Global drivers of eukaryotic plankton biogeography in the sunlit ocean

Authors: Sommeria-Klein, Guilhem[1*]; Watteaux, Romain[2]; Iudicone, Daniele[2]; Bowler, Chris[1]; Morlon, Hélène[1]

[1]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS (UMR 8197, INSERM U1024), 46 rue d'Ulm, F-75005 Paris, France
[2]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

*Corresponding author: guilhem.sk@gmail.com

Short abstract: Eukaryotic plankton are a core component of marine ecosystems with exceptional taxonomic and ecological diversity. Yet how their ecology interacts with the environment to drive global distribution patterns is poorly understood. Here, we use *Tara* Oceans metabarcoding data covering all the major ocean basins combined with a probabilistic model of taxon co-occurrence to compare the biogeography of 70 major groups of eukaryotic plankton. We uncover two main axes of biogeographic variation. First, more diverse groups display stronger biogeographic structure. Second, large-bodied consumers are structured by oceanic basins, mostly via the main currents, while small-bodied phototrophs are structured by latitude, with a comparatively stronger influence of biotic conditions. Our study highlights striking differences in biogeographies across plankton groups and disentangles their determinants at the global scale.

One-sentence summary: Eukaryotic plankton biogeography and its determinants at global scale reflect differences in ecology and body size.

Main text: Marine plankton communities play key ecological roles at the base of oceanic food chains, and in driving global biogeochemical fluxes (Field, Behrenfeld, Randerson, & Falkowski, 1998; Worden et al., 2015). Understanding their spatial patterns of distribution is a long-standing challenge in marine ecology that has lately become a key part of the effort to model the response of oceans to environmental changes (Beaugrand & Kirby, 2018; Raes et al., 2018; Righetti, Vogt, Gruber, Psomas, & Zimmermann, 2019; Tittensor et al., 2010). Part of the difficulty lies in the constant mixing of water masses and hence plankton communities by ocean currents (Jönsson & Watson, 2016). Recent planetary-scale ocean sampling expeditions have revealed that eukaryotic plankton are taxonomically and ecologically extremely diverse, possibly even more so than prokaryotic plankton (de Vargas et al., 2015). Eukaryotic plankton range from pico-sized (0.2-2 mm) to meso-sized (0.2-20 mm) organisms and larger, thus covering an exceptional range of sizes. Eukaryotic plankton also cover a wide range of ecological roles, from phototrophs (e.g., Bacillariophyta, Haptophyta, Mamiellophyceae) to parasites (e.g., Marine Alveolates or MALVs), and from heterotrophic protists (e.g., Diplonemida, Ciliophora, Acantharea) to metazoans (e.g., Arthropoda and Chordata, respectively represented principally by Copepods and Tunicates). Understanding how these body size and ecological differences modulate the influence of oceanic currents and local environmental conditions on geographic distributions is needed if we want to predict how eukaryotic communities, and therefore the trophic interactions and global biogeochemical cycles they participate in, will change with changing environmental conditions.

Previous studies suggested that all eukaryotes up to a size of approximately 1 mm are globally dispersed and primarily constrained by abiotic conditions (Finlay, 2002). While this view has been revised, the influence of body size on biogeography is manifest (Villarino et al., 2018, Richter et al. 2019). Interestingly, a recent study found that the turnover in community composition along currents slows down, rather than speeds up, with increasing body size (Richter et al, 2019). This suggests that, rather than influencing biogeography through its effect on abundance and ultimately dispersal capacity (i.e., larger organisms are more dispersal-limited; Finlay, 2002; Villarino et al., 2018), body size influences biogeography through its relationship with ecology and ultimately the sensitivity of communities to environmental conditions as they drift along currents. Under this scenario, the distribution of large long-lived generalist predators such as Copepods (Arthropoda) is expected to be stretched to the scale of currents systems through large-scale transport and mixing by main currents (Hellweger, van Sebille, & Fredrick, 2014; Lévy, Jahn, Dutkiewicz, & Follows, 2014; Madoui et al., 2017; Richter et al., 2019), and to be patchy as a result of small-scale turbulent stirring (Abraham, 1998). These contrasted views illustrate that little is known on how the interplay between body size, ecology, currents and the local environment shapes biogeography (Oziel et al., 2020).

Here we study plankton biogeography across all major eukaryotic groups in the sunlit ocean using 18S rDNA metabarcoding data from the *Tara* Oceans global survey, including recently released data from the Arctic Ocean (Ibarbalz et al., 2019). The data encompass 250,057 eukaryotic Operational Taxonomic Units (OTUs) sampled globally at the surface and at the Deep Chlorophyl Maximum (DCM) across 129 stations. We use a probabilistic model that allows identification of a number of 'assemblages', each of which represents a set of OTUs that tend to co-occur across samples (Sommeria-Klein et al., 2019; Valle, Baiser, Woodall, & Chazdon, 2014; Methods). Each local planktonic community can then be seen as a sample drawn in various proportions from the assemblages.

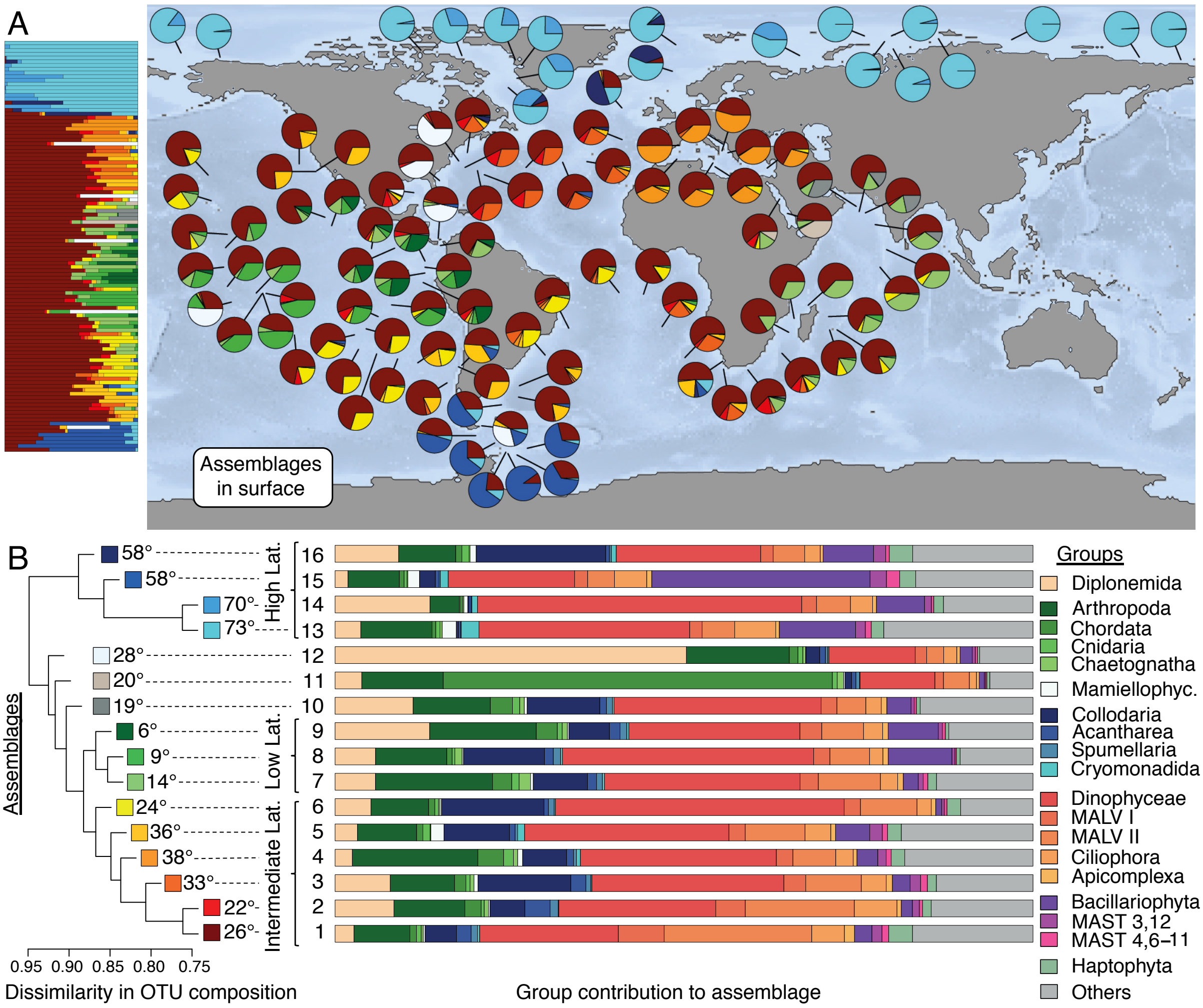

Figure 1: Global surface biogeography of eukaryotic plankton. The biogeography of all eukaryotic OTUs across *Tara* Oceans stations is characterized by 16 assemblages of co-occurring OTUs, each represented by a

distinct color (in A and the left panel in B) and identified by a number from 1 to 16 (in B). (**A**) Relative contribution of the 16 assemblages to surface plankton community in *Tara* Oceans stations, represented as pies on the world map and as stacked bars vertically ordered by latitude on the left-hand side of the map. (**B**) Left panel: dendrogram of assemblage dissimilarity with respect to their composition in OTUs (Simpson dissimilarity). The mean absolute latitude at which each assemblage is found is indicated. Three clusters can be distinguished: a high-latitude cluster — the most distinctive — in shades of blue, an intermediate-latidude cluster in shades from yellow to red, and a low-latitude cluster in shades of green. Right panel: barplot displaying the contribution of major eukaryotic groups (deep-branching monophyletic groups) to assemblages. The 19 groups shown in the barplot are those tallying more than 1,000 OTUs, grouped by phylogenetic relatedness.

