern Ocean stations: 82, 83, 84, 85, 86, 87, 88, 89.

397  
398 **Ocean circulation simulations**  
399 We derived travel times from the MITgcm Darwin simulation<sup>50</sup> based on an optimized global ocean  
400 circulation model from the ECCO2 group<sup>51</sup>. The horizontal resolution of the model was approximately  
401 18 km, with 1,103,735 total ocean cells. We ran the model for six continuous years in order to smooth  
402 anomalies that might occur during any single year. We used surface velocity simulation data to  
403 compute trajectories of floats originating in ocean cells containing all *Tara* Oceans stations, and  
404 applied the following stitching procedure to generate a large number of trajectories for each initial  
405 position. (The use of surface velocity data implies that Ekman transport also influences trajectories  
406 within the simulation.)

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

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

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

424 We excluded stations that were not from open ocean locations from correlation analyses to avoid  
425 sites impacted by coastal processes (those numbered 54, 61, 62, 79, 113, 114, 115, 116, 117, 118, 119,  
426 120, and 121). In analyses where Southern Ocean (including Antarctic) stations were considered  
427 independently from other stations, the following were considered Southern Ocean (including  
428 Antarctic) stations: 82, 83, 84, 85, 86, 87, 88, 89. We calculated rank-based Spearman correlations  
429 between β-diversity,  $T_{\min}$  and environmental parameters (either differences in temperature or the  
430 Euclidean distance composed of differences in NO<sub>2</sub>NO<sub>3</sub>, PO<sub>4</sub> and Fe, see above) for surface samples  
431 with a Mantel test with 1,000 permutations and a nominal significance threshold of  $p < 0.01$ . For the  
432 correlations presented in Fig. 2a, Fig. 3 and Supplementary Fig. 9 correlation values were derived from  
433 pairs of stations connected by  $T_{\min}$  up to the value on the x-axis. We calculated partial correlations of  
434 metagenomic and OTU dissimilarity and  $T_{\min}$  by controlling for differences in temperature and for  
435 differences in nutrient concentrations, and partial correlations of dissimilarity with temperature or  
436 nutrient variation by controlling for  $T_{\min}$ .

#### 437 **Community turnover in the North Atlantic**

438 *Tara* Oceans stations numbered 72, 76, 142, 143, 144, and all stations from 146 to 151 were located  
439 along the main current system connecting South Atlantic and North Atlantic oceans and continuing to  
440 the strait of Gibraltar. In addition, we included stations 4, 7, 18, and 30 located on the main current  
441 system in the Mediterranean Sea (Supplementary Fig. 10). As the *Tara* Oceans samples within the  
442 subtropical gyre of the North Atlantic and in the Mediterranean Sea were all collected in winter,  
443 seasonal variations should not play a role in the variability in community composition that we  
444 observed (see Supplementary Table 2). We calculated genomic e-folding times (the time after which  
445 the detected genomic similarity between plankton communities changes by 63%) over scales from  
446 months to years based on an exponential fit of metagenomic dissimilarity to  $T_{\min}$  with the form  $y = C_0$   
447  $e^{-x/\tau}$  (where  $C_0$  is a constant and  $\tau$  the folding time). Exponential fits for size fractions 0-0.2 μm and 5-  
448 20 μm were not calculated due to an insufficient number of sampled stations in the North Atlantic  
449 (Supplementary Information 6).

450 The synthetic map (Supplementary Fig. 10a) was generated with the R packages maps\_3.0.0.2,  
451 mapproj 1.2.4, gplots\_2.17.0 and mapplots\_1.5. We derived dynamic sea surface height from the *Csiro*  
452 *Atlas of Regional Seas* (CARS2009, <http://www.cmar.csiro.au/cars>) using the hgt2000\_cars2009a.nc  
453 file and plotted with the R package RNetCDF.

456 **References**

- 457
- 458 1. Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* 4, 102–112 (2006).
- 459 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
- 460 3. Longhurst, A. *Ecological Geography of the Sea*. (Academic Press, 2006).
- 461 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).
- 462 5. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537, 689–693 (2016).
- 463 6. McGowan, J. A. & Walker, P. W. Structure in the Copepod Community of the North Pacific Central Gyre. *Ecol. Monogr.* 49, 195–226 (1979).
- 464 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
- 465 8. Villarino, E. et al. Large-scale ocean connectivity and planktonic body size. *Nat. Commun.* 9, 142 (2018).
- 466 9. Watson, J. R. et al. Realized and potential larval connectivity in the Southern California Bight. *Mar. Ecol. Prog. Ser.* 401, 31–48 (2010).
- 467 10. Moore, C. M. et al. Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.* 6, 701–710 (2013).
- 468 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).
- 469 12. Armbrust, E. V. The life of diatoms in the world's oceans. *Nature* 459, 185–192 (2009).
- 470 13. Pittman, S.J. (ed.). *Seascape Ecology*. (Wiley-Blackwell, 2017).
- 471 14. Tagliabue, A. et al. The integral role of iron in ocean biogeochemistry. *Nature* 543, 51–59 (2017).
- 472 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).
- 473 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).
- 474 17. Karsenti, E. et al. A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol.* 9, e1001177 (2011).
- 475 18. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* 348, 1261359 (2015).
- 476 19. de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348, 1261605–1261605 (2015).
- 477 20. Talley, L. D., Pickard, G. L., Emery, W. J. & Swift, J. H. *Descriptive Physical Oceanography: An Introduction*. (Elsevier, 2011).
- 478 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).
- 479 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
- 480 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).
- 481 24. Jönsson, B. F. & Watson, J. R. The timescales of global surface-ocean connectivity. *Nat. Commun.* 7, 11239 (2016).
- 482 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).
- 483 26. Sarmiento, J. L. & Gruber, N. *Ocean Biogeochemical Dynamics*. (Princeton University Press, 2006).
- 484 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
- 485 28. Beaugrand, G. Reorganization of North Atlantic Marine Copepod Biodiversity and Climate. *Science* 296, 1692–1694 (2002).
- 486 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).
- 487 30. Pesant, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* 2, 150023 (2015).

- 512 31. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans  
513 expedition. *Sci. Data* 4, 170093 (2017).
- 514 32. Morel, A. et al. Examining the consistency of products derived from various ocean color sensors in  
515 open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sens. Environ.* 111, 69–88  
516 (2007).
- 517 33. Benoit, G. et al. Multiple comparative metagenomics using multiset k -mer counting. *PeerJ Comput.*  
518 *Sci.* 2, e94 (2016).
- 519 34. R Core Team, T. R: A language and environment for statistical computing. (R Foundation for Statistical  
520 Computing, 2017).
- 521 35. Warnes, G. R. et al. R package gplots: Various R Programming Tools for Plotting Data. (2015).
- 522 36. Brum, J. R. et al. Patterns and ecological drivers of ocean viral communities. *Science* 348, 1261498  
523 (2015).
- 524 37. Oksanen, J. et al. R package vegan: Community Ecology Package. (2019).
- 525 38. Legendre, P. & Legendre, L. *Numerical Ecology*. (Elsevier, 2012).
- 526 39. Kreft, H. & Jetz, W. A framework for delineating biogeographical regions based on species  
527 distributions. *J. Biogeogr.* 37, 2029–2053 (2010).
- 528 40. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical  
529 clustering. *Bioinformatics* 22, 1540–1542 (2006).
- 530 41. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J.*  
531 *Comput. Appl. Math.* 20, 53–65 (1987).
- 532 42. Maechler, M., Rousseeuw, P. J., Struyf, A., Hubert, M. & Hornik, K. R package cluster: Cluster Analysis  
533 Basics and Extensions. (2015).
- 534 43. Bostock, M., Ogievetsky, V. & Heer, J. D3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* 17,  
535 2301–2309 (2011).
- 536 44. Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P. & Deckmyn, A. R package maps: Draw  
537 Geographical Maps. (2018).
- 538 45. McIlroy, D., Brownrigg, R., Minka, T. P. & Bivand, R. R package mapproj: Map Projections. (2015).
- 539 46. Gerritsen, H. R package mapplots: Data Visualization on Maps. (2014).
- 540 47. Ridgway, K. R., Dunn, J. R. & Wilkin, J. L. Ocean Interpolation by Four-Dimensional Weighted Least  
541 Squares—Application to the Waters around Australasia. *J. Atmospheric Ocean. Technol.* 19, 1357–1375 (2002).
- 542 48. Reygondeau, G. et al. Dynamic biogeochemical provinces in the global ocean. *Glob. Biogeochem.*  
543 Cycles 27, 1046–1058 (2013).
- 544 49. Oliver, M. J. & Irwin, A. J. Objective global ocean biogeographic provinces. *Geophys. Res. Lett.* 35,  
545 L15601 (2008).
- 546 50. Clayton, S. et al. Biogeochemical versus ecological consequences of modeled ocean physics.  
547 *Biogeosciences Discuss.* 1–20 (2016). doi:10.5194/bg-2016-337
- 548 51. Menemenlis, D. et al. ECCO2: High resolution global ocean and sea ice data synthesis. *Mercat. Ocean*  
549 *Q. Newsl.* 31, 13–21 (2008).
- 550 52. Rattray, A. et al. Geographic distance, water circulation and environmental conditions shape the  
551 biodiversity of Mediterranean rocky coasts. *Mar. Ecol. Prog. Ser.* 553, 1–11 (2016).
- 552 53. Carradec, Q. et al. A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9, 373 (2018).
- 553 54. Wu, S., Xiong, J. & Yu, Y. Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass  
554 Copepoda. *PLoS ONE* 10, e0131498 (2015).
- 555 55. Vannier, T. et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. *Sci. Rep.* 6,  
556 37900 (2016).
- 557 56. Piganeau, G., Eyre-Walker, A., Grimsley, N. & Moreau, H. How and Why DNA Barcodes Underestimate  
558 the Diversity of Microbial Eukaryotes. *PLoS ONE* 6, e16342 (2011).
- 559 57. Worden, A. Z. et al. Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine  
560 Picoeukaryotes Micromonas. *Science* 324, 268–272 (2009).
- 561 58. Seeleuthner, Y. et al. Single-cell genomics of multiple uncultured stramenopiles reveals  
562 underestimated functional diversity across oceans. *Nat. Commun.* 9, 310 (2018).
- 563 59. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical  
564 clustering. *Bioinformatics* 31, 3718–3720 (2015).
- 565 60. Sokal, R. R. & Rohlf, F. J. The Comparison of Dendograms by Objective Methods. *Taxon* 11, 33–40  
566 (1962).
- 567 61. Sneath, P. H. A. & Sokal, R. R. *Numerical taxonomy. The principles and practice of numerical*  
568 *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. Madou, 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

