Meta-omics and environmental meta-data : towards a new comprehension of the biological carbon pump

Anne-Sophie Benoiston

▶ **To cite this version:**

Anne-Sophie Benoiston. Meta-omics and environmental meta-data : towards a new comprehension of the biological carbon pump. Biodiversity and Ecology. Sorbonne Université, 2019. English. NNT : 2019SORUS182 . tel-03481925

HAL Id: tel-03481925
https://theses.hal.science/tel-03481925

Submitted on 15 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sorbonne Université

Ecole doctorale 515 Complexité du Vivant

Muséum National d'Histoire Naturelle

Institut de Systématique, Evolution, Biodiversité - UMR 7205

Méta-omique et méta-données environnementales : vers une nouvelle compréhension de la pompe à carbone biologique

Présentée pour obtenir le grade de

Docteur de Sorbonne Université par

ANNE-SOPHIE BENOISTON

Directrice : Lucie Bittner

Co-encadrant : Lionel Guidi

Devant un jury composé de :

Cécile Lepère	MCU, Université Clermont-Auvergne	Rapportrice
Ramiro Logares	CR, Institut de Ciències del Mar	Rapporteur
Emma Rochelle-Newall	DR, IRD	Examinatrice
Stéphane Blain	PR, Sorbonne Université	Examinateur
Christian Tamburini	DR, CNRS	Examinateur
Lucie Bittner	MCU, Sorbonne Université	Directrice

Soutenance prévue à Paris le 26 septembre 2019

Abstract

La pompe à carbone biologique comprend la production primaire de matière organique dans la zone euphotique, son export vers les profondeurs et sa reminéralisation. Les acteurs les plus fréquemment cités sont les diatomées en raison de leur contribution à la production primaire et à l'export de carbone et les copépodes pour la production de pelotes fécales. Cependant, la pompe biologique est le résultat d'interactions complexes entre organismes plutôt que de leurs actions indépendantes. En outre, bien qu'il ait été montré que la distribution de taille et la composition minérale du phytoplancton en surface ont une influence significative sur l'intensité de l'export de carbone, on ne sait pas si les données méta-omiques peuvent prédire efficacement les processus de la pompe à carbone biologique. Dans cette thèse, je propose d'abord de revisiter l'étude de la pompe à carbone biologique dans l'océan oligotrophe en définissant des états biogéochimiques de l'océan sur la base de la contribution relative de la production primaire, de l'export de carbone et de l'atténuation du flux dans les stations d'échantillonnage *Tara* Océans. L'analyse des états en termes de composition et d'interactions microbiennes inférées à partir de données de métabarcoding a révélé que les associations plutôt que la composition microbienne semblent caractériser les états de la pompe à carbone biologique. Ensuite, en utilisant les données méta-omiques et environnementales des expéditions *Tara* Oceans, je propose pour la première fois de prédire ces états biogéochimiques à partir d'abondances biologiques dérivées d'ADN environnemental, dans l'objectif de fournir une liste de biomarqueurs.

Mots-clés: pompe à carbone biologique, plancton, méta-omique, diatomées, réseaux d'associations microbiennes, machine learning

The biological carbon pump encompasses a series of processes including the primary production of organic matter in the surface ocean, its export to deeper waters and its remineralization. The common highlighted actors are diatoms because of their contribution to primary production and carbon export and copepods for their production of fecal pellets. However, the biological pump is the result of complex interactions among organisms rather than their independent actions. Besides, although size distribution and

mineral composition of phytoplankton in surface was shown to significantly influence the strength of carbon export, it is unknown whether meta-omic data can efficiently predict the processes of the biological carbon pump. In this thesis, I first propose to revisit the study of the biological carbon pump in the oligotrophic ocean by defining biogeochemical states of the ocean based on the relative contribution of primary production, carbon export and flux attenuation in *Tara* Oceans sampling stations. The analysis of the states in terms of microbial composition and interactions inferred from metabarcoding data revealed that variation in associations rather than lineages presence seems to drive the states of the biological carbon pump. Then, by using meta-omics and environmental parameters from the *Tara* Oceans expeditions, I propose the first study trying to predict biogeochemical states from biological abundances derived from environmental DNA, with the goal of providing a list of biomarkers.

Keywords: biological carbon pump, plankton, meta-omics, diatoms, microbial association networks, machine learning

Acknowledgements

Cette aventure a commencé il y a déjà presque quatre ans, lorsque j'ai passé un entretien pour un stage. J'ai alors rencontré Lucie Bittner et Lionel Guidi, qui m'ont permis de travailler avec eux pendant six mois, qui se sont transformés en trois ans et demi. Mes premiers remerciements leur reviennent donc. Merci de m'avoir fait confiance, plus que je n'ai confiance en moi. Merci pour votre enthousiasme et votre sympathie. Merci de m'avoir encouragée dans les moments de doute, tout en me laissant de grandes libertés. Je souhaite également remercier Stéphane Le Crom qui m'a accueillie dans son équipe et m'a aidée à me fixer des objectifs et organiser mon travail. Merci aussi pour sa participation à mes deux premiers comités de thèse en tant que directeur "officiel".

Tout cela n'aurait pu se concrétiser sans l'aide de Chris Bowler. Merci d'avoir financé mon stage de Master 2, de m'avoir permis de travailler avec toi sur une revue de littérature, d'avoir toujours répondu présent quand je t'ai sollicité et enfin de me permettre de rejoindre ton équipe en septembre. Je remercie également l'école doctorale Complexité du Vivant et sa directrice, Emmanuèle Mouchel-Vielh, pour m'avoir permis de poursuivre mon travail de recherche en thèse.

Je tiens à remercier les membres de mon jury : Cécile Lepère et Ramiro Logares pour avoir accepté d'évaluer mon travail, en espérant qu'il soit à la hauteur de vos exigences ; Stéphane Blain pour les discussions et précieux conseils apportés en tant que membre de mon comité de thèse, ainsi que Emma Rochelle-Newall et Christian Tamburini.

Ces trois dernières années ne se seraient pas aussi bien passées sans les doctorants et stagiaires de l'unité Evolution Paris Seine. Je remercie d'abord Arnaud et Quentin pour m'avoir chaleureusement accueillie dès mon stage de Master 2. Merci pour votre aide, vos conseils, pour m'avoir fait découvrir la meilleure boulangerie de Jussieu qui m'a nourrie un certain nombre de midis. Merci ensuite à Gabriel. Merci pour ton écoute, les intéressantes discussions sur les hamsters, les chinchillas, les crêpes à la pâte à crêpe, la paléontologie, les relations humaines. Merci pour les jours passés en Vaucluse sur les fouilles et notre week-end à Londres. En résumé, merci pour ton amitié. Merci ensuite à Thomas pour ton ouverture d'esprit, de m'avoir incitée à m'intéresser à la politique

■

pendant les dernières présidentielles, pour ta désorganisation chronique pendant ta thèse (sauf pour les "SCEP" endiablées !), pour tes réflexions sur la recherche et la thèse. Et merci aussi pour m'avoir fait rencontrer celui qui me supporte désormais tous les jours ! Merci à Emile pour tes explications et conseils avisés en statistique multidimensionnelle, pour ta relecture et pour ta décontraction naturelle en (presque) toutes circonstances ! Merci également à Juliette pour ta gentillesse et les soirées passées dans ton "château", et aux doctorants et post-doctorants membres de l'équipe AIRE : Jananan, Raphaël, Andrew et Romain. Merci enfin à tous ceux qui ont participé à l'animation du stand que j'ai organisé pour la fête de la science à Jussieu : Emile, Arnaud, Juliette et Gabriel. A ce propos, je me dois aussi de remercier Mehdi et Isabelle pour votre aide dans l'organisation et la mise en place de nos ateliers.

Je voudrais également remercier les personnes que j'ai rencontrées lors de la campagne océanographique MOOSE-GE. Merci Laurent et Pierre de m'avoir permis de participer à cette mission qui était une première pour moi. Merci au "quart de l'ambiance" : Armelle, Amélie, Emile, Marylou, Joana, Maxime, Alix, Ronan et Benoit pour votre bonne humeur et les razzias dans le frigo à 4h du matin ! Merci aussi à Fabrice, Magali et Loic pour votre patience et vos explications pour l'échantillonnage du plancton.

Merci aux Nantais du LS2N : Damien, Samuel et Erwan pour votre expertise, vos encouragements, votre optimisme et votre énergie. Vous m'avez poussée à aller vers l'inconnu, à prendre des risques, à imaginer. Sans vous je ne serais certainement pas allée si loin. Je remercie particulièrement Erwan pour ton aide précieuse et nos discussions éclairantes sur les méthodes d'inférence de réseau. Je souhaite aussi remercier Sakina Dorothée-Ayata, Eric Pelletier et Federico Ibarbalz qui m'ont apporté leur expertise tout au long de ma thèse.

Merci également à Eric, Karen et Muriel pour leur gentillesse, les pépites dans les copies d'étudiants, le tournage des vidéos de thèse et tous les repas partagés au RU de Jussieu. Je tiens aussi à remercier Danielle. Merci d'avoir souvent été réactive lorsqu'il s'agissait de mettre en place des ordres de mission ou réserver des voyages. Merci aussi à Nora qui m'a aidée à rendre possible la venue de mon jury de thèse.

Après être passée par l'unité Evolution de l'UPMC, j'ai pu rejoindre l'atelier de bioinformatique pour mes derniers mois de thèse. Un grand merci à toute l'équipe, ainsi qu'à Philippe Grandcolas pour votre accueil chaleureux. Je suis ravie d'avoir partagé ces derniers mois avec vous. Merci particulièrement à Joël et Sophie pour leur aide technique lors de notre installation, merci à Guillaume et Mathilde de nous avoir permis de rejoindre l'ABI, et merci à Elise et Martin pour l'organisation des soirées dégustation de bière !

Je clôture ces remerciements en adressant un immense merci à ceux qui partagent ma vie en dehors du labo et qui m'ont accompagnée de près ou de loin ces trois dernières années. Pour commencer, merci à ma famille qui m'a toujours soutenue et qui a cru en moi (bien plus que moi !) : mes parents, ma soeur, mes grands-parents. Merci ensuite à mes amies qui me suivent depuis très longtemps : Léna, Elodie et Coralie. Et les copains/copines que j'ai rencontrés plus tard : Elisabeth, Chloé, Charlotte, Solen, François, Guillaume, Marie, Christophe, Caroline, Ulysse, Dimitri, Victor, Jennifer, Xavier et Kenzo. J'adresse un grand merci à Clara : j'ai adoré travailler avec toi en projet, et merci pour ta relecture et tes suggestions. Merci aussi à JB, Nicolas, Sonia et Robert : les moments passés sur les fouilles étaient géniaux ! (et le seront encore évidemment). Enfin, merci à Nicolas pour m'avoir accompagnée ces trois dernières années, je ne suis pas sûre que je serai parvenue au bout sans toi.

This manuscript is based on the LaTeX template created by John Liaperdos (2015) and available on www.overleaf.com.

Contents

List of Figures

List of Tables

Abbreviations

DCM Deep chlorophyll maximum. 47, 109

DIC Dissolved inorganic carbon. 17, 20, 21, 38

DNA Deoxyribonucleic acid. 64, 66, 67, 78, 108, 118, 131

DOC Dissolved organic carbon. 21

DOM Dissolved organic matter. 21, 33, 34

ESD Equivalent sperical diameter. 30

mtry Number of tested predictors at each split by random forest. 8, 112, 114, 115, 118

ntree Number of trees in a random forest. 8, 112, 114, 115, 118

OOB Out-of-bag. 8, 9, 111–118

OTU Operational taxonomic unit. 8, 9, 11, 67–70, 74, 78, 80–91, 93, 95, 99, 108–112, 115–117, 119–122, 126–128

Pg C Petagrams of carbon (10^{15} grams of carbon). 15, 16, 21, 22, 35

POC Particulate organic carbon. 33, 35, 37, 130, 131

POM Particulate organic matter. 21, 33

PSD Particle size distribution. 31, 79

RF Random forest. 4, 8, 9, 107, 110–120, 122

RNA Ribonucleic acid. 66, 67, 109

rRNA Ribosomal ribonucleic acid. 67

TEP Transparent exopolymer particle. 28, 37, 41, 42

UVP Underwater Video Profiler. 7, 30, 35, 79

Chapter 1

General introduction

1.1 The carbon cycle on Earth

1.1.1 The reservoirs of carbon

The carbon cycle corresponds to the fluxes and exchanges of carbon between the differents spheres of the Earth: biosphere, lithosphere, hydrosphere and atmosphere (figure 1.1, Ciais et al., 2013). Most of the carbon is contained in the lithosphere, in inorganic form in limestone rock and in organic form in fossil fuels (~15,000,000 Pg C, Siegenthaler and Sarmiento, 1993; Sigman and Boyle, 2000; Ciais et al., 2013). The ocean contains around 38,000 Pg C, mainly in the deep ocean (Sigman and Boyle, 2000; Ciais et al., 2013). The atmosphere contains relatively few carbon compared to the other reservoirs (around 830 inorganic Pg C, Ciais et al., 2013). Biosphere is made of organic carbon. It is estimated that it contains 650 Pg C (Ciais et al., 2013; Houghton, 2014).

1.1.2 Exchanges of carbon between the reservoirs

On Earth, carbon is present in organic and inorganic form. Organic carbon is synthesized by living organisms and composes the alive and dead biomass. It is associated to other elements such as hydrogen, oxygen, nitrogen and phosphorus. On the contrary, inorganic carbon comes from minerals or results from organic carbon recycling.

Carbon is exchanged from a reservoir to another by physico-chemical and biological processes. The carbon cycle maintains an equilibrium between the reservoirs, allowing the Earth's temperature to remain stable. The exchanges of carbon occur on different time scales, we refer to the "slow" and "fast" carbon cycles. The slow carbon cycle implies geological processes that act on thousands to millions of years. This cycle starts by the dissolution of atmospheric CO_2 in the ocean, supplemented by the river supply of calcium

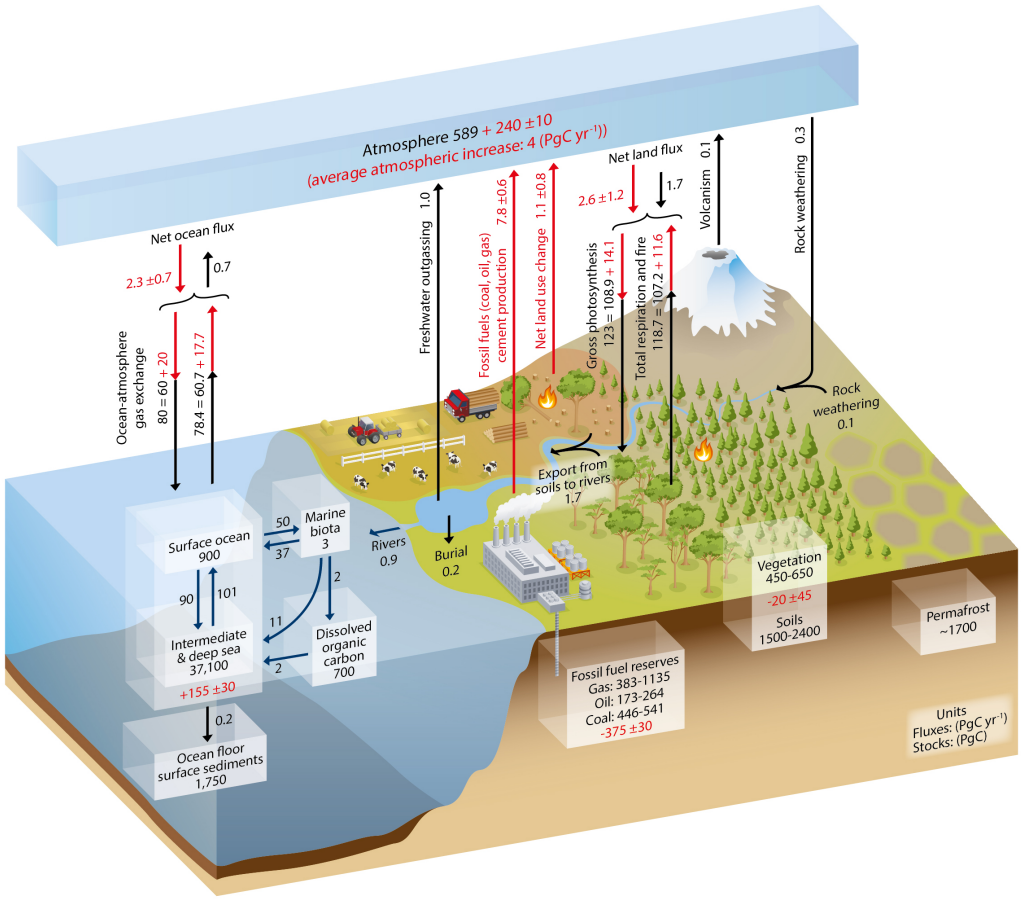


Figure 1.1 - Diagram of the Earth carbon cycle. Numbers represent reservoir mass, also called "carbon stocks" in Pg C and annual carbon exchange fluxes (in Pg C yr^{-1}). Black numbers and arrows indicate reservoir mass and exchange fluxes estimated for the time prior to the Industrial Era, about 1750. Red numbers in the reservoirs denote cumulative changes of anthropogenic carbon over the Industrial Period 1750-2011 (a positive cumulative change means that a reservoir has gained carbon since 1750). Red arrows and numbers indicate annual "anthropogenic" fluxes averaged over the 2000-2009 time period (Ciais et al., 2013).

whose presence in rivers results from the weathering of terrestrial rocks by acid rain. In the ocean, the calcium ions Ca^{2+} react with carbonate ions $CO_3{}^{2-}$ to form calcium carbonate $CaCO_3$ that constitutes the calcareous sketeleton of numerous marine organisms such as coccolithophores and foraminifera. After their death, these organisms sink towards the deep ocean where a part of them will be stored as limestone rocks. In some oxygen-lean environments, the accumulation of organic matter (mainly from plants) can form hydrocarbons such as coal, oil and natural gas. The carbon trapped in sedimentaty rocks is released in the atmosphere as CO_2 through volcanic emissions.

The fast carbon cycle is the result of exchanges between the atmosphere, the ocean and living organisms (time scales of days; 90 and 120 Pg C are exchanged each year between the atmosphere and the surface ocean and between the atmosphere and land, respectively, Houghton, 2014). The production of organic matter is at the basis of this

cycle. Phytoplankton and plants use the solar energy to fix CO_2 and produce organic matter and oxygen (details are given in section 1.3.1). This organic carbon then undergoes a series of transformations in the food chain, releasing again carbon dioxide in the ocean or the atmosphere.

1.2 The carbon pumps in the ocean

Dissolved CO_2 exists in several forms in the ocean, constituting dissolved inorganic carbon (DIC):

$$DIC = CO_2\,(\text{dissolved}) + H_2CO_3 + HCO_3^- + CO_3^{2-} \tag{1.1}$$

where H_2CO_3, HCO_3^- et CO_3^{2-} are carbonic acid, bicarbonate and carbonate ion, respectively. In the literature, the terms "total dissolved inorganic carbon" (C_T) and "total CO_2" (TCO_2 or $\sum CO_2$) can be found to designate the DIC (Legendre et al., 2015). Less than 1% of DIC is in the form of dissolved CO_2, most is in the form of bicarbonate and carbonate ions (Houghton, 2014).

The distribution of DIC in the ocean is characterized by a vertical concentration gradient: deep waters are richer in DIC than surface waters (on average 2,284 $\mu mol.kg^{-1}$ beyond 1,200 meters and 2,012 $\mu mol.kg^{-1}$ in surface, figure 1.2). The process responsible for this gradient is the carbon pump, defined by Volk and Hoffert in 1985 as "a process that depletes the surface ocean of CO_2 concentration relative to the deep-water CO_2 concentration". They distinguish three components: the **solubility pump**, the **carbonate pump** and the **soft-tissue pump** [1]. The carbon at the origin of these three pumps comes from the atmosphere. Atmospheric CO_2 dissolved in the surface waters reacts with water to form bicarbonate (HCO_3^-), carbonate ions (CO_3^{2-}) and protons (H^+):

$$CO_2 + H_2O \longleftrightarrow HCO_3^- + H^+ \longleftrightarrow CO_3^{2-} + 2\,H^+ \tag{1.2}$$

It is the disequilibrium induced by the difference between the partial pressure of CO_2 in the atmosphere and the ocean that leads to a flow of carbon from the atmosphere to the ocean. The higher is this difference, the more intense are the fluxes between the two environments. This disequilibrium varies according to regions (figure 1.3) because the

[1] The *soft-tissue pump* and the *carbonate pump* are the result of the action of the living organisms that use the inorganic carbon to produce organic matter or calcium carbonate, that is why they are sometimes referred to as the *biological carbon pump*. However, some authors use this expression to name only the organic component of the carbon pump. In this manuscript, we will use the term *biological carbon pump* to refer to the *soft-tissue pump* exclusively, consistent with the Glossary in IPCC (2013).

partial pressure of CO_2 depends on its concentration and solubility, and for gases like CO_2, solubility depends on temperature and pressure. Note that, as shown in equation 1.2, the dissolution of CO_2 adds H^+ ions in sea water, which increases the ocean acidity.

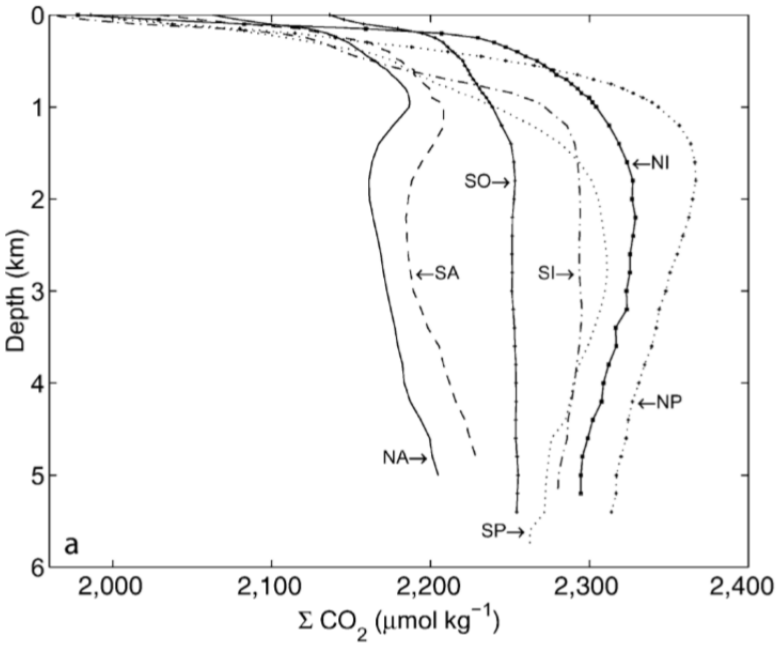


Figure 1.2 - Mean vertical distribution of dissolved inorganic carbon ($\sum CO_2$) in the oceans. NA/SA: North/South Atlantic, SO: Southern Ocean, NI/SI: North/South Indian Ocean, NP/SP: North/South Pacific Ocean (Zeebe and Wolf-Gladrow, 2009).

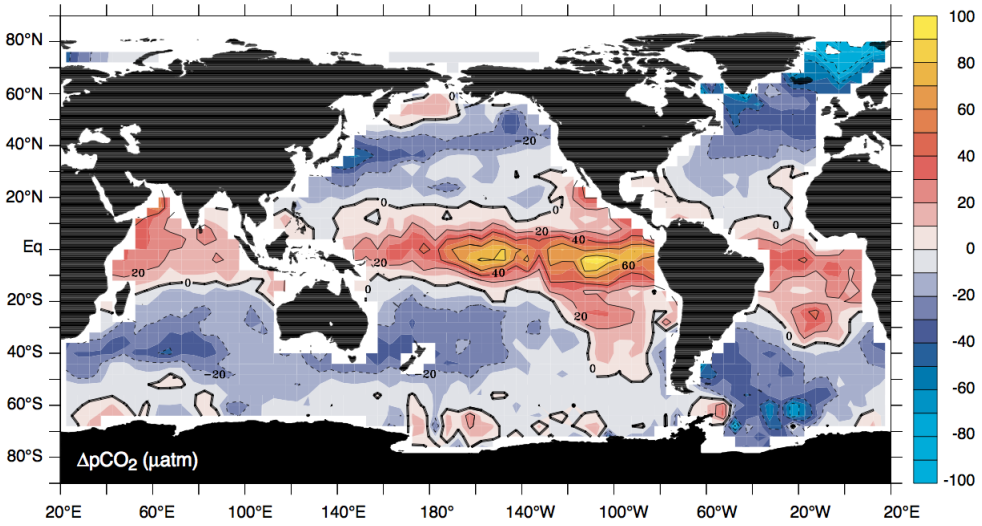


Figure 1.3 - Map of the climatological carbon dioxide partial pressure difference ΔpCO_2 across the air-sea interface. Supersaturation is indicated by positive numbers (colors from light red to yellow) and undersaturation by negative numbers (colors form dark to light blue) (Sarmiento and Gruber, 2006).

1.2.1 The solubility pump

The solubility pump is closely linked to the thermohaline circulation. It is driven by two processes: the solubility of CO_2 is negatively correlated to sea water temperature; and the high density of cold surface waters at high latitudes leads to the formation of deep water. As a result, CO_2-rich surface waters carry dissolved CO_2 in areas of deep convection, such as the North Atlantic Nordic Seas and the Weddell Sea in the Southern Ocean. The solubility pump can store carbon for centuries as deep water is exposed at the ocean surface roughly every 1,000 years (DeVries and Primeau, 2011).

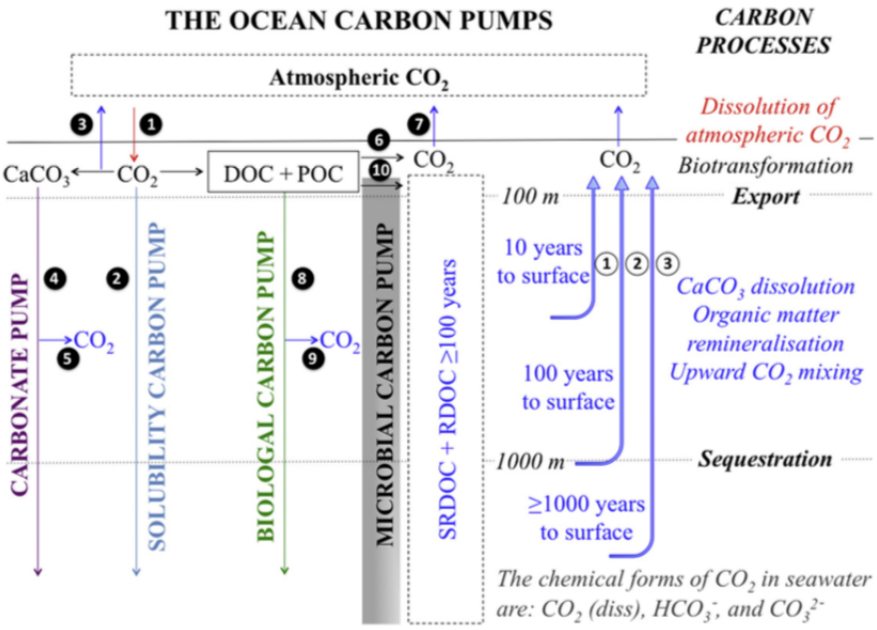


*Figure 1.4 - The four carbon pumps in the ocean: the **carbonate pump** (in purple on the left), i.e. precipitation of $CaCO_3$ accompanied by the release of CO_2 (3) and followed by the sink of bio-mineral particles towards the depth where carbon is sequestered (4); the **solubility pump** (in blue), i.e. dissolution of atmospheric CO_2 in surface waters (1), followed by the deep mixing of the CO_2-rich water and sequestration (2); the **biological carbon pump** (in green), i.e. use of dissolved CO_2 by phytoplankton for the production of organic matter, its transformation by the food web (6) and loss to the atmosphere (7), followed by the transfer of organic carbon in deep waters where it is sequestered (8); and the **microbial pump** (in gray) that produces refractory dissolved organic carbon (RDOC), thereby sequestering carbon (10). During its transfer to the deep, CO_2 is released in the water column by dissolution of part of the sinking $CaCO_3$ (5) and remineralization of part the sinking organic matter (9). The dissolved CO_2 is released in the atmosphere on different time scales depending on the depth (white numbers 1 to 3 on the right). (Legendre et al., 2015).*

1.2.2 The biological pump

The biological pump corresponds to a series of biologically mediated processes allowing to trap carbon on timescale it takes the ocean to bring deep water to the surface (i.e. between 500 and 1500 years depending on the ocean basin, DeVries and Primeau, 2011) or even on geological time scales (i.e. up to millions of years) in some cases. These processes

include the production of organic matter in the surface ocean, its export to the deep and its transformation by the pelagic food web.

In surface waters, phytoplankton take up DIC and nutrients to produce organic matter through photosynthesis (i.e. *primary production*). This newly produced organic matter faces grazing by zooplankton and respiration by zooplankton and heterotrophic bacteria. The organic matter that escaped *remineralization* is *exported* to deeper waters and potentially to the sediments (figure 1.5). These three processes and their control factors are detailed in section 1.3.

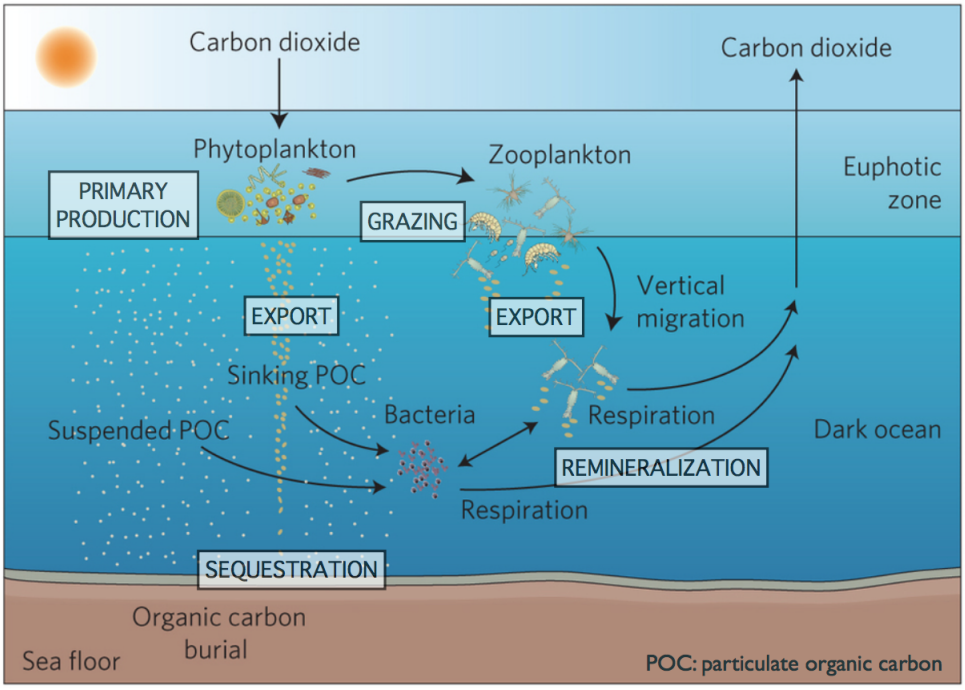


Figure 1.5 - Diagram of the biological carbon pump modified from Herndl and Reinthaler (2013). Phytoplankton fix carbon dioxide in the euphotic zone using solar energy (primary production). They are grazed on by herbivorous zooplankton, or consumed directly or indirectly by heterotrophic microbes feeding on solubilized remains of phytoplankton. A part of the primary production is exported out of the euphotic zone. The organic matter that escapes remineralization in the water column is sequestered.

1.2.3 The carbonate pump

Some planktonic organisms bear calcareous skeletal structures consisting mainly of $CaCO_3$. They include photosynthetic cells (e.g. coccolithophores), protozoans (e.g. foraminifera) and metazoans (e.g. pteropods). They contribute to the export of carbon in the form of $CaCO_3$, but the carbonate pump is also called carbonate counter-pump because carbonate precipitation decreases the alkalinity of sea water, which releases CO_2 to the surrounding waters and from there to the atmosphere (figure 1.4):

$$Ca_2{}^{+} + 2\,HCO_3{}^{-} \longrightarrow CaCO_3 + CO_2 + H_2O \tag{1.3}$$

The release of CO_2 in the atmosphere reduces DIC in surface waters. However, the dissolution of the sinking $CaCO_3$ at depth releases $HCO_3{}^{-}$, which increases the DIC concentration in deep waters. Together, these two effects contribute to the vertical concentration gradient of DIC. Besides, as the skeleton of calcifying organisms makes them denser, they sink relatively fast (Klaas and Archer, 2002), which increases their chances to be preserved an buried in sediments.

1.2.4 The microbial pump

Phytoplankton is thought to be the main source of dissolved organic matter (DOM) in the ocean, that contains about 660 Pg C in the form of DOM. Phytoplankton secrete DOM, but it is also released by zooplankton grazing, viral lysis and solubilization of particulate organic matter (POM) by heterotrophic bacteria (Arnosti, 2010; Buchan et al., 2014; Mühlenbruch et al., 2018), thereby providing substrate to heterotrophic bacteria. A small fraction of the DOM pool is not remineralized and is transformed into resistant DOM that accumulates in the ocean as biologically refractory DOM and creates most of the large reservoir of dissolved organic carbon (DOC) (more than 95% of the total DOC in the ocean is refractory DOC, Jiao et al., 2010). It is the successive processing of DOC by microorganisms that transforms reactive DOC in refractory DOC. Different fractions of DOC co-occur in the ocean. They are defined based on their lifetime (time required to decrease the concentration of the fraction to $1/e$ of its initial value where e is the Napierian constant $e \approx 2.71828$, so $1/e = 0.37$): labile (LDOC; average lifetime: hours to days), semi-labile (SLDOC; ca. 1.5 years), semi-refractory (SRDOC; ca. 20 years), refractory (RDOC; ca. 16,000 years) and ultra-refractory (URDOC; 40,000 years) (Hansell, 2013; Legendre et al., 2015). LDOC is readily available for microorganisms, while refractory fractions of DOC are resistant to microbial decomposition and can be stored in the ocean for up to thousand years. The suite of microbially mediated processes that leads to the creation of refractory DOC has been called the "microbial carbon pump" (Jiao et al., 2010). Whereas the solubility, carbonate and biological pumps rest upon the vertical transport of carbon from the surface to the deep ocean, the microbial pump is based on the production of RDOM at any depth of the water column.

1.3 The processes of the biological carbon pump

1.3.1 Primary production

The first step of the biological carbon pump is the primary production. It corresponds to the production of organic matter by terrestrial and aquatic autotrophic organisms. From an ecological point of view, primary production is the accumulation of solar energy by plants that is available for other trophic levels. This process depends on oxygenic photosynthesis that probably appeared more than 2.45 billion years ago (Buick, 2008). Photosynthetic organisms require CO_2, water (H_2O) and sunlight to produce organic molecules and oxygen (as shown by the following simplified equation of the oxygenic photosynthesis) as well as nutrients.

$$CO_2 + H_2O + \text{light} \longrightarrow CH_2O + O_2 \tag{1.4}$$

From a biogeochemical perspective, primary production is a carbon flux from the atmosphere to the biosphere. Gross primary production (GPP) is the total energy fixed by photosynthetic organisms via photosynthesis, while net primary production (NPP) is GPP minus phytoplankton respiration (R) necessary to the plant's metabolism (Roxburgh et al., 2005):

$$NPP = GPP - R \tag{1.5}$$

Net primary production is expressed in units of carbon per unit area (or volume) per unit time. Note that primary production can also be the result of other types of autotrophy such as chemoautotrophy in which the source of energy is mineral instead of light.

1.3.1.1 Global primary production estimates

Global oceanic NPP is between 45 and 50 Pg C per year and mostly performed by phytoplankton (Antoine et al., 1996; Field et al., 1998). It corresponds to 45% of total primary production on Earth, although phytoplankton biomass is ~1 Pg C, which is only 0.2% of the photosynthetically active carbon biomass of Earth (Field et al., 1998). This high productivity relative to phytoplankton total biomass is explained by higher turnover rates in the ocean than on land (Falkowski and Raven, 2007) and by a higher photosynthetic biomass proportion in phytoplankton than in terrestrial plants that are mainly composed of stems and roots, which respire and generally do not photosynthesize (Field et al., 1998).

1.3.1.2 Patterns in time and space

Globally, primary production is highest in coastal upwelling regions (figure 1.6) where mean primary production is around 2000 g C m^{-1} $year^{-1}$, while it is about 440 g C m^{-1} $year^{-1}$ in the open ocean. However, 80% of primary production occurs in the open ocean because of its much larger surface (Chavez et al., 2010). Primary production is subject to strong seasonal variations outside equatorial and tropical areas. Peaks of production (blooms) occur in spring at temperate latitudes, due to higher light and nutrients availability (Field et al., 1998; Uitz et al., 2010). Phytoplankton thrive and deplete nutrients in the euphotic zone, which inhibits its growth. In some areas, a small bloom can occur when light levels are still high enough and nutrients are injected from winter convection and overturning (Koeve, 2001).

1.3.1.3 Contribution of different types of phytoplankton

Diatoms contribute to a significant proportion of the ocean's primary production (Nelson et al., 1995). Using time series data of surface chlorophyll from satellite observations with SeaWiFS (Sea-viewing Wide Field-of-view Sensor), primary production of microphytoplankton was estimated to be 70% in coastal upwelling systems and 50% in temperate and subpolar regions during the spring-summer season (Uitz et al., 2010). The rest of primary production is due to smaller phytoplankton. Uitz et al. (2010) estimated the contribution of nanophytoplankton (primarily including prymnesiophytes) and picophytoplankton (mainly cyanobacteria) to be 44% and 24%, respectively. Using quantitative niche models, Flombaum et al. (2013) found consistent estimates: they predicted the abundant cyanobacteria *Prochlorococcus* and *Synechococcus* to be responsible for 8.5% and 16.7% of ocean net primary production, respectively.

The spatial distribution of these different types of phytoplankton reflects variations in physical properties, illumination, nutrient availability and grazers (De La Rocha and Passow, 2014). Diatoms require higher nutrient concentrations than coccolithophores (that are part of nanophytoplankton). Thus, they thrive preferentially in eutrophic coastal and subpolar areas (figure 1.6) while coccolithophores prefer more stratified conditions and are more abundant in the oligotrophic open ocean (Quéré et al., 2005). Thanks to their higher surface to volume ratio, picophytoplankton have an advantage over larger cells in nutrient-limited conditions (Raven and Falkowski, 1999), which explains their abundance in oligotrophic subtropical gyres where they play a substantial role in primary production (Uitz et al., 2010).

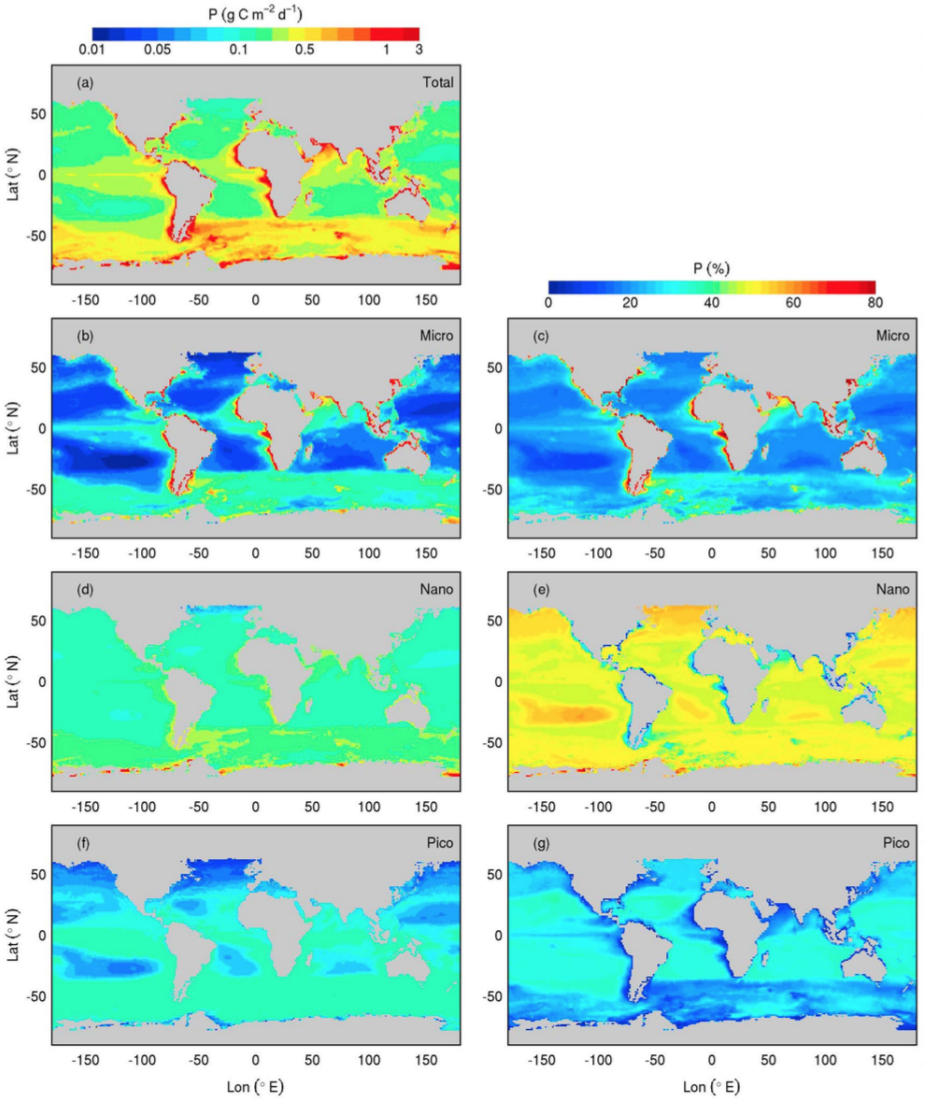


Figure 1.6 - Total and class-specific (micro-, nano- and picophytoplankton) primary production in the ocean for the December-February period (1998-2007). Daily primary production is given in absolute units (in grams of carbon per square meter per day) on the left panel and contribution (in percentage) of the phytoplankton classes on the right panel (Uitz et al., 2010).

1.3.1.4 Methods of estimating primary production

Historically, ocean primary production has been estimated with 14 C-based field measurements (Steemann Nielsen, 1952). Although the method helped oceanographers increasing temporal and spatial variability of primary production, estimations were made on discrete time points and do not cover the ocean's surface (Chavez et al., 2010). Optical methods have been developed to overcome this issue. These models include sea surface chlorophyll and irradiance because of their correlation to primary productivity (Smith et al., 1982; Eppley et al., 1985; Platt, 1986; Falkowski, 1981), such as the VGPM (Vertically Generalized Production Model) (Behrenfeld and Falkowski, 1997). It estimates net primary

production from chlorophyll concentration, chlorophyll efficiency to fix carbon and the amount of light received by the euphotic zone. The calculation of net primary production can be summarized by this relationship:

$$\mathrm{NPP} = \mathrm{chl} \cdot \mathrm{pb_{opt}} \cdot \mathrm{day\ length} \cdot \mathrm{f(par)} \cdot \mathrm{z_{eu}} \tag{1.6}$$

where *NPP* is the net primary production production (expressed in mg of carbon fixed per day per unit volume), *chl* is the chlorophyll concentration, pb_{opt} is the maximum daily net primary production in a given water column (expressed in mg of carbon fixed per mg of chlorophyll per hour), *day length* is the number of hours of day light, *f(par)* is the ratio of realized water column integrated NPP to the maximum potential NPP if photosynthetic rates were maintained at maximum levels throughout the water column (f(par) = 0.66125 * par / (par + 4.1) where *par* is the photosynthetically active radiation), and z_{eu} is the depth of the euphotic zone.

Remote estimations of primary productivity have been possible thanks to the use of radiometers aboard satellites since the end of the 1970's. The Moderate Resolution Imaging Spectroradiometer (MODIS) is currently flying on NASA's Terra and Aqua satellites. It captures a wide range of wavelengths, allowing to provide measurements of the ocean's color that translates the concentration of organisms in the surface waters, including chlorophyll-containing organisms which are of interest for the estimation of primary production. Figure 1.6 gives an example of maps that can be achieved from satellite data.

1.3.2 Carbon export

Carbon export corresponds to the transport of photosynthetically-produced organic matter to the deep. This material is exported as particles as a result of gravity or by active transport by zooplankton, or as dissolved organic carbon through vertical mixing or advection. The export flux (or export production) is defined as the quantity of carbon that leaves the euphotic zone or the mixed layer depth (Passow and Carlson, 2012). Although primary production sets the upper limit of export flux, most of it is recycled within the euphotic zone and thus escapes export (Buesseler, 1998).

1.3.2.1 Spatial and temporal variability of carbon export

Between 5 and 25% of primary production is exported out of the euphotic zone (De La Rocha and Passow, 2007). However, export intensity is highly variable depending on the regions. Generally, primary production and export flux are tightly linked (Buesseler, 1998). This

is particularly the case in coastal upwelling systems that support high primary productivity (figures 1.6 and 1.7). However, carbon export is not always proportional to the local primary production levels (Buesseler, 1998; Buesseler and Boyd, 2009). Seasonality also greatly influences carbon export efficiency. Following blooms of large cells such as diatoms, episodic export pulses of 50% or higher of primary production have been observed (Buesseler, 1998).

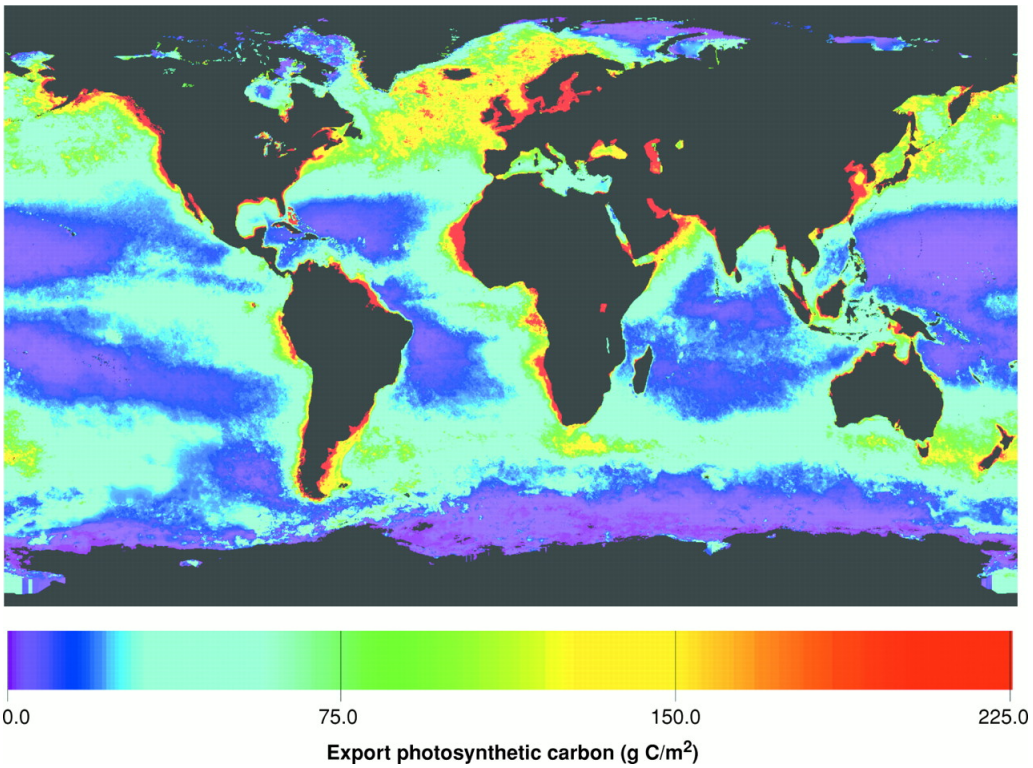


Figure 1.7 - Global map of annual mean export production for the world oceans estimated from the relation between total and export production of Eppley and Peterson (1979) and monthly mean total production maps produced from Coastal Zone Color Scanner (satellite radiometer) chlorophyll distributions according to the algorithm of Behrenfeld and Falkowski (1997) (Falkowski et al., 1998).

1.3.2.2 Composition of sinking particles

Marine snow

Our knowledge about the composition about sinking particles is mostly based upon the analysis of the content of sediment traps (see next section for details about these instruments). Organic material preferentially sinks as large particles (known as "marine snow") rather than individual cells. Marine snow is composed of aggregated phytodetritus, appendicularian mucus feeding structures, fecal matter, prokaryotic cells and miscellaneous detrital particles (Alldredge and Silver, 1988; Simon et al., 2002). They can be responsible for a large part of the vertical flux. For example, Shanks (2002) estimated the contribution of marine snow to be >90% of particle flux. The higher sinking velocities

of marine snow compared to unaggregated material enhances export flux, despite their elevated rates of decomposition.

Zooplankton

Although highly variable, contribution of zooplankton fecal pellets to carbon flux has been shown to be substantial. The proportion of vertical flux due to fecal pellets ranges from less than 1% to more than 90% and is dependant on phytoplankton and zooplankton biomass and composition (e.g. Dubischar and Bathmann, 2002). Producers include metazoans such as copepods, euphausiids, salps, appendicularians and tunicates, as well as protistan microzooplankton (Turner, 2015). Zooplankton fecal pellets have long been considered as major vectors of carbon to the deep sea, but their importance appear to have been overestimated. Indeed, degradation by bacteria and coprophagy (ingestion), coprorhexy (fragmentation) and coprochaly (dispersal of pellet content into the water after disruption of their peritrophic membranes) by copepods and protozoans participate to the destruction of fecal pellets (Iversen and Poulsen, 2007), thus transforming large fast-sinking particles into smaller suspended ones and retaining fecal material in the euphotic zone (e.g. Poulsen and Kiørboe, 2006; Iversen and Poulsen, 2007). Sinking velocities of fecal pellets amount in hundreds of meters per day (Turner, 2002). The sinking velocities of fecal pellets and aggregates can be increased according to their phytoplankton content, which may contain biominerals such as opal and coccoliths (Armstrong et al., 2002; Klaas and Archer, 2002). Zooplankton also actively contribute to the export flux to depth through their diurnal vertical migrations[2] (see references in Turner, 2002 and Turner, 2015). Zooplankton feed in surface at night and respire and excrete dissolved and particulate material at depth during the day. It can represent between 10 and 50% of total vertical flux (Bollens et al., 2011) and can even exceed the carbon flux due to fecal pellets and other particles such as in the subarctic North Pacific during winter (Kobari et al., 2013). In some environments, mucus feeding structures produced by appendicularians and pteropods appear to significantly contribute to particle flux (Alldredge, 2005). These sticky structures, used to collect food, are discarded by their owners once they become clogged with material and continue to collect particles while sinking in the water column, which increases their sinking velocities.

[2]Zooplankton diurnal vertical migration are also referred to as the "mesopelagic-migrant pump" that is part of additional export pathways that inject particles to depth termed "particle-injection pumps" (PIPs) and that are not reviewed here. These multi-faceted pumps are physically and/or biologically mediated and are three dimensional mechanisms that require specific investigation approaches (Boyd et al., 2019). Taking them into account may partly explain the carbon budget deficits reported in the mesopelagic zone (Dall'Olmo et al., 2016).

Phytodetritus

Pulsed export of phytodetritus is an important component of export fluxes. They are mostly seasonal as they occur following spring phytoplankton blooms in temperate waters, or austral summer blooms in the Southern Ocean. They are commonly composed of diatoms, haptophytes such as coccolithophorids and *Phaeocystis* spp., and dinoflagellates (Beaulieu, 2003). Intact phytoplankton cells have even been observed in the deep dark ocean (Agusti et al., 2015). Sedimentation of phytodetritus is enhanced by aggregation that can result from biological and physical processes. Indeed, aggregation requires the collision of particles and their subsequent attachment. The probability of collision of particles depends on their concentration, density, size and shape, but also on shear and differential settling (Simon et al., 2002), while the probability of attachment is enhanced by the presence of transparent exopolymer particles (TEP), produced by phytoplankton and bacteria or derived from dissolved precursors released by phytoplankton. TEP are sticky gels that were found to form the matrices of marine aggregates, thus promoting sedimentation of particles (Passow, 2002). Bacteria are also known to produce TEP and may stimulate TEP production by phytoplankton, which contributes to increase aggregation (Passow et al., 2001).

1.3.2.3 Quantifying carbon export

Different methods have been used to estimate export flux in the water column. The oldest method is based on the estimation of "new" production in the euphotic zone through the measurement of nutrient uptakes using ^{15}N-labeled compounds (Dugdale and Goering, 1967). Contrary to regenerated production which is supported by recycled nutrients in the euphotic zone, new production is supported by nitrogen brought to the euphotic zone through upwelling, river input, atmospheric deposition or N_2-fixation. In the context of this method, nitrogen is considered to limit phytoplankton growth and the quantity of exported material is assumed to be equal to new production in the euphotic zone.

More direct measurements of particle flux include particle-reactive nuclides and sediment traps. Particle-reactive nuclides are used as tracers of particle flux. This method is based on the properties and half-lives of radioisotope pairs in the uranium decay series. The ^{238}U-^{234}Th pair has been extensively used to determine export fluxes (e.g. Buesseler et al., 1992; Coale and Bruland, 1985). The parent nuclide ^{238}U is soluble while the daughter nuclide ^{234}Th is particle-reactive and is thus rapidly scavenged by particulate material. If none of the isotopes were physically removed, their activities would be in secular equilibrium (i.e. identical). However, the scavenging of ^{324}Th by particles followed

by their vertical sinking results in a lower concentration of ^{234}Th in surface waters. Fluxes of ^{324}Th can thus be inferred from its vertical distribution.

Sediment traps have been extensively deployed during the last 40 years to characterize the nature and magnitude of sinking particles (Honjo et al., 2008). These funnel-shaped instruments collect particles and trap them in small containers (figure 1.8) that are often poisoned to avoid microbial decomposition. They can be ballasted and thus immobilized at one location, or floating to drift with the currents. Time series can be obtained by installing a rotating mechanism allowing to change the container collecting material at programmed time intervals. There exist several issues that question the ability of sediment traps to represent the actual flux of particles, such as advection that may bring particles that do not come from above, swimmers that may also be retained in traps, remineralization of trapped material and seasonal variations in export flux (Buesseler, 1991; Siegel and Deuser, 1997). However, buoyant sediment traps, poisoned containers, and rotating mechanisms all contribute to limit these known biases.

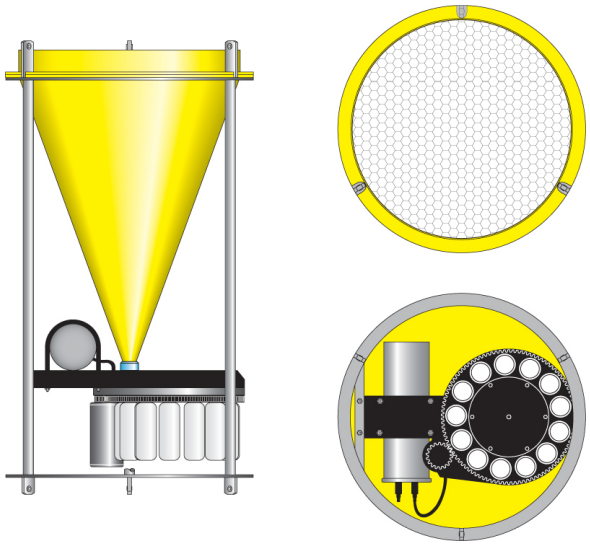


Figure 1.8 - Diagram of an automated time-series sediment trap, consisting of a broad funnel with collecting containers at the bottom. A baffle (at top right) keeps out large objects that would clog the funnel. The circular tray (at bottom right) holds collection vials. On a preprogrammed schedule (every 5 days to 1 month) the instrument seals one vial and rotates the next one into place. Scientist retrieve the samples up to a year later to analyze the collected sediment (illustration by Jane Doucette, WHOI, https://divediscover.whoi.edu).

From observations obtained with free-floating sediment traps deployed in the framework of VERTEX (Vertical Transport and Exchanges) in the North Pacific, one of the earliest international programs that explored the particulate organic matter fluxes, Martin et al. (1987) fitted their observations to the following power law function:

$$F_{(z)} = F_{100}\left(\frac{z}{100}\right)^{-b} \tag{1.7}$$

where z is the depth, b is a dimensionless factor (estimating remineralization or export efficiency), F_{100} is the particulate organic carbon flux at 100 m. The mean value of b estimated from the flux values was 0.86 Martin et al. (1987). This value has long been assumed to be uniform in space an time, although it is instead a spatially variable value (Henson et al., 2012; Lutz et al., 2007; Guidi et al., 2015).

More recently, optical approaches were developed thanks to methodological advances: satellite-based models (Siegel et al., 2014) and *in-situ* video systems such as the Underwater Video Profiler (Picheral et al., 2010) that is described below.

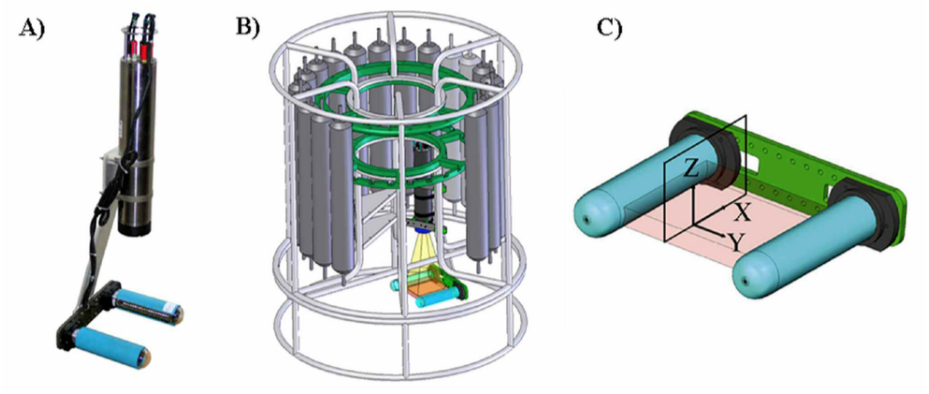


Figure 1.9 - Illustration of the UVP (version 5) (A) alone and (B) mounted on a Niskin bottle rosette frame. (C) Schematic diagram of the UVP light system where the illuminated volume of water is colored in pink (Picheral et al., 2010).

The Underwater Video Profiler

The Underwater Video Profiler (UVP) is an optical instrument that records images of ocean particles (figure 1.9). This instrument includes a camera that records the images in a volume of water illuminated by diodes emitting a collimated red light.

The UVP converts the measured area of particles in equivalent spherical diameter ESD to estimate particle size (Jennings et al., 1988). The area in pixels (S_p) of the objects captured by the UVP is converted in area in mm^2 (S_m) with the following relationship:

$$S_m = AS_p^B \tag{1.8}$$

where A and B are constants. These constants are estimated so that log-transformed differences ΔS between S_p and S_m are minimized:

$$\Delta S = \sum_{i=1}^{n} [log(S_{p,i} - log(S_{m,i})]^2 \tag{1.9}$$

Estimation of carbon flux from the particles' size distribution

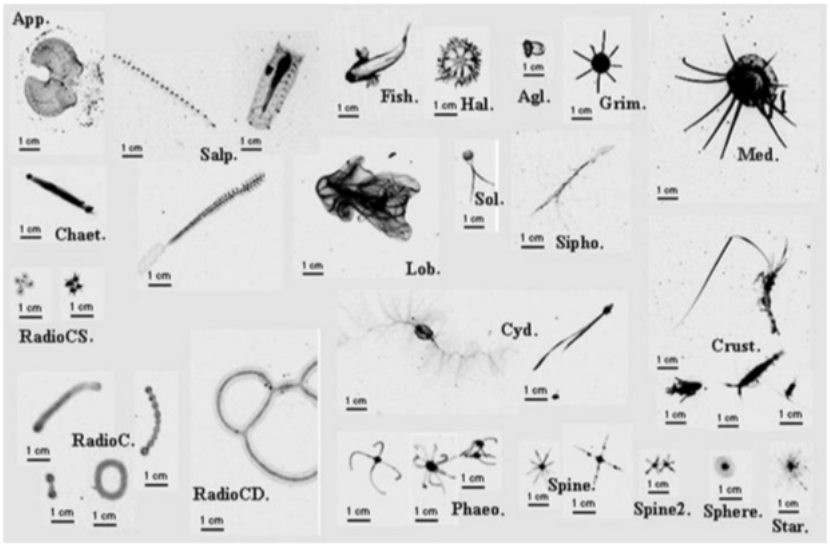


Figure 1.10 - UVP images of zooplankton, such as appendicularians (App.), salps (Salp.) and custraceans (Crust.) (Stemmann et al., 2008)

The particle size distribution (PSD) $n(s)$ is calculated in terms of particle concentration (ΔC, number by volume unit) in a given size range:

$$n(s) = \Delta C / \Delta s \tag{1.10}$$

The PSD follows a decreasing power law from the micrometer to the millimeter scale (McCave, 1984; Sheldon et al., 1972):

$$n(d) = a d^{-b} \tag{1.11}$$

where a and b are constants and d is the particles diameter (Sheldon et al., 1972; McCave, 1984; Jackson et al., 1997). The exponent b corresponds to the slope of particle size distribution, which is often estimated from the equation $ln(n(d)) = ln(a) - b * ln(d)$, where ln is natural logarithm. This slope b is often used as a descriptor of PSD (Guidi et al., 2009).

The particles mass $m(d)$ and their settling speed $w(d)$ (calculated with the Stokes law) are functions of d. Thus the total particles mass in the Δd interval is $n(d)m(d)\ \Delta d$ and the mass flux in this interval is $n(d)m(d)w(d)\ \Delta d$ (Guidi et al., 2008). The total carbon flux of particles F is the the mass flux spectrum integrated over all particle sizes (Guidi et al., 2008):

$$F(d) = \int_{d_{min}}^{d_{max}} n(d)m(d)w(d)\ \Delta d \tag{1.12}$$

where d_{min} and d_{max} are the minimum and maximum particles size.

1.3.3 Flux attenuation

Flux attenuation refers to the decreasing of organic matter vertical flux throughout the water column. Less than 50% of net primary production is exported below the euphotic zone while a few percent (less than 3%, De La Rocha and Passow, 2007) is sequestered (i.e. reaches deep waters, below 1000 m) (Andersson et al., 2004; Boyd and Trull, 2007; Buesseler and Boyd, 2009; Lutz et al., 2002; Martin et al., 1987). Carbon flux decreases exponentially with depth[3], with the most important decrease occurring in the euphotic zone (figure 1.11), implying that degradation processes are particularly active in the upper few hundred meters (De La Rocha and Passow, 2007).

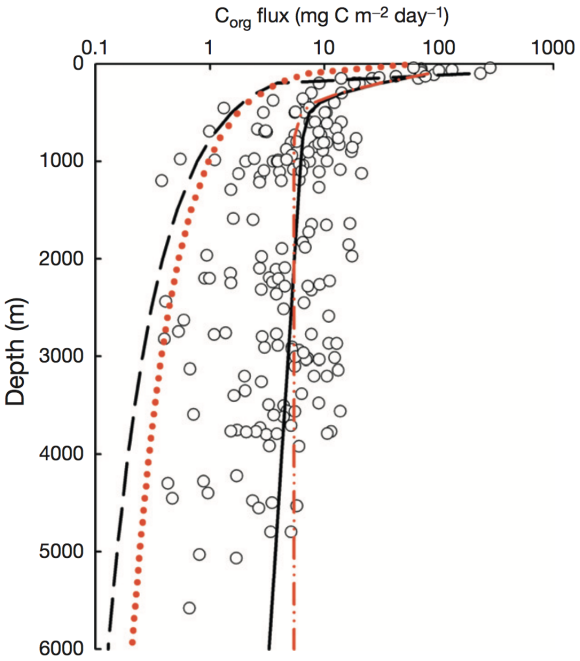


Figure 1.11 - Carbon fluxes as a function of depth, estimated from sediment traps and regionalized estimates of export production. The figure is from De La Rocha and Passow (2014), who used data from Lutz et al., 2002 (black points) and equations from Suess, 1980 (black dashed line), Martin et al., 1987 (red dotted line), Lutz et al., 2002 (red dashed and dotted line) and Andersson et al., 2004 (black solid line).

[3]Flux attenuation is often estimated from particle flux profiles measured using surface and deeper sediment traps and assumes that export from the surface does not significantly change in the time it takes material to reach the deepest trap. However, the slow sinking rates of the sinking material may lead to misleading flux profiles interpretations in regions with a strong seasonality in export, such as the temperate and polar regions (Giering et al., 2017).

1.3.3.1 Degradation of particulate organic matter

Our understanding of the processes determining the degradation of organic matter is much less developed than our knowledge about primary production (Sarmiento and Gruber, 2006). However, two planktonic groups are recognized to actively participate to the degradation of particulate organic matter: heterotrophic bacteria and zooplankton. Bacteria solubilize particulate organic carbon (POC) to dissolved organic carbon and oxidize it to CO_2, thereby removing most of POC in the upper layers of the ocean, whereas zooplankton contribute to flux attenuation by consuming organic matter, including fecal pellets (see section 1.3.2.2), and fragmenting particles while swimming.

Bacterial transformation of organic matter

Until the 1970's, importance of bacteria (and archaea) in the ocean carbon cycle was largely ignored (Steele, 1974). Consumption of organic matter by bacteria was considered as negligible. However, the discovery of the domination of bacteria en terms of abundance, diversity and metabolic activity changed our vision of a trophic chain implying the transfer of most of the primary production towards zooplankton and bigger animals (Azam, 1998; Azam and Malfatti, 2007). Until half of the ocean primary production is processed by bacteria through the "microbial loop" (Azam et al., 1983), making the flux of organic matter towards bacteria a major biogeochemical process.

Remineralization of organic matter by bacteria is initiated by extracellular enzymes. They hydrolyze POM to release lower molecular weight DOM (Arnosti, 2010), which provides substrates for bacterial growth in the surrounding water (bacteria are obligate osmotrophs). A fraction of this matter is incorporated as biomass while the other is respired (Arnosti, 2010). Because they are rich in resources, aggregates are considered "hotspots" of microbial remineralization in the ocean (Azam, 1998). After their formation (e.g. during bloom collapse), they are readily colonized by heterotrophic bacteria (Smith et al., 1992). Bacterial concentration is higher in aggregates that in free fractions (Caron et al., 1986; Turley and Mackie, 1994), as is their hydrolytic enzymatic activity (Simon et al., 2002; Grossart et al., 2007; Ziervogel and Arnosti, 2008). The phytoplanktonic composition also influences the molecular composition of organic matter (fatty acids, sugars, proteins, nucleic acids) and thus the growth and metabolic activity of bacteria (Buchan et al., 2014). Bacteria-mediated dissolution of silica from diatom frustules has also been reported, which contributes to upper ocean silicon regeneration (Bidle and Azam, 1999). Bacterial groups are known to be associated with aggregates, such as Flavobacteria, Gammaproteobacteria, Rhodobacteraceae and Altermonadaceae (Grossart and Ploug, 2001; Buchan et al., 2014).

Impact of zooplankton on flux attenuation

Zooplankton contribute to decreasing vertical particle flux by breaking up aggregates into smaller particles. The shear stress induced by euphausiids swimming disaggregates marine snow into smaller particles that have lower sinking velocities than the original aggregates (Dilling and Alldredge, 2000; Goldthwait et al., 2004, 2005). Besides, disaggregation may release DOM that is carried along the interstices of sinking marine snow and would have sunk with particles otherwise (Alldredge, 2000; Goldthwait et al., 2005). The slower sinking rates resulting from fragmentation also increase their residence time in surface waters and thus the likelihood of aggregate remineralization (Goldthwait et al., 2005).

The relative contribution of bacteria and zooplankton to the respiration of net primary production is currently poorly known (De La Rocha and Passow, 2007; Steinberg et al., 2008). Estimations attribute to microzooplankton the respiration of 35 to 59% of primary production (Calbet and Landry, 2004) and to mesozooplankton 17 to 32% (Hernández-León and Ikeda, 2005), while other authors consider that the bulk of respiration in the ocean is due to bacteria (Cho and Azam, 1988; Rivkin and Legendre, 2001). Yet others suggest that particle flux attenuation in the upper mesopelagic is driven by zooplankton fragmentation and solubilization rather than by microbial respiration (Belcher et al., 2016). However, higher abundance of zooplankton in highly productive areas may increase flux attenuation compared to lower productive areas that are rather dominated by small cells and where flux attenuation is mostly caused by bacterial activity (Guidi et al., 2015). In addition to heterotrophic bacteria and zooplankton, viral cell lysis may be responsible for a significant part of flux attenuation (Proctor and Fuhrman, 1990; Fuhrman, 1999; Rohwer et al., 2009; Lara et al., 2017) although their role is currently poorly known due to a paucity of data (Lara et al., 2017).

1.3.3.2 Measurement of vertical flux attenuation

Vertical flux attenuation can be inferred from particle flux profiles that are estimated from sediment traps and the thorium isotope technique (see section 1.3.2.3). Other methods exist to measure oxygen consumption and thus remineralization rates.

The distribution of oxygen in the water can inform us about primary production and remineralization. In the surface waters, oxygen concentrations are close to saturation, whereas below the euphotic zone, oxygen is mainly supplied by advection and mixing, and removal is due to remineralization. Taking into account these physical processes and oxygen distribution, oxygen utilization rates and particulate organic material oxidation

below the euphotic zone can be estimated. Basically, it means estimating the change in O_2 since the last time the water mass was in contact with the atmosphere (which is equivalent to the age of the water mass). Thus, oxygen utilization rates (OUR) can be obtained by calculating the ratio between the apparent oxygen utilization (AOU, i.e. the difference between the saturation oxygen concentration and the observed oxygen concentration) and the age of the water mass (Jenkins, 1982). Estimates of oxygen consumption can also be obtained from measurements (by Winkler titration) of bacterial respiration in dark seawater incubations (Obernosterer et al., 2008).

In this thesis, flux attenuation was estimated with flux values calculated from the particle size distribution (itself being derived from the UVP images, see section 3.2.3.2).