Across the *Tara* Oceans samples and considering all eukaryotic OTUs together, we identified 16 geographically structured assemblages, each composed of OTUs covering the full taxonomic range of eukaryotic plankton (Fig. 1, S1; Appendix 1). Local planktonic communities often cannot be assigned to a single assemblage, as would be typical for terrestrial macro-organisms on a fixed landscape (Ficetola, Mazel, & Thuiller, 2017; Wallace, 1876), but are instead mixtures of assemblages (Fig. 1A). This is consistent with previous findings suggesting that neighbouring plankton communities are continuously mixed and dispersed by currents (Lévy et al., 2014; Richter et al., 2019). Nevertheless, three assemblages are particularly represented and most communities are dominated by one of them (Fig. 1A). The most prevalent assemblage represents a set of OTUs (about one fifth of the total) that are globally ubiquitous except in the Arctic Ocean (assemblage 1, in dark red). This assemblage typically accounts for about half the number of OTUs in non-Arctic communities, and is particularly rich in parasitic groups such as MALV (Fig. 1B). The two others dominate, respectively, in the Arctic Ocean (assemblage 13, in cyan) and in the Southern Ocean (assemblage 15, in marine blue), and are particularly rich in diatoms (Fig. 1B). Based on similarity in their OTU composition, the assemblages cluster into three main categories corresponding to low, intermediate and high latitudes (Fig. 1B). The transition between communities composed of high-latitude and lower-latitude assemblages is fairly abrupt, and occurs around 45° in the North Atlantic and -47° in the South Atlantic, namely at the latitude of the subtropical front, where the transition between cold and warm waters takes place (Fig. 1A&B; Talley, 2011).

This global analysis hides a strong heterogeneity across the 70 most diversified deep-branching groups of eukaryotic plankton (Table S1). Comparing the biogeography of these major groups using a normalized information-theoretic metric of dissimilarity (Meila, 2006; Methods), we found high pairwise dissimilarity values (ranging between 0.64 and 0.97; Fig. S2). This heterogeneity can be decomposed into two main interpretable axes of variation (Fig. 2; Methods). The first axis reflects the *amount* of biogeographic structure: group position on this axis is positively correlated to short-distance spatial autocorrelation (Pearson's correlation coefficient $\rho = 0.91$ at the surface; Fig. S3A), which measures the tendency for close-by communities to be composed of the same assemblages (Methods). Groups scoring low on this axis are characterized by strong local variation, or "patchiness". The second axis reflects the *nature* of the biogeographic structure: group position on this axis is positively correlated to the scale of biogeographic organization, which we measured as the characteristic distance at which spatial autocorrelation vanishes ($\rho = 0.53, p = 10^{-6}$ at the surface; Fig. S3B) and which ranges from ~7,000 to ~14,400 km across groups. Group position on the second axis is also positively correlated to within-basin autocorrelation ($\rho = 0.56, p = 10^{-7}$ at the surface; Fig. S3C), which measures the tendency for communities from the same oceanic basin (e.g., North Atlantic, South Atlantic, Mediterranean, Southern Ocean) to be composed of the same assemblages, and negatively correlated with latitudinal autocorrelation ($\rho = -0.49, p = 10^{-5}$ at the surface; S3D), which measures the tendency for communities at the same latitude on both sides of the Equator to be composed of the same assemblages (Methods). Results are similar at the DCM, although less pronounced (Fig. S4). The 70 groups of eukaryotic plankton cover the full spectra of biogeographies (Fig. 2, Fig. S5, Table S1), from those with weak spatial organization, or high patchiness (i.e., scoring low on the first axis, such as

Collodaria or Basidiomycota), to those organized at large spatial scale by oceanic basin (i.e., scoring high on both axes, such as Chordata or Arthropoda), and those organized at smaller spatial scale and according to latitude (i.e., scoring high on the first and low on the second axis, such as Mamiellophyceae, Haptophyta or MAST 3,12). These striking differences across planktonic groups suggest that accounting for their specificities is crucial to understanding their biogeography.

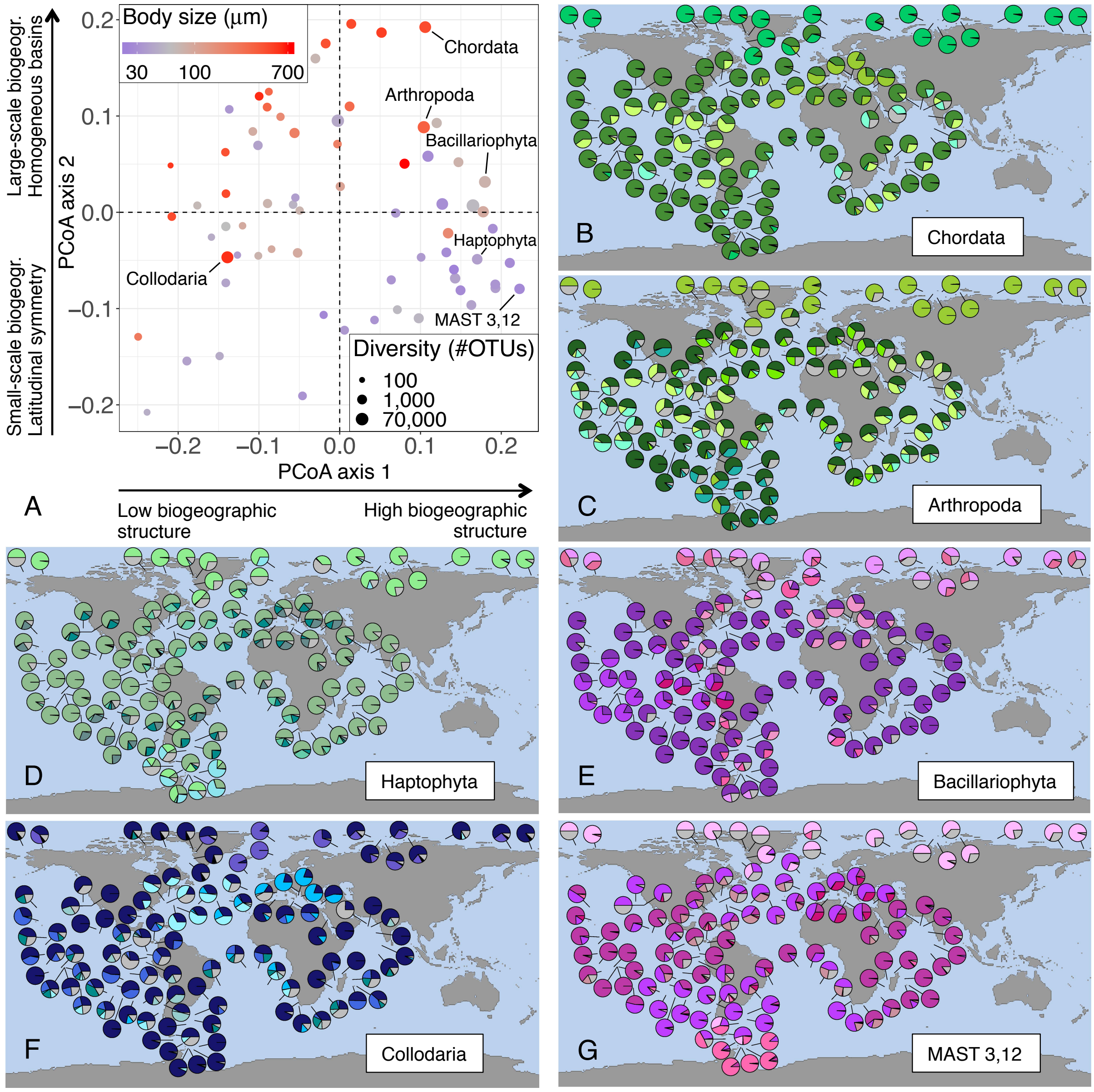

Figure 2: Biogeographic heterogeneity across major eukaryotic plankton groups. (**A**) Principal Coordinate Analysis (PCoA) of the biogeographic dissimilarity between 70 major groups of eukaryotic plankton. Each dot corresponds to the projection of a specific plankton group onto the first two axes of variation. Position along the first axis reflects the amount of biogeographic structure displayed by the group, from a patchy distribution with weak short-distance spatial autocorrelation on the left to a structured distribution with strong short-distance spatial autocorrelation on the right. Position along the second axis reflects the nature of biogeographic structure, from a biogeography structured by latitude at the bottom to a biogeography structured by oceanic basins at the top, as well as the scale of biogeographic organization, from small to large scale. Dot size is proportional to the log diversity of the corresponding group, and dot color represents its mean log body-size. (**B**-**G**) Surface biogeography of six major eukaryotic plankton groups. The relative contribution of the 5 to 7 most prevalent assemblages is shown in color, and that of the remaining assemblages is shown in gray; the color used for the most prevalent assemblage corresponds to the color used in Fig. 1B for the corresponding group.