622 **Author Contributions**

623

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

635

636

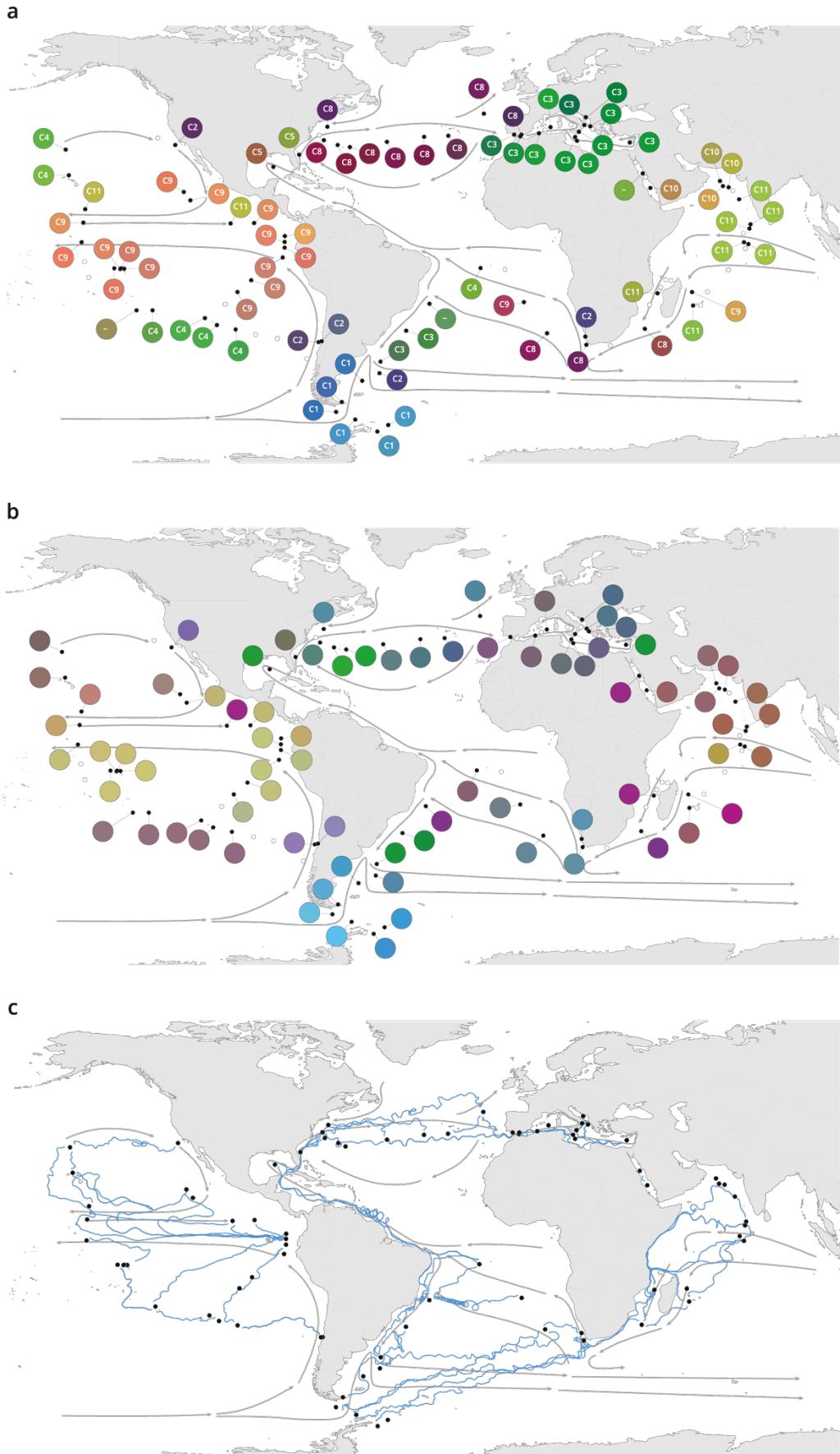
637 **Author Information**

638

639 The authors declare that all data reported herein are fully and freely available from the date of  
640 publication, with no restrictions, and that all of the samples, analyses, publications, and ownership  
641 of data are free from legal entanglement or restriction of any sort by the various nations in whose  
642 waters the *Tara* Oceans expedition sampled. Metagenomic and metabarcoding sequencing reads  
643 have been deposited at the European Nucleotide Archive under accession numbers provided in  
644 Supplementary Table 1. Contextual metadata of *Tara* Oceans stations are available in Supplementary  
645 Table 2. Metagenomic dissimilarity, OTU community dissimilarity, simulated travel times and  
646 geographic distances are provided in Supplementary Tables 3-16. All Supplementary Tables, in  
647 addition to tables of 18S V9 barcodes and OTUs and the V9 reference database are available on  
648 FigShare at the following URL: <http://doi.org/10.6084/m9.figshare.11303177>

649 The authors declare no competing financial interests.

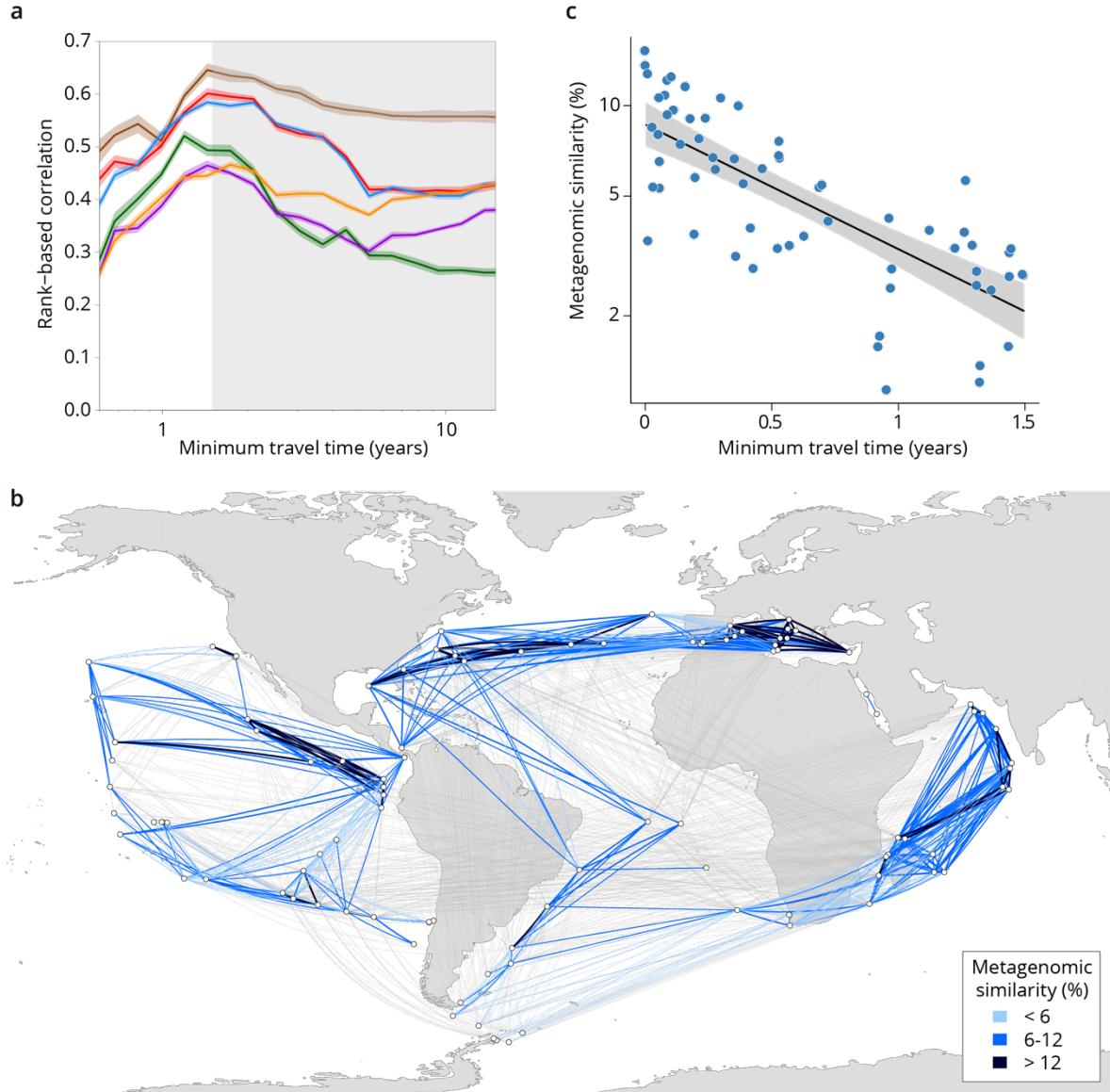
650 Correspondence and requests for materials should be addressed to Olivier Jaillon, Daniele Iudicone,  
651 Maurizio Ribero d'Alcalà, Colomban de Vargas.



