



INSTITUT DE FRANCE  
Académie des sciences

# *Comptes Rendus*

---

## *Biologies*


Paul Frémont, Marion Gehlen and Olivier Jaillon

**Plankton biogeography in the 21st century and impacts of climate change: advances through genomics**

Volume 346, issue 1 (2023), p. 13-24

Published online: 2 February 2023

<https://doi.org/10.5802/crbio.107>

 This article is licensed under the  
CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.  
<http://creativecommons.org/licenses/by/4.0/>



*Les Comptes Rendus. Biologies* sont membres du  
Centre Mersenne pour l'édition scientifique ouverte  
[www.centre-mersenne.org](http://www.centre-mersenne.org)  
e-ISSN : 1768-3238



---

Review Article / Article de revue

# Plankton biogeography in the 21st century and impacts of climate change: advances through genomics

## *Biogéographie du plancton au XXI<sup>e</sup> siècle et impacts du changement climatique : avancées par la génomique*

Paul Frémont<sup>✉</sup> \*, *a, b*, Marion Gehlen<sup>✉</sup> *c* and Olivier Jaillon<sup>✉</sup> *a, b*

<sup>a</sup> Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, Evry, France

<sup>b</sup> Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans, Paris, France

<sup>c</sup> LSCE-IPSL, CEA, CNRS, Université Paris-Saclay, Gif-sur-Yvette, France

*E-mails:* [pfremont@genoscope.cns.fr](mailto:pfremont@genoscope.cns.fr) (P. Frémont), [mgehlen@lsce.ipsl.fr](mailto:mgehlen@lsce.ipsl.fr) (M. Gehlen), [ojaillon@genoscope.cns.fr](mailto:ojaillon@genoscope.cns.fr) (O. Jaillon)

**Abstract.** This article summarizes recent advances in our knowledge of plankton biogeography obtained by genomic approaches and the impacts of global warming on it. Large-scale comparison of the genomic content of samples of different plankton size fractions revealed a partitioning of the oceans into genomic provinces and the impact of major oceanic currents on them. By defining ecological niches, these provinces are extrapolated to all oceans, with the exception of the Arctic Ocean. By the end of the 21st century, a major restructuring of these provinces is projected in response to a high emission greenhouse gas scenario over 50% of the surface of the studied oceans. Such a restructuring could lead to a decrease in export production by 4%. Finally, obtaining assembled sequences of a large number of plankton genomes defining this biogeography has allowed to better characterize the genomic content of the provinces and to identify the species structuring them. These genomes similarly enabled a better description of potential future changes of plankton communities under climate change.

**Résumé.** Cet article résume des avancées récentes dans nos connaissances sur la biogéographie du plancton obtenues par des approches de génomiques ainsi que les impacts du réchauffement climatique sur celle-ci. La comparaison à large échelle du contenu génomique d'échantillons de différentes fractions de taille de plancton a révélé un partitionnement des océans en provinces génomiques ainsi que l'impact des principaux courants océaniques sur celles-ci. En définissant des niches écologiques, ces provinces sont extrapolées à l'ensemble des océans à l'exception de l'océan Arctique. D'ici la fin

---

\* Corresponding author.

du XXI<sup>e</sup> siècle, une restructuration majeure de ces provinces est projetée en réponse à un scénario de fortes émissions de gaz à effet de serre sur 50 % de la surface des océans étudiés. Une telle restructuration pourrait engendrer une diminution de 4 % de la production exportée. L'obtention de séquences assemblées d'un grand nombre de génomes de plancton définissant cette biogéographie a permis de mieux caractériser le contenu génomique des provinces et d'identifier les espèces les structurant. Ces génomes ont aussi permis de mieux caractériser les futurs changements potentiels de communautés de plancton sous changement climatique.

**Keywords.** Plankton biogeography, Genomics, Climate change, Ocean, 21st century.

**Mots-clés.** Biogéographie du plancton, Génomique, Changement climatique, Océan, 21<sup>e</sup> siècle.

**Note.** Paul Frémont is 2022 recipient of the Grandes avancées en biologie Prize.

**Note.** Paul Frémont est lauréat 2022 du Prix des Grandes avancées en biologie.

*Published online: 2 February 2023*

## La version française de l'article est disponible à la suite de la version anglaise

The German naturalist Alexander Von Humboldt introduced the concept of biogeography in the 19th century. Following his expedition to South America, he described numerous plant species in his book *Kosmos: Entwurf einer physischen Weltbeschreibung* [1] and for the first time made the link between their geographical distribution and the associated climate [1]. Biogeography is thus the study of the spatial and temporal distribution of species in the present, through geological time and more recently in response to climate change. At the heart of this discipline, we find the notion of ecological niche, which we owe the intuition to Humboldt. It was mathematized by the zoologist George Evelyn Hutchinson in the 1950s [2] and is at the center of current research in contemporary ecology [3]. The ecological niche of a species or community of species is thus conceptualized as the hypervolume of environmental parameters allowing the life of this or these species [2].

Although it was first defined for terrestrial organisms, an ecological niche can be defined for marine organisms and in particular for the plankton that interests us here. But why are we interested in plankton? Several important factors lead us today to study these aquatic microorganisms. First of all, phytoplankton carries out photosynthesis and is the main contributor to the primary production in the oceans, which is almost equivalent in mass to the terrestrial production [4]. Thus, all food webs depend on this primary production, including remineralizing bacteria and large marine mammals (whales) but also humans (fisheries) [5]. Secondly, plankton drives the biological carbon pump, i.e. the export of or-

ganic carbon to the ocean floor and its storage in the sediments [6, 7]. This phenomenon occurs via multiple physical (particle aggregation) and biological processes (blooms, zooplankton grazing and fecal pellets) [6, 7].

Finally, the emergence of a spatial structuring of plankton in the oceans in the context of an environment in perpetual movement, which is the seascape [8], is a crucial and fundamental scientific question.

This question, linked to that of the ecological niches of plankton, is still very much debated among the scientific community. Today, in the context of global warming, these pivotal processes for the Earth's climate are endangered and it is more necessary than ever to better understand them to better anticipate and manage future global changes. Several 19th century expeditions included already scientific objectives [9]. However, marine ecosystems have remained the least understood and studied, which has led relatively recently to the launch of contemporary expeditions [10–12]. These recent expeditions are concomitant with the omics revolution that allows massive sequencing of plankton DNA, either from culture or from filtered water samples, or not, by metagenomics. This approach has greatly enriched our knowledge of marine plankton [10–13].

In this article, we summarize recent advances in our knowledge of plankton biogeography, which have been made possible by the production and analysis of metagenomic data from the *Tara* Oceans expeditions (2009–2013) through the results of three papers published this year 2022 [14–16].