We investigated how biogeographic differences among major groups relate to their diversity, body size, and ecology, coarsely defined as either phototroph, phagotroph, metazoan or parasite (Methods). We found that the amount of biogeographic structure (group position on the first axis) is strongly correlated to diversity ($\rho = 0.77, p = 10^{-13}$ below 2,000 OTUs; Fig. 3A). This suggests that geographic structure could play a role in generating and maintaining eukaryotic plankton diversity over ecological and possibly evolutionary scales, for example by promoting allopatric speciation and endemism. This relationship vanishes however for groups larger than about 2,000 OTUs, and two of the most diverse groups (Diplonemida, 38,769 OTUs and Collodaria, 17,417 OTUs) exhibit comparatively weak biogeographic structure. The amount of biogeographic structure is weakly anticorrelated to body size ($\rho = -0.32, p = 0.007$; Fig. S6A), and after accounting for differences in diversity across groups, is lower for metazoans than for phototrophs (ANCOVA t-test: $p = 0.035$, Fig. S6B), in agreement with the expectation of a higher local patchiness in their distribution induced by turbulent stirring (Abraham, 1998; Bertrand et al., 2014). In contrast, the nature of biogeographic structure (group position on the second axis) is strongly correlated to body size ($\rho = 0.61, p = 10^{-8}$; Fig. 3B) and ecology (ANOVA F-test: $p = 10^{-6}$, Fig. 3C), and only weakly to diversity ($\rho = 0.25, p = 0.033$; Fig. S6C). Metazoan groups score high on the second axis of variation (with the notable exception of Porifera sponges, probably at the larval stage) and phototrophs score low, while phagotrophs occupy an intermediate position, spanning a comparatively wider range of biogeographies (Fig. 3C). Parasites are just below metazoans, which suggests that their biogeography is influenced by that of their hosts. While body size covaries with ecology (phagotrophs are larger than phototrophs on average, and metazoans significantly larger than other plankton types; Fig. S7), the positive relationship between group position on the second axis and body size still holds within each of the four ecological categories (ANCOVA F-test: $p = 0.004$; Fig. S8). Diatoms (Bacillariophyta) are a striking example: of all phototrophs, they have the largest body size and also score highest on the second axis of variation. Conversely, ecology significantly influences group position on the second axis even after accounting for body size differences (ANCOVA F-test: $p = 0.035$). Collodaria, which we did not assign to an ecological category, score lower than expected from their large body size, but close to the average for phagotrophic groups (Fig. 2, Table S1). These results suggest that biogeographic patterns are influenced by both body size and ecology. To summarize, diversity-rich groups are biogeographically structured, with large-bodied heterotrophs (metazoans such as Copepods and Tunicates) exhibiting biogeographic variations at the scale of oceanic basins or larger, and small-bodied phototrophs (such as Haptophyta) at smaller spatial scale and following latitude (Fig. 2).

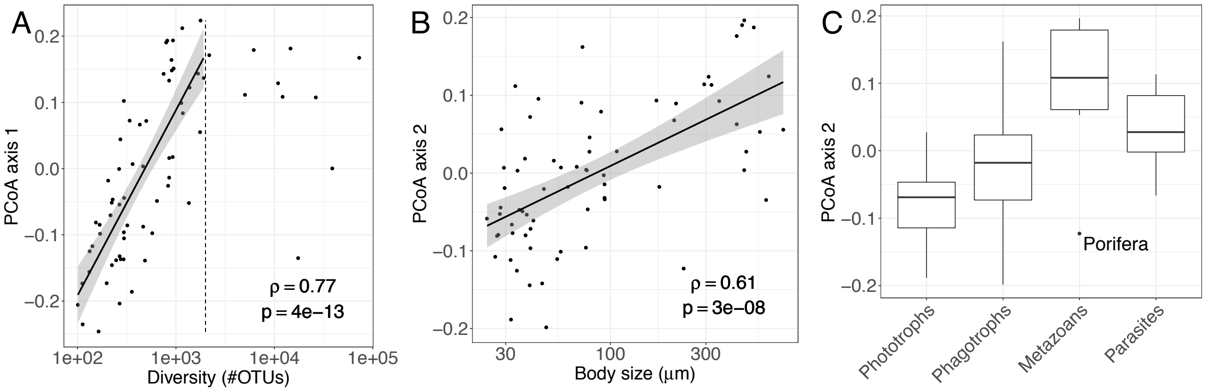

Figure 3: Relationship between biogeography and diversity, mean body size and ecology across major eukaryotic plankton groups. (**A**) The position of the 70 plankton groups along the first axis of biogeographic variation, indicative of the amount of biogeographic structure, increases sharply with log diversity (number of OTUs in the group) up to approximately 2,000 OTUs, but not beyond. (**B**) The position of the 70 plankton groups along the second axis, indicative of the nature and spatial scale of biogeographic structure, increases with log mean body size, indicating that large-bodied plankton is organized at larger spatial scale and according to

oceanic basins rather than latitude. (**C**) Positions along the second axis of plankton groups binned into four broad ecological categories. Pairwise differences are all significant except between Phagotrophs and Parasites.

A global biogeography matching oceanic basins suggests that communities respond to environmental variations slowly enough to be homogenised by ocean circulation at the basin scale (i.e., gyres; Richter et al., 2019), but have little ability to disperse between basins, either due to the comparatively limited connectivity by currents or to environmental barriers, and therefore that their biogeography is primarily shaped by the main ocean currents (Hellweger et al., 2014). Conversely, a biogeography matching latitude, symmetric with respect to the Equator, suggests a faster response of communities to environmental variations within basins (which are structured by latitude and currents, e.g. the cross-latitudinal influence of the Gulf Stream), low cross-basin dispersal limitation, and therefore a comparatively more important role of local environmental filtering in shaping biogeography. We investigated the ability of transport by currents and local environmental conditions to explain the global biogeography of major taxonomic groups. We compared biogeographic maps to maps of connectivity by currents and environmental conditions. We transformed minimum transport times between pairs of stations, previously computed from a global ocean circulation model (Methods; Clayton et al., 2017; Richter et al., 2019), into a set of connectivity maps describing patterns of connectivity by currents at different temporal scales (Methods; Fig. S9, S10). These connectivity maps can be interpreted as the geographic patterns that would be expected for plankton transported by currents; more precisely, each map corresponds to a specific time scale, and can be interpreted as the geographic patterns that would be expected for plankton which temporal variation along currents match this scale. We estimated local abiotic conditions using yearly-averaged measurements of temperature, nutrient concentration and oxygen availability (World Ocean Atlas 2013; Boyer et al., 2013; cf. Methods). Because biotic interactions (predation, competition, parasitic and mutualistic symbiosis) are thought to be important determinants of plankton community structure (Lima-Mendez et al., 2015), we also quantified local biotic conditions using the relative read counts of major eukaryotic groups (excluding the focal group; cf. Methods). Biotic conditions, similarly to abiotic ones, have a latitudinal structure, and we refer here to them collectively as 'environmental conditions' (Fig. S11, S12). The resulting environmental maps can be interpreted as the geographic patterns that would be expected for organisms that are strongly responsive to local environmental conditions but whose dispersal by currents is not limiting. Hence, a biogeography matching connectivity maps better than environmental maps suggest that the constraints imposed by oceanic currents (the transport of the plankton across those regions, modulated by mixing, ecological drift and speciation, but also by responses to nutrient supplies and temperature variations) dominate over those imposed by local environmental conditions.