652  
653  
654

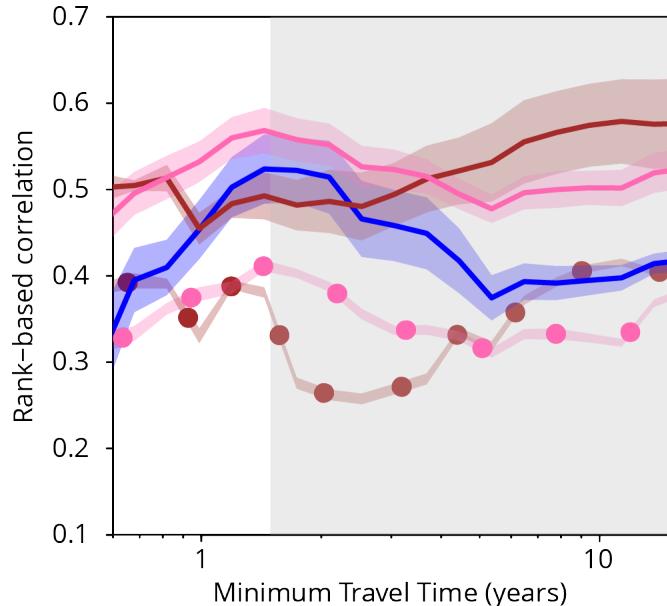
**Figure 1 | Plankton biogeography, environmental variation and ocean transport among *Tara Oceans* stations.** Major currents are represented by solid arrows. **a**, Genomic provinces of *Tara Oceans* surface

655 samples for the 0.8-5  $\mu\text{m}$  size fraction, each labeled with a letter prefix ('C' represents the 0.8-5  $\mu\text{m}$  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  $\text{NO}_2\text{NO}_3$  to  $\text{PO}_4$  (replaced  
660 by  $10^{-6}$  for 3 stations where the measurement of  $\text{PO}_4$  was 0) and of  $\text{NO}_2\text{NO}_3$  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.



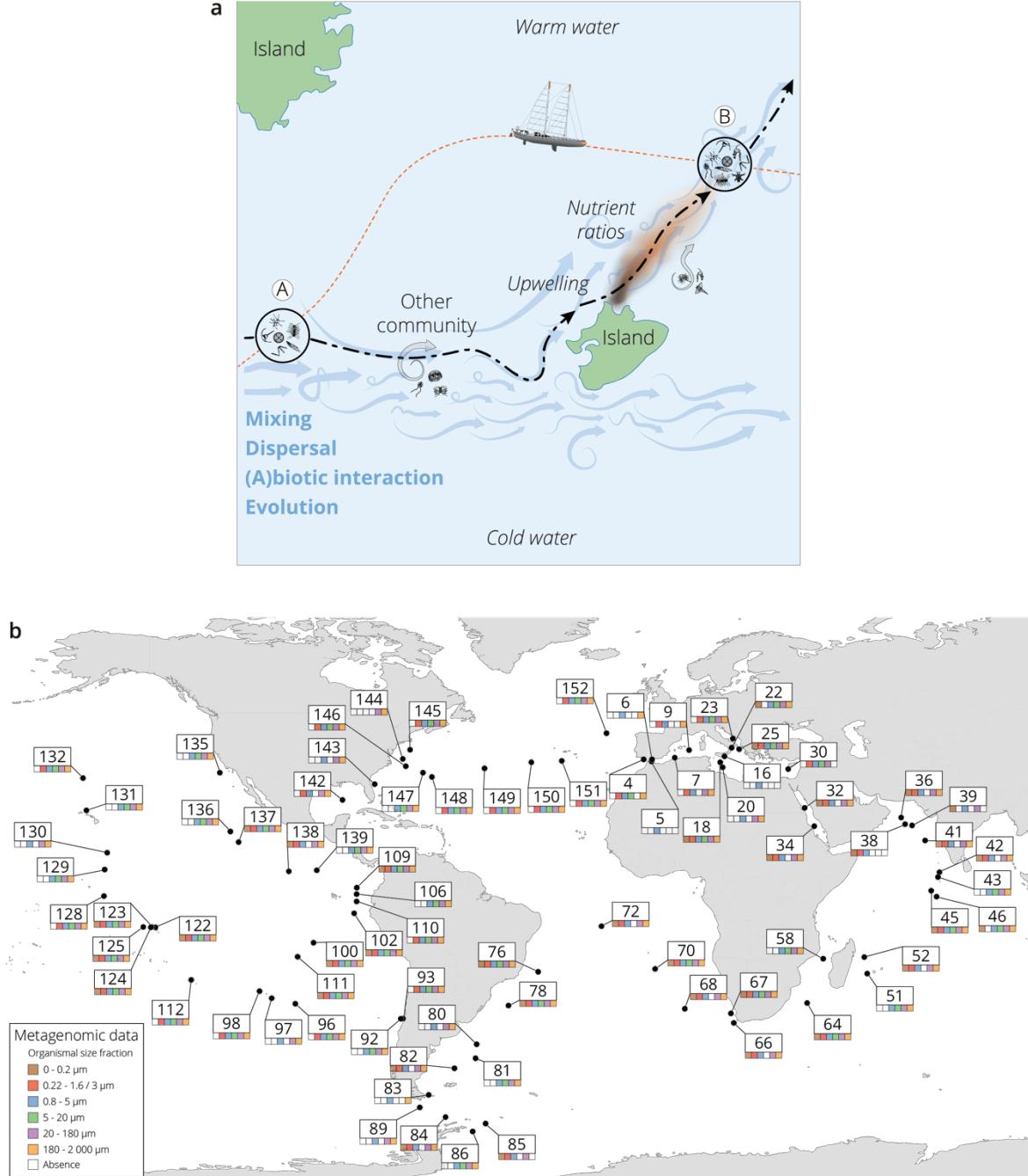
664  
665 **Figure 2 | Metagenomic dissimilarity and travel time of plankton are maximally correlated up to ~1.5 years.**  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675

**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: 0.2-1.6/3 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).

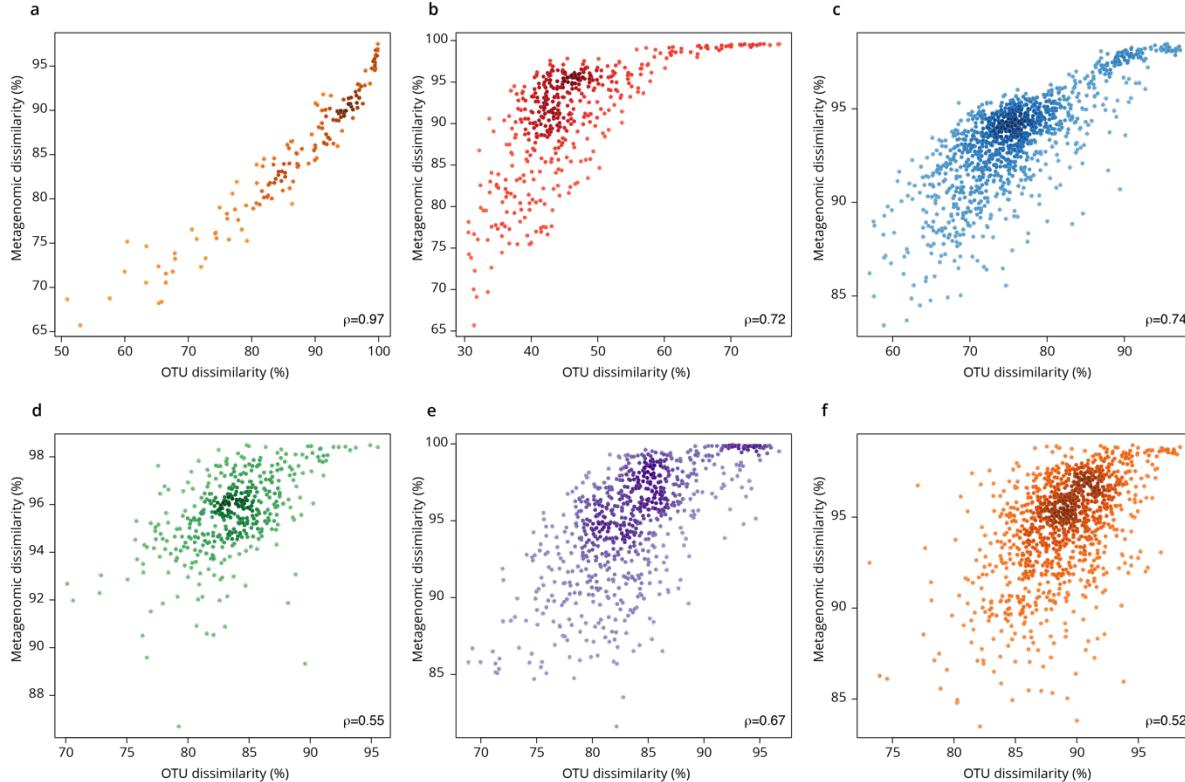


676  
677  
678  
679  
680  
681  
682  
683  
684  
685

**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  $\text{NO}_2\text{NO}_3$ ,  $\text{PO}_4$  and Fe (pink), metagenomic dissimilarity and differences in temperature (red),  $T_{\min}$  and differences in  $\text{NO}_2\text{NO}_3$ ,  $\text{PO}_4$  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  $\mu\text{m}$  and 5-20  $\mu\text{m}$  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.

