Restructuring of plankton genomic biogeography in the surface ocean under climate change

Paul Frémont, Marion Gehlen, Mathieu Vrac, Jade Leconte, Tom O. Delmont, Patrick Wincker, Daniele Iudicone, Olivier Jaillon

nature climate change

ARTICLES

https://doi.org/10.1038/s41558-022-01314-8

Restructuring of plankton genomic biogeography in the surface ocean under climate change

Paul Frémont[1,2]✉, Marion Gehlen[3]✉, Mathieu Vrac[3], Jade Leconte[1,2], Tom O. Delmont[1,2], Patrick Wincker[1,2], Daniele Iudicone[4] and Olivier Jaillon[1,2]✉

The impact of climate change on diversity, functioning and biogeography of marine plankton remains a major unresolved issue. Here environmental niches are evidenced for plankton communities at the genomic scale for six size fractions from viruses to meso-zooplankton. The spatial extrapolation of these niches portrays ocean partitionings south of 60° N into *climato-genomic* provinces characterized by signature genomes. By 2090, under the RCP8.5 future climate scenario, provinces are reorganized over half of the ocean area considered, and almost all provinces are displaced poleward. Particularly, tropical provinces expand at the expense of temperate ones. Sea surface temperature is identified as the main driver of changes (50%), followed by phosphate (11%) and salinity (10%). Compositional shifts among key planktonic groups suggest impacts on the nitrogen and carbon cycles. Provinces are linked to estimates of carbon export fluxes which are projected to decrease, on average, by 4% in response to biogeographical restructuring.

Planktonic communities are composed of complex and heterogeneous assemblages of small animals, single-celled eukaryotes (protists), bacteria, archaea and viruses. They contribute to the regulation of the Earth system through primary production via photosynthesis[1] and carbon export to the deep ocean[2,3] and are at the base of the food webs that sustain the whole trophic chain in the oceans[4].

The composition of planktonic communities varies over time at a given site with daily[5] to seasonal fluctuations[6] following environmental variability[7] (for example, temperature, nutrients). Overlying these relatively short-scale spatiotemporal variations, a more macroscale partitioning of the ocean has been revealed by different combinations of biological and physicochemical data[8–10] and recently documented at the resolution of community genomics[11]. The basin-scale biogeographical structure has been proposed to result from a combination of multiple biophysicochemical processes named the seascape[7]. These processes include both abiotic and biotic interactions[12], neutral genetic drift[13], natural selection[14,15], temperature variations, nutrient supply but also advection and mixing along currents[11,13].

Today, knowledge of global-scale plankton biogeography at the DNA level is in its infancy. We lack understanding and theoretical explanations for the emergence and maintenance of biogeographical patterns at genomic resolution. Omics data (that is, the DNA/RNA sequences representative of the variety of coding and non-coding sequences of organisms) provide the appropriate resolution to track and record global biogeographical features[11], modulation of the repertoire of expressed genes in a community in response to environmental conditions[2,16,17] and eco-evolutionary processes[13–15]. Moreover, as recently demonstrated by global expeditions[2,18–21], metagenomic sequencing can be consistently analysed across plankton organisms. The strong links between plankton and environmental conditions suggest potentially major consequences of climate change on community composition and biogeography[22,23]. Time series observations have highlighted recent changes in the planktonic ecosystem such as changes in community composition[24] or poleward shifts of some species[25], mainly attributed to temperature increase caused by anthropogenic greenhouse gas emissions. These changes are expected to intensify with ongoing global warming[26] and could lead to major reorganization of plankton community composition[22] with a potential decline in diversity[27–29]. Another major consequence of global reorganization of the seascape on biological systems would be a decrease of primary production at mid-latitudes and an increase at higher latitudes[26].

Here we report the global structure of plankton biogeography south of 60° N based on metagenomic data using niche models and its putative modifications under climate change. First, we define environmental niches[30], that is, the envelope of environmental parameters suitable for an organism or a population at the scale of genomic provinces across size organism size fractions representing major plankton groups from nano- (viruses) to meso-zooplankton (small metazoans). Then, we spatially extrapolate their niches into *climato-genomic* provinces to derive the structure of plankton biogeography for each size fraction individually and for all combined. Next, considering the same niches, we assess the spatial reorganization of these provinces under climate change by the end of the century. We study compositional changes among planktonic groups known to be critical for two major biogeochemical cycles: nitrogen with nitrogen-fixing bacteria and carbon with phototrophs and copepods. By correlating our provinces with carbon export fluxes, these compositional changes would lead to a reduction of approximately 4% of carbon export. Finally, we quantify the relative importance of the environmental drivers explaining projected changes.

Niche models and signature genomes from genomic provinces

We define and validate environmental niches using four machine learning techniques for 27 previously defined genomic provinces[11].

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, Evry, France. [2]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans, Paris, France. [3]LSCE-IPSL, CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France. [4]Stazione Zoologica Anton Dhorn, Naples, Italy. ✉e-mail: pfremont@genoscope.cns.fr; marion.gehlen@lsce.ipsl.fr; ojaillon@genoscope.cns.fr

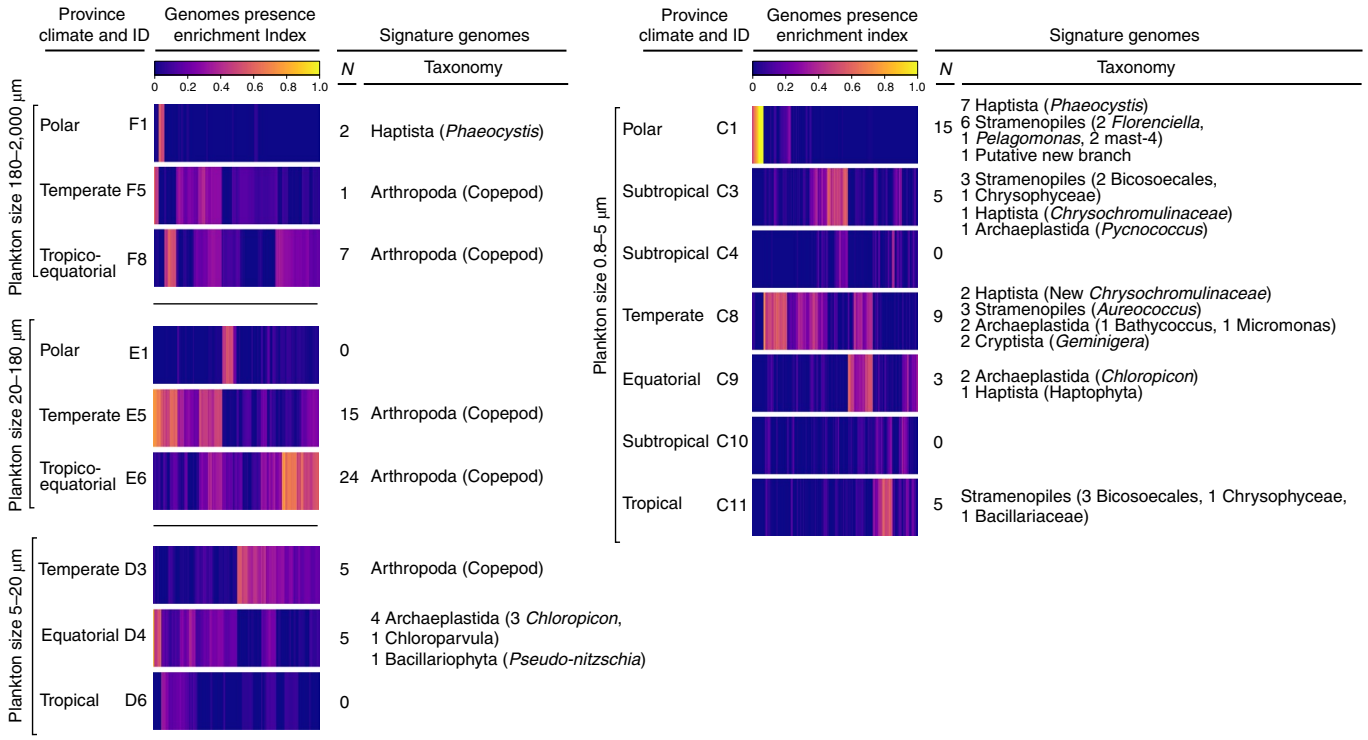


Fig. 1 | Eukaryotic signature genomes of provinces of eukaryote enriched size classes. For each plankton size class, indexes of presence enrichment (equation (1)) for 713 genomes of eukaryotic plankton[31] in corresponding provinces are clustered and represented in a colour scale. Signature genomes (Methods) are found for almost all provinces; their number (*N*) and taxonomies are summarized (detailed list in Supplementary Table 6). For each size class, a genome is considered to be signature of a province if its presence enrichment index is superior to 0.5 for that province and inferior to 0.1 for other provinces.

These provinces cover major ocean regions except the Arctic for which no data were available at the time of this study. They correspond to 529 metagenomes for six size fractions (ranging from 0 µm to 2,000 µm) sampled at 95 sites (Supplementary information 1, Extended Data Fig. 1 and Supplementary Fig 1). Niche definition relies on predictor variables that significantly differentiate genomic provinces[11]: sea surface temperature (SST), salinity, dissolved silica, nitrate, phosphate and iron, plus a seasonality index of nitrate (equation (2)).

The signal of ocean partitioning is probably due to abundant and small genomes whose geographical distributions closely match provinces. Within a collection of 1,778 bacterial, 110 archaeal and 713 eukaryotic environmental genomes[31,32] characterized from *Tara* Oceans samples without cultivation, we find a total of 324 signature genomes covering all but four provinces and displaying a taxonomic signal coherent with the size fractions (Fig. 1 for eukaryotes and Extended Data Fig. 2 for Bacteria and Archaea). Some of the signature genomes correspond to unexplored lineages with no cultured representatives, highlighting the knowledge gap for organisms that structure plankton biogeography and the strength of a rationale devoid of any a priori on reference genomes or species.

Structure of present-day biogeography of plankton

Niches are extrapolated to a global ocean biogeography for each size fraction at 1°×1° spatial resolution using environmental predictors from 2006–2013 World Ocean Atlas 2013 (WOA13) climatology[33]. At each grid point, provinces with the highest probability of presence computed from the mean projection of four machine learning models are retained (Fig. 2 and Supplementary Fig. 2). They are referred to as *dominant* provinces and manually annotated based on their climate (Supplementary Table 1).

In agreement with previous observations[11], provinces of large size fractions (>20 µm) are wider and partially decoupled from those of smaller size fractions, probably due to differential responses to oceanic circulation and environmental variations, different life cycle constraints, lifestyles[7,11] and trophic network positions[34]. Biogeographies of small metazoans that enrich the largest size fractions (180–2,000 µm and 20–180 µm) are broadly aligned with latitudinal bands (tropico-equatorial, temperate and subpolar) dominated by a single province (Fig. 2a,b). A more complex oceanic structuring emerges for the smaller size fractions (<20 µm) (Fig. 2c–f) with several provinces per large geographical region. For size fraction 0.8–5 µm enriched in small protists (Fig. 2d), distinct provinces are identified for tropical oligotrophic gyres and for the nutrient-rich equatorial upwelling region. A complex pattern of provinces, mostly latitudinal, is found for the bacteria- (Fig. 2e, 0.22–3 µm) and the virus enriched size classes (Fig. 2f, 0–0.2 µm), although less clearly linked to large-scale oceanographic regions. A single province extending from temperate to polar regions emerges for size fraction 5–20 µm enriched in protists (Fig. 2c) for which fewer samples were available (Supplementary Fig. 1b,c), a likely source of bias. A consensus map combining all size fractions, built using the PHATE algorithm[35], summarizes the main characteristics of the biogeographies described above (Supplementary Discussion 2 and Supplementary Fig. 3).

Finally, we compare genomic biogeographies and existing ocean partitionings[8–10]. Though each of them is unique (Supplementary Figs. 6–8), common borders highlight a global latitudinal partitioning independent of the type of data (Supplementary Discussion 3).

Future changes in plankton biogeography structure

We assess the impacts of climate change on plankton biogeography at the end of the century following the Representative Concentration

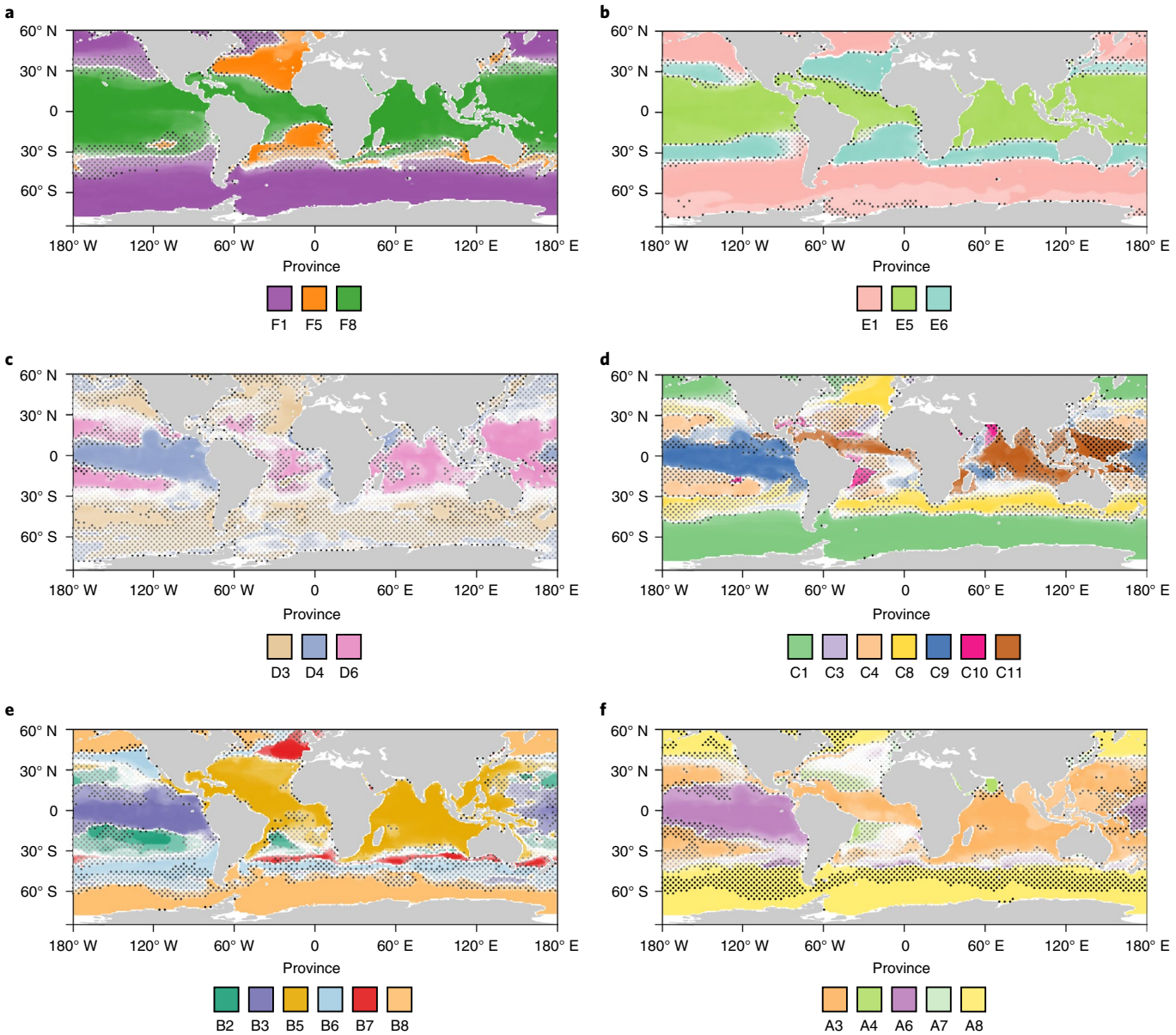


Fig. 2 | Global biogeographies of size-structured plankton provinces projected on WOA13 dataset. a–f, Maps showing biogeographies of metazoans enriched (180–2,000 µm) (**a**), small metazoans enriched (20–180 µm) (**b**), protist enriched (5–20 µm) (**c**), protist enriched (0.8–5 µm) (**d**), bacteria enriched (0.22–3 µm) (**e**) and viruses enriched (0–0.2 µm) (**f**) size fraction. At each grid point of the maps, the *dominant* province is represented using a darkness of colour proportional to its presence probability. Dots represent areas of uncertainty (where the delta of probability between the *dominant* and another province is inferior to 0.5). Simple biogeographies are observed in large size fractions (>20 µm) with a partitioning in three major oceanic areas: tropico-equatorial, temperate and polar. More complex geographic patterns and patchiness are observed in smaller size fractions. Maps generated using R package 'maps'[51].