We found that the total variance in surface community composition that can be explained by connectivity patterns and local environmental conditions (abiotic and biotic) averages 34% across groups (min. 8% and max. 65%) and is, as expected, tightly correlated to the amount of biogeographic structure ($\rho = 0.91$; Fig. 4A; Methods). The variance purely explained by connectivity patterns is for most groups larger than that purely explained by the local environment (40% versus 22% of explained variance on average at the surface; Fig. 4B-D, S13A), and is primarily contributed by between-basin connectivity patterns (Fig. S10 & S14). This supports a prominent role of transport by the main current systems and of the processes occurring along those pathways in shaping eukaryotic plankton biogeography, both by extending the distribution of some taxa beyond their optimal range (Dutkiewicz et al., 2019) and by constraining long-distance dispersal. We note that unmeasured environmental variations along currents likely contribute to this role of ocean circulation. As expected from our previous results, the ratio of the fractions of variance purely explained by connectivity patterns and the local environment, which reflects their relative contributions to biogeography, increases with group position on the second axis of variation ($\rho = 0.32$,

$p = 0.008$; Fig. 4B). Accordingly, the relative contribution of connectivity by currents also increases with average group body size ($\rho = 0.42, p = 3.10^{-4}$; Fig. 4C) and depends on ecology (ANOVA F-test: $p = 0.037$; Fig. 4D). These results indicate that metazoans are closer to freely drifting tracers strongly influenced by currents, and constrained in particular by limited between-basin connectivity, while phototrophs are more strongly coupled with environmental factors and disperse more readily between basins. The difference in sensitivity to local environmental conditions can be explained by differences in ecological requirements and community dynamics. Why there is a difference in between-basins dispersal is less clear. All basins are connected by currents within a few years of transport time (Jönsson & Watson, 2016), and small phototrophs may have a higher ability to disperse through environmental barriers by forming spores or dormant states (Finlay, 2002). Alternatively, the looser environmental coupling and slower dynamics of metazoan communities might make them more sensitive to the smaller between-basin compared to within-basin water flow. Finally, within the variance explained by the local environment, the contribution of pure biotic conditions largely dominates that of pure abiotic conditions for most groups (47% versus 16% on average at the surface; Fig. S13B), irrespective of their body size, ecology, diversity or biogeography (Fig. S15). Results are similar at the DCM, but are far less pronounced (Fig. S16, S17). Although we cannot exclude the possibility that local biotic conditions reflect the indirect effect of local abiotic factors that are not accounted for in our study, such as fluxes of nutrients, which are often more relevant to planktonic organisms than instantaneous nutrient concentrations (Dutkiewicz et al., 2019), these results indicate an additional role for interspecific interactions in shaping community composition (Lima-Mendez et al., 2015; Vincent & Bowler, 2020).

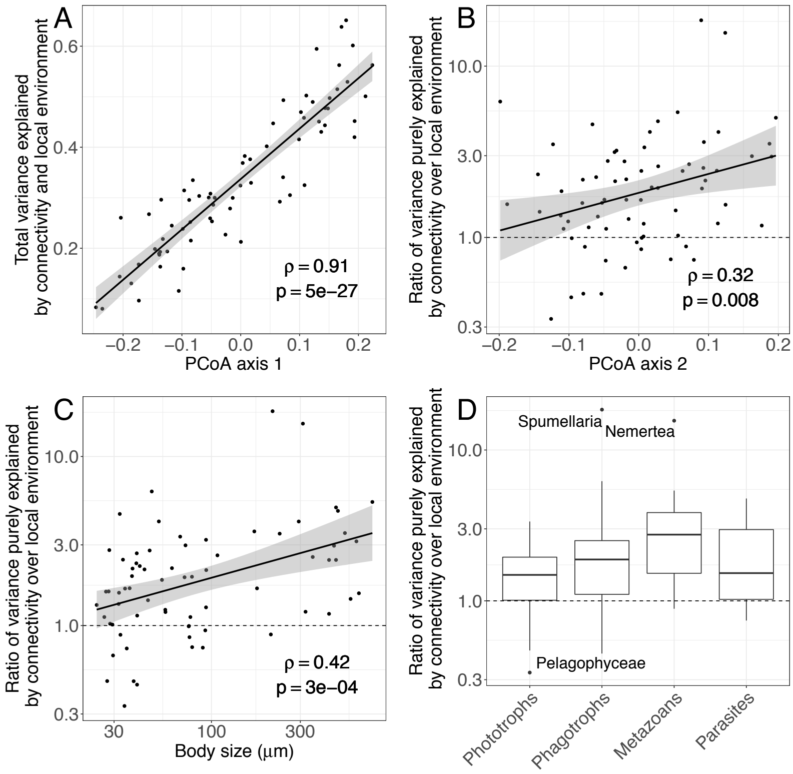

Figure 4: Drivers of surface biogeography across major eukaryotic plankton groups. (**A**) The total variance in surface biogeography that can be explained by the combination of connectivity by currents and (abiotic and biotic) local environmental conditions increases with the position of plankton groups on the first axis of biogeographic variation. (**B-D**) Across major plankton groups, the log ratio of the variance explained purely by connectivity over the variance explained purely by (abiotic and biotic) local environmental conditions (B) increases with group position on the second axis of variation, (C) increases with mean body size, and (D) varies across broad ecological categories (only the pairwise difference between Phototrophs and Metazoans is

significant). The ratio is higher than 1 for most groups, reflecting an overall stronger influence of connectivity by currents compared to local environmental conditions on plankton biogeography at the surface.

Our study clarifies the patterns and processes underlying the global biogeography of the main groups of eukaryotic plankton in the sunlit ocean. Consistent with the recently proposed concept of seascape (Kavanaugh et al., 2016), we find that community variation along currents is slow enough to allow currents to be the dominant driver of global-scale biogeography (Richter et al., 2019). The continuous movement of water masses generates biogeographic patterns that are better represented by overlapping taxa assemblages than by the well-delineated biomes characteristic of terrestrial systems. Our comparison of eukaryotic plankton groups reveals several additional results. First, the geographic structuring induced by currents may have favored the generation and maintenance of eukaryotic plankton diversity. Second, plankton ecology matters beyond body size differences, and reciprocally body size matters beyond ecological differences. Third, body size and ecology influence primarily the nature of biogeographic patterns, namely their spatial scale of organization and whether they are organized by oceanic basins or latitude, and only secondarily the amount of biogeographic structure, namely local patchiness. Fourth, biotic conditions appear to be a more important driver of biogeography than local abiotic conditions. Our results reconcile the views that larger-bodied organisms are more dispersal-limited (Finlay, 2002; Villarino et al., 2018) and yet display a slower compositional turnover along currents than smaller organisms (Richter et al., 2019): at the global scale, organisms of larger sizes are indeed more dispersal-limited; however at the regional scale, they have wider spatial distributions, presumably linked to their specific ecologies, longer lifespan and reduced sensitivity to local environmental variations. At the two extremes, metazoan heterotrophs are structured at the scale of oceanic basins following the main currents, while small phototrophs are structured latitudinally with a comparatively larger influence of local environmental conditions, predominantly biotic ones. Together, our results suggest that predictive modeling of plankton communities in a changing environment (Ibarbalz et al., 2019; Lotze et al., 2019) will critically depend on our ability to model the impact of changes in ocean currents and to develop niche models accounting for both species ecology and interspecific interactions.

Acknowledgements:

We are grateful to Federico Ibarbalz for his essential help with the data. We thank Olivier Jaillon and Colomban de Vargas for feedback and early discusions on the project. We thank Mick Follows and Oliver Jahn for sharing MITgcm simulation results for the Arctic Ocean. We thank Florian Hartig and Odile Maliet for their precious advice on Bayesian inference, Leandro Arístide and Felipe Delestro for their kind assistance with the figures, and Carmelo Fruciano and Benoît Perez for their guidance on statistics. We thank Leandro Arístide, Fabio Benedetti, Julien Clavel, Carmelo Fruciano, Elena Kazamia, Sophia Lambert, Eric Lewitus, Odile Maliet, Marc Manceau, Olivier Missa, Silvia de Monte, Isaac Overcast, Benoît Perez, Ignacio Quintero, Enrico Ser-Giacomi, Ana Catarina Silva and Flora Vincent for suggestions and fruitful discussions.

This work was supported by *European Research Council* grants (ERC 616419-PANDA, to H.M.; ERC 835067-DIATOMIC, to C.B.), grants from the French *Agence Nationale de la Recherche* (MEMOLIFE, ref. ANR-10-LABX-54, to G.S.K., H.M. and C.B.; OCEANOMICS, ref. ANR-11-BTBR-0008, to C.B.) and funds from CNRS. C.B. and H.M. are members of the Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE. This article is contribution number XXX of *Tara* Oceans.

References:

Abraham, E. R. (1998). The generation of plankton patchiness by turbulent stirring. *Nature*, *391*(6667), 577–580. doi: 10.1038/35361

Beaugrand, G., & Kirby, R. R. (2018). How Do Marine Pelagic Species Respond to Climate Change? Theories and Observations. *Annual Review of Marine Science*, *10*(1), 169–197. doi: 10.1146/annurev-marine-121916-063304

Bertrand, A., Grados, D., Colas, F., Bertrand, S., Capet, X., Chaigneau, A., … Fablet, R. (2014). Broad impacts of fine-scale dynamics on seascape structure from zooplankton to seabirds. *Nature Communications*, *5*(1), 1–9. doi: 10.1038/ncomms6239

Boyer, T. P., Antonov, J. I., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., … O'Brien, T. D. (2013). *World Ocean Database 2013.*

Clayton, S., Dutkiewicz, S., Jahn, O., Hill, C., Heimbach, P., & Follows, M. J. (2017). Biogeochemical versus ecological consequences of modeled ocean physics. *Biogeosciences*, *14*(11), 2877–2889. doi: 10.5194/bg-14-2877-2017

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., … Tara Oceans, C. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237). (WOS:000354877900034). doi: 10.1126/science.1261605

Dutkiewicz, S., Cermeno, P., Jahn, O., Follows, M. J., Hickman, A. E., Taniguchi, D. A. A., & Ward, B. A. (2019). Dimensions of Marine Phytoplankton Diversity. *Biogeosciences Discussions*, 1–46. doi: https://doi.org/10.5194/bg-2019-311

Ficetola, G. F., Mazel, F., & Thuiller, W. (2017). Global determinants of zoogeographical boundaries. *Nature Ecology and Evolution*. doi: 10.1038/s41559-017-0089

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, *281*(5374), 237–240.