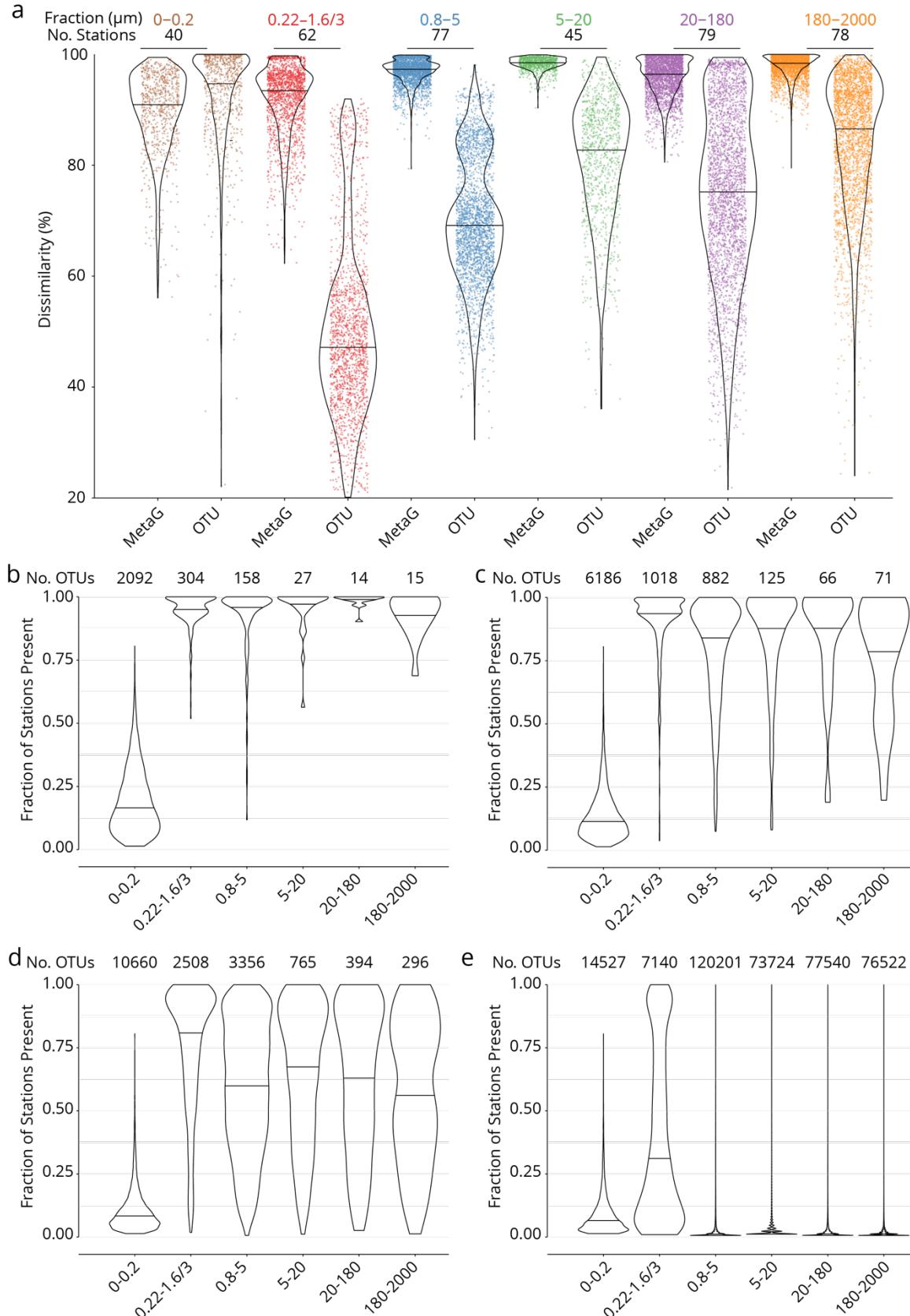


686  
687 **Supplementary Figure 1 | The seascape, plankton transport and community metagenomic samples of Tara**  
688 **Oceans stations. a,** A community sampled at a given location (A) changes over time as it travels along ocean  
689 currents (dashed bold line) to a second location (B). It is affected by numerous external processes, including  
690 mixing with water containing other communities and changes in local nutrient concentration, and by internal  
691 processes, such as biotic interactions. In this study, the *Tara* schooner followed a sampling route (orange  
692 dashed line) leading to an elapsed time between the 2 sampling sites A and B that was independent of  
693 plankton travel time. **b,** Location, station number, and sequenced surface metagenomic samples.



694  
695  
696  
697  
698  
699  
700

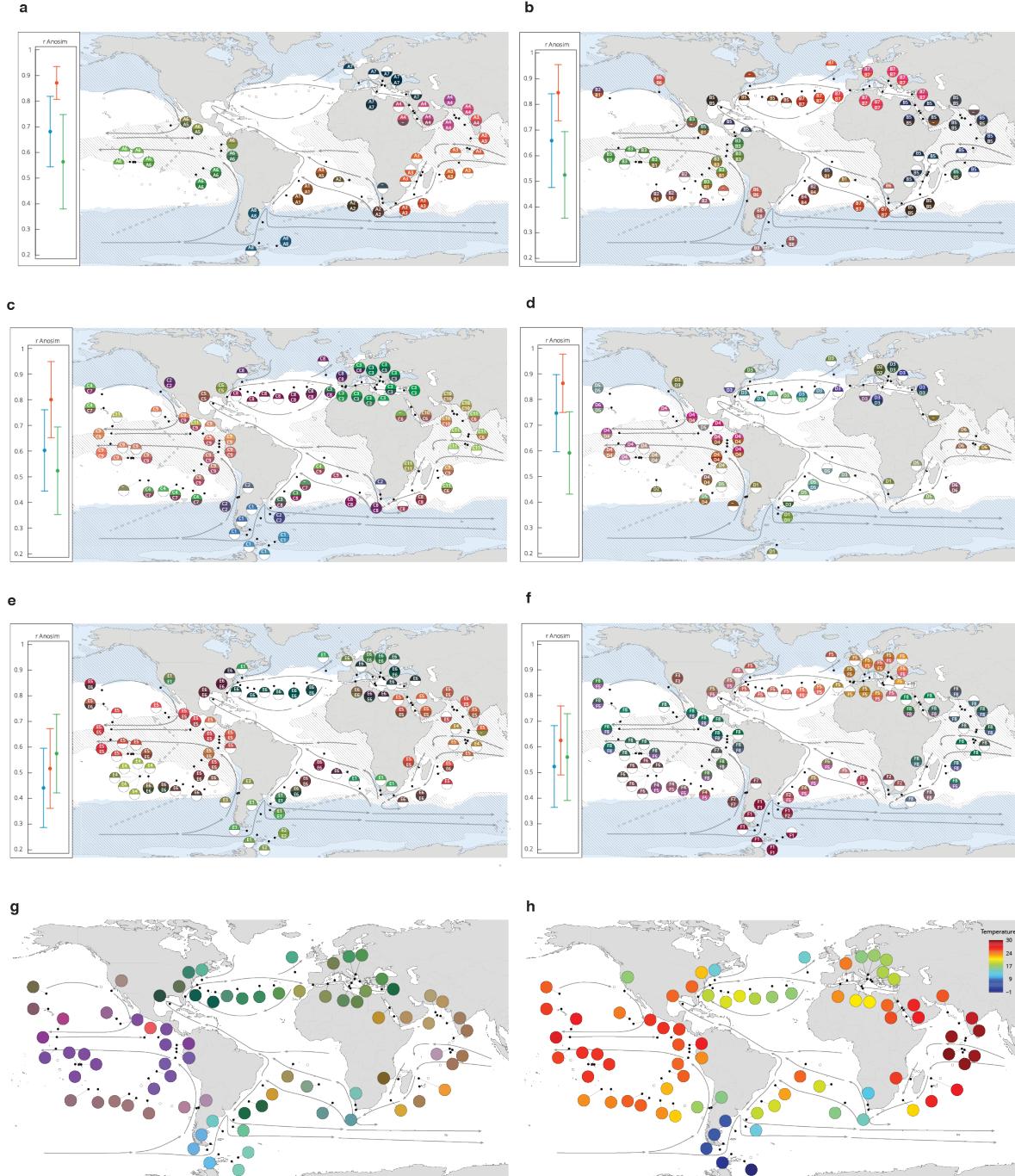
**Supplementary Figure 2 |  $\beta$ -diversity estimates from metagenomic and OTU-based dissimilarity are correlated.** Scatter plots of metagenomic dissimilarity versus OTU community dissimilarity for six organismal size fractions. Each point represents a pairwise comparison between two samples. **a**,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