1.3.3.3 Global rates and regional and vertical variations of remineralization

Global annual estimates of marine respiration (including that of phytoplankton) are between 55 and 76 Pg C $year^{-1}$ (del Giorgio and Duarte, 2002). Remineralization is known to vary spatially in the oceans (Buesseler et al., 2007; Guidi et al., 2015), although remineralization rates has been thought to be uniform (Martin et al., 1987).

Using buoyant sediment traps, Buesseler et al. (2007) observed a high variability at two contrasting sites in the Pacific Ocean. Transfer efficiency of sinking particulate organic carbon between 150 and 500 m where of 20 and 50%. Going further, Guidi et al. (2015) estimated remineralization in different biogeochemical provinces from data obtained with the UVP, sediment traps and $^{238}U/^{234}Th$ disequilibrium. They found estimates that range between -50% and +100% of the commonly used globally uniform remineralization value.

1.3.4 Sequestration

The sequestration flux has to be distinguished from the export flux. While export is the flux that leaves the euphotic zone, sequestration refers to the flux that is removed from the atmosphere for 100 years or more (Passow and Carlson, 2012). The sequestration depth is often defined at the bottom of the mesopelagic zone (i.e. approximately 1000 m depth) but, as it depends of the ventilation depth, Antia et al. (2001) proposed to use the depth of winter mixing.

Annual POC sequestration is estimated at 0.86 Pg C yr^{-1} (Passow and Carlson, 2012). Of this amount, only about 0.002-0.16 Pg C are preserved each year in the sediments, which represents only 0.01-0.3% of the net primary productivity (Hedges and Keil, 1995). More recently, Guidi et al. (2015) found a lower estimation of 0.33 Pg C yr^{-1} (which

may be explained by the lowest depth considered for sequestration, i.e. 2000 m) and mapped regional variations of sequestration flux. In fact, like primary production, export and remineralization, and because it is dependant on these three processes (Passow and Carlson, 2012), sequestration of organic carbon varies geographically (Hedges and Keil, 1995; Dunne et al., 2007). Notably, Dunne et al. (2007) suggested that continental shelves may account for 48% of the global flux of organic carbon to the seafloor, although great variations exist between estimations (see Hedges and Keil, 1995).

1.3.5 Summary of currently known factors influencing the efficiency the biological carbon pump

As listed by Lemaitre (2017), control factors of the biological carbon pump are divided in chemical, biological and physical parameters.

1.3.5.1 Chemical factors

Chemical factors include the macro- and micronutrients availability. The difference between these two types of nutrients lays in their concentration in seawater and organisms, macronutrients being more concentrated than micronutrients (about 10^3 higher, Sarmiento and Gruber, 2006). Phytoplankton require these elements to form organic carbon, however some of them are limiting in the ocean. Nitrogen, phosphorus and iron are thought to be the main limiting nutrients. Much of the low-latitude oceans are depleted in surface nitrogen and phosphorus (Moore et al., 2013). These areas are referred to as low-nutrient low-chlorophyll (LNLC). On the contrary, regions such as the Southern Ocean have relatively higher concentrations of nitrogen and phosphorus. However, productivity is not as high as expected given these higher concentrations. These so-called high-nutrients low-chlorophyll (HNLC) regions are mostly limited by iron (which is an important component of electron transport proteins involved in photosynthesis and respiration), as shown by iron fertilization experiments in the Southern Ocean (Boyd et al., 2000; Blain et al., 2007). Besides, silicic acid may be limiting for silicifying planktonic organisms like diatoms, silicoflagellates and radiolarians, which use it in to build their shell (Moore et al., 2013).

In addition to primary production, chemical elements impact the settling speed of particles. The biogenic and lithogenic mineral content of sinking particles has been shown to increase their sinking velocity. For example, diatoms are considered effective organic carbon exporters because their frustule acts as a ballast, and the calcium carbonate shell of coccolithophorids also provides ballast for sinking of organic matter (Klaas and Archer,

2002; Armstrong et al., 2002; Francois et al., 2002).

1.3.5.2 Biological controls

Much of the variation in the efficiency[4] is explained by biological processes (i.e. the activities of phytoplankton, zooplankton and bacteria) (De La Rocha and Passow, 2007).

As suggested in sections 1.3.1.3 and 1.3.2.2, the composition of phytoplankton communities controls the intensity of primary production and particulate organic carbon flux. More precisely, the size structure of these communities has a key role in regulating the efficiency of the biological carbon pump (Boyd and Newton, 1999; Guidi et al., 2009). High POC export has been related to the dominance of microphytoplankton (mostly diatoms) relative to nano- and picophytoplankton (Guidi et al., 2009), which can also be explained by their mineral content (see the previous section). However, nanoplanktonic diatoms that are generally overlooked play a role in spring blooms and carbon export (Leblanc et al., 2018).

Zooplankton can have positive and negative impact on particle flux. By feeding upon phytoplankton cells, they repackage organic matter into denser and faster sinking particles (Turner, 2015). On the contrary, they may be responsible for a significant part of remineralization through respiration and contribute to the fragmentation of organic aggregates by swimming (Dilling and Alldredge, 2000; Goldthwait et al., 2004). Some zooplankton also feed on fecal pellets, thereby reducing the downward flux of organic matter (Iversen and Poulsen, 2007).

Together with zooplankton, bacteria remineralize most of particles in the water columnn. Thus, their effect on the biological pump efficiency is mostly negative. However, they also produce TEP that promote aggregation of particles and thereby enhance carbon flux (Passow, 2002).

1.3.5.3 Influence of physical parameters

Physical dynamics influence both primary production and carbon transport to the deep. Eddies, as well as upwellings and fronts, bring episodic pulses of new nutrients into the euphotic zone or initiates stratification, increasing primary production (McGillicuddy and Robinson, 1997; Bidigare et al., 2003; Mahadevan et al., 2012) and thus carbon export efficiency. A large part of dissolved inorganic carbon is also exported through physical

[4]Efficiency refers to the capacity of the biological pump to sequester as much carbon as it could be. Two definitions of efficiency exist: it can be either the ratio between export flux and primary production or the ratio between sequestration and export flux.

mixing and diffusion (Burd et al., 2010). Flux attenuation is also affected by physical controls. Remineralization rates are influenced by pressure: as pressure increases in the ocean interior, less bacteria are observed on particles and a shift in microbial communities is observed (Tamburini et al., 2009). Besides, remineralization has been shown to be temperature dependant (White et al., 2012; Laufkötter et al., 2017).

1.4 Relationships between the biological carbon pump and the Earth's climate

1.4.1 The biological carbon pump in the past

Cyclic variations of climate and glaciations happened during the last two millions years. The pace of glacial and interglacial periods successions is explained by changes of orbital parameters of the Earth (known as the "Milankovitch cycles"), with characteristic frequencies of 100, 41 et 23 thousand years (Hays et al., 1976; Berger, 1988). Although these orbital variations explain much of the glacial/interglacial oscillations, they do not account for their amplitude and rapid transitions demonstrated by palaeoclimatic and palaeoceanographic records. As a consequence, a positive feedback from the climate system must amplify these cycles.

From the measurement of CO_2 concentration in air bubbles trapped in ice cores (Petit et al., 1999), it was found that during interglacial periods, the atmospheric partial pressure of CO_2 was lower, near 280 parts per million by volume (p.p.m.v.), than during peak glacial times when it lay between 180 and 200 p.p.m.v. (figure 1.12, Sigman and Boyle, 2000). As CO_2 is a greenhouse gas, changes in its concentration may play a significant role in the energetics of glacial/interglacial oscillations. The identification of the processes responsible for these variations motivated intensive research (reviewed in Sigman and Boyle, 2000 and Turner, 2015). Among them, changes in efficiency of the biological carbon and their impact on the $CaCO_3$ cycle has retained the attention of palaeoclimatologists and paleoceanographers.

The deep sea stores ten times more carbon than the terrestrial biosphere, soil, atmospheric and warm upper ocean carbon reservoirs combined (figure 1.1). Because deep water is exposed to the surface only roughly every 1,000 years, changes in concentration of atmospheric CO_2 driven by anything else than the ocean would be diluted into the large reservoir of deep DIC, which would attenuate the sharp CO_2 concentration changes over glacial and interglacial times. From these considerations, Broecker (1982a,b) concluded that changes in CO_2 concentration must be due to oceanic processes.

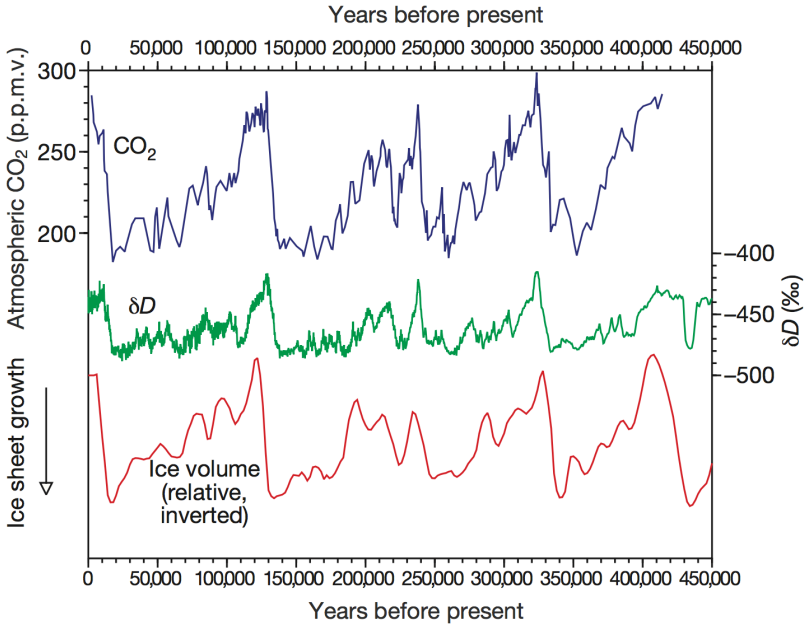


Figure 1.12 - History of atmospheric CO_2 concentration during the last 420 thousand years, measured in the Vostok ice cores (Antarctica). δD is the ratio of deuterium to hydrogen in ice and gives a proxy for air temperature over Antarctica. Global ice volume is based on benthic foraminiferal oxygen isotope data from deep-sea sediment cores. It is plotted as relative sea level, so peak glacial times appear as sea level minima. These results show that atmospheric CO_2 was one of the earliest parameters to change at the end of glacial maxima, roughly in step with Southern Hemisphere warming and preceding the decline in the Northern Hemisphere ice volume (Sigman and Boyle, 2000).

Several potential explanations have been proposed: changes in terrestrial carbon storage, ocean temperature, ocean alkalinity and changes in the marine calcium carbonate budget (Sigman and Boyle, 2000). However, none of these propositions seem to be sufficient to explain the 80 to 100 p.p.m.v. difference in atmospheric CO_2 concentration between ice ages and interglacial periods. In 1982, Broecker hypothesized that a strengthened biological pump during glacial times would be the main cause of the lower CO_2 levels. Changes in nitrate and phosphate concentrations would be at the origin of this increase. Two mechanisms could enhance primary production and export flux to the deep ocean. Nitrate and phosphate concentrations could have been increased in low- and mid-latitude surface waters where the current low levels limit the fixation of dissolved inorganic carbon for primary production of organic matter. Another mechanism, supported by Sigman and Boyle (2000), involves a more complete consumption of nitrate and phosphate in high-latitudes waters compared to the present.

A higher N_2 fixation during glacial periods, due to an increased airborne iron supply to the open ocean (Falkowski, 1997), and a decrease in water column denitrification (Ganeshram et al., 1995) have been proposed as the causes of the increase of the oceanic nitrate reservoir, which would have led to a greater export flux and potentially lower CO_2 levels during glacial periods. However, as phosphorus is considered to be a limiting

nutrient on glacial/interglacial timescales, this hypothesis implies that marine organisms must be able to deviate from the Redfield ratio (C:N:P = 106:16:1) to deal with the lack of phosphorus compared to nitrogen.

In high-latitude oceans, the thermocline (that separates the upper mixed layer from the deep waters below) is very shallow to non existent. The nutrient- and CO_2-rich deep waters brings nutrients and CO_2 to the surface and return to the subsurface before all nutrients are utilized by phytoplankton for production of organic matter. This incomplete consumption of nutrients allows CO_2 to be released in the atmosphere. Sigman and Boyle (2000) hypothesized that changes in CO_2 atmospheric levels during glacial times was the result of an enhanced nutrient utilization in the Southern Ocean. This would be due to an increase in carbon export (possibly to the input of iron from dust) and a decrease in the exposure of deep waters at the surface (which would be the result of a northward shift and decrease in strength of eastward winds). In their review of 2010, Sigman et al. concluded that changes in efficiency of the biological carbon pump in the Southern Ocean were an important factor of glacial/interglacial changes in atmospheric CO_2.

Another explanation for a strengthened biological pump during ice ages involves diatoms. These micro-algae, identified as key players of the biological pump, require silica for their growth. Since the Southern Ocean is a hub for the global circulation, the dynamics of silicic acid in the Southern Ocean can have important consequences in diatom dynamics and thus on the biological pump and atmospheric CO_2 levels (Dugdale and Wilkerson, 2001; Sarmiento et al., 2004; Benoiston et al., 2017). The potentially reduced Si:C uptake ratio of siliceous producers during ice ages (Pichevin et al., 2009) under conditions of increased iron availability from enhanced dust input, together with nutrient-rich upwelling waters in the Southern Ocean (Sarmiento et al., 2004) may have supported diatom growth and thus invigorated the biological pump.

1.4.2 The biological carbon pump in the Anthropocene

Since the beginning of the industrial era, humans disrupt the carbon cycle by burning fossil fuels, manufacturing cement and changing land use (i.e. deforestation), which emit CO_2 in the atmosphere (figure 1.1, Ciais et al., 2013; Le Quere et al., 2009). The ocean is the largest sink for anthropogenic CO_2: between 20 and 40% of emitted CO_2 has been absorbed by the ocean during the last two centuries (Ciais et al., 2013; Khatiwala et al., 2013; DeVries, 2014). As CO_2 is a greenhouse gas, the remaining CO_2 in the atmosphere contributes to the global warming of the Earth. Besides, the dissolution of atmospheric CO_2 drives the ocean more acid (it creates H_2CO_3 that increases water's acidity). How

the biological pump will respond to these perturbations remains uncertain (Passow and Carlson, 2012).

1.4.2.1 Global warming

Different scenarios were reviewed by Passow and Carlson (2012) and Turner (2015). One possible consequence of ocean warming is an increased stratification of the ocean's surface layer. This may lead to a decreased input of nutrients from the deeper layers, which would decrease primary production and carbon export, particularly in the tropical and subtropical ocean (Bopp et al., 2001; Doney, 2006). This decrease in available nutrients could result in a shift in phytoplanktonic communities from diatoms to coccolithophorids (Cermeño et al., 2008) or from diatoms to small microflagellates and cyanobacteria (Falkowski and Oliver, 2007). Using niche models, Flombaum et al. (2013) projected increases in cell numbers and in area of the cyanobacteria *Prochlorococcus* and *Synechococcus* as a result of global warming by the end of the 21st century. Zooplanktonic communities could also be affected by the ocean warming: a shift from ecosystems by large zooplankton towards communities dominated by microzooplankton may reduce the export of particulate detrital food to depth (Smith et al., 2008). Warming could also accelerate heterotrophic microbial decomposition, thus reducing carbon export efficiency (Riebesell et al., 2009).

On the contrary, warming may also increase wind and/or storm frequency, which would promote injection of nutrients from below and increase primary production (Peters, 2008). Evidences of the ability of some diatoms to grow in stratified environments thanks to diverse strategies could also promote diatom blooms as observed in some environments (Dore et al., 2008; Kemp and Villareal, 2013), which could enhance the biological pump. In polar oceans, reductions in sea ice coverage may have implications on biological communities and the biological pump. A decline in krill has been observed in the Southern Ocean. They are replaced by salps that produce large and fast-sinking fecal pellets that enhance export flux (Loeb et al., 1997). Enhanced formation of marine snow and other organic aggregates may also be a result of warming. Both warming and acidification have been found to increase TEP formation (Engel et al., 2004). As TEP favours aggregation of particles, this could lead to faster sedimentation rates and greater export to depth (Egge et al., 2009).

1.4.2.2 Acidification

Contrary to enhanced TEP formation, a decrease in the ocean's pH is expected to have a positive feedback to global warming by reducing calcification of calcareous plankton, such as coccolithophorids (e.g. Bach et al., 2012), thereby reducing the export flux. Ocean acidification may also result in changes in the nitrogen cycle. They include increased nitrogen fixation by cyanobacteria (Hutchins et al., 2009), greater denitrification rates due to the expansion of suboxic habitats (Deutsch and Weber, 2012) and decreased nitrification (Beman et al., 2011).

This review of possible outcomes of global warming and ocean acidification, although not exhaustive, demonstrates that the responses of the biological carbon pump to climate change are diverse and lean sometimes towards increased carbon flux, sometimes towards decreased carbon flux. These contrasting results may be due to the fact that the biological carbon pump is the result of multiple interacting parameters (e.g. temperature, nutrients availability, planktonic composition, and even pollution) that affect the efficiency of carbon export, flux attenuation and sequestration, not to mention other anthropogenic perturbations such as pollution (e.g. Rochelle-Newall et al., 2008) that may affect carbon fluxes.

1.5 Marine plankton

In sections 1.3 and 1.4, we reviewed the role of marine plankton in controlling the processes of the biological carbon pump but their role is not limited to primary production, remineralization and contribution to vertical carbon fluxes. Instead, they contribute to most biogeochemical cycles (see section 1.5.2) and are at the basis of the oceanic trophic chain. Most of them are invisible to the naked eye but they represent 95% of the total marine biomass. This highly diverse group of organisms is present in all aquatic environments (e.g. marine, freshwater) and is of vital importance for these ecosystems.

1.5.1 Plankton diversity

The word "plankton" comes from the greek πλανκτός (planktós) meaning "to wander". It designates organisms that are adrift on the currents. Plankton do not constitute a monophyletic group (i.e. all planktonic lineages do not derive from the same direct common ancestor). Plankton gather highly diverse organisms in terms of size (figure 1.13), morphology, taxonomy, physiology and trophic strategies. They comprise viruses, bacteria, archaea, protists (unicellular eukaryotes) and metazoa (i.e. multicellular eukaryotes, in-

cluding eggs and larval stages). The size spectrum of these organisms ranges from the tenth of a micrometer like viruses, to several meters, such as siphonophores and jellyfishes.

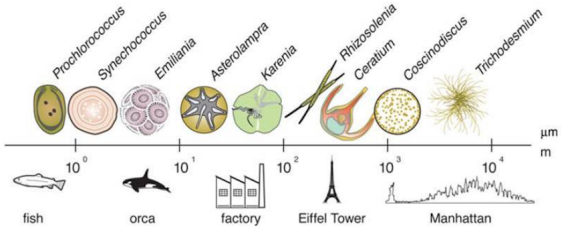


Figure 1.13 - Phytoplankton size's diversity (Finkel et al., 2010).

Historically, plankton have been divided in *phytoplankton* and *zooplankton*. This partition refers to distinct trophic strategies. Phytoplankton correspond to organisms that are able to produce their own organic matter: they are photoautotrophic. In addition to numerous unicellular eukaryotic lineages like diatoms, coccolithophorids and dinoflagellates, phytoplankton also includes photosynthetic bacteria, such as cyanobacteria. All these lineages combined are responsible for almost half of Earth's primary production (Falkowski et al., 1998). Zooplankton refers to heterotrophic organisms and are the main consumers of phytoplankton. It includes members of metazoa but also protozoans. Some of them spend their entire life in the water column (the *holoplankton*) while others are part of plankton only as larvae (the *meroplankton*). The meroplankton leave its planktonic existence by growth or metamorphosis to belong to nekton or benthos. Finally, more and more studies are showing since the last decades that the majority of planktonic eukaryotic organisms are not strictly phototrophs or heterotrophs, but mixotrophs. *Mixotrophy* is defined as the ability to combine the use of autotrophy and heterotrophy in a single organism (Caron, 2016). This underestimated trophic strategy (Faure et al., 2019, see appendix A) exists in originally photosythetic organisms capable of phagotrophy (e.g. many dinoflagellate lineages), but also in heterotrophic organisms that acquired the ability to photosynthesize (through kleptoplasty or symbiosis with photosynthetic organisms).

1.5.2 Biogeochemical importance

As exemplified by the carbon cycle presented in section 1.1, elements cycles are the result of continuous exchanges and interactions between biological and geological components. Biological fluxes of elements essential to life (hydrogen, carbon, nitrogen, phosphorus, oxygen and sulphur) are mainly catalyzed by microbes. For this reason, microbial life is considered as the "engine that drives Earth's biogeochemical cycles" (Falkowski et al., 2008).

Carbon forms the structure of all organic molecules on the planet, representing around 50% of dry biomass. The role of plankton in the carbon cycle is critical. Phytoplankton, that lives in the euphotic zone of the ocean, use dissolved inorganic carbon to produce organic matter through photosynthesis. Under favourable conditions (i.e. higher temperatures and day length, met in spring and summer), phytoplankton thrive and constitute blooms as illustrated on figure 1.14. Oceanic photosynthesis produces half the oxygen we breathe. Dissolved CO_2 is also used by some micro-organisms to build their calcareous shell, such as coccolithophorids. This newly produced biomass serves as the basis of oceanic life as it is consumed by zooplankton and heterotrophic bacteria, thereby recycling (i.e. remineralizing) essential nutrients (up to half of primary production is recycled by bacteria through the microbial loop, Azam et al., 1983). The organic matter that escaped remineralization sinks to the depth and eventually reaches the ocean floor where it can be stored on geological time scales (i.e. over millions of years). All these processes are referred to as the biological carbon pump (see section 1.3).

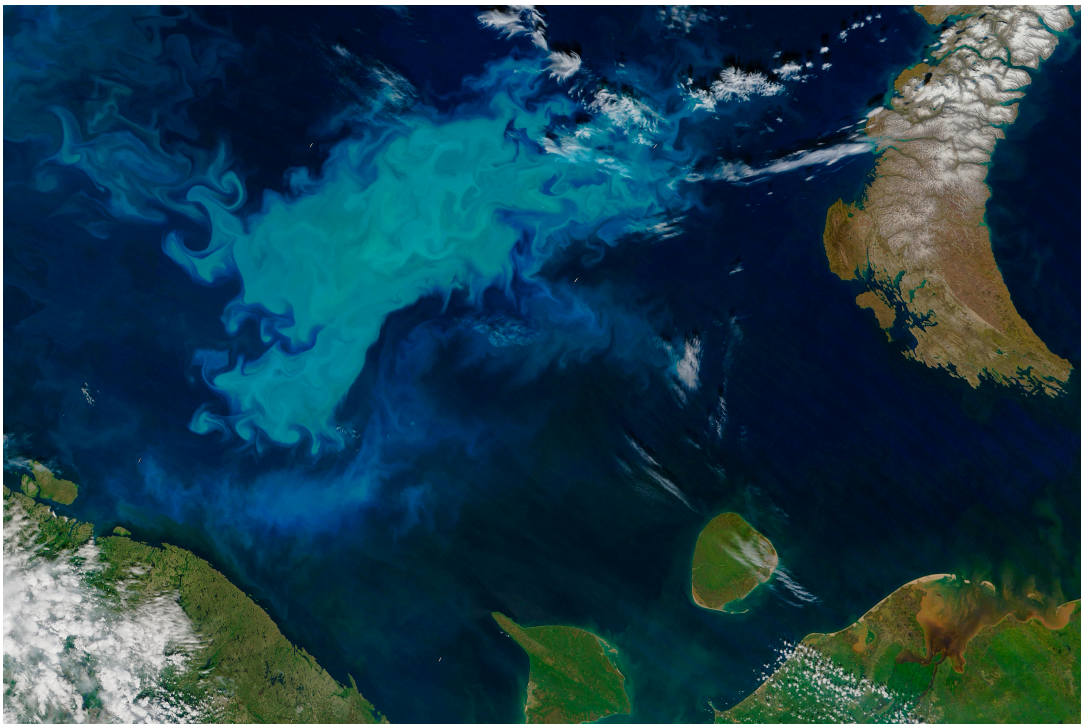


Figure 1.14 - A phytoplankton bloom in the Barents Sea (marginal sea of the Arctic Ocean). This satellite picture was taken on July 6, 2016 by the Moderate Resolution Imaging Spectroradiometer (MODIS). The milky colour of this bloom suggests that it might be constituted of coccolithophores that tend to thrive in the Barents Sea from July to September (NASA, 2016).

Another cycle in which microorganisms have important roles is the nitrogen cycle (figure 1.15). Together with phosphorus, nitrogen is an important constituent of biomass. Outside high-nutrient low-chlorophyll areas (HNLC, where chlorophyll *a* levels are lower than expected given nitrate and phosphate concentrations), nitrogen is a limiting factor of pri-

mary production. The largest pool of nitrogen is in the form of dinitrogen (N_2). To be usable by autotrophic plankton, N_2 needs to be transformed into ammonium (NH_4^+) by N_2-fixing microorganisms (or diazotrophs), such as the filamentous cyanobacteria *Trichodesmium* (Capone et al., 2005). However, most of the nitrogen required by phytoplankton is supplied by the remineralization of organic matter by heterotrophic bacteria that releases NH_4^+. This molecule can be directly assimilated but most of ammonium is thought to be converted into nitrite (NO_2^-) and and then to nitrates (NO_3^-) through nitrification, each step being performed by specialized groups of microorganisms. Among them, members of Proteobacteria and archaea from the genus *Thaumarchaeota* were found to be major players in oceanic ammonia oxidation (Francis et al., 2005). Finally, some microorganisms can convert nitrite and nitrate back to N_2 through denitrification. This process leads to nitrogen loss and occurs in oxygen minimum zones (OMZ) that are predicted to expand due to global warming (Stramma et al., 2008). In these regions, nitrogen loss can also occur through anammox, which is the anaerobic oxidation of NH_4^+ to N_2 (Kuypers et al., 2003).

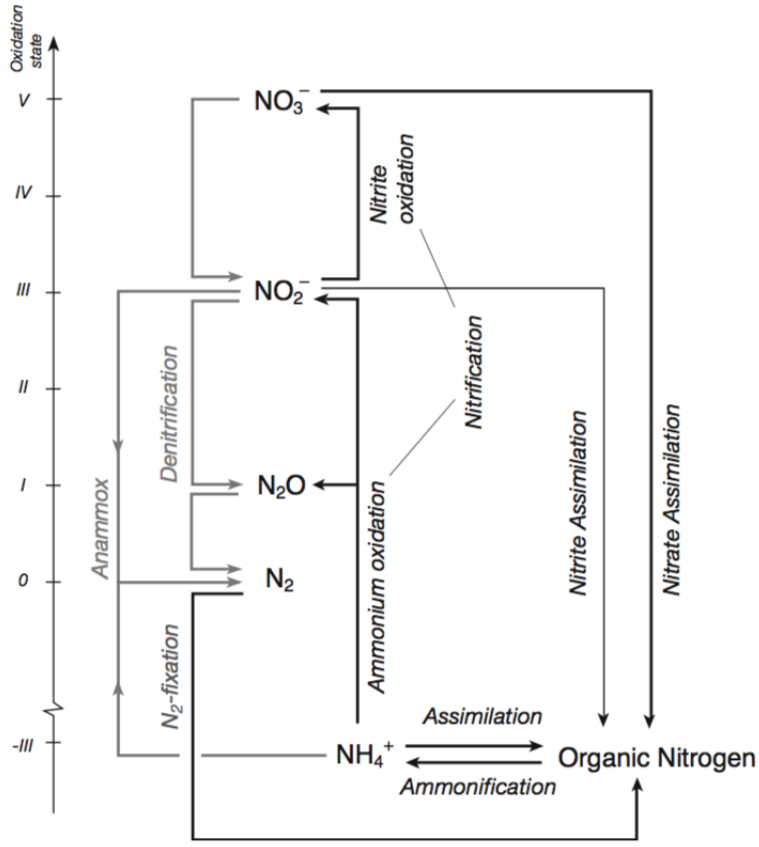


Figure 1.15 - The nitrogen cycle in the ocean (Sarmiento and Gruber, 2006).

Two other cycles worth mentioning are those of phosphorus and iron. Microorganisms actively participate to the phosphorus cycle. In some environments, phosphorus is a

limiting nutrient for primary production. The main source of phosphorus in the ocean is river runoff. Once dissolved in seawater, it is incorporated in organic material by phytoplankton. Microorganisms are the main actors of the remineralization of organic phosphorus compounds. Unlike phosphorus, the largest external source of iron for the ocean is from aeolian dust deposition. This micronutrient is important in regulating ocean primary productivity. In HNLC regions, iron is the limiting factor for primary production. The role of iron is crucial for photosynthesis, respiration and nitrogen fixation as it acts as a co-factor of many cellular enzymes.

1.6 The *Tara* Oceans expeditions

1.6.1 Objectives

Since the *Challenger* expedition (1872–1876) that laid the foundations of modern oceanography and led to the description of planktonic forms by the German naturalist Ernst Haeckel, many expeditions allowed scientists to decipher marine ecosystems. Recently, the Global Ocean Survey (2003-2010) launched by Craig Venter aboard the *Sorcerer II* showed that large environemental sampling coupled with new molecular technologies could improve our knowledge on marine microbial diversity (Venter et al., 2004). Later, the *Malaspina* expedition (2010-2011) explored microbial biodiversity in the deep ocean (Duarte, 2015).

Inspired by the voyage of the *Beagle* described by Charles Darwin in 1839, the *Tara* Oceans expeditions were initiated by Eric Karsenti, then director of the cellular biology and biophysics department of the EMBL (European Molecular Biology Laboratory). The goal of the project was to understand the spatio-temporal structure of planktonic ecosystems and coral reefs on a global scale (Karsenti and Di Meo, 2012). Important efforts were also made to raise awareness amongst general public, especially children, in all countries *Tara* visited.

1.6.2 Sampling methods

The *Tara* Oceans circumglobal expedition (2009-2013) travelled across the oceans to collect a wide variety of organisms spanning five orders of magnitude (from viruses to metazoans) and measure associated environmental data (Karsenti et al., 2011). The sampling route went through the Mediterranean and Red Seas, the Indian Ocean, South Atlantic, Antarctic and Pacific, and came back through North Pacific and North Atlantic. The last part of the voyage consisted in a circumnavigation of the Arctic ocean (figure

1.16). In total, 210 stations were sampled in 20 biogeographic provinces, collecting around 40,000 samples (Pesant et al., 2015).

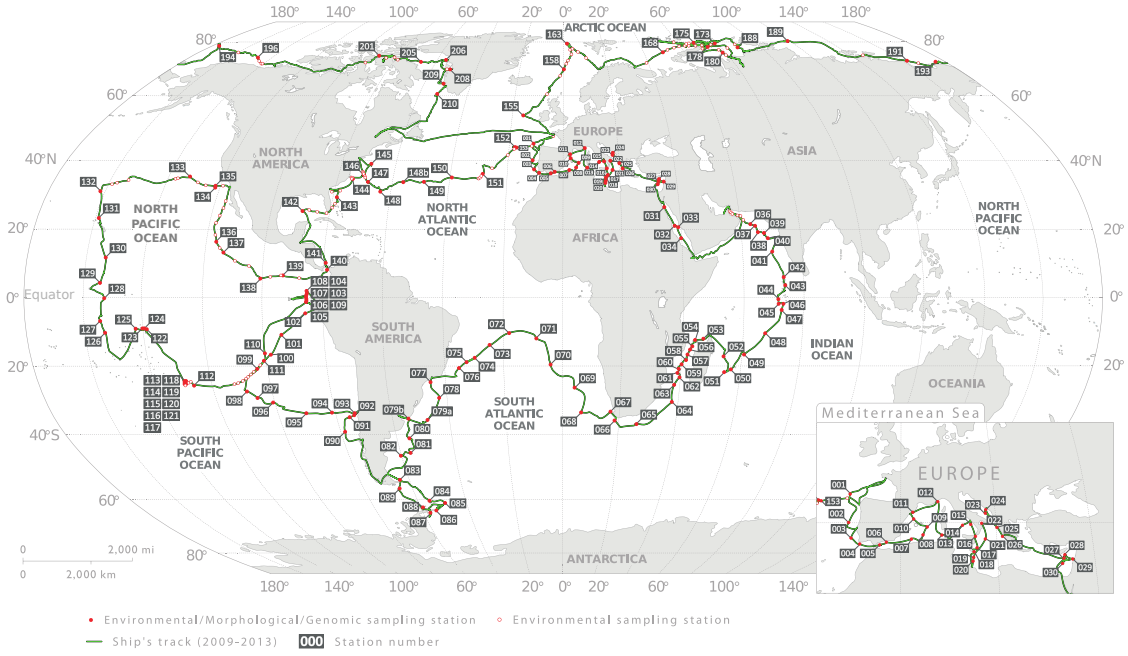


Figure 1.16 - Sampling route and stations of the Tara Oceans Expeditions that crossed the oceans between 2009 and 2013 (Pesant et al., 2015).

Plankton was collected from three depths. In the ocean's surface (~0-200 m), it was collected between 3 and 7 meters below the surface (these samples were labelled "SUR" or "SRF") and in the deep chlorophyll maximum ("DCM") which was determined with a chlorophyll fluorometer. Depending on the stations, plankton was also sampled in the mesopelagic zone (~200-1000 m), labelled "MES"[5]. Various sampling methods were used to capture the diversity of plankton, including Niskin bottles and plankton net tows (figure 1.17A and Pesant et al., 2015). Organisms were separated into 10 fractions, ranging from <0.2 μm for viruses to 2,000 μm for large unicellular eukaryotes and metazoans. Prokaryotes ((eu)bacteria and archaea), that are the subject of the research work presented in chapters 3 and 4, were collected in the 0.2-1.6 μm (up to station #52) and 0.2-3 μm (from station #56) size fractions (figure 1.17A and B) (Alberti et al., 2017).

1.6.3 Subsequent analyses and results

Following the sampling, high throughput quantitative imaging and sequencing were performed. Imaging instruments included on-board and on-land FlowCams and ZooScans,

[5]The research work presented in chapters 3 and 4 focused on SRF and DCM samples.

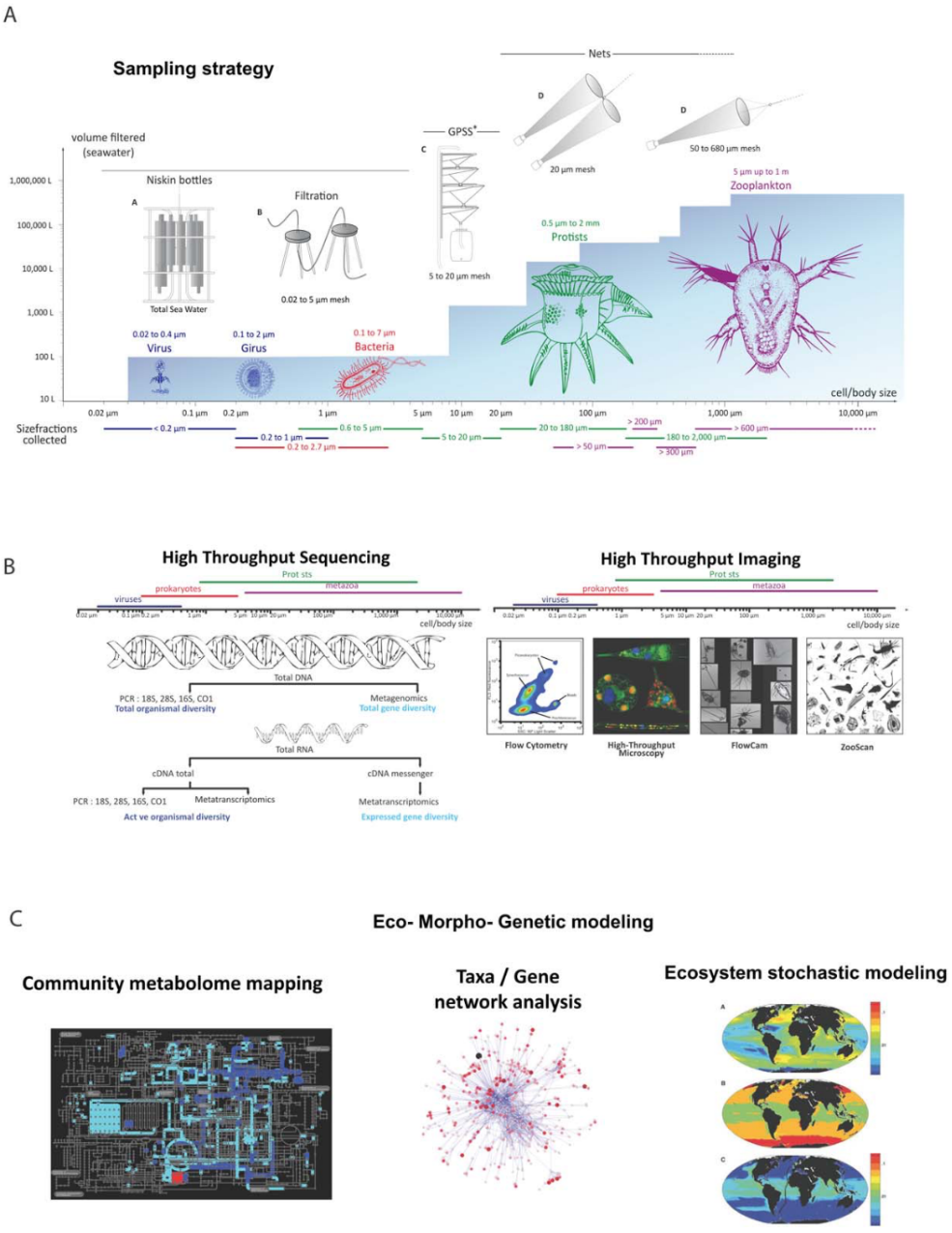


Figure 1.17 - Philosophy of The Tara Oceans project (from sampling to analyses). (A) Methods for sampling organisms by size classes and abundance. The blue background indicates the filtered volume required to obtain sufficient organism numbers for analysis. (B) Methods for analyzing samples. Data on the right are from Tara Oceans sampling stations. (C) Models that will benefit from Tara Oceans data. High throughput genome sequencing and quantitative image analysis provide evolution, metabolic, and interaction data to build community metabolome maps, taxa/gene networks, and spatial ecosystem models (Karsenti et al., 2011).

flow cytometry, light sheet, confocal and electron microscopes (figure 1.17B). High throughput sequencing was performed at the Genoscope (integrated in the French Alternative Energies and Atomic Energy Commission). The sequencing strategy relied on metabarcoding, metagenomic, single-cell genomic and metatranscriptomic approaches (figure 1.17B). Detailed sequencing methods are described in Alberti et al. (2017). Data derived from these

analyses led to major publications (e.g. Sunagawa et al., 2015; Lima-Mendez et al., 2015; Vargas et al., 2015; Brum et al., 2015; Guidi et al., 2016; Vincent et al., 2018) that revealed the unexplored marine microbial diversity. All domains of life (viruses, prokaryotes and eukaryotes) were investigated, giving insights into biodiversity, biogeographic patterns and environmental drivers of planktonic communities. The complex ecological interplay between microorganisms was also investigated through taxa network analysis (figure 1.17C) that showed that associations within plankton are not randomly distributed and abiotic factors are incomplete predictors of planktonic community structure (Lima-Mendez et al., 2015).

1.7 Research questions and thesis outline

In this general introduction, we reviewed the role of plankton in biogeochemical cycles. In particular, we highlighted their influence on the ocean biological carbon pump. For millions of years, the biological carbon pump appears to have had a substantial effect on the Earth's climate. Today, the ocean buffers the CO_2 anthropogenic emissions, a process in which plankton is actively involved by sequestering a part of it in the deep ocean. Decades of research on the biological pump pointed out major contributors to this phenomena such as diatoms and copepods. However, processes of the biological carbon pump are rather the result of intricate planktonic relationships, which are still poorly understood. Besides, the rapidly changing ocean may strongly impact these relationships and plankton biodiversity, abundance and biogeography. The application of environmental omics to the study of these problems is currently expanding our knowledge in these areas. Consequently, the following research questions have driven the analyses presented in chapters 2, 3 and 4:

1. How can environmental omics (meta-omics) improve our knowledge on the ecological and biogeochemical functions of marine microbes, particularly diatoms because of their large contribution to primary production and carbon export?

2. How can primary production, carbon export and flux attenuation processes be integrated to revisit the study of the biological carbon pump?

3. Are these processes characterized by distinct microbial association networks?

4. Can the health state of the biological carbon pump be predicted by meta-omics?

Attempts to answer this questions are presented in the following chapters:

Chapter 2 is a reprint of the review article "The evolution of diatoms and their biogeochemical functions" published in *Philosophical Transactions of the Royal Society B* on which I am co-first author. Diatoms are key players in primary production and export of carbon and silica. This review highlighted the benefit of omics to study of the origin, evolution and diversification of diatoms, but also to reveal their ecological and biogeochemical functions.

Chapter 3 is composed of a presentation of microbial association networks and of a draft manuscript that will be submitted to *The ISME Journal*. In this draft manuscript, we continued the work of Guidi et al. (2016) by defining a new framework to study the biological carbon pump, including the three processes inherent in this phenomenon: primary production, carbon export and flux attenuation. Basically, we defined states of the biological carbon pump, corresponding to situations where one of the processes is dominant compared to the others. Beyond the simple compositional analysis of the samples studied, interactions within prokaryotic plankton communities were explored to test whether the states are characterized by differing microbial association networks. Overall, results suggest that the states are defined by contrasting microbial associations rather than microbial composition.

Chapter 4 presents an overview of machine learning techniques followed by a draft manuscript in which we assessed whether environmental genomics can be used to predict the state of the biological carbon pump and to highlight environmental biomarkers. We tested this hypothesis with random forests using prokaryotic metabarcoding data. Globally, although the classification error rate of the samples included in the analysis is high, we show that, on condition that improvements are added to our model, random forests can be a useful tool to identify biomarkers of the state of the biological carbon pump.

Chapter 5 concludes with a discussion of the assumptions and shortcomings of the methods used in this thesis (i.e. microbial networks inference and machine learning) and lays out perspectives for the study of microbial interactions and their involvement in the carbon cycle.

Chapter 2

Contribution of diatoms to the carbon cycle

This chapter is a reprint of a *Philosophical Transactions of the Royal Society B* article on which I am co-first author.

This review, which was part of the themed issue "The peculiar carbon metabolism in diatoms", presented the evolution and contribution to biogeochemical cycles of a group of eukaryotic microalgae, the diatoms. Indeed, diatoms are key players in primary production and export of carbon and silica. They thrive in upwelling regions at high latitudes and their silicified cell wall serves as a ballast that makes them important contributors to export production. This review highlighted the benefit of genomics to the study of the origin, evolution and diversification of diatoms, but also to reveal their ecological and biogeochemical functions.

I was in charge of producing figure 1 on the major evolutionary and biogeochemical events during the history of life on Earth and figure 2 on the biological carbon pump. I did the bibliography research, extracted data, designed and produced the figures.

2.1 Article 1 (Benoiston et al. 2017): The evolution of diatoms and their biogeochemical functions

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Benoiston A-S, Ibarbalz FM, Bittner L, Guidi L, Jahn O, Dutkiewicz S, Bowler C. 2017 The evolution of diatoms and their biogeochemical functions. *Phil. Trans. R. Soc. B* **372**: 20160397.
http://dx.doi.org/10.1098/rstb.2016.0397

Accepted: 24 March 2017

One contribution of 16 to a theme issue 'The peculiar carbon metabolism in diatoms'.

Subject Areas:
ecology, evolution, genomics

Keywords:
biogeochemistry, carbon export, diatom, geological record, genomics, photosynthesis

Author for correspondence:
Chris Bowler
e-mail: cbowler@biologie.ens.fr

†These authors contributed equally to this study.

THE ROYAL SOCIETY PUBLISHING

The evolution of diatoms and their biogeochemical functions

Anne-Sophie Benoiston[1,†], Federico M. Ibarbalz[2,†], Lucie Bittner[1], Lionel Guidi[3,4], Oliver Jahn[5], Stephanie Dutkiewicz[5] and Chris Bowler[2]

[1]Sorbonne Universités, UPMC Univ. Paris 06, Univ. Antilles, Univ. Nice Sophia Antipolis, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS), 75005 Paris, France
[2]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR8197, INSERM U1024, 46 rue d'Ulm, 75005 Paris, France
[3]Sorbonne Universités, UPMC Univ. Paris 06, CNRS, Laboratoire d'Océanographie de Villefranche (LOV) UMR7093, Observatoire Océanologique, 06230 Villefranche-sur-Mer, France
[4]Department of Oceanography, University of Hawaii, Honolulu, HI 96822, USA
[5]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 54-1514 MIT, Cambridge, MA 02139, USA

CB, 0000-0003-3835-6187

In contemporary oceans diatoms are an important group of eukaryotic phytoplankton that typically dominate in upwelling regions and at high latitudes. They also make significant contributions to sporadic blooms that often occur in springtime. Recent surveys have revealed global information about their abundance and diversity, as well as their contributions to biogeochemical cycles, both as primary producers of organic material and as conduits facilitating the export of carbon and silicon to the ocean interior. Sequencing of diatom genomes is revealing the evolutionary underpinnings of their ecological success by examination of their gene repertoires and the mechanisms they use to adapt to environmental changes. The rise of the diatoms over the last hundred million years is similarly being explored through analysis of microfossils and biomarkers that can be traced through geological time, as well as their contributions to seafloor sediments and fossil fuel reserves. The current review aims to synthesize current information about the evolution and biogeochemical functions of diatoms as they rose to prominence in the global ocean.

This article is part of the themed issue 'The peculiar carbon metabolism in diatoms'.

1. Introduction

Microscopic photosynthetic plankton (phytoplankton) provide the organic biomass on which almost all ocean life depends and fuel a range of essential biogeochemical processes, ranging from the generation of oxygen, the recycling of elemental nutrients, and the removal of carbon dioxide from the atmosphere. They are responsible for around 45% of global primary production and yet represent only 1% of Earth's photosynthetic biomass [1], due to their rapid proliferation times and because all cells are photosynthetically active, unlike multicellular plants. Our appreciation of the roles of these microscopic organisms in the ocean has been transformed over the last decades by improved methods to explore the chequered history of life on Earth and by new DNA sequencing technologies. Scientists are using these resources to address the feedbacks between plankton and the climate system, because planktonic organisms can both influence climate and be affected by climate change [2]. As a major component of plankton communities in today's oceans diatoms are now key to their functioning, yet they rose to prominence only quite recently. Through photosynthesis they provide large amounts of organic material that sustains marine ecosystems as well as contributing to Earth's carbon cycle, and play major roles in the biogeochemical cycling of other nutrients such as nitrogen and silicon [3–5]. Their evolution can be traced back to the origin of photosynthesis.

© 2017 The Author(s) Published by the Royal Society. All rights reserved.

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

2. Photosynthesis as the engine of life

Oxygenic photosynthesis is arguably the most important process in nature. It boosted the remarkable history of life on Earth following its appearance at least 2.4 billion years ago [6] (figure 1*a*). In spite of its early evolution it represents the most complex energy transduction system known; its water oxidizing machine has no analogues elsewhere and its functioning is still poorly understood [16]. The oxidizing or 'splitting' of water was made possible by the coupling of two photosystems that enabled oxygenic photosynthetic bacteria to use light energy to generate oxygen from water and reducing power in the form of NADPH. The oxygen generated from the process subsequently accumulated in the atmosphere and is one of Earth's distinguishing features, because molecular oxygen is extremely rare in the Universe [17]. The utilization of light energy to split water in oxygenic photosynthesis also allows the fixation of CO_2 into organic matter that fuels the food chain.

Oxygenic photosynthesis first evolved in the cyanobacteria, which remain the only prokaryotes capable of performing it. Oxygen initially began to accumulate only slowly in the atmosphere because it was first consumed in oxidation reactions with abundant compounds that contained reduced forms of iron, sulfur, carbon, nitrogen, and other abundant materials, and because it was consumed in the biological process of respiration, which evolved after photosynthesis [8].

Following the evolution of oxygenic cyanobacteria it took around 2 billion years before complex multicellular animal life evolved (figure 1*a*). During this time, eukaryotic organisms appeared bearing the first mitochondria derived from the endosymbiosis of a proteobacterium in an Archaean-like cell, in which respiratory processes could occur [18]. Unambiguous fossils of eukaryotes have been found in shales as old as 1.65–1.85 billion years [19]. Subsequently, chloroplasts evolved following the invasion or engulfment of a cyanobacterium into the prototypic eukaryote. Photosynthetic eukaryotes are considered to have evolved around 1.2 billion years ago [12] although the forms that dominate today's ocean are predominantly derived from additional or 'secondary' endosymbiotic events in which eukaryotic green or red algae were incorporated a second time into a eukaryotic cell [20]. The timing of these events is not well resolved but it certainly happened prior to the appearance of multicellular lifeforms during the Cambrian explosion and preceded a major increase in atmospheric oxygen to levels similar to those found today, from around 1–5% to about 20% (figure 1*a*). The reason why the rise of photosynthetic eukaryotes stimulated such a dramatic increase in oxygen may be a consequence of carbon export to the seafloor, because their larger cells were more strongly ballasted and therefore more likely to sink than cyanobacteria [2] (figure 2). The consequent burial of carbon sequestered it away from the carbon cycle and so it could not be remineralized back to CO_2 by oxidative respiration. Alternatively (or additionally), photosynthetic activity may have increased significantly following the evolution of extensive planktonic ecosystems, e.g. fuelled by increased nutrient availability during this period. Regardless of the cause, atmospheric CO_2 levels dropped significantly during this period, which may have contributed to one or more of the Snowball Earth events that have been documented to have occurred [13] (figure 1*a*,*b*), because CO_2 is a powerful greenhouse gas. Furthermore, the increase in molecular oxygen was probably instrumental in permitting multicellular life to evolve during more temperate periods because it allowed the development of more complex organisms less constrained by oxygen acquisition from a low oxygen environment.

Atmospheric oxygen concentrations have remained relatively stable at around 20% since the early Cambrian period. The emergence of land plants during the Devonian around 400 million years ago (Ma) likely led however to a further large increase in oxygen concomitant with CO_2 drawdown from the atmosphere (figure 1*b*) [23]. Although it did not persist, the elevated oxygen concentrations may have led to the evolution of giant insects and other large animals. Atmospheric oxygen in today's world, while being similar to concentrations prior to the evolution of land plants, is now likely to be maintained principally by terrestrial plants that release oxygen directly to the atmosphere rather than by photosynthetic plankton because the oxygen generated within the water column is likely to be consumed by other organisms rather than being outgassed [24,25]. The release of biogenic oxygen from the ocean may nonetheless be significant in some regions and is likely to be sensitive to temperature changes [26].

The detailed analysis of the geological record left by dead eukaryotic plankton sinking to the seafloor over the last hundreds of millions of years, either based on biomarker molecules or microfossils, has revealed their history during major Earth transitions [27]. By likely underpinning the rise of oxygen that led to the evolution of multicellular organisms, they may have promoted the development of ever more complex lifeforms, not only in the ocean, but also on land as well. Besides the process of photosynthesis, the later appearance of calcification and silicification in some phytoplanktonic organisms (e.g. in coccolithophorids and diatoms, respectively), in addition to more ancient organisms such as foraminifers and radiolarians, permitted the precipitation of hard materials to the ocean interior, as well as organic carbon (figure 1*a*,*b*). A rich amount of data from microfossils, biomarkers, and molecular clocks using conserved marker genes indicate that these processes appeared in photosynthetic organisms around 200 Ma and permitted atmospheric CO_2 to be further sequestered into the deep ocean in the form of organic carbon and calcium carbonate, which over time contributed to the formation of sedimentary rocks such as limestones and cherts, as well as our oil and gas reserves [12,28–30] (figures 1 and 2). This, together with increased weathering and changes in ocean circulation, is believed to have initiated a period of declining atmospheric CO_2 concentrations, contributing to the switch from a greenhouse climate in the Mesozoic to an icehouse climate in the Cenozoic [31]. The concomitant increase in atmospheric O_2 (figure 1*a*,*b*) almost certainly contributed to the evolution of large animals, including placental mammals that have very high metabolic demands [29,32,33].

3. The rise of the diatoms

The composition of eukaryotic phytoplankton in the modern ocean is dominated by diatoms, dinoflagellates and coccolithophores [12]. Through photosynthesis and calcification these organisms make a small but significant contribution

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

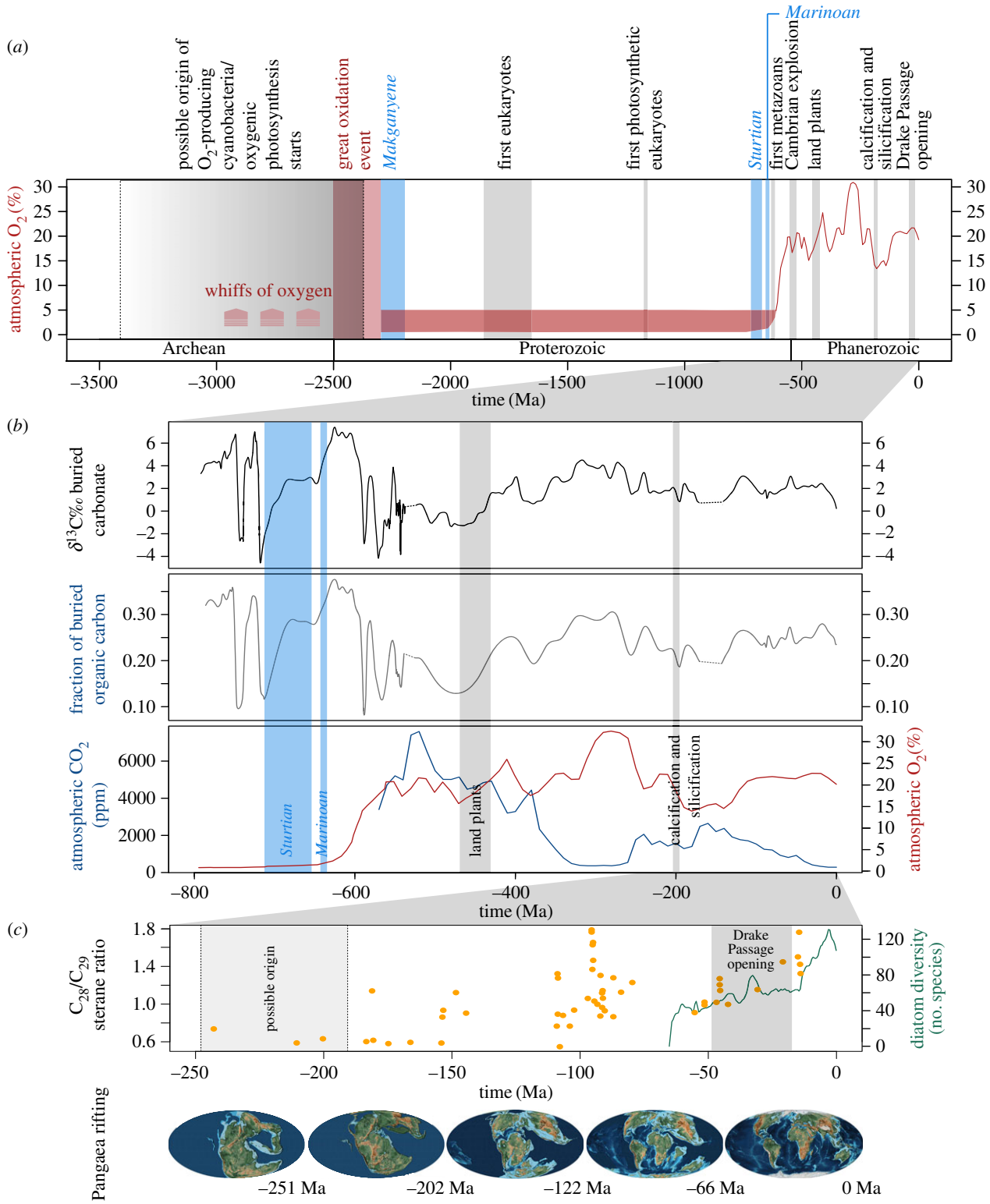


Figure 1. Major evolutionary and biogeochemical events during the history of life on Earth. (*a*) Trends since the evolution of oxygenic photosynthesis, (*b*) trends during the last 800 million years, and (*c*) diatom diversity and abundance data with respect to Pangaea rifting during the last 260 million years. Atmospheric O_2 was modified from Holland [7] according to Lyons *et al.* [8]; it is compared to $\delta^{13}C$ of carbonates [9], fraction of buried organic carbon [9], atmospheric CO_2 [10], diatom diversity [11] and C_{28}/C_{29} sterane ratios [12], which is a geochemical proxy for diatom abundance. Snowball Earth events are shown in light blue and were taken from Hoffman and Kopp *et al.* [13,14]. Pangaea rifting is illustrated with maps taken from the PALEOMAP Project [15]. The grey ranges on the plots represent the estimated span of the events cited in the text. Note that because the oldest direct measurements of atmospheric O_2 come from Pleistocene ice cores, all the detail in the Phanerozoic curve is based on models. Prior to that we have represented the views of Lyons *et al.* [8]: no stable O_2 trends before the Great Oxidation Event, some atmospheric O_2 (1–5%) through most of the Proterozoic, and then a rise to more or less modern values from the Ediacaran to the Silurian. The case is strong that pO_2 during the Carboniferous was higher than today's, but other details in the Phanerozoic curve are conjectural.

(probably around 10%) to the regulation of the partial pressure of carbon dioxide in the upper ocean [34,35]. The other 90% of oceanic carbon is derived from the physico-chemically regulated solubility of CO_2, which generates carbonate ions in the upper ocean [36]. The biological drawdown of atmospheric CO_2 through the activity of photosynthetic organisms in the ocean is known as the biological carbon pump which results in the generation of

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

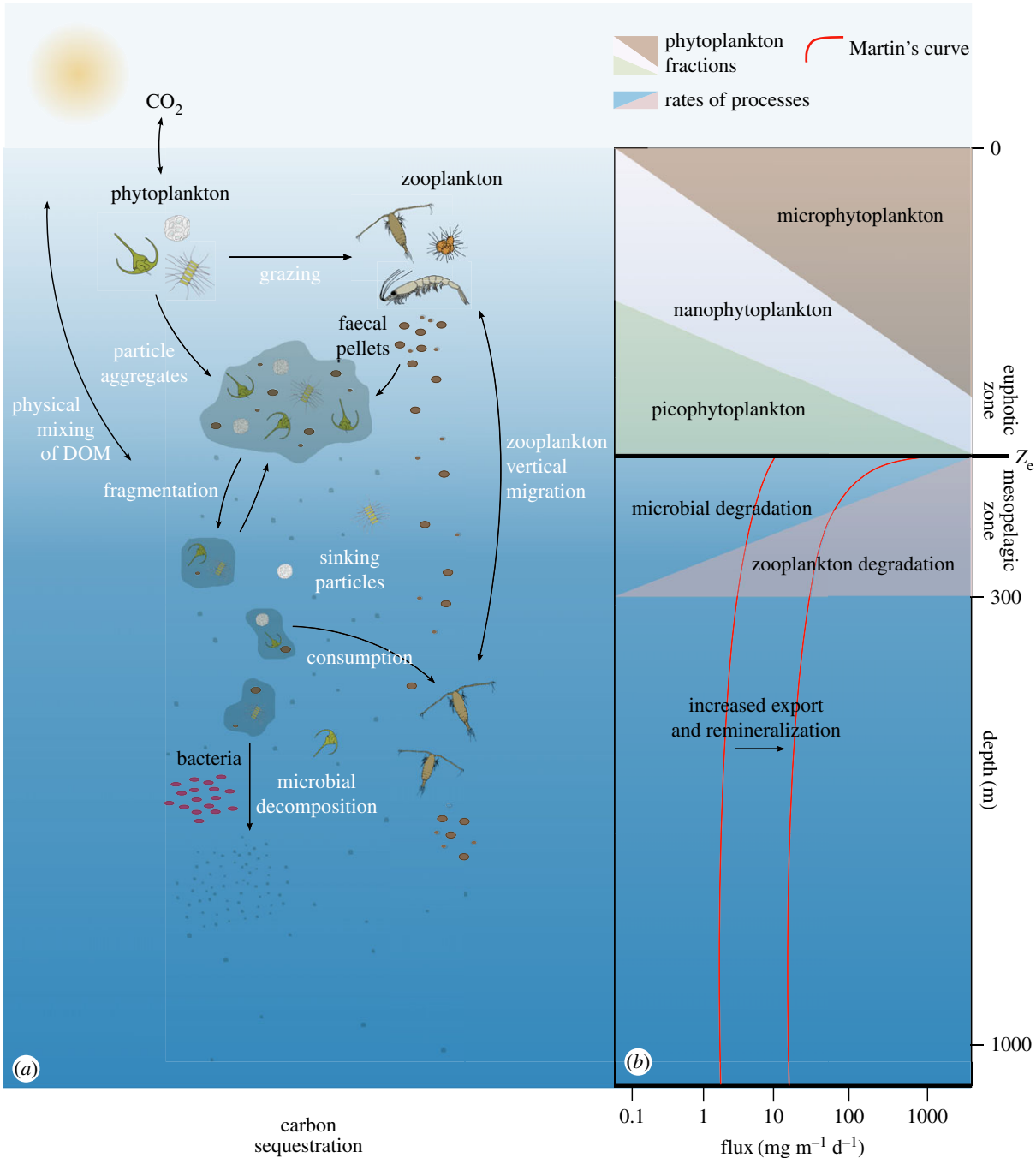


Figure 2. The biological carbon pump in the ocean. (*a*) Inorganic atmospheric carbon (CO_2) is transformed into organic carbon by phytoplankton in the euphotic zone. This carbon is then grazed on by heterotrophic organisms. A fraction of it is exported out of the surface layer as particulate organic carbon (such as dead organic material, faecal pellets produced by the zooplankton and aggregates of these materials) that sinks in the water column. Two other major processes help with the transfer of carbon below the surface layer: physical mixing of dissolved organic material (DOM) and transport by zooplankton vertical migration. (*b*) Different processes that can affect the decrease in the flux of particles in the ocean (adapted from [21]). The dimensions of the different areas represent the relative importance of phytoplankton fractions or rates of processes. The variation of the estimated flux with depth was modelled by fitting the Martin power relationship [22]. Carbon export is influenced by the phytoplankton composition in the euphotic zone: export is high when microphytoplankton (including diatoms) dominate the plankton community in the euphotic zone, while low export values correspond to systems dominated by picophytoplankton. Ze = depth of the euphotic zone. Note that depth, organisms and particle sizes are not to scale.

organic matter that can be consumed by other organisms, as well as calcium carbonate (figure 2). The biological carbon pump exports approximately 5–12 PgC yr^{-1} from the surface to the mesopelagic layer, from which approximately 0.2 PgC yr^{-1} is stored in sediment for millennia [34,35], thus contributing to the vertical gradient of carbon in the ocean. The process also results in biological feedback on atmospheric CO_2 and thus the Earth's climate [37,38]. This structuring of the carbon cycle in the ocean appears to have been established as the three phytoplankton groups rose to prominence in the Mesozoic Era, perhaps as a consequence of the availability of ecological space populated previously by taxa that did not survive the Permian–Triassic mass extinction event, which was Earth's most severe extinction event (resulting in the loss of around 96% of all marine species) [39].

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

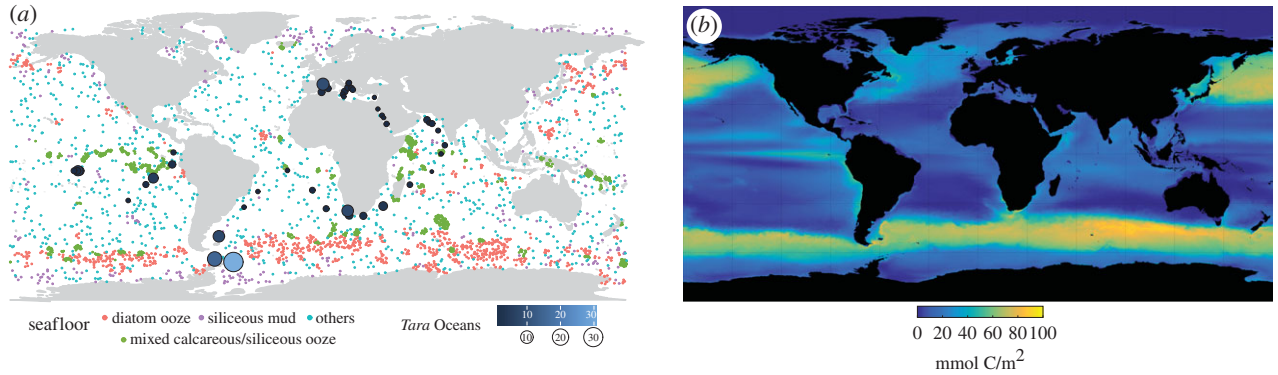


Figure 3. Extent of diatom-rich sediments compared with the distribution of modern diatoms in the ocean. Biosiliceous oozes are present in regions that, still today, are largely dominated by diatoms, in particular the Southern Ocean. (*a*) Small dots represent seafloor sediment samples defined as containing predominantly diatom ooze, siliceous mud, mixed calcareous/siliceous ooze, or others. Circles of varying size and blue colour correspond to diatom relative abundances determined by the *Tara* Oceans survey (modified from Dutkiewicz *et al.* and Malviya *et al.* [47,48]). (*b*) Water column inventory of diatom biomass (mmol C/m^2) from a biogeochemical/ecosystem simulation (modified version of Dutkiewicz *et al.* [49]).

The fossil record left behind by the elaborate siliceous shells of diatoms indicates that they remained minor components in the ocean until the Cretaceous [31,40], when the supercontinent Pangaea began to break apart into the continents we know today and the major ocean basins were formed (figure 1*c*). As well as creating more space in marine ecosystems, the rifting of Pangaea was accompanied by the delivery of more nutrients to the oceans because it was concomitant with continental elevation. The increase in nutrients favoured the selection of large-celled phytoplankton that lived along the continental margins such as diatoms [31,41–43]. Following the mass extinction event at the Cretaceous/Paleogene boundary (65 Ma), the diatoms continued to expand and further populate the oceans. In contrast to dinoflagellates and coccolithophores, diatom diversity continued to increase through the Cenozoic; in particular two pulses of diversification occurred at the Eocene/Oligocene boundary interval (33.9 Ma) and the middle to late Miocene (5–20 Ma) [44] (figure 1*c*). Environmental changes such as sea-level rise, silicate bioavailability, predation, ocean chemistry, increased latitudinal thermal gradients and circulation all likely played a role in driving such diversification [31,41,42]. As one case in point, correlations between increased diatom abundance and carbon export to the deep ocean with reductions in atmospheric CO_2 and reduced temperatures during the opening of the Drake Passage between 19 and 49 Ma suggest that the resulting Antarctic Circumpolar Current may have generated a highly favourable environment for diatom proliferation in the Southern Ocean, that today is still characterized by diatom-rich plankton communities [45,46].

Diatoms today are found throughout the world's oceans, wherever there is sufficient light and nutrients (figure 3). They typically dominate well-mixed coastal and upwelling regions, where the organic carbon they generate supports productive fisheries such as in the Peruvian and Benguela upwelling systems. They appear well adapted to surviving long periods of nutrient and light limitation and often dominate oceanic spring blooms because they can divide more rapidly than other phytoplankton when conditions become favourable for growth, at least as long as silicon is not limiting [50]. They also dominate at high latitudes and in polar environments, in particular along the sea-ice edge where other photosynthetic organisms are rare, making the Arctic and Southern Ocean ecosystems especially dependent on them [3,4] (figure 3). Their importance for the biogeochemistry of these regions over geological time periods is evidenced by the enormous deposits of siliceous mud and oozes more than 1 km thick in places [47] (figure 3*a*). The rise of diatoms in the last few millions of years is accompanied by the establishment of the main petroleum source rocks, derived from carbon export. The often spatial coincidence of silica and fossil fuels, together with the worldwide survey of biomarkers (such as 24-norcholestane or C_{28}-C_{29} steranes) in sediments and source rocks, indicate a crucial role of diatoms in the formation of today's reserves [44]. Moreover, several petroleum basins overlap with regions where diatoms thrive, such as oceanic coastal environments and the Arctic Ocean [51]. Although previous assessments suggest that petroleum source rocks are relatively low in abundance in the Southern Ocean [52], this region may hold significant resources as well.

4. Diatom evolution through the lens of genomics

While sedimentary rocks and the biomarkers within them provide a coarse-grained record of the intertwined histories of life, geology and climate, the evolutionary trajectories of different organisms can best be found by finding remnants of them in their genome sequences (for example, see [53]).