Finlay, B. J. (2002). Global Dispersal of Free-Living Microbial Eukaryote Species. *Science*, *296*(5570), 1061–1063. doi: 10.1126/science.1070710

Hellweger, F. L., van Sebille, E., & Fredrick, N. D. (2014). Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science.*

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., … Zinger, L. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, *179*(5), 1084-1097.e21. doi: 10.1016/j.cell.2019.10.008

Jönsson, B. F., & Watson, J. R. (2016). The timescales of global surface-ocean connectivity. *Nature Communications*, *7*, 11239. doi: 10.1038/ncomms11239

Kavanaugh, M. T., Oliver, M. J., Chavez, F. P., Letelier, R. M., Muller-Karger, F. E., & Doney, S. C. (2016). Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation. *ICES Journal of Marine Science*, *73*(7), 1839–1850. doi: 10.1093/icesjms/fsw086

Lévy, M., Jahn, O., Dutkiewicz, S., & Follows, M. J. (2014). Phytoplankton diversity and community structure affected by oceanic dispersal and mesoscale turbulence. *Limnology and Oceanography: Fluids and Environments*, *4*(1), 67–84. doi: 10.1215/21573689-2768549

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., … Tara Oceans, C. (2015). Determinants of community structure in the global plankton interactome. *Science*, *348*(6237), 9. (WOS:000354877900035). doi: 10.1126/science.1262073

Lotze, H. K., Tittensor, D. P., Bryndum-Buchholz, A., Eddy, T. D., Cheung, W. W. L., Galbraith, E. D., … Worm, B. (2019). Global ensemble projections reveal trophic amplification of ocean biomass declines with climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(26), 12907–12912. doi: 10.1073/pnas.1900194116

Madoui, M.-A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L., … Wincker, P.

(2017). New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Molecular Ecology*, *26*(17), 4467–4482. doi: 10.1111/mec.14214

Mahé, F., Rognes, T., Quince, C., Vargas, C. de, & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. doi: 10.7717/peerj.593

Meila, M. (2006). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873–895.

Oziel, L., Baudena, A., Ardyna, M., Massicotte, P., Randelhoff, A., Sallée, J.-B., … Babin, M. (2020). Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. *Nature Communications*, *11*(1), 1–8. doi: 10.1038/s41467-020-15485-5

Raes, E. J., Bodrossy, L., van de Kamp, J., Bissett, A., Ostrowski, M., Brown, M. V., … Waite, A. M. (2018). Oceanographic boundaries constrain microbial diversity gradients in the South Pacific Ocean. *Proceedings of the National Academy of Sciences*, *115*(35), E8266–E8275. doi: 10.1073/pnas.1719335115

Richter, D. J., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G., … Coordinators, T. O. (2019). Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *BioRxiv*, 867739. doi: 10.1101/867739

Righetti, D., Vogt, M., Gruber, N., Psomas, A., & Zimmermann, N. E. (2019). Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Science Advances*, *5*(5), eaau6253. doi: 10.1126/sciadv.aau6253

Sommeria-Klein, G., Zinger, L., Coissac, E., Iribar, A., Schimann, H., Taberlet, P., & Chave, J. (2019). Latent Dirichlet Allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13109

Talley, L. D. (2011). *Descriptive physical oceanography: an introduction*. Academic press.

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. V., & Worm, B. (2010). Global patterns and predictors of marine biodiversity across taxa. *Nature*, *466*(7310), 1098–1101. doi: 10.1038/nature09329

Valle, D., Baiser, B., Woodall, C. W., & Chazdon, R. (2014). Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. *Ecology Letters*, *17*(12), 1591–1601. (WOS:000345216200012). doi: 10.1111/ele.12380

Villarino, E., Watson, J. R., Jönsson, B., Gasol, J. M., Salazar, G., Acinas, S. G., … Chust, G. (2018). Large-scale ocean connectivity and planktonic body size. *Nature Communications*, *9*(1), 142. doi: 10.1038/s41467-017-02535-8

Vincent, F., & Bowler, C. (2020). Diatoms Are Selective Segregators in Global Ocean Planktonic Communities. *MSystems*, *5*(1). doi: 10.1128/mSystems.00444-19

Wallace, A. R. (1876). *The geographical distribution of animals: with a study of the relations of living and extinct faunas as elucidating the past changes of the earth's surface* (Vol. 1). Cambridge University Press.

Worden, A. Z., Follows, M. J., Giovannoni, S. J., Wilken, S., Zimmerman, A. E., & Keeling, P. J. (2015). Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, *347*(6223), 1257594. doi: 10.1126/science.1257594

Methods:

DNA data processing

Planktonic organisms were sampled in 129 stations of the open ocean (no lagoon or costal waters) covering the Arctic, Atlantic, Indian, East Pacific and Southern Oceans as well as the Mediterranean and Red Seas. Samples were collected from subsurface mixed-layer waters (henceforth referred to as 'surface', about 5 m deep). In about half of the stations, samples were additionally collected at the Deep Chlorophyll Maximum ('DCM', ranging from 20 m to 190 m deep, most commonly around 40 m deep). At both depth levels, four different fractions of organisms' body size were collected: 0.8-5 mm, 5-20 mm (or 3-20 mm in some stations, which we treated as equivalent), 20-180 mm, and 180-2000 mm. In Arctic stations, a small size fraction without upper size limit (0.8 mm – infinity) was collected in place of the 0.8-5 mm size fraction. We treated both fractions as equivalent, since they were found to be of similar composition in stations where both were collected (indeed, small organisms greatly outnumber larger ones).

Whole DNA was extracted from these samples, then the V9 region of the gene coding for the eukaryotic 18S rRNA was PCR-amplified and the resulting amplicons were sequenced by Illumina sequencing. Sequencing reads were trimmed for quality, length and fidelity of primer sequences, then clustered into Operational Taxonomic Units (henceforth 'OTUs') using the SWARM unsupervised algorithm (Mahé, Rognes, Quince, Vargas, & Dunthorn, 2014). OTUs were given taxonomic assignations by matching their most abundant sequence to a custom database derived from the Protist Ribosomal Reference (PR2; Guillou et al., 2013). OTUs with less than 80% similarity to the closest reference sequence were discarded, as well as OTUs matching non-eukaryotic reference sequences. This pipeline resulted in a list of OTUs and their associated read count for each sample. See de Vargas et al. (2015) for further detail on the sampling, wetlab and bioinformatics protocols. Taxonomic assignations of OTUs were then used to obtain ecological annotations based on literature, from which OTUs could be broadly classified into parasites, phototrophs, phagotrophs and metazoans (Ibarbalz et al., 2019).

For every station and depth, we pooled the results obtained for the four size fractions into a single aggregated sample (henceforth simply referred to as a 'sample'). We discarded the samples where one or more size fractions were missing so as not to bias the results. This treatment resulted in retaining 113 stations, broken down into 110 surface samples and 62 DCM samples and encompassing 250,057 OTUs.

Characterizing samples as mixtures of assemblages using Latent Dirichlet Allocation

To capture the spatial patterns of OTU co-occurrence across samples, we used a model-based algorithm of dimensionality reduction, Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003). We considered that an OTU occurs in a sample when it is represented by at least one sequence read, and we discarded read count information. The method consists in fitting a so-called mixed membership model to the list of OTU occurrences in each sample (i.e., the community matrix). Even though the model formally assumes that OTUs can be observed several times in each sample (i.e., it assumes discrete abundance data rather than presence-absence data), this does not impair model fitting and interpretation for presence-absence data (Sommeria-Klein et al., 2019). The model assumes that OTU occurrences are sampled from a mixture of several (unobserved) assemblages. Each assemblage represents a set of OTUs that tend to co-occur across samples. The fitting process consists in inferring the K most likely assemblages from the data, where the number K of assemblages is fixed beforehand. Assemblages are defined by their OTU composition, both in terms of OTU identity and relative prevalence. The relative prevalence of an OTU in an assemblage is proportional to its number of occurrences across the samples where the assemblage is present. Assemblages may share OTUs, and samples may contain a mixture of coexisting assemblages. As a consequence the model is able to capture spatial patterns despite the presence of many ubiquitous OTUs, a typical trait of microbial communities, and to accommodate gradual changes in taxonomic composition across space. The model is little influenced by OTUs of rare occurrence, since those OTUs contribute little co-occurrence information. Symmetric Dirichlet priors are put on the mixture of assemblages in samples and on the mixture of OTUs in assemblages, with respective control parameters a and d.