The first article Richter *et al.* [14], the founding article, focuses on biogeographical patterns at the scale of the world's oceans (with the exception of the Arctic Ocean<sup>1</sup>) observed from metagenomic, metabarcoding and imaging data from the Tara Oceans expeditions from 2009 to 2012. The paper focuses on the link between biogeographical patterns and large-scale currents and, secondarily, environmental parameters.

The observed biogeographical patterns are based on estimates of beta-diversity, i.e. the level of genomic similarity between two communities. Estimates were made at 3 different scales between each pair of samples of the same size fraction of plankton from the expedition: at the metagenomic level, at the taxonomic level by the metabarcoding approach, and taxonomically by morphometric image classification. For metagenomics, a specific algorithm called Simka [17] was developed by a team from INRIA in Rennes to estimate beta-diversity by comparing the  $k$ -mer content ( $k$ -nucleotide subpart of the DNA sequence) of sample pairs. These beta-diversity data were compared with the minimum travel times followed by currents between stations calculated from the MITgcm circulation model [18]. By analyzing the correlation cumulatively, i.e., comparing samples only connected by a time  $t \leq t_0$  (gradually increasing  $t_0$ ), a maximum correlation at 1.5 years of transport ( $t_0 = 1.5$  years) between beta-diversity and transport time appeared for each of the six plankton size fractions ranging from viruses (0–0.2  $\mu\text{m}$ ) to small metazoans (180–2000  $\mu\text{m}$ ). This travel time of 1.5 years is consistent with the characteristic travel time of the surface current, and thus plankton, within studied oceanic basins. The beta-diversity of small size fractions (viruses 0–0.2  $\mu\text{m}$ , bacteria 0.22–3  $\mu\text{m}$  and protists 0.8–5 and 5–20  $\mu\text{m}$ ) was found to be more strongly correlated with these transport times supporting the hypothesis of a greater dependence of these organisms on currents (although this does not mean that these organisms are transported at different speeds).

The metabarcoding data were found to be relatively less well correlated with currents than the metagenomic data and even less so the imaging data.

This result demonstrates the ability of metagenomics (random sequencing of DNA fragments) to detect biogeographical patterns at high resolution. Metagenomics has a higher resolution than other methods because it is based on the comparison of millions of  $k$ -mers, compared to metabarcoding with a few hundred thousand Operational Taxonomic Units (OTUs) and whose resolution is very unequal according to the taxa with the risk of confusing some species. Finally, imaging techniques (limited to large size fractions) resolves only on a few hundred taxa with a significant risk of confusing species. In spite of the progress in sequencing, metagenomics most certainly provides a very small proportion of the DNA of communities, and therefore provides information that comes mainly from the most abundant organisms. As a result, very high-resolution information from a small number of abundant organisms provides a finer biogeographical signal than a fuzzy taxonomic description of much of the existing diversity.

The paper also links beta-diversity to environmental factors, showing a strong correlation between nutrients levels and biogeography of small organisms up to 1.5 years of transport while the influence of temperature is measured at larger scales. Finally, by grouping the sampling sites according to their beta-diversity using the UPGMA algorithm [19] (Unweighted Pair group Method with Arithmetic mean), the study defined genomic provinces for each plankton size fraction discussed above. These provinces are thus geographical areas, not necessarily continuous, characterized by assemblages of dominant organisms which have a very important genomic similarity and are most probably identical species. This result is strong and novel because it demonstrates for the first time a partitioning of the oceans based on genomics. This partitioning is likely stable because of the way the sampling was done and despite the fact that the samples studied were from a single temporal sample at each site. Indeed, while there is a good correlation between beta diversity and transport by currents, it should be noted that the transport times are model results, while the boat's path did not follow the currents either in time or space. Thus, the provinces have a relatively large average coverage in seasons and months (3 seasons and 5 months on average for a province). It should also be noted that this observation does not exclude a certain seasonal variability of the provinces in their biological content and

---

<sup>1</sup>The Arctic was not included in this study because the samples from the Tara Polar Circle expedition were not yet sequenced.

topology. The provinces are the basis of the following paper that extrapolates these genomic provinces to all oceans (with the exception of the Arctic Ocean). Interestingly, this first study also shows that this partitioning is variable according to plankton size fractions with smaller provinces for small organisms than for large organisms.

The second paper by Frémont *et al.* [15] extrapolates the genomic provinces of the first paper to the whole ocean by defining their environmental niches, studies the impact of climate change on them (displacement, expansion/shrinkage) and finally analyzes the potential ecological and biogeochemical impacts of this reorganization.

From a subset of 38 of the 48 original genomic provinces of the first article, this second study has first demonstrated, using machine learning techniques (random forest [20], boosted regression trees [21], single hidden layer neural network [22] and generalized additive models [23]), that 27 of them correspond to statistically defined environmental niches in a robust manner. This result was made possible by the statistical association between the provinces and the environmental parameters characterizing them: temperature, salinity, three macronutrients (nitrate, phosphate and dissolved silica), iron and a nitrate seasonality index. The choice of these parameters is based on the analyses of the first paper [14], although other parameters such as the mixed layer depth or the day length could also have been used. In this paper, the provinces were thus called climato-genomic provinces because they are characterized by their genomic content and their physico-chemical climate. A second major result is the identification of signature genomes (from the third article described below for eukaryotes [16] and another article for prokaryotes [24]) for these provinces whose biogeographies strongly coincide with the provinces themselves.

This structuring constitutes a strong and previously unknown result. Indeed, it shows that the structuring into relatively abstract genomic provinces (based only on the comparison of  $k$ -mers from metagenomes) can be explained by the presence of genomes, documenting the nature of metagenomic similarities. Using the global climatology from the World Ocean Atlas 2013 [25] of the different environmental parameters chosen, the different environmental niches were then extrapolated to the en-

tire oceans (with the exception of the Arctic Ocean) defining their partitioning for the six plankton size fractions used (Figure 2 from Frémont *et al.* [15]). Extrapolation to the whole ocean confirmed the differences in biogeographies of small and large organisms, implying trophic decoupling, and also showed a significant difference between genomic partitioning and known biogeographical partitioning [26–28] (supplementary of the article). Nevertheless, this result must be nuanced in the sense that the biogeographical structuring by the described climato-genomic provinces is only derived from a partial sampling of the oceans with an average distance of 300 km between two proximal samples. This result calls for new studies on the biogeography of plankton with repeated sampling over time and at increased spatial resolution as could be possible for example with sampling from future expeditions [29].