Already prior to the advent of genome sequencing, biochemical and ultrastructural data had provided persuasive evidence that diatoms were derived from a secondary endosymbiotic event involving a red alga that had occurred sometime between 1200 and 700 Ma (figures 1*a* and 4) and that was common to all stramenopiles, the phylum in which diatoms sit, as well as the chromalveolate supergroup of eukaryotes that includes dinoflagellates and coccolithophores [55–57]. The diatom genome sequences analysed to date do provide support for a red algal endosymbiont [58,59], but the abundance of genes apparently derived from a green algal source has led to the controversial hypothesis that a green algal endosymbiont preceded the red alga and that many of its genes were retained prior to the arrival

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

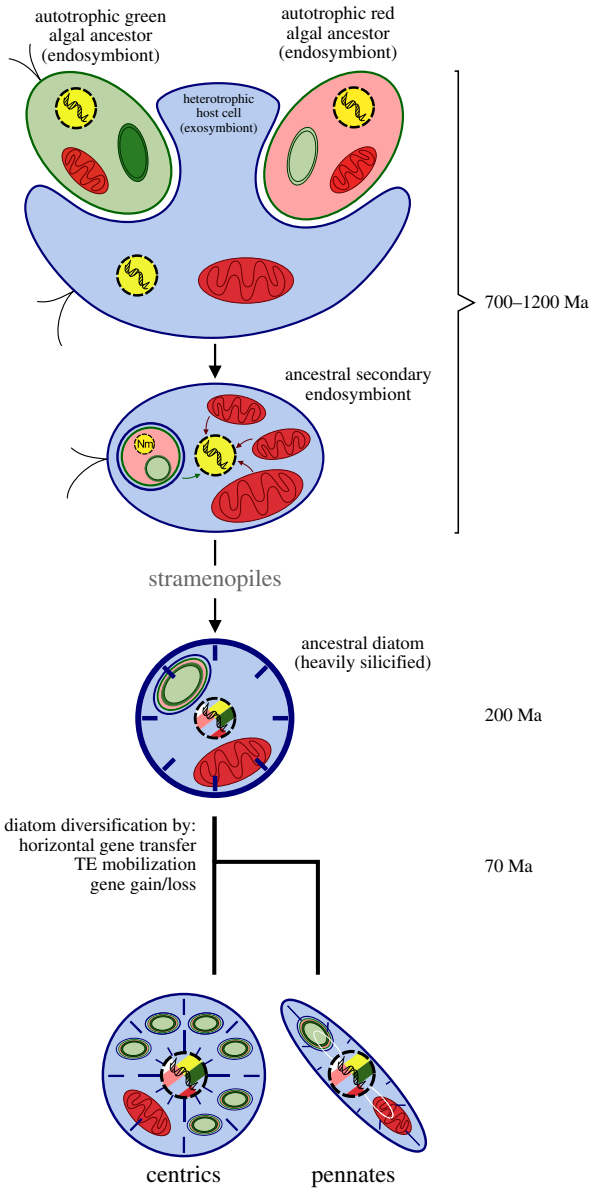


Figure 4. Diatom evolution through the lens of genomics. Major events during the evolution of diatoms from secondary endosymbionts are shown, together with approximate dates. TE, transposable element. The estimated time of separation between pennate and centric diatoms at around 70 million years (70 Ma) is based on Chacón-Baca *et al*. [54]. Figure modified from Bowler *et al*. [3].

of the red alga, whereas the red algal genes that were acquired later were not [60,61] (figure 4). In such a scenario, diatoms (and other photosynthetic chromalveolates) bear red algal-derived chloroplasts driven to a significant extent by green algal genes encoded in the nucleus, which may have provided a selective advantage in ocean environments and thus underlie why such organisms have come to dominate in the ocean whereas photosynthetic organisms harbouring green algal-derived plastids dominate terrestrial habitats [62,63].

An additional feature is the presence of several hundreds of bacterial genes scattered throughout diatom genomes [59], representing around 5% of total gene content. Many such genes appear to have ancient origins because they are shared among several diatoms, and encode functions essential for diatom biology [3]. Diatom-specific transposable elements additionally appear to have been instrumental in generating the rich diversity of species found today [3,64] (figure 4).

The chimeric nature of diatom genomes has brought together unique combinations of genes that collectively encode non-canonical pathways of nutrient assimilation and metabolite management, including for a urea cycle that is integral to nitrogen metabolism [65], and a novel configuration coupling photosynthesis and respiration between diatom chloroplasts and mitochondria [66]. The combined findings have profound and unanticipated implications for our understanding of the role of diatoms in biogeochemical cycles, and highlight the utility of genome sequences for revealing an organisms' metabolic potential. Diatom genomes have furthermore been found to encode large numbers of cyclins [67], key regulators of cell division, that may underlie their impressive proliferative capacity during oceanic blooms, as well as specialized stress-responsive light-harvesting chlorophyll-binding proteins that may be of particular importance for survival in polar-adapted diatoms [68,69].

The peculiarities of the diatom toolbox used to manage silicon metabolism and to generate their silicified cell walls are also being revealed (e.g. [70]), and it is emerging that such processes are deeply integrated within diatom primary metabolism, e.g. for the generation of frustule-localized long chain polyamines as offshoots of the urea cycle [65,71,72]. Notwithstanding, genomics has yet to reveal anything about what ecological or physiological advantages are associated with frustule biogenesis.

The extension of findings from genomics to natural environments will likely reveal further innovations [73]. Evidence is already emerging that some diatoms may have evolved permanent genome-level adaptations to certain conditions (e.g. related to iron bioavailability [74]) whereas others have retained the ability to acclimate to a wider range of conditions through more flexible responses at the transcriptional level [75]. The recent evaluation of the importance of epigenetic processes mediated at the level of DNA methylation or chromatin structural changes [76,77] will reveal whether diatoms have retained or acquired other features from their ancestors that permit additional opportunities for responding to a fluctuating environment over shorter timescales than are operative over macroevolutionary timescales [78].

5. Diatoms in the contemporary oceans

For decades, morphological studies have revealed diatoms to be one of the most ecologically important groups of phytoplankton in the modern oceans and one of the largest components of marine biomass [30,79,80]. More recent environmental omics studies have confirmed this. In particular, in the metabarcoding survey based on the V9 hypervariable region of 18S rDNA performed as part of the *Tara* Oceans global plankton sampling campaign, diatoms are the most abundant group of obligate photosynthetic eukaryotes and the fifth most abundant group of marine eukaryotes [48,81]. Moreover, in some Antarctic stations they represent more than 25% of the sequenced metabarcodes. Metabarcoding studies have allowed a refinement of the diversity estimation and the biogeographic distribution of diatoms even at the genus and species level. Meanwhile, metagenomics and metatranscriptomics data (unpublished results from the *Tara* Oceans consortium) will deepen our knowledge about the role of diatoms in the modern ocean.

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

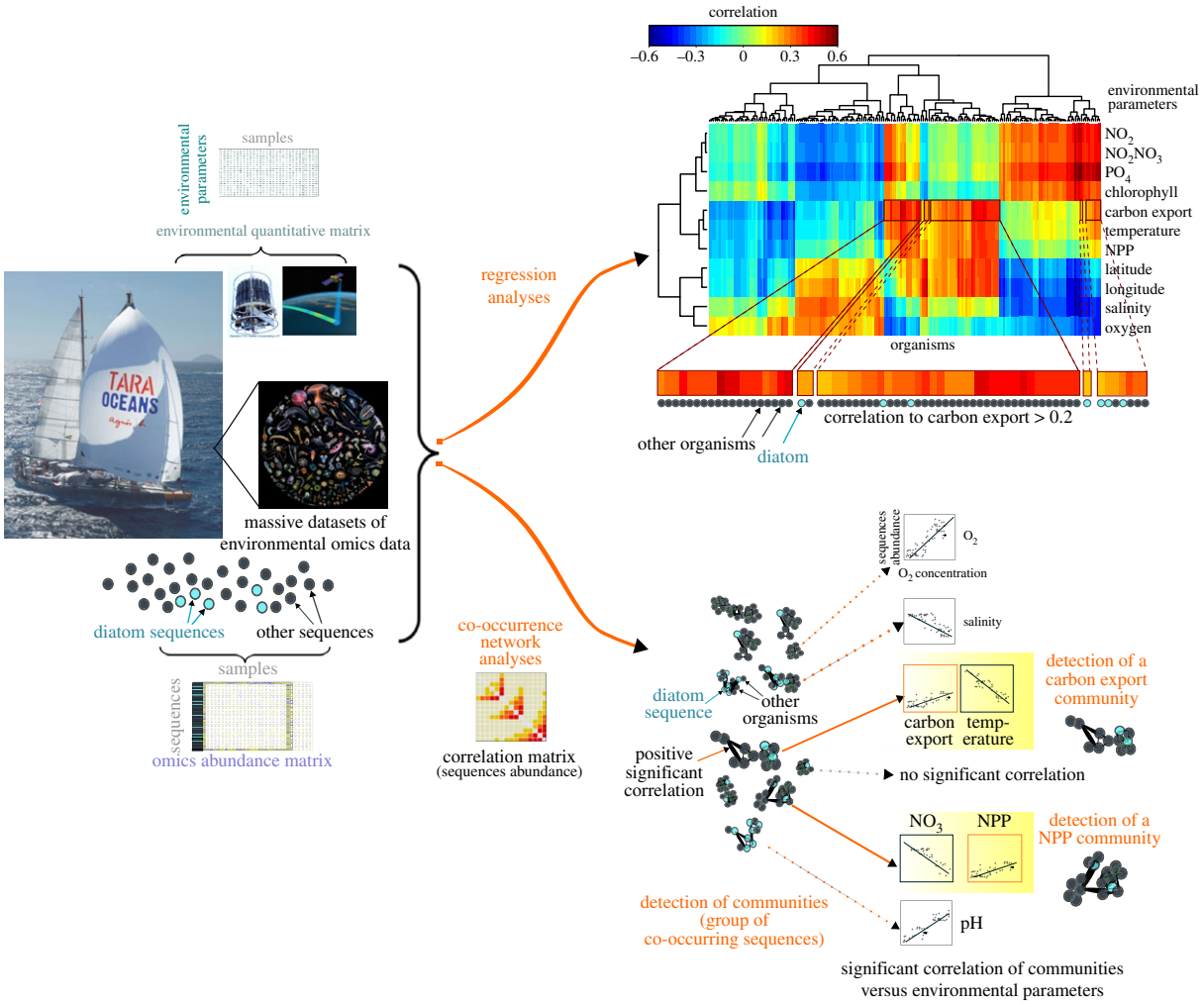


Figure 5. A new context for marine ecosystems biology. High-throughput sequencing technologies nowadays allow the production of massive omics datasets (meta-genomics, meta-transcriptomics, meta-barcoding datasets) from microbial planktonic communities. Such massive datasets can be analysed in parallel with large quantitative matrices of environmental data, and have notably produced a first global picture of microbial organisms correlating with carbon export in oligotrophic oceans [83]. In the framework of the biological carbon study, bioinformatics analyses thus help to establish on one hand the list of the most correlated lineages to NPP (Net Primary Production) or carbon export (e.g. using regression analyses such as sparse partial least square (sPLS) analysis; upper part of the figure), or to detect the communities linked to NPP and carbon export (e.g. using co-occurrence network analyses such as weighted correlation network analyses; lower part of the figure; see more details on methods in Guidi *et al.* [83]). Today, the roles of diatoms in the oceans can thus be considered in a global and integrative context.

In terms of their biogeochemical roles, diatoms are believed to be the principal contributors of primary production and carbon export among all photosynthetic organisms in the modern oceans, in particular because of their dominance in highly productive regions [1,5]. Estimates based on time-series of surface chlorophyll from the SeaWiFS Project indicate that microphytoplankton (mostly diatoms) may contribute up to 70% of the net primary production in coastal upwelling systems and 50% in temperate and sub-polar regions during the spring-summer seasons [82]. Overall, diatoms are estimated to contribute around 40% of the total primary production in the oceans, and therefore around one fifth of all the photosynthesis on Earth, similar to all terrestrial rainforests combined [1]. Similarly, both carbon export and remineralisation variations at global scale seem to be partially explained by the phytoplankton community where diatoms and their resting spores may play critical roles [21,83,84]. Diatoms are also a key component in the biogeochemical cycling of silicon (reviewed extensively in [85]).

The combination of genomics data collected during the *Tara* Oceans expedition with ancillary environmental data allows a new framework, summarized in figure 5, to pinpoint the importance of individual planktonic groups in specific processes, in a holistic context of the entire plankton communities that they are part of. Such network-based methods have already been used to disentangle the key players in euphotic zone communities related to carbon export to deeper layers in the oligotrophic ocean [83] (figure 5). Regression-based analyses on the entire eukaryotic metabarcoding dataset currently available from *Tara* Oceans [81] reveal the dominant roles of diatoms in contributing to net primary production and carbon export, in particular in areas characterized by low temperatures, high oxygen and nutrient concentrations (figure 6). It should therefore be possible to test the robustness of these results by bioinformatic analysis and to further disentangle the roles of diatoms in marine ecosystems using more extended datasets. Such studies could also be performed in the context of different climate simulations to better understand how diatoms affect the carbon cycle and climate regulation.

6. Perspectives

Incontrovertible evidence shows that the Earth's climate has begun to change markedly over the last decades as a

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

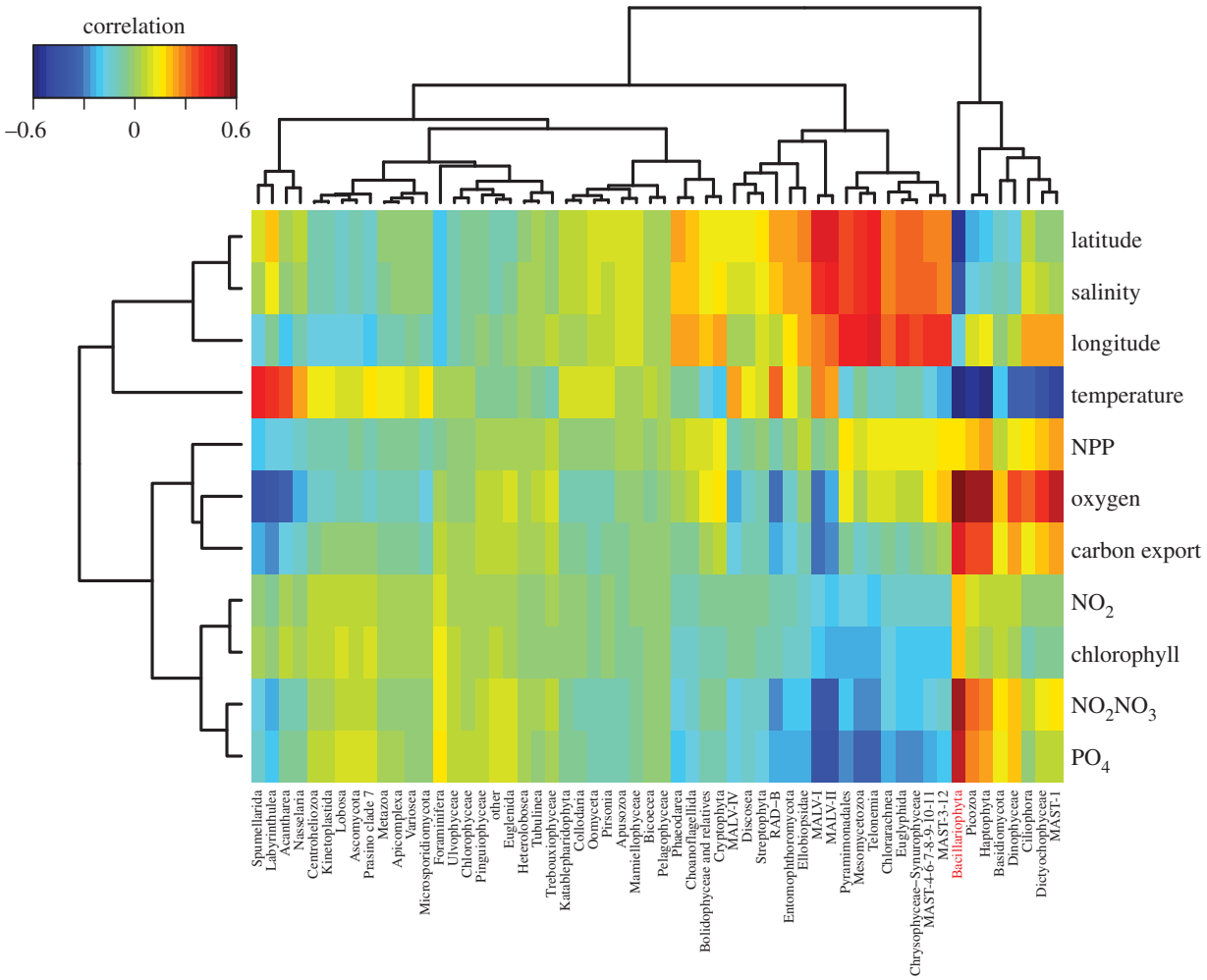


Figure 6. Diatoms and their role in the biological carbon pump as revealed by high-throughput DNA sequence-based datasets. Eukaryotic lineages associated to environmental parameters assessed by standard methods for regression-based modelling (sPLS analysis). Correlations between lineages and environmental parameters are depicted as a clustered heat map. This plot has been created using the *Tara* Oceans metabarcoding dataset (providing an abundance matrix based on 18s rDNA ribotypes (V9 region) from oligotrophic stations as well as a few Southern Ocean stations [48,81] and the associated environmental parameters [86]. With respect to other eukaryotic lineages, diatoms (*Bacillariophyta*) show significant correlations with NPP and chlorophyll, and the highest positive correlation to carbon flux (more than 0.46), supporting the hypothesis that diatoms play a major role in the biological carbon pump at a global scale.

consequence of CO_2 released into the atmosphere from the burning of fossil fuels. The overprint of human activities on Earth's biogeochemical cycles is evident from the simple fact that we are currently burning the equivalent of around 1 million years of buried carbon derived from diatoms and other plankton each year [2]. While we can be confident that the oceans will continue for some time to be the major sink absorbing excess heat and CO_2, and will consequently warm, acidify, and deoxygenate in the coming centuries [87,88], we have no consistent view about how the life support system of the oceans, the plankton, will fare. Regarding diatoms, we can expect shifts in several aspects of diatom diversity and biogeography, which could not only affect biogeochemical cycles but may also pose a challenge for the functioning of marine food webs, in which diatoms are intensely grazed. Given the rise of diatoms to global importance in marine ecosystems over the last tens of millions of years it is crucial that future research addresses their capabilities to adapt to changing environments, both by investigation of the geological record and by the exploration of diatom genomes.

Authors' contributions. C.B. conceived and designed the manuscript, and all authors contributed significantly to the writing and the creation of the figures. A.-S.B., F.M.I. and C.B. worked on the sections of evolutionary and biogeochemical events and diatom genomics. A.-S.B., L.B. and L.G. presented the biological carbon pump. F.M.I., O.J. and S.D. compared the marine sediments with modern distributions. L.B. and L.G. were in charge of the eco-systems biology and high-throughput sequencing sections.

Competing interests. We have no competing interests.

Funding. C.B. acknowledges funding from the ERC Advanced Award 'Diatomite', the Louis D Foundation, the Gordon and Betty Moore Foundation, and the French Government 'Investissements d'Avenir' programmes MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-1253 11-IDEX-0001-02), and OCEANOMICS (ANR-11-BTBR-0008). C.B. also thanks the Radcliffe Institute of Advanced Study at Harvard University for a scholars fellowship during the 2016-2017 academic year. A.-S.B. is funded by the PhD programme from the Ecole Doctorale Complexité du Vivant. L.B. and L.G. acknowledge funding from the French national programme EC2CO-LEFE (FunOmics project). S.D. and O.J. acknowledge funding from the Gordon and Betty Moore Foundation, and from NASA (NNX16AR47G).

Acknowledgements. The authors thank Andrew H. Knoll for his valuable contributions to this manuscript. This article is contribution number 55 of *Tara* Oceans.

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

References

1. Field CB, Behrenfeld MJ, Randerson JT, Falkowski PG. 1998 Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240. (doi:10.1126/science.281.5374.237)
2. Falkowski PG. 2015 *Life's engines: how microbes made Earth habitable*. Princeton, NJ: Princeton University Press.
3. Bowler C, Vardi A, Allen AE. 2010 Oceanographic and biogeochemical insights from diatom genomes. *Ann. Rev. Mar. Sci.* **2**, 333–365. (doi:10.1146/annurev-marine-120308-081051)
4. Armbrust EV. 2009 The life of diatoms in the world's oceans. *Nature* **459**, 185–192. (doi:10.1038/nature08057)
5. Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B. 1995 Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochem. Cycles* **9**, 359–372. (doi:10.1029/95GB01070)
6. Fischer WW, Hemp J, Johnson JE. 2016 Evolution of oxygenic photosynthesis. *Annu. Rev. Earth Planet. Sci.* **44**, 647–683. (doi:10.1146/annurev-earth-060313-054810)
7. Holland HD. 2006 The oxygenation of the atmosphere and oceans. *Phil. Trans. R. Soc. B* **361**, 903–915. (doi:10.1098/rstb.2006.1838)
8. Lyons TW, Reinhard CT, Planavsky NJ. 2014 The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315. (doi:10.1038/nature13068)
9. Hayes JM, Strauss H, Kaufman AJ. 1999 The abundance of ^{13}C in marine organic matter and isotopic fractionation in the global biogeochemical cycle of carbon during the past 800 Ma. *Chem. Geol.* **161**, 103–125. (doi:10.1016/S0009-2541(99)00083-2)
10. Berner RA, Kothavala Z. 2001 GEOCARB III: a revised model of atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **301**, 182–204. (doi:10.2475/ajs.301.2.182)
11. Lazarus D, Barron J, Renaudie J, Diver P, Türke A. 2014 Cenozoic planktonic marine diatom diversity and correlation to climate change. *PLoS ONE* **9**, e84857. (doi:10.1371/journal.pone.0084857)
12. Knoll AH, Summons RE, Waldbauer JR, Zumberge JE. 2007 The geological succession of primary producers in the oceans. In *Evolution of primary producers in the sea* (eds PG Falkowski, AH Knoll), pp. 133–163. Burlington, MA: Elsevier.
13. Hoffman PF. 2016 Cryoconite pans on Snowball Earth: supraglacial oases for Cryogenian eukaryotes? *Geobiology* **14**, 531–542. (doi:10.1111/gbi.12191)
14. Kopp RE, Kirschvink JL, Hilburn IA, Nash CZ. 2005 The Paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc. Natl Acad. Sci. USA* **102**, 11 131–11 136. (doi:10.1073/pnas.0504878102)
15. Scotese CR. 2014 PALEOMAP Atlas for ArcGIS, PALEOMAP Project. See http://www.scotese.com/.
16. Nelson N, Ben-Shem A. 2004 The complex architecture of oxygenic photosynthesis. *Nat. Rev. Mol. Cell Biol.* **5**, 971–982. (doi:10.1038/nrm1525)
17. Goldsmith PF *et al.* 2011 HERSCHEL measurements of molecular oxygen in Orion. *Astrophys. J.* **737**, 96. (doi:10.1088/0004-637X/737/2/96)
18. Martin W, Russell MJ. 2003 On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Phil. Trans. R. Soc. Lond. B* **358**, 59–85. (doi:10.1098/rstb.2002.1183)
19. Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006 Eukaryotic organisms in Proterozoic oceans. *Phil. Trans. R. Soc. B* **361**, 1023–1038. (doi:10.1098/rstb.2006.1843)
20. Reyes-Prieto A, Weber APM, Bhattacharya D. 2007 The origin and establishment of the plastid in algae and plants. *Annu. Rev. Genet.* **41**, 147–168. (doi:10.1146/annurev.genet.41.110306.130134)
21. Guidi L, Stemmann L, Jackson GA, Ibanez F, Claustre H, Legendre L, Picheral M, Gorsky G. 2009 Effects of phytoplankton community on production, size, and export of large aggregates: a world-ocean analysis. *Limnol. Oceanogr.* **54**, 1951–1963. (doi:10.4319/lo.2009.54.6.1951)
22. Martin JH, Knauer GA, Karl DM, Broenkow WW. 1987 VERTEX: carbon cycling in the northeast Pacific. *Deep Sea Res. Part A Oceanogr. Res. Pap.* **34**, 267–285. (doi:10.1016/0198-0149(87)90086-0)
23. Dahl TW *et al.* 2010 Devonian rise in atmospheric oxygen correlated to the radiations of terrestrial plants and large predatory fish. *Proc. Natl Acad. Sci. USA* **107**, 17 911–17 915. (doi:10.1073/pnas.1011287107)
24. Plattner G-K, Joos F, Stocker TF. 2002 Revision of the global carbon budget due to changing air-sea oxygen fluxes. *Global Biogeochem. Cycles* **16**, 12–43. (doi:10.1029/2001GB001746)
25. Keeling RF, Shertz SR. 1992 Seasonal and interannual variations in atmospheric oxygen and implications for the global carbon cycle. *Nature* **358**, 723–727. (doi:10.1038/358723a0)
26. Keeling RF, Garcia HE. 2002 The change in oceanic O_2 inventory associated with recent global warming. *Proc. Natl Acad. Sci. USA* **99**, 7848–7853. (doi:10.1073/pnas.122154899)
27. Miller KG *et al.* 2005 The Phanerozoic record of global sea-level change. *Science* **310**, 1293–1298. (doi:10.1126/science.1116412)
28. Falkowski PG, Fenchel T, Delong EF. 2008 The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039. (doi:10.1126/science.1153213)
29. Falkowski PG, Katz ME, Milligan AJ, Fennel K, Cramer BS, Aubry MP, Berner RA, Novacek MJ, Zapol WM. 2005 The rise of oxygen over the past 205 million years and the evolution of large placental mammals. *Science* **309**, 2202–2204. (doi:10.1126/science.1116047)
30. Smetacek V. 1999 Diatoms and the ocean carbon cycle. *Protist* **150**, 25–32. (doi:10.1016/S1434-4610(99)70006-4)
31. Katz ME, Finkel ZV, Grzebyk D, Knoll AH, Falkowski PG. 2004 Evolutionary trajectories and biogeochemical impacts of marine eukaryotic phytoplankton. *Annu. Rev. Ecol. Evol. Syst.* **35**, 523–556. (doi:10.1146/annurev.ecolsys.35.112202.130137)
32. Knoll AH, Carroll SB. 1999 Early animal evolution: emerging views from comparative biology and geology. *Science* **284**, 2129–2137. (doi:10.1126/science.284.5423.2129)
33. Sperling EA, Frieder CA, Raman AV, Girguis PR, Levin LA, Knoll AH. 2013 Oxygen, ecology, and the Cambrian radiation of animals. *Proc. Natl Acad. Sci. USA* **110**, 13 446–13 451. (doi:10.1073/pnas.1312778110)
34. Bopp L, Bowler C, Guidi L, Karsenti E, de Vargas C. 2015 The ocean: a carbon pump. See http://www.ocean-climate.org.
35. Ciais P *et al.* 2013 Carbon and other biogeochemical cycles. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds TF Stocker *et al.*), pp. 465–570. Cambridge, UK: Cambridge University Press.
36. Siegenthaler U, Sarmiento JL. 1993 Atmospheric carbon dioxide and the ocean. *Nature* **365**, 119–125. (doi:10.1038/365119a0)
37. Volk T, Hoffert M. 1985 Ocean carbon pumps: analysis of relative strengths and efficiencies in ocean-driven atmospheric CO_2 changes. In *The carbon cycle and atmospheric CO_2: natural variations Archean to Present* (eds ET Sundquist, WS Broecker), pp. 99–110. American Geophysical Union.
38. Frankignoulle M, Canon C, Gattuso J-P. 1994 Marine calcification as a source of carbon dioxide: positive feedback of increasing atmospheric CO_2. *Limnol. Oceanogr.* **39**, 458–462. (doi:10.4319/lo.1994.39.2.0458)
39. Benton M. 2005 *When life nearly died: the greatest mass extinction of all time*. London, UK: Thames & Hudson.
40. Sims PA, Mann DG, Medlin LK. 2006 Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia* **45**, 361–402. (doi:10.2216/05-22.1)
41. Kooistra WHCF, Gersonde R, Medlin LK, Mann DG. 2007 The origin and evolution of the diatoms: their adaptation to a planktonic existence. In *Evolution of primary producers in the sea* (eds PG Falkowski, AH Knoll), pp. 207–249. Burlington, MA: Elsevier.
42. Katz ME, Fennel K, Falkowski PG. 2007 Geochemical and biological consequences of phytoplankton evolution. In *Evolution of primary producers in the sea* (eds PG Falkowski, AH Knoll), pp. 405–430. Burlington, MA: Elsevier.
43. Cermeño P, Falkowski PG, Romero OE, Schaller MF, Vallina SM. 2015 Continental erosion and the Cenozoic rise of marine diatoms. *Proc. Natl Acad. Sci. USA* **112**, 4239–4244. (doi:10.1073/pnas.1412883112)

Downloaded from http://rstb.royalsocietypublishing.org/ on July 17, 2017

44. Cermeño P. 2016 The geological story of marine diatoms and the last generation of fossil fuels. *Perspect. Phycol.* **3**, 53–60. (doi:10.1127/pip/2016/0050)
45. Siegenthaler U *et al.* 2005 Stable carbon cycle—climate relationship during the late Pleistocene. *Science* **2**, 1313–1317. (doi:10.1126/science.1120130)
46. Rabosky DL, Sorhannus U. 2009 Diversity dynamics of marine planktonic diatoms across the Cenozoic. *Nature* **457**, 183–186. (doi:10.1038/nature07435)
47. Dutkiewicz A, Müller RD, O'Callaghan S, Jónasson H. 2015 Census of seafloor sediments in the world's ocean. *Geology* **43**, 795–798. (doi:10.1130/G36883.1)
48. Malviya S *et al.* 2016 Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl Acad. Sci. USA* **113**, E1516–E1525. (doi:10.1073/pnas.1509523113)
49. Dutkiewicz S, Hickman AE, Jahn O, Gregg WW, Mouw CB, Follows MJ. 2015 Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model. *Biogeosciences* **12**, 4447–4481. (doi:10.5194/bg-12-4447-2015)
50. Winder M, Cloern JE. 2010 The annual cycles of phytoplankton biomass. *Phil. Trans. R. Soc. B* **365**, 3215–3226. (doi:10.1098/rstb.2010.0125)
51. Gautier DL *et al.* 2009 Assessment of undiscovered oil and gas in the Arctic. *Science* **324**, 1175–1179. (doi:10.1126/science.1169467)
52. Masters CD, Root DH, Dietzman WD. 1983 *Distribution and quantitative assessment of world crude oil reserves and resources*. Open-File Report 83-728, US Geological Survey, http://pubs.er.usgs.gov/publication/ofr83728.
53. Edwards D, Kenrick P. 2015 The early evolution of land plants, from fossils to genomics: a commentary on Lang (1937) 'On the plant-remains from the Downtonian of England and Wales'. *Phil. Trans. R. Soc. B* **370**, 20140343. (doi:10.1098/rstb.2014.0343)
54. Chacón-Baca E, Beraldi-Campesi H, Cevallos-Ferriz SRS, Knoll AH, Golubic S. 2002 70 Ma nonmarine diatoms from northern Mexico. *Geology* **30**, 279–281. (doi:10.1130/0091-7613)
55. Gibbs SP. 1981 The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Ann. N Y Acad. Sci.* **361**, 193–208. (doi:10.1111/j.1749-6632.1981.tb54365.x)
56. Parker MS, Mock T, Armbrust EV. 2008 Genomic insights into marine microalgae. *Annu. Rev. Genet.* **42**, 619–645. (doi:10.1146/annurev.genet.42.110807.091417)
57. Cavalier-Smith T. 1981 Eukaryote kingdoms: seven or nine? *Biosystems* **14**, 461–481. (doi:10.1016/0303-2647(81)90050-2)
58. Armbrust EV *et al.* 2004 The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86. (doi:10.1126/science.1101156)
59. Bowler C *et al.* 2008 The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244. (doi:10.1038/nature07410)
60. Moustafa A, Beszteri BB, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009 Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**, 1724–1726. (doi:10.1126/science.1172983)
61. Deschamps P, Moreira D. 2012 Reevaluating the green contribution to diatom genomes. *Genome Biol. Evol.* **4**, 683–688. (doi:10.1093/gbe/evs053)
62. Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJR. 2004 The evolution of modern eukaryotic phytoplankton. *Science* **305**, 354–360. (doi:10.1126/science.1095964)
63. Falkowski PG, Oliver MJ. 2007 Mix and match: how climate selects phytoplankton. *Nat. Rev. Microbiol.* **5**, 813–819. (doi:10.1038/nrmicro1792)
64. Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, Grandbastien M-A, Bowler C. 2009 Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* **10**, 624. (doi:10.1186/1471-2164-10-624)
65. Allen AE *et al.* 2011 Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203–207. (doi:10.1038/nature10074)
66. Bailleul B *et al.* 2015 Energetic coupling between plastids and mitochondria drives CO_2 assimilation in diatoms. *Nature* **524**, 366–369. (doi:10.1038/nature14599)
67. Huysman MJJ *et al.* 2010 Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biol.* **11**, R17. (doi:10.1186/gb-2010-11-2-r17)
68. Bailleul B, Rogato A, de Martino A, Coesel S, Cardol P, Bowler C, Falciatore A, Finazzi G. 2010 An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. *Proc. Natl Acad. Sci. USA* **107**, 18 214–18 219. (doi:10.1073/pnas.1007703107)
69. Mock T *et al.* 2017 Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* **541**, 536–540. (doi:10.1038/nature20803)
70. Kotzsch A, Pawolski D, Milentyev A, Shevchenko A, Scheffel A, Poulsen N, Shevchenko A, Kröger N. 2016 Biochemical composition and assembly of biosilica-associated insoluble organic matrices from the diatom *Thalassiosira pseudonana*. *J. Biol. Chem.* **291**, 4982–4997. (doi:10.1074/jbc.M115.706440)
71. Kröger N, Poulsen N. 2008 Diatoms—from cell wall biogenesis to nanotechnology. *Annu. Rev. Genet.* **42**, 83–107. (doi:10.1146/annurev.genet.41.110306.130109)
72. Prihoda J, Tanaka A, de Paula WBM, Allen JF, Tirichine L, Bowler C. 2012 Chloroplast-mitochondria cross-talk in diatoms. *J. Exp. Bot.* **63**, 1543–1557. (doi:10.1093/jxb/err441)
73. Caron DA *et al.* 2016 Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat. Rev. Microbiol.* **15**, 6–20. (doi:10.1038/nrmicro.2016.160)
74. Peers G, Price NM. 2006 Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature* **441**, 341–344. (doi:10.1038/nature04630)
75. Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, Morales R, Allen AE, Armbrust EV. 2012 Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc. Natl Acad. Sci. USA* **109**, E317–E325. (doi:10.1073/pnas.1118408109)
76. Veluchamy A *et al.* 2013 Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nat. Commun.* **4**, 2091. (doi:10.1038/ncomms3091)
77. Veluchamy A *et al.* 2015 An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum*. *Genome Biol.* **16**, 102. (doi:10.1186/s13059-015-0671-8)
78. Tirichine L, Bowler C. 2011 Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *Plant J.* **66**, 45–57. (doi:10.1111/j.1365-313X.2011.04540.x)
79. Leblanc K *et al.* 2012 A global diatom database—abundance, biovolume and biomass in the world ocean. *Earth Syst. Sci. Data Discuss.* **5**, 147–185. (doi:10.5194/essdd-5-147-2012)
80. Brun P, Vogt M, Payne MR, Gruber N, O'Brien CJ, Buitenhuis ET, Le Quéré C, Leblanc K, Luo Y-W. 2015 Ecological niches of open ocean phytoplankton taxa. *Limnol. Oceanogr.* **60**, 1020–1038. (doi:10.1002/lno.10074)
81. de Vargas C *et al.* 2015 Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605. (doi:10.1126/science.1261605)
82. Uitz J, Claustre H, Gentili B, Stramski D. 2010 Phytoplankton class-specific primary production in the world's oceans: seasonal and interannual variability from satellite observations. *Global Biogeochem. Cycles* **24**, GB3016. (doi:10.1029/2009GB003680)
83. Guidi L *et al.* 2016 Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470. (doi:10.1038/nature16942)
84. Rembauville M, Manno C, Tarling GA, Blain S, Salter I. 2016 Strong contribution of diatom resting spores to deep-sea carbon transfer in naturally iron-fertilized waters downstream of South Georgia. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **115**, 22–35. (doi:10.1016/j.dsr.2016.05.002)
85. Tréguer PJ, De La Rocha CL. 2013 The world ocean silica cycle. *Ann. Rev. Mar. Sci.* **5**, 477–501. (doi:10.1146/annurev-marine-121211-172346)
86. Pesant S *et al.* 2015 Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023. (doi:10.1038/sdata.2015.23)
87. Pörtner H-O, Karl DM, Boyd PW, Cheung WWL, Lluch-Cota SE, Nojiri Y, Schmidt DN, Zavialov PO. 2014 Ocean systems. In *Climate change 2014: impacts, adaptation, and vulnerability. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds CB Field *et al.*), pp. 411–484. Cambridge, UK: Cambridge University Press.
88. Allison EH, Bassett HR. 2015 Climate change in the oceans: human impacts and responses. *Science* **350**, 778–782. (doi:10.1126/science.aac8721)

Chapter 3

Revisiting the study of the biological carbon pump through the use of microbial association networks

The goal of this chapter is to investigate with new eyes the biological carbon pump by defining biogeochemical states from net primary production, carbon export and flux attenuation estimates, and building microbial association networks for each of these states.

The chapter consists in a presentation of the usual steps to infer microbial networks and the potentially relevant network properties for the analysis of microbial association networks.

This introduction is followed by a draft manuscript of a work in collaboration with my supervisors (Lucie Bittner and Lionel Guidi) and colleagues from the Laboratoire des Sciences du Numérique de Nantes (LS2N), Damien Eveillard, Samuel Chaffron and Erwan Delage. The study consisted in (1) delineating biogeochemical states of the biological carbon pump, (2) analyzing their taxonomic composition, (3) inferring microbial association networks for these states and (4) comparing the networks properties. The main body of the article is supplemented by additional figures and tables.

3.1 Introduction to microbial association networks

Biodiversity research often focused on species richness and neglected interactions or assumed that they were homogeneously distributed (Bascompte, 2009). However, from the macro to the microbial world, organisms are engaged in various types of relationships. Therefore, it becomes necessary to take into account the patterns of interactions between them rather than the list of species composing the community (Bascompte, 2009). As

presented in chapter 1, microorganisms are of great importance for biogeochemical cycles (Falkowski et al., 2008) which require the cooperation of billions of them, forming a highly complicated and intricate system. Despite this importance, microorganisms and their interactions have often been overlooked, mainly because of the difficulty to identify them (between 85 and 99% of bacteria and archaea cannot be cultured in the lab, Vacher et al., 2016). Microbial interactions have long been identified using co-culture experiments (e.g. Long and Azam, 2001; Long et al., 2005, 2013; Sher et al., 2011; Biller et al., 2014), which are time-consuming and tedious tasks. Moreover, they introduce biases inherent in laboratory techniques (i.e. number of actors, representativeness of the natural environment), making difficult the extrapolation of the results to natural conditions. For these reasons, many scientists have chosen to use environmental DNA to identify microbial lineages and their interactions (although other techniques may reveal *in situ* associations, e.g. Lepère et al., 2016). The advent of high-throughput sequencing and associated bioinformatic pipelines has revolutionized the identification of microbial communities in the last 10 years (e.g. Bik et al., 2012; Zinger et al., 2012), providing a comprehensive view of environmental communities. Using meta-omic data (any kind of sequences present in an environment), it is now possible to infer co-occurrence networks using statistical tools (Faust and Raes, 2012), allowing to detect *in situ* associations between micro-organisms (e.g. Chaffron et al., 2010; Lima-Mendez et al., 2015). These newly produced networks are mined to rediscover already known associations (which help to confirm the validity of the approach), but are mainly composed of yet unknown relations.

While the study of microbial associations based on these networks is rapidly spreading (Röttjers and Faust, 2018), few examples of studies linking biogeochemical cycles to microbial association networks are available to date (e.g. Guidi et al., 2016; Mandakovic et al., 2018) and they are only developing concerning the ocean carbon cycle and particularly the biological carbon pump. In 2016, Guidi et al. used environmental and metagenomic data gathered during the *Tara* Oceans expeditions to delineate specific plankton networks (prokaryotic, eukaryotic and viral) correlated to carbon export in the oligotrophic ocean. In particular, their analyses allowed them to identify unexpected strongly associated lineages such as Radiolaria, alveolate parasites and the cyanobacteria *Synechococcus* and its phages, suggesting that network approaches could effectively help improve our knowledge on the carbon cycle through the use of microbial association networks. In this chapter, we propose to improve these results by integrating the other components of the biological carbon pump (i.e. net primary production and flux attenuation) and inferring microbial association networks corresponding to biogeochemical states that we defined. The results are presented in this chapter in the form of a draft manuscript preceded by

an introduction to microbial association networks.

3.1.1 Application of graph theory to ecological networks

Graph theory appears to have its origin in the city of Königsberg, then the capital of western Prussia. In 1736, the Swiss mathematician Leonard Euler gave the first representation of a graph that modelled the islands and bridges of the city (Euler, 1736). Since then, many disciplines, from technological sciences to sociology and biology, modelled systems with networks.

The potential first graphical representation of an ecological network is attributed to Lorenzo Camerano (Camerano, 1880). He represented a food web with groups of species linked by feeding relations (figure 3.1; Cohen, 1994). Two central ideas emerge from Camerano's 1880 essay. First, the equilibrium in the species abundances is maintained by feeding relations, and second, if a perturbation (i.e. a change in the abundance of one of the components of the community) occurs in a natural community, it propagates along the food chain. These ideas and the graphical representation of food webs was visionary, but it probably stayed unnoticed for some time as it doesn't resemble any known representation from later zoologists (Egerton, 2007). After Camerano, many food web diagrams were published (Pierce et al., 1912; Shelford, 1913; Petersen, 1915), until Elton generalized these diagrams and coined the terms *food chain* and *food cycle* (Summerhayes and Elton, 1923; Elton, 1927). Following Elton, numerous hypothetical and empirical trophic networks were described. However, quantitative, comparative research on potential generalities in the network structure appeared only in the 1970's (Dunne, 2009).

3.1.1.1 From abundance data to microbial association networks

Another type of ecological network, that may be linked to inter-species trophic relations (Morales-Castilla et al., 2015), is based on the species co-occurrence. The co-occurrence of two species is their simultaneous presence at the same place. Whereas trophic networks focus on trophic relations between species, co-occurrence networks aims at translating a wider range of interactions. These interactions can be classified based on the positive, negative or neutral outcome for the species involved (figure 3.2). Positive relationships for both partners are known as mutualism. Examples from the microbial world include syntrophy (i.e. cross-feeding) where two species are dependant on the metabolites produced by the other (Morris et al., 2013) and cooperation in bacterial biofilms (Nadell et al., 2009). Commensalism occur when one partner benefits while the other is unaffected. In the mi-

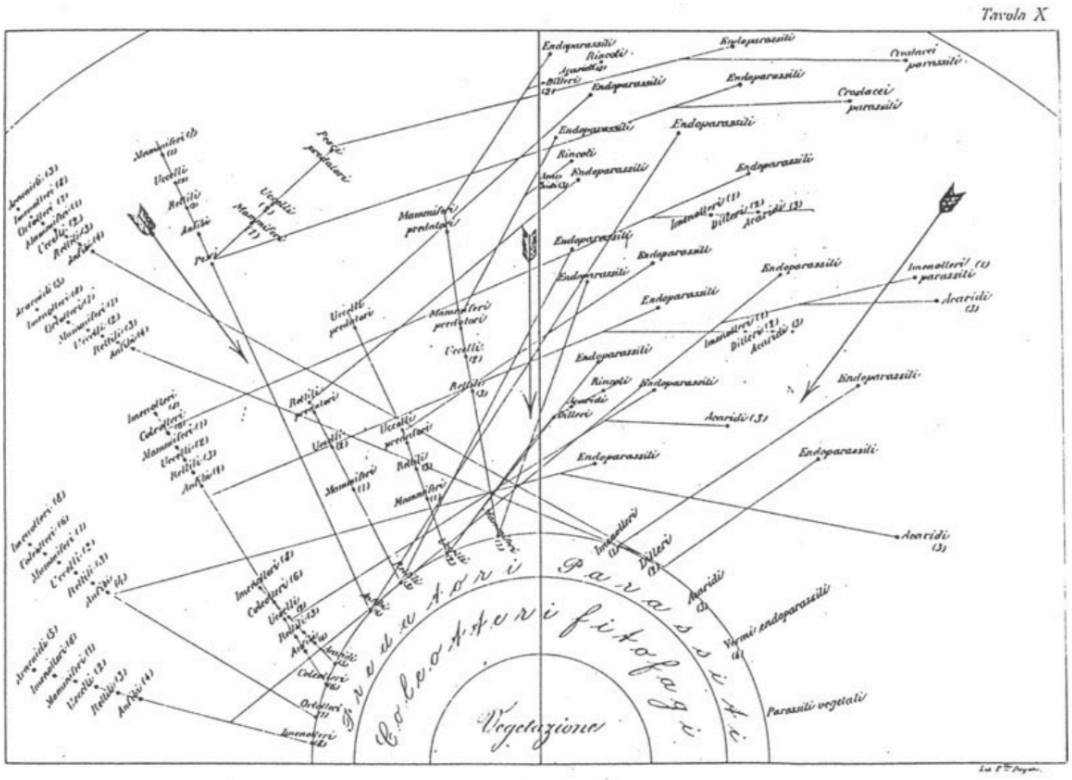


Figure 3.1 - First representation of a trophic network by Lorenzo Camerano, illustrating the "enemies of phytophagous Coleoptera and the enemies of those enemies" (Camerano, 1880).

crobial world, commensalism occurs when a species produces metabolites that are used by other species, with no gain for the first species (e.g. between denitrifiers and annamox bacteria). When one of the partners benefit from the relationship while having a negative effect on the other, we refer to prey-predator relationships and parasitism (the parasite takes advantage of the host to shelter, reproduce or feed). Amensalism correspond to biological interactions that turns out to be negative for one of the partners, without benefit for the other. Finally, competition results in a negative outcome for both species involved, which exist in bacterial populations through the production of antibiotics (Nadell et al., 2009).

Co-occurrence can be inferred from incidence (i.e. presence-absence patterns) or abundance data. To overcome the limitations of culture-dependant methods, abundance and incidence data of microbial communities are now typically derived from high-throughput sequencing. The advent of these approaches revolutionized the taxonomic identification of microbial communities and the characterization of their functions. The collection of environmental samples and extraction of DNA or RNA is followed by amplification and sequencing. Species are generally identified through the sequencing of DNA barcodes (also called metabarcodes). These barcodes allow for the taxonomic characterization of organisms based on short distinctive DNA sequences and are different depending on the

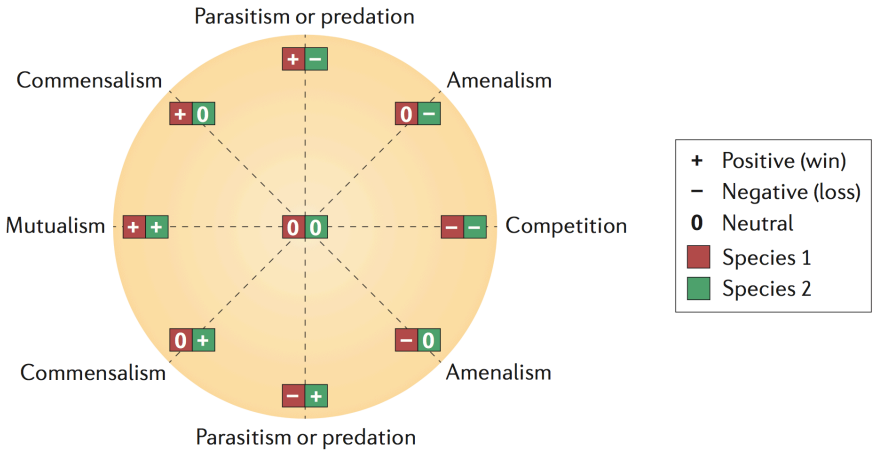


Figure 3.2 - Classification of ecological pairwise interactions, based on the outcome for both species involved. Interaction can result in positive (+), negative (-) or neutral (0) outcomes for the partners. For example, commensalism result in a positive outcome for one of the species involved, while the other doesn't take any advantage from the interaction, but is not harmed (Faust and Raes, 2012).

taxonomic groups. The DNA coding for the 16S ribosomal RNA (rRNA) is regarded as the "gold standard" for characterizing prokaryotic communities (Sun et al., 2013). One or more of the 16S hypervariable regions (from V1 to V9) are sequenced. Other barcode genes are available for eukaryotes. *Tara* Oceans eukaryotic organisms were identified with the V9 region of the 18S rRNA coding gene (Vargas et al., 2015; Alberti et al., 2017), whereas other environmental studies focuses on the V4 region (e.g. Stoeck et al., 2010; Massana et al., 2015). However, barcodes adapted to specific eukaryotic groups are also used (e.g. the internal transcribed spacer (ITS) region of the nuclear ribosomal repeat unit for fungal species and a part of the cytochrome *c* oxidase 1 (CO1) mitochondrial gene for animals).

After quality control steps to check potential errors from the sequencing process, and sometimes after sequences corrections (e.g. denoising, chimera checking), metabarcodes can be either directly studied with their corresponding abundance (e.g. Thompson et al., 2017) or they can be clustered into operational taxonomic units (OTU). Metabarcodes are often clustered when they are at least 97% similar (Westcott and Schloss, 2015), which is totally debatable. Moreover, protocols, approaches and tools to cluster sequences into OTUs are numerous (for a broader review and comparison of the approaches, see Westcott and Schloss, 2015). Metabarcodes are compared to curated reference databases (e.g. SILVA for prokaryotes, Quast et al., 2013; PR2 for microbial eukaryotes, Guillou et al., 2013). Finally, the analysis of the metabarcodes classically leads to the building of an

abundance matrix of the metabarcodes (or the OTUs) in the samples. Abundance tables of metabarcodes or OTUs are used as input for network inference methods, based on the idea that abundances are shaped by ecological interactions. This conception, although debated, originates from Jared Diamond who suggested that competition among birds of a same island resulted in mutual exclusion (Diamond, 1975). From this assumption, positive relationships between two species are inferred when they co-occur and negative relationships are deduced when they co-exclude each other. For instance, co-exclusions may reflect competition or amensalism, whereas co-occurrence may be due to commensalism. Parasitism and predation can be more difficult to predict because the consumer rely on its prey but can also make its population decrease. However, caution is required because the co-occurrence of two species doesn't necessarily imply that they interact. The co-presence of two species in a same environment can be due to cross-feeding for example. However, it can also result from shared environmental preferences (Chaffron et al., 2010) or because of a third factor such as an unreported abiotic driver or a species not accounted in the data set (Röttjers and Faust, 2018). Besides, unless the detected interactions were previously described in the literature or further experimentally validated, it is virtually impossible to confirm the underlying nature of the relationships behind positive and negative associations.

The ecological networks resulting from network inference are referred to as microbial co-occurrence networks (Kara et al., 2013), microbial association networks (Faust and Raes, 2012; Kurtz et al., 2015), microbial correlation networks (Duran-Pinedo et al., 2011; Friedman and Alm, 2012), or networks of co-existing microbes (Chaffron et al., 2010).

3.1.1.2 Methods for inferring microbial associations

Microbial association networks can be inferred with a wide range of methods based on metagenomic data. These methods rely on diverse metrics and models, the simplest and fastest ones using pairwise dissimilarity measures and the more complex ones using multiple regression and Gaussian graphical models (Layeghifard et al., 2017). Some are more used than others because of their speed and ease of use (Layeghifard et al., 2017), yet others, such as probabilistic graphical models (e.g. Kurtz et al., 2015), may show higher accuracy.

Two groups of inference network methods can be distinguished: the ones based on statistical inference and the ones using probability theory. The first group gathers methods based on dissimilarity, correlation and regression while the second includes more recently developed methods which are known as probabilistic graphical models. Although little

applied to microbial networks at present, logic-based machine learning algorithms may also be used (Vacher et al., 2016).

Dissimilarity and **correlation**-based methods have been implemented in many tools such as CoNet (Faust et al., 2012), SparCC (Friedman and Alm, 2012), WGCNA (Langfelder and Horvath, 2008) and CCREPE (Schwager et al., 2019). The Bray-Curtis and Kullback Leibler indices are often used to compute pairwise dissimilarity scores and their significance is tested through a permutation test (Faust et al., 2012; Layeghifard et al., 2017). Pairwise correlations between OTUs are computed with the Pearson's product moment or Spearman's nonparametric rank correlation coefficients. Correlation-based methods are popular and have been used to infer microbial associations in diverse environments, such as human gut (e.g. Jackson et al., 2018), soil (e.g. Mandakovic et al., 2018) and ocean (e.g. Lima-Mendez et al., 2015; Guidi et al., 2016). However, caution is needed when using correlation: spurious correlations may occur among low-abundance OTUs when data are sparse (i.e. significant correlations can be detected when computed on many matching zeros, while the taxa involved may vary randomly below the detection limit) and this type of metric is sensitive to compositionality (Layeghifard et al., 2017; Röttjers and Faust, 2018). That's why some precautions are needed when inferring microbial association networks from 16S data, such as applying a prevalence filter to remove rare taxa and revome outlier samples. Besides, correlation methods were shown to considerably vary in sensitivity and precision, making comparisons between studies difficult (Weiss et al., 2016).

Regression-based methods are an alternative to classical pairwise dissimilarity and correlation metrics. Indeed, they allow to capture more complex forms of relationship involving more than two partners by inferring the abundance of an OTU from the combined abundances of other OTUs with multiple regression. However, the meaning of regression results may be more difficult to interpret (Faust et al., 2012; Layeghifard et al., 2017) and, like correlation and dissimilarity, detected associations don't necessarily imply that there exist an underlying biological explanation for this. The high number of available predictors may also lead to overfitting, although sparse regression and cross-validation may overcome this problem. A second approach may be used to capture complex interactions: **association rule mining** (Agrawal et al., 1993). This approach consists in finding significant rules from presence-absence data (e.g. "if species A and B are present, then species C is absent"). Networks resulting from multiple regression or association rule mining are directed hypergraphs (Faust and Raes, 2012).

Probabilistic graphical models represent dependencies between random variables. They use graphs to measure joint probability distributions and represent sets of conditional

dependance and independance (Layeghifard et al., 2017). SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Assoiation Inference) (Kurtz et al., 2015) is an example of this type of methods. It builds microbial networks using either sparse inverse covariance or neighborhood selection. It attempts to avoid the spurious associations resulting from the application of correlation measures. In short, a link between two OTUs is inferred if their abundances are not conditionnally independant and if there is a relationship between them that cannot be better explained by another link.

Logic-based machine learning algorithms may also be used to learn microbial association networks from OTU occurrence or abundance data combined with background knowledge (e.g. about the species and their environment). They have been used successfully to infer trophic networks (Bohan et al., 2011; Tamaddoni-Nezhad et al., 2013) but could be applied to build microbial association networks (Vacher et al., 2016).

3.1.2 Overview of graph theory and useful metrics for microbial association networks analysis

The easiest way to represent microbial associations is in the form of a graph where nodes are species and edges represent their interactions. Microbial association networks analysis borrows many concepts and tools from graph theory to retrieve community properties that are encoded in the network structure, such as keystone species and ecological niches. To introduce the analyses presented in section 3.2, a brief overview of graph theory and its contribution to microbial association networks is given here.

3.1.2.1 Definition of a graph

A graph is a structure that models the relationships between objects (called *nodes* or *vertices*) connected by links (called *edges*). In mathematical language, a graph is denoted as $G = (V, E)$, where V is the set of vertices and E is the set of edges of the graph G. The terms *graph* and *network* are often used interchangeably in the scientific literature. However, subtle differences exist between them: while the term network refers to real systems, graphs are the mathematical representation of networks (Barabási, 2016) (real networks of different nature can have the same graph representation).

Graphs are divided in different classes, based on the properties of their edges: *directed* or *undirected*, *weighted* or *unweighted* (figure 3.3A), *signed* or *unsigned*, *cyclic* or *acyclic*. The edges of directed graphs are associated with a direction. For example, food webs are directed graphs where edges point from predators to preys. On the contrary, protein-protein interaction (PPI) networks describe physical interactions between proteins of an

organism, thus they are modelled as undirected graphs (Mason and Verwoerd, 2006). Weighted graphs have a weight assigned to their edges. Continuing with the example of food webs, the weight of the relationship between two species could be the amount of energy transferred from the prey to the predator. In unweighted graphs, all edges have the same weight (in practice, it is set to 1). Edges of signed graphs can bear a positive or negative sign. In social networks, positive edges would model the relationship between friends while negative edges would represent the one between enemies. Finally, a graph is cyclic if it contains a cycle, which is a path that has the same start and end node. In organic chemistry, molecules like sugars and aromatic compounds can be represented as cyclic graphs.

3.1.2.2 Interpretation of the structural properties of a graph

After a graph is obtained, many properties can be calculated, related to individual nodes and edges (such as centrality measures, see the box below), but also to the whole graph (such as the diameter). From these properties can be inferred characteristics of microbial association networks but they should be interpreted with care because networks are simplified representations of the system (Faust and Raes, 2012).

The **node degree distribution** is often calculated on microbial networks. It is the distribution of the number of direct neighbours a node has in a network. It provides information on the topology of the networks (i.e. the way in which the nodes and edges are arranged within a network). For example, scale-free graphs (i.e. graphs that have a power-law degree distribution) have many low degree nodes and few highly connected nodes (see figure 3.3B). Scale-free graphs are recognized to be typical of biological networks (Barabási and Albert, 1999), however, some microbial networks have also been found to be random graphs (e.g. Mandakovic et al., 2018). The degree distribution of random graphs follows a binomial or a Poisson distribution, in which node degrees are clustered around the mean degree (figure 3.3B). This type of graph is characterized by nodes randomly connected to each other (Barabási, 2016).

The **diameter** is the distance between the two furthest nodes of a graph (i.e. the longest shortest path). The diameter of biological and social networks is typically small in comparison to the networks' size (i.e. its number of nodes). This property is often referred to as the small world property. This measure gives information on how fast an information can be transmitted in a network. Thus, a short diameter suggests that few intermediate interactions are necessary to transfer information.

Centrality measures and clustering coefficient

The calculation of centrality measures varies according to the types of networks cited in the previous paragraph. For the sake of brevity, the following measures apply only to unweighted and undirected networks. The degree centrality is a local centrality measure while global centrality measures include the **betweenness** centrality and the **closeness** centrality. The degree, betweenness centrality and clustering coefficient are illustrated in figure 3.3C.

The **degree** centrality (DC) of a node is the number of its direct neighbours (i.e. the number of edges a node has).

The **betweenness** centrality (BC) of a node k is the fraction of shortest paths[a] going through a given node (Freeman, 1977):

$$BC(k) = \sum_i \sum_j \frac{\rho(i,k,j)}{\rho(i,j)}, i \neq j \neq k \quad (3.1)$$

where $\rho(i,k,j)$ is the number of shortest paths between nodes i and j that pass through node k, and $\rho(i,j)$ is the number of shortest paths between nodes i and j. Betweenness centrality quantifies the influence of a node on the network.

The **closeness** centrality (CC) of a node k is the reciprocal of the sum of topological distances from all other nodes in the network (Bavelas, 1950):

$$CC(k) = \frac{1}{\sum_j d(k,j)} \quad (3.2)$$

where the distance $d(k,j)$ from the node k to another node j is defined as the number of links in the shortest path from one to the other. A node close to the other nodes of the network can typically communicate rapidly with them.

The **clustering coefficient** of a node is the fraction of connections among all possible connections between its direct neighbors (Watts and Strogatz, 1998). Thus, it measures the degree to which nodes in a graph tend to cluster together.

[a]A path is the distance between two nodes, measured as the number of edges that separates them.

From the network topology can be delineated **modules**. Modules are densely clustered subgraphs that can be detected manually or by using a dedicated algorithm (e.g. Clauset et al., 2004). Modules have been interpreted as niches in some studies (e.g. Chaffron et al., 2010).

Finally, highly central nodes according to different measures (see the box) have often

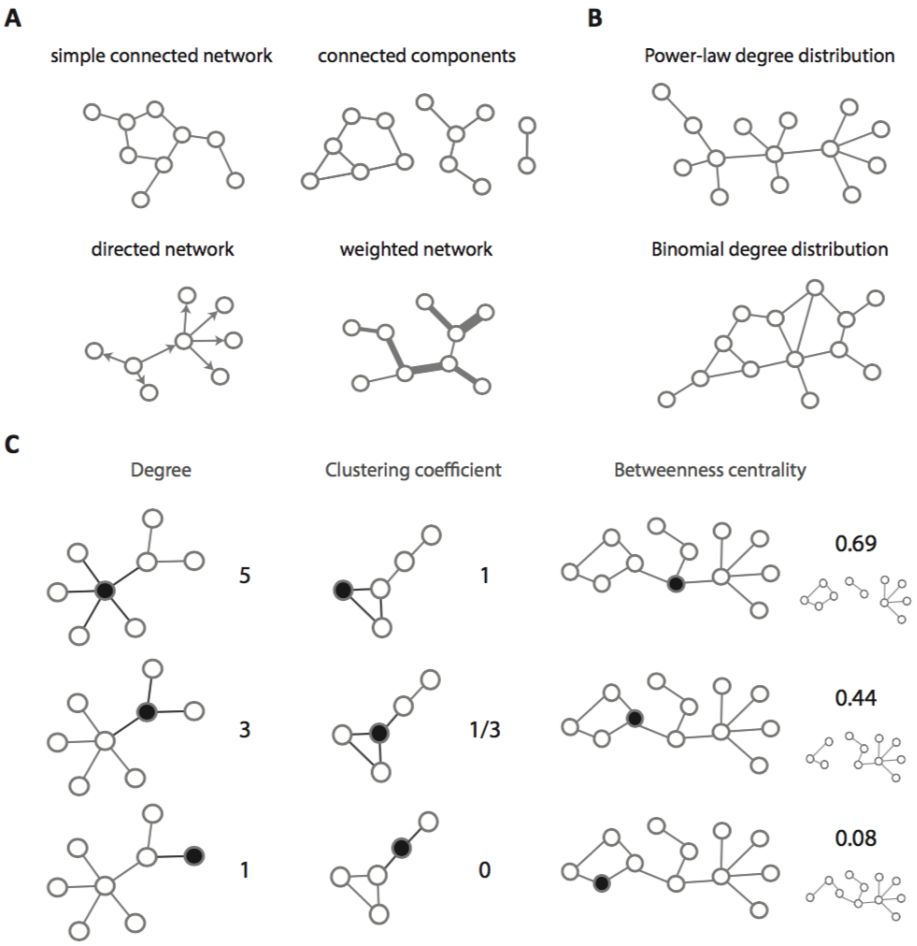


Figure 3.3 - Overview of graph types and graph theory metrics. On all diagrams, nodes are represented as circles and edges as links between circles. (A) Classes of graphs, based on the properties of their edges, (B) Two graph topologies that differ in their degree distribution and (C) Illustration of centrality measures (degree and betweenness) and the clustering coefficient (Perez, 2015).

been considered as **keystone species**. The keystone species concept originates with Paine (1969) who observed a reduction of species richness in a rocky shore community in California after removal of the top predator (a starfish). Keystone species are commonly defined as species "whose effect on its community or ecosystem is large, and disproportionately large relative to its abundance" (Power et al., 1996). The term refers to the structure of an arch that would collapse if the keystone was removed. In macro-ecology, predators are often considered keystone species because they control the population dynamics of their preys. For example in the microbial world, nitrogen-fixing bacteria make nitrogen available to other plants and animals, thus playing a central role in the nitrogen cycle (Robidart et al., 2014). Since this concept was introduced in microbial ecology, the identification of keystone microbial species is a critical issue given the complexity of microbial communities, their high diversity and the difficulty to cultivate most microbes (Lupatini et al., 2014).

3.1.3 Guidi et al. (2016): a first study of the carbon export through the lens of meta-omic data

In their study published in 2016, Guidi et al. initiated research linking omic data and biogeochemical processes. The objective was to relate the planktonic community structure to carbon export in the oligotrophic ocean. For this purpose, they applied a systems biology approach known as weighted gene correlation network analysis (WGCNA, Langfelder and Horvath, 2008) to detect significant associations between the *Tara* Oceans genomic data and carbon export. This method delineates subnetworks of highly correlated OTUs or genes and extract the ones that are associated to environmental variables. From these subnetworks, they emphasized key nodes using partial least square regression (PLSr). This network-based approach was applied to eukaryotic, prokaryotic and viral metabarcoding and to prokaryotic metagenomic datasets. It revealed unexpected taxa such as Radiolaria and alveolate parasites, as well as Synechococcus and their phages, as lineages most strongly associated with carbon export. The relative abundance of few bacterial and viral genes was also showed to predict a significant fraction of the variability in carbon export in the oligotrophic ocean.

In the following article, we propose to improve these results by integrating the other components of the biological carbon pump (i.e. net primary production and flux attenuation) and inferring microbial association networks corresponding to biogeochemical states of the biological carbon pump.

3.2 Article 2 (Benoiston et al., in prep.): The microbial drivers of the biological carbon pump

Target journal for the following draft manuscript: The ISME Journal

The microbial drivers of the biological carbon pump

Anne-Sophie Benoiston[1,2], Damien Eveillard[3], Samuel Chaffron[3], Erwan Delage[3], Lionel Guidi[2], Lucie Bittner[1]

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, Sorbonne Université, CNRS, EPHE, Université des Antilles, CP 50, 57 rue Cuvier, 75005 Paris, France.

[2]Sorbonne Université, CNRS, Laboratoire d'océanographie de Villefranche, LOV, 06230 Villefranche-sur-Mer, France.

[3]LS2N UMR6004 CNRS, Université de Nantes, Centrale Nantes, IMTA, Nantes, France.

3.2.1 Abstract

The biological carbon pump consists in a series of processes that encompasses the production of organic matter by phytoplankton, its transport to the deep and its degradation and physical fragmentation by heterotrophic bacteria and zooplankton. While key players have been identified, the microbial relationships driving the biological pump have been poorly investigated. Here we propose to integrate the processes of the biological pump (i.e. primary production, carbon export and flux attenuation) to define biogeochemical states dominated by one of the processes. Microbial association networks were inferred from prokaryotic association networks. Through the analysis of these networks, we show that variation in microbial associations rather than lineages drive the states of the biological carbon pump.

3.2.2 Introduction

The world ocean is characterized by a strong vertical gradient of concentration of dissolved inorganic carbon (DIC). Concentrations of DIC are higher in the deep (below 1200 m) than in surface waters (i.e. 2284 $\mu mol.kg^{-1}$ and 2012 $\mu mol.kg^{-1}$, respectively, Volk and Hoffert, 1985). This vertical gradient is maintained by the three ocean carbon pumps defined by Volk and Hoffert (1985) in their seminal article, namely the solubility pump, the carbonate pump and the biological pump. In the surface, dissolved inorganic materials are converted to organic matter through photosynthesis (i.e. primary production process). The bulk of this newly produced biomass is recycled in the surface waters by respiration or heterotrophy (i.e. remineralisation process) while the organic material that escapes recycling is transported by sinking to the deep waters (i.e. export process), or even to the sediments, where it is sequestered.

Because of its capacity to trap atmospheric carbon (Falkowski et al., 1998; Boyd and Trull, 2007; Buesseler and Boyd, 2009) and its impact on the Earth's climate (Sigman and Boyle, 2000), the biological carbon pump has been the subject of particular attention by oceanographers for four decades. First sediment traps were deployed in the 1960s-1970s (Wiebe et al., 1976; Berger and Soutar, 1967; Honjo, 1976; Soutar et al., 1977) and allow to capture sinking particles. Installed at various depths (generally between the surface and 3000 m, sometimes deeper, Honjo et al., 2008) they give access to particle fluxes in surface and at depth. In particular, carbon export (i.e. the quantity or carbon that leaves the euphotic zone) and flux attenuation along the water column can be calculated from these fluxes.

Fluxes through the mesopelagic zone (from about 200 to 1000 m depth) are influenced

by planktonic organisms at the surface (Suess, 1980; Berger et al., 1989; Tréguer et al., 2003; Guidi et al., 2009; Buesseler and Boyd, 2009). The common highlighted actors are large phytoplankton (such as diatoms) (Allen et al., 2005; Agusti et al., 2015; Benoiston et al., 2017) because of their significant contribution to primary production and carbon export, and zooplankton by the production of fecal pellets (i.e copepods) (Turner, 2002, 2015) and mucus feeding structures (i.e appendicularians) (Alldredge, 2005). Phytoplankton composition in surface notably influences the strength of carbon export (Boyd and Newton, 1995; Boyd et al., 2008). In particular, Guidi et al. (2009) showed that carbon export is linked to the size of sinking particles, the export flux being more efficient when microphytoplankton (especially diatoms ballasted by their siliceous skeleton) dominates the euphotic zone (i.e. the surface ocean water layer where light intensity is sufficient for photosynthesis), compared to a community dominated by picophytoplankton (e.g. Cyanobacteria).

Since already 30 years, the identification of micro-organisms and their diversity in the environment is based on molecular data (e.g. Zinger et al., 2012). With the advent of the high-throughput sequencing, microbial communities and their molecular functions are revealed comprehensively in a multitude of ecosystems (e.g. Qin et al., 2010; Delmont et al., 2011; Sunagawa et al., 2015; Vargas et al., 2015; Thompson et al., 2017; Carradec et al., 2018). However, metabarcoding or metagenomic analyses focusing on free-living and particle-attached micro-organisms collected in sediment traps are rare and restricted to small or single geographical location to date (LeCleir et al., 2014; Fontanez et al., 2015).

In 2016, exploiting the meta-omic sets from the *Tara* Oceans expedition in light of their *in situ* environmental data, Guidi et al. (2016) published the first microbiology-driven carbon export study at the global scale. They highlighted molecular planktonic communities significantly associated to carbon export, as well as their key lineages and functions, which potential importance had not been revealed until then. These findings suggest that the biological carbon pump is the result of complex interactions among organisms rather than their independent actions, and that in general, meta-omics and statistical analyses are offering new insights into the understanding of elements cycling.

In the present study, from the metagenomic and *in situ* environmental data from the *Tara* Oceans expedition, we propose to revisit the study of carbon cycling in the oligotrophic ocean at a global scale by integrating for the first time the three processes of the biological carbon pump (i.e. primary production, export and flux attenuation, used as a proxy for remineralization). At a first step, we defined biogeochemical states in our data set based on the relative contribution of the following quantifiable parameters: net primary production (NPP), carbon export (CE) and flux attenuation (FA). A state corresponds to samples in

which one of these parameters is relatively high compared to the two other ones (e.g. the NPP state corresponds to samples in which NPP is relatively high compared to CE and FA). In a second step, each of these three biogeochemical states were analysed to determine its underlying network organisation, both in terms of taxonomic composition and lineage associations. In a third step, we compared these networks in order to highlight the common and specific actors and associations between the three biogeochemical states. Finally, we revealed that variation in associations rather than lineages presence seems to drive the states of the biological carbon pump.