We fitted the model to all samples simultaneously, making no distinction between surface and DCM samples. We used the Gibbs sampling algorithm of Phan et al. (2008), wrapped in the R package 'topicmodels' (Grün & Hornik, 2011), with control parameters $\alpha = 0.1$ and $\delta = 0.1$. Values of a and d lower than 1 favor low spatial overlap and few shared OTUs between assemblages, respectively. Model output is chiefly influenced by d: values of d close to 1 or higher led to solutions where very few widely distributed assemblages shared the bulk of OTUs. These solutions were associated with lower predictive power on held-out data (as measured by perplexity; see next paragraph) and lower posterior probability compared to lower d values. We ran the MCMC (Markov Chain Monte Carlo) chains for 3,000 iterations starting from random assemblages. After the first 2,000 iterations (burn-in), we recorded samples every 25 iterations for the last 1,000 iterations (i.e., 40 MCMC samples per chain). MCMC samples are sets of values for all the model's latent variables, which follow the model's posterior distribution given the data once the chain has converged. The associated likelihood values are computed as part of the algorithm. Among the 40 MCMC samples, we picked that with likelihood closest to the mean across samples, as a proxy for the set of latent variable values maximizing the posterior distribution.

We selected the optimal number K of assemblages by cross-validation. We partitioned the data into random sets of 10 samples, and fitted the model on the data while successively holding out each 10-sample validation set. We then measured the predictive power of each fitted model on the corresponding validation set. We measured it using perplexity, a decreasing function of predictive power defined as the geometric mean of the likelihood across OTU occurrences (*perplexity* function in R package 'topicmodels'; Grün & Hornik, 2011). We compared the mean perplexity across validation sets for K between 2 and 35, and picked the minimum value after smoothing the curve with a 6-degree-of-freedom spline (function *smooth.spline*, R package 'stats'; R Core Team, 2018). For large datasets, the mean perplexity as a function of K may enter a plateau after an initial decrease (Fig. S1). As a heuristic means to select the K value corresponding to the onset of the plateau, we first fitted the model to the whole dataset for the K value with minimum mean perplexity, and used the number of assemblages obtained after removing all the assemblages with a cumulative prevalence across the dataset of less than one sample. We then fitted the model again for the number of assemblages thus obtained.

Once we had selected the K value, we ran 100 independent MCMC chains on the whole dataset from random initial conditions. To check for potential insufficient mixing along the chains, we measured the similarity in the spatial distribution of assemblages across the chains (Table S1), using the metric defined in Sommeria-Klein et al. (2019). We picked the chain with posterior probability closest to the mean across chains for the final interpretation.

Comparing assemblages

Each assemblage is characterized by a list of OTUs and their relative prevalence. When running LDA on the whole eukaryotic data set, we measured the pairwise dissimilarity between assemblages as the Simpson dissimilarity of their composition in OTUs. We then built an UPGMA tree out of the dissimilarity matrix to obtain a hierarchical clustering of assemblages (function *agnes*, R package 'cluster').

Major eukaryotic groups

After having first considered all eukaryotic OTUs combined, we sought to compare biogeographic patterns across major groups of eukaryotic plankton. To this end, we classified OTUs into deep-branching monophyletic groups based on taxonomic assignations, as in de Vargas et al. (2015), and we discarded those tallying less than 100 OTUs. We obtained 70 groups tallying between 101 to 72,769 OTUs (Dinophyceae), for a total of 241,020 OTUs.

We classified eukaryotic groups into four broad ecological categories based on the dominant ecology of their constituent OTUs: parasites, phototrophs, phagotrophs and metazoans. All groups fell entirely or mostly into one of these categories, except Dinophyceae (various ecological functions, including many mixotrophs) and Collodaria (mostly phagotrophic photohosts), which we did not classify and thus excluded from our statistical comparisons to ecology.

We estimated the mean body size of each group based on the distribution of the corresponding sequence reads over the four size fractions and across samples. Specifically, we computed the mean body size $\langle d_G \rangle$ of group G across samples as:

$$\langle d_G \rangle = \frac{1}{S} \sum_{i=1}^{S} \frac{\sum_{f=1}^{4} \sum_{t \in G} p_{t,f,i} d_f}{\sum_{f=1}^{4} \sum_{t \in G} p_{t,f,i}}$$

where S is the number of samples, d_f the mid-range body size of fraction f (i.e., respectively 2.9 mm, 12.5 mm, 100 mm, and 1,090 mm for the four size fractions), and $p_{t,f,s} = n_{t,f,i} / \sum_t n_{t,f,i}$ the relative abundance of OTU t in fraction f of sample i, as inferred from the number $n_{t,f,i}$ of sequence reads assigned to it. Groups' mean body size ranges from 24 mm (Cryptophyta) to 731 mm (Chaetognatha).

Groups diversity and body size are independent from each other ($p = 0.25$), but variation in body size partly overlaps with ecological categories: all pairs of ecological categories have significantly distinct body size except parasites and phagotrophs (Fig. S7).

Amount of biogeographic structure

To quantify the amount of biogeographic structure exhibited by a planktonic group, we computed, separately for surface and DCM samples, the short-distance spatial autocorrelation I_k in the global distribution of each assemblage k across stations. We measured I_k using Moran's index (function Moran.I, R package 'ape'; Paradis & Schliep, 2018), defined as:

$$I_k = \frac{S}{\sum_{i=1}^{S} \sum_{j=1}^{S} w_{ij}} \frac{\sum_{i=1}^{S} \sum_{j=1}^{S} w_{ij} \left(\theta_i^k - \langle \theta^k \rangle\right)\left(\theta_j^k - \langle \theta^k \rangle\right)}{\sum_{i=1}^{S} \left(\theta_i^k - \langle \theta^k \rangle\right)^2}$$

where S is the number of stations, θ_i^k the proportion of assemblage k in station i (i.e., $\sum_{k=1}^{K} \theta_i^k = 1$), $\langle \theta^k \rangle = \sum_{i=1}^{S} \theta_i^k / S$ its mean over stations, and $w_{ij} = w\left(d_{ij}\right)$ is a weight function that decreases with the spatial distance $d_{\square j}$ between stations i and j. We defined the spatial distance between two stations as the shortest path between them that follows Earth's surface without crossing land (Dijkstra's algorithm; Richter et al., 2019). We chose an inverse-square weight function satisfying $w\left(maxd_{ij}\right) = 0$ and $w\left(mind_{ij}\right) = 1$:

$$w_{ij} = w\left(d_{ij}\right) = \frac{\left(\frac{\max d_{ij}}{d_{ij}}\right)^2 - 1}{\left(\frac{\max d_{ij}}{\min d_{ij}}\right)^2 - 1}$$

where $mind_{ij}$ is about 100 km and $maxd_{ij}$ 23,500 km. We then computed the overall short-distance spatial autocorrelation I in the biogeography as the weighted mean of I_k over assemblages, using the mean assemblage proportions $\langle \theta^k \rangle$ as weights, separately for the surface and the DCM:

$$I = \sum_{k=1}^{K} \langle \theta^{\square} \rangle I_k$$

Scale of biogeographic organization

We quantified the scale of biogeographic organization as the characteristic distance at which spatial autocorrelation vanishes. We measured this distance in surface and at the DCM by computing Moran's I with a step weight function taking value $w_{ij} = 1 if d_{ij} < d$ and $w_{ij} = 0$ otherwise, and by varying d linearly between $mind_{ij}$ and $maxd_{ij}$ over 20 increments: $d^n = mind_{ij} + n\left(maxd_{ij} - mind_{ij}\right)/20$ for *n* between 1 and 20. Moran's I decreases first linearly with spatial distance d and then vanishes asymptotically. We smoothed the $I(d)$ curve with a 5-degree-of-freedom spline, and then performed a linear regression (function *lm*, R package 'stats') on its linear domain. We defined the characteristic distance at which spatial autocorrelation vanishes as the x-axis intercept of the linear regression (i.e., $-b/a$, where *a* and *b* are the slope and y-axis intercept, respectively).

Autocorrelation within oceanic basins

We measured the spatial autocorrelation within oceanic basins by computing Moran's I with a step weight function taking value $w_{ij} = 1$ when stations i and j belong to the same oceanic basin and $w_{ij} = 0$ otherwise, separately at the surface and the DCM. We defined as separate oceanic basins the Arctic Ocean, North Atlantic Ocean, South Atlantic Ocean, Mediterranean Sea, Red Sea, Indian Ocean, North Pacific Ocean, South Pacific Ocean and Southern Ocean. We expect a correlation between short-distance and within-basin spatial autocorrelation, since both are computed as Moran's I using different weight functions. To take this into account, we divided for each group the within-basin autocorrelation by the short-distance autocorrelation in statistical analyses.