701  
702  
703  
704  
705

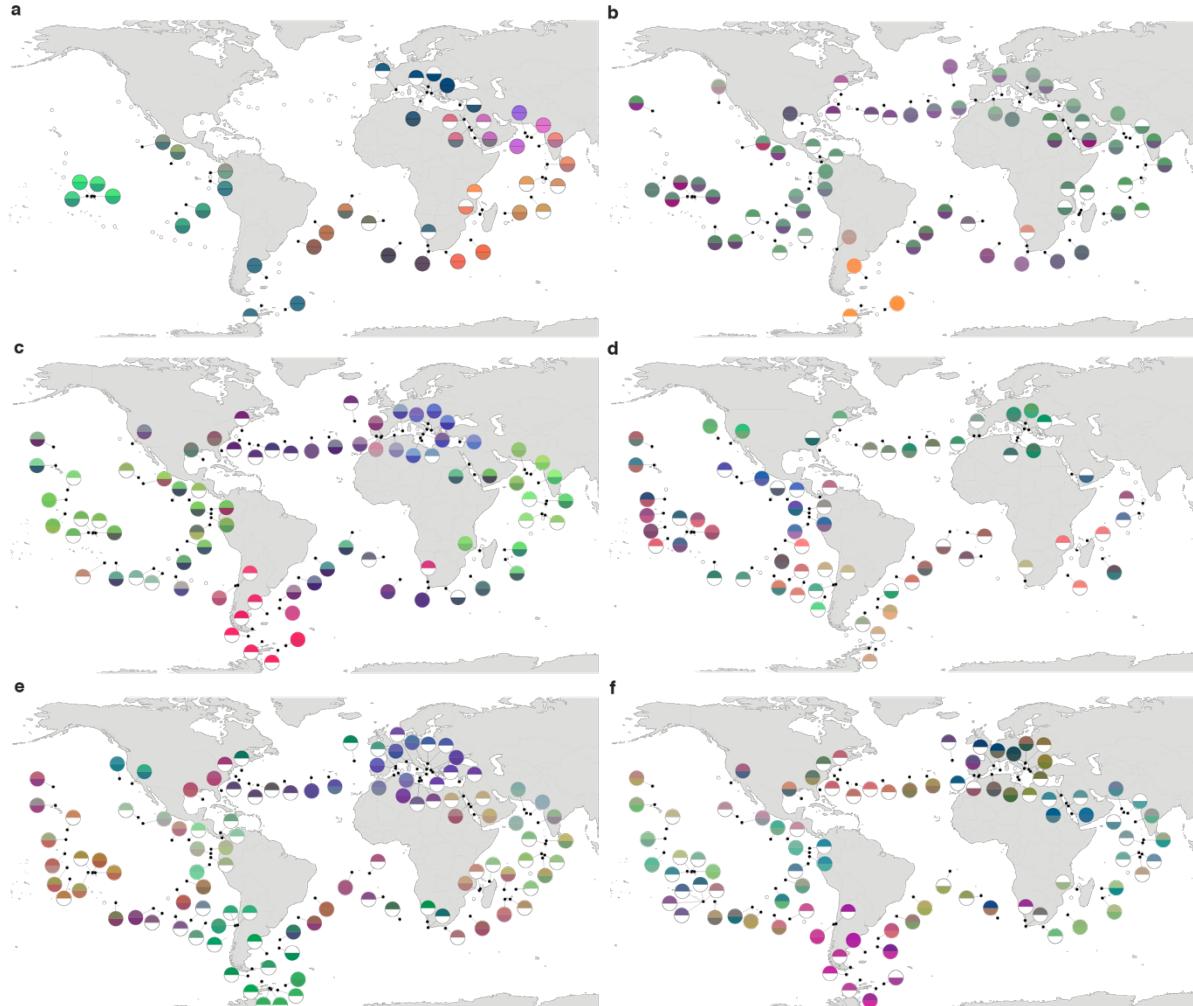
**Supplementary Figure 3 | Global dissimilarity and OTU occupancy.** **a**, Dissimilarity distributions for six organismal size fractions (measured either as metagenomic or OTU dissimilarity; see Supplementary Information 1). One colored point represents one pair of stations. Violin plots (horizontal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

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



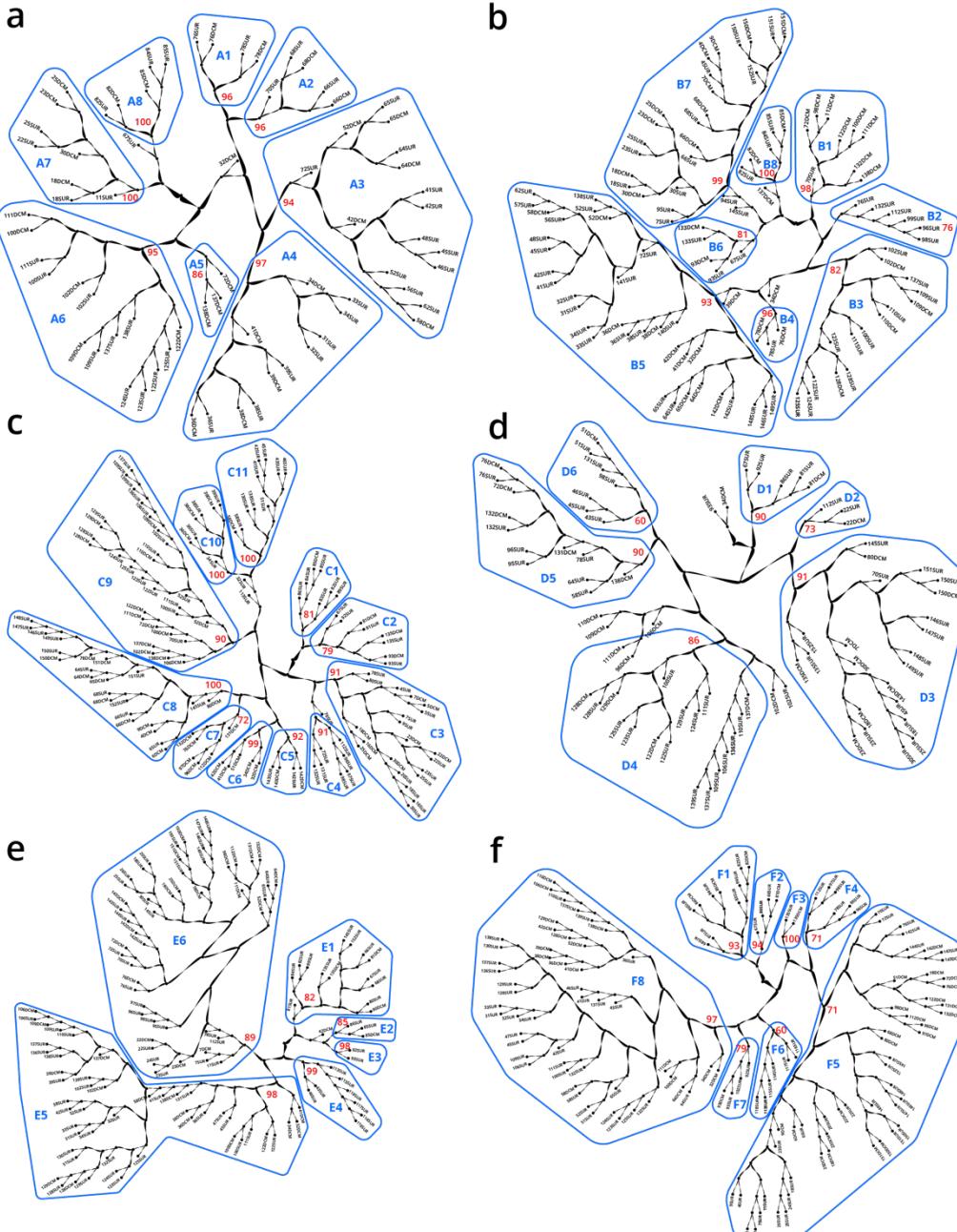
712  
 713 **Supplementary Figure 4 | Genomic provinces in comparison to previous ocean divisions, and ordination**  
 714 **maps of environmental parameters.** a-f, Geographical maps of genomic provinces by organismal size fraction  
 715 (see Supplementary Information 2). Circles denote stations with data available for the size fraction and contain  
 716 the corresponding genomic province identifiers (one letter prefix per size fraction (A-F); stations not assigned  
 717 to genomic provinces are shown as '-'). The top portion of each circle represents samples collected at the  
 718 surface and the bottom portion represents the deep chlorophyll maximum (stations missing metagenomic  
 719 data for one of the two depths are drawn as half circles). Colors are based on PCoA-RGB (Methods) and do not  
 720 correspond among size fractions. Major currents are shown with solid black arrows, wind transport with  
 721 dashed grey arrows. Blue zones indicate temperature < 14 °C. Hashed zones indicate phosphate concentration  
 722 > 0.4 mmol. Hierarchical dendograms that were used to build genomic provinces are shown in Supplementary  
 723 Fig. 6. Maps with colors based on OTU dissimilarity are shown in Supplementary Fig. 5. a, 'A' prefix, 0-0.2 µm  
 724 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.  
 725 f, 'F' prefix, 180-2000. **Insets**, Results of ANOSIM to determine, independently for each size fraction, the ability  
 726 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<sub>2</sub>NO<sub>3</sub>, PO<sub>4</sub> 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).