3.2.3 Materials and methods

3.2.3.1 Sample collection and taxonomic profiling

The *Tara* Oceans circumglobal expedition sampled plankton and collected environmental data at 210 sites across all major oceanic provinces from 2009 to 2013 (Pesant et al., 2015). Plankton was collected in the surface water layer, the deep chlorophyll maximum and the mesoplagic zone. Metagenomic DNA from prokaryote-enriched size fraction filters (i.e. the 0.22-1.6 μm fraction up to station #52 and the 0.22-3 μm fraction from station #56) was extracted as described in Logares et al. (2014) and sequenced on Illumina sequencing machines. 16S fragments directly identified in Illumina-sequenced metagenomes (16S $_{mi}$tags) were identified as described in Logares et al. (2014). 16S $_{mi}$tags were mapped to 16S reference sequences from the SILVA database (Quast et al., 2013), release 115: SSU Ref NR 99) with a threshold of 97% sequence identity using USEARCH v.6.0.30759 (Edgar, 2010). OTUs abundances were calculated by counting the number of 16S $_{mi}$tags clustering into the same OTU. From the resulting abundance matrix published by Sunagawa et al. (2015), we removed OTUs detected with sequence abundance < 2 to reduce sequencing artefacts. Finally, we selected samples collected in the surface water layer and the deep chlorophyll maximum. The final matrix included 26 281 OTUs and 104 samples and is publicly available on https://figshare.com/s/23798e4046a2c21a9103.

3.2.3.2 Environmental parameters calculation and definition of the biogeochemical states of the biological carbon pump

Primary production

Net primary production (NPP) was calculated from satellite measurements using the vertically generalized production model (VGPM) (Behrenfeld and Falkowski, 1997) at the sampling location, at a period of 8 days around the sampling date (Chaffron et al., 2014).

Carbon export

The Underwater Video Profiler (UVP) (Picheral et al., 2010) was used to estimate particle concentration and particle size distributions (PSDs). The PSD is often calculated in terms of concentration ΔC (number of particles per unit volume) in a given size range Δs: $n(s) = \Delta C/\Delta s$. Although any measure of particle size can be used for s, the particle diameter d was used for the measure of particle size. Assuming that the mass $m(d)$ and sinking speed $w(d)$ are functions of d, then the total mass of particles in the size range Δd is $n(d)m(d)\ \Delta d$ and the mass flux in that interval is $n(d)m(d)w(d)\ \Delta d$ (Guidi et al., 2008). The total carbon flux of particles F corresponds to the flux spectrum integrated over all particle sizes:

$$F = \int_{d_{min}}^{d_{max}} n(d)m(d)w(d)\ \Delta d \tag{3.3}$$

where d_{min} and d_{max} are the minimum and maximum particle diameters, $n(d)$ is the particle size spectrum, $m(d)$ is the mass (here carbon content) of a spherical particle and $w(d)$ the settling rate calculated using Stokes law. The combined mass and settling rates of particles were described as power law functions of their diameter (of the form yd^x) fitted by comparing image-derived PSD with sediment traps estimate of mass flux (Guidi et al., 2008). If both $m(d)$ and $w(d)$ are given by power relationships, then the resulting combined quantity is $wm = Ad^b$. Hence, the particle carbon flux can be approximated using equation (3.3) over a finite number of small logarithmic intervals of diameters d from 250 μm to 1.5 mm (particles <250 μm and >1.5 mm are not considered, as presented in Guidi et al., 2008) (Guidi et al., 2009; Picheral et al., 2010) such that:

$$F = \sum_{i=1}^{x} n_i A_i^B\ \Delta d_i \tag{3.4}$$

where A=12.5 ± 3.40 and B=3.81 ± 0.70 corresponds to the best set of parameters that minimized the log-transformed differences between particle flux in sediment traps and PSDs from the UVP images (Guidi et al., 2008). Carbon export (CE) was measured at 150m to make sure that it was measured below the euphotic zone Z_e and the maximum mixed layer depth Z_{MLD} (mean Z_e in Tara Oceans stations = 64 m, SD = 38 m; mean Z_{MLD} in Tara Oceans stations = 48 m, SD = 44 m). Besides, measuring carbon export at this depth was consistent with the average deployment depth of conventional sediment traps.

Flux attenuation

Carbon flux data, calculated for each *Tara* Oceans stations every 5 meters below the surface, were smoothed (symmetrical moving average on 40 meters) to avoid transitory fluctuations (R package `stats`, function `filter`). Flux attenuation (FA) was computed as

follows:

$$FA = \frac{F_{max}(20..150)}{F_{+300}} \tag{3.5}$$

where FA is the flux attenuation, $F_{max}(20..150)$ is the maximum flux value between 20 and 150 m and F_{+300} is the flux value 300 m below $F_{max}(20..150)$. The difference of 300 m was chosen because the majority of flux attenuation occurs above 300m depth below the euphotic zone (Buesseler and Boyd, 2009).

Definition of the NPP, CE and FA states

To define biogeochemical states of the biological carbon pump, NPP, CE and FA measurements were normalized as follows:

$$z_i = \frac{x_i}{x_{max}} \tag{3.6}$$

where x_i are the absolute values and x_{max} is the absolute maximum value of each measurement, setting the maximum of each variable to 1. Once normalized, the percentage of each variable was computed for each sample (i.e. for one sample, when NPP, CE and FA measurements were all available, they were normalized in order to set the sum of the three variables to 1). Absolute and normalized values of NPP, CE and FA estimations are publicly available on `https://figshare.com/s/f67bcb072aea125039d3`. These steps allowed us to classify the samples either in the NPP, CE or FA states (e.g. samples in which NPP is relatively high compared to CE and FA were classified in the NPP state) and to create the corresponding three OTUs abundance matrices.

3.2.3.3 Analysis of states differenciation based on environmental parameters and community composition

A Mantel test was performed to compare Bray-Curtis environmental similarity (based on the contribution of net primary production, carbon export and flux attenuation) and geographic distances with the function *mantel()* of the R package *vegan*. Non-metric multidimensional scaling was performed on prokaryotic samples based on 16S $_{mi}$tag relative abundances at different taxonomic levels with the function *metaMDS()* of the R package *vegan*. To test whether there is a significant difference between the states on basis of community composition, we computed analyses of similarity (ANOSIM) with the function *anosim()* from the R package *vegan*.

3.2.3.4 Association networks inference

OTUs abundance matrices of each state were filtered following these two steps: (1) OTUs that occurred in less than 70% of the samples of each subset were removed, and

(2) only the 20% most variable OTUs were retained for network reconstruction (figure 3.9). These matrices are available on `https://figshare.com/s/cf2ed998f7796daa368a`, `https://figshare.com/s/283f073035987abdd6d7` and `https://figshare.com/s/a9d4d2f3be5cce4bbe20`. Each of the reduced OTUs abundance matrix was then analyzed with SPIEC-EASI (R package developped by Kurtz et al., 2015) in order to build association networks (in which a node corresponds to an OTU, and an edge corresponds to a positive or negative association). SPIEC-EASI estimates ecological networks from meta-omic data using a graphical model (neighborhood selection or sparse inverse covariance selection). In both selection methods, the tuning parameter λ controls the sparsity of the final model. This final model is selected by subsampling the original dataset and estimating the graph for several λ values. For each graph, a stability metric is calculated and the λ value minimizing the graph variability is selected. Model selection is performed with the StARS method (Stability Approach to Regularization Selection) (Liu et al., 2010).

SPIEC-EASI first applies the centered log-ratio transformation to the abundance matrix to alleviate compositionality bias (Aitchison, 1981) and then builds the network. The networks were inferred using the neighborhood selection (Meinshausen and Bühlmann, 2006). The SPIEC-EASI parameters that generate the sparser graphs were determined through testing of multiple values for each parameter. Thus, the scaling factor that determines the minimum sparsity (`lambda.min.ratio`) was set at 0.001 while the number of tested lambda (`nlambda`) was determined as 20. The number of StARS subsamples (`rep.num`) was set to 20.

3.2.3.5 Networks properties and metrics

We assessed the topological properties of the microbial association networks with the R package *igraph* (Csardi and Nepusz, 2006). The following networks properties potentially relevant for the analysis of microbial association networks were calculated (density, diameter and average path length are calculated on the whole graph while centrality measures and the clustering coefficient are calculated on nodes):

- density: a node can be connected to all other nodes of the network. Possible connections are therefore equal to $n*(n-1)/2$. The density of a network is the proportion of actual connections among all possible connections.
- diameter: length of the longest among all shortest paths (i.e. path between two nodes that has the fewest number of edges) between node pairs.
- average path length: average path length is calculated by finding the shortest path

between all pairs of nodes, adding them up, and then dividing by the total number of pairs.

- degree centrality (DC): number of edges linked to a node.
- betweenness centrality (BC): the betweenness of a node is equal to the fraction of shortest paths between all other nodes that are passing through this node. A version of the BC exists for edges: it is the fraction of shortest paths that go through this edge.
- closeness centrality (CC): reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.
- clustering coefficient: the clustering coefficient of a node is the fraction of connections among all possible connections between its neighbours.

The mean degree, betweenness, closeness and clustering coefficient was calculated from all the nodes belonging to a network. The goodness of the fit of the node degree distribution with the power law and Poisson distributions was calculated with a Kolmogorov-Smirnov test (function *fit_power_law()* of the *igraph* package) and a chi-squared test, respectively.

For each network, keystone OTUs were defined as nodes displaying high degree, high betweenness and high closeness. The top ten keystone OTUs were selected in each network by computing the following centrality score:

$$CS_i = \frac{DC_i}{max(DC)} + \frac{BC_i}{max(BC)} + \frac{CC_i}{max(CC)}$$

where CS_i is the centrality score of node i, DC is the degree centrality, BC is the betweenness centrality and CC is the closeness centrality.

3.2.4 Results

3.2.4.1 Spatial structure of the net primary production, carbon export and flux attenuation states

The relative contribution of net primary production, carbon export and flux attenuation to the biological carbon pump differs between samples. From these variations, we defined the net primary production (NPP), the carbon export (CE) and the flux attenuation (FA) biogeochemical states (a state corresponds to samples in which one of these parameters is relatively high compared to the two other ones) (figure 3.4A). For each state, the dominating process (NPP, CE or FA) represents from 40% to 100% of the relative contribution.

NPP, CE and FA measurements were all available for only 61 out of 104 samples. The resulting NPP subset involves 14 samples, the CE subset involves 32 samples and the FA subset involves 18 samples, with overlaps between the subsets.

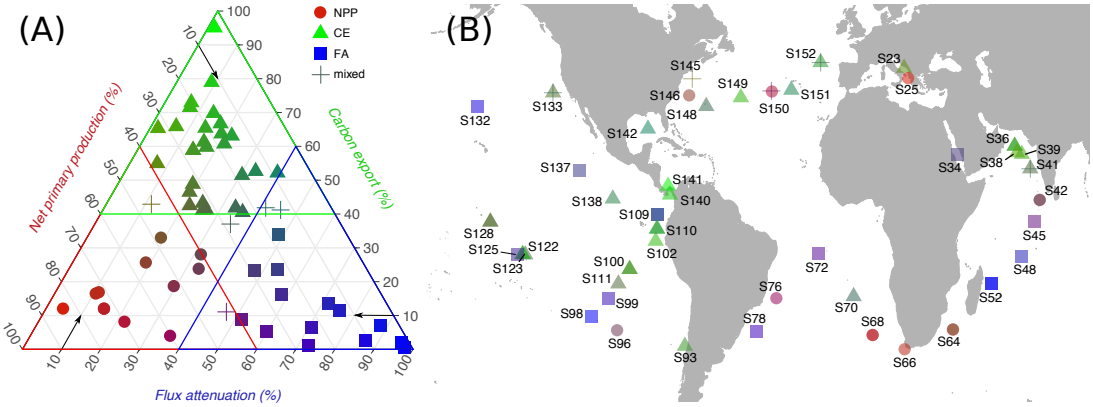


Figure 3.4 - Contribution of net primary production (NPP), carbon export (CE) and flux attenuation (FA) to the biological carbon pump in Tara Oceans samples. (A) Ternary plot of the relative contribution (in %) of NPP, CE and FA in Tara Oceans samples and delimitation of the three biogeochemical states. The arrows indicate how to read the values on the axes, the base referring to the axis and the head to the direction of reading. The 14 samples corresponding to the NPP state are framed by a red triangle, the 32 samples corresponding to the CE state are framed by a green triangle and the 18 samples corresponding to FA are framed by a blue triangle. Crosses correspond to mixed samples (i.e. samples with equal relative contribution of two or three processes). (B) Map of Tara Oceans samples, represented as in the ternary plot (A).

Samples related to each state were projected on a world map (figure 3.4B). No geographical structure is observed on the basis of the contribution of NPP, CE and FA (Mantel test *Spearman's* r=0.036, p=0.144). However, some samples exhibit a strong contribution of one of the three variables (NPP for stations 25, 66 and 68, CE for stations 140 and 141, FA for station 52).

3.2.4.2 Taxonomical composition of the three states differs at the level of orders, families and OTUs

We compared the taxonomical composition of each biogeochemical state, from the phylum to the OTU level. At the phylum level, the three states show similar composition (Figure 3.5): a dominance of Alphaproteobacteria, Cyanobacteria and Betaproteobacteria. To investigate the potential link between taxonomical composition and states, non-metric multidimensional scaling (NMDS) and analysis of similarity (ANOSIM) were performed (figures 3.10 and 3.11). A significant link (at the *alpha* = 0.05 level) was revealed at the levels of orders (*ANOSIM* r=0.105, p=0.035), families (*ANOSIM* r=0.098, p=0.044) and OTUs (*ANOSIM* r=0.109, p=0.034) (figure 3.11). A community structuring effect appears within the FA samples, whereas the NPP and CE samples composition show an over-

lap, especially at the level of OTUs. Besides, five lineages display significantly different abundances at the *alpha* = 0.05 level: Bacteroidetes (*Kruskal-Wallis test* p=0.029), Gracilibacteria (*Kruskal-Wallis test* p=0.011), Chlorobi (*Kruskal-Wallis test* p=0.013), Gemmatimonadetes (*Kruskal-Wallis test* p=0.015) and Thaumarcheota (*Kruskal-Wallis test* p=0.005).

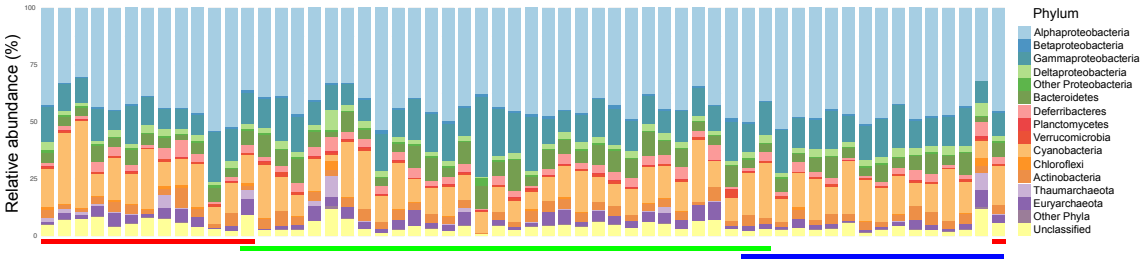


Figure 3.5 - Phylum-level (class-level for Proteobacteria) taxonomic composition of the states. Colored bars under the barplots indicate the state each sample belongs to (red: NPP, green: CE, blue: FA).

3.2.4.3 Association networks differ in their properties

Before association networks building, OTUs were filtered for each state in order to focus our study on the most prevalent and variable ones. Network metrics related to nodes and edges of the three states association networks are reported in table 3.1. Node degree distribution of networks did not fit the power law according to the Kolmogorov-Smirnov test (p=0 for all degree distributions). The NPP and FA node degree distribution were found to follow the Poisson distribution (*Goodness-of-fit* p=5.172e-4 for NPP and p=1.529e-5 for FA) which is typical of random graphs (Erdös and Rényi, 1959), where each pair of nodes have the same probability to be linked by an edge. The NPP network involves the highest number of nodes and edges. Networks are mostly composed of positive edges (between 66% and 73%), the network having the fewest positive edges percentage (PEP) being the FA network and the one having the highest PEP being the CE network (table 3.1).

The degree, betweenness, closeness and clustering coefficient distributions differ between the states (supplementary figure 3.12 and table 3.3). More precisely, the degree, betweenness and closeness are significantly different between the NPP and CE networks. The FA and NPP networks differed significantly in terms of degree, betweenness, closeness and clustering coefficient, while the closeness and clustering coefficient were significantly different between the CE and FA networks.

Table 3.1 - Networks characteristics and metrics. Standard deviations are given in brackets.

	NPP state	CE state	FA state
number of samples	14	32	18
number of nodes	425	371	363
number of edges	1 991	1 570	1 581
number of positive edges (%)	1 401 (70%)	1 149 (73%)	1 048 (66%)
number of negative edges (%)	590 (30%)	421 (27%)	533 (34%)
network density	0.022	0.023	0.024
network diameter	6	6	6
average degree	9.369 (SD=2.442)	8.464 (2.578)	8.711 (2.198)
average clustering coefficient	0.1263 (SD=0.0620)	0.131 (0.094)	0.1107 (0.0687)
average path length	3.24	3.29	3.18
average betweenness	474.86 (SD=308.607)	423.3 (276.270)	394.0 (218.2669)
average closeness	0.0007322 (SD=5.396e-5)	0.0008274 (6.535e-5)	0.0008795 (5.228e-5)

3.2.4.4 Specific *vs* core OTUs within the states

We compared the nodes metrics from the core and specific OTUs of each state. At the OTU level, the highest proportion of OTUs (*i.e.* 37.5% or 216 OTUs) corresponds to core OTUs (*i.e.* OTUs shared by the three states) (figure 3.6). The NPP state is showing the highest proportion of specific OTUs (19.8% or 114 OTUs) whereas CE and FA showed only 8.5% (49 OTUs) and 8% (46 OTUs) of specific OTUs. Significantly higher degree, betweenness and closeness were observed for specific nodes of the NPP and the CE networks. On the contrary, no difference in terms of centrality between core and specific nodes was observed in the FA graph (table 3.2). Our results suggest that specific OTUs were often located in more central positions than core OTUs within two (i.e. NPP and CE) out of the three networks.

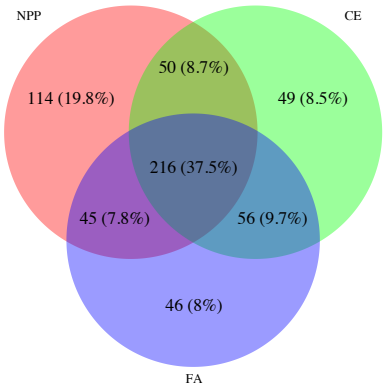


Figure 3.6 - Venn diagram showing overlapping and specific nodes (OTUs) of the NPP, CE and FA networks.

Table 3.2 - Results of Wilcoxon-Mann-Whitney U one-sided tests comparing core and specific nodes (OTUs) attributes. Significant results mean that the metric is higher for specific nodes.

Nodes attributes	NPP		CE		FA	
	U	p-value	U	p-value	U	p-value
degree	10 902	0.04245*	4 098	0.006519**	5 616.5	0.9196
between-ness	10 576	0.01761*	4 336	0.02427*	5 113	0.6224
closeness	9 609	5.205e-4***	4 059	0.00547**	4 574.5	0.1998

*** p<0.001, ** p<0.01, * p<0.05

3.2.4.5 Associations between and within the states

The specific and shared edges (OTU-OTU associations) within the three networks were listed. While more than a third of OTUs are common to all networks (figure 3.6), only very few edges (1.8%) are shared by the three networks (figure 3.7). This is also true when taking into account positive and negative edges separately. However, we observe that the networks share more positive (2.7%) than negative edges (0.2%). Interestingly, the networks that shared the most edges are the CE and the FA networks. The network that has the most specific edges is the NPP one. These tendencies are also observed for positive and negative edges separately.

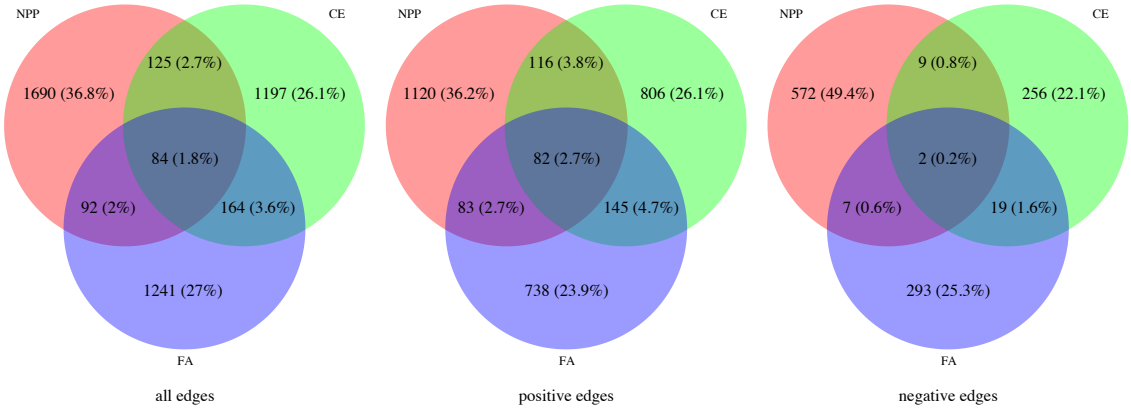


Figure 3.7 - Venn diagrams considering all, only positive and only negative edges of the NPP, CE and FA networks.

The positive edge percentage (PEP) between specific OTUs only, between core OTUs only and between core and specific OTUs (i.e. edges involving a core OTU on one side and a specific OTU on the other side) was computed. The PEP among specific OTUs is similar to the global PEP: it ranges from 71% to 75% but is higher among core OTUs (between 76% and 85% depending on the network, supplementary table 3.4). On the contrary, the PEP between core and specific edges is much lower as it lies between 50% and 57%, suggesting that OTUs preferentially connect to OTUs of the same "core nature". We further tested this hypothesis by measuring the assortativity coefficient of the networks as defined in

Newman (2003): it measures the preference for a network's nodes to attach to others that are from the same category (here core or specific). We computed the assortativity coefficient on edges involving core and specific nodes only. We observed that the three networks are assortatively mixed with respect to the core or specific nature of the nodes (when calculated on all edges: NPP network r=0.2, CE network r=0.25, FA network r=0.1; when calculated on positive edges only: NPP r=0.21, CE r=0.23, FA r=0.12), meaning that specific nodes preferentially connect to specific nodes and that core nodes are more likely to connect to core nodes, although this tendency seems to be less strong for the FA network.

3.2.4.6 Keystone OTUs and associations

Degree, betwenness, closeness and clustering coefficient were computed for all nodes. We first determined relationships between these four measures of centrality. Degree, betweenness and closeness appeared to be strongly and positively correlated, while they were negatively correlated to clustering coefficient (figure 3.13).

Hence, we defined keystone OTUs as having high degree, high betweenness and high closeness, thus they are the most connected nodes and play the role of bridges along the shortest path between many other nodes. Their high closeness entails that they are also close to all other nodes of the network. Among the top 10 keystone OTUs of the NPP network, the highest centrality score was assigned to *Synechococcus*, followed by a *Marine Group II* archaea, an undetermined cyanobacteria, and *Prochlorococcus* (table 3.5). Other keystone OTUs of this network include *Marinicella*, *Magnetospira* and representatives of the SAR11 and SAR86 clades. The top keystone OTUs of the CE network belong to the *Marinicella* genus and a *Marine Group II* archaea (same OTU as in the NPP network) (table 3.6). Among the eight following most central OTUs, three fall within the SAR86 clade, two within the SAR11 clade, one belong to the SAR324 clade, one to the NS9 marine group of Flavobacteria and an uncultured *Rhodobacteraceae*. As for the NPP network, the top one keystone OTUs of the FA nodes was assigned to *Synechococcus* (table 3.7). Another OTU taxonomically assigned to the *Synechococcus* genus is also present among the top ten keystone species of this network. Other keystone OTUs include representatives of the SAR11 and SAR86 clades, as well as an OTU belonging to the *Rickettsiales* order.

To assess the importance of edges in the networks, we calculated the edge betweenness. Like node betweenness, edge betweenness is the fraction of shortest paths going through a given edge. Consequently, we defined keystone associations as the top 10 edges having a high betweenness centrality (supplementary tables 3.8, 3.9 and 3.10). A low propor-

tion of edges display a high betweenness in the networks while a high proportion of edges have a low betweenness in the three networks (figure 3.8). Most keystone edges involve OTUs assigned to the SAR11 clade. Other common partners include *Synechococcus*, *Prochlorococcus* and SAR86. Keystone edges involve OTUs defined as keystones such as OTUs taxonomically assigned to SAR86 (ref. EF572127.1.1520) and Marine Group II (ref. DQ156348.18592.20063) in the NPP network (supplementary tables 3.5 and 3.8) and OTUs assigned to *Synechococcus* (ref. AF098371.1.1444) and SAR86 (refs. EU802400.1.1495 and FJ45180.1.1374) in the FA network (supplementary tables 3.7 and 3.10). Interestingly, the most central OTU of the FA network (assigned to *Synechococcus*) is involved in three out of the top 10 keystone associations of the same network.

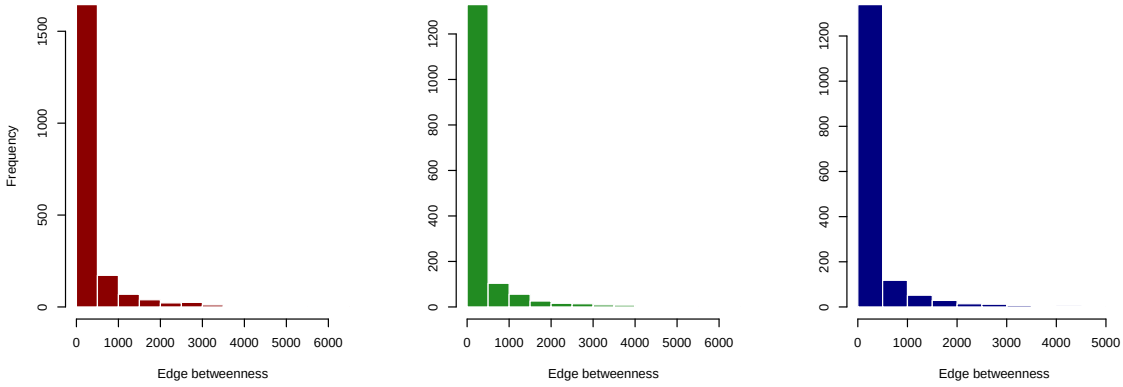


Figure 3.8 - Histogram of edge betweenness in the NPP, CE and FA networks (from left to right).

3.2.5 Discussion

3.2.5.1 A first attempt to empirically define states of the biological carbon pump

In this work, we revisited the study of the biological carbon pump by considering it as a system that can be dominated either by net primary production, carbon export or flux attenuation. We defined biogeochemical states from the relative contribution of each process in samples of the *Tara* Oceans expedition. Hence, with this methodology the states and the classification of the corresponding samples depend on the input data set. Other NPP, CE or FA estimations may fall outside the framework presented here, so to improve, confirm or refute our current results, more data would be needed to better represent the variability of the processes studied. For instance, adding *in situ* measurements and meta-omic samples from polar regions or from coastal upwelling regions could help to take into account a higher variability of NPP. However, as defining the states required measures from the three interest variables in a sufficient number of sampling stations, the

Tara Oceans data set satisfied this need, owing to the extensive suite of marine plankton collection and associated environmental data sampled in contrasting ecosystems (Pesant et al., 2015; Alberti et al., 2017). The original methodology developed in this manuscript is however directly transposable to any other or improved data sets. A further development of our study could be the implementation of our methodology to the understanding of the biological pump processes throughout the water column and notably in the mesopelagic areas (Duarte, 2015; Boeuf et al., 2019), and/or to marine ecosystem sampled regularly in order to investigate potential seasonal effect (e.g. Boeuf et al., 2019; Faust et al., 2015; Ai et al., 2019; Cram et al., 2014) or even punctual events such as blooms (Caputi et al., 2019).

3.2.5.2 Highlight of communities involved in each state of the biological carbon pump

The comparison of the taxonomic composition from the NPP, CE and FA state revealed variation at the levels orders, families and OTUs. To go beyond this basic community composition comparison, we inferred association networks for each of the state. Following a similar development to the one from Guidi et al. (2016), we revealed, for each state of the biological carbon pump, the corresponding prokaryotic community and its associations. The inference of microbial co-occurrence networks from meta-omic data became more and more widespread recently (e.g. Chaffron et al., 2010; Steele et al., 2011; Lima-Mendez et al., 2015; Mandakovic et al., 2018), allowing to infer widely distributed *in situ* relationships among micro-organisms. Whereas microbial association networks are often reported to follow a power-law distribution (meaning that they are scale-free) (e.g. Chaffron et al., 2010; Zhou et al., 2010; Steele et al., 2011; Ma et al., 2016), none of the networks inferred in this study were found to display this distribution. The degree distribution of the NPP and FA networks follows the Poisson distribution, which is typical of random networks (Barabási, 2016) and has previously been observed in microbial networks (e.g. Mandakovic et al., 2018).

Networks were inferred with SPIEC-EASI (Kurtz et al., 2015). This method relies on the inference of graphical models using the concept of conditional independence. It avoids wiring indirectly connected OTUs and thus produces sparse networks (compared to other inference methods like SparCC, Friedman and Alm, 2012, and CCREPE, Schwager et al., 2019). As a consequence, SPIEC-EASI has a high precision (percentage of predicted edges that are true positives), meaning that most predicted edges are true positives (Kurtz et al., 2015; Röttjers and Faust, 2018). Although the precision of SPIEC-EASI declines as the number of samples is reduced and that its sensibility (percentage of edges from the real

network that are predicted) is low (meaning that an important proportion of interactions existing in the real network are not predicted by the method, Röttjers and Faust, 2018), this method was shown to outperform classical correlation-based methods in terms of precision and capacity to remove indirect edges.

Considering that the samples we used to reconstruct networks are from different environments, associations predicted by network methods may reflect the influence of habitat filtering (Chaffron et al., 2010). To mitigate this effect, we applied a prevalence filter to remove taxa with low prevalence (as recommended in Berry and Widder, 2014; Weiss et al., 2016 and Röttjers and Faust, 2018). Consequently, only OTUs occurring in many samples were used to reconstruct networks and potentially important OTUs, for example highly specialized species that may have a low abundance, were not included in the networks.

3.2.5.3 Communities shared more than one third of common actors but differ on their associations

Although general characteristics of the three inferred networks are similar, centrality metrics vary between the states. For example, the NPP network displays higher node degree and betweenness compared to the CE and FA networks.

More than a third of the OTUs are shared between the three networks (37.5%), whereas very few associations are preserved (1.8%). Interestingly, the CE and FA networks have the highest number of common associations compared to the NPP network. Thus, the NPP network showed relatively more singularities compared to the CE and FA networks, but this trend has to be interpreted with caution as this network involves more nodes.

A large number of OTUs correspond to the same taxonomy, including *Synechococcus*, *Prochlorococcus*, SAR11 and SAR86. For instance, *Synechococcus* corresponds to 36 OTUs, and their repartition in the states is not homogeneous: 2 OTUs are specific to NPP, 10 OTUs are specific to CE, 6 OTUs are specific to FA, whereas 6 OTUs are core. These variations may reflect the involvement of distinct ecotypes (Vergin et al., 2013; Farrant et al., 2016) and definitely point out the need to investigate relationships between prokaryotic lineages via more resolutive genomic markers, in order to highlight precisely their different ecological associations (Sher et al., 2011).

Nowadays, network inference studies generate thousands of hypothetical associations, but only few of these interactions are validated, either by microscopy (Lima-Mendez et al., 2015; Mordret et al., 2016; Vincent et al., 2018) or by comparison with literature-curated gold standard databases (Poelen et al., 2014; Li et al., 2016; Vincent and Bowler, 2019).

Pairwise culture of bacterial strains highlighted positive (Sher et al., 2011; Biller et al., 2014) and antagonistic interactions among marine pelagic bacteria (Grossart et al., 2004; Long and Azam, 2001). We recovered a positive association observed in the lab between the genera *Prochlorococcus* and *Alteromonas* (Sher et al., 2011; Biller et al., 2014). In the lab, *Prochlorococcus* was found to support the growth of heterotrophic marine bacteria such as *Alteromonas* through the release of vesicles that was the only carbon source (Biller et al., 2014). This association is present only in the NPP network that include three edges involving *Prochlorococcus* on one side and *Alteromonas* on the other side. This association gives an example of the transport of carbon fixed by photoautotrophs into the microbial food web. Although the experimental validation of microbial interactions is necessary, this is a long a tedious task. As the integration of external data provides additional support for hypotheses based on microbial networks, the recurrent observations of the same association through distinct *in silico* studies might also be a path to consider for network validation.

Among the keystone OTUs was highlighted the cyanobacteria *Synechococcus* in the NPP and FA networks. This result confirms once again and via a distinct inference network method (compared to Guidi et al., 2016), the importance of *Synechococcus* in the biological carbon pump (Morán et al., 2004). SAR11 was also detected among highly central OTUs in all networks. This group is highly widespread and abundant (approximately 25% of all plankton, Giovannoni, 2017). The central position of SAR11 and SAR86 in all networks may be explained by their essential function of oxydizing carbon and nutrients in oligotrophic conditions (Giovannoni, 2017; Dupont et al., 2012). In addition to eubacteria, poorly known yet highly abundant Euryarchaeota from the Marine Group II (Zhang et al., 2015) display central positions in the NPP and CE networks. Like SAR11 and SAR86, they are potentially important players in the global carbon cycle because of their unique patterns of organic carbon degradation (Zhang et al., 2015).

3.2.6 Conclusion and perspectives

To our knowledge, this study constitutes the first attempt to define biogeochemical states of the biological carbon pump by integrating its three components (primary production, carbon export and flux attenuation) based on *in situ* measurements. The building and comparison from association networks of each state revealed the variability of the bacterioplanktonic communities involved. However, the associations rather than the lineages seem to distinguish the states. Besides, hub OTUs and associations that we considered as keystones of microbial communities were highlighted. The consistence of these hypotheses will have to be confirmed in the future by the study of additional data sampled

at a broader geographic scale and at a thinner and more regular time. Finally, a global comprehension of the carbon cycle in the ocean will also imply the study of the entire planktonic community, from viruses to metazoans. Our study constitutes a new and necessary step towards linking genes to ecosystems, and towards the use of "ecogenomics sensors" (sensu Ottesen, 2016) to monitor the health of the carbon cycle at a global scale.

3.2.7 Supplementary information

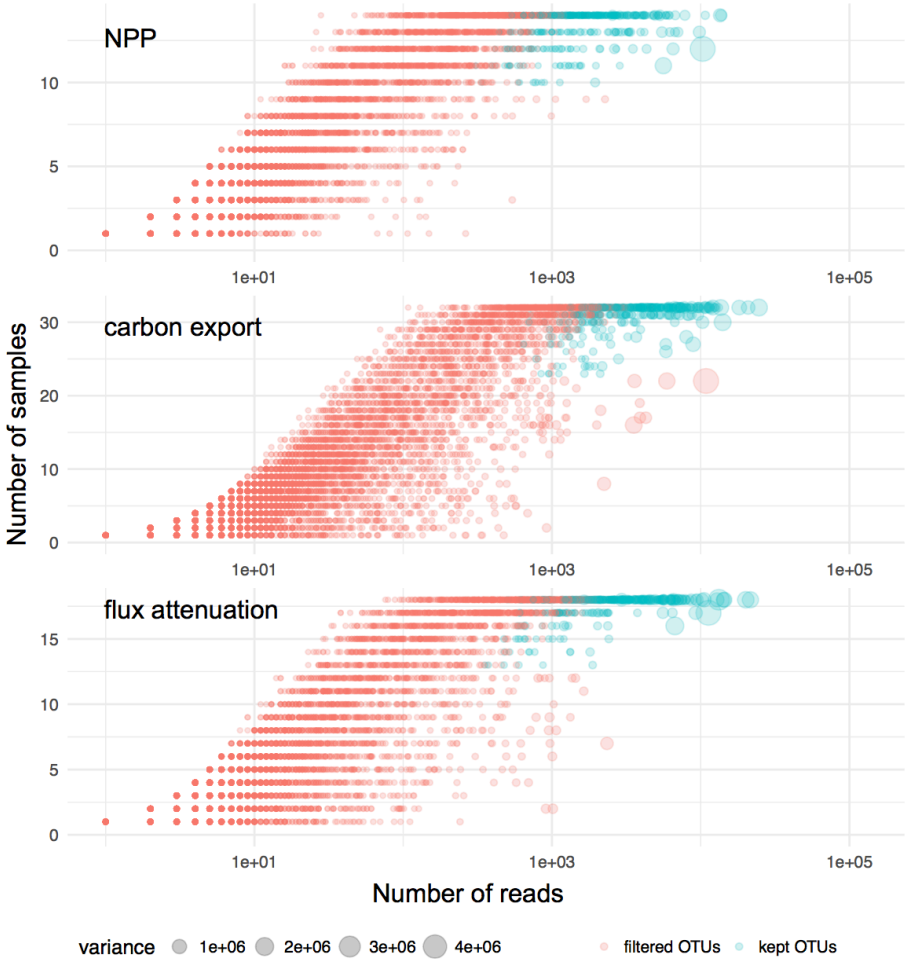


Figure 3.9 - Filtering of OTUs before association network building. On the biplots, each point represents an OTU. They are plotted according to their abundance and the number of samples in which they appear. OTUs kept for association networks building are displayed in turquoise. Discarded OTUs are displayed in red. The points' size translates their abundance's variance. OTUs were filtered (1) on their prevalence (at least 70%) and (2) on the variance of their abundance (20% most variant OTUs).

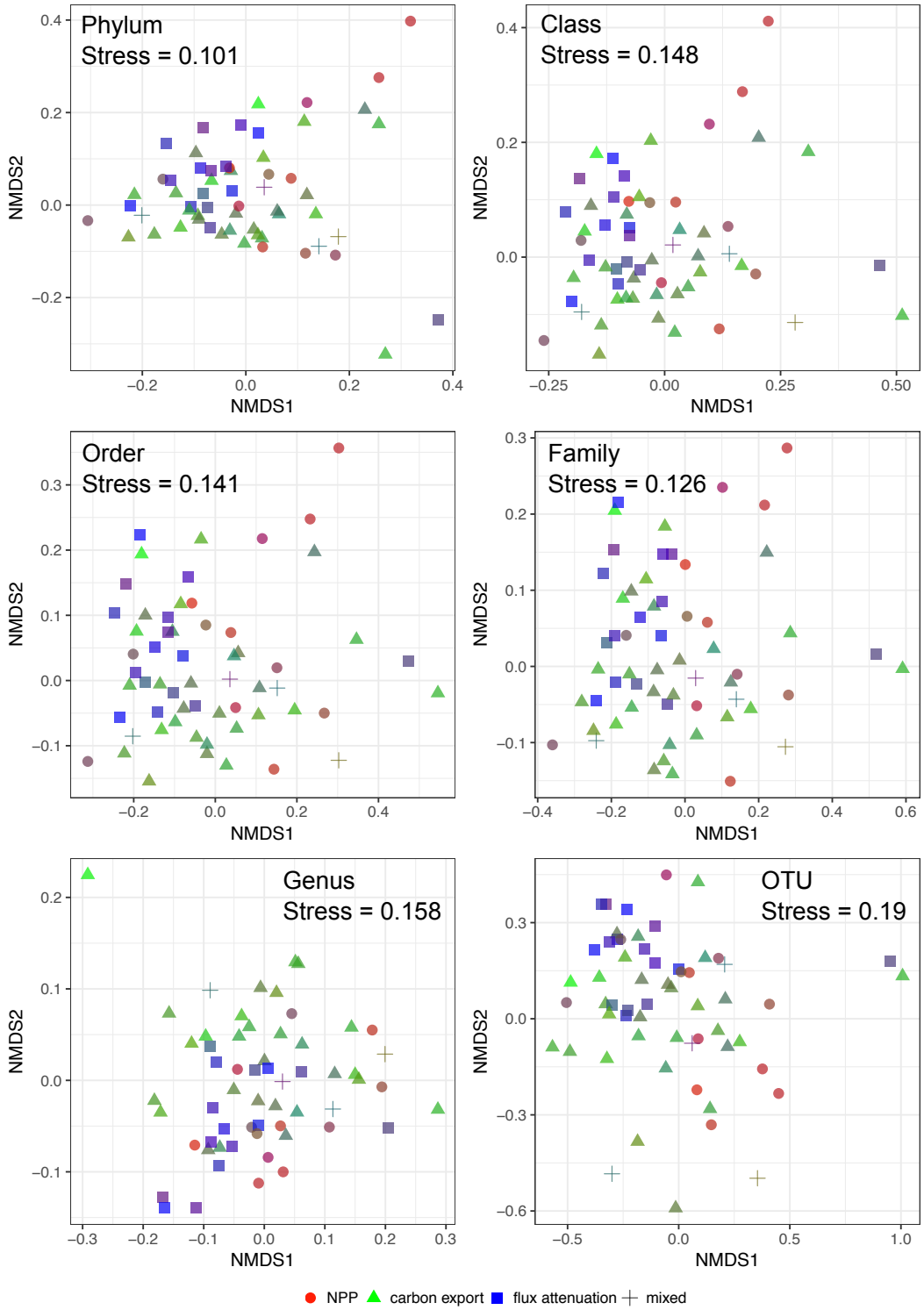


Figure 3.10 - Non-metric multidimensional scaling performed on prokaryotic samples based on 16S $_{mi}$tag relative abundances at different taxonomic levels. Colors refer to the relative contribution of net primary production (red), carbon export (green) and remineralization (blue). Shapes refer to the state each sample belong to (circles for NPP, triangles for CE, squares for FA and crosses for mixed samples). Samples 93SRF and 93DCM were excluded from the representation because they appeared as outliers and made difficult the visualization of other samples on the map.

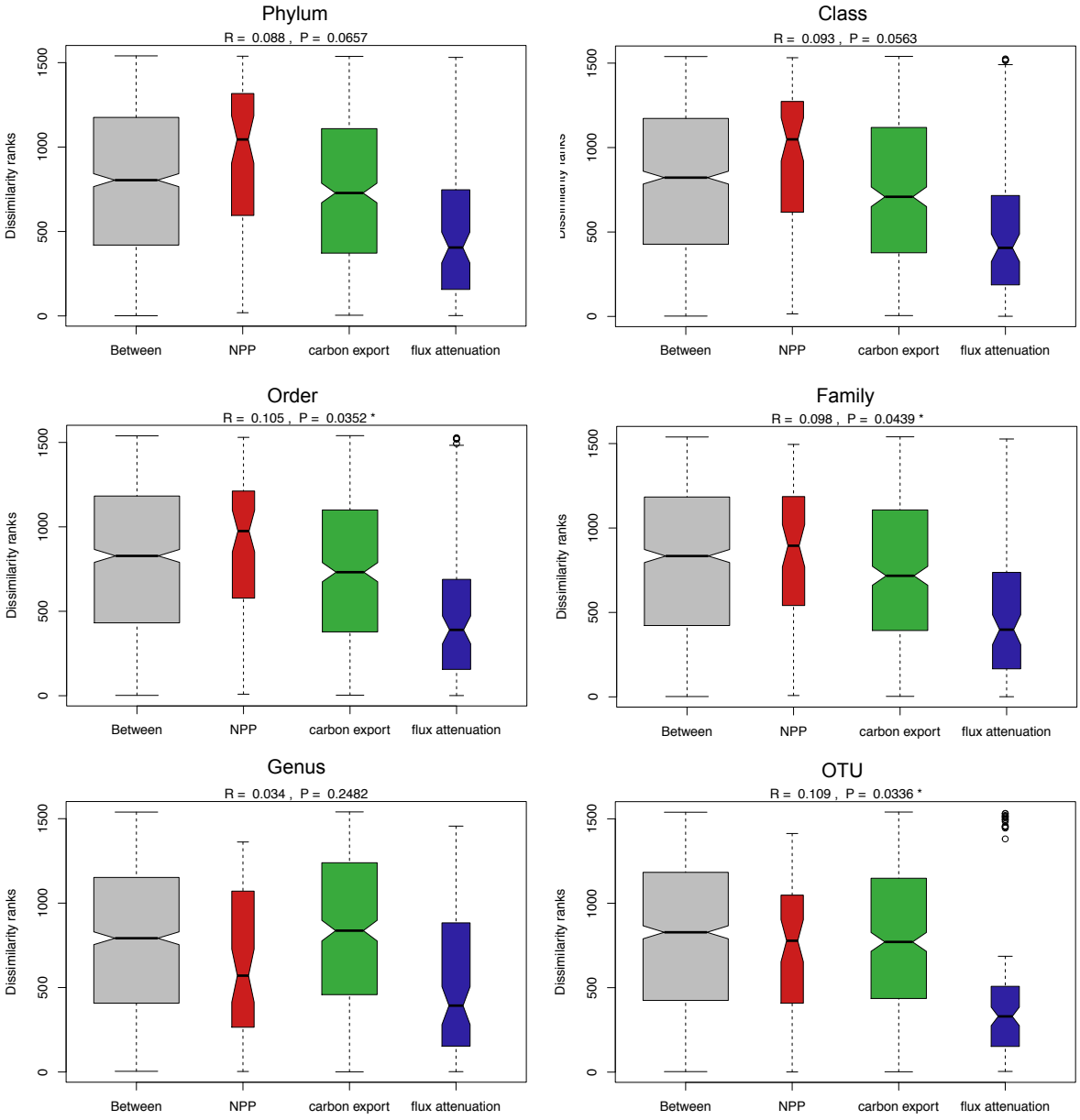


*Figure 3.11 - Analysis of Similarity (ANOSIM) comparing the three states, based on 16S $_{mi}$tag abundances. Analysis has been performed at different taxonomic levels (from the phylum to the OTU levels). P-values are reported and labelled with a * when <0.05.*

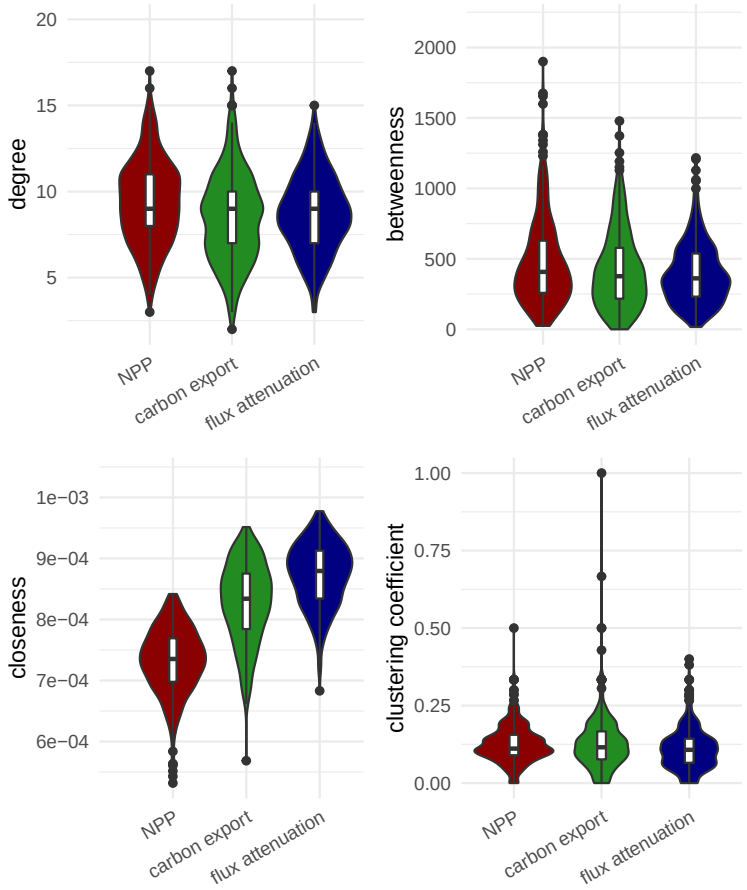
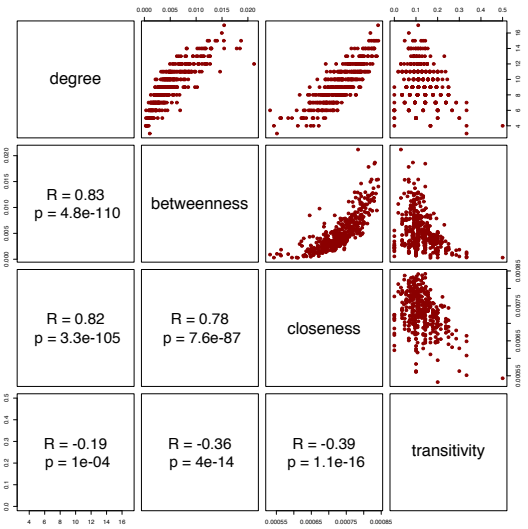


Figure 3.12 - Violin plots of degree, betweenness, closeness and clustering coefficient, showing differences between the networks.

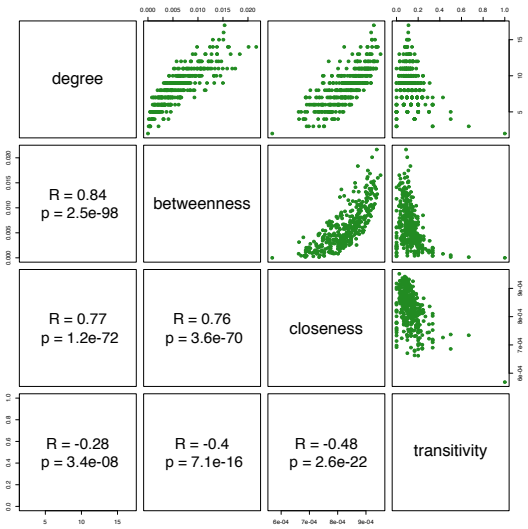
Table 3.3 - Results of Kruskal-Wallis followed by a pairwise Wilcoxon rank sum test comparing nodes attributes between networks. P-values were adjusted with the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

	Kruskal-Wallis tests		**Pairwise Wilcoxon rank sum tests**		
	χ^2	p-value	NPP vs. CE p-value	CE vs. FA p-value	FA vs. NPP p-value
degree	28.38	6.876e-7***	1.5e-6***	0.12046	0.00028***
betweenness	9.7104	7.788e-3**	0.0431*	0.5087	0.0079*
closeness	621.95	<2.2e-16***	<2.2e-16***	<2.2e-16***	<2.2e-16***
clustering coefficient	17.763	1.389e-4***	0.740019	0.00186**	0.00016***

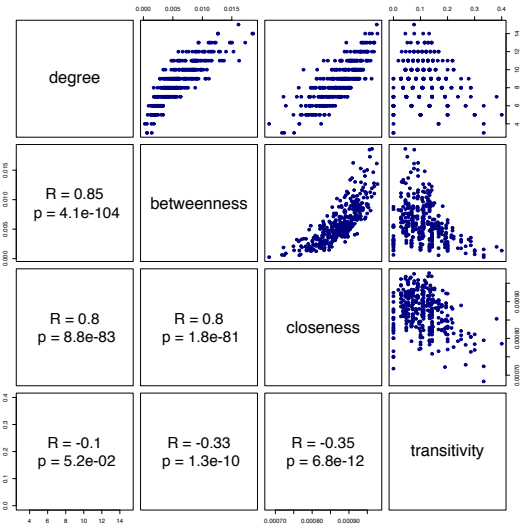
*** p<0.001, ** p<0.01, * p<0.05



(a) Net primary production



(b) Carbon export



(c) Flux attenuation

Figure 3.13 - Scatterplot matrices of centrality metrics measured on the three networks (transitivity = clustering coefficient).

Table 3.4 - Positive edge percentage (PEP) according to the core or specific "nature" of the nodes

edges	group	number	percentage
all	NPP	1401	**70**
	CE	1149	**73**
	FA	1048	**66**
spe—spe	NPP	150	**75**
	CE	41	**71**
	FA	24	**73**
core—core	NPP	484	**85**
	CE	493	**85**
	FA	447	**76**
core—spe	NPP	251	**57**
	CE	92	**50**
	FA	102	**50**

Table 3.5 - Top ten keystone OTUs of the NPP network.

NPP keystone OTUs	degree	betweenness	closeness	clustering coefficient	score
AY664189.1.1285;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	15	1674.96	0.0008319468	0.09	2.75
DQ156396.1.1264;Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II	17	1381.61	0.0008410429	0.11	2.73
FJ826491.1.1449;Bacteria;Cyanobacteria;Chloroplast	14	1654.97	0.0008305648	0.14	2.68
EU802725.1.1445;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus	12	1900.04	0.0007836991	0.03	2.64
EF572127.1.1520;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	14	1599.36	0.0008136697	0.07	2.63
EU010164.1.1471;Bacteria;Proteobacteria;Gammaproteobacteria;Order Incertae Sedis;Family Incertae Sedis;Marinicella	16	1340.70	0.0008190008	0.07	2.62
AY663895.1.1212;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	15	1375.92	0.0008368201	0.05	2.60
HQ673514.1.1438;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	15	1258.94	0.0008417508	0.14	2.54
GU061987.1.1504;Bacteria;Proteobacteria;Skagenf62	14	1382.25	0.0008223684	0.07	2.53
JN018991.1.1360;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;Magnetospira	15	1229.80	0.0008278146	0.10	2.51

Table 3.6 - Top ten keystone OTUs of the CE network.

CE keystone OTUs	degree	betweenness	closeness	clustering coefficient	score
JQ347414.1.1492;Bacteria;Proteobacteria;Gammaproteobacteria;Order Incertae Sedis;Family Incertae Sedis;Marinicella	14	1478.15	0.0009380863	0.09	2.81
DQ156396.1.1264;Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II	14	1372.92	0.0009216590	0.10	2.72
EU394556.1.1501;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	17	1045.71	0.0009276438	0.11	2.68
EU802368.1.1440;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	16	1032.12	0.0009182736	0.11	2.60
GU061386.1.1471;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	14	1091.03	0.0009416196	0.08	2.55
EU802603.1.1489;Bacteria;Proteobacteria;Deltaproteobacteria;SAR324 clade(Marine group B)	15	1029.15	0.0009208103	0.10	2.55
EF572791.1.1441;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	13	1128.09	0.0009514748	0.03	2.53
EU802548.1.1445;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured	13	1252.62	0.0008710801	0.13	2.53
FQ032825.6410.7925;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;NS9 marine group	14	1103.94	0.0009057971	0.10	2.52
EU802670.1.1499;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	15	969.53	0.0009319664	0.11	2.52

Table 3.7 - Top ten keystone OTUs of the FA network.

FA keystone OTUs	degree	betweenness	closeness	clustering coefficient	score
AF098371.1.1444;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	14	1216.29	0.0009624639	0.04	2.92
EU802400.1.1495;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	14	1206.94	0.0009560229	0.08	2.90
FJ745180.1.1374;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	15	1053.63	0.0009756098	0.08	2.86
EF573597.1.1440;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	13	1128.57	0.0009569378	0.05	2.77
AF382132.1.1496;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	13	982.94	0.0009496676	0.09	2.65
DQ009161.1.1974;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 3	13	957.48	0.0009661836	0.06	2.64
AY726846.1.1215;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	13	967.60	0.0009416196	0.12	2.63
EF574651.1.1468;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;S25-593	14	826.42	0.0009775171	0.13	2.61
EU802893.1.1441;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	14	825.72	0.0009442871	0.11	2.58
AACY020398456.2130.3630;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	12	998.44	0.0009319664	0.08	2.57

Table 3.8 - Top ten keystone edges of the NPP network.

Node 1	Node 2	edge betweenness
EU802434.1.1365;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	GQ337163.1.1448;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Deep 1	6147
JN547444.1.1442;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus	JQ226757.1.1346;Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II	5229
EU802434.1.1365;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	AF098371.1.1444;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	4994
EF572127.1.1520;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	GQ337163.1.1448;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Deep 1	4951
EU802534.1.1442;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured	EU802355.1.1369;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SAR116 clade	4749
JN547444.1.1442;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Prochlorococcus	EU802874.1.1450;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured	4336
EU802874.1.1450;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured	DQ009427.1.2076;Bacteria;Deferribacteres;Deferribacteres;Deferribacterales;SAR406 clade(Marine group A)	3929
AY534100.1.1466;Bacteria;Chloroflexi;SAR202 clade	AM747382.1.1283;Bacteria;Cyanobacteria;Chloroplast	3830
DQ156348.18592.20063;Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II	AY534100.1.1466;Bacteria;Chloroflexi;SAR202 clade	3479
EU802844.1.1359;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2	EU799655.1.1454;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;OM75 clade	3462

Table 3.9 - Top ten keystone edges of the CE network.

Node 1	Node 2	edge betweenness
EU802768.1.1494;Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OCS155 marine group	EF573194.1.1484;Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Marinoscillum	6410
GU940897.1.1343;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	EU802768.1.1494;Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OCS155 marine group	6077
EF573115.1.1498;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	GQ349357.1.1386;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;NS5 marine group	5045
AY664083.1.1206;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	EF573194.1.1484;Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Marinoscillum	4970
AY664083.1.1206;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	GU061825.1.1499;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	4944
EF573115.1.1498;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	FJ826317.1.1453;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae;Magnetospira	4912
EF572745.1.1469;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;S25-593	JQ222710.1.1336;Archaea;Euryarchaeota;Thermoplasmata;Thermoplasmatales;Marine Group II	4245
GU234848.1.1327;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	GQ349357.1.1386;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;NS5 marine group	4165
EU802434.1.1365;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	FJ745176.1.1372;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Cryomorphaceae;Owenweeksia	3942
EF572745.1.1469;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;S25-593	FJ745172.1.1390;Bacteria;Verrucomicrobia;Opitutae;Puniceicoccales;Puniceicoccaceae;marine group	3926

Table 3.10 - Top ten keystone edges of the FA network.

Node 1	Node 2	edge betweenness
DQ009253.1.1978;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	AF098371.1.1444;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	5447
EF571982.1.1487;Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;OCS155 marine group	EF572587.1.1496;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	4946
EF572587.1.1496;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	AACY020398456.2130.3630;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	4754
EU800517.1.1373;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	DQ009253.1.1978;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	4378
EU800517.1.1373;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	FJ745032.1.1322;Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;uncultured	4310
EU802684.1.1366;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	AF098371.1.1444;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	4281
JX017005.1.1403;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 2	FJ745032.1.1322;Bacteria;Actinobacteria;Acidimicrobiia;Acidimicrobiales;uncultured	4155
EU802434.1.1365;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade	AF098371.1.1444;Bacteria;Cyanobacteria;Cyanobacteria;SubsectionI;FamilyI;Synechococcus	4144
EU802894.1.1450;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;SAR116 clade	EU802400.1.1495;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	3799
EU805254.1.1438;Bacteria;Proteobacteria;Alphaproteobacteria;SAR11 clade;Surface 1	FJ745180.1.1374;Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;SAR86 clade	3735

Chapter 4

Random forest-based estimates of the biological pump processes from meta-omics

The goal of this chapter is to test if statistical learning methods are possible tools to predict from meta-omics the states of the biological carbon pump, and to highlight environmental biomarkers.

The chapter starts with an overview of statistical learning describing the basic terminology and concepts, mainly based on the books *An Introduction to Statistical Learning* by James et al. (2013) and *Introduction au Machine Learning* by Azencott (2018).

The chapter continues with a draft manuscript of the work I initiated with Damien Eveillard, Associate Professor at the Laboratoire des Sciences du Numérique de Nantes (LS2N), France. This work has been led in collaboration with Marie Soret, a Master 2 student from Sorbonne Université, Paris (Master "Image et Son pour les systèmes intelligents"), that Lucie Bittner and I supervised for a 6 months internship from April to September 2018. We introduced Marie to high-throughput sequencing, marine metagenomics and to the problematic of the biological carbon pump. She brought her expertise in machine learning and in MATLAB coding. I defined the different steps of the study: (1) tests with different machine learning algorithms, (2) choice of the best algorithm, (3) run and optimisation. Marie Soret implemented these steps and I complemented her analyses during these past months. I conclude the chapter with a discussion on our current results, on pitfalls and perspectives, including potential improvements for futures studies.

4.1 Introduction to machine learning techniques

Machine learning refers to the ability, for a computer program, to learn without being programmed. Basically, it consists in searching for a predictive function based on data. The data used by the algorithms to learn are of two types: the *input* data (that consists in a data matrix of samples, also called observations or examples[1] for which predictors, also called features, attributes or simply variables[2], are available), which are used to predict, and the *output* data (i.e. the labels), which are the variables we want to predict (figure 4.2a). Machine learning is used in a wide range of subjects, for example to set up anti-spam filters, to recommend books, movies or other products adapted to our tastes, to identify faces on pictures and to diagnose diseases. Machine learning problems are divided into two different categories: the *supervised* and *unsupervised* learning (figure 4.1). In the case of supervised learning, the value of the variable to predict is known. The goal of supervised learning is to make predictions from a *labeled* set of observations, in order to predict the label of upcoming unlabeled observations. The labels can be *classes* or *real numbers*. In the first case, the problem consists in *classifying* the observations while in the second case, it is a *regression* problem. From a usually large number of predictors, the *training* consists in identifying the ones that explain the best the labels of the training observations. The model trained on a subset of possible situations is then used to make predictions on upcoming observations. In the case of unsupervised learning, data are not labeled. The problem is not to predict but rather to better understand the data by finding underlying structures in the data or extracting groups of observations displaying common characteristics. When labeling the whole training set requires too much time and efforts, a third intermediate category can be used: the *semi-supervised* machine learning, which learns from a partially labeled training set.

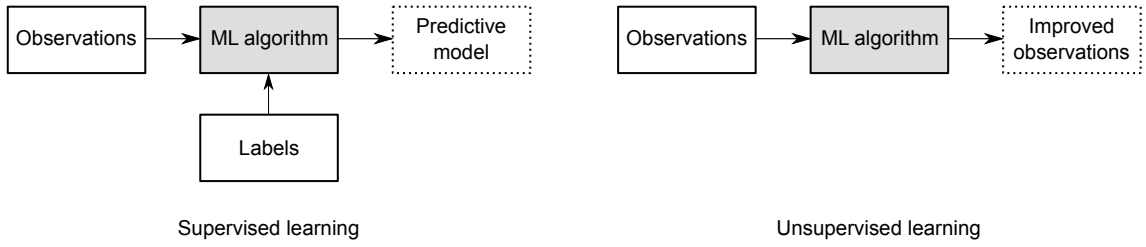


Figure 4.1 - Conceptual diagram of supervised and unsupervised machine learning algorithms, modified from Azencott (2018).

As suggested in the previous paragraph, in the case of supervised machine learning the data set is divided into two subsets: the *training* set and the *test* set (figure 4.2b). The

[1] In this chapter, we will prefer the term "observation" to "sample" and "example" but all terms are synonyms.

[2] Similarly, we will prefer the term "predictor".

training set is used to train the machine learning model. The objective of the training step is to search for a model whose predictions are as close as possible to the true labels. In other words, it means minimizing the prediction error on the training set (i.e. the *empirical error*). However, minimizing the empirical error does not ensure that the error will be minimized on all possible data (i.e. the *generalization error*). Indeed, in some cases, the model may be overfitted and the error on upcoming observations not used for training will be underestimated. Yet, the interest of a supervised machine learning algorithm is to perform well on unknown observations. To evaluate a model, it is thus essential to use labeled observations that were not used to train it. The simplest way to achieve this goal is to keep a part of the observations to evaluate the model: this is the *test* set.

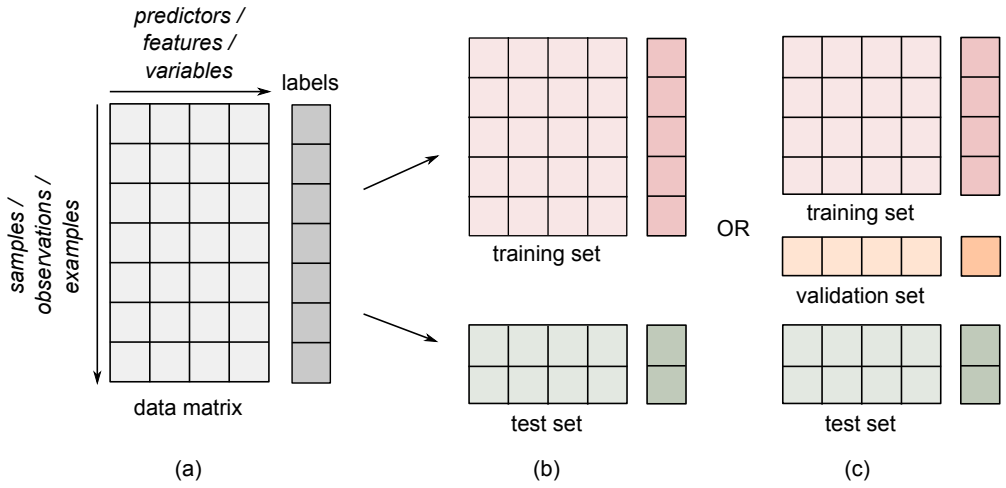


Figure 4.2 - Organisation of data in a supervised learning problem. The data consist in a matrix of observations and predictors and a label vector (a). It is divided in a training set, that is used in the training step and a test set, that is used to evaluate the performances of the model (b). To choose between several models, a validation test is created in order to keep test observations on which the generalization error can be calculated once the model has been chosen (c).

To evaluate the predictive performances of a supervised model, different criteria can be used depending on whether the situation is a *classification* or a *regression problem*. In the case of a classification problem, the prediction error is assessed by calculating the *classification error rate* (i.e. the percentage of misclassified test observations), which is the complementary to 1 of the *accuracy* (i.e. the percentage of correctly classified test observations). However, not all errors are equal. Let's take the example of a disease diagnosis: a false positive, that may be disconfirmed by further analysis, is preferable to a false negative, where the disease is not detected and won't be treated. Consequently, the performances of a classification model are often summarized in a *confusion matrix* (figure 4.3) from which numerous evaluation criteria may be derived, such as the *precision* (proportion of correct positive predictions among all positive predictions, i.e. TP/(TP+FP)), the *sensitivity* (the proportion of correctly identified positive predictions, i.e. TP/(TP+FN)) and

the *specificity* (the proportion of correctly identified negative predictions, i.e. TN/(FP+TN)).

		Actual class	
		0	1
Predicted class	0	true negatives (TN)	false negatives (FN)
	1	false positives (FP)	true positives (TP)

Figure 4.3 - Example of a confusion matrix for binary classification. Here the classes are 0 and 1. If we take the example of disease diagnosis, 0 may be the healthy patients while 1 may be the ill patients. True positives (TN) are correctly classified positive observations; false positives (FP) are negative observations misclassified as positive; false negative (FN) are positive observations incorrectly classified as negative and true negatives (TN) are correctly classified positive observations.

In regression problems, it is delicate to say whether a prediction is correct or not because of numerical imprecision. Therefore, we prefer to quantify the performance of a model as a function of the distances between the true and predicted values. The most commonly used measure is the *mean squared error* (MSE), which is the sum of the squared differences between the true (y_i) and predicted ($\hat{y}_i$) values, divided by the number of test observations (n) used to compute the error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.1}$$

Other criteria may be used, such as the *root mean squared error* (RMSE) which allows for measuring the error in the same units of the predicted variable.

In the case where one would like to choose between a set of models, the initial data matrix may be divided in a third set of observations: the *validation* set (figure 4.2c). Of course, we could calculate the error on the test set and then choose the one that has the smallest error. However, the test set would no longer represent an independent set on which we could assess the generalization error. Therefore, the solution is to split the initial data set in three parts: the training set on which the different algorithms are trained, the validation set on which the models are evaluated to select the best one and the test set on which we evaluate the generalization error. To ensure the representativeness of the training and validation sets, several procedures have been conceived. Classically, we use the *cross-validation* that consists in dividing the data set several times and compute the mean of the obtained results. There exist several procedures of cross-validation. The leave-one-out cross validation (LOOCV) uses single observations for the validation set and the remaining observations make up the training set. The k-fold cross-validation (k-fold CV) is an alternative to LOOCV, described in section 4.2.3.2. The evaluation of a model can also be achieved with a bootstrap, a procedure that creates a given number of

sets, obtained by resampling data with replacement. The model is trained on each of the bootstrapped sets and is evaluated on the remaining observations.

Predictive machine learning techniques encompass a number of methods. The simplest and oldest ones are the *linear* and *logistic regressions*, which is based on the method of least squares (i.e. the minimization of the squared differences between the actual and predicted values). In the same period was developed the *linear discriminant analysis* by Fisher (1936). By the end of the 1970's, non-linear methods (i.e. methods that are able to predict non-linear relationships between predictors and response variables), that had not been addressed yet due to computational limitations, began to be developed, such as *regression* and *classification trees*, which led to the development of random forests that will presented in section 4.2.3.3. Another type of algorithm has been at the origin of the numerous recent successes of artificial intelligence: the artificial neural networks, initially created to model the information processing by biological neuronal networks. This PhD manuscript will not describe all these methods in detail, but will rather briefly present the ones that have been used for the following research article of section 4.2, with a special emphasize on random forests.

4.2 Article 3 (Benoiston et al., in prep.): Machine learning and meta-omics: are we ready to predict ecosystem processes from omics?

Target journals for the following draft manuscript: Biogeosciences, Journal of Geophysical Research, Frontiers in Marine Science, Journal of Marine Systems

Machine learning and environmental genomics: are we ready to predict ecosystem processes from omics?

Benoiston A.S.[1,2], Soret M.[1], Eveillard D.[3], Guidi L.[2], Bittner, L.[1]

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, Sorbonne Université, CNRS, EPHE, Université des Antilles, CP 50, 57 rue Cuvier, 75005 Paris, France.

[2]Sorbonne Université, CNRS, Laboratoire d'océanographie de Villefranche, LOV, 06230 Villefranche-sur-Mer, France.

[3]LS2N UMR6004 CNRS, Université de Nantes, Centrale Nantes, IMTA, Nantes, France.

4.2.1 Abstract

Global change reshapes biodiversity and ecosystems services. Their variations are often evaluated by using biomonitoring indicators (biotic and abiotic). New generation sequencing techniques now provide relative abundance of microbial lineages and functions from environmental samples. Machine learning could be used to identify sets of environmental sequences as bioindicators. By using meta-omics and environmental parameters from the *Tara* Oceans expeditions, we propose the first study trying to predict biogeochemical states from biological abundances derived from environmental DNA. A prokaryotic metabarcoding data set was analyzed with random forests (RF), a machine learning technique that has proven his efficiency in research and engineering. The relevance of the predictions are discussed, and a list of bioindicators / predictors (here Operational Taxonomic Units) is proposed.