The second part of the article consisted in the analysis of the impact of climate change on the organization of genomic provinces using the average output from 6 Earth system models for the end of the century according to the RCP8.5 (Representative Concentration Pathway) scenario of high greenhouse gas emissions. Earth system models allow to project, according to physical and chemical laws, the dynamics of the climate system and of the major biogeochemical cycles (e.g. carbon, nitrogen, phosphorus) in response to different greenhouse gas emission scenarios. According to the RCP8.5 scenario, the atmosphere would reach CO<sub>2</sub> concentrations above 1000 ppm (parts per million) by 2100. These projections allow us to understand the evolution of the physico-chemical conditions of the ocean under climate change. Using these projections, it was possible to extrapolate the current genomic provinces to 2100 and compare their present and future distributions. The rates of displacement of the provinces found are comparable to those found in the literature for phytoplankton in particular [30] and the tropical provinces are expanding at the expense of the temperate provinces. Approximately 50% of the area of the considered surface oceans is projected to undergo a significant change in plankton community. This figure should be nuanced; it could probably be much higher. Indeed, in many areas, the environmental parameters projected by the end of the century exceed the current extreme values (tropics and Antarctic in particular, see supplementary

of the article) and are therefore outside the validity of machine learning techniques (our methods cannot project changes in these areas). Thus, this figure of 50% is relative to the current observed provinces, that is to say that 50% of the considered ocean is projected as being able to be colonized in the future by plankton communities existing today. The biological response in areas beyond the current conditions seems very uncertain and new selection pressures are expected. Importantly, fisheries and exclusive economic zones are largely included in the reorganization of genomic provinces. The restructuring of provinces is also accompanied by compositional changes in plankton community size and taxonomy (analysis based on the relative abundance of genomes from the third paper [16]). The analysis was restricted to three major plankton groups: copepods, phototrophs, and diazotrophs (organisms capable of metabolizing atmospheric di-nitrogen ( $N_2$ )). Projected changes in these groups could create new prey/predator trophic interactions by the end of the century.

The study also focused on the potential impacts of the reorganization of plankton biogeography on the biogeochemical cycles of carbon and nitrogen. First, an increase in diazotrophs in the tropico-equatorial Pacific Ocean is projected in agreement with other studies projecting an increase in diazotrophy in these areas [31]. Finally, biogeographical structuring into provinces and particularly the assemblage of these provinces (across the six size fractions) has also been shown to be a good statistical indicator of carbon export (at 100 m depth) to the ocean floor. Thus, an average decrease of 4% of the biological pump is projected for the year 2100 and is linked to community changes. This strong result is in agreement with current models [32] and would have a positive feedback effect (i.e. aggravating) on global warming.

To summarize, this study has allowed, for the first time, to generate a genomic-based plankton biogeography at the global scale (with the exception of the Arctic Ocean). Moreover, it allowed evaluating the consequences of climate change on plankton communities by 2100 under a high greenhouse gas emission scenario. Genomics is in line with many projections of current mechanistic models, demonstrating its potential and supporting the strength of these models. Some of the analysis in this study was en-

abled by the study in the third paper, which we will now present.

The third paper Delmont *et al.* [16] presents a collection of eukaryotic Metagenome-Assembled Genomes (MAGs) from the metagenomic data of the Tara Oceans expeditions (2009–2013). The main result of this paper is the creation of a collection of 683 eukaryotic MAGs and 30 SAGs (single cell genomes) from the metagenomic data of 5 eukaryotic enriched size fractions of the Tara Oceans expeditions (from 0.8  $\mu\text{m}$  to 2 mm). The assembly by the MEGAHIT algorithm [33] used a total of 280 billion Illumina sequencing reads from 939 metagenomes sampled across 143 sites and covering a large part of the world's oceans (including the Arctic Ocean samples from the *Tara* Polar Circle expedition). However, the sequence of these genomes is often incomplete. The genomes generated are estimated to be  $40 \pm 24\%$  complete on average (up to 93.7%), with sizes ranging from 10 Mbp (Mega base pair) to 1.3 Gbp (Giga base pair). They correspond in total to 39.1 billion of the 280 billion reads submitted to the assembler (14.0%). As the assembler could not absorb all the metagenomic data at once (memory limit), the MAGs were obtained by 11 subsets of metagenomic data that were each co-assembled (each containing between 6 and 38 billion reads). These sets correspond to geographically coherently distributed collection sites.

In this collection of MAGs, a set of 10 million genes has been identified. In a novel way, it covers a set of poorly characterized eukaryotic lineages and broadly extends the cultivation efforts of mainly coastal marine microorganisms. A phylogeny of the collection was generated from the DNA encoding RNA polymerase and includes most of the known major groups of marine microbial eukaryotes, such as the Opisthokonts (mainly copepods), Archaeplastida, Stramenopiles, Alveolates, Rhizaria (the last three forming the SAR supergroup), Cryptista, Excavata, Haptista, and a new putative group. However, we note the absence of the Dinoflagellates, a very important and abundant group but with very large genomes [34]. A much larger sequencing effort would be required for genome assembly of these organisms. A classification of the genomes has also been carried out on the basis of the functional annotation of the genes constituting them. This classification revealed functional convergences between phylogenetically

distant groups such as the diatoms and other algae such as the Mamiellales. Finally, the modeling of the ecological niches of the 713 genomes was tested and, in total, 374 statistically robust niches (52%) could be defined. The same methodology and environmental parameters as in the second paper [15] were used for this analysis.

All the results of this analysis and the climate projections at the end of the century have been compiled on a website ([http://end.mio.osupytheas.fr/Ecological\\_Niche\\_database/](http://end.mio.osupytheas.fr/Ecological_Niche_database/)). It is hosted at the Oceanographic Institute of Marseille (MIO). Interestingly, the MAG whose niche expands the most by the end of the century corresponds to a tropico-equatorial copepod, as does the MAG whose niche shrinks the most. These results are consistent with a strong potential reorganization of the marine food chain in the global oceans. The site also includes a tab presenting the climatologies of the different environmental parameters as well as a tab allowing the user to define virtual environmental niches and visualize them at the present time and at the end of the century.