Pathway 8.5 (RCP8.5)[36] greenhouse gas concentration trajectory. To consistently compare projections of present and future biogeographies, we use the bias-adjusted mean of six Earth System Model (ESM) climatologies (Supplementary Table 2 and Supplementary Fig. 9). The highest SST warming (7.2 °C) is located off the east coast of Canada in the North Atlantic while complex patterns of salinity and nutrient variations are projected in all oceans (Supplementary Fig. 10). Following this trajectory, future temperature at most sampling sites will be higher than the mean and maximum contemporary temperature within their current province (Extended Data Fig. 3).

Our projections indicate multiple large-scale changes in biogeographical structure, including expansions, shrinkages and shifts in plankton organism size-dependent provinces (Fig. 3a–d, Extended Data Fig. 4 and Supplementary Figs. 11 and 12). A change in the *dominant* province in at least one size fraction would occur over 60.1% of the ocean surface studied, ranging from 12% (20–180 µm) to 31% (0.8–5 µm) (Fig. 3 and Table 1). Consistent with previous studies[22,37], the majority of shifts are poleward (Supplementary Discussions 2, 4 and 5).

We calculate a dissimilarity index (equation (4)) at each grid point between probabilities of future and present *dominant* provinces for all size fractions combined (Fig. 4a). Large dissimilarities are obtained over northern (25° to 60°) and southern (−25° to −60°) temperate regions (mean of 0.29 and 0.24, respectively), mostly reflecting the poleward shift of temperate provinces (red arrows in Fig. 4a). In austral and equatorial regions, despite important environmental changes (Supplementary Fig. 10) and previously

projected changes in diversity[27–29] and biomass[38], the contemporary provinces remain the most probable at the end of the century (mean dissimilarities of 0.18 and 0.02, respectively).

To further study the future decoupling between provinces of different size fractions, we analyse the assemblages of *dominant* provinces of each size fraction. Whether a threshold is applied to the dissimilarity index (>1/6; Methods), from 45.3% to 57.1% of the ocean surface, mainly located in temperate regions, would be inhabited in 2090–2099 by assemblages that exist elsewhere in 2006–2015 (Fig. 4b versus Fig. 4c). Contemporary assemblages would disappear on 3.5% to 3.8% of the surface, and, conversely, novel assemblages not encountered today would cover 2.9% to 3.0% of the surface. Changes of assemblages impact key economic zones with over 41.8% to 51.8% of the surface of the main fisheries and 42.2% to 55.2% of exclusive economic zones for which the future assemblage would differ from the contemporary one (Extended Data Fig. 5).

Impact on plankton groups driving carbon and nitrogen cycles

To assess the potential biogeochemical impact of biogeographical restructuring, we focus on compositional changes among nitrogen-fixing bacteria (diazotrophs), copepods and phototrophs, three groups considered important for the carbon and nitrogen cycles[39,40] that are well represented among the *Tara* Oceans environmental genomes. Focusing on marine areas where *dominant* provinces are projected to be replaced, we compare the present and future distribution of environmental genomes corresponding to 27 diazotrophs, 198 copepods and 231 phototrophs[31,32].

We project statistically significant changes in composition for all three groups (Fig. 3e,f and Extended Data Figs. 6–9, Wilcoxon test, $p < 0.05$). Here we describe only changes in diazotrophs for which the genome collection most likely encompasses the whole diversity of the group (Supplementary Discussion 6 addresses changes in copepods and phototrophs).

Diazotrophy, the biotic fixation of atmospheric nitrogen, is an important process for both nitrogen and carbon cycles. It supports biological productivity in the nitrogen-limited tropical oceans[41]. Marine diazotrophs include cyanobacteria (for example, *Trichodesmium*) described as the principal nitrogen fixers[42] and various heterotrophic bacterial diazotrophs (HBDs) that lack cultured representatives or in situ imaging[32,41]. Forty-eight of the bacterial environmental genomes are diazotrophs (eight cyanobacteria and 40 HBDs). These genomes encapsulate 92% of the metagenomic signal of the known marker gene responsible for microbial nitrogen fixation[32] (nifH, the nitrogenase subunit H). Twenty-seven are found in at least five samples of a given size class and have been used for the analysis.

In size fraction 0.8–5 µm, we project significantly higher relative abundances in cyanobacteria in the tropico-equatorial regions of the Pacific Ocean (C9 to C11 and C4; Fig. 3f), which might point towards an increase in nitrogen fixation in this region as previously suggested by other models[40]. Supporting this result, we find similar significant compositional changes towards an increase in cyanobacteria in size fraction 5–20 µm (Extended Data Fig. 7c). We also project significant compositional changes for some clades of HBDs (for example, increase in gammaproteobacteria: C8 to C3; Fig. 3f) though no global trend can be identified here. In the other size classes (20–180 µm, 180–2,000 µm and 0.22–3 µm), compositional changes are not significant (Extended Data Fig. 8). To summarize, although we cannot estimate nitrogen fixation rates using genomic data alone, genomic measurements of nitrogen-fixing cyanobacteria are in agreement with a potential increase in nitrogen fixation in the tropics, echoing results from other models[40].

Linking assemblages of provinces and carbon export fluxes

Carbon export refers to the processes by which organic carbon is transferred from the surface ocean to depth mainly by sinking particles that are derived from surface ocean plankton ecosystem processes[43]. Surface plankton communities are reported to be important in determining carbon export fluxes through the water column[2]. We test this hypothesis by calculating mean fluxes of carbon export that can be associated to the assemblages of our *climato-genomic* provinces from three datasets of extrapolated carbon fluxes[3,43,44]. We find significant differences of carbon fluxes for 46% of pairs of province assemblages (pairwise Wilcoxon test, Holm correction $p < 0.05$). Total carbon export values are comparable to those computed from the reference datasets over the same regions (from 2.6 using Henson et al.[3] to 6.7 Pg C per year using Eppley et al.[43]; Supplementary Table 9). Our projections of the reorganization of *climate-genomic* provinces correspond with a decrease in carbon export of 4.0% on average by the end of the century (from 3.3% for Laws et al.[44] to 4.4% for Eppley et al.[43]; Fig. 4d).

We use the Apriori algorithm[45] to identify associations between changes in the genome composition of communities and changes in carbon export among four main latitudinal zones of the ocean: equatorial, subtropical north/south and temperate/subpolar (Extended Data Fig. 10). In temperate/subpolar regions, no significant association rules are found (Extended Data Fig. 10a). In northern and southern subtropical regions, decreases in carbon export are associated with global decreases in nano-diatoms, nano-algae and pico-algae (0.8–5 µm and 5–20 µm; Extended Data Fig. 10b,c). In equatorial regions, the decrease in carbon export is mainly associated with decreases in nano-diatoms and nano-algae (5–20 µm and 0.8–5 µm) and increases in diazotrophs (0.8–5 µm and 5–20 µm; Extended Data Fig. 10d). No significant rules are found for carbon export increase or changes in copepods and micro-algae.

Drivers of plankton biogeography reorganization

We quantify the relative importance of environmental predictors (SST, salinity, dissolved silica, phosphate, nitrate, iron and seasonality of nitrate) into niche definitions and in driving future changes of the structure of plankton biogeography (equation (6)). Among these environmental properties, temperature is the first influential parameter for niche definition (for 19 out of 27 niches) but with a relative contribution of only 22.6% on average (Supplementary Fig. 16a).

For each site highly dissimilar between 2006–2015 and 2090–2099 (Bray-Curtis > 1/6), the relative impact of each environmental parameter in driving this change is calculated (equation (6) (ref. [22]); Fig. 5a). Overall, comparing the relative contribution of each parameter to the reorganization of the provinces, SST would be

Fig. 3 | Modelled present and future global biogeographies of the three and seven provinces for the small metazoan enriched size fraction (20–180 µm) and the protist enriched size fraction (0.8–5 µm) and associated compositional shifts in marine hexanauplia and diazotrophs. a–d, Presence probability and uncertainties are represented as in Fig. 2. **e–f**, At each grid point of the maps, the dominant province is represented using a darkness of colour proportional to its presence probability. Dots represent areas of uncertainty (where the delta of probability between the dominant and another province is inferior to 0.5). Top: locations of *dominant* province shifts in size fraction 20–180 µm (**e**) and 0.8–5 µm (**f**). Bottom: Summary of significant compositional shifts in marine Hexanauplia (copepods) (**e**) classified by size ('not classified' when no preferential size class was found) or bacterial diazotrophs (**f**). Each type of transition is coloured according to the one on the map or in grey (when less frequent). Bar plots represent mean relative abundances based on genome abundance in the given province. Statistically significant compositional changes in a type of genome are represented by triangles of the associated transition colour (Wilcoxon test, $p < 0.05$). Maps generated using R package 'maps'[51].

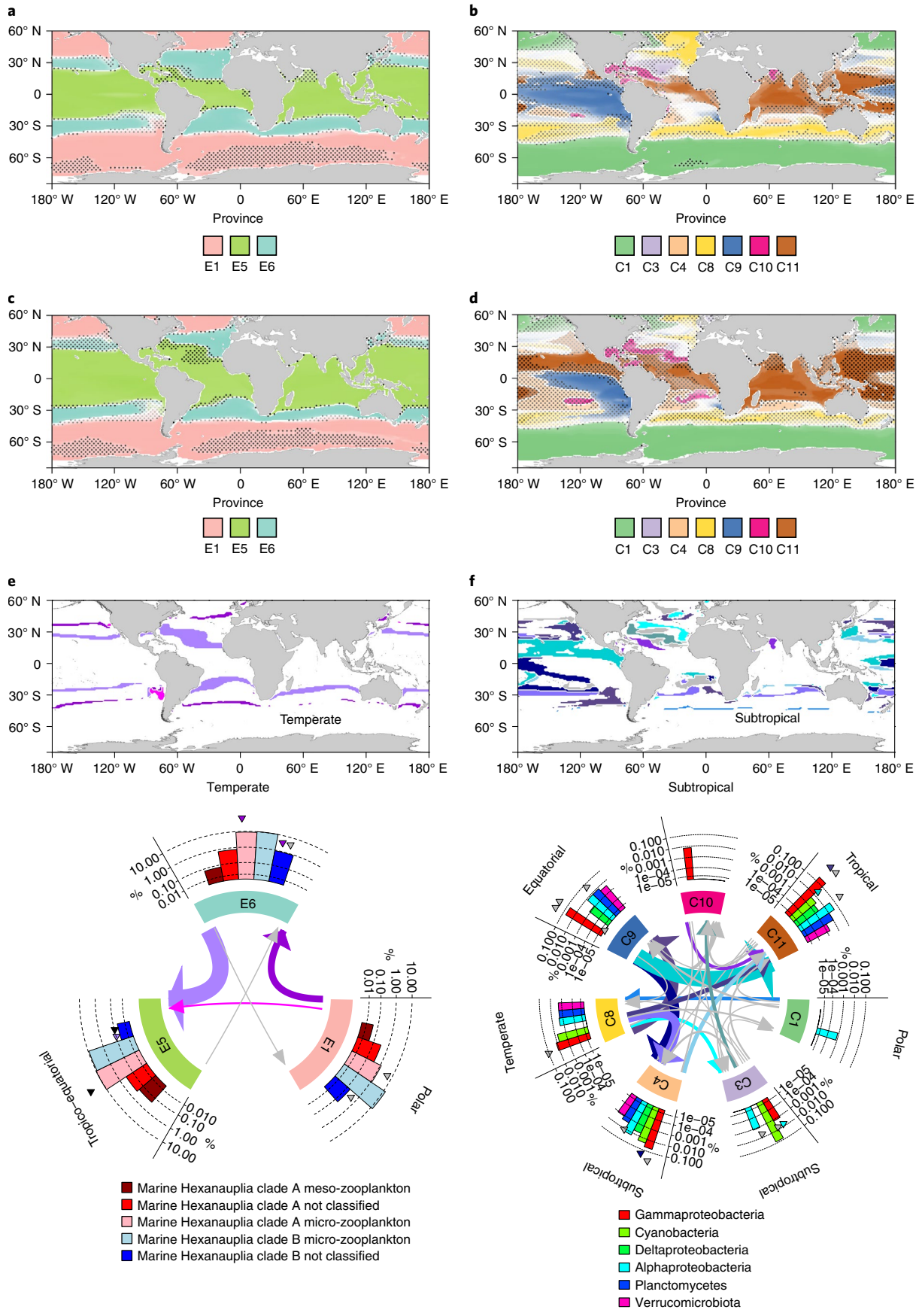


Table 1 | Global statistics of covered areas and province changes and transitions

Size fraction (µm)	Present-day covered area (%)	End-of-century covered area (%)	% area with a change of *dominant* province	Most frequent transition	%	Second-most frequent transition	%
180–2,000	74	74	13	temperate → tropico-equatorial (F5 → F8)	67	polar → temperate (F1 → F5)	14
20–180	78	77	12	temperate → tropico-equatorial (E6 → E5)	67	polar → temperate (E1 → E6)	29
5–20	45	49	22	temperate → equatorial (D3 → D4)	47	equatorial → tropical (D4 → D6)	25
0.8–5	56	59	31	equatorial → tropical (C9 → C11)	22	temperate → equatorial (C8 → C9)	15
0.22–3	60	61	15	temperate → tropical (B7 → B5)	22	polar → temperate (B8 → B6)	16
0–0.2	64	66	16	equatorial → tropical (A6 → A3)	32	temperate → equatorial (A8 → A6)	31
Total			**60**				

From 12% to 31% of the total covered area is estimated to be replaced by a different province at the end of the century compared with present day depending on the size fraction. In total, considering all size fractions, this represents 60% of the total covered area with at least one predicted change of *dominant* province across the six size fractions.