4.2.2 Introduction

In the context of a rapidly changing ocean, it becomes urgent to increase our capacities of observation and prediction (Claustre et al., 2009; Gruber et al., 2010; Sauzède et al., 2017). It is necessary to predict future changes in the carbon cycle, particularly in the biological carbon pump (Siegel et al., 2016), whose modifications due to global warming and ocean acidification are uncertain (Passow and Carlson, 2012). Marine plankton play an essential role in the biological carbon pump which contributes to the ocean's capacity to buffer anthropogenic carbon dioxide emissions (Ciais et al., 2013; Khatiwala et al., 2013; DeVries, 2014), and changes in their biodiversity, abundance and biogeography may strongly impact oceanic ecosystems. Therefore, monitoring planktonic populations is of foremost importance to predict future changes occurring in the ocean. Current biogeochemical models often include plankton as large "boxes" (e.g. NPZ (nutrient, phytoplankton, zooplankton) models, Franks, 2002), although improvements have been made by including, for example, plankton functional types (Quéré et al., 2005). Beyond the hypothetical inclusion of a more realistic view of planktonic diversity in these models, a better understanding of the link between plankton and biogeochemistry requires to identify bioindicators of the health status of the ocean.

Among the available techniques to identify these bioindicators are machine learning techniques, also called statistical learning (James et al., 2013). Since the last decade, they are increasingly applied in many science fields. For example in ecology, random forests algorithms have been used to predict land cover from satellite data (e.g. Immitzer et al., 2012; Rodriguez-Galiano et al., 2012). In oceanography, estimates of water-column nu-

trient concentrations and carbonate system parameters were predicted from a set of *in situ* measurements (e.g. temperature, salinity, hydrostatic pressure, salinity, latitude and longitude) and remote sensing imagery using neural networks (Sauzède et al., 2017; Friedrich and Oschlies, 2009). Machine learning techniques have also found various applications in genomics (e.g. to recognize specific DNA regions, to assign functional annotations to genes or to understand the mechanisms underlying gene expression, Libbrecht and Noble, 2015). In medicine, the identification of genomic biomarkers from machine learning techniques is growing, especially in the context of precision medicine (Ziegler et al., 2012), enabling diagnosis, prognosis, and selection of targeted therapies. Recently, machine learning was applied on environmental DNA barcoding for biomonitoring purposes (Cordier et al., 2017, 2018) and notably to define biomarkers tracing the origin of ballast water (Gerhard and Gunsch, 2019). To overcome the limitations of morphology-based assessment of biodiversity, Cordier et al. (2017, 2018) demonstrated that supervised machine learning could be used to predict accurate biotic indices (i.e. continuous biological metrics that classify an environment from "poor" to "high" ecological quality, based on taxonomic richness, composition, abundance, and functions) with the advantage of pointing out a top list of sequences, which can be taxonomically assigned or not. All these studies show that machine learning offers promising techniques to investigate genomic, transcriptomic and metabarcoding data.

Benoiston et al. (in prep.) defined three biogeochemical states of the biological carbon pump from estimations of net primary production (NPP), carbon export (CE) and flux attenuation (FA). They demonstrated that microbial association networks corresponding to each of these states had specific interactions and specific keystone Operational Taxonomic Units (OTUs) (i.e. that are highly central in microbial association networks). However, it is unknown whether microbial abundances estimated by metabarcoding (here OTUs) can be good predictors of these biogeochemical states. The goal of the present study is to use supervised machine learning techniques in order: (1) to predict the state of the biological carbon pump from environmental DNA, and (2) to identify biological predictors (here OTUs) that could be used as biomarkers for future biomonitoring of the health state of the pump (Bohan et al., 2017; Cordier et al., 2019). We address these questions by training predictive models upon prokaryotic abundances estimated in the *Tara* Oceans samples and by measuring their performances.

4.2.3 Materials and methods

4.2.3.1 Data set building

Omics and environmental parameters data were retrieved from the *Tara* Oceans expedition (Pesant et al., 2015; Karsenti et al., 2011; Alberti et al., 2017).

For 104 stations and two depths (surface water layer (SUR) and deep chlorophyll maximum (DCM)), we focused on the prokaryote-enriched size fraction filters (i.e. the 0.22-1.6 μm fraction up to station #52 and the 0.22-3 μm fraction from station #56, Sunagawa et al., 2015). In the Illumina-sequenced metagenomes, 16S ribosomal RNA gene fragments were directly identified (so called 16S $_{mi}$tags, Logares et al., 2014) and mapped to operational taxonomic units (OTUs) based on clustering of reference sequences from the SILVA database (Quast et al., 2013) at 97% sequence identity. 16S $_{mi}$tags counts were normalized by the total sum for each sample. The resulting abundance matrix of the OTUs (26 281 OTUs) in the 104 samples (a sample corresponds to a community sampled at a station and a given depth) can be downloaded at `https://figshare.com/s/23798e4046a2c21a9103`.

From the analysis of *Tara* Oceans environmental data, Benoiston et al. (in prep.) defined three biogeochemical states of the biological carbon pump. The *Tara* Oceans samples were classified in the net primary production (NPP), in the carbon export (CE) or in the flux attenuation state (FA) (`https://figshare.com/s/f67bcb072aea125039d3`; e.g. samples in which NPP absolute values is relatively high compared to CE and FA were classified in the NPP state). The NPP, CE and FA classes involves 14, 32 and 18 samples, respectively. We removed the samples for which the classes were overlapping (see figure 3.4 of Benoiston et al., in prep.), thus our final set for this prediction study involves 12 NPP, 15 CE and 29 FA samples.

4.2.3.2 Tests and algorithms comparison

Comparing the performances of different machine learning methods

The performances of six supervised machine learning algorithms were tested in order to choose the best method as the one to predict the states of the biological carbon pump. The tested algorithms are single decision trees (`ftree`), pattern recognition network (`fpat`), discriminant analysis (`fdiscr`), discriminant analysis ensemble (`fdens`), random forests (`fbag`) and forests trained on RUSBoost algorithm (`frus`). Here the input data of the algorithms is the OTUs abundance matrix, in which the samples correspond to the observations and the OTUs to the predictors. The goal is to assign the observations to the correct class (NPP, CE or FA). It is a classification problem.

Principle of the six machine learning methods tested

A **decision tree** consists of nodes and branches. At each *internal node* (including the first one, that is the *root node*), the observations are segmented in two subsets according to a splitting rule until they reach a *terminal node* (or *leaf*) where the decision to assign the observation to a class is made (see figure 4.4). MATLAB function: `fitctree`

Linear discriminant analysis (LDA) is a method used to find linear combinations of predictors (here OTUs) that allow to separates observations in classes. LDA models the distribution of the predictors separately in each of the classes and then uses Bayes' theorem to estimate the probability of an observation to belong to a given class, knowing the distribution of the predictors. MATLAB function: `fitcdiscr` and `fitcensemble` for ensembles of LDA

Pattern recognition networks are feedforward artificial neural networks. These networks are based on multiple layers (i.e. input, hidden and output layers) composed of neurons that are basically transfer functions. The neurons of one layer are interconnected with the neurons of the subsequent layer. The weights of the connections are adjusted by minimizing a cost function during the training phase through backpropagation. MATLAB function: `pattnet`

Random UnderSampling Boosting (RUSBoost) is a tree-based method (Seiffert et al., 2008) that works in a similar way to bagging (see section 4.2.3.3) as it uses trees as building blocks to predict output variables. It is especially effective at classifying imbalanced data. Indeed, as its name implies, classes with more observations are under sampled: to grow each tree of the ensemble, N (the number of observations in the class with the fewest observations) observations of every class are used. MATLAB function: `fitcensemble`, method: `RUSBoost`

Random forests (RF) is a tree-based method, which is detailed in section 4.2.3.3. MATLAB function: `fitcensemble`, method: `bag`.

For each of the six supervised methods, the set of observations was randomly divided in three subsets: a *training set* used to fit the model (training step), a *validation set* used to estimate the actual performances of the models and a *test set*. As our data set is limited to relatively few samples (59 samples), we used the *k-fold cross validation* (*k*-fold CV). *k*-fold CV consists in dividing the data set into *k* subsets, or *folds*, of approximately equal size. One of the *k* folds is used as a validation set and the *k-1* others constitute the training set. The error rate is calculated on the observations of the validation set. The algorithm is ran *k* times, each time a different *fold* is used as the validation set. The process results in *k* estimates of the validation error rate that are averaged to obtain the *k*-fold CV estimate (James et al., 2013). For each of the six algorithms, the prior distribution of the classes

was set to be uniform to overcome the issues due to the statistical over-representation of the CE class.

4.2.3.3 Random forest development

Principle of random forests

Random forests (RF) are a tree-based method that can be applied to both classification and regression problems (Breiman, 2001). RF are ensembles of decision trees. A decision tree consists of nodes (a node corresponds to a subset of observations or to a unique observation) and branches (a branch corresponds to a classification decision) (figure 4.4). The RF algorithm performs in two steps: *training* (using a training set) and *prediction* (using a validation or test data set).

First, the ***training* step** consists in creating N decision trees (formed by nodes and branches). At each node (white nodes in figure 4.4), the observations (here the *Tara* Oceans samples) are segmented according to a splitting rule until they reach a terminal node (a *leaf* corresponding to a green/red/blue node in figure 4.4) where is made the decision to assign the observation(s) to a class (here NPP, CE or FA). Only a random set of predictors (here the OTUs) is available at each split and the predictor that discriminates the best the classes is chosen. Usually, RF use the Gini Index as a measure of best split criterion (Breiman et al., 1984). The best predictor is chosen by minimizing the heterogeneity of the child nodes that result from the test (because we want the observations belonging to distinct classes to be at best discriminated by the test). Each tree is built independently on a random bootstrapped training subset of observations with replacement: this technique is called *bootstrap aggregating* or *bagging*. In our study, for each of the N trees, 2/3 of the observations were randomly sampled to build the training subset.

Second, the RF algorithm performs the ***prediction* step**. The observations not included in the training set are called *out-of-bag* (OOB) observations. OOB observations are used to evaluate the performance of the RF: each of the OOB observations is run on the decision trees (built during the *training* step) for which it is OOB, and thus classified according to the training structure (path of grey nodes leading to a colored leaf in figure 4.4). As a result, for each observation X, which could have appeared several times in the OOB observations (but not necessarily in all OOB subsets), a majority vote is applied in order to assign it to a final class (e.g. an observation X might be classified once in the NPP class, 0 time in the CE class and 10 times in the FA class so as a result with a majority vote, it will be classified in the FA class). The proportion of misclassifications among the total number of OOB observations (the OOB predictions are compared to their true classes)

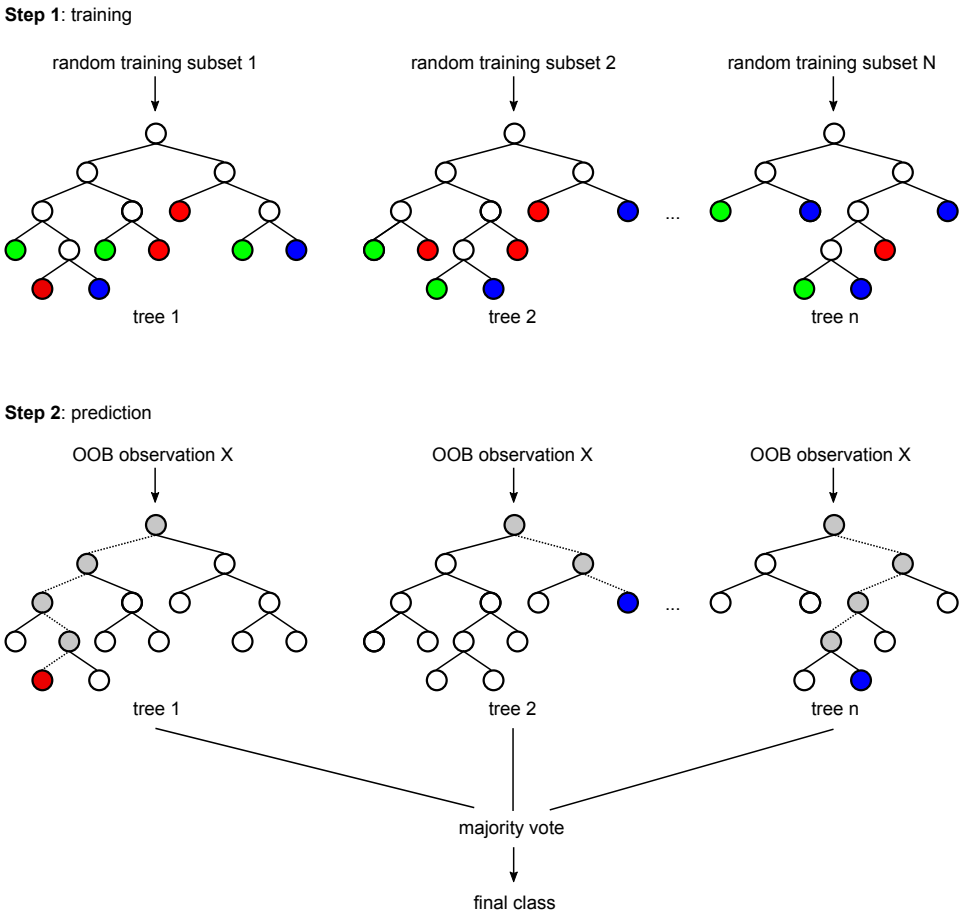


Figure 4.4 - Conceptual diagram of the random forest algorithm. N trees are grown from N random bootstrapped subsets of observations (samples) (step 1: training). At each split, a random subset of predictors (OTUs) is selected and the predictor that discriminate the classes the best is chosen. Each of the remaining observations that were not used to fit a given tree (i.e. the out-of-bag OOB observations) are one after the other used for prediction (step 2). For an observation X, for each of the trees in which this sample was OOB at step 1, a class is assigned following the decision trees. Finally, the majority vote is used to infer a final class to the observation. White nodes correspond to subsets of observations, green/red/blue leaves correspond to a final class, gray nodes correspond to the decision paths followed by an OOB sample leading to its classification during step 2.

gives an estimate of the prediction error rate of the model. In return, the accuracy of the model (percentage of well classified observations) can be calculated from the OOB observations.

Optimization of random forests

Two parameters were set to improve the RF prediction: the number of trees in each forest (`ntree`) and the number of predictors to try at each split (`mtry`). Besides, to avoid the problem arising from the over-representation of the CE class over the NPP and FA classes, we enforced the prior probability of each class to represent a uniform distribution.

Predictors importance measure

RF allows to measure the importance of the predictors for the classification process. This

feature is particularly interesting if the number of predictors is very high. In this way, it is possible to assess how each predictor influences the model and a selection of the most interesting predictors according to the model is possible. The Gini Index and the OOB error rate were used to assess the importance of predictors. The measure of the importance of a given predictor is computed by permuting it while leaving all the others unchanged, and measuring the variation in the OOB error rate or the Gini Index that occurred.

The comparison of the different statistical learning algorithms tested was performed with MATLAB while the final predictions with the RF were done with the R package `randomForest` (R version 3.5.2, `randomForest` version 4.6-14).

4.2.4 Results

4.2.4.1 Tests to select a machine learning algorithm

Six supervised machine learning algorithms were tested on our data set for a classification purpose. For each algorithm, the accuracy (percentage of well classified observations from the validation set) was computed for 10 runs (i.e. 10 models were built for each algorithm). The accuracy of the models is available on figure 4.5, showing that, on average, RF gave the best recognition rate.

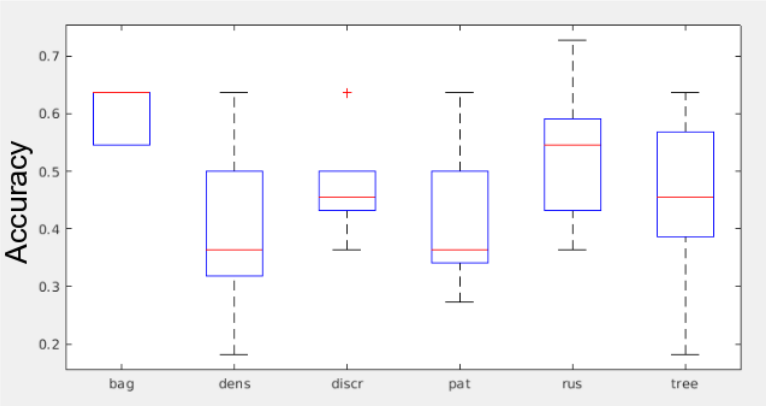


Figure 4.5 - Predictive performances (accuracy) of the six machine learning algorithms tested for the prediction of the biological carbon pump states (classes). Here the comparison of the distributions (summarized by boxplots) of the accuracy showed, on average, a better performance with RF. `bag`*: random forest,* `dens`*: discriminant analysis ensemble,* `discr`*: discriminant analysis,* `pat`*: pattern recognition network,* `rus`*: forest train on RUSBoost algorithm,* `tree`*: single decision tree.*

4.2.4.2 Effect of the number of trees and the number of tested predictors at each split on the model's accuracy

To assess the optimal number of trees in the forest (`ntree`) and the number of tested predictors at each split (`mtry`), RF were created with variable number of trees, ranging from 1 to 2,000 and variable number of predictors. As suggested by Breiman (2001), we tried the default value of `mtry` implemented in the function `randomForest()` from the `randomForest` R package (i.e. 4), half of the default and twice the default. We also tested higher values as it may give better performance in the case of having a large number of predictors but expecting only very few to be "important" (Liaw and Wiener, 2001). To sum up, the values 2, 4, 8, 16, 32, 64 and 128 were tested for `mtry`. As a result, 14,000 different RF prediction models were tested for the classification. Figure 4.6 shows the OOB error rate as a function of `ntree`. The OOB error rate is very high for very low values of `ntree`, and as `ntree` increases until around 100, the OOB error rate sharply drops, regardless the value of `mtry`. The OOB error rate then continues to be highly variable until it becomes more stable from approximately 1,600 trees. The common strategy to select `ntree` is to take the value from which the OOB error rate stabilizes (e.g. Immitzer et al., 2012). For this reason, we set `ntree` to 1,600. The tested values of `mtry` yield variable classification errors, especially before reaching the stabilization point. From this point, the OOB error rate converges around 48% for all the values of `mtry`. Thus, the value of `mtry` was observed to have little effect on the OOB error rate. However, running the RF with a number of tested predictors at each split equal to 128 seemed to give slightly lower and more stable OOB error rates. This value was therefore used to train the final model.

4.2.4.3 Predictive performances of RF for the prediction of the states of biological carbon pump

We evaluated the performances of the RF to predict the state of the biological carbon pump (i.e. the classes) by running the RF 10 times. With the parameters used for the analysis (i.e. `ntree` = 1,600 and `mtry` = 128), the median of the OOB error rate is 44.64%, meaning that only 55.36% of the observations were correctly classified. The confusion matrix (figure 4.7) and the class error show that, beyond this global predictive performance, the observations belonging to the CE class are better classified than the others. 75.86% of the CE observations are well classified while only 46.67% of the FA observations and 16.67% of the NPP observations are correctly classified. As we tuned the distribution of the classes to be uniform, other factors must come into play to give these imbalanced class errors. To disentangle the confusion between the classes, we performed the classification

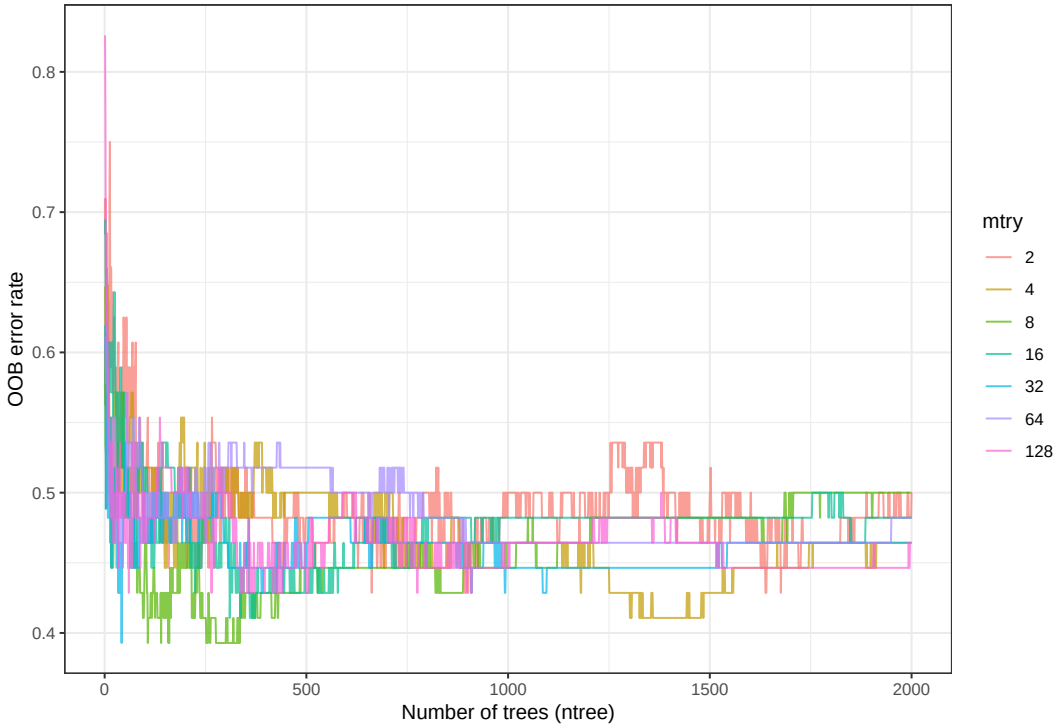


Figure 4.6 - Effect of the number of trees (`ntree`) and random split predictors (`mtry`) on the out-of-bag (OOB) error rate with RF.

on pairs of classes (i.e. the input data included observations of only two classes out of three). Lower OOB rates are observed when running pairwise classifications. When selecting observations belonging to the CE and NPP classes only, the OOB rate drops to 36.59%. It is even smaller when considering the FA and CE classes (22.73%) and the FA and NPP classes (18.52%). The differences between the observed OOB rates indicate that the NPP and CE classes appear to be more often confused than the NPP and FA classes on one hand and the FA and CE classes on the other hand. Indeed, most NPP observations (91.67%) are classified as CE observations. However, fewer CE observations (13.79%) are classified as NPP observations (figure 4.8). A high proportion of FA observations (53.33%) are also mistaken for CE observations.

4.2.4.4 Best predictors identification

The importance of the predictors for classifying samples in the NPP, CE or FA class was assessed with the mean decrease in accuracy and the mean decrease in node heterogeneity measured with the Gini Index. Figure 4.9 shows the contribution of each variable to the classification model generated using the prokaryotic OTU relative abundances. The top 30 best predictors according to the mean decrease in accuracy (values ranging from 1.47e-03 to 4.22e-03) share 16 OTUs with the top 30 best predictors according to the

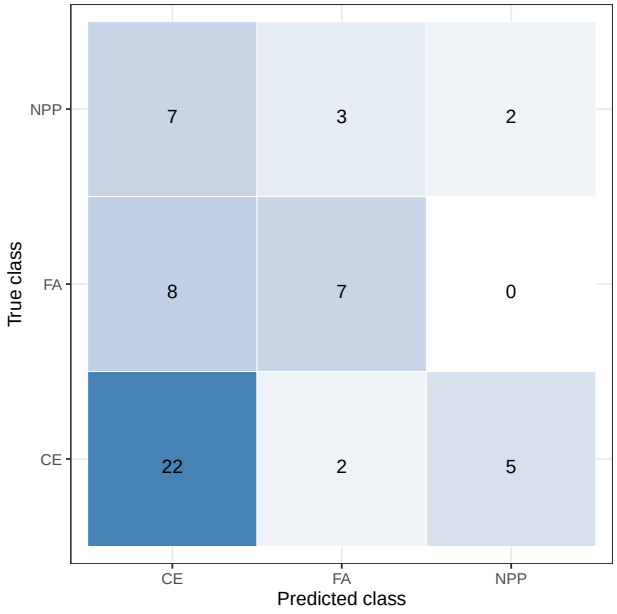


Figure 4.7 - Confusion matrix comparing the RF predictions to the true classes of the OOB observations. Elements on the diagonal of the matrix represent out-of-bag observations whose class was correctly predicted, while off-diagonal elements represent observations that were misclassified. Squares are colored in blue when numbers are high and white when numbers are low. A good classifier is expected to to give a confusion matrix with a blue diagonal (from the left bottom to the right top) and white off-diagonal elements.

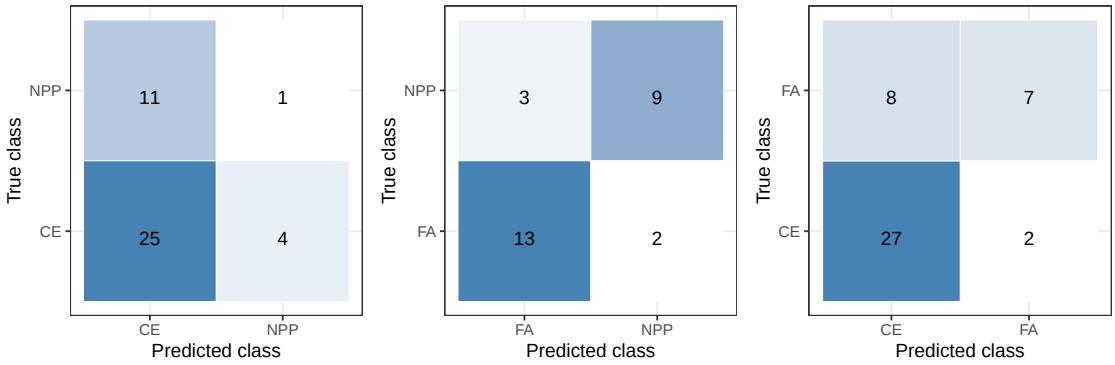


Figure 4.8 - Confusion matrices comparing the RF predictions to the true classes of the OOB observations for pairs of classes (from left to right: CE vs NPP, FA vs NPP and CE vs FA).

mean decrease in the Gini Index (values ranging from 0.135 to 0.066). The most important OTU according to both measures is an Actinobacteria belonging to the family OCS155 Marine Group. Regarding the mean decrease in accuracy, the following OTUs with a high contribution to the RF model are another Actinobacteria of the family OCS155 Marine Group (order Acidimicrobiales) and a Proteobacteria of the family S25-593 (order Ricksettiales). According to the mean decrease in the Gini Index, the two Actinobacteria belonging to the family OCS155 Marine Group detected as the most important by the first measure is also important for prediction.

Overall, almost all of the top 30 predictors according to the mean decrease in accuracy

are bacteria, with 53% belonging to the Proteobacteria phylum, followed by Actinobacteria (13%) and Cyanobacteria (13%). According to the mean decrease in Gini Index, 50% of them are Proteobacteria, 13% are Actinobacteria, 10% are Cyanobacteria and 7% are Chloroflexi. Few of them are annotated to the genus level. However, we can note the presence of the Cyanobacteria *Procholoroccus*, the Alphaproteobacteria *Tateyamaria*, the Gammaproteobacteria *Alteromonas* and the Flavobacteria *Mangrovimonas*.

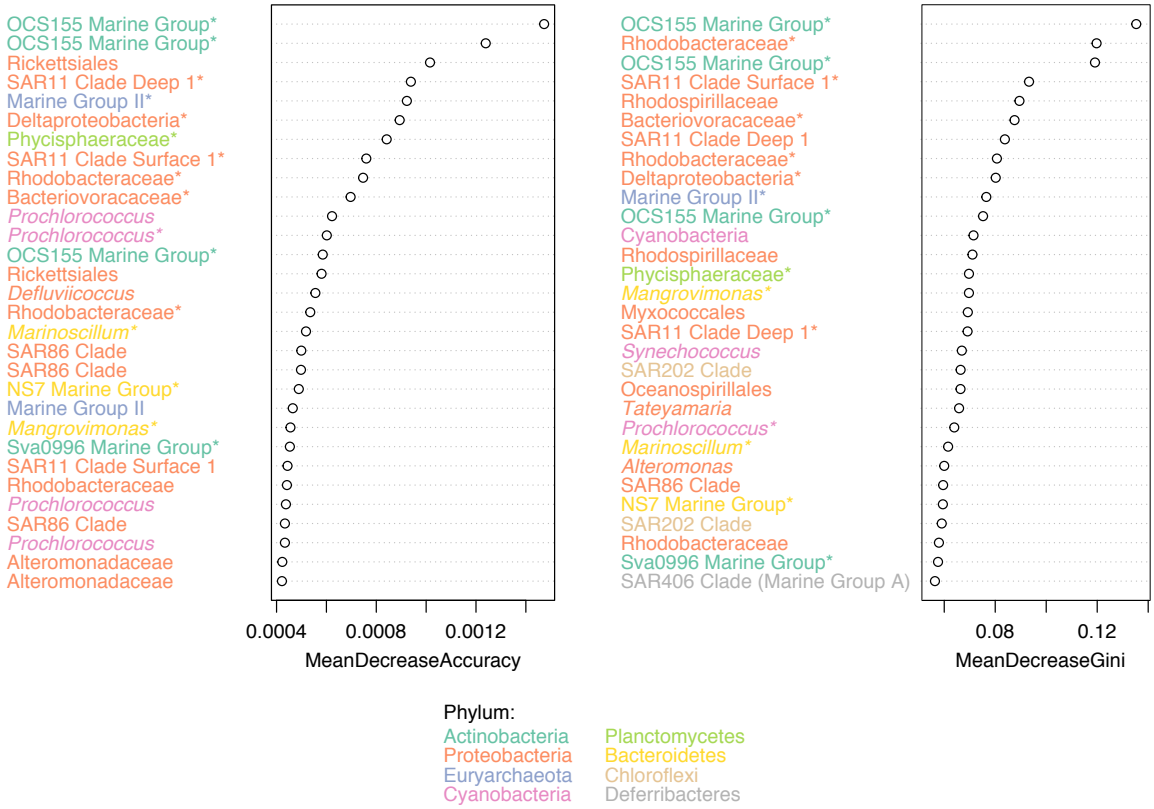


Figure 4.9 - Contribution of the 30 most important OTUs in terms of mean decrease in accuracy and Gini Index. The finest taxonomic annotation is given for each OTU and colors correspond to their phylum, as indicated in the legend. Common OTUs between the top 30 most important OTUs according to both measures are highlighted with an asterisk.

4.2.5 Discussion

In this study, we presented the first evaluation of the performance of machine learning for the classification of ocean samples differentiated on biogeochemical features, based on the analysis of prokaryotic metabarcoding data. The results show that RF outperform five alternative approaches such as single decision trees or discriminant analysis. RF have many advantages over other statistical learning algorithms: it (1) can be used efficiently in large databases that have many more predictors than observations, (2) does not overfit, (3) is robust to noise (i.e. performs well even when most predictors are noise), (4) generates an internal estimate to monitor error (i.e. the OOB error rate), (5) can extract important

predictors acting in the classification, and (6) few parameters are necessary to tune to achieve good performance (i.e. the number of trees in the forest and the number of input variables tried at each split).

Two parameters were tuned to optimize the performance of RF: the number of trees (`ntree`) in the forest and the number of tried predictors at each split (`mtry`). As `ntree` increases, the OOB error rate decreases, which is in accordance with the general trend observed with many data sets (Breiman, 2001). Besides, our results show that the value of `mtry` has little effect on the OOB error rate, which is consistent with the results of Breiman (2001).

Although RF showed better performances compared to other machine learning algorithms, the error rate obtained when classifying observations in the three classes NPP, CE and FA is relatively high (almost half of the observations are misclassified, i.e. 44.64%) compared, for example, to the study of Gerhard and Gunsch (2019), who used environmental DNA to predict the origin of water samples (i.e. ocean, harbor or ballast) and obtained an OOB error rate of 11.94%. Whereas the distribution of the classes was set to be uniform, it appears that an important proportion of observations belonging to the NPP and FA classes are mistaken for the CE class which was initially representing almost half of the observations. In future studies, other parameters could be tuned in order improve the performance of the model. For example, stratified sampling could be a good solution to prevent misclassification due to unbalanced classes. Stratified sampling consists in dividing observations into homogeneous groups before sampling, and sampling an equal number (or another proportion if desired) of representatives of each group to grow each tree. A consequence of stratified sampling is that the most prevalent class is undersampled, but increases the influence of the rare classes. Indeed, as RF uses majority vote to classify observations, unless the rarest classes are easily distinguishable from the most prevalent one, they are unlikely to "win" when running the prediction.

The two classes that appear to be the most confused are the NPP and the CE classes. Beyond the problem of class imbalance, the confusion between the two classes may be due to the fact that net primary production and carbon export are correlated measures (Pearson's R = 0.53, p = 1.93e-07 in our data set). Thus, samples in the NPP class may display high values of carbon export and, likewise, samples in the CE class may display high values of net primary production. As a result, the distinction between these two classes may be unclear. On the contrary, the pairwise classification of NPP and the FA samples appears to give the best OOB error rate, which may be explained by the weaker relationship between the two measures (Pearson's R = -0.16, p = 0.222 in our data set).

We identified the best predictors for classifying the samples into the three biological carbon pump's classes by calculating the mean decrease in accuracy and the mean decrease in the Gini Index. These two measures are often used to select important predictors and possibly to rerun the RF only with these best predictors (selected after a first application of RF). Most OTUs identified in the top 30 predictors belong to the Proteobacteria, Actinobacteria and Cyanobacteria phyla. According to both measures, the OTU that leads to the highest mean decrease is an OTU belonging to the OCS155 Marine Group (phylum Actinobacteria). This group, discovered in the Oregon coastal waters (Rappé et al., 2000) has been found at various places in the ocean but its function and its ecological role is still unknown (Liu et al., 2015a). This bacterial group is not observed among the top 10 keystone OTUs of Benoiston et al. (in prep.). However, an OTU assigned to the family Rhodobacteraceae and identified as an important predictor according to the mean decrease in accuracy (25th most important OTU on figure 4.9) is also amon the top 10 keystone OTUs of the CE microbial association network of Benoiston et al. (in prep.).

The decrease of the mean decrease accuracy or Gini Index for the best predictors is rather smooth. As a result, no OTU seems to clearly distinguish from the others. Besides, the mean decrease in accuracy and in the Gini Index is very low for all OTUs. The one identified as the best predictor by both measures leads to a mean decrease in accuracy of less than 0.0014 and a mean decrease in the Gini Index of less than 0.13, which is very low compared to other results obtained with RF on metagenomic data (e.g. Dinsdale et al., 2013). These low values could be explained by the fact that the relative abundances of many OTUs are probably highly correlated. Indeed, RF selects at each split the predictor that discriminates the best the classes. When a data set has correlated predictors, any of them can be selected as they all lead to similar Gini Indices. Although random selection of predictors at each split may reduce this effect, if many predictors are correlated, the importance of a given OTU may appear to be smaller if many others remove the same amount of node heterogeneity. In our case, selecting uncorrelated predictors may allow to better disentangle good predictors and reduce the complexity of the model. This could be done by computing correlation between relative abundances of OTUs and extracting groups of highly correlated OTUs. This type of analysis is implemented in the method WGCNA (Langfelder and Horvath, 2008), which delineates clusters of highly correlated genes or OTUs and computes a sort of principal component (called "module eigen value") for each cluster. These eigen values could be used as predictors in the RF model and overcome the problems arising from many correlated predictors. Another solution could be to randomly select one OTU among each of these clusters.

4.2.6 Conclusion

This assessment of the use of random forests from metabarcoding data shows that prokaryotic OTUs relative abundances may not be highly efficient to predict biogeochemical states of the biological carbon pump as we defined them. However, we suggest to perform further analyses to confirm our findings such as tuning other parameters of the RF method and pre-selecting uncorrelated predictors. As the three environmental measures (net primary production, carbon export and flux attenuation) used to define the states seem to be more or less strongly correlated, it would also be interesting to apply RF in the case of regression to predict absolute values of the environmental measures. If better performances could be obtained, important predictors could be tested as biomarkers of the processes of the biological carbon pump, and possibly integrated in biogeochemical models to improve predictions.

Chapter 5

Discussion and perspectives

5.1 Summary of the main results

The general objective of this work was to improve our understanding of the biological carbon pump through the use of environmental omic data in the oligotrophic ocean. More precisely, we were interested in how microbial association networks and machine learning techniques could give insights into this process.

Chapter 2 presented a literature review on the **evolution of diatoms and their biogeochemical functions**. In this review, we highlighted that they are an important group of eukaryotic plankton that **contributes to about 20% of total primary production on Earth** and **significantly contribute to the export of carbon** and silicon to the deep ocean, therefore being highly important in the ocean carbon cycle. Moreover, these microalgae may have been active in controlling past climate changes since their diversification and proliferation in the Southern Ocean since the Cretaceaous. In the last part of this chapter, we focused on the **benefit of omics** for the study of the evolutionary history of diatoms and their metabolism.

In **chapter 3**, **biogeochemical states of the biological carbon pump were defined to integrate its three components**: primary production, carbon export and flux attenuation. To our knowledge, this is the first time biogeochemical states are defined in this way. This allowed us to classify samples of the *Tara* Oceans expedition according to these states that correspond to situations where one of the processes is dominating compared to the others. The metabarcoding data associated to these samples give access to the bacterioplankton communities that show variability according to the states. In particular, **significant differences were highlighted at the levels of prokaryotic OTUs, families and orders**. However, the most significant differences appear to come from the properties of microbial association networks that were inferred for each state based on microbial

abundances. Indeed, the **networks display significantly different centrality metrics**, such as the degree, betweenness and closeness centrality, and clustering coefficient. Besides, we highlighted hub OTUs and associations that we considered as keystones of microbial communities. The **highly connected and central OTUs are different in the three networks**, except for one of them (an OTU taxonomically assigned to the archaeal clade Marine Group II). Despite these differences, many keystone OTUs have the same taxonomical assignation, potentially indicating that they represent ecotypes of the same species.

Finally, **chapter 4** focused on testing whether the biogeochemical states of the biological carbon pump could be predicted from the relative abundances of prokaryotic plankton estimated with metabarcoding techniques. After having tested several algorithms, the **random forests appeared to be the most efficient** one for this purpose. However, the assessment of random forests showed that prokaryotic OTUs relative abundances may not be highly efficient to predict biogeochemical states of the biological carbon pump as we defined them. Actually, the main issue we detected is that **the primary production and the carbon export classes are often confused** by the random forests model we built. This may be explained by the fact that **these two processes are tightly linked in many oceanic regions**, therefore preventing from effectively distinguish them, and by the prevalence of the carbon export class. **Potential biomarkers were also highlighted**, based on their importance for predicting the states. Most of them belonged to the phyla Proteobacteria, Actinobacteria and Cyanobacteria. However, these predictors considered independently had a weak impact on the predictive power of the model, probably because of the correlation of their relative abundances. Despite these results, **we provide in the discussion several avenues that could improve the current model**. These include tuning other parameters of the RF method to decrease the influence of the most prevalent class and pre-selecting uncorrelated predictors.

5.2 Limits of microbial association networks inference

During this thesis, I used networks inference methods to deduce microbial associations from metabarcoding data. These methods have been more and more applied the last decade and gave insights into previously unknown microbial interactions and community organization (Röttjers and Faust, 2018) in various environments (e.g. ocean, soil, gut). Microbial networks are valuable tools to visualize microbial interactions and may represent emergent properties of microbial communities (i.e. properties that arise from the connective structure of communities that would not be observed in individual microbes,

Aderem, 2005). The structure of microbial networks may translate habitat preferences (Chaffron et al., 2010) and distinguish hub species, potentially relevant for the stability of the networks (Berry and Widder, 2014). However, in general, interpreting these networks is not straightforward (Vacher et al., 2016; Röttjers and Faust, 2018).

5.2.1 Biases related to high-throughput sequencing data

When inferring networks from high-throughput sequencing data, limitations arising from 16S marker genes may be problematic. These include: low resolution, varying sequencing depth, primer bias, 16S copy number and sparsity (Hong et al., 2009; Louca et al., 2018; Röttjers and Faust, 2018).

Species or strains may be distinguishable with difficulty from marker 16S sequences. Depending on the taxonomic groups considered, the sequence similarity threshold allowing to discriminate them may vary: in some cases, close bacterial species have near identical 16S gene sequences (e.g. Liu et al., 2015b) while in others, the high intraspecies variability of 16S genes overestimates biodiversity (Sun et al., 2013). Then, microbial abundances are often rarefied (i.e. normalized to the same total sum per sample) because of varying sequencing depth according to samples. Consequently, microbial associations are inferred from relative abundances (i.e. compositional data) which are not independent (changes in the relative abundance of one taxon will necessarily influence the relative abundance of the others). This step may lead to spurious associations depending on the inference method chosen (Gloor et al., 2017). Some of them, such as SparCC (Friedman and Alm, 2012) and SPIEC-EASI (Kurtz et al., 2015), use special data transformations that make them are robust to compositionality. Polymerase chain reaction primers bias may also lead to underestimate microbial richness by missing a high number of taxa (Hong et al., 2009) that are consequently not taken into account when inferring microbial associations. Counts of 16S gene sequences are used to estimate bacterial and archaeal taxa abundances in environmental samples. Yet, taxa have varying 16S gene copy number in their genome. As a consequence, taxa with more copies will appear more abundant (Louca et al., 2018). Finally, microbiome data are often sparse (i.e. contain many zeros) because large number of low-abundant taxa are detected in few samples and it is difficult to assess whether this reflects true absences of if taxa are not detected due to sampling or sequencing limitations. In conclusion, future research should continue to improve the qualitative and quantitative reliability of high-throughput data (Vacher et al., 2016).

5.2.2 Limits of inference methods

The most popular methods for inferring microbial networks are based on correlation measures (e.g. Friedman and Alm, 2012; Faust et al., 2012; Schwager et al., 2019). Although it is a useful measure identify apparent interdependencies among many variables, correlation may lead to spurious results regarding microbial associations inference. Correlation may pose problem when applied on sparse matrices because correlation on many zeros may be highly significant, although this issue may be tackled by applying a prevalence filter that removes rare taxa (Berry and Widder, 2014; Röttjers and Faust, 2018). Correlated but indirectly connected taxa may be inferred by these methods (e.g. if taxon A and B are linked by cross-feeding relationships, and that a third taxon C is directly associated to taxon B, then it will be indirectly associated to taxon A). However, in the resulting association network, these associations, represented by edges, won't be distinguishable. Inferred edges may also be caused by species not accounted in the 16S data set or when two taxa are both affected by a same environmental factor. However, inference methods such as SPIEC-EASI (Kurtz et al., 2015) and FlashWeave (Tackmann et al., 2018) were developed to attempt to address this last issue by utilizing the concept of conditional independence, thus reducing the number of uninformative spurious indirect relationships inferred from the data.

Another important aspect of network inference methods to point out is that they are not based on prior knowledge about how microbes interact in reality. Actually, as explained in chapter 3, most methods infer positive associations when taxa co-occur and infer negative associations when they exclude each other. Yet the fact remains that we do not know how most microbes interact and how it can be translated in their abundance patterns. As a consequence, the predictive performances of network inference methods to retrieve associations are tested on simulated data and then compute sets of measures such as sensitivity, precision and AUC (Area Under The Curve) of ROC (Receiver Operating Characteristics) curves (e.g. Kurtz et al., 2015). Simulated data depend on assumptions about microbial populations dynamics. For example, Berry and Widder (2014) simulated microbial communities with generalized Lotka-Volterra dynamics but this model may not reflect the true dynamics of microbes. However, no large-scale experimentally validated microbial ecological network exists, thus it is the only current means to assess the performance of these methods to recover microbial interactions. Experimentally tested interactions in microcosms exist but they are difficult and time-consuming (Vandermeer (1969) and Friedman et al. (2017) are notable exceptions). Thus, the current challenge in the field of microbial network inference is to build a well-resolved empirical network as a

reference (Vacher et al., 2016).

5.2.3 Interpreting network properties from a biological point of view

Thanks to the inference of microbial networks, emergent properties may be detected in microbial communities. However, it is unclear whether the abstract concepts borrowed from graph theory translates experimental observations.

Defining keystone taxa is often an issue when analyzing association networks. Because hub nodes (i.e. nodes that have the highest degree) are connected to a high number of other nodes, they are often considered as potential keystone taxa. Indeed, highly connected nodes of a network have been hypothesized to be important for the survival and stability of an ecosystem: contrary to the removal of random species, the disappearance of well-connected species may lead to a rapid collapse of the entire network (Paine, 1969; Bascompte, 2009). Other types of centrality such as betweenness are used as a proxy for node importance. Betweenness centrality of a node is the fraction of shortest paths from all nodes to all other nodes that pass through the node (Freeman, 1977). Thus, a node with a high betweenness centrality is potentially an intermediary between a high number of nodes couples. Defining the importance of a node on its betwenness assumes that species interact with each other via the shortest path. However, as the way microbes interact is generally unknown, it is difficult to prefer a metric over another to define keystone species and hub species they may not share the same biological implications as keystones (Röttjers and Faust, 2018). Berry and Widder (2014) tested this hypothesis by simulating multi-species microbial communities and inferring co-occurrence networks from simulated abundances, which showed that some nodes topological features (particularly high degree, closeness and clustering coefficient) could be used as predictors to correctly classify nodes as keystones. Although this type of study is a first step towards a standardized method to identify keystones in microbial networks, it is unknown whether the findings of Berry and Widder (2014) apply to real-world data sets. Yet, further research is needed in that field to allow for results comparison because measures to identify keystones are still highly inconsistent across studies.

Other association patterns may be informative as well. For example, motifs (e.g. triad motifs are the association of three nodes) may translate special patterns of communication between microbes (e.g. to coordinate their behaviour) or constitute biomarkers (Ma and Ye, 2017). Clustering coefficient and modularity (that quantifies the extent to which a network can be broken up into smaller components) may indicate degradation pathways, habitat filtering or cross-feeding (Chaffron et al., 2010; Röttjers and Faust, 2018). Finally,

the node degree distribution has often be used as a proxy for network robustness (that quantifies the resistance of a network to random or targeted node removal). In biological networks whose distribution follows a power law, the few hub nodes are not sensitive to random node removal but are sensitive to node removal targeting hub nodes. However, node removal may not reflect the way microbial communities respond to perturbations. Besides, this property is difficult to assess in networks that are not inferred from time series.

In conclusion, inferred microbial association networks have to be interpreted with care considering all these facts and, most important, a series of recommendations, such as those proposed by Berry and Widder (2014) and Röttjers and Faust (2018) (i.e. filtering out infrequent taxa, sequence communities with highly uneven composition more deeply, include as many samples as possible, use absolute abundance or data transformations that are robust to compositionality, use sequencing data at the highest resolution possible) should be applied as frequently as possible.

Efforts for the publication of interaction databases based on exhaustive literature surveys (e.g. Thompson et al., 2012; Poelen et al., 2014; Gao et al., 2017; Bjorbækmo et al., 2019; Vincent and Bowler, 2019) should be more than ever strongly encouraged. Besides, as mentioned by Carr et al. (2019), bioinformatically inferred associations are extremely useful for reducing the number of potential hypotheses that might be tested, but will never preclude the necessity for experimental validation.

5.3 Topological graph alignment of association networks

When comparing microbial association networks, global measures such as mean centrality measures, diameter and average path length are often calculated. In the article presented in section 3.2, we compared microbial networks based on these measures, as well as hub OTUs considered as keystones. However, although they can give insights on the network global characteristics, other methods allow to examine changes in network structure and allow for visual representation and interactive examination of important network attributes. These methods are referred to as network alignment methods. Initially developed to compare protein-protein interaction networks, network alignment may reveal valuable information, such as evolutionary conserved pathways (Kelley et al., 2003; Kuchaiev et al., 2010) and protein complexes or functional orthologs (Bandyopadhyay et al., 2006). This type of analysis have been successfully applied for the first time to microbial association networks by (Mandakovic et al., 2018) to examine changes in network structure, providing a comprehensive way to understand topological shifts among mem-

bers from two networks. Thus, network alignment could complement classical network comparison.

Network alignment consists in finding a node-to-node mapping (also called an *alignment*) between two networks. The objectives of network alignment algorithms are to (i) maximize the number of mapped proteins (nodes) that are evolutionarily or functionally related and (ii) maximize the number of common interactions (edges) between the networks (Malod-Dognin and Pržulj, 2015). Because the problem of network alignment is NP-complete, no efficient algorithm is known for solving it (Kuchaiev et al., 2010) and several network alignment heuristics (i.e. approximate aligners) have been proposed (Malod-Dognin and Pržulj, 2015). There exist *local* and *global* network aligners. Local network aligners align sub-networks called motifs that share similarities. Since these highly conserved motifs can overlap, local aligners typically result in one-to-many or many-to-many node mappings between the two input networks. On the other hand, global aligners aims to find the best overall alignment, resulting in one-to-one node mappings.

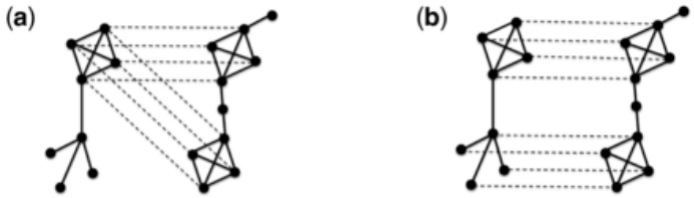
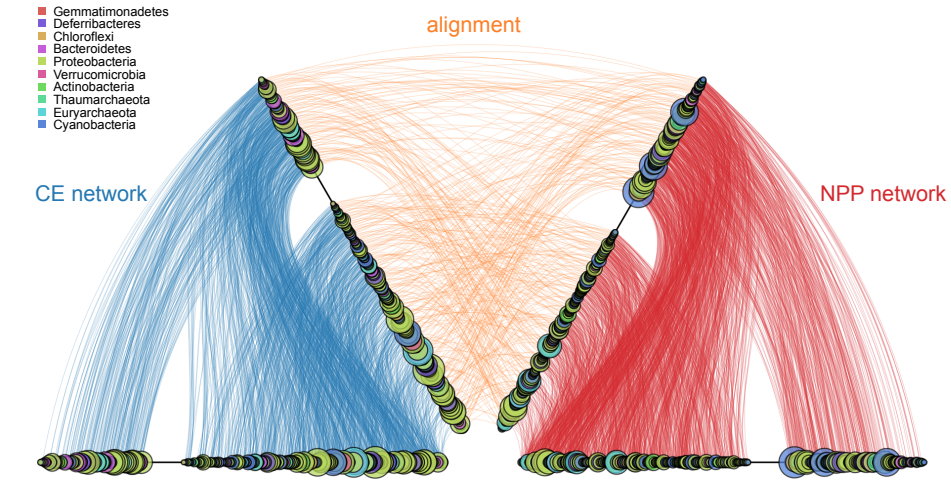
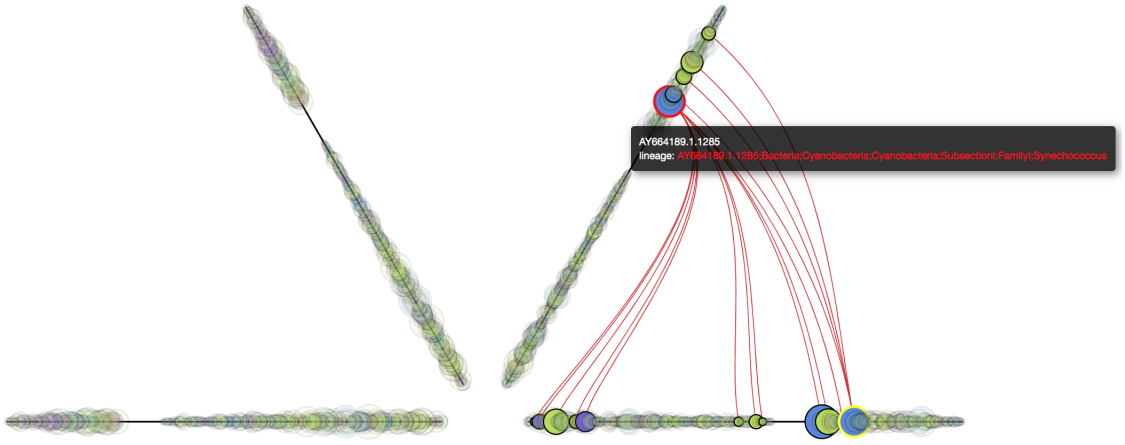


Figure 5.1 - Comparison of local and global network alignments. (a) Local network alignment specifying two different alignments of highly conserved sub-networks, each node of the first graph having ambiguous mappings under the different alignments. (b) Global alignment showing the best overall alignment, at the expense of local conserved regions (Meng et al., 2016).

As an example, we applied the global alignment tool L-GRAAL (Lagrangian GRAphlet-based network ALigner, Malod-Dognin and Pržulj, 2015) to the networks build and presented in section 3.2. This method aligns networks by taking into account both sequence similarity between nodes (i.e. similarity between reference sequences of the OTUs) and network topology. The balance between sequence and topology similarities is set by a parameter *alpha* that varies from 0 (topological information only) to 1 (sequence information only). The alignments were performed with values of *alpha* varying from 0 to 1 using a step size of 0.1. For each value of *alpha*, two measures of topological similarity were computed: the edge correctness and (EC) and symmetric sub-structure (S3). Given two aligned networks G and H, the EC is the percentage of edges from G that align to edges from H (Kuchaiev et al., 2010), while the S3 takes into account the unique edges in the composite graph created by the overlap in the two networks (Saraph and Milenković, 2014). Alignments were visualized in the form of hive plots with the tool HiveAlign (Mandakovic et al., 2018, https://gitlab.univ-nantes.fr/erwan.delage/HiveAlign).



(a) Complete alignment



(b) Focus on the top keystone OTU of the NPP network

Figure 5.2 - Network alignment between the CE (carbon export) and NPP (net primary production) networks inferred in chapter 3. The best consensus between topological, sequence similarity, EC and S3 was found for alpha = 0.6. On the axes are represented the nodes (nodes of the CE network on the left axes and nodes of the NPP network on the the right axes). Their color correspond to their phylum while their size translate their betweenness centrality. On each axis, the two sets of nodes correspond to common and specific nodes to each network. The order of each set of nodes translates their cluster coefficient.

The alignment is given in figure 5.2. The edge correctness of this alignment indicate that 28.98 % of the edges from the CE network are aligned to the edges of the NPP network, which informs on the global similarity of the network. The visual representation gives additional insights according to the tuned parameters (e.g. size of the nodes, order on the axis). Besides, interactive representations allow to focus on specific nodes, such as on figure 5.2b where the top keystone OTU of the NPP network (taxonomically assigned to *Synechococcus*) is highlighted, showing that it aligns to no other node of the CE network. It suggests that its position in the NPP network is unique, and that this feature could be considered as a characteristic of the microbial association network of the NPP state.

5.4 Limits of the use of machine learning for biological problems

Machine learning techniques are more and more popular for predicting and classifying biological data (Webb, 2018). In particular, it can mine interesting information from omic data. They can identify biomarkers for different purposes such as disease diagnosis or biomonitoring (Ziegler et al., 2012; Cordier et al., 2019), thus enabling decision making (e.g. provide adequate treatments to patients, take action so as to improve environmental health). However, several issues may limit their use for biological research.

The first one is known as the "Black Box" issue. Indeed, although machine learning trained models can be highly effective at predicting, they are often highly complex and the way the algorithms build the models is seen as being opaque. Yet, complex decision-making requires to improve human interpretability of the models (Miotto et al., 2017; Cordier et al., 2019). Algorithms such as random forests can help to provide more interpretable models by identifying variables that matter the most important for prediction (see chapter 4), which can be used by biologists to establish biomarkers.

Then, the models produced by machine learning algorithms are primarily dependent on the data we feed them. In particular, the limiting factor is the quantity of available data. Machine-learning algorithms require large quantities of data to learn. This may be a problem when collecting data is expensive or requires many efforts and is highly time-consuming, which is often the case in natural sciences. Besides, the models may be skewed if the training data do not account for the potential natural variability of the data. Thus, efforts should be continued to increase the sampling to develop larger databases in diverse ecosystems.

5.5 Perspectives for the study of microbial interactions and their involvement in the ocean carbon cycle

Several questions could not be addressed in the framework of this thesis but could be the subject of future research:

In chapter 3, we inferred association networks from archaeal and bacterial abundances. Thus, we didn't take into account their interactions with eukaryotes and viruses. Yet **cross-domain interactions** are numerous in the ocean. For example, many interactions between diatoms and bacteria have been reported (Amin et al., 2012), ranging from synergistic (e.g. production of vitamins required by diatoms and utilization of the diatoms

extracellular products by bacteria, Amin et al., 2012) to parasitic interactions (e.g. production of algicidal compounds by bacteria, Furusawa et al., 2003; Mayali and Azam, 2004). Protists are also important grazers of bacteria such as *Synechococcus* (Apple et al., 2011) and bacteria face viral lysis (Proctor and Fuhrman, 1990; Fuhrman, 1999; Rohwer et al., 2009). As a consequence, it seems necessary to take all these interactions into account. Most reconstructed microbial association networks include only bacteria but some authors have performed cross-domain networks, thereby recovering known symbiosis in the ocean between diatoms and Flavobacteria, and between dinoflagellates and members of Rhodobacterales (Lima-Mendez et al., 2015), and showing that fungi stabilize connectivity in the lung and skin microbial ecosystems (Tipton et al., 2018). Producing cross-domain analysis of the communities associated with the states of the biological carbon pump could reveal important associations because its functioning is highly dependent on cross-domain interactions (i.e. grazing, microbial decomposition, viral lysis, e.g. Grossart et al., 2005, 2006, 2007).

Our study of prokaryotic interactions focused on small size fractions (i.e. 0.22-1.6 μm and 0.22-3 μm) in which free prokaryotes are usually found. Thus, the samples didn't encompass **bacteria associated to larger eukaryotes and aggregate-attached bacteria**. Yet, microorganisms interact at the nanometre to millimetre scale (Azam and Malfatti, 2007) and eukaryotic phytoplankton such as diatoms have associated to them bacteria in their phycosphere (i.e. microscale region surrounding a phytoplankton cell that is rich in extracellular products) (Amin et al., 2012; Seymour et al., 2017). The analysis of fecal pellets may also provide insights into the feeding interactions between zooplankton, protists and bacteria. Besides, aggregates are "hotspots" of remineralization (Azam, 1998), which may reduce aggregates sinking rates, and display phylogenetically distinct assemblages from free prokaryotic populations (DeLong et al., 1993).

Although most of the flux attenuation occurs in the euphotic zone, **remineralization in the mesopelagic zone** is not anecdotal (90% of the annual quantity of exported carbon is respired back to CO_2, Robinson et al., 2010) and may be equal or exceed particulate organic carbon export (Lemaitre et al., 2018), thus affecting carbon sequestration by the biological carbon pump. However, our knowledge of the ecology of the "twilight zone" is lower in relation to the euphotic zone (Robinson et al., 2010) and would require investigation.

In this thesis, we focused on samples collected in temperate to tropical oceans. The **polar oceans** were thus excluded from the analysis, but these areas have unique characteristics that influence the magnitude, nature and timing of primary production and vertical POC fluxes. In the Arctic Ocean, sea ice cover attenuates solar irradiance, disturbs the mixing

of nutrients by winds, and induces stratification in the upper water, thereby influencing the primary production rates by phytoplankton (mostly diatoms). The extreme seasonality (including photoperiod and ice conditions) controls the spring bloom dynamics which determines POC export (Sakshaug, 2004; Carmack et al., 2006; Tremblay et al., 2018). Besides, sympagic species (i.e. associated with sea ice) form mats under sea ice. The release of this biomass during ice melt induces a rapid export of POC and provides food to pelagic and benthic communities (Forest et al., 2007). Variability in ice edge position relative to the coast may also influence carbon flux on arctic shelves by controlling the spatial co-occurence of primary production and grazers (Loeng et al., 2005). Besides, the data collected *Tara* Polar Circle expedition showed that viral, eukaryotic microbes and prokaryotic communities strongly differ from those of the temperate and tropical oceans (Gregory et al., 2019 and unpublished data) These unique features strongly suggest that microbial associations and potential biomarkers of the biological carbon pump are highly different in polar oceans compared to the rest of the ocean.

A criticism that can be made to the *Tara* Oceans expedition is its "snapshot" sampling strategy: ocean samples were collected at single time point and in sparse stations relative to the global ocean, which is inherent in global ocean studies (Karsenti et al., 2011). Time series microbial data would allow to complement the spatial coverage of the *Tara* Oceans sampling (Cram et al., 2014; Fuhrman et al., 2015; Faust et al., 2015; Ai et al., 2019). Indeed, it would **include seasonal and interannual variation** of the biological carbon pump processes that may be significant (Lohrenz et al., 1992; Karl et al., 1996). For example, monitoring a phytoplankton bloom phenomenon with environmental DNA sequencing and associated biogeochemical variables may be considered (Martin et al., 2011).

Further investigations of **prokaryotic functions** should be made to better understand their ecological role and molecular underpinnings in the biological carbon pump (Worden et al., 2015). In their 2016 study, Guidi et al. identified (from metagenomic data) sets of prokaryotic genes significantly correlated to carbon export in the oligotrophic ocean. These sets included a significant proportion of transmembrane sugar transporters and functions specific to the *Synechococcus* accessory photosynthetic apparatus or involved in carbohydrate breakdown, highlighting the potential roles of bacteria in primary production and the formation and degradation of marine aggregates. In addition, the application of metagenomics and metaproteomics in more geographically localized areas gave insights into the successive decomposition of algal-derived organic matter by bacteria (Teeling et al., 2012; Georges et al., 2014).

Finally, progress in the subjects cited above would allow for a better **monitoring of the**

ocean carbon cycle. By using machine learning approaches on omic data, new potential biomarkers of ecosystems health may progressively be discovered (e.g. Cordier et al., 2018). Some authors are already dreaming of "next-generation biomonitoring" (Bohan et al., 2017) by using "ecogenomic sensors" that would detect molecular markers indicative of the state of the ocean ecosystems (Scholin, 2009; Armbrust, 2014). Such instruments (free-drifting *in situ* robotic samplers) have already been used to sample plankton for further transcriptional analysis (Ottesen et al., 2013, 2014; Robidart et al., 2012, 2014), showing differences between the field and the laboratory. Microbial association network reconstruction would also benefit from these autonomous samplers by allowing a high spatial and temporal resolution (Bohan et al., 2017 and figure 5.3). As ecological networks structure determines ecosystem functioning, modifications would indicate response to climate change and other anthropogenic environmental perturbations. Networks would therefore constitute an adequate tool for observing and predicting the effects of environmental change.

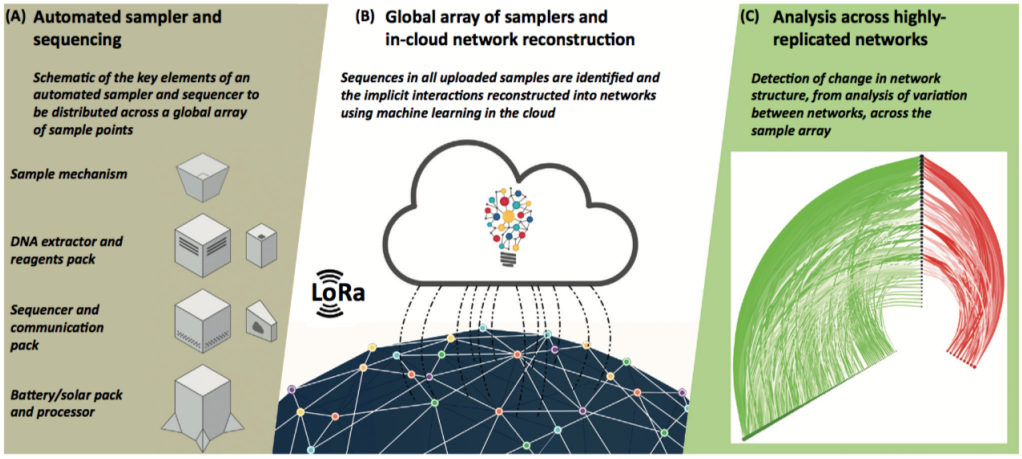


Figure 5.3 - Next-generation biomonitoring using automated sensors. The workflow includes (A) the design of automated sensors for sampling and sequencing, (B) the deployment of the sensors from which the sequence data are uploaded and ecological networks are reconstructed, and (C) the analysis of the networks (Bohan et al., 2017).

Bibliography

Aderem, A. (2005). Systems Biology: Its Practice and Challenges. *Cell*, 121(4):511–513.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *In: Proceedings of the 1993 Acm Sigmod International Conference on Management of Data, Washington Dc (usa*, pages 207–216.

Agusti, S., González-Gordillo, J. I., Vaqué, D., Estrada, M., Cerezo, M. I., Salazar, G., Gasol, J. M., and Duarte, C. M. (2015). Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nature Communications*, 6:7608.

Ai, D., Li, X., Pan, H., Chen, J., Cram, J. A., and Xia, L. C. (2019). Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. *BMC Genomics*, 20(S2).

Aitchison, J. (1981). A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13(2):175–189.

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.-M., Belser, C., Bertrand, A., Cruaud, C., Silva, C. D., Dossat, C., Gavory, F., Gas, S., Guy, J., Haquelle, M., Jacoby, E., Jaillon, O., Lemainque, A., Pelletier, E., Samson, G., Wessner, M., Team, G. T., Bazire, P., Beluche, O., Bertrand, L., Besnard-Gonnet, M., Bordelais, I., Boutard, M., Dubois, M., Dumont, C., Ettedgui, E., Fernandez, P., Garcia, E., Aiach, N. G., Guerin, T., Hamon, C., Brun, E., Lebled, S., Lenoble, P., Louesse, C., Mahieu, E., Mairey, B., Martins, N., Megret, C., Milani, C., Muanga, J., Orvain, C., Payen, E., Perroud, P., Petit, E., Robert, D., Ronsin, M., Vacherie, B., Acinas, S. G., Royo-Llonch, M., Cornejo-Castillo, F. M., Logares, R., Fernández-Gómez, B., Bowler, C., Cochrane, G., Amid, C., Hoopen, P. T., Vargas, C. D., Grimsley, N., Desgranges, E., Kandels-Lewis, S., Ogata, H., Poulton, N., Sieracki, M. E., Stepanauskas, R., Sullivan, M. B., Brum, J. R., Duhaime, M. B., Poulos, B. T., Hurwitz, B. L., Coordinators, T. O. C., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Vargas, C. D., Follows, M., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Raes, J., Sardet, C., Sieracki, M. E., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Pesant, S., Karsenti, E., and Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Scientific Data*, 4:sdata201793.

Alldredge, A. (2005). The contribution of discarded appendicularian houses to the flux of particulate organic carbon from oceanic surface waters. In Gorsky, G., editor, *Response of marine ecosystems to global change: Ecological impact of appendicularians*, pages 309–26. Contemporary Publishing International Paris, France.

Alldredge, A. L. (2000). Interstitial dissolved organic carbon (DOC) concentrations within sinking marine aggregates and their potential contribution to carbon flux. *Limnology and Oceanography*, 45(6):1245–1253.

Alldredge, A. L. and Silver, M. W. (1988). Characteristics, dynamics and significance of marine snow. *Progress in Oceanography*, 20(1):41–82.

Allen, J. T., Brown, L., Sanders, R., Mark Moore, C., Mustard, A., Fielding, S., Lucas, M., Rixen, M., Savidge, G., Henson, S., and Mayor, D. (2005). Diatom carbon export enhanced by silicate upwelling in the northeast Atlantic. *Nature*, 437(7059):728–732.

Amin, S. A., Parker, M. S., and Armbrust, E. V. (2012). Interactions between Diatoms and Bacteria. *Microbiology and Molecular Biology Reviews : MMBR*, 76(3):667–684.

Andersson, J. H., Wijsman, J. W. M., Herman, P. M. J., Middelburg, J. J., Soetaert, K., and Heip, C. (2004). Respiration patterns in the deep ocean. *Geophysical Research Letters*, 31(3).

Antia, A. N., Koeve, W., Fischer, G., Blanz, T., Schulz-Bull, D., Schölten, J., Neuer, S., Kremling, K., Kuss, J., Peinert, R., Hebbeln, D., Bathmann, U., Conte, M., Fehner, U., and Zeitzschel, B. (2001). Basin-wide particulate carbon flux in the Atlantic Ocean: Regional export patterns and potential for atmospheric CO2 sequestration. *Global Biogeochemical Cycles*, 15(4):845–862.

Antoine, D., André, J.-M., and Morel, A. (1996). Oceanic primary production: 2. Estimation at global scale from satellite (Coastal Zone Color Scanner) chlorophyll. *Global Biogeochemical Cycles*, 10(1):57–69.

Apple, J. K., Strom, S. L., Palenik, B., and Brahamsha, B. (2011). Variability in Protist Grazing and Growth on Different Marine Synechococcus Isolates. *Applied and Environmental Microbiology*, 77(9):3074–3084.

Armbrust, E. V. (2014). Taking the pulse of ocean microbes. *Science*, 345(6193):134–135.

Armstrong, R. A., Lee, C., I. Hedges, J., Honjo, S., and Wakeham, S. (2002). A new, mechanistic model for organic carbon fluxes in the ocean based on the quantitative association of POC with ballast minerals. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49:219–236.

Arnosti, C. (2010). Microbial Extracellular Enzymes and the Marine Carbon Cycle. *Annual Review of Marine Science*, 3(1):401–425.

Azam, F. (1998). Microbial Control of Oceanic Carbon Flux: The Plot Thickens. *Science*, 280(5364):694–696.

Azam, F., Fenchel, T., Field, J., Gray, J., Meyer-Reil, L., and Thingstad, F. (1983). The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series*, 10:257–263.

Azam, F. and Malfatti, F. (2007). Microbial structuring of marine ecosystems. *Nature Reviews Microbiology*, 5(10):782–791.

Azencott, C.-A. (2018). *Introduction au Machine Learning*. Dunod.

Bach, L. T., Bauke, C., Meier, K. J. S., Riebesell, U., and Schulz, K. G. (2012). Influence of changing carbonate chemistry on morphology and weight of coccoliths formed by *Emiliania huxleyi*. *Biogeosciences*, 9(8):3449–3463.

Bandyopadhyay, S., Sharan, R., and Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16(3):428–435.

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Bascompte, J. (2009). Disentangling the Web of Life. *Science*, 325(5939):416–419.

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Beaulieu, S. E. (2003). Accumulation and fate of phytodetritus on the sea floor. In *Oceanography and Marine Biology, An Annual Review, Volume 40*, pages 179–217. CRC Press.