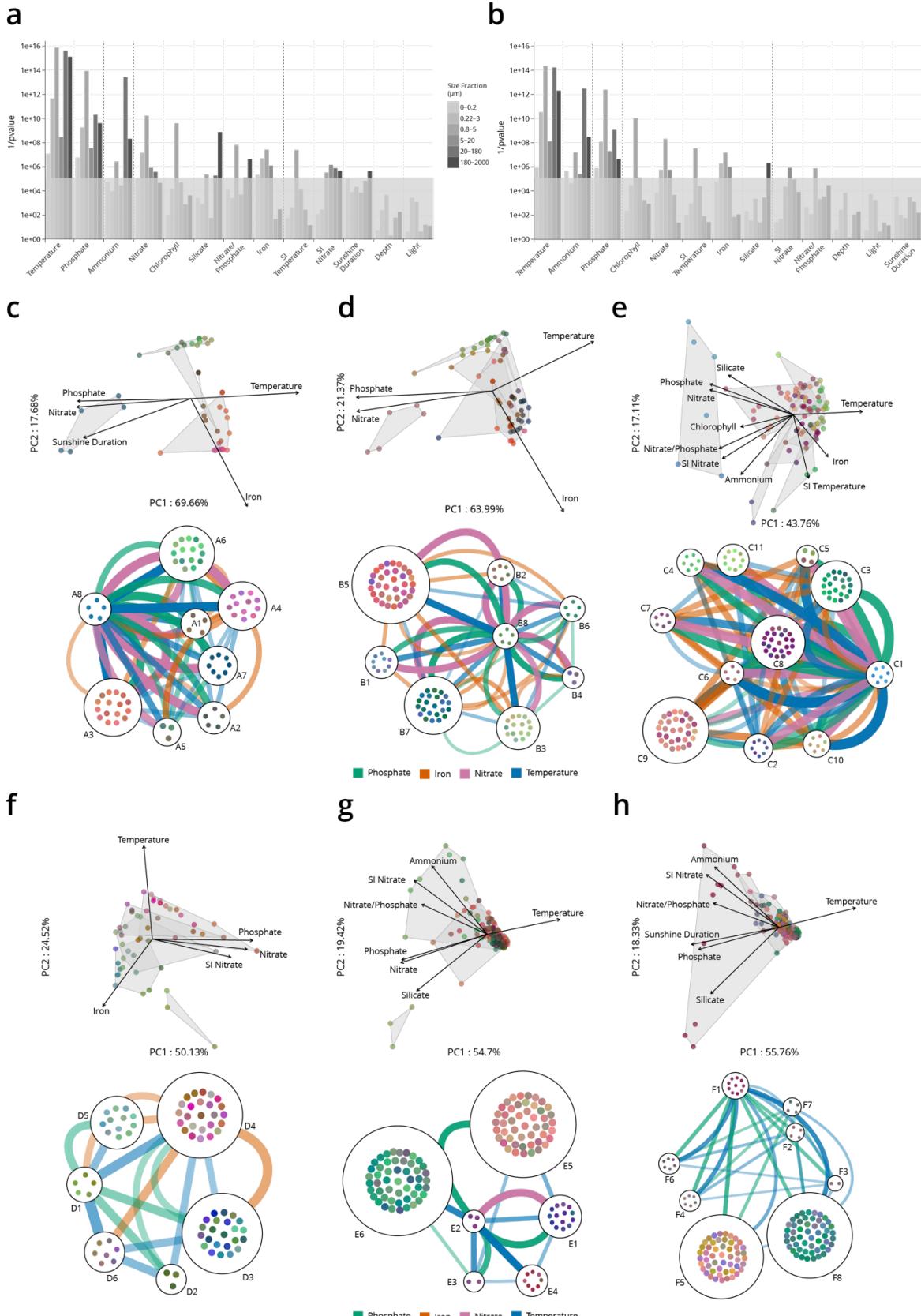


732  
733  
734  
735  
736  
737  
738

**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  $\mu\text{m}$  size fraction. **b**, 0.22-1.6/3  $\mu\text{m}$ . **c**, 0.8-5  $\mu\text{m}$ . **d**, 5-20  $\mu\text{m}$ . **e**, 20-180  $\mu\text{m}$ . **f**, 180-2000  $\mu\text{m}$ .



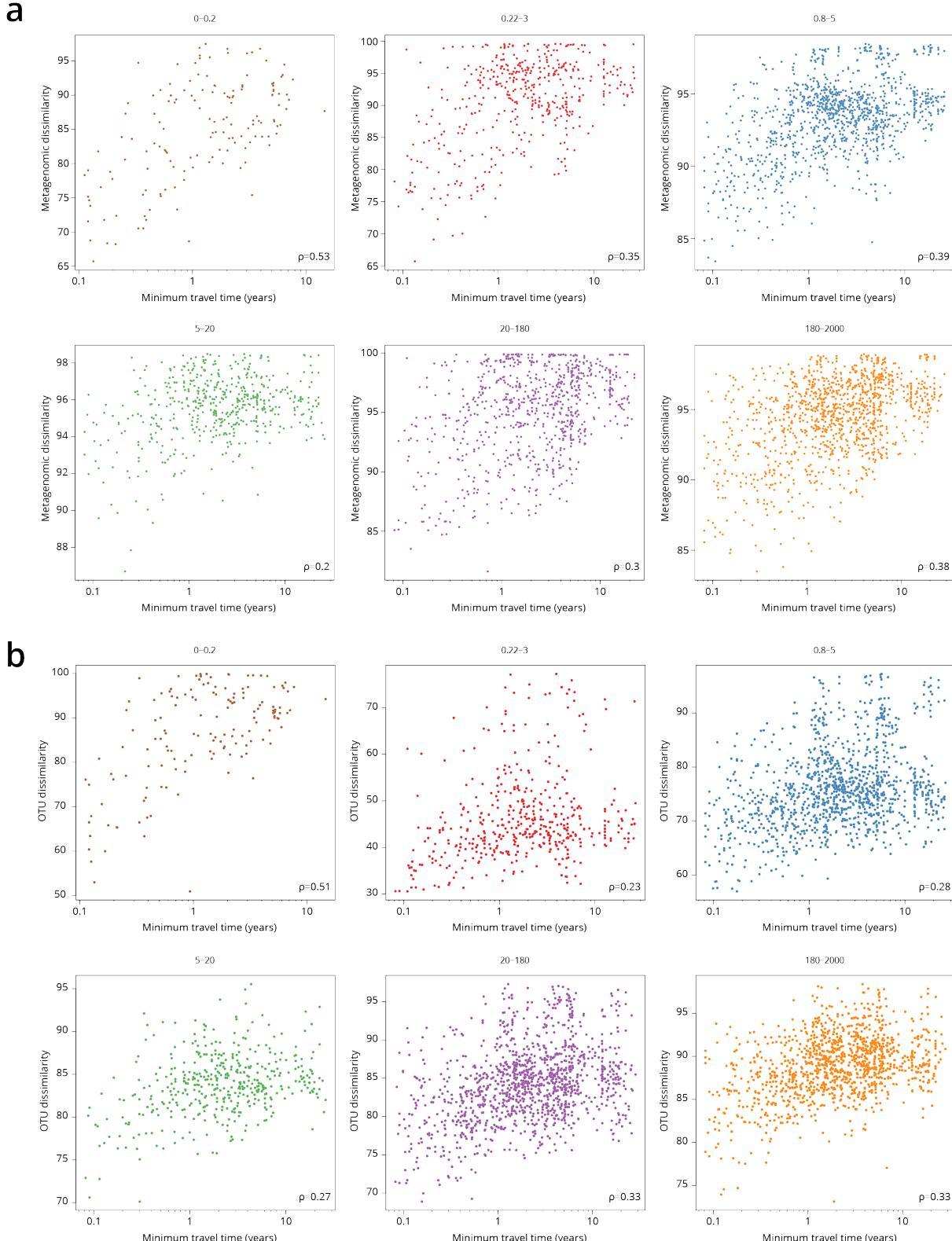
739  
740 **Supplementary Figure 6 | Hierarchical trees illustrating how samples were partitioned into genomic**  
741 **provinces.** Dendograms resulted from UPGMA clustering. Each sample (SUR: surface, DCM: deep chlorophyll  
742 maximum) is shown as a leaf. Genomic provinces are shown with their identifiers in blue polygons; identifiers  
743 are composed of one letter prefix per size fraction (A-F) and a number. Bootstrap values in red show the  
744 support at the key nodes that separate genomic provinces from one another. See also Supplementary  
745 Information 2 on the robustness of genomic provinces. **a**, 'A' prefix, 0-0.2 μm size fraction. **b**, 'B' prefix, 0.22-  
746 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.