To summarize, these three papers based on the same dataset have significantly incremented the knowledge of plankton biogeography through metagenomics. The genomes and provinces spatially structuring plankton in the oceans were previously

### ***Version française***

La notion de biogéographie a été introduite par le naturaliste allemand Alexander Von Humboldt au XIX<sup>e</sup> siècle. Il décrit, suite à son voyage en Amérique du Sud, dans son ouvrage *Kosmos: Entwurf einer physischen Weltbeschreibung* [1] de nombreuses espèces végétales et fait pour la première fois le lien entre leur répartition géographique et le climat associé [1]. La biogéographie consiste ainsi en l'étude de la répartition spatiale et temporelle des espèces aussi bien actuelle, à travers les temps géologiques et plus récemment en réponse au changement climatique. Au cœur de cette discipline, on retrouve la notion de niche écologique dont on doit l'intuition à Humboldt. Elle est mathématisée par le zoologiste George Evelyn Hutchinson dans les années 1950 [2] et se retrouve au centre des recherches actuelles de l'écologie contemporaine [3]. La niche écologique

unknown. This result is new because of its fine resolution linked to genomics and the spatial decoupling between the biogeographies of the different plankton size fractions. This implies a decoupling at the trophic level between prey (phytoplankton) and predators (zooplankton) at the scale of the global oceans. The potential impact of climate change on the biogeography of the different planktonic organisms has been evaluated by the end of the century in a scenario of high greenhouse gas emissions. It is projected that the current climate-genomic provinces could reorganize over about 50% of the oceans (excluding the Arctic Ocean) by 2100. In addition, tropical provinces would expand at the expense of temperate provinces. Many future ocean areas could reach climatic extremes exceeding current maximum (or minimum) values of temperature and nutrient levels. In these areas, the current provinces are maintained, however, we can expect the disappearance of some niches and the appearance of new communities under selection pressure unknown to date. Finally, the restructuring of the biogeography of plankton could lead to a decrease of 4% of the biological carbon pump leading to an increase in global warming.

### **Conflicts of interest**

Authors have no conflict of interest to declare.

d'une espèce ou communauté d'espèces est ainsi conceptualisée comme l'hypervolume des paramètres environnementaux permettant la vie de cette ou ces espèces [2].

Bien qu'elle ait d'abord été définie pour les organismes terrestres, on peut définir une niche écologique pour les organismes marins et en particulier pour le plancton qui nous intéresse ici. Mais pourquoi s'intéresser au plancton? Plusieurs facteurs d'importance nous poussent aujourd'hui à étudier ces micro-organismes aquatiques. Tout d'abord, le phytoplancton, réalise la photosynthèse et contribue ainsi fortement à la production primaire réalisée dans les océans, quasi équivalente en masse à la production terrestre [4]. Ainsi, l'ensemble des réseaux trophiques dépend de cette production primaire, incluant les bactéries reminéralisatrices, les

grands mammifères marins (baleines) mais aussi les hommes (pêcheries) [5]. Ensuite, le plancton est en partie à l'origine de la pompe à carbone biologique c'est-à-dire l'export de carbone organique vers les fonds océaniques et son stockage dans les sédiments [6, 7]. Ce phénomène se produit *via* de multiples processus physiques (agrégations particulières) et biologiques (floraisons, broutage par le zooplancton et production de boulettes fécales) [6, 7].

Enfin, l'émergence d'une structuration spatiale du plancton dans les océans dans le contexte d'un environnement en perpétuel mouvement, qu'est le *seascapes* [8], est une question scientifique cruciale et fondamentale.

Cette question liée à celle des niches écologiques du plancton est encore très débattue parmi la communauté scientifique. Aujourd'hui dans le contexte du changement climatique, ces processus pivots pour le climat terrestre sont mis en danger et il est plus que jamais nécessaire de mieux les comprendre pour mieux anticiper et gérer les changements globaux à venir. Certaines expéditions du XIX<sup>e</sup> siècle ont inclus des objectifs scientifiques [9]. Cependant, les écosystèmes marins sont restés les moins compris et étudiés, ce qui a conduit relativement récemment au lancement d'expéditions contemporaines [10–12]. Ces récentes expéditions sont concomitantes de la révolution omique qui permet un séquençage massif d'ADN de plancton, soit à partir de culture soit à partir de prélèvements d'eau filtrée, ou non, par métagénomique. Cette approche a largement enrichi nos connaissances sur le plancton marin [10–13].

Dans cet article, nous faisons la synthèse d'avancées récentes sur nos connaissances de la biogéographie du plancton, qui ont été permises par la production et l'analyse des données métagénomique des expéditions *Tara Océans* (2009–2013) à travers les résultats de trois articles publiés cette année 2022 [14–16].

Le premier article Richter *et al.* [14], fondateur, porte sur les structures biogéographiques à l'échelle de l'ensemble des océans du globe (à l'exception de l'Arctique<sup>2</sup>) observés à partir des données de métagénomiques, de métabarcoding et d'imagerie des expéditions *Tara Océans* de 2009 à 2012. L'article est centré sur le lien entre les structures biogéographiques

et les courants de large échelle et, de manière plus secondaire, les paramètres environnementaux.

Les structures biogéographiques observées sont basées sur des estimations de diversité bêta c'est-à-dire le niveau de similarité génomique entre deux communautés. Des estimations ont été réalisées à 3 échelles différentes entre chaque paire d'échantillons d'une même fraction de taille de plancton de l'expédition : au niveau métagénomique, au niveau taxonomique par l'approche de métabarcoding, et taxonomique par classification morphométrique d'images. Pour la métagénomique, un algorithme spécifique appelé Simka [17] a été développé par une équipe de l'INRIA de Rennes pour estimer la diversité bêta par comparaison du contenu en  $k$ -mer (sous partie de  $k$  nucléotides de la séquence d'ADN) des paires d'échantillons. Ces données de diversité bêta ont été comparées avec les temps de parcours minimaux suivis par les courants entre stations calculés à partir du modèle de circulation MITgcm [18]. En analysant la corrélation de façon cumulative, c'est-à-dire en comparant les échantillons uniquement connectés par un temps  $t \leq t_0$  (en augmentant petit à petit le  $t_0$ ), un maximum de corrélation à 1,5 an de transport ( $t_0 = 1,5$  an) entre la diversité bêta et le temps de transport est apparu pour chacune des six fractions de taille de plancton allant des virus (0–0,2  $\mu\text{m}$ ) aux petits métazoaires (180–2000  $\mu\text{m}$ ). Ce temps de parcours de 1,5 an correspond au temps caractéristique de parcours du courant de surface, et donc du plancton, au sein des bassins océaniques étudiés. La diversité bêta des petites fractions de taille (virus 0–0,2  $\mu\text{m}$ , bactéries 0,22–3  $\mu\text{m}$  et protistes 0,8–5 et 5–20  $\mu\text{m}$ ) s'est avérée être corrélée plus fortement à ces temps de transport appuyant l'hypothèse d'une plus grande dépendance de ces organismes aux courants (sans pour autant signifier que ces organismes soient transportés à des vitesses différentes).