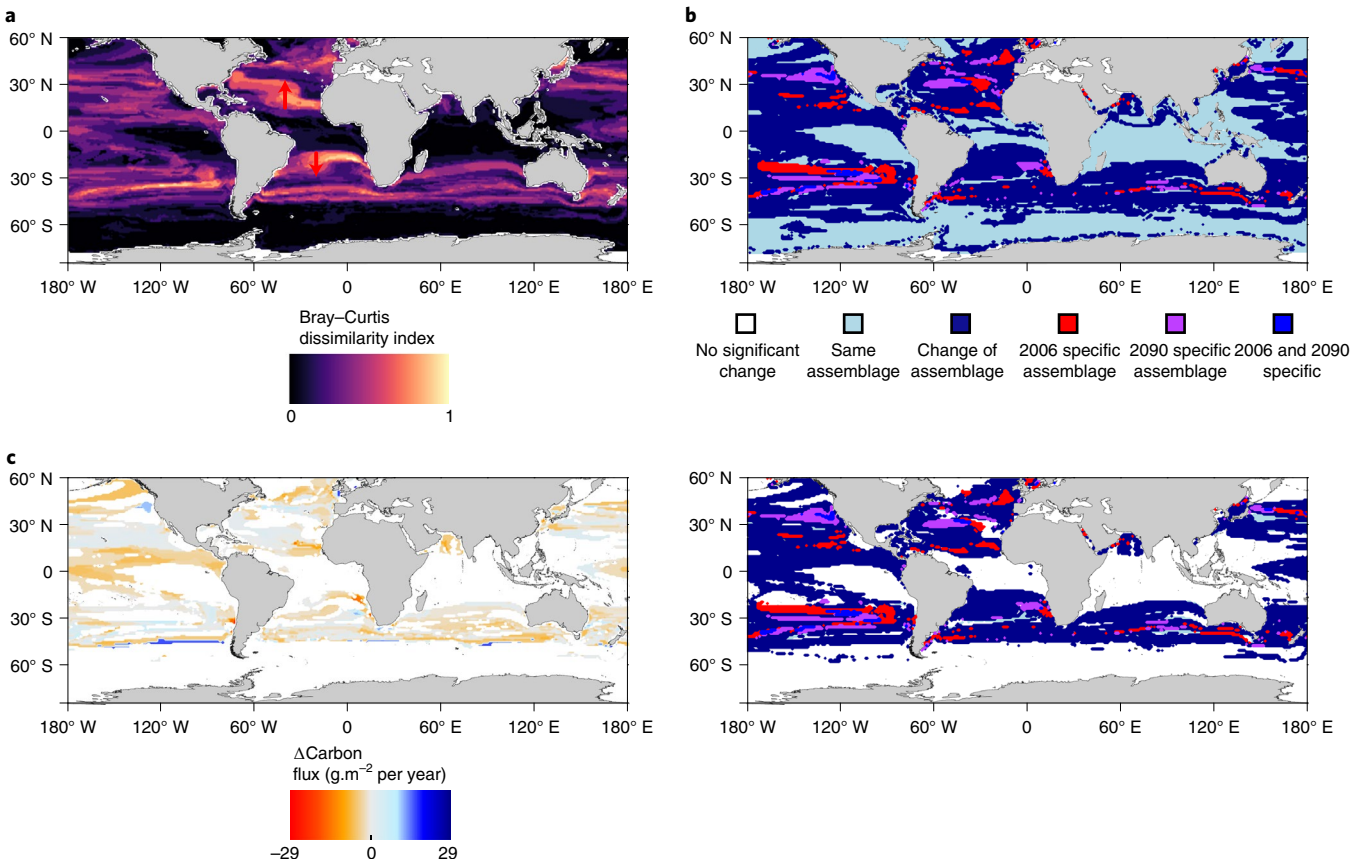


Fig. 4 | Global impact of climate change on plankton community assemblages and carbon export. a, Bray–Curtis dissimilarity index (equation (4)) is calculated by integrating all the *dominant* provinces presence probabilities over the six size fractions. **b**, An assemblage is the combined projected presence of the dominant province of each size class. Assemblage reorganization is either mapped on all considered oceans (top) or with a criterion on the Bray–Curtis index (Bray–Curtis >1/6) (bottom). New assemblages are expected to appear in 2090 (purple + blue), whereas some 2006 specific assemblages are projected to disappear (red + blue). **c**, Difference in carbon export by the end of the century based on present day mean export within each assemblage (Henson et al.[3]). Maps generated using R package 'maps'[51].

responsible for 50% on average followed by phosphate (11%) and salinity (10.3%) (Supplementary Fig. 17). Over the majority of the ocean, SST is the primary driver of the reorganization (Fig. 5a). In some regions, salinity (for example, eastern North Atlantic) and phosphate (for example, equatorial region) dominate (Fig. 5a). When excluding the effect of SST, salinity and phosphate become

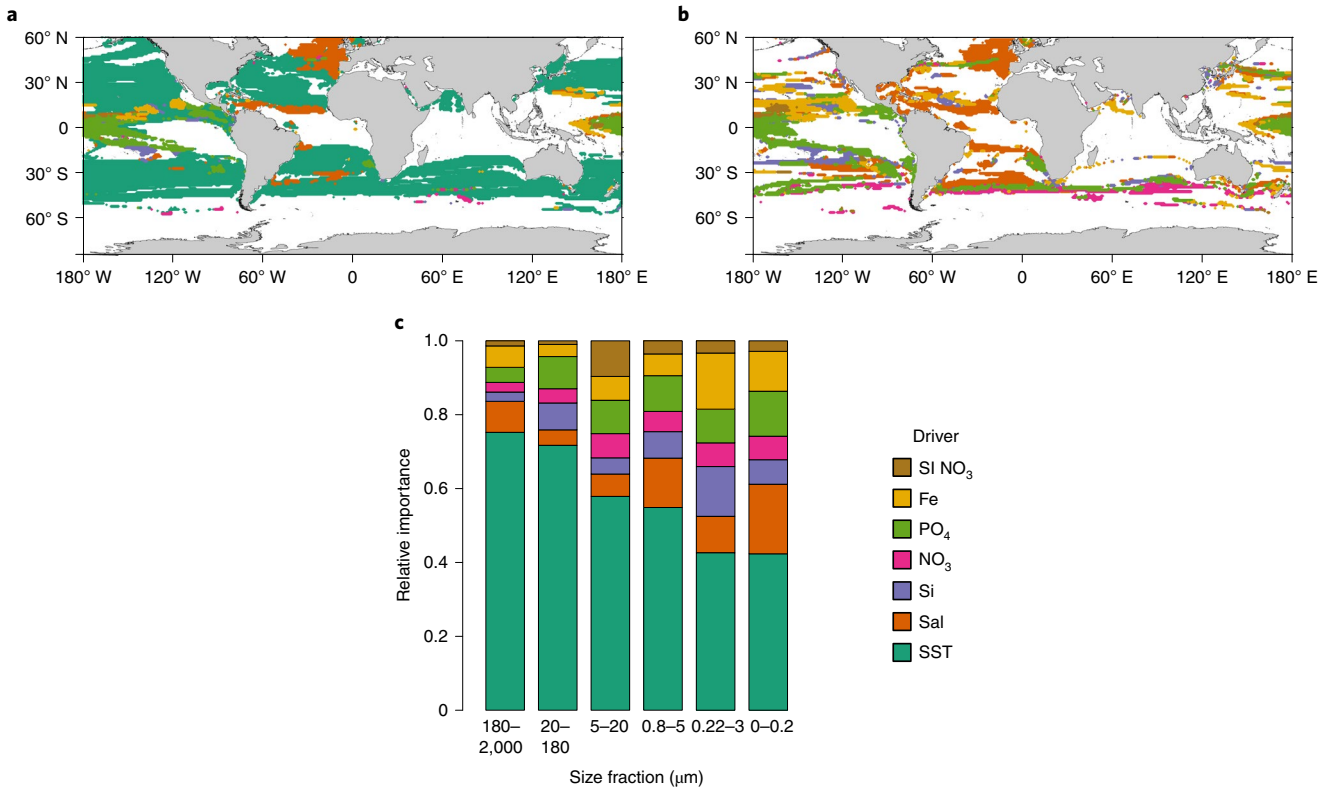


Fig. 5 | Drivers of plankton biogeographical restructuring in response to climate change. a, Map of most impacting drivers on *dominant* province changes. Temperature appears as the top impacting driver on the majority of the ocean with a significant change of province. **b**, Most impacting driver without considering temperature change. Salinity (Sal) and phosphate are found to be the second and third drivers of province reorganization; notably, they dominate at tropical and subpolar latitudes. **c**, Relative importance of the drivers in the different size fractions. Sea Surface Temperature (SST) is found to be the most important driver for all size classes but has a more important impact in large size classes (>20 µm). SI NO_3 is a seasonality index of nitrate (Methods). Maps generated using R package 'maps'[51].

the primary drivers of the reorganization of the provinces (Fig. 5b). The impact of SST varies across size classes with a significantly higher contribution in large size classes (>20 µm) compared with the small ones (mean of ~73% versus ~49%, *t*-test $p < 0.05$; Fig. 5c). Though the contribution of combined nutrients to niche definition is similar for small and large size classes (mean of ~56% versus ~61%; Supplementary Fig. 16 and Supplementary Table 3), their future projected variations have a higher relative impact on the reorganization of biogeographies of small organisms (mean of ~39% versus ~20%, *t*-test $p < 0.05$; Supplementary Fig. 16 and Supplementary Table 3). For instance, in the tropical zone, the shrinkage of the equatorial province C9 (size fraction 0.8–5 µm; Fig. 2b,d and Supplementary Fig. 18e) is 24% driven by a reduction of dissolved phosphate concentrations and 25% by a SST increase. In contrast, SST drives 56% of the shrinkage of the temperate province F5 (size fraction 180–2,000 µm; Supplementary Fig. 18d). Finally, non-poleward shifts are found only within small size fractions (<20 µm) (Extended Data Fig. 4 and Supplementary Fig. 11), highlighting differential responses to nutrients and SST changes between large and small organisms, the latter being enriched in phytoplankton that directly rely on nutrient supplies.

Discussion

We propose a novel partitioning of the ocean into plankton size-dependent *climato-genomic* provinces, complementing previous efforts based on other biophysicochemical data[9–11]. Though initially built at genomic scale, our biogeographies paradoxically reveal basin-scale provinces that are larger than Longhurst's biogeochemical provinces[10] and Fay and McKingley biomes[9]. We propose that this apparent paradox emerges from the combination of the scale, nature and resolution of sampling. First, two proximal samples from the *Tara* Oceans expedition are separated by ~300 km on average, sampled over three years. This relatively large spatiotemporal scale overlies shorter-scale compositional variations previously observed[5,6]. At the global scale, the limited effect of seasons on the frontiers of biogeochemical provinces[10] does not affect their global topology which might be the case for the global structuring of the genomic biogeography. Second, our estimates of plankton community dissimilarities are highly resolutive as they are computed at genomic scale with billions of small DNA fragments[11,18], thus smoothing out the more discrete species-level signals. Together, from these combinations of processes and patterns occurring at multiple scales emerge basin-scale provinces associated with coherent environmental niches and signature genomes.

These *climato-genomic* provinces are structured in broad latitudinal bands with smaller organisms (<20 µm) displaying more complex patterns and partially decoupled from larger organisms. This decoupling is the result of distinct statistical links between provinces based on organism size fractions and environmental parameters and could reflect their respective trophic modes[11,34].

Complex changes of the parameters defining the niches are projected under climate change, leading to the reorganization of size-dependent provinces. Assuming a constant relationship to environmental drivers that define the *climato-genomic* provinces, climate change is projected to restructure them over approximately 50% of surface oceans south of 60° N by the end of the century (Fig. 4). The largest reorganization is detected in subtropical and temperate regions in agreement with other studies[29,37] and is accompanied

by appearance and disappearance of size-fractionated provinces' assemblages. In the tropical-equatorial and southern regions, environmental conditions are projected to exceed current extremes by the end of the century. While some studies extrapolate important diversity and biomass changes in these zones[27–29,38], here we project shifts of their boundaries and maintenance of their climatic label. The present approach does not account for putative changes in community composition or the emergence of novel niches over these regions for which novel environmental selection pressure is expected.

Despite these limitations, genomic data allow us to project compositional shifts due to the reorganization of the provinces at a high phylogenetic and functional resolution. We find significant changes in composition of diazotrophs, copepods and phototrophs including changes in size, a key trait for carbon export. These changes include increase of large copepods in subpolar regions (180–2,000 µm), increase in nitrogen-fixing cyanobacteria in subtropical regions in agreement with modelling studies[40] and complex changes in phototrophs. These changes might lead to novel prey–predator interactions, for example, in regions where new assemblages of communities are projected. Critically, carbon export fluxes are significantly different among *climato-genomic* province assemblages in agreement with theory[3,26] and previous genomic studies[2]. By the end of the century, changes in these assemblages are linked to a global decrease of the carbon pump of ~4%, a consistent estimation compared with most models[26,46]. This decrease is linked to decreases in nano-algae and diatoms and an increase in diazotrophs but not to changes in copepods or micro-algae. These results are to be taken with caution, especially for copepods and phototrophs, as our genome database does not account for the total diversity of these groups and is biased towards small and the most abundant genomes. Nevertheless, the presented association of metagenomic data with physicochemical data paves the way for the integration of multiple global-scale data for a better understanding of biogeochemical cycles.

Overall, our projections for the end of the century do not take into account possible future changes of major biophysicochemical factors such as the dynamics of community mixing, trophic interactions through transport[47], the dynamics of the genomes[13–15] (adaptation or acclimation) and biomass variations[38]. New sampling in current and future expeditions[48] and ongoing technological improvements in biophysicochemical characterization of seawater samples[31,48,49] will soon refine functional[16,17], environmental (micronutrients[50]) and phylogenetic[31] characterization of plankton ecosystems for various biological entities (genotypes, species or communities) and spatiotemporal scales[48]. Ultimately, integrating this varied information will allow a better understanding of the conditions of emergence of ecological niches in the seascape and their response to a changing ocean.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41558-022-01314-8.