Latitudinal autocorrelation

To measure whether the same assemblages tend occur at the same absolute latitude on both sides of the Equator, we computed, separately at the surface and the DCM, Moran's I with a weight function taking value $w_{ij} = e^{-(|l_i|-|l_j|)^2/\sigma^2}$ when $sign(l_i) = -sign(l_j)$ and $w_{ij} = 0$ otherwise, where l_i is the latitude of station *i* in degrees. We used $\sigma^2 = 25$, the value that maximized latitudinal autocorrelation in the surface biogeography of all eukaryotic OTUs combined. As for within-basin autocorrelation, we divided for each group the latitudinal autocorrelation by the short-distance autocorrelation in statistical analyses.

Comparing biogeography across groups

We applied our LDA decomposition pipeline (see above) separately to each of the major groups. To compare the resulting biogeography across groups, we computed a measure of biogeographic dissimilarity between pairs of groups. We used the relative mutual information between the spatial distribution of assemblages, an information theoretic quantity closely related to the Variation of Information (Meila, 2006) but normalized by total entropy so as to make it insensitive to differences in number of assemblages between groups.

We note $\theta_1 = \left(\theta_{1,i}^{k_1}\right)_{i\in[\![1,S]\!]}^{k_1\in[\![1,K_1]\!]}$ and $\theta_2 = \left(\theta_{2,i}^{k_2}\right)_{i\in[\![1,S]\!]}^{k_2\in[\![1,K_2]\!]}$ the spatial distribution over the S stations of the respectively K_1 and K_2 assemblages in the biogeographies of groups 1 and 2, with $\sum_{k_1=1}^{K_1} \theta_{1,i}^{k_1} = 1$ and $\sum_{k_2=1}^{K_2} \theta_{2,i}^{k_2} = 1$ for every station *i*. We computed the entropy $H(\theta_j)$ and the mutual information $I(\theta_1, \theta_2)$ between θ_1 and θ_2 as:

$$H(\theta_j) = -\sum_{k_j=1}^{K_j} \langle \theta^{k_j} \rangle log \langle \theta^{k_j} \rangle$$

$$I(\theta_1, \theta_2) = \sum_{\{k_1, k_2\} \in [\![1, K_1]\!] \times [\![1, K_2]\!]} \langle \theta_1^{k_1} \theta_2^{k_2} \rangle log \frac{\langle \theta_1^{k_1} \theta_2^{k_2} \rangle}{\langle \theta_1^{k_1} \rangle \langle \theta_2^{k_2} \rangle}$$

where $\langle . \rangle$ stands for the mean over the S stations. The relative mutual information between θ_1 and θ_2 is then defined as:

$$\mathcal{I}(\theta_1, \theta_2) = \frac{I(\theta_1, \theta_2)}{H(\theta_1) + H(\theta_2) - I(\theta_1, \theta_2)}$$

The similarity index $\mathcal{I}(\theta_1, \theta_2)$ varies between 0 and 1, and can be transformed into a dissimilarity index by taking $1 - \mathcal{I}(\theta_1, \theta_2)$.

We performed a Principal Coordinate Analysis (function *pcoa.all*, Legendre 2007) on the $1 - \mathcal{I}$ dissimilarity matrix between the 70 major groups, resulting in 69 PCoA axes. We performed multivariate linear regressions (function 'lm') of the projections of groups onto the PCoA axes against six explanatory variables: the amount of biogeographic structure, the scale of biogeographic organization, the within-basin autocorrelation, the latitudinal autocorrelation, the logarithm of group diversity and the logarithm of group body size. Each of these explanatory variables explained a significant part of the variance in the groups' projections onto all PCoA axes ($p < 10^{-3}$). When considering each PCoA axis separately, groups' projections onto the first two PCoA axes could be well predicted by the combination of these six explanatory variables ($R^2_{adj.} = 0.86, p = 10^{-25}$ for the first axis, $R^2_{adj.} = 0.69$, $p = 10^{-15}$for the second axis), while this was not the case for subsequent PCoA axes ($R^2_{adj.} < 0.17, p \gtrsim 10^{-2}$). Therefore the first two PCoA axes carry most of the interpretable biogeographic variation across groups, and as a consequence we focused on the ordination of the groups along those two axes.

Disentangling the effect of body size, diversity and ecology

We assessed correlations between continuous variables using Pearson's correlation coefficient and associated t-test (function *cor.test*). We tested the effect of ecology (with four factor levels: phototrophs, phagotrophs, metazoans and parasites) on a continuous variable (i.e., group position on the first two PCoA axes, or a ratio of explained variances) by an Analysis of Variance (ANOVA), and the respective effects of ecology and a continuous covariate (either log body size or log diversity) by an Analysis of Covariance (ANCOVA; functions *lm* and *anova*). We considered the t-tests between pairs of ecological categories only when the F-test was significant, and grouped ecological categories together when this improved the model. We used a 5% significance threshold.

Abiotic environmental variables

For each sample, we used as local abiotic conditions the mean annual values measured at the approximate location and depth of the sample for temperature, nitrate, phosphate and silicate concentrations, dissolved oxygen concentration, oxygen saturation and apparent oxygen utilization (World Ocean Atlas 2013; Boyer et al., 2013). We also used iron concentration values derived from model simulations (Menemenlis et al., 2008). We conducted a Principal Component Analysis (PCA) on these abiotic environmental variables, separately for surface and DCM samples, after centering and standardization (function dudi.pca, R package 'ade4'; Chessel, Dufour, & Thioulouse, 2004). We retained the first three axes for further analysis (axes with eigenvalue larger than 0.8).

For surface samples, the first axis amounts to 44% of the total variance (eigenvalue = 3.5), and corresponds to variation in temperature as well as in nitrate, phosphate, silicate and dissolved oxygen concentrations. The second axis amounts to 26% of variance (eigenvalue =

2.1) and corresponds to variation in oxygen saturation and utilization. The third axis amounts to 16% of variance (eigenvalue = 1.3) and is mostly driven by iron concentration (Fig. S11).

For DCM samples, the first axis amounts to 51% of the total variance (eigenvalue = 4.1), and corresponds mostly to variation in phosphate and nitrate concentration, as well as oxygen utilization and saturation. The second axis amounts to 27% of variance (eigenvalue = 2.2), and corresponds mostly to variation in temperature and dissolved oxygen concentration. The third axis amounts to 10% of variance (eigenvalue = 0.84) and is driven by iron concentration.

Biotic environmental variables

We used the relative abundances in the community of the 70 major groups of eukaryotic plankton under study as proxy for local biotic conditions. We estimated the local relative abundance $a_{G,i}$ of a group in sample i as the mean of its relative read count in the four size fractions:

$$a_{G,i} = \frac{\sum_{f=1}^{4} \sum_{t \in G} p_{t,f,i}}{\sum_{f=1}^{4} \sum_{t} p_{t,f,i}}$$

where, as defined previously for the calculation of body size, $p_{t,f,i}$ is the relative read count of OTU t in fraction f of sample i. The quantity $a_{G,i}$ is not directly a measure of the relative number of individuals in group G, because it is obtained by summing over size fractions, and both the density of individuals per volume of water and the sampled volume of water differ widely among size fractions. It can nevertheless be used to characterize the variation in community composition across stations.

We conducted a Principal Component Analysis (PCA) on relative abundances a_G across groups, separately for surface and DCM samples, after centring and standardization (function *dudi.pca*, R package ‘ade4’; Chessel et al., 2004), and we retained the axes with eigenvalue larger than 0.8 as biotic environmental variables for further analysis (the first 28 axes for surface samples; the first 23 axes for DCM samples; Fig. S12). To avoid using the abundance of the group under study as an explanatory variable, we performed 70 separate PCAs, each time removing the focal group.

Transport times along currents

To quantify the role of transport by currents in generating the observed biogeographies, we compared them with connectivity maps, known as Moran Eigenvector Maps (MEMs), obtained by decomposing the matrix of pairwise minimum transport times between stations using Principal Coordinate Analysis (PCoA), as described below (Legendre & Legendre, 2012). In terrestrial ecology, similar maps are obtained by decomposing the matrix of pairwise geographic distances between sampled sites, and are classically used to assess the effect of dispersal limitation by distance on the distribution of species.

Here, we measure the connectivity of stations using minimum transport times between stations, in line with previous studies using Lagrangian transit times to explain the spatial distribution of marine plankton (Jönsson & Watson, 2016; Watson et al., 2011; Wilkins, van Sebille, Rintoul, Lauro, & Cavicchioli, 2013). This measure of connectivity is more robust than physical connectivity (i.e. the number of particles exchanged between stations), which strongly depends on the number of particles considered in the simulation as well as on the method used to reconstruct the trajectories of particles between stations. When seeking to explain patterns of taxon presence-absence for planktonic organisms, the minimum transport time between stations appears more relevant than the mean transport time, since only a few individuals are required to ‘seed’ a location with a given taxon (Jönsson & Watson, 2016; Wilkins et al., 2013). Moreover, mean transport times are not well-defined in the global ocean in the absence of a physically motivated upper time-scale (Jönsson & Watson, 2016). Finally, minimum transport time has been shown to be a good predictor of the average amount of change in global plankton community composition that takes place along currents over a timescale of a year (i.e. a few thousands km), as a result of mixing, environmental variations, internal biotic interactions, behaviour and random compositional drift (Richter et al., 2019).