Les données de métabarcoding se sont avérées être relativement moins bien corrélées aux courants que les données de métagénomique et encore moins celle d'imagerie. Ce résultat montre la capacité de la métagénomique (séquençage aléatoire de fragments d'ADN) à détecter à haute résolution des structures biogéographiques. La métagénomique a une plus grande résolution que les autres méthodes car elle se base sur la comparaison de millions de  $k$ -mers, le métabarcoding sur quelques centaines de milliers d'Unités Taxonomiques Opérationnelles (OTUs) et

---

<sup>2</sup>L'Arctique n'a pas été inclut à cette étude car les échantillons de l'expédition *Tara Polar Circle* n'étaient pas encore séquencés.



dont la résolution est très inégale selon les taxons avec le risque de confondre certaines espèces, et l'imagerie (ici les larges fractions de taille) seulement sur quelques centaines de taxons résolus avec un risque important de confusion d'espèces. Malgré les progrès de séquençage, la métagénomique fournit très certainement une très petite proportion de l'ADN des communautés, et donc fournit de l'information qui provient majoritairement des organismes les plus abondants. Il en résulte qu'une information à très haute résolution d'un petit nombre d'organismes abondants apporte un signal biogéographique plus fin qu'une description taxonomique floue d'une grande partie de la diversité existante.

L'article fait aussi le lien entre diversité bêta et facteurs environnementaux, montrant une forte corrélation entre les concentrations en nutriments et la biogéographie des petits organismes jusqu'à 1,5 année de transport tandis que l'influence de la température se mesure à plus large échelle. Enfin, en groupant les sites d'échantillonnages en fonction de leur diversité bêta à l'aide de l'algorithme UPGMA [19] (*Unweighted Pair group Method with Arithmetic mean*), l'étude a permis de définir des provinces génomiques pour chaque fraction de taille de plancton évoquée précédemment. Ces provinces sont donc des zones géographiques, non nécessairement continues, contenant des organismes parmi les plus abondants et qui ont une similarité génomique très importante, très probablement des espèces identiques. Ce résultat est fort et nouveau car il démontre un partitionnement des océans basés pour la première fois sur la génomique. Ce partitionnement est probablement stable en raison de la façon dont l'échantillonnage a été réalisé et malgré le fait que les échantillons étudiés ne soient issus que d'un unique prélèvement temporel à chaque site. En effet, alors qu'il existe une bonne corrélation entre diversité bêta et transport par les courants, il faut noter que les temps de transport sont des résultats de modèles, alors que le parcours du bateau n'a pas suivi les courants ni dans le temps ni l'espace. Ainsi, les provinces ont une couverture moyenne en saisons et en mois relativement importante (3 saisons et 5 mois en moyenne pour une province). A noter aussi que ce constat n'exclut pas une certaine variabilité saisonnière des provinces dans leur contenu biologique et dans leur topologie. Les provinces sont à la base de l'article suivant qui extrapole

ces provinces génomiques à l'ensemble des océans (à l'exception de l'Arctique). De manière intéressante, cette première étude montre aussi que ce partitionnement est variable selon les fractions de taille de plancton avec de plus petites provinces pour les petits organismes que pour les grands organismes.

Le second article Frémont *et al.* [15] extrapole à l'ensemble des océans les provinces génomiques du premier article en définissant les niches environnementales de celles-ci, étudie l'impact du changement climatique sur celles-ci (déplacement, expansion/rétrécissement) et enfin analyse les potentiels impacts écologiques et biogéochimiques de cette réorganisation.

A partir d'un sous-ensemble de 38 des 48 provinces génomiques originales du premier article, cette seconde étude a dans un premier temps mis en évidence, à l'aide de techniques de machine learning (*random forest* [20], *boosted regression trees* [21], *single hidden layer neural network* [22] et *generalized additive models* [23]), que 27 d'entre elles correspondent à des niches environnementales définies statistiquement de manière robuste. Ce résultat a été permis par l'association statistique entre les provinces et les paramètres environnementaux les caractérisant : la température, la salinité, trois macronutriments (nitrate, phosphate et silice dissous), le fer et un indice de saisonnalité du nitrate. Le choix de ces paramètres est basé sur les analyses du premier article [14], bien que d'autres paramètres tels que la profondeur de la couche de mélange ou la durée du jour auraient aussi pu être utilisés. Dans cet article les provinces ont ainsi été appelées provinces climato-génomiques du fait qu'elles sont caractérisées par leur contenu génomique et leur climat physico-chimique. Un second résultat majeur est la mise en évidence de génomes signatures (issus du troisième article décrit plus bas pour les eucaryotes [16] et d'un autre article pour les procaryotes [24]) pour ces provinces dont les biogéographies coïncident fortement avec les provinces elles-mêmes.

Cette structuration constitue un résultat fort et inconnu auparavant. En effet, il montre que la structuration en provinces génomiques relativement abstraites initialement (car basées uniquement sur la comparaison des *k*-mers issus des métagénomiques) s'explique par la présence de génomes permettant de documenter la nature des similarités

métagénomiques. A l'aide des climatologies globales du *World Ocean Atlas* 2013 [25] des différents paramètres environnementaux choisis, les différentes niches environnementales ont ensuite été extrapolées à l'ensemble des océans (excepté l'arctique) définissant leur partitionnement pour les six fractions de tailles de plancton utilisées (Figure 2 de Frémont *et al.* [15]). L'extrapolation à l'ensemble des océans a confirmé les différences des biogéographies des petits et grands organismes, impliquant un découplage trophique, et a aussi permis de montrer une différence importante entre les partitionnements génomiques et les partitionnements biogéographiques connus [26–28] (suppléments de l'article). Néanmoins, ce résultat est sûrement à nuancer au sens où la structuration biogéographique par les provinces climato-génomiques décrites n'est issue que d'un échantillonnage partiel des océans avec une distance moyenne de 300 km entre deux échantillons proximaux. Ce résultat appelle donc à de nouvelles études sur la biogéographie du plancton avec des prélèvements répétés dans le temps ainsi qu'à des échelles plus résolutive spatialement comme pourrait le permettre par exemple l'échantillonnage issu de futures expéditions [29].

Le second volet de l'article a consisté en l'analyse de l'impact du changement climatique sur l'organisation des provinces génomique à l'aide de sorties moyennées d'un ensemble de modèles système Terre à l'horizon de la fin du siècle selon le scénario RCP8.5 (*Representative Concentration Pathway*) d'émission importante de gaz à effet de serre. Les modèles système Terre permettent de projeter, selon des lois physiques et chimiques, la dynamique du système climatique et des grands cycles biogéochimiques (par exemple du carbone, de l'azote et du phosphore) en réponse à différents scénarios d'émission de gaz à effet de serre. Selon le scénario RCP8.5, l'atmosphère atteindrait des concentrations en CO<sub>2</sub> supérieures à 1000 ppm (parties par millions) d'ici 2100. Ces projections permettent d'appréhender l'évolution des conditions physico-chimiques de l'océan sous changement climatique. A l'aide de ces projections, il a été possible d'extrapoler les provinces génomiques actuelles à l'horizon 2100 et comparer leurs distributions présentes et futures. Les taux de déplacement des provinces trouvés sont comparables à ceux existant dans la littérature pour le phytoplancton notamment [30] et les provinces tropicales s'étendent