747  
748  
749  
750  
751

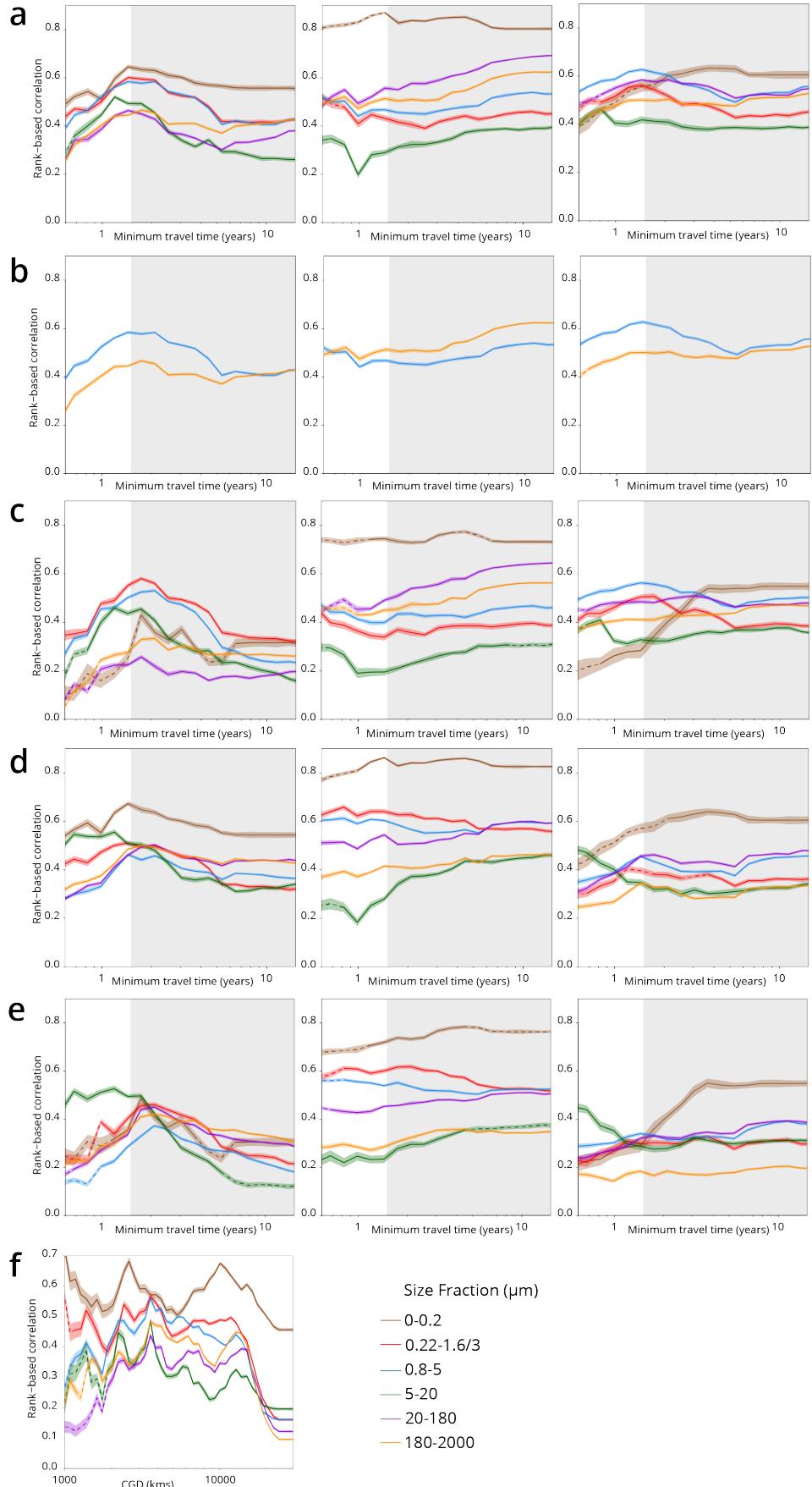
**Supplementary Figure 7 | Environmental parameters that distinguish genomic provinces.** **a-b**, Environmental parameters that significantly differentiate among genomic provinces (Kruskal-Wallis test, grey box indicates  $p$  values  $> 10^{-5}$ ). SI = Seasonality Index. **a**, all stations. **b**,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

752 provinces. **c-h**, Two types of visualizations of the relationships between genomic provinces and environmental  
753 parameters. Sample colors are those from Supplementary Fig. 4. **Top plots within panels c-h**: principal  
754 components analysis-based visualization. Samples, and environmental parameters differing significantly ( $p \leq$   
755  $10^{-5}$ ) among genomic provinces, are projected onto the first two axes of variation. Grey polygons enclose  
756 different genomic provinces. **Bottom plots within panels c-h**: network-based visualization. Each genomic  
757 province is represented as a node, with the individual samples composing the province within the node. Edges  
758 between nodes represent differences in temperature, nitrate, phosphate and iron that significantly  
759 differentiate ( $p \leq 10^{-5}$ ) among genomic provinces, that are statistically significantly different between  
760 individual pairs of genomic provinces (*post hoc* Tukey test,  $p < 0.01$ ) and whose difference in median  
761 parameter values is  $\geq 1$  standard deviation (calculated from the parameter values of all samples in the size  
762 fraction). Thicker edges represent larger differences. **c**, 0-0.2  $\mu\text{m}$  size fraction. **d**, 0.22-1.6/3  $\mu\text{m}$ . **e**, 0.8-5  $\mu\text{m}$ . **f**,  
763 5-20  $\mu\text{m}$ . **g**, 20-180  $\mu\text{m}$ . **h**, 180-2000  $\mu\text{m}$ .



764  
765  
766  
767

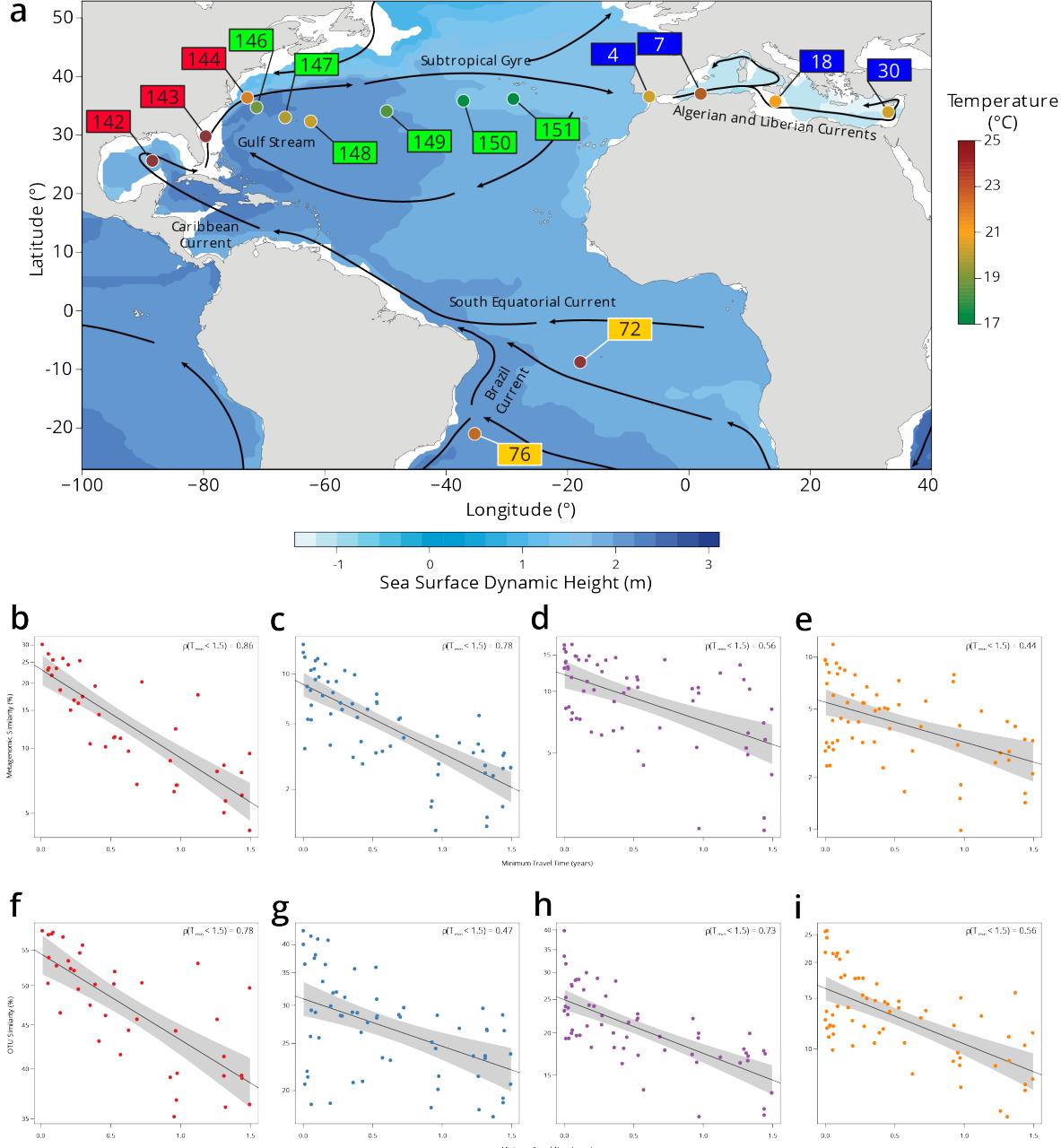
**Supplementary Figure 8 | Global correlations of dissimilarity with minimum travel time ( $T_{\min}$ ).** Scatter plots of dissimilarity versus  $T_{\min}$ . One point represents a pair of samples. **a**, metagenomic dissimilarity. **b**, OTU dissimilarity. Global Spearman correlation values are indicated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



768  
769  
770

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

771 separately by organismal size fraction. Non-significant correlations ( $p > 0.01$ ) are shown with dashed lines. **a-e**,  
772 Correlations for pairs of *Tara Oceans* samples separated by a minimum travel time less than the value of  $T_{\min}$   
773 on the x axis.  $T_{\min} > 1.5$  years is shaded in grey. Left panels: correlation of dissimilarity with  $T_{\min}$ ; middle panels,  
774 dissimilarity with temperature; right panels: dissimilarity with differences in  $\text{NO}_2\text{NO}_3$ ,  $\text{PO}_4$  and Fe. **a-c**,  
775 metagenomic dissimilarity. **d-e**, OTU dissimilarity. There is a maximum correlation of dissimilarity with  $T_{\min}$   
776 (and, for most size fractions, of dissimilarity with nutrients) for  $T_{\min} < \sim 1.5$  years, but the correlation between  
777 dissimilarity and temperature does not display a similar maximum. **b** displays only the 0.8-5  $\mu\text{m}$  (blue) and 180-  
778 2000  $\mu\text{m}$  (orange) size fractions from **a**, to highlight that for smaller plankton, correlations with differences in  
779 nutrient concentrations were stronger for  $T_{\min}$  up to  $\sim 1.5$  years, but for larger plankton, correlations were  
780 stronger with temperature variations for  $T_{\min}$  beyond  $\sim 1.5$  years. **c** and **e**, Partial correlations to estimate the  
781 independent effects of  $T_{\min}$  and environmental distances on  $\beta$ -diversity. Left panels: controlling for differences  
782 in temperature and for differences in  $\text{NO}_2\text{NO}_3$ ,  $\text{PO}_4$  and Fe; middle and right panels: controlling for  $T_{\min}$ . Partial  
783 correlations do not affect the maximum correlation of dissimilarity with  $T_{\min}$  for  $T_{\min} < \sim 1.5$  years. **f**, Correlation  
784 of geographic distance (without traversing land) with metagenomic dissimilarity for pairs of *Tara Oceans*  
785 samples separated by a geographic distance less than the value on the x axis.