Behrenfeld, M. J. and Falkowski, P. G. (1997). Photosynthetic rates derived from saætellite-based chlorophyll concentration. *Limnology and Oceanography*, 42(1):1–20.

Belcher, A., Iversen, M., Giering, S., Riou, V., Henson, S. A., Berline, L., Guilloux, L., and Sanders, R. (2016). Depth-resolved particle-associated microbial respiration in the northeast Atlantic. *Biogeosciences*, 13(17):4927–4943.

Beman, J. M., Chow, C.-E., King, A. L., Feng, Y., Fuhrman, J. A., Andersson, A., Bates, N. R., Popp, B. N., and Hutchins, D. A. (2011). Global declines in oceanic nitrification rates as a consequence of ocean acidification. *Proceedings of the National Academy of Sciences*, 108(1):208–213.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Benoiston, A.-S., Eveillard, D., Chaffron, S., Delage, E., Guidi, L., and Bittner, L. (in prep.). The microbial drivers of the biological carbon pump.

Benoiston, A.-S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., and Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Phil. Trans. R. Soc. B*, 372(1728):20160397.

Berger, A. (1988). Milankovitch Theory and climate. *Reviews of Geophysics*, 26(4):624–657.

Berger, W. H., Smetacek, V., and Wefer, G. (1989). Ocean productivity and paleoproductivity - an overview. In Berger, W., Smetacek, V., and Wefer, G., editors, *Productivity of the Oceans present and past: Report of the Dahlem Workshop on Productivity of the Ocean, Berlin, 1988*, pages 1–34. Life sciences research reports 44, Wiley & Sons, Chichester.

Berger, W. H. and Soutar, A. (1967). Planktonic Foraminifera: Field Experiment on Production Rate. *Science*, 156(3781):1495–1497.

Berry, D. and Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5.

Bidigare, R. R., Benitez-Nelson, C., Leonard, C. L., Quay, P. D., Parsons, M. L., Foley, D. G., and Seki, M. P. (2003). Influence of a cyclonic eddy on microheterotroph biomass and carbon export in the lee of Hawaii. *Geophysical Research Letters*, 30(6).

Bidle, K. D. and Azam, F. (1999). Accelerated dissolution of diatom silica by marine bacterial assemblages. *Nature*, 397(6719):508–512.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in ecology & evolution*, 27(4):233–243.

Biller, S. J., Schubotz, F., Roggensack, S. E., Thompson, A. W., Summons, R. E., and Chisholm, S. W. (2014). Bacterial Vesicles in Marine Ecosystems. *Science*, 343(6167):183–186.

Bjorbækmo, M. F. M., Evenstad, A., Røsæg, L. L., Krabberød, A. K., and Logares, R. (2019). The planktonic protist interactome: where do we stand after a century of research? *bioRxiv*, page 587352.

Blain, S., Quéguiner, B., Armand, L., Belviso, S., Bombled, B., Bopp, L., Bowie, A., Brunet, C., Brussaard, C., Carlotti, F., Christaki, U., Corbière, A., Durand, I., Ebersbach, F., Fuda, J.-L., Garcia, N., Gerringa, L., Griffiths, B., Guigue, C., Guillerm, C., Jacquet, S., Jeandel, C., Laan, P., Lefèvre, D., Lo Monaco, C., Malits, A., Mosseri, J., Obernosterer, I., Park, Y.-H., Picheral, M., Pondaven, P., Remenyi, T., Sandroni, V., Sarthou, G., Savoye, N., Scouarnec, L., Souhaut, M., Thuiller, D., Timmermans, K., Trull, T., Uitz, J., van Beek, P., Veldhuis, M., Vincent, D., Viollier, E., Vong, L., and Wagener, T. (2007). Effect of natural iron fertilization on carbon sequestration in the Southern Ocean. *Nature*, 446(7139):1070–1074.

Boeuf, D., Edwards, B. R., Eppley, J. M., Hu, S. K., Poff, K. E., Romano, A. E., Caron, D. A., Karl, D. M., and DeLong, E. F. (2019). Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proceedings of the National Academy of Sciences*, page 201903080.

Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A., and Tamaddoni-Nezhad, A. (2011). Automated Discovery of Food Webs from Ecological Data Using Logic-Based Machine Learning. *PLOS ONE*, 6(12):e29028.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. *Trends in Ecology & Evolution*, 32(7):477–487.

Bollens, S. M., Rollwagen-Bollens, G., Quenette, J. A., and Bochdansky, A. B. (2011). Cascading migrations and implications for vertical fluxes in pelagic ecosystems. *Journal of Plankton Research*, 33(3):349–355.

Bopp, L., Monfray, P., Aumont, O., Dufresne, J.-L., Treut, H. L., Madec, G., Terray, L., and Orr, J. C. (2001). Potential impact of climate change on marine export production. *Global Biogeochemical Cycles*, 15(1):81–99.

Boyd, P. and Newton, P. (1995). Evidence of the potential influence of planktonic community structure on the interannual variability of particulate organic carbon flux. *Deep Sea Research Part I: Oceanographic Research*, 42:619–639.

Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A., and Weber, T. (2019). Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature*, 568(7752):327.

Boyd, P. W., Gall, M. P., Silver, M. W., Coale, S. L., Bidigare, R. R., and Bishop, J. L. K. B. (2008). Quantifying the surface–subsurface biogeochemical coupling during the VERTIGO ALOHA and K2 studies. *Deep Sea Research Part II: Topical Studies in Oceanography*, 55(14):1578–1593.

Boyd, P. W. and Newton, P. P. (1999). Does planktonic community structure determine downward particulate organic carbon flux in different oceanic provinces? *Deep Sea Research Part I: Oceanographic Research Papers*, 46(1):63–91.

Boyd, P. W. and Trull, T. W. (2007). Understanding the export of biogenic particles in oceanic waters: Is there consensus? *Progress in Oceanography*, 72(4):276–312.

Boyd, P. W., Watson, A. J., Law, C. S., Abraham, E. R., Trull, T., Murdoch, R., Bakker, D. C. E., Bowie, A. R., Buesseler, K. O., Chang, H., Charette, M., Croot, P., Downing, K., Frew, R., Gall, M., Hadfield, M., Hall, J., Harvey, M., Jameson, G., LaRoche, J., Liddicoat, M., Ling, R., Maldonado, M. T., McKay, R. M., Nodder, S., Pickmere, S., Pridmore, R., Rintoul, S., Safi, K., Sutton, P., Strzepek, R., Tanneberger, K., Turner, S., Waite, A., and Zeldis, J. (2000). A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature*, 407(6805):695.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton.

Broecker, W. S. (1982a). Glacial to interglacial changes in ocean chemistry. *Progress in Oceanography*, 11(2):151–197.

Broecker, W. S. (1982b). Ocean chemistry during glacial time. *Geochimica et Cosmochimica Acta*, 46(10):1689–1705.

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., Vargas, C. d., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., Schwenck, S. M., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Coordinators, T. O., Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., and Sullivan, M. B. (2015). Patterns and ecological drivers of ocean viral communities. *Science*, 348(6237):1261498.

Buchan, A., LeCleir, G. R., Gulvik, C. A., and González, J. M. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nature Reviews Microbiology*, 12(10):686–698.

Buesseler, K. O. (1991). Do upper-ocean sediment traps provide an accurate record of particle flux? *Nature*, 353(6343):420.

Buesseler, K. O. (1998). The decoupling of production and particulate export in the surface ocean. *Global Biogeochemical Cycles*, 12(2):297–310.

Buesseler, K. O., Bacon, M. P., Kirk Cochran, J., and Livingston, H. D. (1992). Carbon and nitrogen export during the JGOFS North Atlantic Bloom experiment estimated from 234th: 238u disequilibria. *Deep Sea Research Part A. Oceanographic Research Papers*, 39(7):1115–1137.

Buesseler, K. O. and Boyd, P. W. (2009). Shedding light on processes that control particle export and flux attenuation in the twilight zone of the open ocean. *Limnology and Oceanography*, 54(4):1210–1232.

Buesseler, K. O., Lamborg, C. H., Boyd, P. W., Lam, P. J., Trull, T. W., Bidigare, R. R., Bishop, J. K. B., Casciotti, K. L., Dehairs, F., Elskens, M., Honda, M., Karl, D. M., Siegel, D. A., Silver, M. W., Steinberg, D. K., Valdes, J., Mooy, B. V., and Wilson, S. (2007). Revisiting Carbon Flux Through the Ocean's Twilight Zone. *Science*, 316(5824):567–570.

Buick, R. (2008). When did oxygenic photosynthesis evolve? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1504):2731–2743.

Burd, A. B., Hansell, D. A., Steinberg, D. K., Anderson, T. R., Arístegui, J., Baltar, F., Beaupré, S. R., Buesseler, K. O., DeHairs, F., Jackson, G. A., Kadko, D. C., Koppelmann, R., Lampitt, R. S., Nagata, T., Reinthaler, T., Robinson, C., Robison, B. H., Tamburini, C., and Tanaka, T. (2010). Assessing the apparent imbalance between geochemical and biochemical indicators of meso- and bathypelagic biological activity: What the @$#! is wrong with present calculations of carbon budgets? *Deep Sea Research Part II: Topical Studies in Oceanography*, 57(16):1557–1571.

Calbet, A. and Landry, M. R. (2004). Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnology and Oceanography*, 49(1):51–57.

Camerano, L. (1880). Dell'equilibrio dei viventi mercè la reciproca distruzione. *Atti della Reale Accademia della Scienze di Torino*, 15:393–414.

Capone, D. G., Burns, J. A., Montoya, J. P., Subramaniam, A., Mahaffey, C., Gunderson, T., Michaels, A. F., and Carpenter, E. J. (2005). Nitrogen fixation by Trichodesmium spp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochemical Cycles*, 19(2).

Caputi, L., Carradec, Q., Eveillard, D., Kirilovsky, A., Pelletier, E., Karlusich, J. J. P., Vieira, F. R. J., Villar, E., Chaffron, S., Malviya, S., Scalco, E., Acinas, S. G., Alberti, A., Aury, J.-M., Benoiston, A.-S., Bertrand, A., Biard, T., Bittner, L., Boccara, M., Brum, J. R., Brunet, C., Busseni, G., Carratalà, A., Claustre, H., Coelho, L. P., Colin, S., D'Aniello, S., Silva, C. D., Core, M. D., Doré, H., Gasparini, S., Kokoszka, F., Jamet, J.-L., Lejeusne, C., Lepoivre, C., Lescot, M., Lima-Mendez, G., Lombard, F., Lukeš, J., Maillet, N., Madoui, M.-A., Martinez, E., Mazzocchi, M. G., Néou, M. B., Paz-Yepes, J., Poulain, J., Ramondenc, S., Romagnan, J.-B., Roux, S., Manta, D. S., Sanges, R., Speich, S., Sprovieri, M., Sunagawa, S., Taillandier, V., Tanaka, A., Tirichine, L., Trottier, C., Uitz, J., Veluchamy, A., Veselá, J., Vincent, F., Yau, S., Kandels-Lewis, S., Searson, S., Dimier, C., Picheral, M., Bork, P., Boss, E., Vargas, C., Follows, M. J., Grimsley, N., Guidi, L., Hingamp, P., Karsenti, E., Sordino, P., Stemmann, L., Sullivan, M. B., Tagliabue, A., Zingone, A., Garczarek, L., d'Ortenzio, F., Testor, P., Not, F., d'Alcalà, M. R., Wincker, P., Bowler, C., Iudicone, D., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Vargas, C., Follows, M. J., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Krzic, U., Not, F., Ogata, H., Pesant, S., Raes, J., Reynaud, E. G., Sardet, C., Sieracki, M., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Velayoudon, D., Weissenbach, J., Wincker, P., and and (2019). Community-level responses to iron availability in open ocean plankton ecosystems. *Global Biogeochemical Cycles*, 33(3):391–419.

Carmack, E., Barber, D., Christensen, J., Macdonald, R., Rudels, B., and Sakshaug, E. (2006). Climate variability and physical forcing of the food webs and the carbon budget on panarctic shelves. *Progress in Oceanography*, 71(2):145–181.

Caron, D. A. (2016). Mixotrophy stirs up our understanding of marine food webs. *Proceedings of the National Academy of Sciences*, 113(11):2806–2808.

Caron, D. A., Davis, P. G., Madin, L., and Sieburth, J. M. (1986). Enrichment of microbial populations in macroaggregates (marine snow) from surface waters of the North Atlantic. *Journal of Marine Research*, 44:543–565.

Carr, A., Diener, C., Baliga, N. S., and Gibbons, S. M. (2019). Use and abuse of correlation analyses in microbial ecology. *The ISME Journal*.

Carradec, Q., Pelletier, E., Silva, C. D., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M. B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., Vargas,

C. d., Iudicone, D., Bowler, C., and Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1):373.

Cermeño, P., Dutkiewicz, S., Harris, R. P., Follows, M., Schofield, O., and Falkowski, P. G. (2008). The role of nutricline depth in regulating the ocean carbon cycle. *Proceedings of the National Academy of Sciences*, 105(51):20344–20349.

Chaffron, S., Guidi, L., d'Ovidio, F., Speich, S., Audic, S., De Monte, S., Iudicone, D., Picheral, M., Pesant, S., Tara Oceans Consortium, C., and Tara Oceans Expedition, P. (2014). Environmental context of selected samples from the Tara Oceans Expedition (2009-2013). *PANGAEA*.

Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959.

Chavez, F. P., Messié, M., and Pennington, J. T. (2010). Marine Primary Production in Relation to Climate Variability and Change. *Annual Review of Marine Science*, 3(1):227–260.

Cho, B. C. and Azam, F. (1988). Major role of bacteria in biogeochemical fluxes in the ocean's interior. *Nature*, 332(6163):441–443.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., S., P., and Thornton, P. (2013). 2013: Carbon and other biogeochemical cycles. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 465–570. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111. arXiv: cond-mat/0408187.

Claustre, H., Antoine, D., Boehme, L., Boss, E., D'Ortenzio, F., Fanton Dandon, O., Guinet, C., Gruber, N., Handegard, N. O., Hood, M., Johnson, K., Arne, K., Lampitt, R., Traon, P.-Y., Le Quéré, C., Lewis, M., Perry, M.-J., Platt, T., Roemmich, D., and Yoder, J. (2009). Guidelines towards an integrated ocean observation system for ecosystems and biogeochemical cycles. In *OceanObs'09: Sustained Ocean Observations and Information for Society*.

Coale, K. H. and Bruland, K. W. (1985). 234th:238u disequilibria within the California Current1. *Limnology and Oceanography*, 30(1):22–33.

Cohen, J. E. (1994). Lorenzo Camerano's Contribution to Early Food Web Theory. In Levin, S. A., editor, *Frontiers in Mathematical Biology*, Lecture Notes in Biomathematics, pages 351–359. Springer Berlin Heidelberg.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., and Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from

eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology*, 51(16):9118–9126.

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., and Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18(6):1381–1391.

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., and Pawlowski, J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology*, 27(5):387–397.

Cram, J. A., Chow, C.-E. T., Sachdeva, R., Needham, D. M., Parada, A. E., Steele, J. A., and Fuhrman, J. A. (2014). Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *The ISME Journal*, 9(3):563–580.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Dall'Olmo, G., Dingle, J., Polimene, L., Brewin, R. J. W., and Claustre, H. (2016). Substantial energy input to the mesopelagic ecosystem from the seasonal mixed-layer pump. *Nature Geoscience*, 9(11):820–823.

De La Rocha, C. L. and Passow, U. (2007). Factors influencing the sinking of POC and the efficiency of the biological carbon pump. *Deep Sea Research Part II: Topical Studies in Oceanography*, 54(5):639–658.

De La Rocha, C. L. and Passow, U. (2014). 8.4 - The Biological Pump. In Holland, H. D. and Turekian, K. K., editors, *Treatise on Geochemistry (Second Edition)*, pages 93–122. Elsevier, Oxford.

del Giorgio, P. A. and Duarte, C. M. (2002). Respiration in the open ocean. *Nature*, 420(6914):379–384.

Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R., and Vogel, T. M. (2011). Accessing the Soil Metagenome for Studies of Microbial Diversity. *Applied and Environmental Microbiology*, 77(4):1315–1324.

DeLong, E. F., Franks, D. G., and Alldredge, A. L. (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography*, 38(5):924–934.

Deutsch, C. and Weber, T. (2012). Nutrient Ratios as a Tracer and Driver of Ocean Biogeochemistry. *Annual Review of Marine Science*, 4(1):113–141.

DeVries, T. (2014). The oceanic anthropogenic CO2 sink: Storage, air-sea fluxes, and transports over the industrial era. *Global Biogeochemical Cycles*, 28(7):631–647.

DeVries, T. and Primeau, F. (2011). Dynamically and Observationally Constrained Estimates of Water-Mass Distributions and Ages in the Global Ocean. *Journal of Physical Oceanography*, 41(12):2381–2401.

Diamond, J. (1975). Assembly of species communities. In Diamond, J. and Cody, M., editors, *Ecology and Evolution of Communities*, pages 342–344. Harvard University Press, Boston.

Dilling, L. and Alldredge, A. L. (2000). Fragmentation of marine snow by swimming macrozooplankton: A new process impacting carbon cycling in the sea. *Deep Sea Research Part I: Oceanographic Research Papers*, 47(7):1227–1245.

Dinsdale, E. A., Edwards, R. A., Bailey, B., Tuba, I., Akhter, S., McNair, K., Schmieder, R., Apkarian, N., Clark, M., Guan, E., Hernandez, M., Isaacs, K., Peterson, C., Regh, T., and Ponomarenko, V. (2013). Multivariate Analysis of Functional Metagenomes. *Frontiers in genetics*, 4:41.

Doney, S. C. (2006). Plankton in a warmer world. *Nature*, 444(7120):695–696.

Dore, J. E., Letelier, R. M., Church, M. J., Lukas, R., and Karl, D. M. (2008). Summer phytoplankton blooms in the oligotrophic North Pacific Subtropical Gyre: Historical perspective and recent observations. *Progress in Oceanography*, 76(1):2–38.

Duarte, C. M. (2015). Seafaring in the 21st Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin*, 24(1):11–14.

Dubischar, C. D. and Bathmann, U. V. (2002). The occurrence of faecal material in relation to different pelagic systems in the Southern Ocean and its importance for vertical flux. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(16):3229–3242.

Dugdale, R. C. and Goering, J. J. (1967). Uptake of New and Regenerated Forms of Nitrogen in Primary Productivity. *Limnology and Oceanography*, 12(2):196–206.

Dugdale, R. C. and Wilkerson, F. P. (2001). Sources and fates of silicon in the ocean: the role of diatoms in the climate and glacial cycles. *Scientia Marina*, 65(S2):141–152.

Dunne, J. A. (2009). The network structure of food webs. In *Workshop on Theoretical Ecology and Global Change*.

Dunne, J. P., Sarmiento, J. L., and Gnanadesikan, A. (2007). A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor. *Global Biogeochemical Cycles*, 21(4).

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J., Alexander Richter, R., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J. D., Haft, D. H., Halpern, A. L., Lasken, R. S., Nealson, K., Friedman, R., and Craig Venter, J. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME Journal*, 6(6):1186–1199.

Duran-Pinedo, A. E., Paster, B., Teles, R., and Frias-Lopez, J. (2011). Correlation Network Analysis Applied to Complex Biofilm Communities. *PLOS ONE*, 6(12):e28438.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Egerton, F. N. (2007). Understanding Food Chains and Food Webs, 1700–1970. *The Bulletin of the Ecological Society of America*, 88(1):50–69.

Egge, J. K., Thingstad, T. F., Larsen, A., Engel, A., Wohlers, J., Bellerby, R. G. J., and Riebesell, U. (2009). Primary production during nutrient-induced blooms at elevated CO_2 concentrations. *Biogeosciences*, 6(5):877–885.

Elton, C. S. (1927). *Animal ecology*. Sidgwick & Jackson, ltd.

Engel, A., Delille, B., Jacquet, S., Riebesell, U., Rochelle-Newall, E., Terbrüggen, A., and Zondervan, I. (2004). Transparent exopolymer particles and dissolved organic carbon production by Emiliania huxleyi exposed to different CO2 concentrations: a mesocosm experiment. *Aquatic Microbial Ecology*, 34(1):93–104.

Eppley, R. W. and Peterson, B. J. (1979). Particulate organic matter flux and planktonic new production in the deep ocean. *Nature*, 282(5740):677.

Eppley, R. W., Stewart, E., Abbott, M. R., and Heyman, U. (1985). Estimating ocean primary production from satellite chlorophyll. Introduction to regional differences and statistics for the Southern California Bight. *Journal of Plankton Research*, 7(1):57–70.

Erdös, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, 6:290–297.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140.

Falkowski, P. G. (1981). Light-shade adaptation and assimilation numbers. *Journal of Plankton Research*, 3(2):203–216.

Falkowski, P. G. (1997). Evolution of the nitrogen cycle and its influence on the biological sequestration of CO 2 in the ocean. *Nature*, 387(6630):272–275.

Falkowski, P. G., Barber, R. T., and Smetacek, V. (1998). Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science*, 281(5374):200–206.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879):1034–1039.

Falkowski, P. G. and Oliver, M. J. (2007). Mix and match: how climate selects phytoplankton. *Nature Reviews Microbiology*, 5(10):813–819.

Falkowski, P. G. and Raven, J. A. (2007). *Aquatic Photosynthesis*. Princeton University Press, Princeton, 2nd revised edition edition.

Farrant, G. K., Doré, H., Cornejo-Castillo, F. M., Partensky, F., Ratin, M., Ostrowski, M., Pitt, F. D., Wincker, P., Scanlan, D. J., Iudicone, D., Acinas, S. G., and Garczarek, L. (2016). Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proceedings of the National Academy of Sciences*, 113(24):E3365–E3374.

Faure, E., Not, F., Benoiston, A.-S., Labadie, K., Bittner, L., and Ayata, S.-D. (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. *The ISME Journal*, 13(4):1072.

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, 25:56–66.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Computational Biology*, 8(7):e1002606.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, 281(5374):237–240.

Finkel, Z. V., Beardall, J., Flynn, K. J., Quigg, A., Rees, T. A. V., and Raven, J. A. (2010). Phytoplankton in a changing world: cell size and elemental stoichiometry. *Journal of Plankton Research*, 32(1):119–137.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A., and Martiny, A. C. (2013). Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24):9824–9829.

Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M., and DeLong, E. F. (2015). Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Frontiers in Microbiology*, 6.

Forest, A., Sampei, M., Hattori, H., Makabe, R., Sasaki, H., Fukuchi, M., Wassmann, P., and Fortier, L. (2007). Particulate organic carbon fluxes on the slope of the Mackenzie Shelf (Beaufort Sea): Physical and biological forcing of shelf-basin exchanges. *Journal of Marine Systems*, 68(1):39–54.

Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E., and Oakley, B. B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences*, 102(41):14683–14688.

Francois, R., Honjo, S., Krishfield, R., and Manganini, S. (2002). Factors controlling the flux of organic carbon to the bathypelagic zone of the ocean. *Global Biogeochemical Cycles*, 16(4):34–1–34–20.

Franks, P. J. S. (2002). NPZ Models of Plankton Dynamics: Their Construction, Coupling to Physics, and Application. *Journal of Oceanography*, 58(2):379–387.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41.

Friedman, J. and Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, 8(9):e1002687.

Friedman, J., Higgins, L. M., and Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, 1(5):0109.

Friedrich, T. and Oschlies, A. (2009). Neural network-based estimates of North Atlantic surface pCO2 from satellite data: A methodological study. *Journal of Geophysical Research: Oceans*, 114(C3).

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, 399(6736):541–548.

Fuhrman, J. A., Cram, J. A., and Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, 13(3):133–146.

Furusawa, G., Yoshikawa, T., Yasuda, A., and Sakata, T. (2003). Algicidal activity and gliding motility of Saprospira sp. SS98-5. *Canadian Journal of Microbiology*, 49(2):92–100.

Ganeshram, R. S., Pedersen, T. F., Calvert, S. E., and Murray, J. W. (1995). Large changes in oceanic nutrient inventories from glacial to interglacial periods. *Nature*, 376(6543):755–758.

Gao, N. L., Zhang, C., Zhang, Z., Hu, S., Lercher, M. J., Zhao, X.-M., Bork, P., Liu, Z., and Chen, W.-H. (2017). MVP: a microbe–phage interaction database. *Nucleic Acids Research*, 46(D1):D700–D707.

Georges, A. A., El-Swais, H., Craig, S. E., Li, W. K., and Walsh, D. A. (2014). Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. *The ISME Journal*, 8(6):1301–1313.

Gerhard, W. A. and Gunsch, C. K. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International*, 124:312–319.

Giering, S. L. C., Sanders, R., Martin, A. P., Henson, S. A., Riley, J. S., Marsay, C. M., and Johns, D. G. (2017). Particle flux in the oceans: Challenging the steady state assumption. *Global Biogeochemical Cycles*, 31(1):159–171.

Giovannoni, S. J. (2017). SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annual Review of Marine Science*, 9(1):231–255.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8.

Goldthwait, S., Yen, J., Brown, J., and Alldredge, A. (2004). Quantification of marine snow fragmentation by swimming euphausiids. *Limnology and Oceanography*, 49(4):940–952.

Goldthwait, S. A., Carlson, C. A., Henderson, G. K., and Alldredge, A. L. (2005). Effects of physical fragmentation on remineralization of marine snow. *Marine Ecology Progress Series*, 305:59–65.

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., Poulain, J., Tremblay, J.-E., Vik, D., Acinas, S. G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., Vargas, C. d., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Poulton, N., Raes, J., Sardet, C., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Babin, M., Bowler, C., Culley, A. I., Vargas, C. d., Dutilh, B. E., Iudicone, D., Karp-Boss, L., Roux, S., Sunagawa, S., Wincker, P., and Sullivan, M. B. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, 177(5):1109–1123.e14.

Grossart, H.-P., Czub, G., and Simon, M. (2006). Algae–bacteria interactions and their effects on aggregation and organic matter flux in the sea. *Environmental Microbiology*, 8(6):1074–1084.

Grossart, H.-P., Kiørboe, T., Tang, K., Allagier, M., M. Yam, E., and Ploug, H. (2005). Interactions between marine snow and heterotrophic bacteria: Aggregate formation and microbial dynamics. *Aquatic Microbial Ecology*, 42:19–26.

Grossart, H.-P. and Ploug, H. (2001). Microbial degradation of organic carbon and nitrogen on diatom aggregates. *Limnology and Oceanography*, 46(2):267–277.

Grossart, H.-P., Schlingloff, A., Bernhard, M., Simon, M., and Brinkhoff, T. (2004). Antagonistic activity of bacteria isolated from organic aggregates of the German Wadden Sea. *FEMS Microbiology Ecology*, 47(3):387–396.

Grossart, H.-P., Tang, K. W., Kiørboe, T., and Ploug, H. (2007). Comparison of cell-specific activity between free-living and attached bacteria using isolates and natural assemblages. *FEMS microbiology letters*, 266(2):194–200.

Gruber, N., Arne, K., Borges, A., Claustre, H., Doney, S., Feely, R., Hood, M., Ishii, M., Kozyr, A., Monteiro, P., Nojiri, Y., Chris, S., Schuster, U., Wallace, D., and Wanninkhof, R. (2010). Towards an Integrated Observing System for Ocean Carbon and Biogeochemistry at a Time of Change. In *Workshop on Theoretical Ecology and Global Change*, pages 182–196.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J. R., Coelho, L. P., Espinoza, J. C. I., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., Tara Oceans Consortium Coordinators, Stemmann, L., Not, F., Hingamp, P., Speich, S., Follows, M., Karp-Boss, L., Boss, E., Ogata, H., Pesant, S., Weissenbach, J., Wincker, P., Acinas, S. G., Bork, P., de Vargas, C., Iudicone, D., Sullivan, M. B., Raes, J., Karsenti, E., Bowler, C., and Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470.

Guidi, L., Jackson, G. A., Stemmann, L., Miquel, J. C., Picheral, M., and Gorsky, G. (2008). Relationship between particle size distribution and flux in the mesopelagic zone. *Deep Sea Research Part I: Oceanographic Research Papers*, 55(10):1364–1374.

Guidi, L., Legendre, L., Reygondeau, G., Uitz, J., Stemmann, L., and Henson, S. A. (2015). A new look at ocean carbon remineralization for estimating deepwater sequestration. *Global Biogeochemical Cycles*, 29(7):2014GB005063.

Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L., Picheral, M., and Gorsky, G. (2009). Effects of phytoplankton community on production, size, and export of large aggregates: A world-ocean analysis. *Limnology and Oceanography*, 54(6):1951–1963.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., and 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(Database issue):D597–D604.

Hansell, D. A. (2013). Recalcitrant Dissolved Organic Carbon Fractions. *Annual Review of Marine Science*, 5(1):421–445.

Hays, J. D., Imbrie, J., and Shackleton, N. J. (1976). Variations in the Earth's Orbit: Pacemaker of the Ice Ages. *Science*, 194(4270):1121–1132.

Hedges, J. I. and Keil, R. G. (1995). Sedimentary organic matter preservation: an assessment and speculative synthesis. *Marine Chemistry*, 49(2):81–115.

Henson, S. A., Sanders, R., and Madsen, E. (2012). Global patterns in efficiency of particulate organic carbon export and transfer to the deep ocean. *Global Biogeochemical Cycles*, 26(1).

Herndl, G. J. and Reinthaler, T. (2013). Microbial control of the dark end of the biological pump. *Nature Geoscience*, 6(9):718–724.

Hernández-León, S. and Ikeda, T. (2005). A global assessment of mesozooplankton respiration in the ocean. *Journal of Plankton Research*, 27(2):153–158.

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, 3(12):1365–1373.

Honjo, S. (1976). Coccoliths: Production, transportation and sedimentation. *Marine Micropaleontology*, 1:65–79.

Honjo, S., Manganini, S. J., Krishfield, R. A., and Francois, R. (2008). Particulate organic carbon fluxes to the ocean interior and factors controlling the biological pump: A synthesis of global sediment trap programs since 1983. *Progress in Oceanography*, 76(3):217–285.

Houghton, R. A. (2014). The Contemporary Carbon Cycle. In Holland, H. D. and Turekian, K. K., editors, *Treatise on Geochemistry (Second Edition)*, pages 399–435. Elsevier, Oxford.

Hutchins, D. A., Mulholland, M. R., and Fu, F. (2009). Nutrient Cycles and Marine Microbes in a CO2-Enriched Ocean. *Oceanography*.

Immitzer, M., Atzberger, C., and Koukal, T. (2012). Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data. *Remote Sensing*, 4:2661–2693.

IPCC (2013). *Climate change 2013: The physical science basis. Contribution of the working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Iversen, M. H. and Poulsen, L. K. (2007). Coprorhexy, coprophagy, and coprochaly in the copepods Calanus helgolandicus, Pseudocalanus elongatus, and Oithona similis. *Marine Ecology Progress Series*, 350:79–89.

Jackson, G. A., Maffione, R., Costello, D. K., Alldredge, A. L., Logan, B. E., and Dam, H. G. (1997). Particle size spectra between 1 µm and 1 cm at Monterey Bay determined using multiple instruments. *Deep Sea Research Part I: Oceanographic Research Papers*, 44(11):1739–1767.

Jackson, M. A., Bonder, M. J., Kuncheva, Z., Zierer, J., Fu, J., Kurilshikov, A., Wijmenga, C., Zhernakova, A., Bell, J. T., Spector, T. D., and Steves, C. J. (2018). Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. *PeerJ*, 6.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer-Verlag, New York.

Jenkins, W. J. (1982). Oxygen utilization rates in North Atlantic subtropical gyre and primary production in oligotrophic systems. *Nature*, 300(5889):246–248.

Jennings, B. R., Parslow, K., and Ottewill, R. H. (1988). Particle size measurement: the equivalent spherical diameter. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 419(1856):137–149.

Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., Kattner, G., Wilhelm, S. W., Kirchman, D. L., Weinbauer, M. G., Luo, T., Chen, F., and Azam, F. (2010). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology*, 8(8):593–599.

Kara, E. L., Hanson, P. C., Hu, Y. H., Winslow, L., and McMahon, K. D. (2013). A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. *The ISME Journal*, 7(3):680–684.

Karl, D. M., Christian, J. R., Dore, J. E., Hebel, D. V., Letelier, R. M., Tupas, L. M., and Winn, C. D. (1996). Seasonal and interannual variability in primary production and particle flux at Station ALOHA. *Deep Sea Research Part II: Topical Studies in Oceanography*, 43(2):539–568.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., Vargas, C. D., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E. G., Sardet, C., Sieracki, M. E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P., and Consortium, t. T. O. (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology*, 9(10):e1001177.

Karsenti, E. and Di Meo, D. (2012). *Tara océans : Chroniques d'une expédition scientifique*. Actes Sud Editions, Arles.

Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., and Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences*, 100(20):11394–11399.

Kemp, A. E. S. and Villareal, T. A. (2013). High diatom production and export in stratified waters – A potential negative feedback to global warming. *Progress in Oceanography*, 119:4–23.

Khatiwala, S., Tanhua, T., Fletcher, S. M., Gerber, M., Doney, S. C., Graven, H. D., Gruber, N., McKinley, G. A., Murata, A., Ríos, A. F., and Sabine, C. L. (2013). Global ocean storage of anthropogenic carbon. *Biogeosciences (BG)*, 10(4):2169–2191.

Klaas, C. and Archer, D. E. (2002). Association of sinking organic matter with various types of mineral ballast in the deep sea: Implications for the rain ratio. *Global Biogeochemical Cycles*, 16(4):63–1–63–14.

Kobari, T., Kitamura, M., Minowa, M., Isami, H., Akamatsu, H., Kawakami, H., Matsumoto, K., Wakita, M., and Honda, M. C. (2013). Impacts of the wintertime mesozooplankton community to downward carbon flux in the subarctic and subtropical Pacific Oceans. *Deep Sea Research Part I: Oceanographic Research Papers*, 81:78–88.

Koeve, W. (2001). Wintertime nutrients in the North Atlantic—new approaches and implications for new production estimates. *Marine Chemistry*, 74(4):245–260.

Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W., and Pržulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, 7(50):1341–1354.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226.

Kuypers, M. M. M., Sliekers, A. O., Lavik, G., Schmid, M., Jørgensen, B. B., Kuenen, J. G., Damsté, J. S. S., Strous, M., and Jetten, M. S. M. (2003). Anaerobic ammonium oxidation by anammox bacteria in the Black Sea. *Nature*, 422(6932):608.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559.

Lara, E., Vaqué, D., Sà, E. L., Boras, J. A., Gomes, A., Borrull, E., Díez-Vives, C., Teira, E., Pernice, M. C., Garcia, F. C., Forn, I., Castillo, Y. M., Peiró, A., Salazar, G., Morán, X. A. G., Massana, R., Catalá, T. S., Luna, G. M., Agustí, S., Estrada, M., Gasol, J. M., and Duarte, C. M. (2017). Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Science Advances*, 3(9).

Laufkötter, C., John, J. G., Stock, C. A., and Dunne, J. P. (2017). Temperature and oxygen dependence of the remineralization of organic matter. *Global Biogeochemical Cycles*, 31(7):1038–1050.

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, 25(3):217–228.

Le Quere, C., Raupach, M., Canadell, J. G., Marland, G., Bopp, L., Ciais, P., Friedlingstein, P., Viovy, N., Conway, T. J., Doney, S. C., Feely, R. A., Foster, P., House, J. I., Prentice, C. I., Gurney, K., Houghton, R. A., Huntingford, C., Levy, P. E., Lomas, M. R., Woodward, F. I., Majkut, J., Sarmiento, J. L., Metzl, N., Ometto, J. P., Randerson, J. T., Peters, G. P., Running, S., Sitch, S., Takahashi, T., and Van der Werf, G. (2009). Trends in the sources and sinks of carbon dioxide. *Nature Geoscience*, 2(12).

Leblanc, K., Quéguiner, B., Diaz, F., Cornet, V., Michel-Rodriguez, M., Madron, X. D. d., Bowler, C., Malviya, S., Thyssen, M., Grégori, G., Rembauville, M., Grosso, O., Poulain, J., Vargas, C. d., Pujo-Pay, M., and Conan, P. (2018). Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. *Nature Communications*, 9(1):1–12.

LeCleir, G. R., DeBruyn, J. M., Maas, E. W., Boyd, P. W., and Wilhelm, S. W. (2014). Temporal changes in particle-associated microbial communities after interception by nonlethal sediment traps. *FEMS Microbiology Ecology*, 87(1):153–163.

Legendre, L., Rivkin, R. B., Weinbauer, M. G., Guidi, L., and Uitz, J. (2015). The microbial carbon pump concept: Potential biogeochemical significance in the globally changing ocean. *Progress in Oceanography*, 134:432–450.

Lemaitre, N. (2017). *Multi-proxy approach (Thorium-234, excess Barium) of export and remineralization fluxes of carbon and biogenic elements associated with the oceanic biological pump*. PhD thesis, Université de Bretagne Occidentale.

Lemaitre, N., Planquette, H., Planchon, F., Sarthou, G., Jacquet, S., García-Ibáñez, M. I., Gourain, A., Cheize, M., Monin, L., André, L., Laha, P., Terryn, H., and Dehairs, F. (2018). Particulate barium tracing of significant mesopelagic carbon remineralisation in the North Atlantic. *Biogeosciences*, 15(8):2289–2307.

Lepère, C., Ostrowski, M., Hartmann, M., Zubkov, M. V., and Scanlan, D. J. (2016). In situ associations between marine photosynthetic picoeukaryotes and potential parasites – a role for fungi? *Environmental Microbiology Reports*, 8(4):445–451.

Li, C., Lim, K. M. K., Chng, K. R., and Nagarajan, N. (2016). Predicting microbial interactions through computational approaches. *Methods*, 102:12–19.

Liaw, A. and Wiener, M. (2001). Classification and Regression by RandomForest. *Forest*, 23.

Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A. M., Coppola, L., Cornejo-Castillo, F. M., d'Ovidio, F., Meester, L. D., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T. O., Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S. G., Sunagawa, S., Bork, P., Sullivan, M. B., Karsenti, E., Bowler, C., Vargas, C. d., and Raes, J. (2015). Determinants of community structure in the global plankton interactome. *Science*, 348(6237):1262073.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *arXiv:1006.3316 [stat]*.

Liu, J., Fu, B., Yang, H., Zhao, M., He, B., and Zhang, X.-H. (2015a). Phylogenetic shifts of bacterioplankton community composition along the Pearl Estuary: the potential impact of hypoxia and nutrients. *Frontiers in Microbiology*, 6.

Liu, Y., Lai, Q., Göker, M., Meier-Kolthoff, J. P., Wang, M., Sun, Y., Wang, L., and Shao, Z. (2015b). Genomic insights into the taxonomic status of the *Bacillus cereus* group. *Scientific Reports*, 5:14082.

Loeb, V., Siegel, V., Holm-Hansen, O., Hewitt, R., Fraser, W., Trivelpiece, W., and Trivelpiece, S. (1997). Effects of sea-ice extent and krill or salp dominance on the Antarctic food web. *Nature*, 387(6636):897–900.

Loeng, H., Brander, K., Carmack, E., Denisenko, S., Drinkwater, K., Hansen, B., Kovacs, K., and Livingstone, P. (2005). Marine systems. In Sundquist, E. T. and Broecker, W. S., editors, *Arctic Climate Impact Assessment*, pages 454–522. Cambridge University Press.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., Raes, J., Poulain, J., Jaillon, O., Wincker, P., Kandels-Lewis, S., Karsenti, E., Bork, P., and Acinas, S. G. (2014). Metagenomic 16s rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, 16(9):2659–2671.

Lohrenz, S. E., Knauer, G. A., Asper, V. L., Tuel, M., Michaels, A. F., and Knap, A. H. (1992). Seasonal variability in primary production and particle flux in the northwestern Sargasso Sea: U.S. JGOFS Bermuda Atlantic time-series study. *Deep Sea Research Part A. Oceanographic Research Papers*, 39(7):1373–1391.

Long, R. A. and Azam, F. (2001). Antagonistic interactions among marine pelagic bacteria. *Applied and Environmental Microbiology*, 67(11):4975–4983.

Long, R. A., Eveillard, D., Franco, S. L. M., Reeves, E., and Pinckney, J. L. (2013). Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats. *FEMS Microbiology Ecology*, 83(1):74–81.

Long, R. A., Rowley, D. C., Zamora, E., Liu, J., Bartlett, D. H., and Azam, F. (2005). Antagonistic Interactions among Marine Bacteria Impede the Proliferation of Vibrio cholerae. *Appl. Environ. Microbiol.*, 71(12):8531–8536.

Louca, S., Doebeli, M., and Parfrey, L. W. (2018). Correcting for 16s rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6.

Lupatini, M., Suleiman, A. K. A., Jacques, R. J. S., Antoniolli, Z. I., de Siqueira Ferreira, A., Kuramae, E. E., and Roesch, L. F. W. (2014). Network topology reveals high connectance levels and few key microbial genera within soils. *Frontiers in Environmental Science*, 2.

Lutz, M., Dunbar, R., and Caldeira, K. (2002). Regional variability in the vertical flux of particulate organic carbon in the ocean interior. *Global Biogeochemical Cycles*, 16(3):11–1–11–18.

Lutz, M. J., Caldeira, K., Dunbar, R. B., and Behrenfeld, M. J. (2007). Seasonal rhythms of net primary production and particulate organic carbon flux to depth describe the efficiency of biological pump in the global ocean. *Journal of Geophysical Research: Oceans*, 112(C10).

Ma, B., Wang, H., Dsouza, M., Lou, J., He, Y., Dai, Z., Brookes, P. C., Xu, J., and Gilbert, J. A. (2016). Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *The ISME Journal*, 10(8):1891–1901.

Ma, Z. S. and Ye, D. (2017). Trios—promising in silico biomarkers for differentiating the effect of disease on the human microbiome network. *Scientific Reports*, 7.

Mahadevan, A., D'Asaro, E., Lee, C., and Perry, M. J. (2012). Eddy-Driven Stratification Initiates North Atlantic Spring Phytoplankton Blooms. *Science*, 337(6090):54–58.

Malod-Dognin, N. and Pržulj, N. (2015). L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics*, 31(13):2182–2189.

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., Bihouée, A., Jean, G., Díaz, F. P., Fernández-Gómez, B., Cabrera, P., Gaete, A., Latorre, C., Gutiérrez, R. A., Maass, A., Cambiazo, V., Navarrete, S. A., Eveillard, D., and González, M. (2018). Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific Reports*, 8(1):5875.

Martin, J. H., Knauer, G. A., Karl, D. M., and Broenkow, W. W. (1987). VERTEX: carbon cycling in the northeast Pacific. *Deep Sea Research Part A. Oceanographic Research Papers*, 34(2):267–285.

Martin, P., Lampitt, R. S., Jane Perry, M., Sanders, R., Lee, C., and D'Asaro, E. (2011). Export and mesopelagic particle flux during a North Atlantic spring diatom bloom. *Deep Sea Research Part I: Oceanographic Research Papers*, 58(4):338–349.

Mason, O. and Verwoerd, M. (2006). Graph Theory and Networks in Biology. *arXiv:q-bio/0604006*. arXiv: q-bio/0604006.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W. H. C. F., Logares, R., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M. C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R., Simon, N., Stoeck, T., Vaulot, D., Zingone, A., and de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, 17(10):4035–4049.

Mayali, X. and Azam, F. (2004). Algicidal Bacteria in the Sea and their Impact on Algal Blooms. *Journal of Eukaryotic Microbiology*, 51(2):139–144.

McCave, I. N. (1984). Size spectra and aggregation of suspended particles in the deep ocean. *Deep Sea Research Part A. Oceanographic Research Papers*, 31(4):329–352.

McGillicuddy, D. J. and Robinson, A. R. (1997). Eddy-induced nutrient supply and new production in the Sargasso Sea. *Deep Sea Research Part I: Oceanographic Research Papers*, 44(8):1427–1450.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meng, L., Striegel, A., and Milenković, T. (2016). Local versus global biological network alignment. *Bioinformatics*, 32(20):3155–3164.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.

Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E. D., Geider, R. J., Guieu, C., Jaccard, S. L., Jickells, T. D., La Roche, J., Lenton, T. M., Mahowald, N. M., Marañón, E., Marinov, I., Moore, J. K., Nakatsuka, T., Oschlies, A., Saito,

M. A., Thingstad, T. F., Tsuda, A., and Ulloa, O. (2013). Processes and patterns of oceanic nutrient limitation. *Nature Geoscience*, 6(9):701–710.

Morales-Castilla, I., Matias, M. G., Gravel, D., and Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, 30(6):347–356.

Mordret, S., Romac, S., Henry, N., Colin, S., Carmichael, M., Berney, C., Audic, S., Richter, D. J., Pochon, X., de Vargas, C., and Decelle, J. (2016). The symbiotic life of Symbiodinium in the open ocean within a new species of calcifying ciliate (Tiarina sp.). *The ISME Journal*, 10(6):1424–1436.

Morris, B. E. L., Henneberger, R., Huber, H., and Moissl-Eichinger, C. (2013). Microbial syntrophy: interaction for the common good. *FEMS Microbiology Reviews*, 37(3):384–406.

Morán, X. A. G., Fernández, E., and Pérez, V. (2004). Size-fractionated primary production, bacterial production and net community production in subtropical and tropical domains of the oligotrophic NE Atlantic in autumn. *Marine Ecology Progress Series*, 274:17–29.

Mühlenbruch, M., Grossart, H.-P., Eigemann, F., and Voss, M. (2018). Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. *Environmental Microbiology*, 20(8):2671–2685.

Nadell, C. D., Xavier, J. B., and Foster, K. R. (2009). The sociobiology of biofilms. *FEMS Microbiology Reviews*, 33(1):206–224.

NASA (2016). Nasa Earth Observatory. https://earthobservatory.nasa.gov/images/88316/the-barents-sea-abloom.

Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A., and Quéguiner, B. (1995). Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, 9(3):359–372.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, 67(2).

Obernosterer, I., Christaki, U., LefÃšvre, D., Catala, P., Van Wambeke, F., and Lebaron, P. (2008). Rapid bacterial mineralization of organic carbon produced during a phytoplankton bloom induced by natural iron fertilization in the Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 55(5):777–789.

Ottesen, E. A. (2016). Probing the living ocean with ecogenomic sensors. *Current Opinion in Microbiology*, 31:132–139.

Ottesen, E. A., Young, C. R., Eppley, J. M., Ryan, J. P., Chavez, F. P., Scholin, C. A., and DeLong, E. F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6):E488–E497.

Ottesen, E. A., Young, C. R., Gifford, S. M., Eppley, J. M., Marin, R., Schuster, S. C., Scholin, C. A., and DeLong, E. F. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science*, 345(6193):207–212.

Paine, R. T. (1969). A Note on Trophic Complexity and Community Stability. *The American Naturalist*, 103(929):91–93.

Passow, U. (2002). Transparent exopolymer particles (TEP) in aquatic environments. *Progress in Oceanography*, 55(3):287–333.

Passow, U. and Carlson, C. A. (2012). The biological pump in a high CO2 world. *Marine Ecology Progress Series*, 470:249–271.

Passow, U., Shipe, R. F., Murray, A., Pak, D. K., Brzezinski, M. A., and Alldredge, A. L. (2001). The origin of transparent exopolymer particles (TEP) and their role in the sedimentation of particulate matter. *Continental Shelf Research*, 21(4):327–346.

Perez, S. I. E. (2015). *Exploring microbial community structure and resilience through visualization and analysis of microbial co-occurrence networks*. PhD thesis, University of British Columbia.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Bescot, N. L., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., and Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2:150023.

Peters, F. (2008). Diatoms in a future ocean — stirring it up. *Nature Reviews Microbiology*, 6(5):407.

Petersen, C. G. J. (1915). A preliminary result of the investigations on the valuation of the sea. *Rep. Danish. Biol. Sta.*, 23:27–29.

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J.-M., Basile, I., Bender, M., Chappellaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V. M., Legrand, M., Lipenkov, V. Y., Lorius, C., PÉpin, L., Ritz, C., Saltzman, E., and Stievenard, M. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, 399(6735):429.

Picheral, M., Guidi, L., Stemmann, L., Karl, D. M., Iddaoud, G., and Gorsky, G. (2010). The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods*, 8(9):462–473.

Pichevin, L. E., Reynolds, B. C., Ganeshram, R. S., Cacho, I., Pena, L., Keefe, K., and Ellam, R. M. (2009). Enhanced carbon pump inferred from relaxation of nutrient limitation in the glacial ocean. *Nature*, 459(7250):1114–1117.

Pierce, W. D. W. D., Cushman, R. A. R. A., and Hood, C. E. (1912). *The insect enemies of the cotton boll weevil*. Washington, D.C. : U.S. Dept. of Agriculture, Bureau of Entomology.

Platt, T. (1986). Primary production of the ocean water column as a function of surface light intensity: algorithms for remote sensing. *Deep Sea Research Part A. Oceanographic Research Papers*, 33(2):149–163.

Poelen, J. H., Simons, J. D., and Mungall, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24:148–159.

Poulsen, L. K. and Kiørboe, T. (2006). Vertical flux and degradation rates of copepod fecal pellets in a zooplankton community dominated by small copepods. *Marine Ecology Progress Series*, 323:195–204.

Power, M. E., Tilman, D., Estes, J. A., Menge, B. A., Bond, W. J., Mills, L. S., Daily, G., Castilla, J. C., Lubchenco, J., and Paine, R. T. (1996). Challenges in the Quest for Keystones. *BioScience*, 46(8):609–620.

Proctor, L. M. and Fuhrman, J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature*, 343(6253):60.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Melo Minardi, R., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., Zoetendal, E., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue):D590–D596.

Quéré, C. L., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Cotrim Da Cunha, L., Geider, R., Giraud, X., Klaas, C., Kohfeld, K. E., Legendre, L., Manizza, M., Platt, T., Rivkin, R. B., Sathyendranath, S., Uitz, J., Watson, A. J., and Wolf-Gladrow, D. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, 11(11):2016–2040.

Rappé, M. S., Vergin, K., and Giovannoni, S. J. (2000). Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems. *FEMS Microbiology Ecology*, 33(3):219–232.

Raven, J. A. and Falkowski, P. G. (1999). Oceanic sinks for atmospheric CO2. *Plant, Cell & Environment*, 22(6):741–755.

Riebesell, U., Körtzinger, A., and Oschlies, A. (2009). Sensitivities of marine carbon fluxes to ocean change. *Proceedings of the National Academy of Sciences*, 106(49):20602–20609.

Rivkin, R. B. and Legendre, L. (2001). Biogenic Carbon Cycling in the Upper Ocean: Effects of Microbial Respiration. *Science*, 291(5512):2398–2400.

Robidart, J. C., Church, M. J., Ryan, J. P., Ascani, F., Wilson, S. T., Bombar, D., Marin, R., Richards, K. J., Karl, D. M., Scholin, C. A., and Zehr, J. P. (2014). Ecogenomic sensor reveals controls on N2-fixing microorganisms in the North Pacific Ocean. *The ISME Journal*, 8(6):1175–1185.

Robidart, J. C., Preston, C. M., Paerl, R. W., Turk, K. A., Mosier, A. C., Francis, C. A., Scholin, C. A., and Zehr, J. P. (2012). Seasonal Synechococcus and Thaumarchaeal population dynamics examined with high resolution with remote in situ instrumentation. *The ISME Journal*, 6(3):513–523.

Robinson, C., Steinberg, D. K., Anderson, T. R., Arístegui, J., Carlson, C. A., Frost, J. R., Ghiglione, J.-F., Hernández-León, S., Jackson, G. A., Koppelmann, R., Quéguiner, B., Ragueneau, O., Rassoulzadegan, F., Robison, B. H., Tamburini, C., Tanaka, T., Wishner, K. F., and Zhang, J. (2010). Mesopelagic zone ecology and biogeochemistry – a synthesis. *Deep Sea Research Part II: Topical Studies in Oceanography*, 57(16):1504–1518.

Rochelle-Newall, E. J., Delesalle, B., Mari, X., Rouchon, C., Torréton, J.-P., and Pringault, O. (2008). Zinc induces shifts in microbial carbon flux in tropical coastal environments. *Aquatic Microbial Ecology*, 52(1):57–68.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104.

Rohwer, F., Prangishvili, D., and Lindell, D. (2009). Roles of viruses in the environment. *Environmental Microbiology*, 11(11):2771–2774.

Roxburgh, S. H., Berry, S. L., Buckley, T. N., Barnes, B., and Roderick, M. L. (2005). What is NPP? Inconsistent accounting of respiratory fluxes in the definition of net primary production. *Functional Ecology*, 19(3):378–382.

Röttjers, L. and Faust, K. (2018). From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews*, 42(6):761–780.

Sakshaug, E. (2004). Primary and Secondary Production in the Arctic Seas. In Stein, R. and MacDonald, R. W., editors, *The Organic Carbon Cycle in the Arctic Ocean*, pages 57–81. Springer Berlin Heidelberg, Berlin, Heidelberg.

Saraph, V. and Milenković, T. (2014). MAGNA: Maximizing Accuracy in Global Network Alignment. *Bioinformatics*, 30(20):2931–2940.

Sarmiento, J. L. and Gruber, N. (2006). *Ocean Biogeochemical Dynamics*. Princeton University Press, Princeton.

Sarmiento, J. L., Gruber, N., Brzezinski, M. A., and Dunne, J. P. (2004). High-latitude controls of thermocline nutrients and low latitude biological productivity. *Nature*, 427(6969):56–60.

Sauzède, R., Bittig, H. C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J.-P., Legendre, L., and Johnson, K. S. (2017). Estimates of Water-Column Nutrient Concentrations and Carbonate System Parameters in the Global Ocean: A Novel Approach Based on Neural Networks. *Frontiers in Marine Science*, 4.

Scholin, C. (2009). What are "ecogenomic sensors?" – a review and thoughts for the future. *Ocean Science Discussions*, 6:191–213.

Schwager, E., Bielski, C., and Weingart, G. (2019). *ccrepe: ccrepe_and_nc.score*. R package version 1.18.1.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., and Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. In *2008 19th International Conference on Pattern Recognition*, pages 1–4.

Seymour, J. R., Amin, S. A., Raina, J.-B., and Stocker, R. (2017). Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology*, 2(7):17065.

Shanks, A. L. (2002). The abundance, vertical flux, and still-water and apparent sinking rates of marine snow in a shallow coastal water column. *Continental Shelf Research*, 22(14):2045–2064.

Sheldon, R. W., Prakash, A., and Sutcliffe, W. H. (1972). The Size Distribution of Particles in the Ocean1. *Limnology and Oceanography*, 17(3):327–340.

Shelford, V. E. (1913). *Animal communities in temperate America, as illustrated in the Chicago region ; a study in animal ecology,*. The Geographic Society of Chicago by the University of Chicago Press, Chicago, Illinois.

Sher, D., Thompson, J. W., Kashtan, N., Croal, L., and Chisholm, S. W. (2011). Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. *The ISME Journal*, 5(7):1125–1132.

Siegel, D. A., Buesseler, K. O., Behrenfeld, M. J., Benitez-Nelson, C. R., Boss, E., Brzezinski, M. A., Burd, A., Carlson, C. A., D'Asaro, E. A., Doney, S. C., Perry, M. J., Stanley, R. H. R., and Steinberg, D. K. (2016). Prediction of the Export and Fate of Global Ocean Net Primary Production: The EXPORTS Science Plan. *Frontiers in Marine Science*, 3.

Siegel, D. A., Buesseler, K. O., Doney, S. C., Sailley, S. F., Behrenfeld, M. J., and Boyd, P. W. (2014). Global assessment of ocean carbon export by combining satellite observations and food-web models. *Global Biogeochemical Cycles*, 28(3):181–196.

Siegel, D. A. and Deuser, W. G. (1997). Trajectories of sinking particles in the Sargasso Sea: modeling of statistical funnels above deep-ocean sediment traps. *Deep Sea Research Part I: Oceanographic Research Papers*, 44(9):1519–1541.

Siegenthaler, U. and Sarmiento, J. L. (1993). Atmospheric carbon dioxide and the ocean. *Nature*, 365(6442):119–125.

Sigman, D. M. and Boyle, E. A. (2000). Glacial/interglacial variations in atmospheric carbon dioxide. *Nature*, 407(6806):859–869.

Sigman, D. M., Hain, M. P., and Haug, G. H. (2010). The polar ocean and glacial cycles in atmospheric CO_2 concentration. *Nature*, 466(7302):47–55.

Simon, M., Grossart, H.-P., Schweitzer, B., and Ploug, H. (2002). Microbial ecology of organic aggregates in aquatic ecosystems. *Aquatic Microbial Ecology*, 28(2):175–211.

Smith, C. R., De Leo, F. C., Bernardino, A. F., Sweetman, A. K., and Arbizu, P. M. (2008). Abyssal food limitation, ecosystem structure and climate change. *Trends in Ecology & Evolution*, 23(9):518–528.

Smith, D., Simon, M., L. Alldredge, A., and Azam, F. (1992). Intense hydrolytic enzyme activity on marine aggregates and implications for rapid particle dissolution. *Nature*, 359.

Smith, R. C., Eppley, R. W., and Baker, K. S. (1982). Correlation of primary production as measured aboard ship in Southern California Coastal waters and as estimated from satellite chlorophyll images. *Marine Biology*, 66(3):281–288.

Soutar, A., Kling, S. A., Crill, P. A., Duffrin, E., and Bruland, K. W. (1977). Monitoring the marine environment through sedimentation. *Nature*, 266(5598):136–139.

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C.-E. T., Sachdeva, R., Jones, A. C., Schwalbach, M. S., Rose, J. M., Hewson, I., Patel, A., Sun, F., Caron, D. A., and Fuhrman, J. A. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal*, 5(9):1414–1425.

Steele, J. H. (1974). *The structure of marine ecosystems*. Harvard Univ. Press, Massachusetts.

Steemann Nielsen, H. E. (1952). The Use of Radio-active Carbon (C14) for Measuring Organic Production in the Sea. *ICES Journal of Marine Science*, 18(2):117–140.

Steinberg, D. K., Van Mooy, B. A. S., Buesseler, K. O., Boyd, P. W., Kobari, T., and Karl, D. M. (2008). Bacterial vs. zooplankton control of sinking particle flux in the ocean's twilight zone. *Limnology and Oceanography*, 53(4):1327–1338.

Stemmann, L., Youngbluth, M., Robert, K., Hosia, A., Picheral, M., Paterson, H., Ibanez, F., Guidi, L., Lombard, F., and Gorsky, G. (2008). Global zoogeography of fragile macrozooplankton in the upper 100–1000 m inferred from the underwater video profiler. *ICES Journal of Marine Science*, 65(3):433–442.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., and Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, 19 Suppl 1:21–31.

Stramma, L., Johnson, G. C., Sprintall, J., and Mohrholz, V. (2008). Expanding Oxygen-Minimum Zones in the Tropical Oceans. *Science*, 320(5876):655–658.

Suess, E. (1980). Particulate organic carbon flux in the oceans - Surface productivity and oxygen utilization. *Nature*, 288:260–263.

Summerhayes, V. S. and Elton, C. S. (1923). Contributions to the ecology of Spitsbergen and Bear Island. *Journal of Ecology*, 11(2):214–216.

Sun, D.-L., Jiang, X., Wu, Q. L., and Zhou, N.-Y. (2013). Intragenomic Heterogeneity of 16s rRNA Genes Causes Overestimation of Prokaryotic Diversity. *Applied and Environmental Microbiology*, 79(19):5962–5969.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T. O., Bowler, C., Vargas, C. d., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Tackmann, J., Rodrigues, J. F. M., and Mering, C. v. (2018). Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *bioRxiv*, page 390195.

Tamaddoni-Nezhad, A., Milani, G. A., Raybould, A., Muggleton, S., and Bohan, D. A. (2013). Chapter Four - Construction and Validation of Food Webs Using Logic-Based Machine Learning and Text Mining. In Woodward, G. and Bohan, D. A., editors, *Advances in Ecological Research*, volume 49 of *Ecological Networks in an Agricultural World*, pages 225–289. Academic Press.

Tamburini, C., Goutx, M., Guigue, C., Garel, M., Lefèvre, D., Charrière, B., Sempéré, R., Pepa, S., Peterson, M. L., Wakeham, S., and Lee, C. (2009). Effects of hydrostatic pressure on microbial alteration of sinking fecal pellets. *Deep Sea Research Part II: Topical Studies in Oceanography*, 56(18):1533–1546.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., Kassabgy, M., Huang, S., Mann, A. J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F. D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K. H., Glöckner, F. O., Schweder, T., and Amann, R. (2012). Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science*, 336(6081):608–611.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J. T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M. F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N. A., Lefler, J., Brislawn, C. J., Humphrey, G., Owens, S. M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J. A., Clauset, A., Stevens, R. L., Shade, A., Pollard, K. S., Goodwin, K. D., Jansson, J. K., Gilbert, J. A., Knight, R., The Earth Microbiome Project Consortium, Rivera, J. L. A., Al-Moosawi, L., Alverdy, J., Amato, K. R., Andras, J., Angenent, L. T., Antonopoulos, D. A., Apprill, A., Armitage, D., Ballantine, K., Bárta, J., Baum, J. K., Berry, A., Bhatnagar, A., Bhatnagar, M., Biddle, J. F., Bittner, L., Boldgiv, B., Bottos, E., Boyer, D. M., Braun, J., Brazelton, W., Brearley, F. Q., Campbell, A. H., Caporaso, J. G., Cardona, C., Carroll, J., Cary, S. C., Casper, B. B., Charles, T. C., Chu, H., Claar, D. C., Clark, R. G., Clayton, J. B., Clemente, J. C., Cochran, A., Coleman, M. L., Collins, G., Colwell, R. R., Contreras, M., Crary, B. B., Creer, S., Cristol, D. A., Crump, B. C., Cui, D., Daly, S. E., Davalos, L., Dawson, R. D., Defazio, J., Delsuc, F., Dionisi, H. M., Dominguez-Bello, M. G., Dowell, R., Dubinsky, E. A., Dunn, P. O., Ercolini, D., Espinoza, R. E., Ezenwa, V., Fenner, N., Findlay, H. S., Fleming, I. D., Fogliano, V., Forsman, A., Freeman, C., Friedman, E. S., Galindo, G., Garcia, L., Garcia-Amado, M. A., Garshelis, D., Gasser, R. B., Gerdts, G., Gibson, M. K., Gifford, I., Gill, R. T., Giray, T., Gittel, A., Golyshin, P., Gong, D., Grossart, H.-P., Guyton, K., Haig, S.-J., Hale, V., Hall, R. S., Hallam, S. J., Handley, K. M., Hasan, N. A., Haydon, S. R., Hickman, J. E., Hidalgo, G., Hofmockel, K. S., Hooker, J., Hulth, S., Hultman, J., Hyde, E., Ibáñez-Álamo, J. D., Jastrow, J. D., Jex, A. R., Johnson, L. S., Johnston, E. R., Joseph, S., Jurburg, S. D., Jurelevicius, D., Karlsson, A., Karlsson, R., Kauppinen, S., Kellogg, C. T. E., Kennedy, S. J., Kerkhof, L. J., King, G. M., Kling, G. W., Koehler, A. V., Krezalek, M., Kueneman, J., Lamendella, R., Landon, E. M., Lane-deGraaf, K., LaRoche, J., Larsen, P., Laverock, B., Lax, S., Lentino, M., Levin, I. I., Liancourt, P., Liang, W., Linz, A. M., Lipson, D. A., Liu, Y., Lladser, M. E., Lozada, M., Spirito, C. M., MacCormack, W. P., MacRae-Crerar, A., Magris, M., Martín-Platero, A. M., Martín-Vivaldi, M., Martínez, L. M., Martínez-Bueno, M., Marzinelli, E. M., Mason, O. U., Mayer, G. D., McDevitt-Irwin, J. M., McDonald, J. E., McGuire, K. L., McMahon, K. D., McMinds, R., Medina, M., Mendelson, J. R., Metcalf, J. L., Meyer, F., Michelangeli, F., Miller, K., Mills, D. A., Minich, J., Mocali, S., Moitinho-Silva, L., Moore, A., Morgan-Kiss, R. M., Munroe, P., Myrold, D., Neufeld, J. D., Ni, Y., Nicol, G. W., Nielsen, S., Nissimov, J. I., Niu, K., Nolan, M. J., Noyce, K., O'Brien, S. L., Okamoto, N., Orlando, L., Castellano, Y. O., Osuolale, O., Oswald, W., Parnell, J., Peralta-Sánchez, J. M., Petraitis, P., Pfister, C., Pilon-Smits, E., Piombino, P., Pointing, S. B., Pollock, F. J., Potter, C., Prithiviraj, B., Quince, C., Rani, A., Ranjan, R., Rao, S., Rees, A. P., Richardson, M., Riebesell, U., Robinson, C., Rockne, K. J., Rodriguezl, S. M., Rohwer, F., Roundstone, W., Safran, R. J., Sangwan, N., Sanz, V., Schrenk, M., Schrenzel, M. D., Scott, N. M., Seger, R. L., Seguin-Orlando, A., Seldin, L., Seyler, L. M., Shakhsheer, B., Sheets, G. M., Shen, C., Shi, Y., Shin, H., Shogan, B. D., Shutler, D., Siegel, J., Simmons, S., Sjöling, S., Smith, D. P., Soler, J. J., Sperling, M., Steinberg, P. D., Stephens, B., Stevens, M. A., Taghavi, S., Tai, V., Tait, K., Tan, C. L., Taş, N., Taylor, D. L., Thomas, T., Timling, I., Turner, B. L., Urich, T., Ursell,

L. K., van der Lelie, D., Van Treuren, W., van Zwieten, L., Vargas-Robles, D., Thurber, R. V., Vitaglione, P., Walker, D. A., Walters, W. A., Wang, S., Wang, T., Weaver, T., Webster, N. S., Wehrle, B., Weisenhorn, P., Weiss, S., Werner, J. J., West, K., Whitehead, A., Whitehead, S. R., Whittingham, L. A., Willerslev, E., Williams, A. E., Wood, S. A., Woodhams, D. C., Yang, Y., Zaneveld, J., Zarraonaindia, I., Zhang, Q., and Zhao, H. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463.

Thompson, R. M., Brose, U., Dunne, J. A., Hall, R. O., Hladyz, S., Kitching, R. L., Martinez, N. D., Rantala, H., Romanuk, T. N., Stouffer, D. B., and Tylianakis, J. M. (2012). Food webs: reconciling the structure and function of biodiversity. *Trends in Ecology & Evolution*, 27(12):689–697.

Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., Bonneau, R., and Ghedin, E. (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*, 6.

Tremblay, J.-E., Simpson, K., Martin, J., Miller, L., Gratton, Y., Barber, D., and Price, N. M. (2018). Vertical stability and the annual dynamics of nutrients and chlorophyll fluorescence in the coastal, southeast Beaufort Sea. *Journal of Geophysical Research: Oceans*.

Tréguer, P., Legendre, L., Rivkin, R. T., Ragueneau, O., and Dittert, N. (2003). Water Column Biogeochemistry below the Euphotic Zone. In Fasham, M. J. R., editor, *Ocean Biogeochemistry: The Role of the Ocean Carbon Cycle in Global Change*, Global Change — The IGBP Series (closed), pages 145–156. Springer Berlin Heidelberg, Berlin, Heidelberg.

Turley, C. and Mackie, P. (1994). Biogeochemical significance of attached and free-living bacteria and the flux of particles in the NE Atlantic Ocean. *Marine Ecology-progress Series*, 115:191–203.

Turner, J. T. (2002). Zooplankton fecal pellets, marine snow and sinking phytoplankton blooms. *Aquatic Microbial Ecology*, 27(1):57–102.

Turner, J. T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Progress in Oceanography*, 130:205–248.

Uitz, J., Claustre, H., Gentili, B., and Stramski, D. (2010). Phytoplankton class-specific primary production in the world's oceans: Seasonal and interannual variability from satellite observations. *Global Biogeochemical Cycles*, 24(3).

Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Dubois Peyrard, N., Moalic, Y., Sabbadin, R., Schwaller, L., Chiquet, J., Alex Smith, M., Vallance, J., Fievet, V., Jakuschkin, B., and Bohan, D. A. (2016). Learning ecological networks from next-generation sequencing data. In *Ecosystem Services: From Biodiversity to Society, Part 2*, volume 54 of *Advances In Ecological Research*.

Vandermeer, J. H. (1969). The Competitive Structure of Communities: An Experimental Approach with Protozoa. *Ecology*, 50(3):362–371.

Vargas, C. d., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N. L., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T. O., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

Vergin, K. L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A. H., Kilpert, F., Worden, A. Z., and Giovannoni, S. J. (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *The ISME Journal*, 7(7):1322–1332.

Vincent, F. and Bowler, C. (2019). Diatoms structure the plankton community based on selective segregation in the world's ocean. *bioRxiv*, page 704353.

Vincent, F. J., Colin, S., Romac, S., Scalco, E., Bittner, L., Garcia, Y., Lopes, R. M., Dolan, J. R., Zingone, A., Vargas, C. d., and Bowler, C. (2018). The epibiotic life of the cosmopolitan diatom Fragilariopsis doliolus on heterotrophic ciliates in the open ocean. *The ISME Journal*, 12(4):1094–1108.

Volk, T. and Hoffert, M. I. (1985). Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean-driven atmospheric co2 changes. In Sundquist, E. T. and Broecker, W. S., editors, *The Carbon Cycle and Atmospheric CO2: Natural Variations Archean to Present*, pages 99–110. American Geophysical Union.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440.

Webb, S. (2018). Deep learning for biology. *Nature*, 554:555.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J., and Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10(7):1669–1681.

Westcott, S. L. and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16s rRNA gene sequences to operational taxonomic units. *PeerJ*, 3:e1487.

White, A. E., Watkins-Brandt, K. S., Engle, M. A., Burkhardt, B., and Paytan, A. (2012). Characterization of the Rate and Temperature Sensitivities of Bacterial Remineralization of Dissolved Organic Phosphorus Compounds by Natural Populations. *Frontiers in Microbiology*, 3.

Wiebe, P. H., Boyd, S. H., and Winget, C. (1976). Particulate matter sinking to the deep-sea floor at 2000 M in the Tongue of the Ocean, Bahamas, with a description of a new sedimentation trap. *Journal of Marine Research*, 34(03-04):341–354.