au détriment des provinces tempérées. Environ 50 % de l'aire des océans de surface considérés est projetée comme pouvant subir un changement important de communauté planctonique. Ce chiffre est à nuancer, il pourrait être sûrement bien plus élevé. En effet dans de nombreuses zones, les paramètres environnementaux projetés d'ici la fin du siècle dépassent les valeurs extrêmes présentes (tropiques et antarctique notamment, voir suppléments de l'article) donc en dehors du domaine de validité des techniques de machine learning (nos méthodes ne peuvent pas projeter de changements dans ces zones). Ainsi, ce chiffre de 50 % est relatif aux provinces actuelles observées, c'est-à-dire que 50 % de l'océan considéré est projeté comme pouvant à l'avenir être colonisé par des communautés de plancton existantes aujourd'hui. La réponse biologique dans les zones dépassant les extrêmes actuels semble très incertaine et de nouvelles pressions de sélection y sont attendues. De manière importante, les zones de pêches et zones économiques exclusives sont largement comprises dans la réorganisation des provinces génomiques. La restructuration des provinces s'accompagne aussi de changements compositionnels des communautés de plancton en taille et taxonomie (analyse basée sur l'abondance relative des génomes du troisième article [16]). L'analyse a été restreinte à trois grands groupes de plancton : les copépodes, les phototrophes et les diazotrophes (organismes capables de métaboliser le diazote atmosphérique (N<sub>2</sub>)). Les changements projetés dans ces groupes pourraient notamment créer de nouvelles interactions trophiques proie/prédateur d'ici la fin du siècle.

L'étude s'est aussi attachée aux impacts potentiels de la réorganisation de la biogéographie du plancton sur les cycles biogéochimiques du carbone et de l'azote. Tout d'abord, une augmentation des diazotrophes dans l'océan pacifique tropico-équatorial est projetée en accord avec d'autres études projetant une augmentation de la diazotrophie dans ces zones [31]. Enfin, la structuration biogéographique en provinces et particulièrement l'assemblage de ces provinces (à travers les six fractions de taille) s'est aussi avérée être un bon indicateur statistique de l'export de carbone (à 100 m de profondeur) vers les fonds océaniques. Ainsi, une diminution moyenne de 4 % de la pompe biologique est projetée à l'horizon 2100 et est mise en lien avec les changements de communautés. Ce résultat fort est en accord avec les

modèles actuels [32] et aurait un effet de rétroaction positive (c'est-à-dire aggravant) sur le réchauffement climatique.

Pour résumer, cette étude a permis, pour la première fois, de générer une biogéographie génomique du plancton à l'échelle globale (à l'exception de l'océan Arctique). De plus, elle a permis d'évaluer les conséquences du changement climatique sur les communautés de plancton à l'horizon 2100 selon un scénario d'importantes émissions de gaz à effet de serre. De manière intéressante, la génomique va dans le même sens que de nombreuses projections des modèles mécanistiques actuels, démontrant son potentiel et confortant la force de ces modèles. Une partie des analyses de cette étude a été permise par l'étude du troisième article que nous présenterons maintenant.

Le troisième article Delmont *et al.* [16] présente une collection de Génomes Assemblés par Métagénomiques (MAGs) eucaryotes issue des données métagénomiques des expéditions *Tara Océans* (2009–2013). Le principal résultat de cet article est la création d'une collection de 683 MAGs et 30 SAGs (génomomes cellule unique) eucaryotes à partir des données métagénomiques de 5 fractions de taille enrichies en eucaryotes des expéditions *Tara Océans* (de 0.8  $\mu\text{m}$  à 2 mm). L'assemblage par l'algorithme MEGAHIT [33] a utilisé au total 280 milliards de lectures de séquençage Illumina issus de 939 métagénomomes échantillonnés à travers 143 sites et couvrant une grande partie des océans mondiaux (incluant les échantillons de l'océan l'Arctique issu de l'expédition *Tara Polar Circle*). Cependant la séquence de ces génomes est souvent incomplète, les génomes générés sont estimés complets en moyenne à  $40 \pm 24\%$  (allant jusqu'à 93,7 %), avec des tailles allant de 10 Mbp (Méga paire de bases) à 1,3 Gbp (Giga paire de bases). Ils correspondent au total à 39,1 milliards des 280 milliards de lectures soumises à l'assembleur (14,0 %). L'assembleur ne pouvant absorber d'un seul coup l'ensemble des données métagénomique (limite mémoire), les MAGs ont été obtenus par 11 sous-ensembles de données métagénomiques qui ont été chacun co-assemblés (contenant chacun entre 6 et 38 milliards de lectures). Ces ensembles correspondent à des sites de prélèvement répartis de manière cohérente géographiquement.

Dans cette collection de MAGs, un ensemble de 10 millions de gènes a été identifié. De manière nova-

trice, elle couvre un ensemble de lignées eucaryotes mal caractérisées et étendant largement les efforts de culture de micro-organismes marins principalement côtiers. Une phylogénie de la collection a été générée à partir de l'ADN codant pour l'ARN polymérase et regroupe la plupart des grands groupes d'eucaryotes microbiens marins connus, tels que les Opisthocontes (principalement des copépodes), les Archaeplastida, les Straménopiles, les Alvéolés, les Rhizaria (les trois derniers formant le supergroupe SAR), les Cryptista, les Excavata, les Haptista et un nouveau groupe putatif. On note toutefois l'absence des Dinoflagellés, groupe très important et abondant mais aux génomes très grands [34]. Un effort de séquençage bien plus important serait nécessaire pour l'assemblage du génome de ces organismes. Un classement des génomes a aussi été réalisé sur la base de l'annotation fonctionnelle des gènes les constituant. Ce classement a notamment révélé des convergences fonctionnelles entre groupes distants phylogénétiquement comme les diatomées et d'autres algues telles que les Mamiellales. Enfin, la modélisation des niches écologiques des 713 génomes a été testée et, au total, 374 niches robustes statistiquement (52 %) ont pu être définies. La même méthodologie et les mêmes paramètres environnementaux que pour le second article [15] ont été utilisés pour cette analyse.

L'ensemble des résultats de cette analyse et les projections climatiques à la fin du siècle ont été compilés sur un site internet ([http://end.mio.osupytheas.fr/Ecological\\_Niche\\_database/](http://end.mio.osupytheas.fr/Ecological_Niche_database/)). Il est hébergé à l'institut océanographique de Marseille (MIO). Le MAG dont la niche s'étend le plus d'ici la fin du siècle correspond à un copépode tropico-équatorial, de même que le MAG dont la niche se rétrécit le plus. Ces résultats vont dans le sens d'une réorganisation potentielle forte de la chaîne trophique marine des océans globaux. Le site comporte aussi un onglet présentant les climatologies des différents paramètres environnementaux ainsi qu'un onglet permettant à l'utilisateur de définir des niches environnementales virtuelles et de les visualiser au temps présent et à la fin du siècle.