The minimum transport times were computed by Richter et al. (2019) using a numerical simulation of a global oceanic circulation model (MITgcm Darwin; Clayton et al., 2017), as summarized here. In this simulation, particles were released uniformly across the globe and advected for a cycle of 6 years using the horizontal velocity field along with a turbulent diffusivity. A set of 10,000-year trajectories was then constructed using this 6-year master cycle with particles seeded in each sampling station. Transport times between sampled locations were inferred by considering every event when a particle travelled from one sampled location to another, up to a radius of 200 km (see Richter et al., 2019 for more details). Only stations that had exchanged at least 10 particles were considered significantly connected. This computation was performed twice using simulations at 5-m depth and 75-m depth, so as to estimate the minimum transport times at the surface and at the DCM, respectively. We thus obtained two symmetric square matrices, one for surface samples and one for DCM samples, with minimum transport times as entries for connected pairs of stations and missing values for unconnected pairs.

From these two matrices of pairwise minimum transport times, we generated connectivity maps (MEMs) taking one value per station as follows (Legendre & Legendre, 2012). We first computed for each matrix a minimum spanning tree among samples using function *spantree* of R package ‘vegan’ (Oksanen et al., 2018). Following the recommendations of Legendre & Legendre (2012), we truncated the matrix of minimum transport times to retain only those connections necessary to connect all stations together (i.e., to obtain a connex graph), if possible. For surface samples, we found that a single tree connected all stations as long as we retained all minimum transport times below 2.1 years (which corresponds to distances up to a few thousands km, cf. Fig. S9). By doing so, we effectively restricted ourselves to the range of minimum transport times over which minimum transport time increases approximately linearly with the geographic distance between stations. For DCM samples, no single spanning tree connected all stations, and so we chose to retain all minimum transport times below 3.15 years, which led to the Mediterranean, the Red Sea and the Southern Ocean being disconnected from the remaining samples. In both matrices, we set the diagonals and all the elements above the selected threshold to four times the threshold value, and we conducted a PCoA of the resulting truncated connectivity matrices (function *pcoa.all*, Legendre 2007). We obtained 61 eigenvectors associated with strictly positive eigenvalues for the surface connectivity matrix and 35 for the DCM connectivity matrix, which we used as connectivity maps at the surface and the DCM.

The resulting connectivity maps display patterns of connectivity at temporal and spatial scales ranging from a few days and a hundred km (the minimal distance between a pair of stations) up to the global scale, and can therefore be used to assess the influence of transport by currents both within and between ocean basins (Fig. S10), which is difficult to achieve when directly using pairwise transport times between stations. They identify oceanographic features that are known to support high connectivity, such as the North Atlantic gyre system, the eastward flow between Scandinavia and Siberia in the Arctic Ocean, the South Pacific gyre, the Mediterranean Sea cyclonic circulation and the western Indian Ocean gyre system (Fig. S10).

Variation partitioning

To assess the influence of explanatory variables on biogeography, we compared their distribution across stations to that of assemblages through multivariate linear regression, after centering and standardization. We used the adjusted coefficient of multiple determination R_a^2 as a measure of the variance in the distribution of assemblages across stations (i.e., in the biogeography) that can be explained by a set of explanatory variables (function *rda*, R package ‘vegan’; Oksanen et al., 2018). Given a partition of the explanatory variables into two subsets A and B (e.g., connectivity maps and local environmental conditions), we partitioned the explained variance $R^2_{a,A\cap B}$ into the variance explained purely by subsets A and B as well as jointly by both subsets: $R^2_{a,A\cap B} = \cancel{R}^2_{a,A} + \cancel{R}^2_{a,B} + \cancel{R}^2_{a,A\cap B}$. This partitioning can be obtained from the variance independently explained by subsets A and B ($R^2_{a,A}$ and $R^2_{a,B}$) as follows (function *varpart*, R package ‘vegan’):

$$\cancel{R}^2_{a,A\cap B} = R^2_{a,A} + R^2_{a,B} - R^2_{a,A\cap B}$$
$$\cancel{R}^2_{a,A} = R^2_{a,A\cap B} - R^2_{a,B}$$
$$\cancel{R}^2_{a,B} = R^2_{a,A\cap B} - R^2_{a,A}$$

For each taxonomic group, we tested whether each variable individually explained a significant amount of variance in the biogeography (functions *rda* and *anova*), separately for the surface and DCM sets of samples, and we retained only the significant variables in further analyses.

We partitioned the variance explained by the combination of all retained variables into the following three fractions: the variance purely explained by connectivity maps, that purely explained by environmental variables (lumping biotic and abiotic variables together) and finally the variance jointly explained by both sets of variables (function *varpart*). We interpreted the fraction purely explained by connectivity maps as the part of the biogeography that can be attributed to transport by currents, through the homogenization of plankton communities at the local scale and through neutral structuring at the global scale. We interpreted the fraction purely explained by environmental variables as the part of biogeography that can be attributed to the response of community composition to local biotic and abiotic conditions. The jointly explained fraction is the part of the biogeography that is compatible with either of the two mechanisms. Some overlap is indeed to be expected between patterns of connectivity and environmental conditions, since environmental conditions are themselves transported by currents. Finally, the unexplained part of the variance can be interpreted as reflecting the effect of environmental variations along currents between stations, which are not taken into account in our analyses, unmeasured local abiotic and biotic parameters, local fluctuations in community composition, and sampling and measurement noise. We compared across taxonomic groups the following quantities: the total explained variance, the fraction of it purely explained by connectivity maps, the fraction of it purely explained by the local environment, and the ratio of the variance explained by connectivity (both purely and jointly) over that explained by the local environment (both purely and jointly).

We similarly partitioned the variance explained by the local environment into the variance purely explained by abiotic variables, that purely explained by biotic variables, and the variance jointly explained by both sets of variables, and compared them across taxonomic groups.

References:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyer, T. P., Antonov, J. I., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., … O’Brien, T. D. (2013). *World Ocean Database 2013.*

Chessel, D., Dufour, A., & Thioulouse, J. (2004). The ade4 Package - I: One-Table Methods. _R News_, (4(1)), 5–10.

Clayton, S., Dutkiewicz, S., Jahn, O., Hill, C., Heimbach, P., & Follows, M. J. (2017). Biogeochemical versus ecological consequences of modeled ocean physics. *Biogeosciences*, *14*(11), 2877–2889. doi: 10.5194/bg-14-2877-2017

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., … Tara Oceans, C. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237). doi:

10.1126/science.1261605

Grün, B., & Hornik, K. (2011). *topicmodels: an R package for fitting topic models*.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. doi: 10.1093/nar/gks1160

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., … Zinger, L. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, *179*(5), 1084-1097.e21. doi: 10.1016/j.cell.2019.10.008

Jönsson, B. F., & Watson, J. R. (2016). The timescales of global surface-ocean connectivity. *Nature Communications*, *7*, 11239. doi: 10.1038/ncomms11239

Legendre, P., & Legendre, L. (2012). *Numerical Ecology*. Elsevier.

Mahé, F., Rognes, T., Quince, C., Vargas, C. de, & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. doi: 10.7717/peerj.593

Meila, M. (2006). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873–895.

Menemenlis, D., Campin, J.-M., Heimbach, P., Hill, C., Lee, T., Schodlok, M., & Zhang, H. (2008). *ECCO2: High Resolution Global Ocean and Sea Ice Data Synthesis*. 10.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2018). *vegan: Community Ecology Package, version 2.5-2*.

Paradis, E., & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceeding of the 17th International Conference on World Wide Web - WWW '08*, 91. doi: 10.1145/1367497.1367510

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richter, D. J., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G., … Coordinators, T. O. (2019). Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *BioRxiv*, 867739. doi: 10.1101/867739

Sommeria-Klein, G., Zinger, L., Coissac, E., Iribar, A., Schimann, H., Taberlet, P., & Chave, J. (2019). Latent Dirichlet Allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13109

Watson, J. R., Hays, C. G., Raimondi, P. T., Mitarai, S., Dong, C., McWilliams, J. C., … Siegel, D. A. (2011). Currents connecting communities: nearshore community similarity and ocean circulation. *Ecology*, *92*(6), 1193–1200. doi: 10.1890/10-1436.1

Wilkins, D., van Sebille, E., Rintoul, S. R., Lauro, F. M., & Cavicchioli, R. (2013). Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nature Communications*, *4*(1). doi: 10.1038/ncomms3457