Received: 28 October 2020; Accepted: 15 February 2022;
Published online: 4 April 2022

References

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* https://doi.org/10.1126/science.281.5374.237 (1998).
2. Guidi, L. et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* https://doi.org/10.1038/nature16942 (2016).
3. Henson, S. A., Sanders, R. & Madsen, E. Global patterns in efficiency of particulate organic carbon export and transfer to the deep ocean. *Glob. Biogeochem. Cycles* https://doi.org/10.1029/2011GB004099 (2012).
4. Azam, F. et al. The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* https://doi.org/10.3354/meps010257 (1983).
5. Saab, M. A. Day-to-day variation in phytoplankton assemblages during spring blooming in a fixed station along the Lebanese coastline. *J. Plankton Res.* https://doi.org/10.1093/plankt/14.8.1099 (1992).
6. Djurhuus, A. et al. Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nat. Commun.* https://doi.org/10.1038/s41467-019-14105-1 (2020).
7. Kavanaugh, M. T. et al. Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation. *ICES J. Mar. Sci.* https://doi.org/10.1093/icesjms/fsw086 (2016).
8. Longhurst, A. R. *Ecological Geography of the Sea* (Elsevier, 2007).
9. Fay, A. R. & McKinley, G. A. Global open-ocean biomes: mean and temporal variability. *Earth Syst. Sci. Data* https://doi.org/10.5194/essd-6-273-2014 (2014).
10. Reygondeau, G. et al. Dynamic biogeochemical provinces in the global ocean. Glob. Biogeochem. *Cycles* https://doi.org/10.1002/gbc.20089 (2013).
11. Richter, D. J. et al. Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. Preprint at *bioRxiv* https://doi.org/10.1101/867739 (2020).
12. Dutkiewicz, S. et al. Dimensions of marine phytoplankton diversity. *Biogeosciences* https://doi.org/10.5194/bg-17-609-2020 (2020).
13. Hellweger, F. L., Van Sebille, E. & Fredrick, N. D. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* https://doi.org/10.1126/science.1254421 (2014).
14. Laso-Jadart, R. et al. Investigating population-scale allelic differential expression in wild populations of Oithona similis (Cyclopoida, Claus, 1866). *Ecol. Evol.* https://doi.org/10.1002/ece3.6588 (2020).
15. Delmont, T. O. et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* https://doi.org/10.7554/eLife.46497 (2019).
16. Carradec, Q. et al. A global ocean atlas of eukaryotic genes. *Nat. Commun.* https://doi.org/10.1038/s41467-017-02342-1 (2018).
17. Salazar, G. et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* https://doi.org/10.1016/j.cell.2019.10.014 (2019).
18. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Sci. Data* https://doi.org/10.1038/sdata.2017.93 (2017).
19. Pesant, S. et al. Open science resources for the discovery and analysis of *Tara* Oceans data. *Sci. Data* https://doi.org/10.1038/sdata.2015.23 (2015).
20. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol.* https://doi.org/10.1371/journal.pbio.1001177 (2011).
21. Duarte, C. M. Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnol. Oceanogr. Bull.* https://doi.org/10.1002/lob.10008 (2015).
22. Barton, A. D., Irwin, A. J., Finkel, Z. V. & Stock, C. A. Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1519080113 (2016).
23. Benedetti, F., Guilhaumon, F., Adloff, F. & Ayata, S. D. Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea. *Ecography* https://doi.org/10.1111/ecog.02434 (2018).
24. Beaugrand, G. et al. Prediction of unprecedented biological shifts in the global ocean. *Nat. Clim. Change* **9**, 237–243 (2019).
25. Pinsky, M. L., Worm, B., Fogarty, M. J., Sarmiento, J. L. & Levin, S. A. Marine taxa track local climate velocities. *Science* https://doi.org/10.1126/science.1239352 (2013).
26. Bopp, L. et al. Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* https://doi.org/10.5194/bg-10-6225-2013 (2013).
27. Thomas, M. K., Kremer, C. T., Klausmeier, C. A. & Litchman, E. A global pattern of thermal adaptation in marine phytoplankton. *Science* https://doi.org/10.1126/science.1224836 (2012).
28. Ibarbalz, F. M. et al. Global trends in marine plankton diversity across kingdoms of life. *Cell* https://doi.org/10.1016/j.cell.2019.10.008 (2019).
29. Busseni, G. et al. Large scale patterns of marine diatom richness: drivers and trends in a changing ocean. *Glob. Ecol. Biogeogr.* https://doi.org/10.1111/geb.13161 (2020).
30. Hutchinson, G. E. Concluding remarks. *Cold Spring Harb. Symp. Quant. Biol.* **22**, 415–427 (1957).
31. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.15.341214 (2020).

32. Delmont, T. O. et al. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *ISME J.* https://doi.org/10.1038/s41396-021-01135-1 (2021).
33. Boyer, et al. *World Ocean Database 2013, NOAA Atlas NESDIS 72* (National Oceanic and Atmospheric Administration, 2013); https://doi.org/10.7289/V5NZ85MT
34. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* https://doi.org/10.1038/s41579-020-0364-5 (2020).
35. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0336-3 (2019).
36. van Vuuren, D. P. et al. The representative concentration pathways: an overview. *Climatic Change* https://doi.org/10.1007/s10584-011-0148-z (2011).
37. Polovina, J. J., Dunne, J. P., Woodworth, P. A. & Howell, E. A. Projected expansion of the subtropical biome and contraction of the temperate and equatorial upwelling biomes in the North Pacific under global warming. *ICES J. Mar. Sci.* https://doi.org/10.1093/icesjms/fsq198 (2011).
38. Flombaum, P., Wang, W. L., Primeau, F. W. & Martiny, A. C. Global picophytoplankton niche partitioning predicts overall positive response to ocean warming. *Nat. Geosci.* https://doi.org/10.1038/s41561-019-0524-2 (2020).
39. Richardson, A. J. In hot water: zooplankton and climate change. *ICES J. Mar. Sci.* **65**, 279–295 (2008).
40. Wrightson, L. & Tagliabue, A. Quantifying the impact of climate change on marine diazotrophy: insights from Earth system models. *Front. Mar. Sci.* **7**, 635 (2020).
41. Zehr, J. P. & Capone, D. G. Changing perspectives in marine nitrogen fixation. *Science* **368**, eaay9514 (2020).
42. Luo, Y.-W. et al. Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst. Sci. Data* **4**, 47–73 (2012).
43. Eppley, R. W. & Peterson, B. J. Particulate organic matter flux and planktonic new production in the deep ocean. *Nature* **282**, 677–680 (1979).
44. Laws, E. A., Falkowski, P. G., Smith, W. O., Ducklow, H. & McCarthy, J. J. Temperature effects on export production in the open ocean. *Glob. Biogeochem. Cycles* **14**, 1231–1246 (2000).
45. Agrawal, R. & Srikant, R. in *Proceedings of the 20th International Conference on Very Large Data Bases* (eds Bocca, J. B. et al.) 487–499 (Morgan Kaufmann, 1994).
46. Laufkötter, C. et al. Projected decreases in future marine export production: the role of the carbon flux through the upper ocean ecosystem. *Biogeosciences* **13**, 4023–4047 (2016).
47. Iudicone, D. Some may like it hot. *Nat. Geosci.* https://doi.org/10.1038/s41561-020-0535-z (2020).
48. Gorsky, G. et al. Expanding Tara Oceans protocols for underway, ecosystemic sampling of the ocean–atmosphere interface during Tara Pacific expedition (2016–2018). *Front. Mar. Sci.* https://doi.org/10.3389/fmars.2019.00750 (2019).
49. Istace, B. et al. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* https://doi.org/10.1093/gigascience/giw018 (2017).
50. Grand, M. M. et al. Developing autonomous observing systems for micronutrient trace metals. *Front. Mar. Sci.* https://doi.org/10.3389/fmars.2019.00035 (2019).
51. Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P. & Deckmyn, A. maps: Draw geographical maps. R version 3.5.0 https://cran.r-project.org/web/packages/maps/index.html (2021).

Methods

Genomic provinces of plankton. Environmental niches are computed for trans-kingdom plankton genomic provinces from Richter et al.[11]. These provinces consist of the clustering of each of six metagenomic dissimilarity matrices (based on the amount of DNA k-mers shared between pairs of samples) from six available size fractions with sufficient metagenomic data from the *Tara* Oceans dataset. The six size fractions (0–0.2 µm, 0.22–3 µm, 0.8–5 µm, 5–20 µm, 20–180 µm and 180–2,000 µm) were initially sampled to represent major plankton groups[18–20,34]. Two large size classes (180–2,000 µm and 20–180 µm) are enriched in zooplankton dominated by arthropods (mainly copepods) and cnidarians. Size classes 5–20 µm and 0.8–5 µm are enriched in smaller eukaryotic algae such as dynophytes (5–20 µm), pelagophytes and haptophytes (0.8–5 µm). Finally, size classes 0.22–3 µm and 0–0.2 µm are respectively enriched in bacteria and viruses. Within each size fraction (from large to small), there are respectively 8, 8, 11, 6, 6 and 8 (48 in total) provinces defined in Richter et al.[11] formed by *Tara* Oceans metagenomes (644 metagenomes sampled either at the surface or at the deep chlorophyll maximum across 102 sites).

Genome signature of the provinces. We analysed the distribution of 713 eukaryotic and 1,888 prokaryotic genomes[31,32] within the genomic provinces. These genomes are metagenome-assembled genomes (MAGs) obtained from *Tara* Oceans metagenomes. For each size class, we select MAGs that are present (according to a criteria defined in Delmont et al.[31]) in at least five samples. We computed an index of presence enrichment of MAGs within provinces as the Jaccard index (J)[52], defined as follows:

$$J = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} \quad (1)$$

M_{11} is the number of samples where the MAG is present and matches a sample of the province. M_{01} and M_{10} are, respectively, the number of samples where the MAG is not present in a sample of the province and, reciprocally, the number of samples where the MAG is present outside the province. A MAG is considered to be signature of a province if the Jaccard index is superior to 0.5 for this province and inferior to 0.1 for all other provinces of the given size class (Fig. 1 and Extended Data Fig. 2).

WOA13 data. Physicochemical parameters proposed to have an impact on plankton genomic provinces[11] are used to define environmental niches: SST, salinity (Sal), dissolved silica (Si), dissolved nitrate (NO_3), dissolved phosphate (PO_4), dissolved iron (Fe) and a seasonality index of nitrate (SI NO_3). With the exception of Fe and SI NO_3, these parameters are extracted from the gridded WOA13[33]. Climatological Fe fields are provided by the biogeochemical model PISCES-v2[55]. SI NO_3 is defined as the range of nitrate concentrations over the year in one grid cell divided by the maximum range encountered in WOA13 at the *Tara* Oceans sampling stations:

$$\mathrm{SI}\left([\mathrm{NO_3}](x,y,z)\right) = \frac{\max_t\left([\mathrm{NO_3}](x,y,z,t)\right) - \min_t\left([\mathrm{NO_3}](x,y,z,t)\right)}{\max_S\left(\max_t\left([\mathrm{NO_3}](x,y,z,t)\right) - \min_t\left([\mathrm{NO_3}](x,y,z,t)\right)\right)} \quad (2)$$

$[NO_3]$ is the nitrate concentration. The indices t and S, respectively, refer to time (months) and spatial (*Tara* samples) coordinates. All parameters are co-located with the corresponding stations and extracted at the month corresponding to the *Tara* Oceans sampling. To compensate for missing physicochemical samples in the *Tara* Oceans in situ dataset, climatological data (WOA13) are used. The correlation between in situ samples and corresponding values extracted from WOA13 are high (r^2: SST: 0.96, Sal: 0.83, Si: 0.97, NO_3: 0.83, PO_4: 0.89). In the absence of corresponding WOA13 data, a search is done within 2° around the sampling location and values found within this square are averaged.

Nutrients such as NO_3 and PO_4 display a strong collinearity when averaged over the global ocean (correlation of 0.95 in WOA13), which could complicate disentangling their respective contributions to niche definition. However, observations and experimental data allow identification of limiting nutrients at the regional scale, characterized by specific plankton communities[56]. The projection of niches into the future climate would yield spurious results if the present-day collinearity is not maintained[57], but there is up to now no evidence for large-scale changes in global nutrient stoichiometry[58].

ESM and bias correction. Outputs from six ESMs (Supplementary Table 2) are used to project environmental niches under greenhouse gas concentration trajectory RCP8.5[36]. Environmental drivers are extracted south of 60° N for present-day (2006–2015) and end-of-century (2090–2099) conditions for each model, and the multi-model mean is computed. A bias correction method, the cumulative distribution function transform (CDFt)[59], is applied to adjust the distributions of SST, Sal, Si, NO_3 and PO_4 of the multi-model mean to the WOA13 database. CDFt is based on a quantile mapping approach to reduce the bias between modelled and observed data while accounting for climate change. Therefore, CDFt does not rely on the stationarity hypothesis and present and future distributions can be different. CDFt is applied on the global fields of the mean model simulations. By construction, CDFt preserves the ranks of the simulations to be corrected. Thus, the spatial structures of the model fields are preserved.

Environmental niche models: training, validation and projections. Provinces with similar metagenomic content are retrieved from Richter et al.[11]. From a total of 48 initial provinces, 10 provinces are removed either because they are represented by too few samples (7 out of 10) or they are found in environments not resolved by ESMs (for example, lagoons of Pacific Ocean islands, 3 out of 10). This narrows down the number of samples from 644 to 595 metagenomes. Four machine learning methods are applied to compute environmental niches for each of the 38 provinces: gradient boosting machine (gbm)[60], random forest (rf)[61], single hidden layer neural networks (nn)[62] and generalized additive models (gam)[63]. Hyper parameters of each technique (except gam) are optimized. These are (1) for gbm, the interaction depth (1, 3 and 5), learning rate (0.01, 0.001) and the minimum number of observations in a tree node (1 to 10); (2) for rf, the number of trees (100 to 900 with step 200 and 1,000 to 9,000 with step 2,000) and the number of parameters used for each tree (1 to 7); (3) for nn, the number of neurons of the network (1 to 10) and the decay (1.10^{-4} to 9.10^{-4} and 1.10^{-5} to 9.10^{-5}). For gam the number of splines is set to 3, respectively 2 only when not enough points are available (for fraction 0–0.2, 65 points). R packages 'gbm' (2.1.3), 'randomForest' (4.6.14), 'mgcv' (1.8.16) and 'nnet' (7.3.12) are used for gbm, rf, nn and gam models.

To define the best combination of hyper parameters for each model, we perform random cross-validation by training the model on 85% of the dataset randomly sampled and by calculating the area under the curve[64] (AUC) on the 15% remaining points of the dataset. This process is repeated over 30 random subsets of the entire training set for each combination of hyperparameters. We work in a presence/absence framework, that is, for each province separately, the dataset consists of the variable to predict ('presence' or not of the province in the sample) and the predictors consisting of the environmental variables for each sample. A fixed probability threshold of 0.5 for presence/absence detection is used to calculate the AUC. Fixing the probability threshold allows optimization of all models according to this threshold so that the *dominant* province has a reasonably high probability of presence (at least in regions with similar environmental parameters to the training dataset) and for the four types of statistical model we use. The best combination of hyper parameters is the one for which the mean AUC over the cross-validation is the highest. A model is considered valid if at least three out of the four techniques have a mean AUC superior to 0.65, which is the case for 27 out of the 38 provinces (Supplementary Fig. 1a). A climatic annotation is given to the 27 validated niches (Supplementary Table 2). Final models are trained on the full dataset and only the techniques that have a mean AUC higher than 0.65 are considered to make the projections. The vast majority (23) of the 27 validated niches is validated by all 4 models and the remaining 4 niches by only 3 models. Relative influences of each parameter in defining environmental niches are calculated using the 'feature_importance' function from the DALEX R package[65] for all four statistical methods (Supplementary Fig. 16a). To evaluate the consistency and coherence of environmental niche models, we first make global projections on the 2006–2013 WOA2013 climatology. Projections are consistent with sampling regions for provinces encompassing vast oceanic areas. For example, the genomic province sampled in temperate Atlantic regions of size fraction 180–2,000 µm is projected to be present in the north and south temperate Atlantic but also other temperate regions (Supplementary Fig. 2). For model training and projections, physicochemical variables are scaled to have a mean of 0 and a variance of 1 (using means and standard deviations of the training data). This standardization procedure allows for better performance of neural networks models. Finally, as statistical models often disagree on projection sets whereas they give similar predictions on the training set (Supplementary Figs. 4, 5), we use the ensemble model approach for global-scale projections of provinces[66], that is, the mean projections of the validated machine learning techniques.