Pour résumer, ces trois articles basés sur le même jeu de données ont permis d'incrémenter significativement les connaissances sur la biogéographie du plancton grâce à la métagénomique. Ainsi, les génomes et provinces structurant spatialement le plancton dans les océans étaient jusqu'alors

inconnus. Ce résultat est nouveau par sa résolution fine liée à la génomique et le découplage partiel entre les biogéographies des différentes fractions de taille de plancton. Ainsi il existe un découplage au niveau trophique entre proies (phytoplancton) et prédateurs (zooplancton) à l'échelle des océans globaux. L'impact potentiel du réchauffement climatique sur la biogéographie des différents organismes planctoniques a été évalué d'ici la fin du siècle dans un scénario d'émissions importantes de gaz à effet de serre. Il est ainsi projeté que les provinces climato-génomiques actuelles pourraient se réorganiser sur environ 50 % des océans (hors océan Arctique) d'ici 2100. De plus, les provinces tropicales s'étendraient au détriment des provinces tempérées. De nombreuses zones océaniques futures pourraient atteindre des climats extrêmes dépassant les valeurs maximales (ou minimales) actuelles de température et concentrations en nutriments. Dans ces zones, les provinces actuelles sont maintenues, cependant on peut s'attendre à la disparition de certaines niches et à l'apparition de nouvelles communautés sous pression de sélection inconnue à ce jour. Enfin, la restructuration de la biogéographie du plancton pourrait engendrer une diminution de 4 % de la pompe à carbone biologique entraînant un aggravement du réchauffement climatique.

## Conflit d'intérêt

Les auteurs n'ont aucun conflit d'intérêt à déclarer.

## References

- [1] A. von Humboldt, *Kosmos: Entwurf einer physischen Weltbeschreibung*, Cotta, Stuttgart/Tübingen, 1847.
- [2] G. E. Hutchinson, "Concluding remarks", in *Cold Spring Harbor Symposia on Quantitative Biology*, Cold Spring Harbor Laboratory Press, UK, 1957.
- [3] A. Pocheville, "The Ecological Niche: History and Recent Controversies", in *Handbook of Evolutionary Thinking in the Sciences* (T. Heams, P. Huneman, G. Lecointre, M. Silberstein, eds.), Springer, Dordrecht, 2015, p. 547-586.
- [4] C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, "Primary production of the biosphere: Integrating terrestrial and oceanic components", *Science* **281** (1998), no. 5374, p. 237-240.
- [5] F. Azam, T. Fenchel, J. Field, J. Gray, L. Meyer-Reil, F. Thingstad, "The Ecological Role of Water-Column Microbes in the Sea", *Mar. Ecol. Prog. Ser.* **10** (1983), no. 3, p. 257-263.
- [6] R. W. Eppley, B. J. Peterson, "Particulate organic matter flux and planktonic new production in the deep ocean", *Nature* **282** (1979), no. 5740, p. 677-680.
- [7] L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. R. Brum, L. P. Coelho, J. C. I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, S. Kandels-Lewis, M. Picheral, J. Poulain, S. Searson, L. Stemmann, F. Not, P. Hingamp, S. Speich, M. Follows, L. Karp, D. Iudicone, M. B. Sullivan, J. Raes, E. Karsenti, C. Bowler, G. Gorsky, "Plankton networks driving carbon export in the oligotrophic ocean", *Nature* **532** (2016), p. 465-470.
- [8] M. T. Kavanaugh, M. J. Oliver, F. P. Chavez, R. M. Letelier, F. E. Muller-Karger, S. C. Doney, "Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation", *ICES J. Mar. Sci.* **73** (2016), no. 7, p. 1839-1850.
- [9] M. Sears, D. Merriman, *Oceanography: The Past*, 1st ed., Springer-Verlag, New York, 1980.
- [10] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Neelson, R. Friedman, M. Frazier, J. C. Venter, "The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific", *PLoS Biol.* **5** (2007), no. 3, article no. e77.
- [11] C. M. Duarte, "Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition", *Limnol. Oceanogr. Bull.* **24** (2015), p. 11-14.
- [12] E. Karsenti, S. G. Acinas, P. Bork, C. Bowler, C. de Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzonzi, J. M. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, P. Wincker, C. Abergel, D. Arslan, S. Audic, J. M. Aury, N. Babic, L. Beaufort, L. Bittner, E. Boss, C. Boutte, J. Brum, M. Carmichael, R. Casotti, A. Chambouvet, P. Chang, C. Chica, C. Clerissi, S. Colin, F. M. Cornejo-Castillo, C. Da Silva, S. De Monte, J. Decelle, Y. Desdesvres, C. Dimier, J. Dolan, M. Duhaime, X. Durrieu de Madron, F. d'Ortenzio, F. d'Ovidio, I. Ferrera, L. Garczarek, M. J. Garet-Delmas, S. Gasmí, J. M. Gasol, N. Grimsley, R. Heilig, J. Ignacio-Espinoza, J. L. Jamet, L. Karp-Boss, M. Katinka, H. Khalili, Z. Kolber, N. Le Bescot, H. Le Goff, G. Lima-Mendez, F. Mahé, M. G. Mazzocchi, M. Montrosor, P. Morin, B. Noel, C. Pedrós-Alió, E. Pelletier, Y. Perez, M. Picheral, G. PíganEAU, O. Poirot, J. Poulain, N. Poulton, F. Prejger, J. Prihoda, I. Probert, J. Rampal, G. Reverdin, S. Romac, J. B. Romagnan, F. Roullier, C. Rouviere, G. Samson, S. Santini, H. Sarmiento, A. Sciandra, S. Solonenko, L. Stemmann, L. Subirana, S. Sunagawa, A. Tanaka, P. Testor, A. Thompson, V. Tichanné-Seltzer, L. Tirichine, E. Toulza, S. Tozzi, A. Veluchamy, A. Zingone, "A holistic approach to marine Eco-systems biology", *PLoS Biol.* **9** (2011), no. 10, article no. e1001177.
- [13] S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, L. Karp-Boss, E. Karsenti, M. Lescot, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes,