786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799

**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  $\tau = 0.91$  y for 0.22-3 µm,  $\tau = 0.91$  y for 0.8-5 µm,  $\tau = 2.22$  y for 20-180 µm and  $\tau = 1.99$  y for 180-2000 µm. Turnover times using the OTU community similarity are  $\tau = 4.23$  y for 0.22-3 µm,  $\tau = 4.08$  y for 0.8-5 µm,  $\tau = 2.6$  y for 20-180 µm and  $\tau = 2.1$  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.

800 **Supplementary Information**

801

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

803

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

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

832

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

834

835 We assessed the robustness of genomic provinces in five separate ways. First, we tested 5 different  
836 hierarchical clustering algorithms from R-package *pvclust\_1.3-2*<sup>40</sup> (UPGMA - Unweighted Pair Group  
837 Method with Arithmetic mean; McQuitty's method; Complete linkage; Ward's method; Single linkage)  
838 on the metagenomic pairwise dissimilarities produced by Simka separately for the six organismal size  
839 fractions, followed by multiscale bootstrap resampling. We used the cophenetic correlation  
840 coefficient from the R-package *dendextend\_1.5.2*<sup>59</sup> to measure how accurately the dendograms  
841 produced by each method preserved the pairwise distances within the input dissimilarity matrices<sup>60,61</sup>.  
842 The ranking of the cophenetic correlation coefficient for different clustering methods within each size  
843 fraction was consistent with a published large-scale methodological comparison of clustering methods  
844 for biogeography (Supplementary Table 17), which considered UPGMA agglomerative hierarchical  
845 clustering to have consistently the best performance<sup>39</sup>. Second, we compared clustering results among  
846 all size fractions using Baker's Gamma Index<sup>62</sup> from the R-package *corrplot\_0.77*<sup>63</sup>, which is a measure  
847 of association (similarity) between two trees based on hierarchical clustering (dendograms). The  
848 Baker's Gamma Index is defined as the rank correlation between the stages at which pairs of objects  
849 combine in each of the two trees. For each type of correlation, the UPGMA was consistently the most  
850 correlated with other clustering methods (Supplementary Table 18). This allowed us to conclude, in

851 agreement with previous results<sup>39</sup>, that the UPGMA method is likely more robust than the other  
852 methods we tested.

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

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

884

#### 885 ***Supplementary Information 3. Comparison of genomic provinces to previous biogeographical 886 divisions***

887

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

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

904 We tested whether genomic provinces were comparable with these partitionings by performing an  
905 analysis of similarity (ANOSIM; Supplementary Fig. 4, insets; Methods). The four small size classes, 0-  
906 0.2  $\mu\text{m}$ , 0.22-1.6/3  $\mu\text{m}$ , 0.8-5  $\mu\text{m}$ , and 5-20  $\mu\text{m}$  (Supplementary Fig. 4a-d) were more consistent with  
907 Longhurst BGCPs. In contrast, for the two larger size fractions 20-180  $\mu\text{m}$  and 180-2000  $\mu\text{m}$ , the three  
908 biogeographical divisions were not strongly different within the ANOSIM (Supplementary Fig. 4e-f).

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

914

#### 915 **Supplementary Information 4. Differences in genomic province sizes among organismal size** 916 **fractions**

917

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

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

953 conditions. Therefore, they should be affected primarily by large scale, mostly latitudinal, variations  
954 in the environment, leading to larger genomic provinces, whereas smaller plankton are grouped into  
955 smaller provinces more influenced by local environmental conditions. Overall, this difference in  
956 biogeography suggests a size-based decoupling between smaller and larger plankton (which may also  
957 extend to nekton such as tuna and billfish<sup>68</sup>), with implications for the structure and function of  
958 oceanic food webs and other types of biotic interactions.  
959

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

961  
962 As plankton communities are transported by ocean currents, they change over time due to the  
963 various processes that occur in the context of the seascape: variations in temperature, light and  
964 nutrients (where changes in the latter may also be induced by plankton communities), intra- and inter-  
965 individual and species biological interactions, and mixing with neighboring water masses. Thus, a  
966 continuum of composition among nearby samples is expected as a natural consequence of community  
967 turnover within the seascape over time. We observed the effects of continuous turnover in our  
968 biogeographical analyses (Fig. 1a, Supplementary Fig. 4, Supplementary Fig. 5, Supplementary  
969 Information 2) in which nearby samples often reflected gradual, but not complete changes in  
970 community composition.

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

977 The global ocean current system is composed of a series of large-scale main currents and associated  
978 recirculations (which are also referred to as gyres). Therefore, we present the following hypothesis as  
979 a potential explanation of our results: the average global timescale of 1.5 years is comparable to the  
980 crossing time of an ocean gyre (i.e., the amount of time it takes a water parcel to travel from one side  
981 of a gyre to the other), e.g., to cross the North Atlantic basin while riding the Gulf Stream system. This  
982 time scale of 1.5 years is probably an underestimate, since our sparse sampling did not cover all  
983 current systems. Within different systems, the transport by main currents leads to stable, continuous  
984 patterns of changes in community structure and nutrient concentrations, and also explains how  
985 temporally stable genomic provinces can exist in the face of ocean circulation. Within each system we  
986 have thus to expect that a community turnover is long enough to allow for this long range  
987 predictability due to smooth, continuous changes. Significant heterogeneity in environmental  
988 conditions among different circulation patterns means that moving from system to another (and  
989 therefore, in our case here, beyond the 1.5 year timescale; Supplementary Fig. 9c-f) disrupts the  
990 interlinked relationship among local seascape processes, leading to a global delimitation into separate  
991 ecological continua among different gyre-scale current systems.  
992

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

994  
995 In order to characterize the impact of physical and biological processes on changes in metagenomic  
996 composition during travel along currents, we focused on the well-known current systems crossing the  
997 North Atlantic into the Mediterranean Sea (the Gulf Stream and other currents around the subtropical  
998 gyre<sup>20,69–71</sup>; Supplementary Fig. 10a). Across this region, the piconanoplankton (0.8-5 µm) were split  
999 into three genomic provinces, C5, C8 and C3, each less than 5,000 km wide (~11 months of travel time;  
1000 Supplementary Fig. 4c). In contrast, mesoplankton (180-2000 µm) biogeography corresponded to a  
1001 single province, F5, spanning from the Caribbean to Cyprus (> 9,700 km or ~18 months of travel time;  
1002 Supplementary Fig. 4f; see also Supplementary Information 4). Metagenomic dissimilarity and  $T_{\min}$   
1003 were strongly correlated within the region (Spearman's  $\rho$  between 0.44 and 0.86 depending on size

1004 fraction, Supplementary Fig. 10b-e), which allowed us to explore the relationship of genomic province  
1005 size, ocean transport and plankton community turnover over scales from months to years. We  
1006 calculated metagenomic turnover times as e-folding times based on an exponential fit of  
1007 metagenomic dissimilarity to  $T_{\min}$  (ranging from a few months to a few years, Methods). The  
1008 metagenomic turnover time of smaller plankton (< 20  $\mu\text{m}$ ) was approximately one year. In contrast,  
1009 for the larger size fractions, the metagenomic turnover time was approximately two years, suggesting  
1010 that a lower turnover rate for larger plankton may explain their geographically larger genomic  
1011 provinces.

1012 We note that our results on metagenomic turnover time appear different from a recently published  
1013 study that also calculated turnover rates for plankton, which found faster rates for larger organisms<sup>8</sup>.  
1014 This may be explained by two significant differences between our approach and theirs: first, their  
1015 measurements of  $\beta$ -diversity were based on presence/absence (Jaccard) comparisons among either  
1016 morphological species or OTUs, whereas our calculations of turnover time above were based on  
1017 metagenomic sequences. As described above (Supplementary Information 1), there are significant  
1018 differences in resolution between OTU-based and metagenomic data, and we would expect similar  
1019 differences in resolution between organismal observation data and metagenomic sequences. In fact,  
1020 due to these differences in resolution, our estimates of metagenomic time based on OTU rather than  
1021 metagenomic data show a similar trend to those of Villarino *et al.*<sup>8</sup> (Supplementary Fig. 10f-i). Second,  
1022 their turnover rates were calculated separately for individual plankton groups (the 9 main groups were  
1023 prokaryotes, coccolithophores, dinoflagellates, diatoms, all microbial eukaryotes, gelatinous  
1024 zooplankton, mesozooplankton, macrozooplankton and myctophids), whereas our metagenomic data  
1025 represent samples of the full plankton community within each size fraction. Among these, several  
1026 groups (e.g., dinoflagellates or mesozooplankton) would be expected to be found across multiple *Tara*  
1027 Oceans size fractions, blurring potential comparisons. Thus, our study and Villarino *et al.* calculated  
1028 rates of change using broadly similar approaches, but based on very different underlying biological  
1029 substrates.