Combined size class provinces and ocean partitioning comparisons. To combine all size classes' provinces, we use the PHATE algorithm[35,67] from the R package 'phateR'. This algorithm allows visualization of high dimensional data in the requested number of dimensions while best preserving the global data structure[67]. We choose to train PHATE separately on WOA13 projections and present-day and end-of-century projections including presence probabilities of *non dominant* provinces. We use three dimensions and set hyper parameter k-nearest neighbours (knn) and decay, respectively, to 1,000 for WOA13 and 2,000 for model data as in this case there are twice as many points. The hyper parameter knn reflects the degree to which the mapping of PHATE from high to low dimensionality should respect the global features of the data. We argue that 1,000 and 2,000 are good choices as it will be sufficient to have a highly connected graph, conserve global structure, allow visualization of structures of the size of the provinces (mean number of points in a province: 4,867) and have a reasonable computational time. Decay is set to 20 in both cases. Then we cluster the resulting distance matrix using the k-medoïds algorithm[68], and the silhouette average width criteria[69] is used as an indicator of good fit. The silhouette criterion is maximal for two, three and four clusters, and two peaks are found at seven and 14 clusters (not shown). We choose

to present the four and seven cluster geographical patterns as they seem more relevant with respect to the resolutions of our environmental datasets (WOA13 and climate models). We compare the three polar clusters of the seven cluster geographical patterns with Antarctic Circumpolar Currents fronts[70] by overlying them on the map (black lines in Supplementary Fig. 3b).

To visualize the global biogeography structure, the resulting three vectors of PHATE are plotted using an RGB colour code. Each coordinate of each vector is respectively assigned to a given degree of colour component between 0 and 255 (8 bits red, green or blue) using the following formula (Supplementary Figs. 3 and 13):

$$C\_\mathrm{col}(i) = \frac{C(i) - \min(C(1), C(2), C(3))}{\max(C(1), C(2), C(3)) - \min(C(1), C(2), C(3))} \times 255 \quad (3)$$

$C(i)$ is the ith component of the PHATE axes and $C\_\mathrm{col}(i)$ is the corresponding degree of color of $C(i)$. Respectively, components 1, 2 and 3 are assigned to red, green and blue.

To compare the six size fraction biogeographies, the combined size class biogeography and existing biogeochemical partitions of the oceans[9,10], we use the adjusted rand index[71] (Supplementary Figs. 6–8) and overlay their boundaries above our partitions.

Centroids and migration shifts. The centroid of each province is defined as the average latitude and longitude for which the probability of presence is superior to 0.5 and weighted by both the probability of presence at each grid point and the grid cell area. The migration shift is calculated as the distance between the present-day and the end-of-the-century centroids considering the Earth as a perfect sphere of radius 6,371 km. For consistency (that is, to avoid long distance aberrant shifts), it is only calculated for provinces with an area of dominance larger than $10^6\,\mathrm{km}^2$ in the given basin.

Bray–Curtis dissimilarity index. Climate change impact on global projections is calculated at each grid point as the Bray–Curtis dissimilarity index[72,73] defined as follows:

$$\mathrm{BC} = \frac{\sum_n \left| P_n^{\mathrm{future}} - P_n^{\mathrm{present}} \right|}{\sum_n \left| P_n^{\mathrm{future}} + P_n^{\mathrm{present}} \right|} \quad (4)$$

where (P_n^{present} and P_n^{future}) are, respectively, the probability of presence of the province n in present day and at the end of the century. Only the probabilities of *dominant* provinces are non-null, and all others are set to zero. The mask of main fisheries[53] (chosen as the first four deciles) and exclusive economic zones[54] are overlaid on the Bray–Curtis map.

Change in province assemblages. A province assemblage is defined as the assemblage of *dominant* provinces of each size fraction at a given grid point of the considered ocean. We consider two criteria of change in province assemblage between present-day and end-of-the-century conditions. The first one, more straightforward and less stringent, considers that a province assemblage occurs when a change of *dominant* province is found in at least one size fraction. In a more stringent way, a change of assemblage is considered important for $\mathrm{BC} > \frac{1}{6}$ (previous section). This threshold corresponds to an idealized case where each *dominant* province has a probability of 1, and a change of *dominant* province is found in only one size fraction. This criterion allows us to discard assemblage changes for which the changes in probability of presence of *dominant* provinces are very low.

Differential genomic composition across provinces. We characterize the composition in marine Hexanauplia (copepods), marine diazotrophs and phototrophs of provinces by considering the mean relative abundances of groups of MAGs[31,32] characterized taxonomically and by size (for copepods and phototrophs) in each province for all size fractions except the viral one and for size fractions >5 µm (for marine Hexanauplia). Marine Hexanauplia are annotated taxonomically (either as belonging to clade A (109 MAGs) or clade B (105 MAGs)) from which 198 are found in at least five samples that we use (98 clade A and 100 clade B). To attribute a preferential size class to these MAGs, mean relative abundances over all sites of each of them are compared across size fraction 180–2,000 µm, 20–180 µm and 5–20 µm using Welch ANOVA[74] (p value <0.05). When the test is significant, the MAG is annotated to its preferential size class (Meso-zooplankton: 180–2,000 µm, Micro-zooplankton: 20–180 µm or 5–20 µm). When the test is not significant, the MAG is annotated as unclassified. Then for each group of MAG (defined by the preferential size class plus the clade), the mean of the sum over the MAGs from this group is calculated in each province to characterize the province. The same procedure is applied for 27 out of 48 bacterial diazotrophs (present in at least five samples) from the prokaryotic MAG collection[32], distinguishing groups at the phylum level (Gammaproteobacteria $n=8$, Cyanobacteria $n=8$, Deltaproteobacteria $n=2$, Alphaproteobacteria $n=4$, Planctomycetes $n=3$, Verrucomicrobiota $n=2$) and for 231 phototrophs (present in at least five samples) distinguishing algae ($n=172$), diatoms (Bacillariophyta, $n=11$) and cyanobacteria ($n=48$) and their preferential size class (pico: 0.22–3 µm, nano: 0.8–5 µm and 5–20 µm and micro: 20–180 µm and 180–2,000 µm). Finally, significant differential composition between provinces are annotated using Holm corrected[75] pairwise Mann–Whitney U test[76] (p <0.05 for significance).

Estimation of carbon export fluxes in assemblages of provinces. We use three sets of carbon export data at 100 m depth (extrapolated data)[3,43,44] to estimate an average particulate organic carbon (POC) flux for each assemblage of *climato-genomic* provinces. These extrapolations are based on estimates of carbon export efficiency from in situ measurements either using the POC/^{234}Th radioisotope ratio[3] or the *f*-ratio technique[43,44]. They both derive a relation between carbon export efficiency and SST. Then to obtain extrapolations of carbon export, these ratios are multiplied by estimates of primary production from satellite estimates of sea surface chlorophyll concentrations. For each dataset, we associate each extrapolation of POC fluxes to the assemblage of present-day provinces from the nearest point of the flux estimate. An average flux is associated with an assemblage if at least three data points are associated with that assemblage. Each grid point is associated to a given assemblage and is assigned to its average carbon flux. Thus, we calculate carbon export estimates where an assemblage from the set of present-day assemblages can be projected in the future. Next we calculate total carbon export for current and projected end-of-century assemblages, summing fluxes over these areas (Supplementary Table 9). Similar conclusions are reached using median values of carbon export. We verified that using whole assemblages as predictors of carbon export is more robust than using single size fraction dominant provinces (ANOVA[77], Akaike Information Criterion[78]).

Linking changes in carbon export and compositional changes. We use the Apriori algorithm[45] to identify associations between changes in community composition and changes in carbon export. From the relative abundance values of phototroph, diazotroph and copepod MAGs[31,32], we consider only significant changes in their composition among communities by a pairwise Wilcoxon test (Holm correction p <0.05). We also consider only phototrophic and copepod MAGs for which a preferred size class was found. The Apriori algorithm determines association rules between sets of 'transactions', here community changes with carbon export. The algorithm identifies sufficiently frequent associations between transactions and computes a lift as follows:

$$\mathrm{lift}(A \rightarrow B) = \frac{\mathrm{confidence}(A \rightarrow B)}{\mathrm{support}(B)} \quad (5)$$

where $\mathrm{confidence}(A \rightarrow B)$ is the number of transactions where both A and B occur divided by the number of transactions where A occurs, and $\mathrm{support}(B)$ is the percentage of transactions where B occurs in the whole dataset. The lift corresponds to the increase factor in likeliness that B occurs when A occurs. The Apriori algorithm is launched for four latitudinal zones of the ocean: equatorial, subtropical (north and south) and temperate/subpolar zones. Only the directionality of changes is considered (increase/decrease in either carbon flux or a type of organism). If rules overlap perfectly, only the largest association rules are kept (maximum rule length is set to 7). Only association rules pointing towards changes in carbon fluxes are kept (A = change(s) in one or several plankton group and B = change in carbon export in equation (5)). Minimum support is set to 0.05.

Driver analysis. To assess the relative importance of each driver in province changes, the methodology from Barton et al.[22] is adopted. For a set N of n provinces (individual provinces or all provinces together), the probability of presence of each province is recalculated for present-day conditions except for driver d (from the set of drivers D) for which the end-of-the-century condition is used ($P_n^{\text{future for } d\text{th driver only}}$). The set of driver D can be either all drivers (Fig. 5a,c) or all drivers except SST (Fig. 5b). The relative importance of driver d at a given grid point for the set of N of provinces is computed as follows:

$$\mathrm{RI}(d) = \frac{\sum_{n\in N} \left| P_n^{\text{future for } d\text{th driver only}} - P_n^{\mathrm{present}} \right|}{\sum_{d\in D} \sum_{n\in N} \left| P_n^{\text{future for } d\text{th driver only}} - P_n^{\mathrm{present}} \right|} \quad (6)$$

RI(d) is computed at grid cells where $\mathrm{BC} > \frac{1}{6}$ and is calculated with either the set of all drivers (Fig. 5a,c) or all drivers except SST (Fig. 5b). When RI(d) is calculated for individual provinces (Fig. 5c and Supplementary Fig. 18d,e), it is computed only at grid cells where $\mathrm{BC} > \frac{1}{6}$ and the concerned province is either *dominant* in present-day and/or end-of-century conditions.

Maps. We use the continents map background from R package 'maps'[51] for all map figures in this study.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

All data used are available at https://github.com/institut-de-genomique/NCLIM-20102618B. All coordinates of ocean partitionings from this study

are available at https://figshare.com/articles/dataset/Biogeographies_genomic_provinces/19071620[79].

Code Availability

All codes used are available at https://github.com/institut-de-genomique/NCLIM-20102618B.

References

52. Jaccard, P. Distribution comparée de la flore alpine dans quelques régions des Alpes occidentales et orientales. *Bull. Murith.* **31**, 81–92 (1902).
53. Watson, R. A. A database of global marine commercial, small-scale, illegal and unreported fisheries catch 1950–2014. *Sci. Data* https://doi.org/10.1038/sdata.2017.39 (2017).
54. *Maritime Boundaries Geodatabase: Maritime Boundaries and Exclusive Economic Zones (200NM), version 11* (Flanders Marine Institute, 2019); https://doi.org/10.14284/386
55. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L. & Gehlen, M. PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geosci. Model Dev.* https://doi.org/10.5194/gmd-8-2465-2015 (2015).
56. Bibby, T. S. & Moore, C. M. Silicate:nitrate ratios of upwelled waters control the phytoplankton community sustained by mesoscale eddies in sub-tropical North Atlantic and Pacific. *Biogeosciences* https://doi.org/10.5194/bg-8-657-2011 (2011).
57. Brun, P., Kiørboe, T., Licandro, P. & Payne, M. R. The predictive skill of species distribution models for plankton in a changing climate. *Glob. Change Biol.* https://doi.org/10.1111/gcb.13274 (2016).
58. Redfield, A. C. in *James Johnstone Memorial Volume* (ed. Daniel, R. J.) 176–192 (Liverpool Univ. Press, 1934).
59. Michelangeli, P. A., Vrac, M. & Loukos, H. Probabilistic downscaling approaches: application to wind cumulative distribution functions. *Geophys. Res. Lett.* https://doi.org/10.1029/2009GL038401 (2009).
60. Ridgeway, G. gbm: Generalized boosted regression models. R version 1.6–3.1 https://cran.r-project.org/web/packages/gbm/gbm.pdf (2010).
61. Breiman, L. & Cutler, A. randomForest: Breiman and Cutler's random forests for classification and regression. R package 4.1.0 https://www.stat.berkeley.edu/~breiman/RandomForests/ (2012).
62. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* 4th edn (Springer, 2002).
63. Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* https://doi.org/10.1198/016214504000000980 (2004).
64. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* https://doi.org/10.1016/j.patrec.2005.10.010 (2006).
65. Biecek, P. DALEX: explainers for complex predictive models. *J. Mach. Learn. Res.* **19**, 1–5 (2018).
66. Jones, M. C. & Cheung, W. W. L. Multi-model ensemble projections of climate change effects on global marine biodiversity. *ICES J. Mar. Sci.* https://doi.org/10.1093/icesjms/fsu172 (2015).
67. Vallejos, C. A. Exploring a world of a thousand dimensions. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0330-9 (2019).
68. Kaufman, L. and Rousseeuw, P.J. in *Statistical Data Analysis Based on the L1 Norm and Related Methods* (ed. Dodge, Y.) 405–416 (North-Holland, 1987).
69. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
70. Orsi, A. H., Whitworth, T. & Nowlin, W. D. On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep Sea Res. Part I* https://doi.org/10.1016/0967-0637(95)00021-W (1995).
71. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* https://doi.org/10.1007/BF01908075 (1985).
72. Somerfield, P. J. Identification of the Bray–Curtis similarity index: comment on Yoshioka (2008). *Mar. Ecol. Prog. Ser.* https://doi.org/10.3354/meps07841 (2008).
73. Bloom, S. Similarity indices in community studies: potential pitfalls. *Mar. Ecol. Prog. Ser.* https://doi.org/10.3354/meps005125 (1981).
74. Welch, B. L. The generalisation of student's problems when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
75. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
76. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
77. Sthle, L. & Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **6**, 259–272 (1989).
78. Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345–370 (1987).
79. Frémont, P. et al. Biogeographies of genomic provinces from 'Restructuring of plankton genomic biogeography in the surface ocean under climate change'. figshare. https://figshare.com/articles/dataset/Biogeographies_genomic_provinces/19071620 (2022).

Acknowledgements

P.F. was supported by a CFR doctoral fellowship and the NEOGEN impulsion grant from the Direction de la Recherche Fondamentale of the CEA. This study received funding from the European Union's Horizon 2020 Blue Growth research and innovation programme under grant agreement number 862923 (project AtlantECO), ATIGE Genopole postdoctoral fellowship (T.O.D.), HYDROGEN/ANR-14-CE23–0001 (T.O.D.) and ANR-11-IDEX-0004–17-EURE-0006. M.G. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 820989 (project COMFORT). This study benefited from access to high-performance computing resources through GENCI- [TGCC/CINES/IDRIS] and the ESPRI computing and data centre (https://mesocentre.ipsl.fr), which is supported by CNRS, Sorbonne Université, Ecole Polytechnique and CNES and through national and international grants. We thank the commitment of the Research Federation for the Study of Global Ocean Systems Ecology and Evolution (FR2022/TaraGOSEE) and of Stazione Zoologica Anton Dohrn. We thank T. Roy for preparation of the climatic data, S. Henson for providing carbon export data, LAGE (Laboratoire d'Analyses Génomiques des Eucaryotes, CEA) members for stimulating discussions on this project, M. Mariadassou, S.D. Ayata and B.H. Mele for discussions on statistics and climate envelope models, C. Scarpelli and members of the scientific computation team from Genoscope for support on computations, L. Bopp for initial discussions on this project and on climate models and N. Le Bescot (TernogDesign) for help with the figures. We thank all members of the *Tara* Oceans consortium for maintaining a creative environment and for their constructive criticism. *Tara* Oceans would not exist without the *Tara* Ocean Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/).