- C. Sardet, M. Sieracki, S. Speich, L. Stemann, M. B. Sullivan, S. Sunagawa, P. Wincker, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard, H. Ogata, S. Pesant, M. B. Sullivan, P. Wincker, C. de Vargas, "Tara Oceans: towards global ocean ecosystems biology", *Nat. Rev. Microbiol.* **18** (2020), p. 428-445.
- [14] D. J. Richter, R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, O. Da Silva, T. O. Delmont, A. Fernández-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, T. O. Coordinators, S. Pesant, J.-M. Aury, J. R. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, F. Lombard, L. Karp-Boss, C. Bowler, M. B. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. Ribera d'Alcalà, D. Iudicone, O. Jaillon, "Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems", *eLife* **11** (2022), article no. e78129.
- [15] P. Frémont, M. Gehlen, M. Vrac, J. Leconte, T. O. Delmont, P. Wincker, D. Iudicone, O. Jaillon, "Restructuring of plankton genomic biogeography in the surface ocean under climate change", *Nat. Clim. Change* **12** (2022), p. 393-401.
- [16] T. O. Delmont, M. Gaia, D. D. Hingsinger, P. Frémont, C. Vanni, A. Fernandez-Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. Da Silva, M. Wessner, B. Noel, J.-M. Aury, S. Sunagawa, S. G. Acinas, P. Bork, E. Karsenti, C. Bowler, C. Sardet, L. Stemann, C. de Vargas, P. Wincker, M. Lescot, M. Babin, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, O. Jaillon, S. Kandels, D. Iudicone, H. Ogata, S. Pesant, M. B. Sullivan, F. Not, K.-B. Lee, E. Boss, G. Cochrane, M. Follows, N. Poulton, J. Raes, M. Sieracki, S. Speich, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, "Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean", *Cell Genomics* **2** (2022), no. 5, article no. 100123.
- [17] G. Benoit, P. Peterlongo, M. Mariadassou, E. Drezen, S. Schbath, D. Lavenier, C. Lemaitre, "Multiple comparative metagenomics using multisets *k*-mer counting", *PeerJ Comput. Sci.* **2016** (2016), no. 11, article no. e94.
- [18] S. Clayton, S. Dutkiewicz, O. Jahn, C. Hill, P. Heimbach, M. J. Follows, "Biogeochemical versus ecological consequences of modeled ocean physics", *Biogeosciences* **14** (2017), no. 11, p. 2877-2889.
- [19] R. R. Sokal, "A statistical method for evaluating systematic relationships", *Univ. Kans. Sci. Bull.* **38** (1958), no. 22, p. 1409-1438.
- [20] L. Breiman, "Random forests", *Mach. Learn.* **45** (2001), p. 5-32.
- [21] G. Ridgeway, *Generalized Boosted Models: A guide to the gbm package*, vol. 1, 2007.
- [22] W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S*, Springer, New York, 2002.
- [23] S. N. Wood, "Stable and efficient multiple smoothing parameter estimation for generalized additive models", *J. Am. Stat. Assoc.* **99** (2004), no. 467, p. 673-686.
- [24] T. O. Delmont, J. J. P. Karlusich, I. Veseli, J. Fuessel, A. M. Eren, R. A. Foster, C. Bowler, P. Wincker, E. Pelletier, "Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean", *ISME J.* **16** (2022), p. 927-936.
- [25] T. P. Boyer, J. I. Antonov, O. K. Baranova, C. Coleman, H. E. Garcia, A. Grodzky, D. R. Johnson, R. A. Locarnini, A. V. Mishonov, T. D. O'Brien, C. R. Paver, J. R. Reagan, D. Seidov, I. V. Smolyar, M. M. Zweng, T. D. O'Brien, C. R. Paver, J. R. Reagan, D. Seidov, I. V. Smolyar, M. M. Zweng, K. D. Sullivan, in *World ocean atlas database 2013*, NOAA Atlas NESDIS, vol. 72, Sydney Levitus, Ed.; Alexey Mishonoc, Technical Ed., 2013, National Oceanographic Data Center (U.S.), Ocean Climate Laboratory.
- [26] A. R. Longhurst, *Ecological Geography of the Sea*, Elsevier Inc., Amsterdam, Netherlands, 2007.
- [27] G. Reygondeau, A. Longhurst, E. Martinez, G. Beaugrand, D. Antoine, O. Maury, "Dynamic biogeochemical provinces in the global ocean", *Glob. Biogeochem. Cycles* **27** (2013), p. 1046-1058.
- [28] A. R. Fay, G. A. McKinley, "Global open-ocean biomes: Mean and temporal variability", *Earth Syst. Sci. Data* **6** (2014), no. 2, p. 273-284.
- [29] G. Gorsky, G. Bourdin, F. Lombard, M. L. Pedrotti, S. Audrain, N. Bin, E. Boss, C. Bowler, N. Cassar, L. Caudan, G. Chabot, N. R. Cohen, D. Cron, C. De Vargas, J. R. Dolan, E. Douville, A. Elineau, J. M. Flores, J. F. Ghiglione, N. Haëntjens, M. Herta, S. G. John, R. L. Kelly, I. Koren, Y. Lin, D. Marie, C. Moulin, Y. Moucherie, S. Pesant, M. Picheral, J. Poulain, M. Pujo-Pay, G. Reverdin, S. Romac, M. B. Sullivan, M. Trainic, M. Tresol, R. Troublé, A. Vardi, C. R. Voolstra, P. Wincker, S. Agostini, B. Banaigs, E. Boissin, D. Forcioli, P. Furla, P. E. Galand, E. Gilson, S. Reynaud, S. Sunagawa, O. P. Thomas, R. L. V. Thurber, D. Zoccola, S. Planes, D. Allemand, E. Karsenti, "Expanding Tara Oceans protocols for underway, ecosystemic sampling of the ocean-atmosphere interface during Tara Pacific expedition (2016-2018)", *Front. Mar. Sci.* **6** (2019), article no. 750.
- [30] J. J. Polovina, J. P. Dunne, P. A. Woodworth, E. A. Howell, "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.* **66** (2011), no. 6, p. 986-995.
- [31] L. Wrightson, A. Tagliabue, "Quantifying the impact of climate change on marine diazotrophy: Insights from earth system models", *Front. Mar. Sci.* **7** (2020), article no. 635.
- [32] C. Laufkötter, M. Vogt, N. Gruber, O. Aumont, L. Bopp, S. C. Doney, J. P. Dunne, J. Hauck, J. G. John, I. D. Lima, R. Seferian, C. Völker, "Projected decreases in future marine export production: The role of the carbon flux through the upper ocean ecosystem", *Biogeosciences* **13** (2016), no. 13, p. 4023-4047.
- [33] D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph", *Bioinformatics* **31** (2015), no. 10, p. 1674-1676.
- [34] B. Song, S. Chen, W. Chen, "Dinoflagellates, a unique lineage for retrogene research", *Front. Microbiol.* **9** (2018), article no. 1556.