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

Zeebe, R. E. and Wolf-Gladrow, D. A. (2009). Carbon dioxide, dissolved (ocean). In *Encyclopedia of Paleoclimatology and Ancient Environments*, pages 123–127. Springer.

Zhang, C. L., Xie, W., Martin-Cuadrado, A.-B., and Rodriguez-Valera, F. (2015). Marine Group II Archaea, potentially important players in the global ocean carbon cycle. *Frontiers in Microbiology*, 6.

Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., and Zhi, X. (2010). Functional Molecular Ecological Networks. *mBio*, 1(4):e00169–10.

Ziegler, A., Koch, A., Krockenberger, K., and Großhennig, A. (2012). Personalized medicine using DNA biomarkers: a review. *Human Genetics*, 131(10):1627–1638.

Ziervogel, K. and Arnosti, C. (2008). Polysaccharide hydrolysis in aggregates and free enzyme activity in aggregate-free seawater from the north-eastern Gulf of Mexico. *Environmental Microbiology*, 10(2):289–299.

Zinger, L., Gobet, A., and Pommier, T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, 21(8):1878–1896.

Glossary

alveolate group of unicellular eukaryotes, characterized by flattened vesicles packed into a continuous layer supporting the membrane. This group gathers ciliates, dinoflagellates and apicomplexa. 64, 74

appendicularian free-swimming solitary tunicate that is characterized by neoteny and thus resembles the larvae of most tunicates, having a trunk and a tail. Its size is generally less that 1 cm in body length, excluding the tail. Appendicularians are filter-feeders like most tunicates. 26, 27, 31, 77

atmosphere gaseous layer wrapping the Earth, mainly composed of nitrogen, oxygen, argon and carbon dioxide. 15–17, 19–22, 35, 40

autotrophic an autotrophic organism produces organic matter by reducing inorganic matter. It generally uses carbon (in the form of carbon dioxide), nitrogen (in the form of nitrates or dinitrogen), water and mineral nutrients. The source of energy for reducing inorganic matter comes from the light (photoautotrophy) or chemical reactions (chemoautotrophy). 22, 45

benthos aquatic organisms living near or on the seafloor or the bottom of lakes and rivers. 43

biosphere dynamic system shaped by all ecosystems, i.e. all living beings and their environment. 15, 22, 38

bloom an algae bloom is a rapid increase of the population of algae that takes place under certain environmental conditions (e.g. increased temperature and light, nutrient enrichment). 7, 23, 26, 28, 33, 41, 44, 89, 131

coccolithophore single-celled marine algae that possess and external skeleton made of plates of calcium carbonate (coccoliths) which may have complex shapes. Coccolithophores belong to the class Prymnesiophyceae. 16, 20, 23, 44, 165, 168

coccolithophorid synonym of coccolithophore. 28, 36, 41–44

copepod group of small tear-drop shaped crustaceans whose size typically varies from 1 to 2 mm and that possess an exoskeleton, two pairs of large antennae and a single median compound eye at the center of their head. 27, 49, 77, 168

diatom unicellular algae that possess a silica cell wall (the frustule) made of two valves. Diatoms display various forms and can form colonies. 23, 26, 28, 33, 36, 37, 40, 41, 43, 49–51, 77, 121, 129–131

dinoflagellate unicellular eukaryote characterized by unique flagellar insertion and features of their nucleus. This group is recognized to include many mixotrophic species. 28, 43, 130, 165

dissolved organic matter dissolved organic matter is defined as the component of organic matter that passes through a filter of a given pore size, commonly 0.45 μm (although other pore size may be used, typically between 0.22 and 0.7 μm). It is generally non sinking and is transported by ocean circulation and mixing. Besides, the pool size of dissolved inorganic matter is far larger than particulate organic matter: it is nearly equal to total organic matter (Sarmiento and Gruber, 2006). 21, 167

euphausiid shrimp-like crustaceans that have an exoskeleton and stalked, compound eyes. All members of the order Euphausiacea are commonly referred to as euphausiids. 27, 34, 168

euphotic zone the euphotic zone, also called photic, sunlight or sunlit zone, is the uppermost layer of water in the ocean where light intensity is of at least 1%. Its depth is generally located between 100 and 200 m. 20, 23, 25, 27, 28, 32, 34, 35, 37, 44, 76, 77, 79, 80, 130

foraminifera protozoa characterized by an external perforated shell (test) that comprises one or multiple chambers. From the foramens of the test go out pseudopodes (projections of the cell membrane) that allow foraminifera to catch food and move. 16, 20

heterotrophic an heterotrophic organism cannot produce its own food, needing instead to feed on pre-existing organic matter. 20, 21, 33, 34, 41, 43–45, 91

hydrosphere all areas of a planet where water is present in liquid (e.g. ocean, rivers, lakes, ground water), solid (e.g. sea ice, glaciers) and gaseous form (water vapor). 15

kleptoplasty sequestration of algae chloroplasts by host organisms. 43

lithosphere outermost shell of a terrestrial-type planet, characterized by its rigid mechanical properties; it includes the crust and the uppermost mantle. 15

mesopelagic zone zone that extends from approximately 200 to 1000 m below the ocean surface, where light penetrates but is insufficient for photosynthesis. 35, 47, 76, 130

meta-omics refers to molecular techniques that aim at characterizing, quantifying and analyzing large quantities of biological molecules (DNA, RNA, proteins) from environmental samples. It includes metabarcoding (described in section 3.1.1.1), metagenomics (that consist in studying DNA), metatranscriptomics (that consist in studying translated genes, i.e. RNA) and metaproteomics (that consist in studying proteins). 49, 77, 101, 107

microorganism living organism invisible to the naked eye, that can only be observed by means of a microscope and may be single-celled or live in colonies. 21, 44–46, 49, 64, 130

microphytoplankton fraction of phytoplankton whose size is between 20 and 200 μm. 23, 37, 77

nanophytoplankton small phytoplankton whose size is between 2 and 20 μm. 23

nekton refers to all organisms that have the ability to swim and move against the currents (such as fishes, cephalopods and marine mammals). 43

particulate organic matter it is the component of organic matter that is isolated by filtration (see the definition of dissolved organic matter) and for which transport by sinking is important (Sarmiento and Gruber, 2006). 21, 29, 33

pelagic relative to the open sea. 20, 91, 131

phage virus infecting and replicating within bacteria and archaea. 64, 74

phytoplankton planktonic organisms that are able to produce their own organic matter by performing photosynthesis (they are photoautotrophic). 7, 17, 19–24, 27, 28, 35–37, 40, 43–46, 77, 107, 130, 131

picophytoplankton phytoplankton whose size is between 0.2 and 2 μm. 7, 23, 24, 37, 77

protist unicellular eukaryote. 27, 42, 130

protozoa single-celled heterotrophic eukaryote that feeds on other microorganisms, or organic tissues and debris. 20, 27, 43, 166

prymnesiophyte abundant nanoplanktonic group of phytoplankton (class Prymnesiophyceae, division Haptophyta) in many oceanic regions that includes coccolithophores. 23

pteropod free-swimming pelagic opistobranch gastropod whose foot has two large fins and actively swims. 20, 27

Radiolaria unicellular heterotrophic eukaryotes possessing intricate mineral skeletons usually made of silica and needle-like pseudopods. 64, 74, 168

salp planktonic tunicates that can form long and stringy colonies. 27, 31, 41

thermohaline circulation oceanic circulation induced by differences in density of seawater that arise from differences in temperature and salinity of water masses. 19

tunicate subphylum that is part of the Chordates and includes animals possessing a dorsal nerve chord and a notochord. 27, 165, 168

zooplankton heterotrophic plankton that are usually microscopic (e.g. Radiolaria) but can be larger (e.g. copepods, euphausiids), measuring up to meters such as some jellyfishes. 7, 20, 21, 25, 27, 31, 33, 34, 37, 41, 43, 44, 77, 107, 130

Appendices

Appendix A

Article 4 / Co-authored manuscript 1: Faure et al. 2019

The goal of this article was to better understand the environmental diversity of marine mixotrophic protists from metabarcoding and environmental data of the *Tara* Oceans expedition. This study confirmed previous findings showing that mixotrophic protists are ubiquitous in the global ocean while detecting them in biogeographical provinces where no morphological identification had been recorded before. These findings have potentially important implications for future estimations of carbon export in biogeochemical models, as mixotrophic protists are estimated to be responsible of up to 30% of the global carbon export.

For this co-authored manuscript, I was in charge of building the contextual dataset using the environmental variables available in the PANGAEA repository form the *Tara* Oceans expeditions. The original environmental variables were divided in eight files and were related to carbonate chemistry, pigment concentrations, nutrients, sensor data, mesoscale features, water column features, methodological context and sequencing methodology. To keep only one version of each variable that was calculated twice or more using different tools, units and/or formulas, I selected 83 out of the 235 variables available on PANGAEA. As several values were available for a same station and a same depth, I calculated the median of the samples belonging to the same station and depth was calculated to get a unique value for every station/depth. Besides, I calculated carbon export values[1] and added diversity data (Sunagawa et al. 2015, supplementary table W8 available on http://ocean-microbiome.embl.de/companion.html), validated iron data calculated using the Darwin model and CDOM data from Arctic samples calculated by Atsushi Matsuoka.

[1] Carbon export was calculated from flux profiles as the mean flux between 130m and 170m in a radius of 1km around the sampling location and 24h around the sampling date.

The ISME Journal
https://doi.org/10.1038/s41396-018-0340-5

ARTICLE

Mixotrophic protists display contrasted biogeographies in the global ocean

Emile Faure[1,2] · Fabrice Not[3] · Anne-Sophie Benoiston[2] · Karine Labadie[4] · Lucie Bittner[2] · Sakina-Dorothée Ayata[1,2]

Received: 2 July 2018 / Revised: 13 December 2018 / Accepted: 13 December 2018
© International Society for Microbial Ecology 2019

Abstract

Mixotrophy, or the ability to acquire carbon from both auto- and heterotrophy, is a widespread ecological trait in marine protists. Using a metabarcoding dataset of marine plankton from the global ocean, 318,054 mixotrophic metabarcodes represented by 89,951,866 sequences and belonging to 133 taxonomic lineages were identified and classified into four mixotrophic functional types: constitutive mixotrophs (CM), generalist non-constitutive mixotrophs (GNCM), endo-symbiotic specialist non-constitutive mixotrophs (eSNCM), and plastidic specialist non-constitutive mixotrophs (pSNCM). Mixotrophy appeared ubiquitous, and the distributions of the four mixotypes were analyzed to identify the abiotic factors shaping their biogeographies. Kleptoplastidic mixotrophs (GNCM and pSNCM) were detected in new zones compared to previous morphological studies. Constitutive and non-constitutive mixotrophs had similar ranges of distributions. Most lineages were evenly found in the samples, yet some of them displayed strongly contrasted distributions, both across and within mixotypes. Particularly divergent biogeographies were found within endo-symbiotic mixotrophs, depending on the ability to form colonies or the mode of symbiosis. We showed how metabarcoding can be used in a complementary way with previous morphological observations to study the biogeography of mixotrophic protists and to identify key drivers of their biogeography.

Introduction

Marine unicellular eukaryotes, or protists, have a tremendous range of life styles, sizes and forms [1], showing a taxonomic and functional diversity that remains hard to define [2, 3]. This variety of organisms is having an impact on major biogeochemical cycles such as carbon, oxygen, nitrogen, sulfur, silica, or iron, while being at the base of marine trophic networks [4–8]. Hence, they are key actors of the global functioning of the ocean.

Historically, marine protists have been classified into two groups depending on their trophic strategy: the photosynthetic plankton (phytoplankton) and the heterotrophic plankton (zooplankton). It is now clear that mixotrophy, i.e., the ability to combine autotrophy and heterotrophy, has been largely underestimated and is commonly found in planktonic protists [6, 9–13]. Instead of a dichotomy between two trophic types, their trophic regime should be regarded as a continuum between full phototrophy and full heterotrophy, with species from many planktonic lineages lying between these two extremes [10]. Mitra et al. [11] have proposed a classification of marine mixotrophic protists into four functional groups, or mixotypes. The constitutive mixotrophs, or CM, are photosynthetic organisms that are capable of phagotrophy, also called "phytoplankton that eat" [11]. They include most mixotrophic nanoflagellates (e.g., *Prymnesium parvum*, *Karlodinium micrum*). On the opposite, the non-constitutive mixotrophs,

These authors contributed equally: Lucie Bittner, Sakina-Dorothée Ayata

Supplementary information The online version of this article (https://doi.org/10.1038/s41396-018-0340-5) contains supplementary material, which is available to authorized users.

✉ Emile Faure
emile.faure@etu.upmc.fr

[1] Sorbonne Université, CNRS, Laboratoire d'océanographie de Villefranche, LOV, 06230 Villefranche-sur-Mer, France

[2] Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, CP 50, 57 rue Cuvier, 75005 Paris, France

[3] Sorbonne Université, CNRS, UMR7144 Adaptation and Diversity in Marine Environment (AD2M) Laboratory, Ecology of Marine Plankton team, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[4] Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Evry, France, 91057 Evry, France

Published online: 14 January 2019

or "photosynthetic zooplankton", are heterotrophic organisms that have developed the ability to acquire energy through photosynthesis [9]. This ability can be acquired in three different ways: the generalist non-constitutive mixotrophs (GNCM) steal the chloroplasts of their prey, such as most plastid-retaining oligotrich ciliates (e.g., *Laboea strobila*), the plastidic specialist non-constitutive mixotrophs (pSNCM) steal the chloroplasts of a specific type of prey (e.g., *Mesodinium rubrum* or *Dinophysis* spp.), and finally the endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) are bearing photosynthetically active endo-symbionts (most mixotrophic Rhizaria from Collodaria, Acantharea, Polycystinea, and Foraminifera, as well as dinoflagellates like *Noctiluca scintillans*).

As drivers of biogeochemical cycles in the global ocean, and particularly of the biological carbon pump [5, 14, 15], marine protists are a key part of ocean biogeochemical models [7, 16–18]. However, physiological details of mixotrophic energy acquisition strategies have only been studied in a restricted number of lineages [9, 19, 20]. They appear to be quite complex and greatly differ across mixotypes, which makes mixotrophy hard to include in a simple model structure [21–25]. Hence at this time, mixotrophy is not included in most biogeochemical models, neglecting the amount of carbon fixed by non-constitutive mixotrophs through photosynthesis, and missing the population dynamics of photosynthetically active constitutive mixotrophs that can still grow under nutrient limitation [23, 26]. This is most probably skewing climatic models predictions [11, 26], as well as our ability to understand and prevent future effects of global change.

A better understanding of the environmental diversity of marine mixotrophic protists, as well as a description of the abiotic factors driving their biogeography at global scale are still needed, in particular to integrate them in biogeochemical models. Leles et al. [27] attempted to tackle this problem by reviewing about 110,000 morphological identification records of a set of more than 60 mixotrophic protists species in the ocean, taken from the Ocean Biogeographic Information System (OBIS) database. They found distinctive patterns in the biogeography of the three different non-constitutive mixotypes (GNCM, pSNCM, and eSNCM), highlighting the need to better understand such diverging distributions [27]. Environmental molecular biodiversity surveys through metabarcoding have been widely used in the past fifteen years to decipher planktonic taxonomic diversity [2, 28–30]. Here, we exploited the global *Tara* Oceans datasets [31–33], and identified 133 mixotrophic lineages, that we classified into the four mixotypes defined by Mitra et al. [11]. This first ever set of mixotrophic metabarcodes allowed us to investigate the global biogeography of both constitutive and non-constitutive mixotrophs, in relation with in-situ abiotic measurements. We tested (i) if new information on marine mixotrophic protists distribution can be gained in comparison with previous morphological identifications [27]; (ii) if the constitutive mixotrophs, which are not addressed in Leles et al. [27], and the non-constitutive mixotrophs diverge in terms of biogeography; (iii) if the study of diversity and abundance of environmental metabarcodes could lead to the definition of key environmental factors shaping mixotrophic communities.

Materials and methods

Samples collection and dataset creation

Metabarcoding datasets from the worldwide *Tara* Oceans sampling campaigns that took place between 2009 and 2013 [31, 33] (data published in open access at the European Nucleotide Archive under project accession number PRJEB6610) were investigated. We analyzed 659 samples from 122 distinct stations, and for each sample, the V9-18S ribosomal DNA region was sequenced through Illumina HiSeq [32]. Assembled and filtered V9 metabarcodes (cf. details in de Vargas et al. [2]) were assigned to the lowest taxonomic rank possible via the Protist Ribosomal Reference (PR2) database [34]. To limit false positives, we chose to only analyze the metabarcodes (i.e., unique versions of V9 sequences) for which the assignment to a reference sequence had been achieved with a similarity of 95% or higher. This represents 65% of the total dataset in terms of metabarcodes and 84% in terms of total sequences. Our dataset involved 1,492,912,215 sequences, distributed into 4,099,567 metabarcodes assigned to 5071 different taxonomic assignations, going from species to kingdom level precision.

Defining a set of mixotrophic organisms

Among these 5071 taxonomic assignations, we searched for mixotrophic protist lineages, taking into account the four mixotypes described by Mitra et al. [11]: constitutive mixotrophs (CM), generalist non-constitutive mixotrophs (GNCM), endo-symbiotic specialist non-constitutive mixotrophs (eSNCM), and plastidic specialist non-constitutive mixotrophs (pSNCM). We used the table S2 from Leles et al. [27], which is referencing 71 species or genera belonging to three non-constitutive mixotypes (GNCM, pSNCM, and eSNCM), as well as multiple other sources coming from the recent literature on mixotrophy [6, 9–12, 35–47], and inputs from mixotrophic protists' taxonomy specialists (cf. Acknowledgments section). Within the 5071 taxonomic assignations of variable precisions, we identified 5 GNCM, 9 pSNCM, 77 eSNCM, and 42 CM lineages (detailed list available publicly under the https://doi.org/10.

6084/m9.figshare.6715754, and all metabarcodes were tagged with their mixotypes in the PR2 database). Among these 133 taxonomic assignations that we will call "lineages", 92 were defined at the species level, 119 at the genus level, and the last 14 at higher taxonomic levels where mixotrophy is always present (mostly eSNCM groups like Collodaria). In the Chrysophyceae family, metabarcodes assigned to clades B2, E, G, H, and I were included even though we couldn't find a general proof that all species included in these clades have mixotrophic capabilities. However, if we exclude the photolithophic Synurophyceae and genera like *Paraphysomonas* and *Spumella*, which we did, a vast majority of Chrysophyceae are considered mixotrophic [10]. The final dataset included 318 054 metabarcodes assigned to the 133 mixotrophic lineages selected, as well as their sequence abundance in 659 samples (table available publicly under the https://doi.org/10.6084/m9.figshare.6715754).

Environmental dataset

We built a corresponding contextual dataset using the environmental variables available in the PANGAEA repository from the *Tara* Oceans expeditions [33, 48]. The set of 235 environmental variables was reduced to 57 due to several selection steps (Data available publicly under the https://doi.org/10.6084/m9.figshare.6715754; see the details of variable selection in section 1 of Supp. Mat.).

Distribution and diversity of mixotrophic protists

For each mixotype, the number of metabarcodes, the total sequence abundance and the mean sequence abundance by metabarcode was computed (Table 1). Also, we measured each metabarcode's station occupancy, i.e., the number of stations in which it was found, and station evenness, i.e., the homogeneity of its distribution among the stations in which it was detected (Fig. 2). Diversity of mixotrophic protists was investigated through mixotype-specific metabarcode richness per station (Table 1). As the number of samples taken per station can impact the abundance and diversity of detected metabarcodes, richness was computed only at stations for which the maximum number of eight samples were available (40 stations over 122).

Global biogeography of mixotrophic protists

Two statistical analyses were performed to investigate mixotrophic protists biogeography. One at the metabarcode level, and one at the lineage level, i.e., merging the sequence abundance of metabarcodes sharing the same taxonomical assignation. The metabarcodes abundance table was composed of 318,054 rows/metabarcodes, and 659 columns/samples, whereas the lineage abundance table was composed of 133 rows/lineages and 659 columns/samples (both datasets are available publicly under the https://doi.org/10.6084/m9.figshare.6715754). The two analyses led to very similar conclusions, but the biogeography of lineages appeared easier to visually represent and interpret than the one of metabarcodes. Hence, we only present here the results of the lineage-based analysis (See section 3 of Sup. Mat. for metabarcode-level analysis results and discussion).

Our statistical model was designed to identify lineages (or metabarcodes) with contrasted biogeographies, and relate their presence to the environmental context. We normalized the sequence counts from the lineage abundance matrix using a Hellinger transformation [49]. We used the environmental dataset and the mixotrophic lineages' abundance matrix as explanatory and response matrices, respectively, to conduct a redundancy analysis (RDA) [49]. For that, we made a species pre-selection using Escoufier's vectors [50], which allowed to keep only the 62 most significant mixotrophic lineages. This method selects lineages

Table 1 Detailed number of lineages found for each mixotype, as well as the number of metabarcodes, the corresponding total sequence counts over all stations, the mean sequence abundance by metabarcode, and mean metabarcode richness

Mixotypes	CM	eSNCM	pSNCM	GNCM
Number of lineages used in this study	42	77	9	5
Number of V9 metabarcodes	26,015	288,536	2143	1360
Total sequence abundance	3,581,751	86,098,397	208.096	63.622
Mean sequence counts per metabarcode	137.7	298.4	97.1	46.8
Mean metabarcode richness per station[a] (std dev)	2162 (1115)	18502 (9238)	67 (102)	84 (111)
Number of absences/station	0/122	0/122	5/122	3/122

The richness was computed as the number of different metabarcodes present at each station. It was calculated for each mixotype and means are indicated in the fifth column. Absences correspond to the number of stations in which no sequences were detected for the corresponding mixotype

CM constitutive mixotrophs, *GNCM* generalist non-constitutive mixotrophs, *eSNCM* endo-symbiotic specialist non-constitutive mixotrophs, *pSNCM* plastidic specialist non-constitutive mixotrophs

[a]The mean indicated here was calculated using only stations having the maximum number of samples (see main text)

according to a principal component analysis (PCA), sorting them based on their correlation to the principal axes. We then used a maximum model (Y~X) and a null model (Y~1) to conduct a two directional stepwise model selection based on the Akaike information criterion (AIC) [51]. The resulting model contained 28 environmental response variables. More details about statistical analyses are available in section 2 and 3 of the Supplementary Materials. analyses and graphs were realized with the R software version 3.4.3 [52]. All scripts are available on GitHub platform (https://github.com/upmcgenomics/MixoBioGeo).

Results

Global distribution and diversity of marine mixotrophic protists

Mixotrophic protists metabarcodes were detected in all the 659 samples with a total sequence abundance of 89,951,866, representing 12.56% of the total sequence abundance in the 659 samples studied. They represented a mean of 12.64% of the total sequence abundance per sample, with a maximum of 96.96% and a minimum of 0.01%. To avoid any potential overestimation of mixotrophic lineages presence in the following results, we marked all records of less than a hundred sequences as questionable. We found both eSNCM and CM in each of the 122 stations studied (Table 1 and Fig. 1). In only two occasions the number of sequences belonging to CM was questionable, at stations for which only one sample was sequenced. GNCM were found absent in only two stations and their presence was questionable in 39 stations (Fig. 1). pSNCM were absent at five stations (three in the Indian Ocean, and two in the Pacific Ocean) and detected with questionable presence in 54 additional stations, which were mostly located in the central Pacific and the Indian Ocean (Fig. 1). We found significant amounts of sequences corresponding to GNCM in the Central Pacific, Southern subtropical Atlantic, and Indian Ocean. The presence of GNCM in these areas has not yet been recorded through morphological identifications during field expeditions [27]. Also, we detected more than 100 sequences of pSNCM metabarcodes at 11 stations belonging to biogeographical

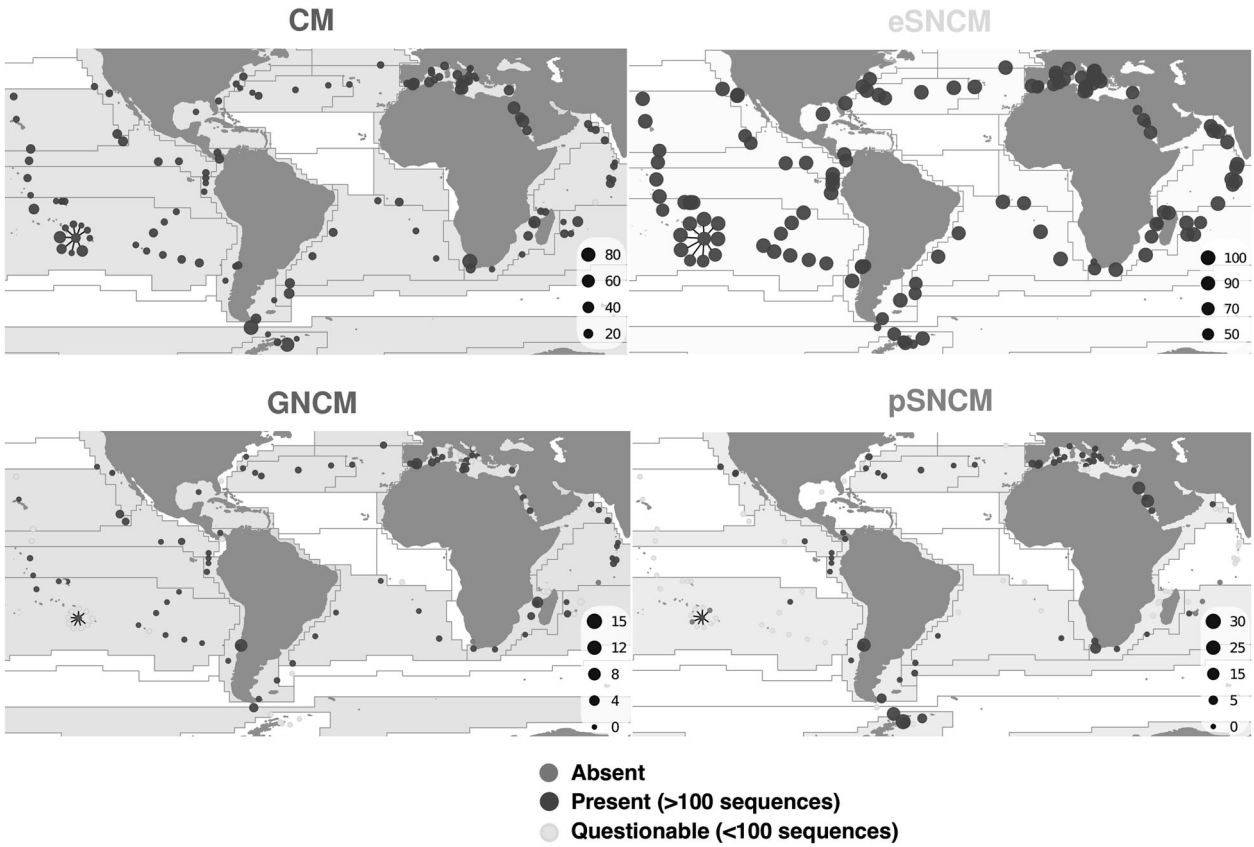


Fig. 1 Global distribution of mixotypes from metabarcoding data. Maps showing for each station the proportion of sequences (in %) belonging to each mixotype over the total number of mixotrophic sequences. Stations in which no sequence was found were marked as absent, ones with less than 100 sequences marked as questionable. Each Longhurst biogeographical provinces [53] is colored in the background if more than 100 sequences are detected in at least one of its stations. A coloured version of this figure can be downloaded at https://doi.org/10.6084/m9.figshare.6715754

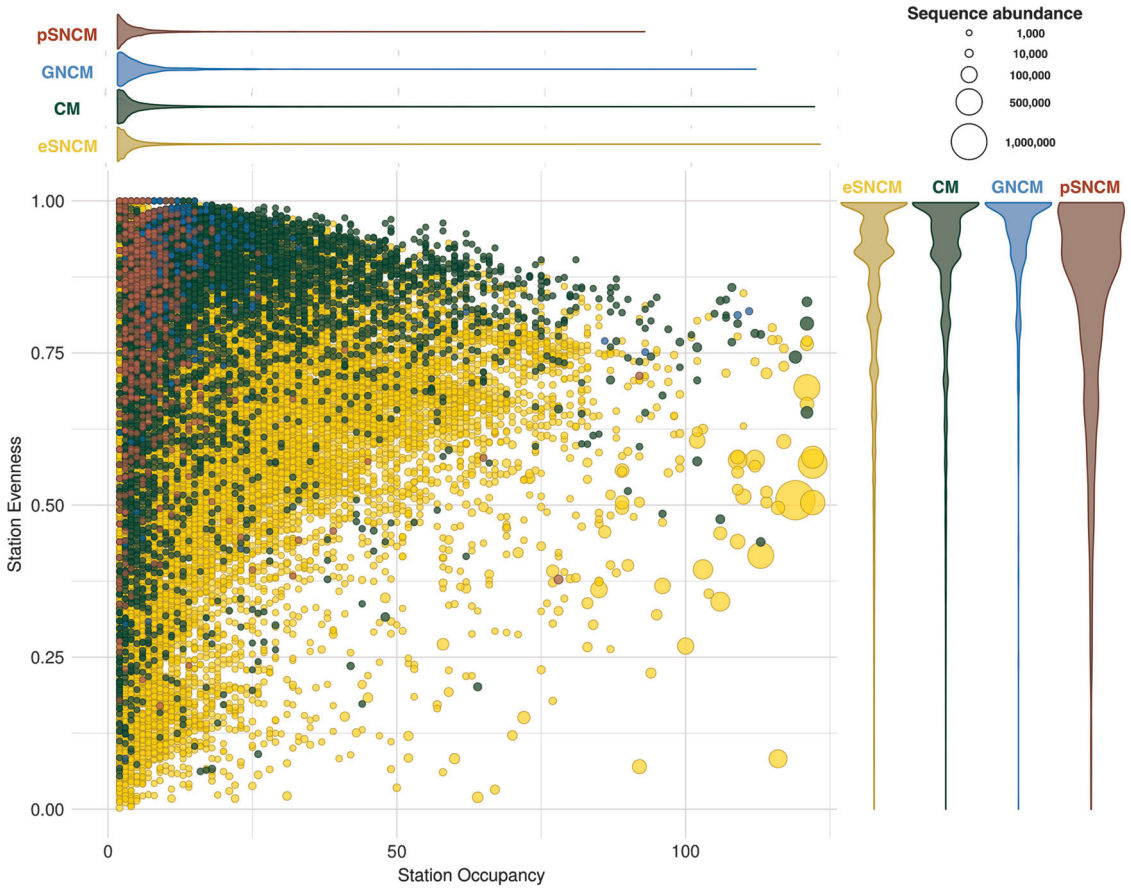


Fig. 2 Sequence abundance, occupancy, and spatial evenness of each mixotrophic metabarcode across sampled stations. Each metabarcode is plotted as a bubble, with its station occupancy, i.e., the number of stations in which it was found, and its station evenness, i.e., the homogeneity of its distribution among the stations in which it was detected, as coordinates. Violin plots were drawn for each mixotype on both the *x* and *y* axes. The size of each bubble is scaled to the sequence abundance found globally for the corresponding metabarcode

provinces in which no morphological identifications had been published [27, 53], mostly in offshore areas of the Atlantic and Pacific Ocean (Fig. 1).

The mean evenness of mixotrophic metabarcodes across stations was of 0.87, and 82.3% of the metabarcodes had a station evenness above 0.5 (Fig. 2). Station occupancy varied a lot depending on the metabarcodes, with a high density of rare metabarcodes leading to a mean of 5.14 stations over a maximum of 122, and a standard deviation of 7.7. However, three eSNCM metabarcodes were found in all the 122 stations, and three CM metabarcodes were detected in 121 stations. The maximum occupancy for a GNCM metabarcode was of 111 stations, while 92 stations was the maximum for a pSNCM metabarcode. CM and GNCM metabarcodes showed a strong tendency towards high evenness values (Fig. 2, means of 0.90 and 0.95, respectively), even for the most sequence abundant metabarcodes. Many eSNCM metabarcodes had high evenness values, but below average values were detected for the most abundant ones (Fig. 2, global mean of 0.87). pSNCM metabarcodes had a similar mean of evenness values (0.87), but a different distribution compared to other mixotypes (Fig. 2). Among the 50 most abundant metabarcodes, 43 corresponded to Collodaria lineages, 47 were eSNCM and 3 were CM, all three assigned to *Gonyaulax polygramma*. GNCM and pSNCM metabarcodes had homogeneously low sequence abundances (Fig. 2 and Table 1).

Main factors affecting the biogeography of mixotrophic protists

The redundancy analysis helped to investigate further the environmental variables responsible for the mixotrophic protists' biogeography. The 62 lineages selected with the Escoufier's vector method corresponded to 20 CM, 34 eSNCM, 3 GNCM, and 5 pSNCM. Even after selection, a significant part of the lineages did not show any response to environmental data in their distribution (Fig. 3, e.g., 19 of the 62 lineages were found between −0.01 and 0.01 on both RDA1 and RDA2). The adjusted R-squared of the RDA was of 34.89% (41.43% unadjusted), with 24.01% of

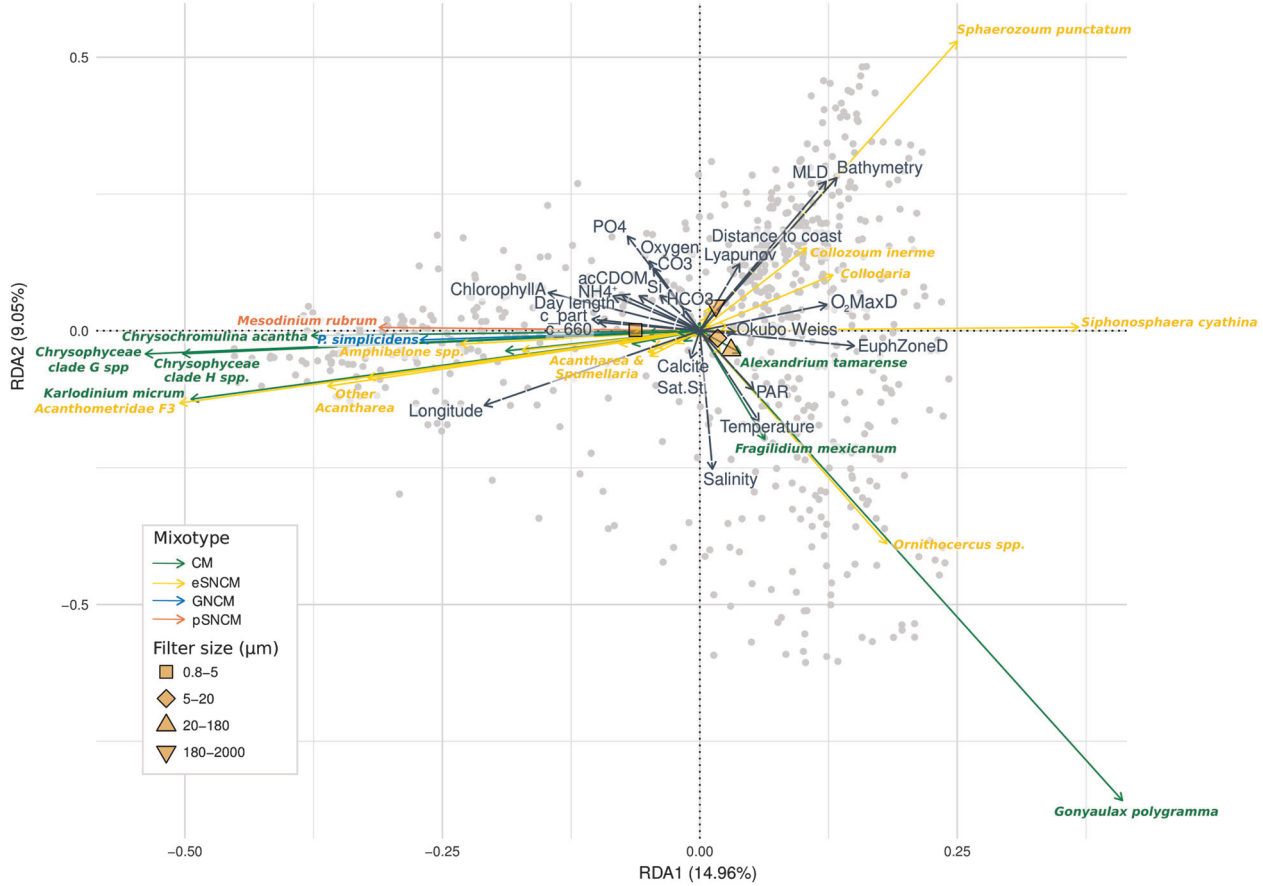


Fig. 3 Impact of environmental variables on the distribution of marine mixotrophs. Triplot of the redundancy analysis (RDA) computed on the 62 Escoufier-selected lineages, after model selection. The adjusted R-squared of the analysis is of 34.89% (41.43% unadjusted). Each gray dot corresponds to a sample, i.e., one filter at one depth at one station. The blue dashed arrows correspond to the quantitative environmental variables. Abbreviations: MLD mixed layer depth, O2MaxD O2 maximum depth, EuphzoneD euphotic zone depth, PAR photosynthetically active radiations, Calcite Sat. St. Calcite Saturation State, c_660 optical beam attenuation coefficient at 660 nm, c_part beam attenuation coefficient of particles, acCDOM absorption coefficient of colored dissolved organic matter. Plain arrows correspond to mixotrophic lineages, colors indicating mixotypes. For more readability, we do not represent all qualitative variables included in the model. That is why only the filter centroids are appearing, even though the sampling depth, season, season moment, i.e., early, middle or late, and biogeographical province were used in the RDA calculation

variance explained on the two first axes (Fig. 3). The first RDA axis (14.96%) marks an opposition between samples from oligotrophic waters with low productivity (RDA1 > 0) and samples from eutrophic and productive water masses (RDA1 < 0). This axis is negatively correlated to chlorophyll concentration, particles density, ammonium concentration, absorption coefficient of colored dissolved organic matter (acCDOM), duration of daylight, silica, CO3, oxygen, and PO4 concentration, as well as longitude. It is positively correlated to bathymetry, deep euphotic zone, deep oxygen maximum, deep mixed layer, as well as to the distance to coast. The second RDA axis (9.05%) is opposing offshore and subpolar samples (RDA2 > 0) to coastal and subtropical ones (RDA2 < 0). The axis is positively correlated to the depth of the mixed layer, the distance to coast, the bathymetry, high maximum Lyapunov exponents as well as high concentrations of PO4, oxygen, CO3 and silica. It is negatively correlated to temperature, salinity, and photosynthetically active radiations (PAR).

Among the 20 CM lineages, seven clearly emerged from the redundancy analysis (Fig. 3) and showed distinct biogeographies related to environmental variables. *Gonyaulax polygramma*, *Alexandrium tamarense*, and *Fragilidium mexicanum*, three Dinophyceae belonging to the Gonyaulacales order, were mainly found in oligotrophic waters with a deep euphotic zone, warm temperature, high salinity, and PAR (RDA1 > 0, RDA2 < 0). The four other CMs (involving all the Chrysophyceae included in the analysis as well as one Dinophyceae from the Kareniaceae family, *Karlodinium micrum*) were found mostly in productive water masses (RDA1 < 0).

eSNCMs can be divided in three groups in the RDA space. The first group (RDA1 < 0) corresponds to eSNCM species dominating rich and productive environments. It

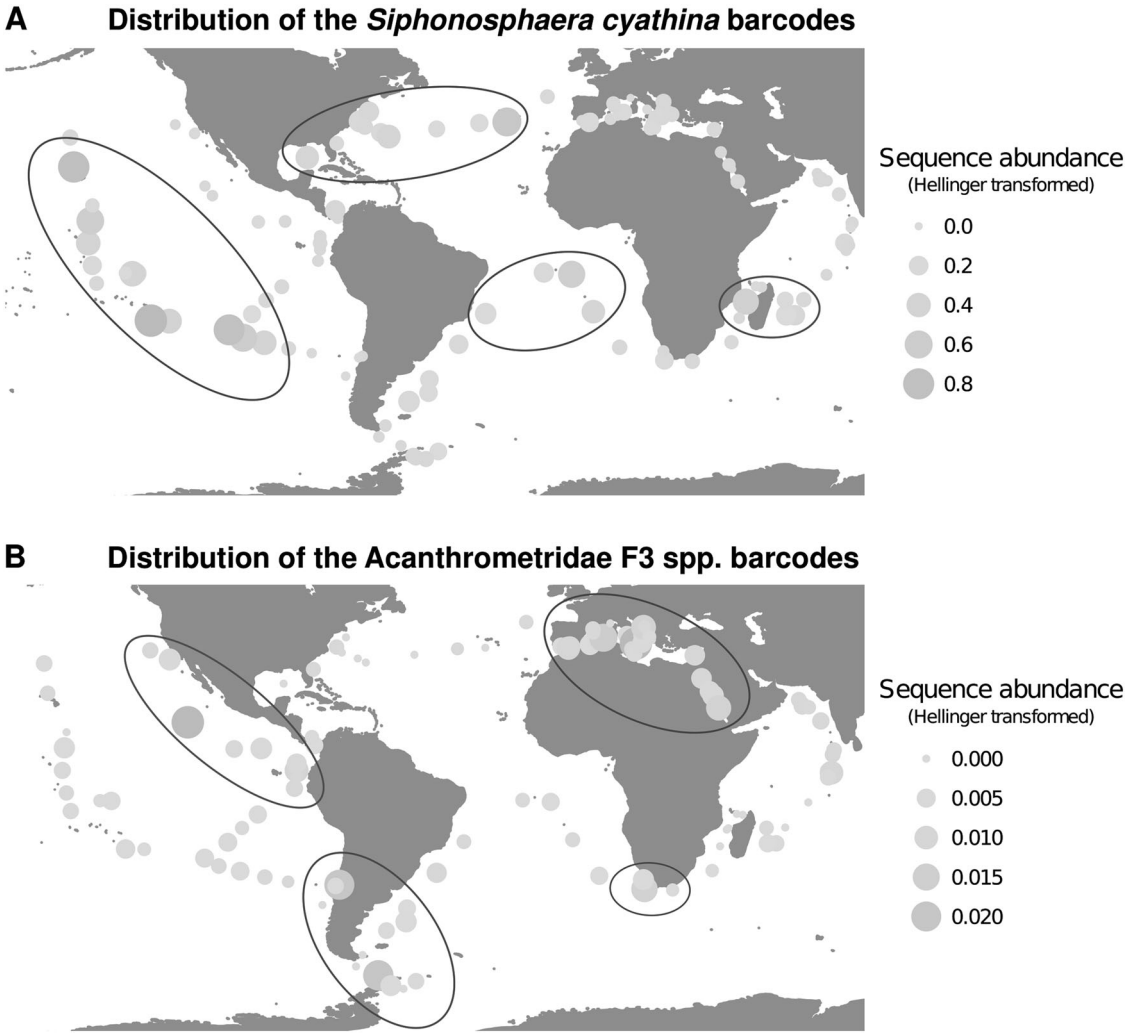


Fig. 4 Contrasted global distributions of metabarcodes corresponding to two eSNCM lineages. Maps of Hellinger-transformed sequence count abundances for metabarcodes assigned to the Collodaria *Siphonosphaera cyathina* **a** and the Acantharia Acanthrometridae F3 spp. **b** These two lineages are opposed on the first RDA axis (Fig. 3 and S1). Size and color both illustrate abundance for better readability. Ellipses were drawn to highlight high abundance zones, and reveal the differences in lineages distribution. A coloured version of this figure can be downloaded at https://doi.org/10.6084/m9.figshare.6715754

includes mainly Acantharia and Spumellaria species. The second group (RDA1 > 0) dominates oligotrophic environments, and includes multiple Collodaria as well as one Dinophyceae genus (*Ornithocercus)*. Within this group, *Ornithocercus* spp. is found mainly in coastal subtropical environments (RDA2 < 0), as opposed to *Sphaerozoum punctatum* that is found mainly in offshore subpolar regions (RDA2 > 0). *Siphonosphaera cyathina* lies between these two trends as it is found only in oligotrophic samples, but isnot influenced by temperature or bathymetry (Figs. 3 and 4). The third group corresponds to the eSNCM lineages that can be interpreted as distributed homogeneously in regards of the environmental data we are using (e.g., lineages with the shortest arrows in Fig. 3). These notably include the 12 Foraminifera lineages present in the RDA. Looking at filters centroids in the RDA space (Fig. 3), we can suppose that eSNCM lineages dominating eutrophic systems (RDA1 < 0) are smaller in size than those dominating oligotrophic ones (RDA1 > 0).

Out of the five pSNCM included in the RDA, only *Mesodinium rubrum*, the most abundant one, is distinctively represented in the RDA space. This suggests that the other pSNCM have homogeneous distributions in response to our environmental variables. *Mesodinium rubrum* dominates eutrophic environments, independently from the bathymetry or the temperature (RDA1 < 0, RDA2 ≈ 0). We find a similar pattern for GNCM, with only *Pseudotontonia simplicidens* well represented in the RDA space out of the three species included in the analysis. Like *M. rubrum*, *Pseudotontonia simplicidens* is the most abundant species in its group and it is mainly found in eutrophic waters (RDA1 < 0).

Discussion

Mixotrophy occurs everywhere in the global ocean

Our metabarcoding survey confirms that marine mixotrophic protists are ubiquitous in the global ocean [27], possibly extending the known range of distribution of two mixotypes (Figs. 1 and 2). Mixotrophic organisms represented more than 12% of the sequences in the complete *Tara* Oceans metabarcoding dataset, showing that they should not be understated. We found contrasted biogeographies among metabarcodes and their corresponding lineages, both within and across mixotypes (Figs. 2–4 and S1, Sup. Mat. section 3). We found constitutive mixotrophs (CM) and endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) metabarcodes at all the 122 stations included in this global study (Table 1 and Fig. 2), verifying that these two mixotypes are the most abundant in the ocean [27, 47, 54, 55]. This dominance of eSNCM and CM in our data is also linked to the relatively high number of metabarcodes available for these two mixotypes in databases. Using 1360 generalist non-constitutive mixotrophs (GNCM) metabarcodes corresponding to only five lineages, we detected them in ten biogeographical provinces [53] where no morphological identification had been recorded before [27]. GNCM metabarcodes had consistently high evenness values, and some had station occupancy records comparable to the most abundant eSNCM and CM metabarcodes (Fig. 2). These results support the hypothesis of a globally ubiquitous distribution of GNCM. Plastidic specialist non-constitutive mixotrophs (pSNCM) were found in five provinces in which no record existed so far from morphological identification field studies [27]. However, these observations were often in a questionable range in terms of sequence abundance (Fig. 1), and the overall distribution of pSNCM in our data appears as very concordant with morphological observations [27]. pSNCM metabarcodes had dominantly low station evenness values, which again supports the conclusions of Leles et al. [27] that identified pSNCM as highly seasonal and spatially restricted in their distribution.

While building our set of mixotrophic lineages, some widespread and potentially mixotrophic genera did not appear, such as *Ceratium* spp., *Tontonia* spp., *Amphisolenia* spp., *Triposolenia* spp., or *Citharistes* spp., mainly because of a poor representation in the PR2 database. Also, we decided to only consider metabarcodes with more than 95% similarity to a reference sequence. This threshold could be too selective for some species and not enough for some others, as single similarity threshold are hardly efficient when studying whole eukaryotic populations [56, 57]. For example, some species appeared with low sequence abundance in the data even though they couldnot have been sampled, such as three lacustrine species, e.g., *Poteriospumella lacustris*. Considering these biases and the sometimes relatively low sequence counts (marked as questionable in Fig. 1), some of the new GNCM and pSNCM records we observed should be considered with care, as they could be over-estimated or even sometimes artefactual. However, the low number of lineages found for these two mixotypes in PR2 and in our dataset are leading us to think that we were unable to capture the whole GNCM and pSNCM communities. This supposes a global underestimation of GNCM and pSNCM abundances in our results.

Tara Oceans metabarcoding dataset is built on snapshot samples taken irregularly during a 3-year cruise, hence allowing no proper seasonal variations investigations. However, morphological identifications of mixotrophic protists revealed seasonal variations in their abundance, with *Mesodinium* biomass blooming in spring in coastal seas for example [27]. As metabarcoding datasets have been successfully applied on time series to detect species successions across gradients of time and space [58–60], it would be interesting to similarly investigate seasonal trends in mixotrophic communities. Our set of mixotrophic lineages and metabarcodes being publicly available, our method will be applicable to any other metabarcoding dataset, including time series. It will also be open to inputs and updates from the global scientific community.

The contrasted biogeographies of marine mixotypes

Constitutive mixotrophs

Constitutive mixotrophs (CM) have very diverse feeding behaviors, with some species requiring phototrophy to grow, others phagotrophy, and some being obligate mixotrophs [9]. They were described in all waters of the global ocean [61–65]. We found them distributed in a range of conditions almost as wide as non-constitutive mixotrophs (Figs. 1 and 3). Among highly abundant lineages, most were dominantly found in eutrophic and shallow habitats. However, a few dinoflagellates were found to be highly dominant in oligotrophic, subtropical waters, showing how wide of a range of conditions constitutive mixotrophs can grow in (Fig. 3). This illustrates how mixotrophy can allow organisms to dominate ecosystems even when environmental conditions are poorly adapted to purely phototrophic or heterotrophic organisms. When taken explicitly into account in biogeochemical models, marine mixotrophs increase carbon export by up to 30% [22]. Hence, their global ubiquity supposes that the carbon export of the biological carbon pump could be underestimated in both oligotrophic and eutrophic areas [26].

Plastidic specialist and generalist non-constitutive mixotrophs (pSNCM and GNCM)

Like Leles et al. [27], we found pSNCM and GNCM to have quite similar biogeographies (Fig. 3, section 3 of Sup. Mat.). Sequence abundance of most of the metabarcodes for these two mixotypes was homogeneously low (Table 1), but the two most abundant species, *Mesodinium rubrum* (pSNCM) and *Pseudotontonia simplicidens* (GNCM), were found mostly in coastal and eutrophic waters, consistently with Leles et al.'s [27] morphological observations (Fig. 3, section 3 of Sup. Mat.). No species-level barcode is available in the PR2 database for the *Tontonia* genus, and only one can be found for *Pseudotontonia* and *Laboea* genera, even though morphological records of these GNCM are numerous [27]. Experiments using meso- and microcosms combined with individual counts and morphological identification have found that GNCM ciliates can represent up to half of the individuals in ciliate communities of the photic zone [11, 66, 67]. A proportion we would have trouble to reach with the five lineages we were able to consider, knowing that there are 8686 different ciliate lineages available in PR2. This highlights the urgent need for supplementing 18S reference databases for mixotrophic ciliates.

Endo-symbiotic specialist non-constitutive mixotrophs (eSNCM)

Endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) is by far the most widespread and abundant non-constitutive mixotype in the global ocean (Figs. 1 and 2) [27, 47, 54]. Their biogeography stands out, with a lot of highly abundant ubiquitous lineages, and some other specialized towards certain types of ecosystems (Fig. 3). They represent 95.7% of the sequence counts in our study and correspond to 90.7% of the metabarcodes (Table 1), which highlights their abundance and diversity. The very high number of rDNA copies present in Rhizaria orders such as Collodaria [47] might lead the eSNCM to appear more abundant in metabarcoding datasets than they ecologically are. However, in oligotrophic open oceans the Rhizaria biomass is estimated to be equivalent to that of all other mesozooplankton [68], and positively correlated to the carbon export [15], showing how ecologically important they can be.

Investigating the divergent biogeographies of Collodaria and Acantharia

Collodaria are living either as solitary large cells or as colonies [47], which explains why they are predominantly found in macro-sized (180–2000 µm) filter samples (Fig. 3). All described Collodaria species so far harbor photosynthetic endo-symbionts, mostly identified as the dinoflagellate species *Brandtodinium nutricula* [47, 69]. These dinoflagellates are able to get in and out of their symbiotic state, which implies a light and/or reversible effect of the Collodarian host on its symbiont metabolism [69]. Based on the same metabarcoding dataset, Collodaria were described as particularly abundant and diverse in the oligotrophic open ocean [47]. In our results, Collodaria dominate oligotrophic, relatively deep waters (Figs. 3 and 4a). These Collodaria appear opposed to another set of Rhizaria (Acantharia and Spumellaria) linked to eutrophic and shallow waters (Figs. 3 and 4b, section 3 of Sup. Mat.). Acantharia are found ubiquitously in the global ocean, but display particularly high sequence abundances in some specific regions [54]. Mixotrophic Acantharia live in symbiosis with the cosmopolitan haptophyte *Phaeocystis*, which is highly abundant and ecologically active in its free-living phase [54]. Unlike the one of Collodaria, this symbiosis is irreversible: an algal symbiont can not go back to its free-living phase [54]. Our results suppose that these specific symbiotic modes could enable Acantharia and Collodaria to dominate different ecosystems (Figs. 3 and 4). Moreover, living in colonies as Collodaria could help to dominate oligotrophic systems, e.g., by accumulating more food and nutrients through their gelatinous extra-cellular matrix [47]. Experiments and modeling studies should help to investigate the contribution of this assumption, comparing food acquisition capacity and growth rates of free individuals versus in colony.

Towards an integration of mixotrophic diversity into marine ecosystem models

The future of marine communities' modeling lies in the integration of omics datasets into modeling frameworks [18, 70–73]. The use of metabolic networks and gene-centric methods has already shown very promising results in modeling prokaryotic ecological dynamics [18, 73]. However, eukaryotic metabolic complexity makes these methods hard to apply on protists for now, and we still lack a universal theoretical framework on how to integrate such methods into concrete modeling [70]. Mixotrophic protists are physiologically complex, and their feeding behavior can vary drastically on short time scales [9]. It will then take a few more years of comparative genomics and transcriptomics studies before being able to model their physiology with purely gene-based approaches. Still, mechanistic models of mixotrophy exist and are quite complex [21, 23], even if the one from Ghyoot et al. [23] could be implemented in a global biogeochemical model [74]. Most models make the choice to represent either one or two (NCM and CM) types of organisms able to play the

role of all mixotypes depending on parameterization. However, no global agreement has been reached on to what extent the different mixotypes should be modeled. This is mainly due to a lack of quantitative and comparative data on the global impact of grazing and carbon fixation by the different mixotypes [75]. With our study, we show how meta-omics data can be used to identify groups of organisms distributed differently in response to the environment. It also allows the identification of ecological traits and environmental factors potentially responsible for these divergences. This information can be used to identify key species or lineages, and design controlled experiments with variations of targeted environmental factors to produce the quantitative data needed by modelers. Considering our results, we propose that host-symbiont dynamics of eSNCM should be investigated as a trait playing a potential role on Rhizaria ability to thrive in oligotrophic conditions. Particularly, the mechanisms behind holobiont formation and its potential reversibility could play major roles on eSNCM carbon fixation in various nutrient conditions. Future experiments comparing responses of Collodaria and Acantharia holobionts to different stresses in terms of grazing and carbon fixation could lead to a better understanding of the physiological differences between their two modes of symbiosis. Also, our results suggest that the metabolic flexibility of CM should allow this mixotype to grow in almost any conditions, with individual species probably spanning continuously between complete autotrophy and complete heterotrophy. The risk is then to create a "perfect beast" mixotroph dominating all systems [21]. To avoid that, we need more comparative data on grazing and carbon fixation of obligate phototrophs versus obligate heterotrophs in response to nutrient depletion and environmental fluctuation. Here again, meta-omics data could help to identify candidates for efficient experiment designs. Finally, the small number of lineages of GNCM and pSNCM in our study makes it hard to come up with strongly supported conclusions on whether they should be differentiated in models or not. They seem to share similar biogeographies using snapshot data (Fig. 3, section 3 of Sup. Mat.), but considering that they have different abilities for conserving stolen chloroplasts over time, it might not be the case when looking at a time series analysis [20, 76, 77].

Our study uses meta-omics data to investigate the global distribution and biogeography of mixotrophic protists in the ocean. Our results, currently based on metabarcoding data, complement morphological records and will be complemented in the near future by metagenomics and metatranscriptomics studies. The latter will allow to distinguish the protists with mixotrophic capabilities from the ones with ongoing mixotrophic activity. This could lead to quantitative estimations of mixotrophic rates in environmental samples, allowing a sharpened study of mixotrophic protists ecology in the global ocean. It could also lead to a metabolic description of complex processes like kleptoplasty and endo-symbiosis, hence facilitating the modeling of mixotrophic behaviors and its incorporation in ocean biogeochemical models.

Acknowledgements We would like to particularly thank Stéphane Pesant and Stéphane Audic for their work on making *Tara* Oceans datasets available. We also thank John Dolan (CNRS, LOV, Villefranche-sur-mer, France), Miguel Mendez-Sandin (Sorbonne Université, Station Biologique de Roscoff, France), and Wei-Ting Chen (National Taiwan Ocean University, Taiwan) for their essential help during the construction of the mixotrophic lineages set. We also thank Florentin Constancias for his help on the metabarcodes clustering tests conducted. Finally, we thank the three anonymous reviewers for their very constructive comments. This article is contribution number #84 of *Tara* Oceans. For the *Tara* Oceans expedition, we thank the commitment of the CNRS (in particular, Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research—Flanders, Rega Institute, KU Leuven, The French Ministry of Research. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Région Bretagne, Lorient Agglomération, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org).

Funding This work was funded by the FunOmics project of the French national program EC2CO-LEFE of CNRS and by the ModelOmics project of the Émergence program of Sorbonne Université, and partly supported byt the project MEGALADOM, part of the MASTODON program from the MITI, CNRS France. Emile Faure acknowledges a 3-year Ph.D. grant from the "Interface Pour le Vivant" (IPV) doctoral program of Sorbonne Université. SD Ayata ackowledges the CNRS for her sabbatical year as visiting researcher at ISYEB.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A. Marine protistan diversity. Annu Rev Mar Sci. 2012;4:467–93.
2. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348:1261605–605.
3. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, et al. CBOL Protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLOS Biol. 2012;10:e1001419.
4. Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nat Rev Microbiol. 2017;15:6–20.

5. Keeling PJ, Campo J del. Marine protists are not just big bacteria. Curr Biol. 2017;27:R541–49.
6. Caron DA. Mixotrophy stirs up our understanding of marine food webs. Proc Natl Acad Sci. 2016;113:2806–08.
7. Le Quéré C, Harrison SP, Colin Prentice I, Buitenhuis ET, Aumont O, Bopp L, et al. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Glob Change Biol. 2005;11:2016–40.
8. Amacher J, Neuer S, Anderson I, Massana R. Molecular approach to determine contributions of the protist community to particle flux. Deep Sea Res Part Oceanogr Res Pap. 2009;56:2206–15.
9. Stoecker DK, Hansen PJ, Caron DA, Mitra A. Mixotrophy in the marine plankton. Annu Rev Mar Sci. 2017;9:311–5.
10. Flynn KJ, Stoecker DK, Mitra A, Raven JA, Glibert PM, Hansen PJ, et al. Misuse of the phytoplankton-zooplankton dichotomy: the need to assign organisms as mixotrophs within plankton functional types. J Plankton Res. 2013;35:3–11.
11. Mitra A, Flynn KJ, Tillmann U, Raven JA, Caron D, Stoecker DK, et al. Defining planktonic protist functional groups on mechanisms for energy and nutrient acquisition: Incorporation of diverse mixotrophic strategies. Protist. 2016;167:106–120.
12. Esteban GF, Fenchel T, Finlay BJ. Mixotrophy in ciliates. Protist. 2010;161:621–41.
13. Selosse M-A, Charpin M, Not F, Jeyasingh P. Mixotrophy everywhere on land and in water: the grand écart hypothesis. Ecol Lett. 2017;20:246–63.
14. Ducklow HW, Steinberg DK, Buesseler KO. Upper ocean carbon export and the biological pump. Oceanogr-Wash DC-Oceanogr Soc. 2001;14:50–8.
15. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks driving carbon export in the oligotrophic ocean. Nature. 2016;532:465–70.
16. Aumont O, Ethé C, Tagliabue A, Bopp L, Gehlen M. PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. Geosci Model Dev. 2015;8:2465–513.
17. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. Emergent biogeography of microbial communities in a model. Ocean Sci. 2007;315:1843–46.
18. Reed DC, Algar CK, Huber JA, Dick GJ. Gene-centric approach to integrating environmental genomics and biogeochemical models. Proc Natl Acad Sci. 2014;111:1879–84.
19. Johnson MD. Acquired phototrophy in ciliates: A review of cellular interactions and structural adaptations. J Eukaryot Microbiol. 2011;58:185–195.
20. Stoecker DK, Johnson MD, de Vargas C, Not F. Acquired phototrophy in aquatic protists. Aquat Microb Ecol. 2009;57:279–310.
21. Flynn KJ, Mitra A. Building the 'perfect beast': modelling mixotrophic plankton. J Plankton Res. 2009;31:965–92.
22. Ward BA, Follows MJ. Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. Proc Natl Acad Sci. 2016;113:2958–63.
23. Ghyoot C, Flynn KJ, Mitra A, Lancelot C, Gypens N. Modeling plankton mixotrophy: A mechanistic model consistent with the shuter-type biochemical approach. Front Ecol Evol. 2017;5:78.
24. Ward BA, Dutkiewicz S, Barton AD, Follows MJ. Biophysical aspects of resource acquisition and competition in algal mixotrophs. Am Nat. 2011;178:98–112.
25. Berge T, Chakraborty S, Hansen PJ, Andersen KH. Modeling succession of key resource-harvesting traits of mixotrophic plankton. ISME J. 2017;11:212–23.
26. Mitra A, Flynn KJ, Burkholder JM, Berge T, Calbet A, Raven JA, et al. The role of mixotrophic protists in the biological carbon pump. Biogeosciences. 2014;11:995–1005.
27. Leles SG, Mitra A, Flynn KJ, Stoecker DK, Hansen PJ, Calbet A, et al. Oceanic protists with different forms of acquired phototrophy display contrasting biogeographies and abundance. Proc R Soc B Biol Sci. 2017;284:20170664.
28. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, et al. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol. 2010;19:21–31.
29. Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol Evol. 2012;27:233–43.
30. Bittner L, Gobet A, Audic S, Romac S, Egge ES, Santini S, et al. Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. Mol Ecol. 2013;22:87–101.
31. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, et al. A holistic approach to marine eco-systems biology. PLoS Biol. 2011;9:e1001177.
32. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan marine plankton nucleotide sequences from the T*ara* Oceans expedition. Sci Data. 2017;4:170093.
33. Pesant S, Not F, Picheral M, Kandels-Lewis S, Bescot NL, Gorsky G, et al. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data. 2015;2:150023.
34. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucl Acids Res. 2013;41:D597–604.
35. Granéli E, Edvardsen B, Roelke DL, Hagström JA. The ecophysiology and bloom dynamics of Prymnesium spp. Harmful Algae. 2012;14:260–70.
36. Liu H, Aris-Brosou S, Probert I, de Vargas C. A time line of the environmental genetics of the haptophytes. Mol Biol Evol. 2010;27:161–176.
37. Hansen P, Moldrup M, Tarangkoon W, Garcia-Cuetos L, Moestrup ø. Direct evidence for symbiont sequestration in the marine red tide ciliate Mesodinium rubrum. Aquat Microb Ecol. 2012;66:63–75.
38. Agatha S, Strüder-Kypke MC, Beran A, Lynn DH. Pelagostrobilidium neptuni (Montagnes and Taylor, 1994) and Strombidium biarmatum nov. spec. (Ciliophora, Oligotrichea): phylogenetic position inferred from morphology, ontogenesis, and gene sequence data. Eur J Protistol. 2005;41:65–83.
39. Jones HLJ, Leadbeater BSC, Green JC. Mixotrophy in marine species of Chrysochromulina (Prymnesiophyceae): ingestion and digestion of a small green flagellate. J Mar Biol Assoc U K. 1993;73:283.
40. Johnsen G, Dalløkken R, Eikrem W, Legrand C, Aure J, Skjoldal HR. Eco-physiology bio-optics and toxicity of the ichtyotoxic Chrysochromulina leadbeateri (Prymnesiophyceae). J Phycol. 1999;35:1465–76.
41. Rhodes L, Burke B. Morphology and growth characteristics of C*hrysochromulina* species (Haptophyceae=Prymnesiophyceae) isolated from New Zealand coastal waters. N Z J Mar Freshw Res. 1996;30:91–103.
42. Hemleben C, Be AWH, Anderson OR, Tuntivate S. Test morphology, organic layers and chamber formation of the planktonic foraminifer Globorotalia menardii (d'Orbigny). J Foraminifer Res. 1977;7:1–25.
43. Fehrenbacher JS, Spero HJ, Russell AD. Observations of living non-spinose planktic foraminifers Neogloboquadrina dutertrei and N. pachyderma from specimens grown in culture. AGU Fall Meet Abstr. 2011;41:PP41A-1724.
44. Spero HJ, Parker SL. Photosynthesis in the symbiotic planktonic foraminifer Orbulina universa, and its potential contribution to oceanic primary productivity. J Foraminifer Res. 1985;15:273–81.
45. Faber WW, Anderson OR, Caron DA. Algal-foraminiferal symbiosis in the planktonic foraminifer Globigerinella aequilateralis;

II, Effects of two symbiont species on foraminiferal growth and longevity. J Foraminifer Res. 1989;19:185–93.
46. Kuile Bter, Erez J. In situ growth rate experiments on the symbiont-bearing foraminifera Amphistegina lobifera and Amphisorus hemprichii. J Foraminifer Res. 1984;14:262–76.
47. Biard T, Bigeard E, Audic S, Poulain J, Gutierrez-Rodriguez A, Pesant S, et al. Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. ISME J. 2017;11:1331–44.
48. Ardyna M, Ovidio F, Speich S, Leconte J, Chaffron S, Audic S, et al. Environmental context of all samples from the Tara OceansExpedition (2009–2013), about mesoscale features at the sampling location. 2017. PANGAEA.
49. Legendre P, Legendre LFJ. Numerical ecology. Elsevier Science, Amsterdam; 1998;197:333.
50. Escoufier Y. Le traitement des variables vectorielles. Biometrics. 1973;29:751.
51. Borcard D, Gillet F, Legendre P. Numerical ecology with R. Springer, New York; 2011;176:177.
52. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
53. Longhurst AR. Ecological geography of the sea. Academic Press, San Diego;1998.
54. Decelle J, Probert I, Bittner L, Desdevises Y, Colin S, de Vargas C, et al. An original mode of symbiosis in open ocean plankton. Proc Natl Acad Sci. 2012;109:18000–5.
55. Le Bescot N, Mahé F, Audic S, Dimier C, Garet M-J, Poulain J, et al. Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. Environ Microbiol. 2016;18:609–26.
56. Wu S, Xiong J, Yu Y. Taxonomic resolutions based on 18S rRNA genes: A case study of subclass Copepoda. PLoS ONE. 2015;10: e0131498.
57. Brown EA, Chain FJJ, Crease TJ, MacIsaac HJ, Cristescu ME. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecol Evol. 2015;5:2234–51.
58. Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B. 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. PloS ONE. 2013;8:e74371.
59. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, et al. The taxonomic and functional diversity of microbes at a temperate coastal site: A 'multi-omic' study of seasonal and diel temporal variation. PLoS ONE. 2010;5:e15545.
60. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, et al. Community genomics among stratified microbial assemblages in the ocean's interior. Science. 2006;311:496–503.
61. Arenovski AL, Lim EL, Caron DA. Mixotrophic nanoplankton in oligotrophic surface waters of the Sargasso Sea may employ phagotrophy to obtain major nutrients. J Plankton Res. 1995;17:801–20.
62. Safi KA, Hall JA. Mixotrophic and heterotrophic nanoflagellate grazing in the convergence zone east of New Zealand. Aquat Microb Ecol. 1999;20:83–93.
63. Moorthi S, Caron DA, Gast RJ, Sanders RW. Mixotrophy: a widespread and important ecological strategy for planktonic and sea-ice nanoflagellates in the Ross Sea, Antarctica. Aquat Microb Ecol. 2009;54:269–77.
64. Unrein F, Gasol JM, Massana R. Dinobryon faculiferum (Chrysophyta) in coastal Mediterranean seawater: presence and grazing impact on bacteria. J Plankton Res. 2010;32:559–64.
65. Sanders RW, Gast RJ. Bacterivory by phototrophic picoplankton and nanoplankton in Arctic waters. FEMS Microbiol Ecol. 2012;82:242–53.
66. Calbet A, Martínez RA, Isari S, Zervoudaki S, Nejstgaard JC, Pitta P, et al. Effects of light availability on mixotrophy and microzooplankton grazing in an oligotrophic plankton food web: Evidences from a mesocosm study in Eastern Mediterranean waters. J Exp Mar Biol Ecol. 2012;424–425:66–77.
67. Dolan JR, PÉrez MT. Costs benefits and characteristics of mixotrophy in marine oligotrichs. Freshw Biol. 2000;45:227–38.
68. Biard T, Stemmann L, Picheral M, Mayot N, Vandromme P, Hauss H, et al. In situ imaging reveals the biomass of giant protists in the global ocean. Nature. 2016;532:504–7.
69. Probert I, Siano R, Poirier C, Decelle J, Biard T, Tuji A, et al. Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. J Phycol. 2014;50:388–99.
70. Stec KF, Caputi L, Buttigieg PL, D'Alelio D, Ibarbalz FM, Sullivan MB, et al. Modelling plankton ecosystems in the meta-omics era. Are we ready? Mar Genom. 2017;32:1–17.
71. Dick GJ. Embracing the mantra of modellers and synthesizing omics, experiments and models. Environ Microbiol Rep. 2017;9:18–20.
72. Mock T, Daines SJ, Geider R, Collins S, Metodiev M, Millar AJ, et al. Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. Glob Change Biol. 2016;22:61–75.
73. Coles VJ, Stukel MR, Brooks MT, Burd A, Crump BC, Moran MA, et al. Ocean biogeochemistry modeled with emergent trait-based genomics. Science. 2017;358:1149–54.
74. Shuter B. A model of physiological adaptation in unicellular algae. J Theor Biol. 1979;78:519–52.
75. Millette NC, Grosse J, Johnson WM, Jungbluth MJ, Suter EA. Hidden in plain sight: The importance of cryptic interactions in marine plankton. Limnol Oceanogr Lett. 2018;3:341–56.
76. Johnson MD, Oldach D, Delwiche CF, Stoecker DK. Retention of transcriptionally active cryptophyte nuclei by the ciliate Myrionecta rubra. Nature. 2007;445:426–8.
77. Schoener DM, McManus GB. Plastid retention, use, and replacement in a kleptoplastidic ciliate. Aquat Microb Ecol. 2012;67:177–87.