This article is contribution number 128 of *Tara* Oceans.

Author contributions

P.F., M.G. and O.J. conceived the study. P.F. computed the results and compiled and analysed the data. M.G., O.J. and J.L. conducted a preliminary study. M.V. wrote the bias correction algorithm. P.F. wrote the initial draft of the paper. T.O.D., P.F., M.G., O.J., D.I., J.L., M.V. and P.W. discussed the results and contributed to writing the paper.

Competing interests

The authors declare no competing interests.

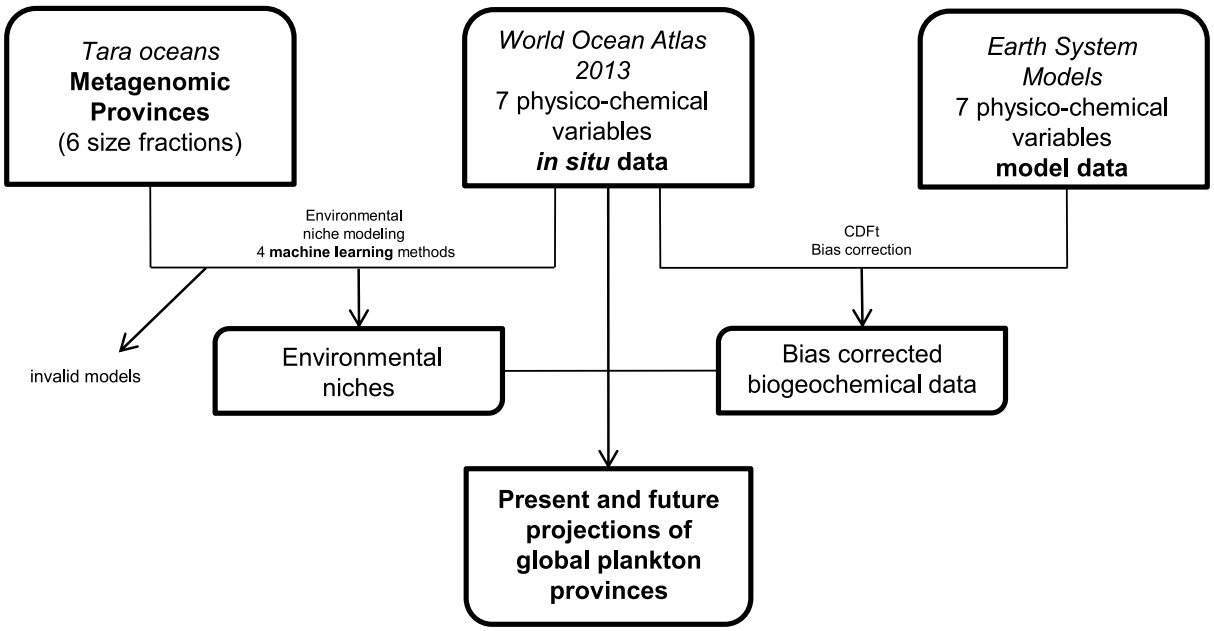
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41558-022-01314-8.

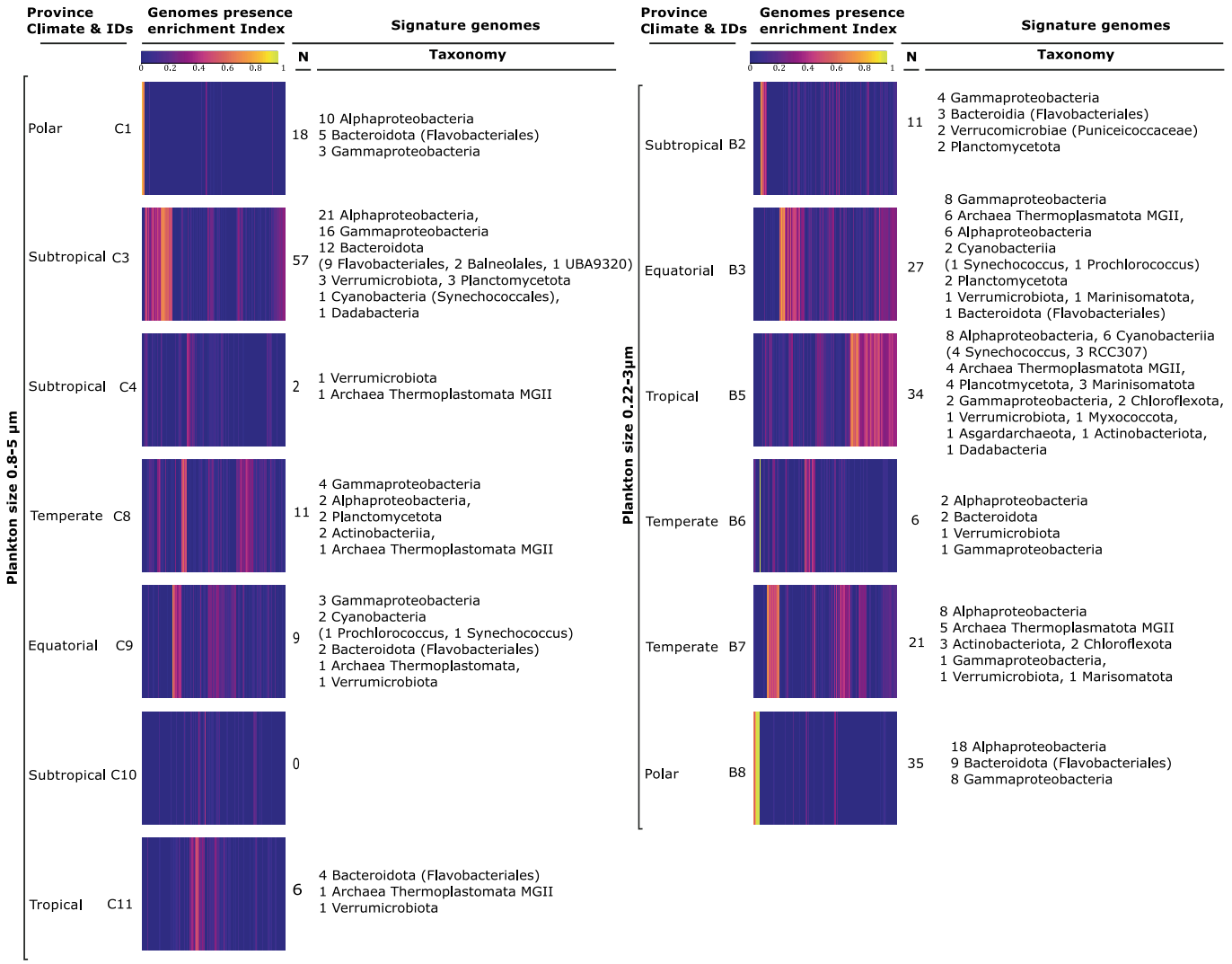
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41558-022-01314-8.

Correspondence and requests for materials should be addressed to Paul Frémont, Marion Gehlen or Olivier Jaillon.

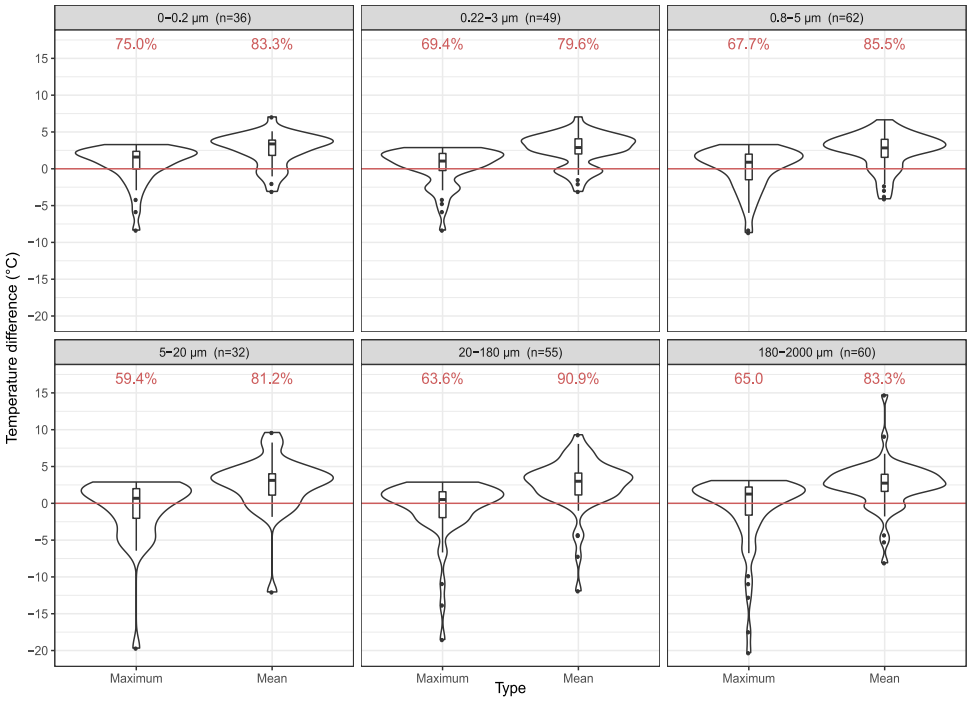
Peer review information *Nature Climate Change* thanks Levente Bodrossy, Robert Ptacnik and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



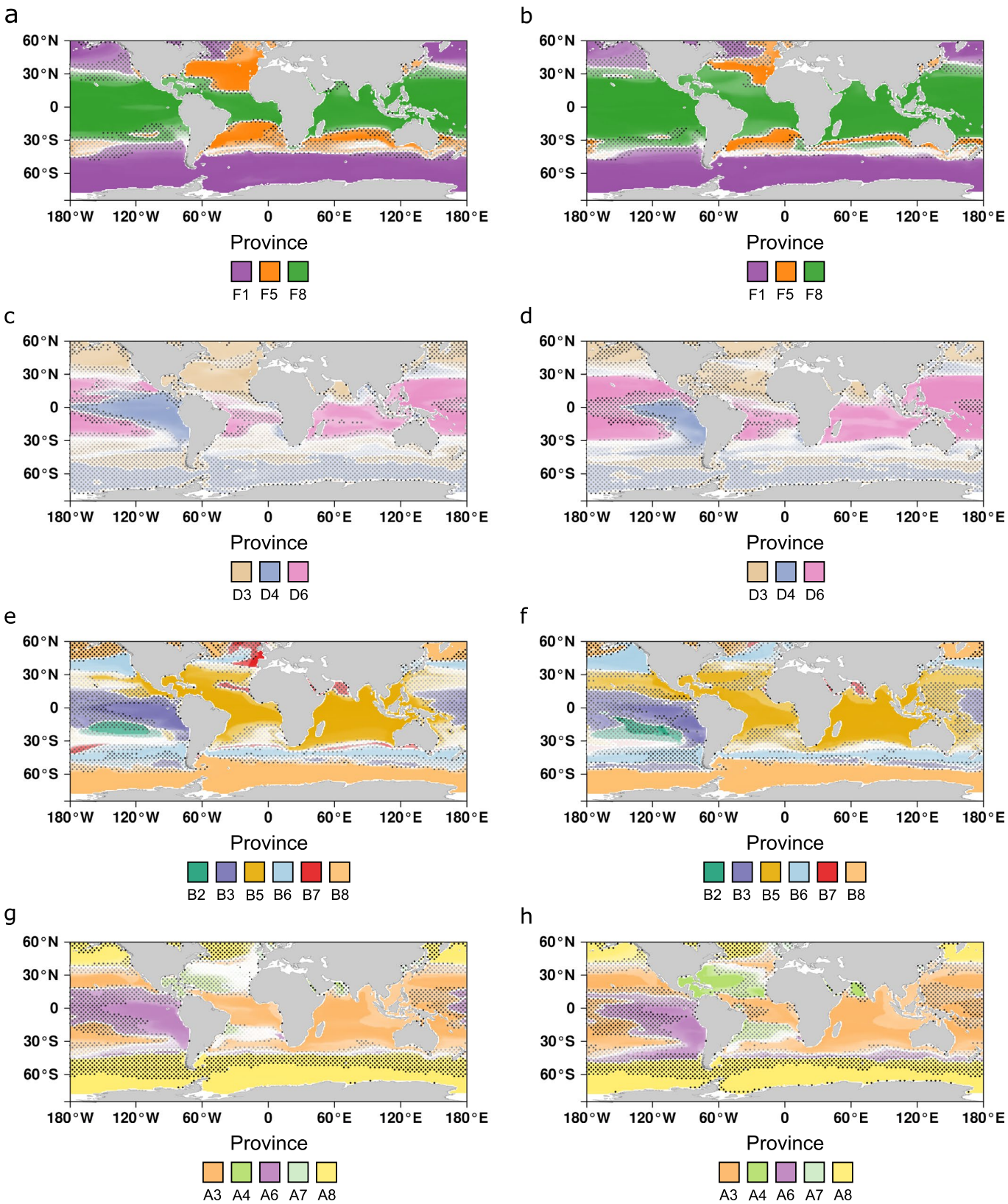
Extended Data Fig. 1 | Study pipeline. Metagenomic data from the 2009–2013 Tara Oceans expedition and in situ measurements of physicochemical variables (World Ocean Atlas 2013, WOA13)[33] are combined to define environmental niches at the plankton community level across 6 size fractions. Bias corrected outputs from a mean model of 6 Earth System Models (Supplementary Table 1) and WOA13 data are then used to project global plankton provinces for present day and end of the century conditions under a high warming scenario (RCP8.5)[36]. Variables are Sea Surface Temperature (SST), Salinity (Sal), Dissolved silica (Si), Nitrate (NO_3), Phosphate (PO_4), Iron (Fe) and a seasonality index of nitrate (SI NO_3).



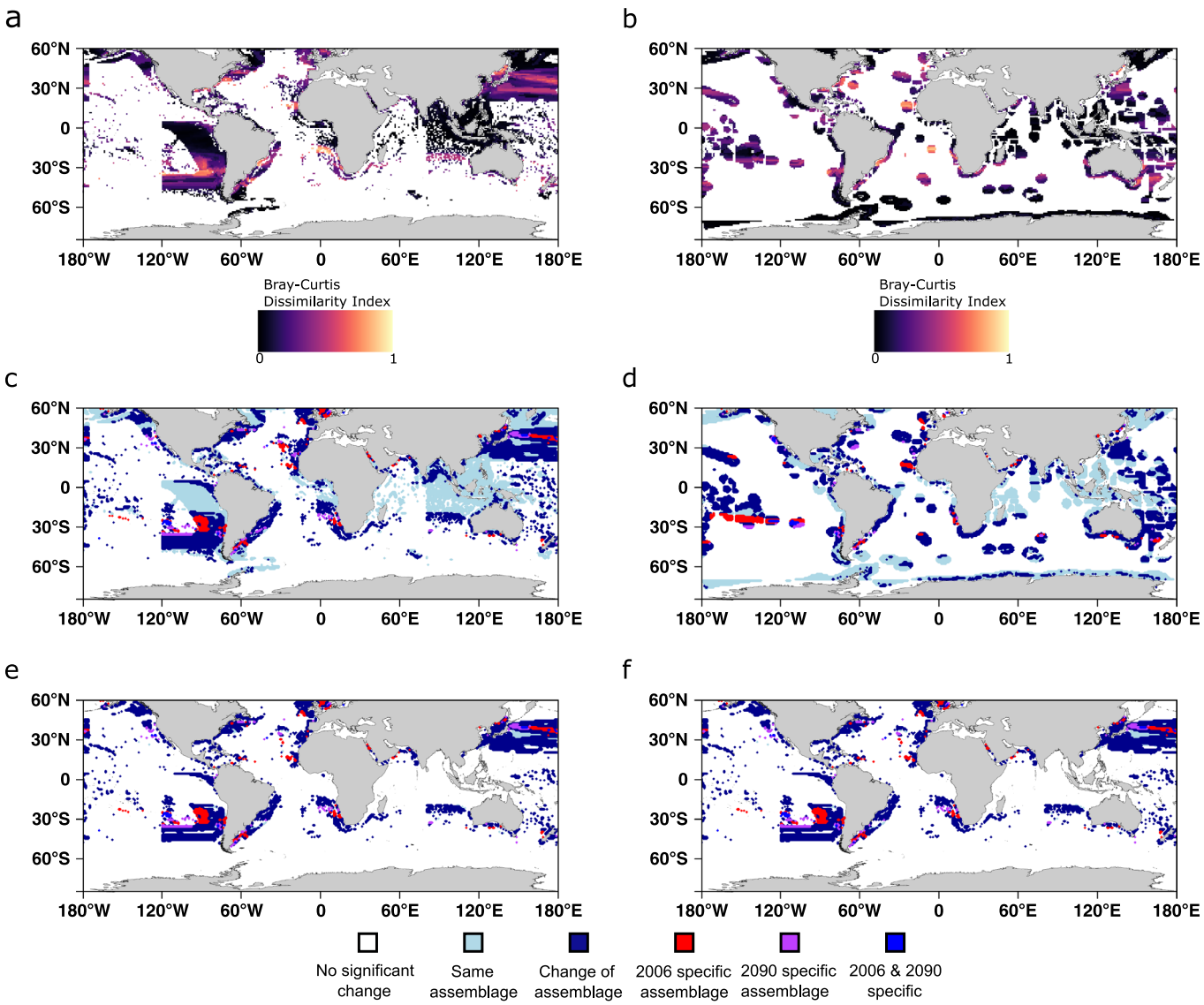
Extended Data Fig. 2 | Prokaryotic signature genomes of provinces of the prokaryote (0.22–3 µm) and protist (0.8–5 µm) enriched size classes. Indexes of presence enrichment[52] for 1888 genomes of prokaryotic plankton[32] in corresponding provinces are clustered and represented in a colour scale. Signature genomes (see Methods) are found for almost all provinces, their number and taxonomies are summarized (detailed list in Supplementary Table 6). A genome is considered to be signature of a province if the presence enrichment index is superior to 0.5 with this province and inferior to 0.1 for all other provinces of the given size class.



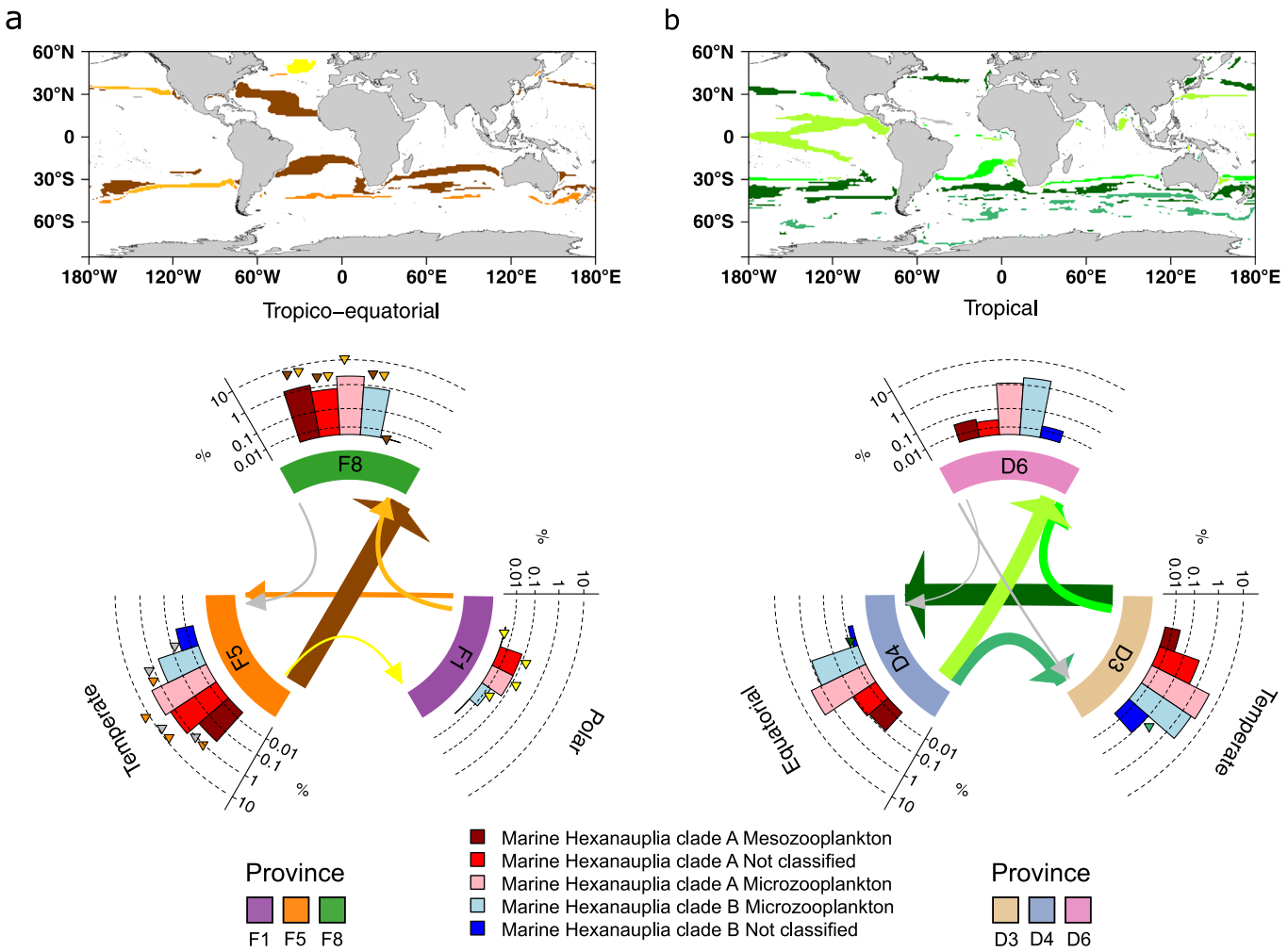
Extended Data Fig. 3 | Distribution of deltas between future temperature at each sampling site (surface) minus either the mean or maximum temperature within their contemporary genomic province. For most of the sites and across size fractions the future temperature projected by the bias adjusted ESM ensemble model is higher than both the maximum and mean contemporary temperature of their genomic province.



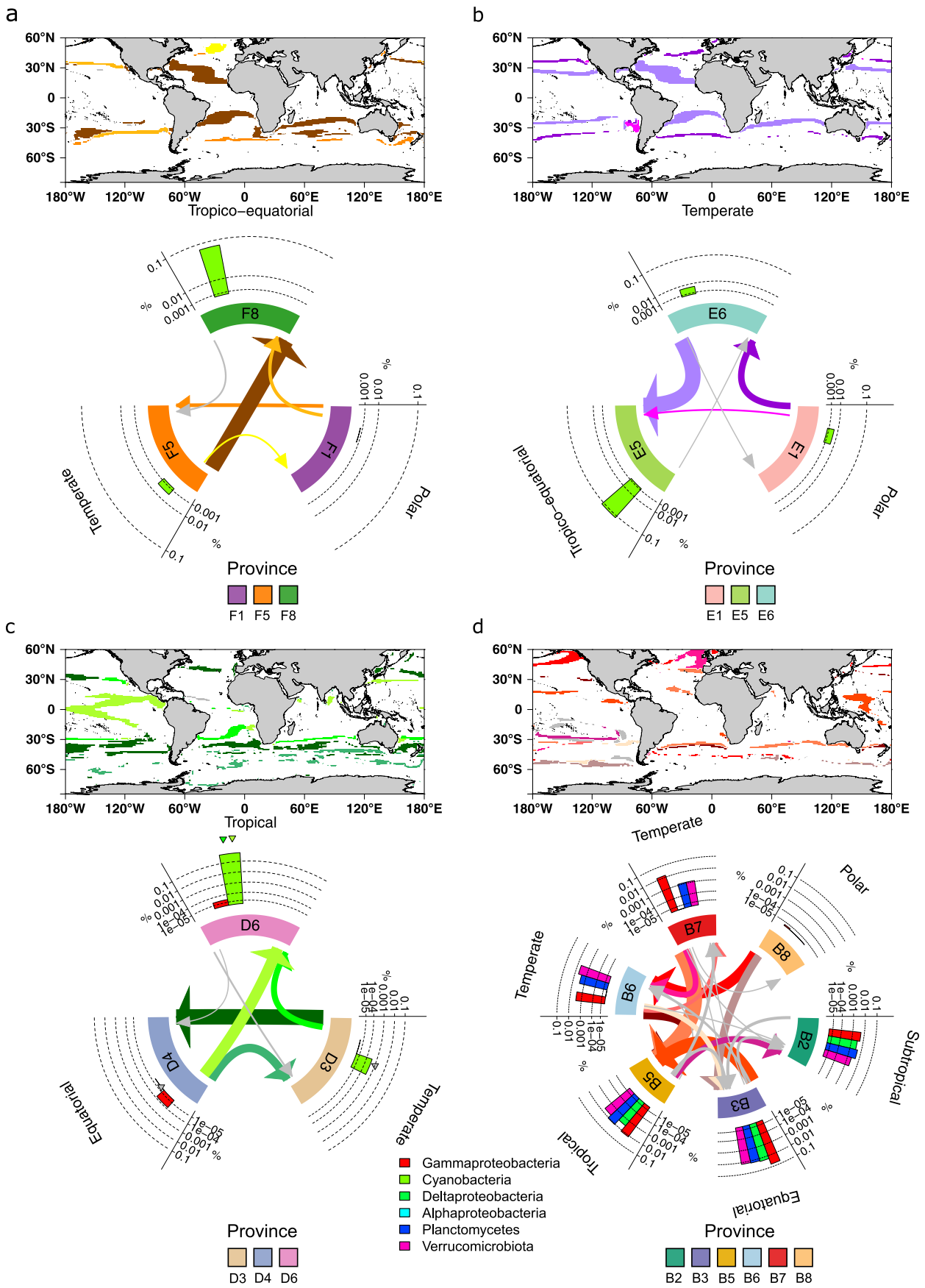
Extended Data Fig. 4 | Global geographical patterns for provinces of four plankton size fractions in present day and at the end of the century. (**a, c, e, g**) Present day and (**b, d, f, h**) end of century biogeographies of size classes 180–2000, 5–20, 0.22–3 and 0–0.2 μm respectively. At each grid point of the maps the *dominant* province is represented using a darkness of colour proportional to its presence probability. Dots represent areas of uncertainty (where the delta of probability between the *dominant* and another province is inferior to 0.5). Expansion of tropical provinces and shrinkage of temperate provinces are consistently projected in all size fractions. We generated these map using R-package maps[51].



Extended Data Fig. 5 | Bray-Curtis dissimilarity index and assemblage change maps comparing present day with end of the century projections of *dominant* provinces in principal fisheries[53] (4 last deciles) and Exclusive Economic Zones[54]. Bray-Curtis dissimilarity index and assemblage changes in (**a, c**) Principal fisheries and (**b, d**) Exclusive Economic Zones. Assemblage changes in (**e**) Principal fisheries and (**f**) Exclusive Economic Zones in areas projected to encounter an important change (Bray-Curtis dissimilarity index superior to 1/6). We generated these map using R-package maps[51].

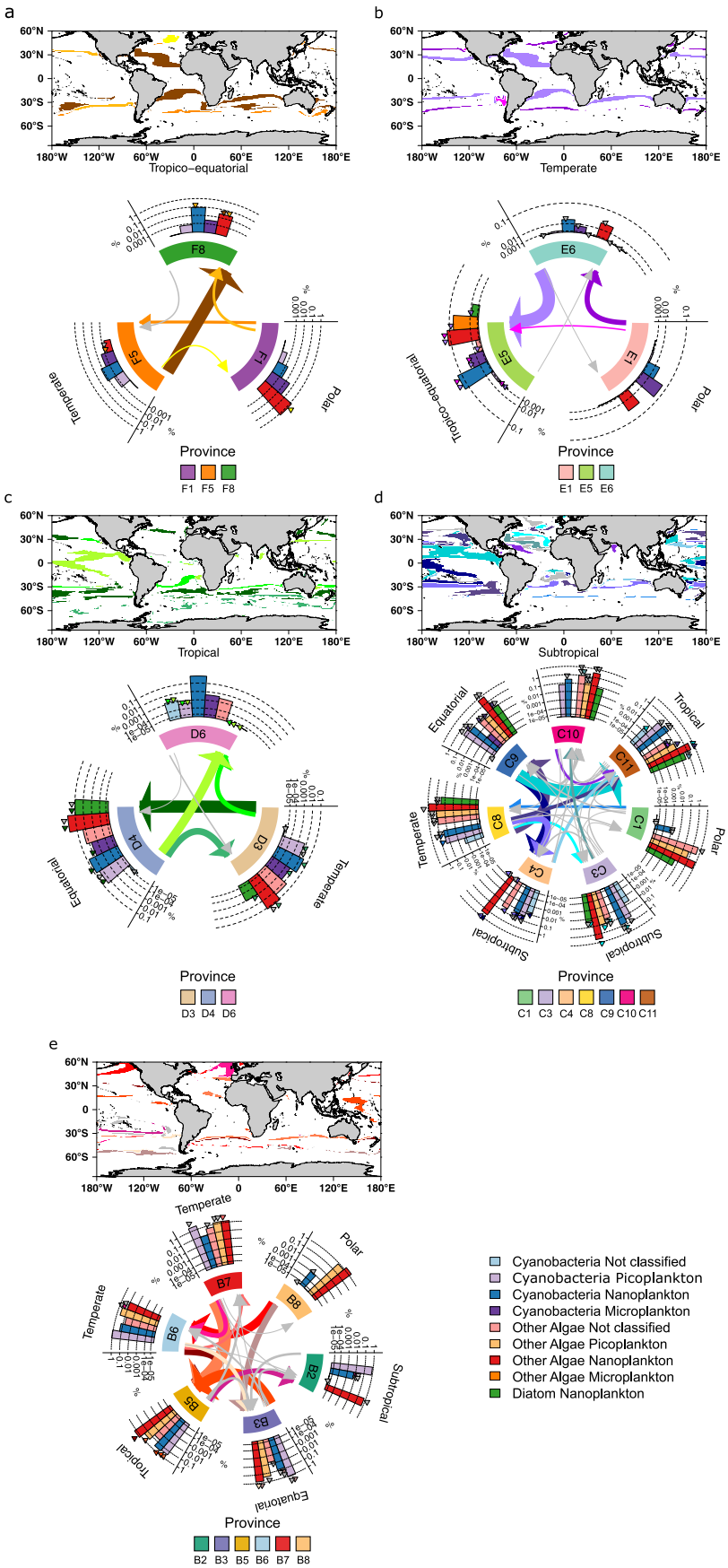


Extended Data Fig. 6 | Projected compositional shifts in marine hexanauplia in areas of *dominant* province change. (**a**) 180–2000 μm and (**b**) 5–20 μm. Top: Locations of *dominant* province change using colours corresponding to the type of province transition. Bottom: Circular plots summarizing significant compositional shifts in marine hexanauplia classified by size ('not classified' when no preferential size class is found). Each type of transition is represented by an arrow coloured according to the map and in grey if they represent less than 2% of the transitions. Barplots represent mean relative abundances of each group of organism. Arrows point towards the end of the century projected province and their widths are proportional to the area of change. Significant compositional changes in a type of organism are represented by triangles of the associated transition colour. We generated these map using R-package maps[51].



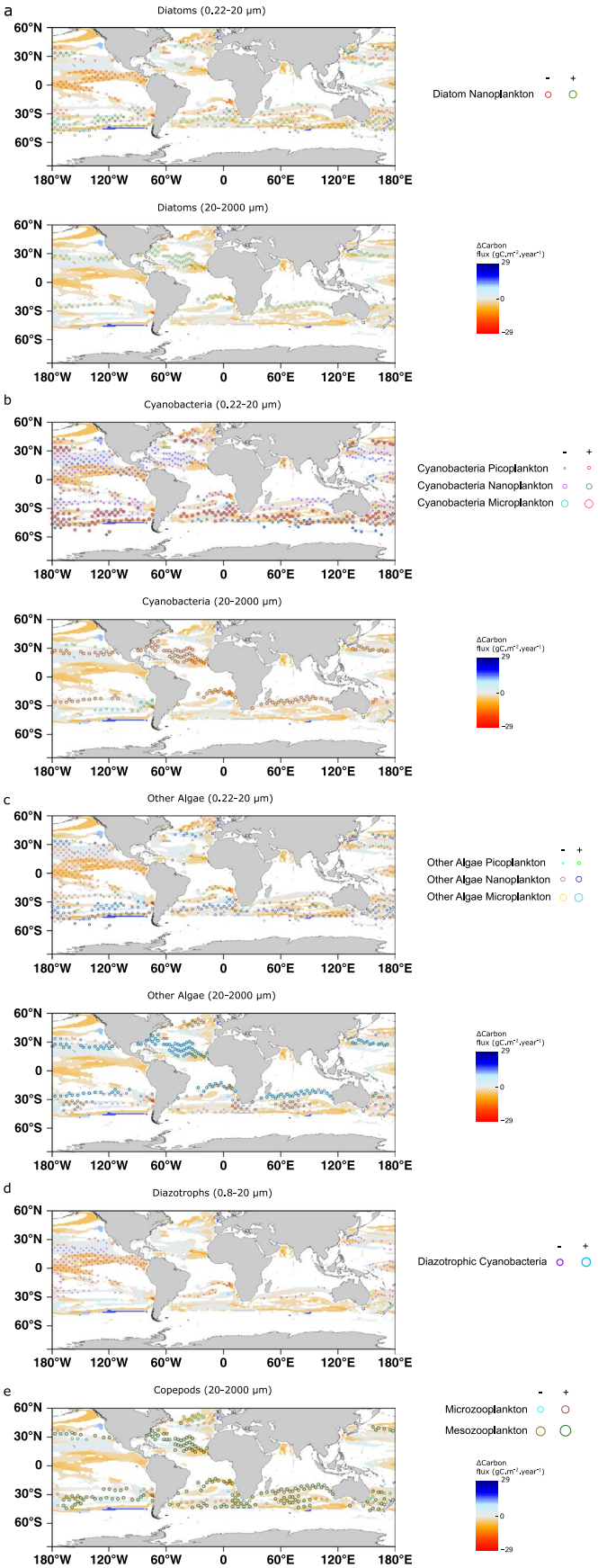
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Projected compositional shifts in bacterial diazotrophs in areas of dominant community change. (**a**) 180–2000 μm (**b**) 20–180 μm (**c**) 5–20 μm and (**d**) 0.22–3 μm. Top: Locations of *dominant* province change using colours corresponding to the type of province transition. Bottom: Circular plots summarizing significant compositional shifts in marine diazotrophs. Each type of transition is represented by an arrow coloured according to the map and in grey if they represent less than 2% of the transitions. Barplots represent mean relative abundances of each group of organism. Arrows points towards the end of the century projected province and their widths are proportional to the area of change. Significant compositional changes in a type of organism are represented by triangles of the associated transition. We generated these map using R-package maps[51].



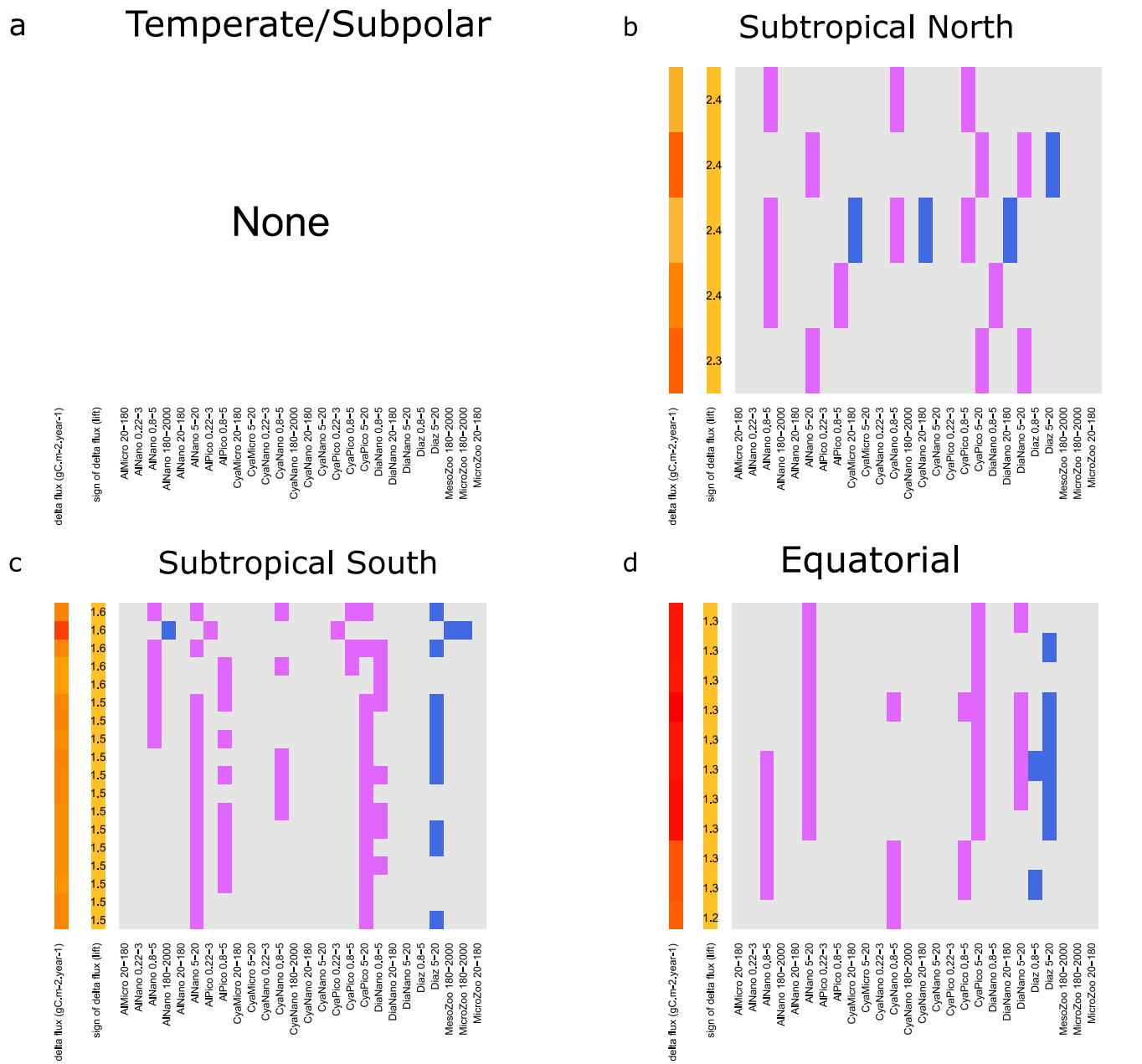
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Projected compositional changes in phototrophs in areas of dominant community change. (**a**) 180–2000 μm (**b**) 20–180 μm (**c**) 5–20 μm (**d**) 0.8–5 μm and (**e**) 0.22–3 μm. Top: Locations of *dominant* province change using colours corresponding to the type of province transition. Bottom: Circular plots summarizing significant compositional shifts in phototrophs classified by size ('not classified' when no preferential size class is found). Each type of transition is represented by an arrow coloured according to the map and in grey if they represent less than 2% of the transitions. Barplots represent mean relative abundances of each group of organism. Arrows point towards the end of the century projected province and their widths are proportional to the area of change. Significant compositional changes in a type of organism are represented by triangles of the associated transition colour. We generated these map using R-package maps[51].



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Maps of carbon export flux changes in link with organisms' projected changes. Significant composition changes based on genomes relative abundances are represented for phototrophs, marine nitrogen fixers (Diazotrophic cyanobacteria) and copepods. For each map, transitions from several characteristic size classes are represented (**a**) Top: Diatoms 0.22–20 µm. Bottom: Diatoms 20–2000 µm (**b**) Top: Cyanobacteria 0.22–20 µm. Bottom: Cyanobacteria 20–2000 µm (**c**) Top: Other Algae 0.22–20 µm. Bottom: Other Algae 20–2000 µm (**d**) Diazotrophs 0.8–20 µm (**e**) Copepods 20–2000 µm. We generated these map using R-package maps[51].



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Association rules between changes in carbon flux and changes in organism relative abundances. Association rules in (**a**) temperate/subpolar (latitude > 40° (<−40°)), (**b**) subtropical North (20° to 40° latitude), (**c**) subtropical South (−20° to −40° latitude) and (**d**) equatorial regions (-20° to 20° latitude). Each line represents an association rule between a change in carbon export found by the Apriori algorithm[45] (first column: mean change in carbon export and second column: sign of the change and lift of the rule (equation 5). The other columns represent the changes in community composition (red: decrease of the given group, green: increase) associated with this change in carbon export.

nature research

Corresponding author(s): Paul Frémont, Marion Gehlen, Olivier Jaillon

Last updated by author(s): Feb 1, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	No custom algorithms or software central to the research. Custom code in R/3.3.1 was used to analyze the data using multiple packages: gbm; randomForest; mdcv; nnet; dismo; FactoMineR; factoextra; readxl; ggplot2;matlab; reshape2; gplots; plotly; stringr; caret; mapproj; mapplots; SDMTools; RColoBrewer; ncdf4; CDFt; plotrix; png; grid; DALEX (R/3.6.0); ggalluvial; stringr; isofor; parallel; scales; Rtsne; sm; scatterplot3d; imager; ingredients; VennDiagramm; tidygraph; ggraph; igraph; animation; VennDiagram

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data and codes used are available at https://github.com/institut-de-genomique/NCLIM-20102618B. All coordinates of ocean partitionings from this study are available at https://figshare.com/articles/dataset/Biogeographies_genomic_provinces/19071620.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study is a multivariate analysis of previously computed 48 clusters (from a related collection of 644 metagenomes) using a niche modelling approach (Neural networks, Random Forest, Gradient Boosting Machine, and Generalized Additive Models) using a set of 6 environmental variables (Temperature, Salinity, Dissolved Silica, Nitrate, Phosphate, modelled iron) and a seasonality index of nitrate. Environmental niches are then projected on WOA13 climatologies and 10 years Earth System Models climatologies in present day (2006-15) and at the end of the century (2090-2099).
Research sample	Collection of 48 clusters of 644 related metagenomes sampled during the Tara Oceans expeditions (2010-2012) representing 6 major plankton groups fractionated by size from 0 to 2000 micrometers (Nature Reviews 378 Microbiology (2020). doi:10.1038/s41579-020-0364-5).
Sampling strategy	Metagenomes used to build the clusters were previously sequenced with Illumina technology at high coverage rates per sample for all size fractions. Described in several other papers recently reviewed (Nature Reviews 378 Microbiology (2020). doi:10.1038/s41579-020-0364-5).
Data collection	Samples were previously collected at 102 sites (Tara Oceans stations) for up to six size fractions (0-0.2, 0.22-3, 0.8-5, 5-20, 20-180, 180-2000 µm) and two depths (subsurface (SUR) and deep chlorophyll maximum (DCM)). Described in several other papers recently reviewed (Nature Reviews 378 Microbiology (2020). doi:10.1038/s41579-020-0364-5). World Ocean Atlas data was collected from https://www.nodc.noaa.gov/cgi-bin/OC5/woa13/woa13.pl at 1° resolution (annual mean and monthly analyzed mean over the period 2005-2012 were used). Earth System Models climatologies were provided by the Laboratoire des Sciences du Climat et de l'Environmment (LSCE).
Timing and spatial scale	Tara Oceans samples were collected during a three years expedition from 2010 to 2012. Sampling stations considered in the present study are on average 300 km apart. They cover the vast majority of temperate, tropical and equatorial oceans and a small part of the antarctic ocean south of South America. Arctic regions were not considered. Described in several other papers recently reviewed (Nature Reviews 378 Microbiology (2020). doi:10.1038/s41579-020-0364-5).
Data exclusions	10 clusters of metagenomes were excluded from the analysis. As mentioned in the article either because they are represented by too few samples (7 out of 10) or they are found in environments not resolved by Earth System Models (e.g. lagoons of Pacific Ocean islands, 3 out of 10). This narrows down the number of samples from 644 to 595 metagenomes.
Reproducibility	This study is purely computational analysis of public data. Source codes are available as stated in the manuscript.
Randomization	Niche modelling validation includes a cross validation analysis by random computational subsampling of the dataset.
Blinding	No prior knowledge of the taxonomic content of the metagenomic clusters was used.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Human research participants
☐	☐ Clinical data
☐	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) — *State the source of each cell line used.*

Authentication — *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination — *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) — *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals — *For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Field-collected samples — *For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics — *Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*

Recruitment — *Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*

Ethics oversight — *Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and*

Sequencing depth	*whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software — *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization — *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template — *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal — *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring — *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings — *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested — *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) — *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction — *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity — *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis — *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis — *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com