Appendix B

Article 5 / Co-authored manuscript 2: Caputi et al. 2019

This article presents a network analysis (with the same methodology used by Guidi et al. (2016), i.e. Weighted Gene Correlation Network Analysis) linking metagenomic and metatranscriptomic data to biogeochemical variables in the *Tara* Oceans sampling stations in order to assess how iron availability shapes plankton communities and how they respond to it. Among other results, Caputi et al. (2019) showed that iron availability is negatively correlated with the expression of marker genes for iron limitation and that plankton response to iron availability is coordinated at subcommunity level. As iron is a limiting micronutrient for primary production, especially for diatoms, knowledge derived from this analysis would improve our understanding of the response of planktonic communities to iron availability variations and the consequences on primary production.

For this article, I was in charge of the generation of eukaryotic metabarcode abundances for different size fractions (i.e. 0.8-5 μm, 5-20 μm, 20-180 μm and 180-2000 μm) from a global abundance matrix gathering all eukaryotic size fractions.

Global Biogeochemical Cycles

RESEARCH ARTICLE
10.1029/2018GB006022

Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems

Luigi Caputi, Quentin Carradec, Damien Eveillard, Amos Kirilovsky, Eric Pelletier, Juan J. Pierella, Karlusich, Fabio Rocha Jimenez Vieira, and Emilie Villar contributed equally to this work.
Chris Bowler and Daniele Iudicone equal coordinating contribution.

Key Points:
- Coherent assemblages of taxa covarying with iron at global level are identified in plankton communities
- Functional responses to iron availability involve both changes in copy numbers of iron-responsive genes and their transcriptional regulation
- Plankton responses to local variations in iron concentrations recapitulate global patterns

Supporting Information:
- Supporting Information S1
- Figure S1
- Figure S2
- Figure S3
- Figure S4
- Figure S5
- Figure S6
- Figure S7
- Table S1
- Table S2
- Table S3

Correspondence to:
M. R. d'Alcalà, P. Wincker, C. Bowler, and D. Iudicone,
pwincker@genoscope.cns.fr;
cbowler@biologie.ens.fr;
iudicone@szn.it;
maurizio@szn.it

Citation:
Caputi, L., Carradec, Q., Eveillard, D., Kirilovsky, A., Pelletier, E., Pierella Karlusich, J. J., et al. (2019). Community-level responses to iron availability in open ocean plankton ecosystems. *Global Biogeochemical Cycles*, *33*. https://doi.org/10.1029/2018GB006022

Received 4 JUL 2018
Accepted 17 DEC 2018
Accepted article online 11 JAN 201

©2019. American Geophysical Union. All Rights Reserved.

Luigi Caputi[1], Quentin Carradec[2,3,4,5], Damien Eveillard[5,6], Amos Kirilovsky[7,8], Eric Pelletier[2,3,4,5], Juan J. Pierella Karlusich[5,7], Fabio Rocha Jimenez Vieira[5,7], Emilie Villar[7,9], Samuel Chaffron[5,6], Shruti Malviya[7,10], Eleonora Scalco[1], Silvia G. Acinas[11], Adriana Alberti[2,5], Jean-Marc Aury[2], Anne-Sophie Benoiston[7,12], Alexis Bertrand[2], Tristan Biard[9], Lucie Bittner[7,9,12], Martine Boccara[7], Jennifer R. Brum[13,14], Christophe Brunet[1], Greta Busseni[1], Anna Carratalà[15], Hervé Claustre[16], Luis Pedro Coelho[17], Sébastien Colin[5,7,9], Salvatore D'Aniello[1], Corinne Da Silva[3,5], Marianna Del Core[18], Hugo Doré[9], Stéphane Gasparini[16], Florian Kokoszka[1,7,19], Jean-Louis Jamet[20], Christophe Lejeusne[1,21], Cyrille Lepoivre[22], Magali Lescot[5,23], Gipsi Lima-Mendez[24,25], Fabien Lombard[5,16], Julius Lukeš[26,27], Nicolas Maillet[1,28], Mohammed-Amin Madoui[2,3,4], Elodie Martinez[29], Maria Grazia Mazzocchi[1], Mario B. Néou[2,3,4], Javier Paz-Yepes[7], Julie Poulain[2,5], Simon Ramondenc[16], Jean-Baptiste Romagnan[30], Simon Roux[14], Daniela Salvagio Manta[18], Remo Sanges[1], Sabrina Speich[5,19], Mario Sprovieri[18], Shinichi Sunagawa[17,31], Vincent Taillandier[16], Atsuko Tanaka[7], Leila Tirichine[7,32], Camille Trottier[6], Julia Uitz[16], Alaguraj Veluchamy[7,33], Jana Veselá[26], Flora Vincent[7], Sheree Yau[34], Stefanie Kandels-Lewis[17,35], Sarah Searson[16], Céline Dimier[7,9], Marc Picheral[5,16], *Tara* Oceans Coordinators, Peer Bork[17,35,36,37], Emmanuel Boss[38], Colomban de Vargas[5,9,39], Michael J. Follows[40], Nigel Grimsley[5,34], Lionel Guidi[5,16,41], Pascal Hingamp[5,23], Eric Karsenti[5,7,35], Paolo Sordino[1], Lars Stemmann[5,16], Matthew B. Sullivan[14], Alessandro Tagliabue[42], Adriana Zingone[1], Laurence Garczarek[9], Fabrizio d'Ortenzio[16], Pierre Testor[43], Fabrice Not[9], Maurizio Ribera d'Alcalà[1], Patrick Wincker[2,3,4,5], Chris Bowler[5,7], Daniele Iudicone[1]

***Tara* Oceans Coordinators: Silvia G. Acinas[11], Peer Bork[17,35,36,37], Emmanuel Boss[38], Chris Bowler[5,7], Colomban de Vargas[5,9,39], Michael J. Follows[40], Gabriel Gorsky[16], Nigel Grimsley[5,34], Pascal Hingamp[5,23], Daniele Iudicone[1], Olivier Jaillon[2,3], Stefanie Kandels-Lewis[17,35], Lee Karp-Boss[38], Eric Karsenti[5,7,35], Uros Krzic[44], Fabrice Not[9], Hiroyuki Ogata[45], Stéphane Pesant[46,47], Jeroen Raes[24], Emmanuel G. Reynaud[48], Christian Sardet[16], Mike Sieracki[49,50], Sabrina Speich[5,29], Lars Stemmann[5,16], Matthew B. Sullivan[14], Shinichi Sunagawa[17,31], Didier Velayoudon[51], Jean Weissenbach[2,3,4], and Patrick Wincker[2,3,4,5]**

[1]Stazione Zoologica Anton Dohrn, Naples, Italy, [2]CEA - Institut François Jacob, Genoscope, Evry, France, [3]CNRS UMR, Evry, France, [4]Université d'Evry Val d'Essonne, Université Paris-Saclay, Evry, France, [5]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/GOSEE, Paris, France, [6]Laboratoire des Sciences du Numérique de Nantes (LS2N) – CNRS, Université de Nantes, École Centrale de Nantes, IMT Atlantique, Nantes, France, [7]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris, Paris, France, [8]INSERM, UMRS1138, Laboratory of Integrative Cancer Immunology, Centre de Recherche des Cordeliers, Paris, France, [9]CNRS, Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, France, [10]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India, [11]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Barcelona, Spain, [12]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS, EPHE, Université des Antilles, Paris, France, [13]Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, USA, [14]Department of Microbiology and Civil, Environmental, and Geodetic Engineering, The Ohio State University, Columbus, OH, USA, [15]Laboratory of Environmental Chemistry, School of Architecture, Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [16]Sorbonne Université, CNRS, UMR 7093, Institut de la Mer de Villefranche sur mer, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-Mer, France, [17]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [18]Institute for Anthropic impacts and Sustainability in the Marine Environment (IAS-CNR), Capo Granitola, Torretta Granitola, Italy, [19]LMD Laboratoire de météorologiedynamique. Ecole normale supérieure de Paris, PSL ResearchUniversity, Paris, France, [20]CNRS/INSU, IRD, MIO UM 110 Mediterranean Institut of Oceanography,

Université de Toulon, Aix Marseille Universités, Université de Toulon, Aix Marseille Universités, La Garde, France, [21]Institut Méditerranéen de Biodiversité et d'Ecologie Marine et Continentale (IMBE), UMR 7263 CNRS, IRD, Aix Marseille Université, Avignon Université, Station Marine d'Endoume, Marseille, France, [22]Information Génomique et Structurale, UMR7256, CNRS, Aix-Marseille Université, Institut de Microbiologie de la Méditerranée (FR3479), ParcScientifique de Luminy, Marseille, France, [23]CNRS, IRD, MIO, Aix Marseille Univ, Université de Toulon, Marseille, France, [24]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium, [25]VIB, Center for the Biology of Disease, Leuven, Belgium, [26]Biology Centre CAS, Institute of Parasitology, České Budějovice, Czech Republic, [27]Faculty of Science, University of South Bohemia, Ceské Budejovice, Czech Republic, [28]USR 3756 IP CNRS, Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, Paris, France, [29]University of Brest, Ifremer, CNRS, IRD, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, Brest, France, [30]IFREMER, Physiology and Biotechnology of Algae Laboratory, rue de l'Iled'Yeu, Nantes, France, [31]Institute of Microbiology, Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, Zurich, Switzerland, [32]Faculté des Sciences et Techniques, Université de Nantes, CNRS UMR6286, UFIP, Nantes, France, [33]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [34]CNRS, Biologie Intégrative des Organismes Marins (BIOM, UMR 7232), Observatoire Océanologique, Sorbonne Universités, UPMC Univ Paris 06, Banyuls sur Mer, France, [35]Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1, Heidelberg, Germany, [36]Max Delbrück Centre for Molecular Medicine, Berlin, Germany, [37]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany, [38]School of Marine Sciences, University of Maine, Orono, ME, USA, [39]Sorbonne University, CNRS, Station Biologique de Roscoff, UMR7144, ECOMAP, Roscoff, France, [40]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA, [41]Department of Oceanography, University of Hawaii, Honolulu, HI, USA, [42]Department of Earth Ocean and Ecological Sciences, School of Environmental Sciences, University of Liverpool, Liverpool, UK, [43]Laboratoire LOCEAN, Sorbonne Universités (UPMC, Univ Paris 06)-CNRS-IRD-MNHN, Paris, France, [44]Cell Biology and Biophysics, European Molecular Biology Laboratory, Heidelberg, Germany, [45]Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan, [46]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany, [47]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany, [48]Earth Institute, University College Dublin, Dublin, Ireland, [49]National Science Foundation, Arlington, VA, USA, [50]Bigelow Laboratory for Ocean Sciences East Boothbay, Boothbay, ME, USA, [51]DVIP Consulting, Sèvres, France

Abstract Predicting responses of plankton to variations in essential nutrients is hampered by limited in situ measurements, a poor understanding of community composition, and the lack of reference gene catalogs for key taxa. Iron is a key driver of plankton dynamics and, therefore, of global biogeochemical cycles and climate. To assess the impact of iron availability on plankton communities, we explored the comprehensive bio-oceanographic and bio-omics data sets from *Tara* Oceans in the context of the iron products from two state-of-the-art global scale biogeochemical models. We obtained novel information about adaptation and acclimation toward iron in a range of phytoplankton, including picocyanobacteria and diatoms, and identified whole subcommunities covarying with iron. Many of the observed global patterns were recapitulated in the Marquesas archipelago, where frequent plankton blooms are believed to be caused by natural iron fertilization, although they are not captured in large-scale biogeochemical models. This work provides a proof of concept that integrative analyses, spanning from genes to ecosystems and viruses to zooplankton, can disentangle the complexity of plankton communities and can lead to more accurate formulations of resource bioavailability in biogeochemical models, thus improving our understanding of plankton resilience in a changing environment.

Plain Language Summary Marine phytoplankton require iron for their growth and proliferation. According to John Martin's iron hypothesis, fertilizing the ocean with iron could dramatically increase photosynthetic activity, thus representing a biological means to counteract global warming. However, while there is a constantly growing knowledge of how iron is distributed in the ocean and about its role in cellular processes in marine photosynthetic groups such as diatoms and cyanobacteria, less is known about how iron availability shapes plankton communities and how they respond to it. In the present work, we exploited recently published *Tara* Oceans data sets to address these questions. We first defined specific subcommunities of co-occurring organisms that co-vary with iron availability in the oceans. We then identified specific patterns of adaptation and acclimation to iron in different groups of phytoplankton. Finally, we validated our global results at local scale, specifically in the Marquesas archipelago, where recurrent phytoplankton blooms are believed to be a result of iron fertilization. By

integrating global data with a localized response, we provide a framework for understanding the resilience of plankton ecosystems in a changing environment.

1. Introduction

Marine plankton play critical roles in pelagic oceanic ecosystems. Their photosynthetic component (phytoplankton, consisting of eukaryotic phytoplankton and cyanobacteria) accounts for approximately half of Earth's net primary production, fueling marine food webs, and sequestration of organic carbon to the ocean interior. Phytoplankton stocks depend on the availability of primary resources such as nutrients that are characteristically limiting in the oligotrophic ocean. For example, high-nutrient low-chlorophyll (HNLC) regions are often lacking the key micronutrient iron and increased bioavailability of iron will typically trigger a phytoplankton bloom (Boyd et al., 2007). Notwithstanding, the community response and its impact on food web structure and biogeochemical cycles are seldom predictable. The composition of blooms when limiting nutrients are supplied as sudden pulses with respect to the pre-existing community has been only poorly explored and is even more elusive when comparing to situations when nutrients are in quasi-steady state. Characterizing these responses is crucial to anticipate future changes in the ocean yet is challenged by community complexity and processes that span from genes to ecosystems. Dissecting these processes would also enhance the robustness of existing biogeochemical models and improve their predictive power (Stec et al., 2017).

In this report we explore the responsiveness of plankton communities to iron and assess the representation of iron bioavailability in biogeochemical models. Using global comprehensive metagenomics and metatranscriptomics data from *Tara* Oceans (Alberti et al., 2017; Bork et al., 2015; Carradec et al., 2018; Guidi et al., 2016), we examine abundance and expression profiles of iron-responsive genes in diatoms and other phytoplankton, together with clade composition in picocyanobacteria and the occurrence of iron-binding sites in bacteriophage structural proteins. These profiles are compared in the global ocean with the iron products from two state-of-the-art biogeochemical models. We further identify coherent subcommunities of taxa covarying with iron in the open ocean that we denote iron-associated assemblages (IAAs). Overall, our findings are congruent with the outputs from the models and reveal a range of adaptive and acclimatory strategies to cope with iron availability within plankton communities. As a further proof of concept, we track community composition and gene expression changes within localized blooms downstream of the Marquesas archipelago in the equatorial Pacific Ocean, where previous observations have suggested them to be triggered by iron (Martinez & Maamaatuaiahutapu, 2004), even though the biogeochemical models currently lack the resolution to detect the phenomenon. Our results indicate that iron does indeed drive the increased productivity in this area, suggesting that a pulse of the resource can elicit a response mimicking global steady state patterns.

2. Materials and Methods

2.1. Iron Concentration Estimates

Due to the sparse availability of direct observations of iron in the surface ocean, iron concentrations were derived from two independent global ocean simulations. The first is the ECCO2-DARWIN ocean model configured with 18-km horizontal resolution and a biogeochemical simulation that resolves the cycles of nitrogen, phosphorus, iron, and silicon (Menemenlis et al., 2008). The simulation resolves 78 virtual phytoplankton phenotypes. The biogeochemical parameterizations, including iron, are detailed in Follows et al. (2007). In brief, iron is consumed by primary producers and exported from the surface in dissolved and particulate organic form. Remineralization fuels a pool of total dissolved iron, which is partitioned between free iron and complexed iron, with a fixed concentration and conditional stability of organic ligand. Scavenging is assumed to affect only free iron, but all dissolved forms are bioavailable. Atmospheric deposition of iron was imposed using monthly fluxes from the model of Mahowald et al. (2005).

PISCES (Aumont et al., 2015) is a more complex global ocean biogeochemical model than ECCO2-DARWIN, representing two phytoplankton groups, two zooplankton grazers, two particulate size classes, dissolved inorganic carbon, dissolved organic carbon, oxygen, and alkalinity, as well as nitrate, phosphate, silicic acid, ammonium, and iron as limiting nutrients. In brief, PISCES accounts for iron inputs from

dust, sediments, rivers, sea ice, and continental margins, and flexible Michaelis-Menten-based phytoplankton uptake kinetics result in dynamically varying iron stoichiometry and drives variable recycling by zooplankton and bacterial activity. Iron loss accounts for scavenging onto sinking particles as a function of a prognostic iron ligand model, dissolved iron levels, and the concentration of particles. Iron loss from colloidal coagulation is also included and accounts for both turbulent and Brownian interactions of colloids. The PISCES iron cycle we use is denoted as "PISCES2" (Tagliabue et al., 2016) performed at the upper end of a recent intercomparison of 13 global ocean models that included iron.

2.2. In Situ Data

To generate a limited data set of observed dissolved iron concentrations for this analysis, we used a dissolved iron database updated from Tagliabue et al. (2012). For this we searched for the nearest available observation at the same depth as the *Tara* Oceans sampling and collected data that were within a horizontal radius of 2° from the sampling coordinates.

2.3. Marquesas Archipelago Sampling

Four stations within the Marquesas archipelago were sampled during the *Tara* Oceans expedition in August 2011 (Bork et al., 2015) using protocols described in Pesant et al. (2015): They were denoted TARA_122, TARA_123, TARA_124, and TARA_125. The sample details and physicochemical parameters recorded during the cruise are available at PANGAEA (http://www. pangaea.de), and nucleotide data are accessible at the ENA archive (http://www.ebi.ac.uk/ena/) (see further details below).

The study was initiated by releasing a glider that characterized the water column until the end of the experiment. First, the mapping of the water column structure via real-time analysis of glider data was conducted. After this initial step, the continuous inspection of near real-time satellite color chlorophyll images and altimetric data revealed a highly turbulent environment, with a mixed layer up to 100-m deep and strong lateral shearing, especially downstream of the islands, which generated an area of recirculation in the wake of the main island (Nuku Hiva). A series of four sampling stations was then planned and executed by performing the full set of measurements and sampling using the *Tara* Oceans holistic protocol (Pesant et al., 2015). Station TARA_122 sampled the HNLC prebloom waters upstream of the islands and thus served as a reference station for the others. This station was located 27-km upstream of the island of Nuku Hiva.

2.4. Oceanographic Observations

The Biogeochemical Argo float deployed in the framework of the Marquesas study (WMO 6900985) was a PROVBIO-1 free-drifter profiler (Xing et al., 2012). It was based on the "PROVOR-CTS3" model, equipped with a standard CTD sensor (to retrieve temperature and salinity parameters) together with bio-optical sensors for the estimation of chlorophyll-*a* concentrations, colored dissolved organic matter, and backscatter at 700 nm. It was also equipped with a radiometric sensor to estimate spectral downward irradiance at three wavelengths (412, 490, and 555 nm) and with a beam transmissometer. The data processing is discussed in Xing et al. (2012). The profiling float was programmed to adopt a modified standard Argo strategy (Freeland & Cummins, 2005). After deployment, it navigated at 700-m depth, to a daily maximum of 1,000 m, and then surfaced a first time, generally early in the morning. It then submerged again to a depth up to 400 m, to again reach the surface approximately at noon. A third profile to 400 m, followed by a subsequent resurfacing, was performed at the end of the day. During all the ascending phases, a complete profile of all the available parameters was collected. At surface, the obtained data were transmitted to land through a satellite connection and the profiler descended again to 1,000 m to start another cycle. The Biogeochemical Argo was deployed on-site at Station TARA_123 on 2 August 2011. It performed 55 profiles in the Marquesas region, before moving westward in early October (then outside the study area), and then southward. It definitively ceased to function in December 2012, approximately 400 km south of the Marquesas islands and after collecting more than 150 profiles.

An autonomous glider was also deployed in the study area. A complete description of glider technology and functioning is available in Testor et al. (2010). This glider was able to reach 1,000-m depths. It was equipped with temperature and salinity sensors, an optode for oxygen concentration measurements, two Wetlab ecopucks with two fluorometers for chlorophyll and colored dissolved organic matter concentrations, and three backscatterometers to estimate backscatter coefficients at three wavelengths (532, 700, and 880 nm). The glider was deployed on 16 July 2011 (approximately 1 month before TARA arrived in the Marquesas

archipelago), close to the position of Station TARA_122. It was recovered on 5 August 2011 by TARA because a malfunction in the tail rudder had been detected. It performed approximately 250 profiles, with 35 dives at 1,000-m depths and 90 dives at 500-m depths.

Analysis of trace metals was performed exclusively at the Marquesas Islands sampling stations (Stations TARA_122-TARA_125) following the methods reported in Scelfo (1997). Dissolved iron was not measured due to lack of technical resources.

2.5. Network Analysis and Correlations With Iron

A co-occurrence network analysis similar to that reported in Guidi et al. (2016) was performed to delineate feature subnetworks of prokaryotic and eukaryotic lineages, as well as viral populations, based on their relative abundance. All procedures were applied on 103 sampling sites (Guidi et al., 2016) after excluding outliers (Stations TARA_82, TARA_84 and TARA_85) on Hellinger-transformed log-scaled abundances. Computations were carried out using the R package WGCNA (Langfelder & Horvath, 2007). After building a co-occurrence weighted graph, a hierarchical clustering was performed. This resulted in the definition of several subnetworks or modules, each represented by its first principal component, called *module eigen value*. Associations between the calculated subnetworks and a given trait were measured by the pairwise Pearson correlation coefficients, as well as with corresponding p values corrected for multiple testing using the Benjamini and Hochberg false discovery rate (FDR) procedure, between the considered environmental trait and their respective principal components. The results are reported in the first 10 columns of the heatmap in Figure S1a in the supporting information. The subnetworks that showed the highest correlation scores are of interest to emphasize a putative community associated with a given environmental trait. In addition to the multiple environmental parameters previously reported (Guidi et al., 2016), we simulated iron bioavailability in *Tara* Oceans stations based on the two different models of iron concentration in the global oceans: the ECCO2-DARWIN model (Menemenlis et al., 2008) and the PISCES2 model (Aumont et al., 2015). Both models performed well in the recent global iron model intercomparison project (Tagliabue et al., 2016), and so we conducted an assessment of model outputs at *Tara* Oceans sampling locations using compilations of iron observations (Tagliabue et al., 2016) augmented by those from the GEOTRACES program (Mawji et al., 2014). ECCO2-DARWIN-derived estimates (57 stations at surface) and PISCES2 model (83 stations at surface, 44 of which also at maximum chlorophyll depth) can be found in Table S1a. For further details on the models and for a comparison of the two, see the supporting information. We then identified eukaryotic, prokaryotic, and viral subnetworks that correlated most strongly with iron bioavailability, denoted IAAs. Four IAAs consisting of eukaryotic metabarcodes (de Vargas et al., 2015) were significantly associated with iron. Similarly, four viral IAAs could be identified by analysis of viral communities. Based on taxonomy, no prokaryotic IAAs with significance could be identified; however, when considering prokaryotic genes (as described in Guidi et al., 2016), five subnetworks of prokaryotic genes could be identified.

In addition to the network analyses, we examined whether the identified subnetworks can be used as predictors of iron bioavailability. Following the protocol described in Guidi et al. (2016), we used *partial least square regression*, which is a dimensionality-reduction method that aims to determine predictor combinations with maximum covariance with the *response variable*. The predictors were ranked according to their *value importance in projection* (VIP) using the R package *pls* (Mevik & Wehrens, 2007). For each eukaryotic IAA, their relative contribution to each sample was estimated by computing the first eigen value.

2.6. Taxonomy Determinations

Taxonomic studies were performed using various methods (photosynthetic pigments, flow cytometry, and optical microscopy for phytoplankton and zooplankton as detailed in Villar et al. (2015); phytoplankton counts using unfiltered bottles or nets as described in Malviya et al. (2016) and Villar et al. (2015); and mesozooplankton samples collected by vertical tows with a WP2 net (200-μm mesh aperture) from 100-m depth to the surface during the day, followed by fixation in buffered formaldehyde (2–4% final concentration), and later analyzed in the laboratory). Data from an Underwater Vision Profiler (UVP) were used to determine particle concentrations and size distributions >100 μm, (Campbell et al., 1994). To have an estimate of biomass variations in the different compartments at the Marquesas Islands, we applied empirical formulas to transform Chlorophyll-*a* (phytoplankton) or body measurements (zooplankton) to biomass. To this

purpose, we used the ratio of phytoplankton biomass to Chlorophyll-*a* (Phyto C: Chl a) in the euphotic zone as previously estimated in an area with similar biogeochemical features (Campbell et al., 1994—See Table S1b). Of note, our aim was not to determine absolute biomass but to estimate variations in biomass between the Marquesas Islands stations. To estimate the total phytoplankton biomass, Chl a concentration from HPLC data was thus used. The relative contribution of microplankton, nanoplankton, and picoplankton to the $[Chla]_{tot}$ was estimated according to Uitz et al. (2006). The biomass of large zooplankton was estimated using previously published conversion factors from body length to carbon content (C:L) in selected zooplankton lineages. Individual body measures were estimated from literature considering similar community composition, with the exception of the Copepoda prosome length, which was herein measured. Zooscan (Bongo net, >300 μm) derived abundance data (ind × m^{-3}) were used to evaluate the total biomass along the water column.

2.7. Genomic Analyses

Eukaryotes larger than 5 μm were collected directly from the ocean using nets with different mesh sizes while smaller organisms and viruses were sampled by peristaltic pump followed by on-deck filtration. Several filtration steps were performed using membranes with different pore sizes to obtain size-fractionated samples corresponding to viruses (<0.1 and 0.1–0.2 μm), prokaryotes (0.2–3 μm), and eukaryotes (0.8–5, 5–20, 20–180, and 180–2,000 μm). In this study, we only used samples collected from the surface water layer. Details about genomics methods are available in Carradec et al. (2018) and in the following publications: virus metagenomes (Roux et al., 2016); prokaryote metagenomes (Sunagawa et al., 2015); eukaryote metabarcoding (de Vargas et al., 2015); and eukaryote metagenomes and metatranscriptomes (Alberti et al., 2017; Carradec et al., 2018). The abundance of individual genes was assessed by normalization to the total number of sequences within the same organismal group (Carradec et al., 2018). Cyanobacterial clade absolute cell abundance was assessed using the *petB* marker gene, as described in Farrant et al. (2016), in combination with flow cytometry counts using the method published by Vandeputte et al. (2017).

Metatranscriptomic and metagenomic unigenes were functionally annotated using PFAM (Finn et al., 2016) as the reference database and search tool (Katoh & Standley, 2013). To detect the presence of genes encoding silicon transporters, ferritin, proteorhodopsin, FBAI, and FBAII among the unigene collection, the profile hidden Markov models of the PFAMs PF03842, PF0210, PF01036, PF00274, and PF01116, respectively, were used, with HMMer v3.2.1 with gathering threshold option (http://hmmer.org/). It is important to note that flavodoxin (PF00253, PF12641, and PF12724), ferredoxin (PF00111), and cytochrome c_6 (PF13442) PFAM families do not discriminate those sequences involved in photosynthetic metabolism from other homologous sequences. The photosynthetic isoforms for flavodoxin, ferredoxin, and cytochrome c_6 were therefore determined by phylogeny, as described below.

To discriminate the photosynthetic isoforms from other homologous sequences, we started with the results from HMMer and then built libraries composed of well-known reference sequences (manually and experimentally curated) from both photosynthetic and nonphotosynthetic pathways. To enrich our libraries, we used the reference sequences to find similar sequences by using BLAST search tool ("tBLASTn" program with an e−5e value threshold) against phyloDB reference database (Dupont et al., 2015). Next, we used MAFFT version 7 using the G-INS-I strategy (Katoh & Standley, 2013). The corresponding phylogenetic reference trees were generated with PhyML 3.0 (Guindon et al., 2010) using the LG substitution model with four categories of rate variation. The starting tree was a BIONJ tree, and the type of tree improvement was subtree pruning and regrafting. Branch support was calculated using the approximate likelihood ratio test with a Shimodaira-Hasegawa-like procedure. We then manually identified the branches containing the photosynthetic versions and those with nonphotosynthetic proteins. We ensured that the approximate likelihood ratio test values of the photosynthetic and nonphotosynthetic branches were higher than 0.7 by retaining only the most conserved matches in our trees. Finally, we realigned and labeled the unigenes against the reference trees depending on the placement of each translated unigene on them.

While HMMer has the highest sensitivity among the classical domain detection approaches, not all the references collected by PFAM are sufficiently rich with HMMer to maintain the same detection (Bernardes et al.,

2016). To deal with the poor representation of *ISIP* genes in the PFAM database and to improve their detection, we adopted a simplified version of the approach presented in Bernardes et al. (2015) to build our own pHMM to detect the different conserved regions represented by ISIP1, ISIP2a, and ISIP3 amino acid sequences. For this, we collected all the sequences in the reference literature (Allen et al., 2008; Chappell et al., 2015; Lommer et al., 2010; Morrissey et al., 2015), all 35 sequences belonging to PFAM PF07692 and the 56 most conserved sequences from PF03713 (all the seeds).

2.8. Data and Code Availability

Sequencing data are archived at ENA under the accession number PRJEB4352 for the metagenomics data and PRJEB6609 for the metatranscriptomics data (Carradec et al., 2018). Environmental data are available at PANGAEA. The gene catalog, unigene functional and taxonomic annotations, and unigene abundances and expression levels are accessible at http://www.genoscope.cns.fr/tara/. Computer codes are available upon request to the corresponding authors.

2.8.1. Accession Numbers of Metagenomics and Metatranscriptomics Data

2.8.1.1. Sample

ERS492651, ERS492651, ERS492650, ERS492669, ERS492669, ERS492662, ERS492662, ERS492658, ERS492658, ERS492650, ERS492740, ERS492742, ERS492742, ERS492751, ERS492751, ERS492763, ERS492763, ERS492757, ERS492740, ERS492757, ERS492825, ERS492825, ERS492824, ERS492824, ERS492829, ERS492852, ERS492852, ERS492846, ERS492846, ERS492829, ERS492897, ERS492897, ERS492895, ERS492895, ERS492912, ERS492912, ERS492909, ERS492909, ERS492904, ERS492904.

2.8.1.2. Experiment

ERX948080, ERX948010, ERX1782415, ERX1782384, ERX1782327, ERX1796912, ERX1796638, ERX1796690, ERX1796805, ERX1782126, ERX1782109, ERX1782245, ERX1796854, ERX1796544, ERX1782292, ERX1782172, ERX1782221, ERX1796700, ERX1796855, ERX1782301, ERX1782464, ERX1782128, ERX1789668, ERX1789366, ERX948029, ERX948074, ERX1796627, ERX1796773, ERX1789369, ERX1789449, ERX1796931, ERX1796605, ERX1789426, ERX1789575, ERX1789524, ERX1796866, ERX1796524, ERX1789649, ERX1789612, ERX1789647, ERX1796596, ERX1796836, ERX1789655, ERX1789574, ERX1789407, ERX1782118, ERX1782283, ERX947973, ERX948088, ERX1789391, ERX1789539, ERX1789587, ERX1796687, ERX1796586, ERX1796703, ERX1789662, ERX1789616, ERX1789589, ERX1796662, ERX1796518, ERX1796678, ERX1796698, ERX1782217, ERX1782352, ERX1796645, ERX1796858, ERX1796924, ERX1789675, ERX1789597, ERX1789700, ERX1789362, ERX1782350, ERX1782418, ERX947994, ERX948064, ERX1789361, ERX1789368, ERX1789532, ERX1796658, ERX1796818, ERX1796632, ERX1789638, ERX1789548, ERX1789579, ERX1796921, ERX1796732, ERX1796741, ERX1789714, ERX1789489, ERX1789628, ERX1796689, ERX1796850, ERX1796523, ERX1782181, ERX1782370, ERX1796607, ERX1796738, ERX1796714, ERX1789437, ERX1789516, ERX1789417.

2.8.1.3. Run

ERR868475, ERR868513, ERR1712182, ERR1712118, ERR1711869, ERR1726556, ERR1726667, ERR1726938, ERR1726688, ERR1712207, ERR1711933, ERR1711897, ERR1726927, ERR1726932, ERR1712069, ERR1712197, ERR1711986, ERR1726883, ERR1726891, ERR1712219, ERR1711929, ERR1711951, ERR1719463, ERR1719159, ERR868466, ERR868469, ERR1726762, ERR1726913, ERR1719393, ERR1719310, ERR1726961, ERR1726522, ERR1719437, ERR1719413, ERR1719343, ERR1726622, ERR1726721, ERR1719297, ERR1719410, ERR1719307, ERR1726770, ERR1726561, ERR1719256, ERR1719298, ERR1719217, ERR1711914, ERR1711917, ERR868363, ERR868489, ERR1719301, ERR1719160, ERR1719214, ERR1726564, ERR1726725, ERR1726569, ERR1719448, ERR1719389, ERR1719194, ERR1726571, ERR1726533, ERR1726892, ERR1726601, ERR1711949, ERR1712155, ERR1726608, ERR1726657, ERR1726763, ERR1719391, ERR1719175, ERR1719381, ERR1719365, ERR1711882, ERR1711999, ERR868382, ERR868352, ERR1719395, ERR1719316, ERR1719207, ERR1726643, ERR1726714, ERR1726846, ERR1719404, ERR1719213, ERR1719459, ERR1726822, ERR1726912, ERR1726691, ERR1719356, ERR1719145, ERR1719293, ERR1726695, ERR1726666, ERR1726903, ERR1712102, ERR1711923, ERR1726745, ERR1726946, ERR1726765, ERR1719295, ERR1719249, ERR1719385.

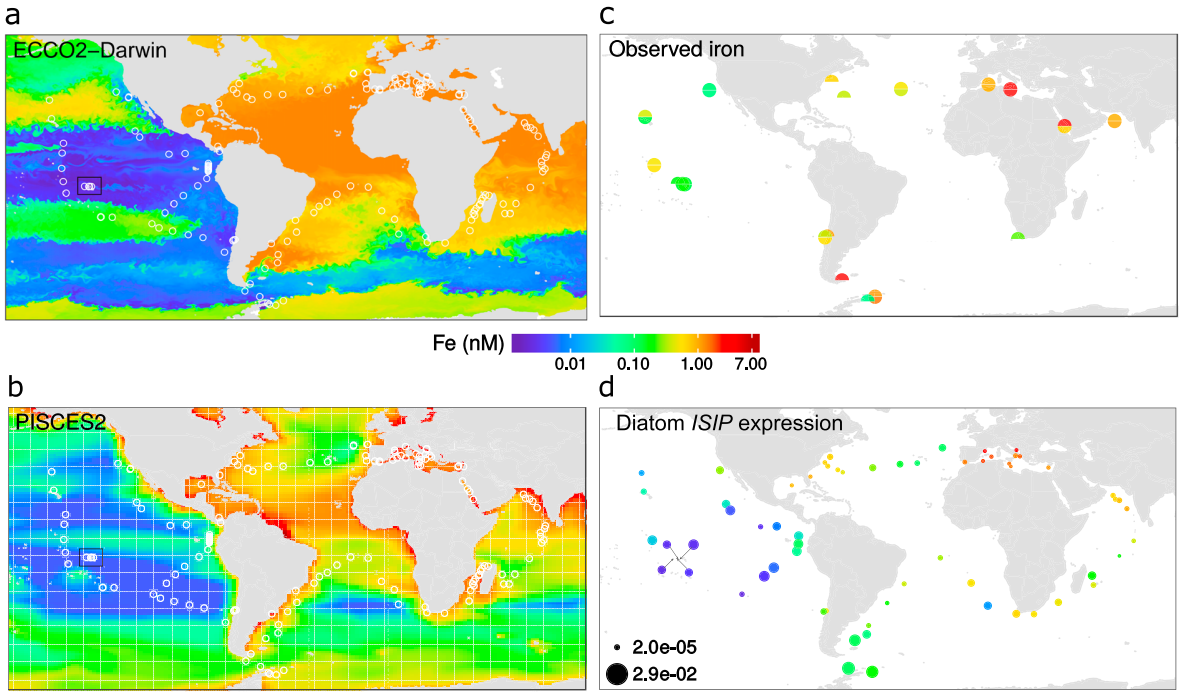


Figure 1. Comparison of ECCO2-DARWIN and PISCES2 iron estimates with observed data and expression of diatom *ISIP* genes at *Tara* Oceans stations. Maps of (a) annual average iron concentrations from the ECCO2-DARWIN model (57 stations at surface), (b) from the PISCES2 model (83 stations at surface, 44 of which also at deep chlorophyll maximum depth), and (c) from the observed data where it was available at less than 2° radius distance from locations of the *Tara* Oceans sampling sites (20 stations at surface, 16 of which also at deep chlorophyll maximum depth). Each circle corresponds to a sampling site, where the upper semicircle is filled according to the surface iron concentration while the lower semicircle is filled according to the deep chlorophyll maximum depth where available. Color scale indicates dissolved iron concentrations expressed in nM. (d) Biogeographical pattern of diatom *ISIP* gene expression. The circle colors represent iron concentration estimates at each *Tara* Oceans sampling site according to PISCES2 model (Table S1a). The abundance of *ISIP* transcripts was normalized by the total abundance of all diatom unigenes at each station, and the corresponding values are represented by the circle area. Boxes indicate the Marquesas Islands sampling area.

3. Results

3.1. Modeled Iron Distributions Are Highly Correlated With the Expression of Marker Genes for Iron Limitation

Iron is a complex contamination-prone micronutrient whose bioavailability is difficult to assess in the ocean (Tagliabue et al., 2017). Rather than using single discrete measurements, we linked observed differences in plankton communities at sites sampled during the *Tara* Oceans expedition (Bork et al., 2015) with the range of iron conditions typical of each location. Specifically, we extracted annual mean iron concentrations and their variability from two state-of-the-art ocean models (ECCO2-DARWIN [Menemenlis et al., 2008] and PISCES2 [Aumont et al., 2015]) and analyzed their correspondence with the best available estimates based upon in situ data (a compilation of iron observations [Tagliabue et al., 2012] merged with GEOTRACES data [Mawji et al., 2014; Tagliabue et al., 2012] in a manner similar to previous studies [Toulza et al., 2012]; Figure 1).

To assess the reliability of the modeled iron distributions, we correlated the expression of diatom *ISIP* genes in metatranscriptomics data sets with the annual means of iron concentrations estimated by the DARWIN model, and with annual and monthly means by the PISCES2 model (Carradec et al., 2018; Table S1a and supporting information S1). These genes have been found in multiple previous studies to be inversely correlated with iron availability (Allen et al., 2008; Chappell et al., 2015; Graff van Creveld et al., 2016; Marchetti et al., 2017; Morrissey et al., 2015). Figure 1 presents a comparison between the estimates of dissolved iron concentrations derived from the annual mean iron field from PISCES2 and ECCO2-DARWIN (Table S1a), with the *Tara* Oceans stations superimposed and best available estimates based upon in situ

measurements (Figure 1). In spite of the evident scarcity of actual iron concentration data (which illustrates the need to use models for estimating iron in the current exercise; Figure 1), both models and *ISIP* mRNA levels describe very satisfactorily the global-scale gradients, with the highest concentrations of iron observed in the Mediterranean and Arabian Seas (both highly impacted by desert dust deposition) and the lowest in the tropical Pacific and Southern Oceans. This demonstrates that the geographical coverage of the *Tara* Oceans expedition is well suited to studies of the role of iron on euphotic planktonic ecosystems. The available data (Figure 1; Table S1a) further indicates that the gradients of iron appear to be better captured by PISCES2, a more complex and recent model (Aumont et al., 2015). This is for instance the case for the North Atlantic Ocean and the Mediterranean Sea, where longitudinal gradients are stronger in PISCES2 and are consistent with *ISIP* gene levels, while ECCO2-DARWIN seems to overestimate iron in the Eastern Atlantic Ocean and underestimate it in the Mediterranean Sea. The opposite is true in the South Atlantic Ocean, where *ISIP* mRNA levels show a clear increase correlated with iron stress between South America and Africa (Figure 1). Overall, in the Atlantic ECCO2-DARWIN has higher concentrations, and thus, a clearer large-scale Atlantic-Pacific gradient is observed.

The Pacific and Southern Oceans (subpolar and polar stations TARA 81–85) are both characterized by low levels of iron, as mentioned above. Notably, PISCES2 has a rather flat distribution in the Pacific Ocean, with very low values, while the other model shows a relatively higher level of iron at the core of the subtropical gyres, that is, close to the Hawaii Islands (Stations TARA_131 and TARA_132) and offshore from South America (TARA_98 and close-by stations) that seems to be in agreement with *ISIP* mRNA levels (at least for the Hawaiian sample—Figure 1). These are very oligotrophic oceanic regions, where nitrate is also a strongly limiting nutrient. Again, the *ISIP* expression pattern in Figure 1 is closer to the PISCES2 model, in that it shows a clear reduction of the stress resulting from iron deprivation within these gyres. Finally, while a significant increase in iron at the Equator may be expected as a consequence of the upwelling in this region, both the models and the *ISIP* levels (at Station TARA_128) suggest that this area is rather characterized by low values of iron. Overall, our analysis indicates that both models correlated very well with *Tara* Oceans transcriptomic data, with no relevant differences among monthly and yearly values, but with annual means from the ECCO2-DARWIN estimates showing the best reliability (Table S1c). This analysis also indicates that metatranscriptomics is now mature enough to provide an independent, biologically based validation of ecosystem models.

3.2. Plankton Response to Iron Availability Is Coordinated at Subcommunity Level

The higher level organization of plankton communities, and its possible relationship with the roles of individual constituents, has been highlighted previously in an analysis of the potential links between community structure and carbon export using data from *Tara* Oceans (Guidi et al., 2016). We here used this approach to explore plankton ecosystem responses to iron bioavailability using an end-to-end approach from genes to communities and from viruses to metazoa to reveal community responses at global scale (see section 2). Known as weighted gene correlation network analysis (WGCNA; see section 2 for further description; Guidi et al., 2016; Langfelder & Horvath, 2007), this approach deciphers subcommunities (modules) of organisms within a global co-occurrence network, and because of the high levels of covariation of individual taxa, it is possible to deduce putative ecological interactions. As proxies for organism abundance we used the relative abundances of eukaryotic lineages (defined as operational taxonomic units; OTUs) derived from 18S-V9 rDNA metabarcoding data (de Vargas et al., 2015). WGCNA generated a total of 31 modules. Each module groups a subset of eukaryotic taxa found in *Tara* Oceans samples whose pairwise relative abundance was highly correlated over all the sampling sites; that is, they have a high probability of co-occurrence and to change their abundance in a coordinated way. Because they react in phase, we can infer that within each subcommunity these organisms have a higher probability of interaction among themselves than with the organisms in other modules.

We found four eukaryotic subnetworks significantly associated with the ECCO2-DARWIN-derived and/or with the PISCES2-derived estimates of iron concentrations in the global ocean (Figures 2a, S1a, and S1b; Table S1d). The Black and Turquoise modules were associated with high significance to the iron concentrations generated by both models whereas the DarkRed and Yellow modules were better associated with ECCO2-DARWIN and PISCES2, respectively, Black (DARWIN: $R = 0.37$, $P = 6 \times 10^{-4}$; PISCES2: $R = 0.38$, $P = 3 \times 10^{-4}$), Turquoise (DARWIN: $R = 0.46$, $P = 1 \times 10^{-5}$; PISCES2: $R = 0.42$, $P = 9 \times 10^{-5}$),

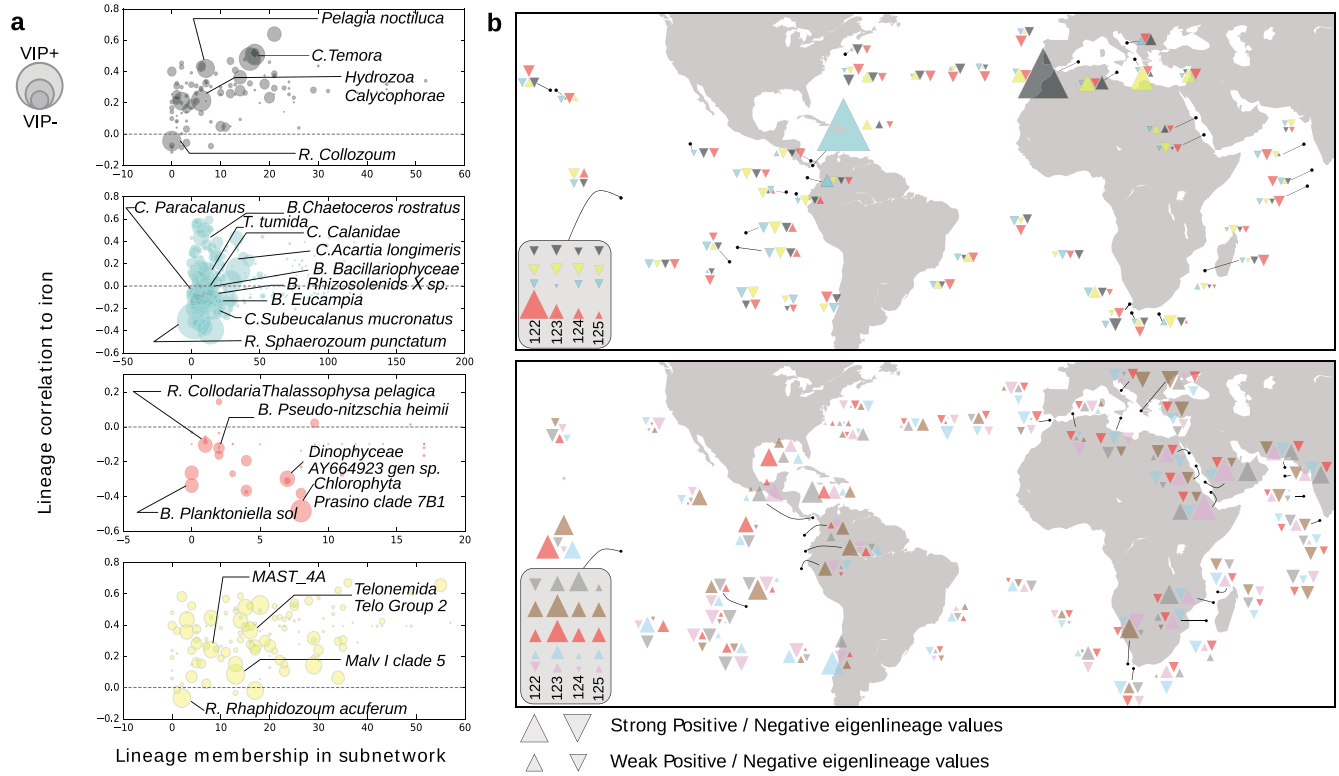


Figure 2. Planktonic iron-associated assemblages (IAAs) in the global ocean and in the Marquesas Islands stations. (a) Description of eukaryotic modules associated with iron. Relative abundances and co-occurrences of eukaryotic lineages were used to decipher modules. Four modules can predict iron with high accuracy: Black, DarkRed, Turquoise, and Yellow. For each IAA, lineages are associated to their score of centrality (*x*-axis), to their correlation with iron concentrations (*y*-axis), and their VIP score (circle area). Representative lineages within each module are emphasized by circles and named (C = Copepoda; B = Bacillariophyta; R = Rhizaria). (b) Upper panel: contribution of *Tara* Oceans stations to the global variance of IAAs of eukaryotic lineages. For each IAA, we represent the projection of stations on the first principal component (upper panel). Lower panel: projection of the relative contribution of the *Tara* Oceans stations to the global variance of iron-associated prokaryotic gene assemblages, as revealed by WGCNA. For each prokaryotic gene module associated with iron (from top to bottom: Grey60, Plum1, Red, SkyBlue, and SaddleBrown), we represent the projection of stations on the first principal component, proportional to triangle sizes for each module. The behavior of each IAA in the Marquesas archipelago stations is shown in the inset.

DarkRed (DARWIN: $R = -0.43$, $P = 5 \times 10^{-5}$; PISCES2: $R = 0.19$, $P = 0.08$), and Yellow (DARWIN: $R = 0.19$, $P = 0.09$; PISCES2: $R = 0.56$, $P = 5 \times 10^{-8}$), and contained between 31 and 591 different OTUs (Tables S1d and S1e). These subnetworks were denoted IAAs. For each IAA subnetwork, WGCNA computes a single representative as a combination of lineages. Such a score, denoted as "module eigengene" score (hereafter termed an eigenlineage score), represents the first eigenvector of the assemblage (Langfelder & Horvath, 2007). Projections of samples on such an eigenvector show the relative importance of samples to the global variance of each IAA. Together with their contribution, in terms of OTU abundance to the total eukaryotic abundance in each station (Table S1f), they provide clues to interpret the link between modules and iron availability. The mismatch in some regions between the two models (see above) is likely the reason why the significance of association of the Yellow module with ECCO2-DARWIN, whose variance and representativeness is particularly significant in the South Adriatic and is minimally present in the Peruvian upwelling area, is much less than that with PISCES2. By contrast, the DarkRed module, which appears to be the best indicator module for the Marquesas area (Figure 2b, upper panel) and is highly relevant in the Peruvian upwelling region, displays a much less significant association and an opposite variation with PISCES2 iron versus ECCO2-DARWIN iron. The IAAs show slightly different, often antagonistic, variance contributions at global scale (Figure 2b, upper panel), with each of them being particularly responsive, in terms of variance, in specific sites, for example, the Yellow module in the Eastern Mediterranean Sea.

We examined the lineage composition of each IAA and the relevance of each taxon within them by determining the relative abundance of each lineage with respect to iron concentration estimates and their centrality within the module (see section 2). The results are reported in Tables S1d and S1e. The IAAs displayed significant differences in terms of numbers of lineages and compositions, with the Turquoise module being the largest and dominated by consumers, predominantly metazoans, and the DarkRed module being the smallest. The Black module displayed the highest proportion of autotrophs, while the DarkRed IAA displayed the highest proportion of diatoms (Bacillariophyta; 57% of all autotrophic protists).

To reduce complexity further, we screened the networks in terms of the VIP score of each node (i.e., the OTUs displaying the highest statistical weight in differentiating sites because of iron availability; section 2; Table S1d; Figures 2a and S1c). Species with high VIP scores can be predicted to be particularly important in reflecting the adjustments of each module via their specific interactions with other members of their subcommunity. Although interpreting why high VIP taxa are related to iron bioavailability is often severely restricted by our knowledge of plankton functional ecology and interorganism interactions, in other cases the role of VIP taxa within the modules is clearer. As an example, identification of an IAA in which several diatoms have the highest VIPs (DarkRed module, eight subnetwork members, −0.337 correlation with iron), commonly found in the most severely iron-limited regions of the world's ocean and often the most responsive groups in mesoscale iron fertilization experiments (Boyd et al., 2007; Marchetti et al., 2006), suggests a strong physiological plasticity of these groups (Greene et al., 1991; Lommer et al., 2012). The fact that *Pseudo-nitzschia* is among the highest scoring VIP genera in the DarkRed module further suggests that this genus tracks regions with low iron bioavailability, being able to profit from it when it becomes available. Other examples concern metazoans: copepods from the genus *Temora* (high subnetwork centrality and strong correlation with iron) are known to be iron-limited (Chen et al., 2011), and the two cnidarian lineages—the class *Hydrozoa* and the genus *Pelagia* (both of which display relatively strong subnetwork centrality and strong correlations with iron)—suggest strong predator-prey links.

Considering the ECCO2-DARWIN-derived VIP scores, lineages with the highest scores (>1) could predict as much as 61.9%, 52.6%, 49.1%, and 38.1% (in the Turquoise, Black, DarkRed, and Yellow IAAs, respectively; leave-one-out cross-validated) of the variability of iron in the oligotrophic ocean. When the PISCES2-derived VIP scores are taken into account, the predictive potential of the IAAs is even higher: 73.2% (Turquoise), 61.9% (Yellow), 59.0% (Black), and 54.4% (DarkRed). More importantly, the VIP scores obtained with the two models for each OTU showed an extremely good covariance (Figure S1d). This confirms the biological coherence and stability of the modules and their components to iron availability despite the occasional mismatch in the predictions of the two models.

Of the photosynthetic groups, autotrophic dinoflagellate taxa were particularly relevant in the Turquoise and Black modules, diatoms were relevant in the DarkRed module, and haptophytes were significantly present in the Yellow module. Metazoans were particularly important in the Black and the Turquoise modules, and marine stramenoplies/marine alveolata groups of phagotrophic and parasitic heterotrophs were relevant in the Black (marine alveolata), Turquoise, and Yellow modules (marine stramenoplies; Figures 2a and S1c; Tables S1d and S1e). This hints at particularly intricate, and still elusive, interactions among organisms that ultimately lead to the observed collective responses.

To further interpret the patterns observed for the IAAs, we chose two additional modules, denoted DarkGrey and Red, because of the different correlations of diatoms within these modules to iron concentrations with respect to the DarkRed module (Figures S1a and S1c). By examining the abundance of the components of each module at different sampling sites (Table S1f), the results suggest that the Turquoise module groups lineages relevant in all of the main oceanic biogeographic regions with the exception of the Mediterranean basin, and with a prominent weight in the Southern Ocean. By contrast, the Black and Yellow modules are of particular importance in the Mediterranean Sea, while other IAAs have minor contributions. The DarkRed module is generally poorly represented; however, in the South Pacific and in particular around the Marquesas Islands, its relevance is high (Figure 2b, upper panel; Table S1f).

Based on all of the above information, we then sketched the ecological profiles of the seven modules, summarized below:

Black IAA: Ubiquitous, but with low abundance except in the Mediterranean basin, and composed principally of heterotrophic organisms (protists and metazoans; Tables S1e S1f). Dinophytes are the autotrophic component of this module while diatoms are poorly represented. Around the Marquesas Islands, its weight is constantly low. Lineages are positively correlated or loosely anticorrelated with iron (Figures 2a and 2b; Table S1d). This module has an intermediate level of internal connectivity and suggests top heavy (pyramidal) trophic interactions. The assemblage resembles a typical pattern in a postbloom phase, with biomass accumulated in the metazoan compartment. No significant differences are seen when the ECCO2-DARWIN-derived and PISCES2-derived VIPs are compared since the module is not relevant in areas where the two models disagree. This pattern is consistent with the differences detected at molecular level.

DarkRed IAA: The module is not particularly significant at global scale in terms of abundance (Figure 2b, upper panel; Table S1e). It contains a small number of lineages with a high relative weight of diatoms and few metazoans but no copepods, with carbon recycling mostly in the protistan compartment. This module is particularly intriguing because, with very few exceptions, all the lineages including diatoms are negatively correlated with iron (Figure S1c). It is particularly responsive in the Marquesas area but is also present in offshore South American upwelling areas. The internal connectivity is of an intermediate level (Table S1d). These features hint at an assemblage in the subtropical ocean driven by the activity of diatoms thriving in regions of low iron availability (while exploiting a higher than average silicon availability), thus showing an inversion of the pattern compared to high iron regions (Figure 2b, upper panel). Significantly, its abundance drops at Station TARA_123 in the Marquesas archipelago (see below).

Turquoise IAA: Ubiquitous, with a general high weight in terms of abundance, and very abundant in the Southern Ocean (in particular in stations TARA_85–88; Table S1f). The module includes relatively few diatoms, but many dinoflagellates (both autotrophic and heterotrophic species; Tables S1d and S1e). Copepods are the most numerous components and show the highest VIP scores. Of note, this module includes the crustacean order *Euphausiacea* (krill), which specifically emerges as having high VIP scores only when the PISCES2-derived iron estimates are used. Both internal connectivity and number of lineages are high (Table S1d). The module as a whole responds in the Marquesas area, especially at TARA_123 (Figure 2b, upper panel; Table S1f).

Yellow IAA: This module is particularly important in South Adriatic and Eastern Mediterranean, as well as in the tropical North Atlantic (Figure 2b, upper panel; Table S1f). It includes relatively few metazoans and diatoms but a notable abundance of haptophytes and heterotrophic protists (Tables S1d and S1e). It displays a weak response in the Marquesas area (TARA_125; Figure 2b, upper panel) and seems to be less dependent on iron availability as compared to the other modules.

DarkGrey: Not an IAA and has a low weight in general, with a slight positive correlation to iron and only low internal connectivity. Diatoms in this module are very relevant (Tables S1d and S1e). It contains a high fraction of metazoans with fewer heterotrophic protists. This module displays a typical bottom heavy (pyramidal) structure with diatoms reacting positively to iron availability.

Red module: Not an IAA, but this module displays a similar response to iron than the DarkGrey module, with the main differences being that it contains few metazoans and the protist compartment is dominated by Dinophyceae. Diatoms are also dominant as autotrophic protists. It is the module that correlates the most with chlorophyll and primary productivity (Figure S1a) and seems to be associated with highly productive areas. It is thus not very relevant globally, with the exception of the South Atlantic Ocean, where it dominates the Benguela upwelling (Station TARA_67), a very rich region that is not iron limited. It is apparently driven by bottom-up flexible responses to iron availability, most probably by macronutrient availability (Tables S1d and S1e). It displays variable correlations of its members to iron and has also a bottom heavy pyramidal trophic structure.

Overall, our analysis strongly suggests that different subassemblages of co-occurring lineages can be pinpointed within communities that respond differently to resource limitation, mostly without marked geographical preferences albeit with high plasticity to iron availability. Particularly remarkable is the contrasting role shown by diatoms, with different lineages covering the full range of correlations with iron (Figure S1c), possibly linked to their different strategies for responding to the lack of a crucial resource. In some cases their communities share a similar response while in others the structure of the assemblage is modified. The further observation that co-occurrence of IAAs can show biogeographical patterns (Figure 2b, upper panel) that are not clearly emphasized by analysis of single eukaryotic groups

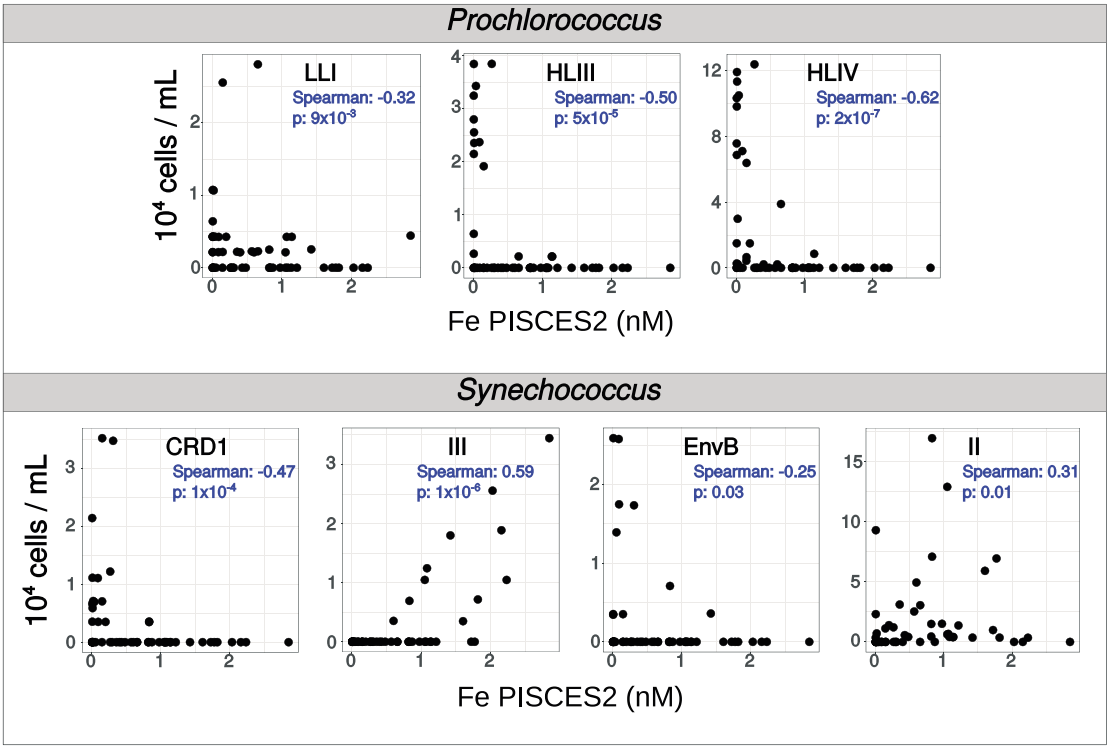


Figure 3. Correlation analysis between absolute cell abundance of marine picocyanobacterial clades and iron concentration estimates from PISCES2 model in surface waters. Only statistically significant correlations are displayed (p value < 0.05). Spearman correlation coefficients and p values are indicated. The cell abundance for each cyanobacterial clade was assessed combining *petB* marker gene counts with flow cytometry determinations using the method published by Vandeputte et al. (2017).

(Figure S1c) is suggestive of a compartmentalization of communities in subcommunities or modules. Our analysis also infers that it is the module as a whole that responds to perturbation, reinforcing the need to dissect plankton responses to iron bioavailability at community scale, while investigating the physiological responses of key species.

In addition to eukaryotes, WGCNA analysis was also performed on prokaryotic communities, as well as on prokaryotic genes from the Ocean Microbial Reference Gene Catalog (Alberti et al., 2017; Sunagawa et al., 2015). Using relative abundances of prokaryotic 16S rDNA miTags, no subnetwork could be associated significantly with iron (maximum $r = 0.19$, $P < 10^{-2}$). However, following the same procedure but using the relative abundances of prokaryotic genes rather than taxa, five subnetworks were significantly associated with iron (ECCO2-DARWIN iron data; Figure 2b, lower panel; Table S1g; $P < 10^{-5}$): Grey60 ($r = 0.38$, $P = 6.\ 10^{-5}$), Plum1 ($r = 0.54$, $P = 3.10^{-9}$), Red ($r = -0.42$, $P = 10^{-5}$), SkyBlue ($r = -0.44$, $P = 2.10^{-6}$), and SaddleBrown ($r = -0.47$, $P = 6.10^{-7}$). VIPs obtained from each of the two models displayed high correlations (Grey60 = 0.99, Plum1 = 0.94, Red = 0.99, SkyBlue = 0.96, SaddleBrown = 0.98). The VIP genes of the SaddleBrown subnetwork represent 25% ($N = 41$) of the total number of genes, and several genes that could be functionally identified encode proteins associated with iron transport, saccharopine dehydrogenase, aminopeptidase N, and ABC-type transporters (Table S1g). The Plum1 subnetwork is a small subnetwork of around 100 genes that is solely associated with iron concentration variability, and 30% of its VIP genes encode principally specialized functions defined as noncore functions in a previous study of the *Tara* Oceans Global Ocean Microbiome (Sunagawa et al., 2015; Table S1g). Not surprisingly, 75% of the genes within this subnetwork encode proteins with unknown functions, although some known functions are linked to iron, such as ferredoxin and regulation of citrate/malate metabolism. The contribution to the global variance by stations located within the Red Sea (Stations TARA_31–34) is particularly high (Figure 2b, lower panel). The Red subnetwork is very large, composed of 3,059 genes. However, only 9%

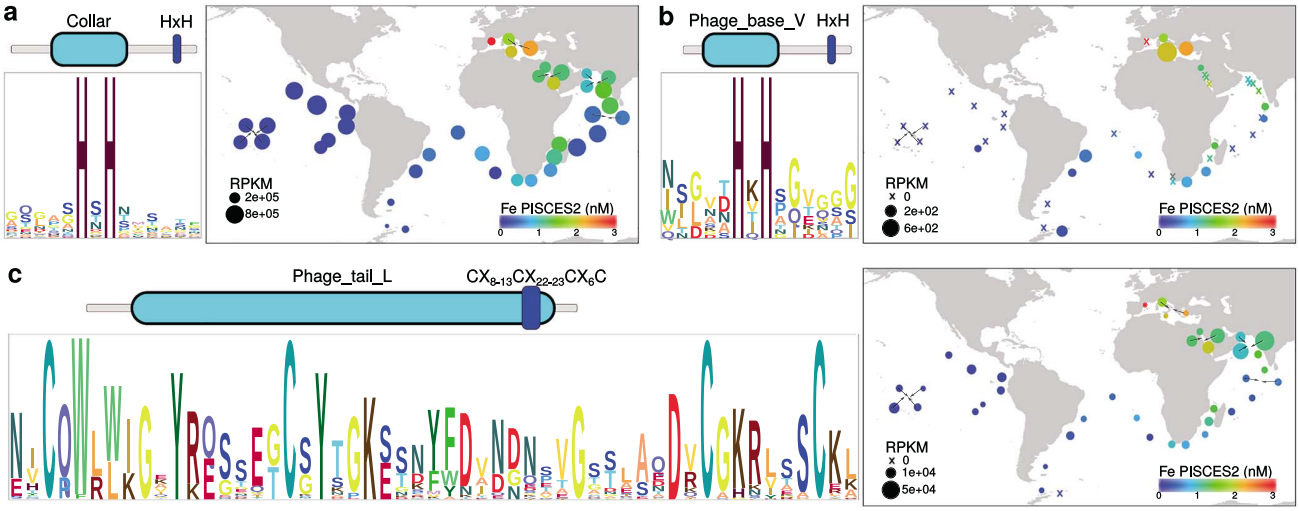


Figure 4. *Tara* Oceans metagenome survey in surface waters for oceanic phages containing putative iron-containing structural proteins. (a) Representation of protein domain architecture of viral tail proteins with putative iron-binding HxH motifs, the HMM logos for the HxH motifs identified in the corresponding *Tara* Oceans viral unigenes, and the biogeographical distribution of the corresponding viral contigs. In the map, the circle colors represent iron concentration estimates at each sampling site according to PISCES2 biogeochemical model (Table S1a), and the circle areas represent the cumulated normalized coverage of the viral contigs of interest. (b) Equivalent analysis for viral spike proteins with putative iron-binding HxH motifs. (c) Equivalent analysis for viral tail tip proteins with $CX_{8-13}CX_{22-23}CX_6C$ motif involved in 4Fe-4S cluster binding.

represent high scoring VIPs, among which functions related to iron are evident (e.g., ABC-type Fe^{3+} siderophore transport system, putative heme iron utilization protein, metalloendopeptidase—Table S1g). Finally, the SkyBlue subnetwork is a small subnetwork (172 genes) containing 33% of VIPs whose functions are generally unknown (Table S1g). The global variance of this gene subnetwork can be correlated principally with several oligotrophic regions of the Pacific Ocean (e.g., Stations TARA_93, 100, 112, and 128).

In summary, association of prokaryotes with iron is detectable at the functional level (gene abundance) but not at the taxonomic level, which would suggest a low level of specialization, at least with the resolution allowed by the 16S marker. To further analyze this aspect, we focused on *Prochlorococcus* and *Synechococcus*, the two most abundant and widespread bacteriophytoplankton in the global ocean, and for which a higher-resolution genetic marker is available. Combining the information from the taxonomic marker *petB*, which encodes cytochrome b_6 (Farrant et al., 2016), with flow cytometry cell counts, we estimated the absolute cell abundance of the picocyanobacterial clades and found that many of them have a strong correlation with predicted iron levels from PISCES2 (Figure 3) and ECCO2-DARWIN models (not shown). *Prochlorococcus* HLIII and IV ecotypes showed the highest anticorrelation with iron, in agreement with previous descriptions that they are the dominant populations in HNLC areas (Rusch et al., 2010; West et al., 2011). *Prochlorococcus* LLI, a minor component in surface waters, also showed anticorrelation with iron. In the case of *Synechococcus*, the strongest positive correlation was found for clade III, whereas a weaker pattern is displayed by clade II. On the contrary, CRD1 showed the highest negative correlation with iron, consistent with it being reported as the major *Synechococcus* clade in HNLC regions (Farrant et al., 2016; Sohm et al., 2016). In addition, clade EnvB also displayed a negative correlation with estimated iron concentrations.

These results demonstrate that iron affects picocyanobacterial community composition and raise the question of whether the lack of correlation with taxonomic networks depends on a poor taxonomic resolution or to being more pronounced for autotrophs with respect to heterotrophs.

Finally, we used relative abundance of viral populations (Brum et al., 2015) to apply WGCNA and tentatively explore whether the viral module subnetworks display any kind of association to the same suite of environmental factors used above for prokaryotes and eukaryotes (data not shown). In spite of the fact that we found four viral IAAs significantly associated with iron using the ECCO2-DARWIN iron estimates (data not

shown), our current knowledge of marine viruses is not advanced enough to discuss our results in the view of the impact on viruses of global iron biogeochemistry. This lack of knowledge is aggravated by the fact that the vast majority of viruses in the IAAs have unknown host ranges.

Viruses are thought to impact oceanic iron during host lysis; however, there is a current discussion about their potential role in complexing iron (Bonnain et al., 2016). To explore this latter point, we surveyed the *Tara* Oceans metagenomes for genes encoding viral structural proteins with putative iron-binding sites. Specifically, we searched for paired histidine residues (H × H motifs) in tail proteins (Bartual et al., 2010) and baseplate assembly proteins (Browning et al., 2012) because this motif has been experimentally implicated in the octahedral coordination of iron. We also analyzed the presence of four conserved cysteine residues involved in the coordination of a 4Fe-4S cluster in tail tip proteins (Tam et al., 2013). Remarkably, these potential iron-binding motifs are present in 87% unigenes encoding viral tail proteins, 47% of baseplate assembly proteins, and 12% in those coding for tip proteins (Figures 4a–4c). The corresponding viral contigs are distributed ubiquitously and with high abundance (Figures 4a–4c), suggesting that a significant fraction of colloidal iron may be associated with viruses in the ocean, a factor that is not currently considered in the modeling of ocean biogeochemistry. The question is then how substantial this contribution could be. Bonnain et al. (2016) made a broad estimation based on the number of iron ions experimentally determined in tails of nonmarine phages, and the amount of tailed viruses typically found in marine surface waters. They thereby suggested that between 6% and 70% of the colloidal iron fraction from surface waters could be bound to tail fibers of phages. In this context, the recent "Ferrojan Horse Hypothesis" posits that iron ions present in phage tails enable phages to exploit their bacterial host's iron-uptake mechanism, where the apparent gift of iron leads to cell lysis (Bonnain et al., 2016). Although our analysis does not allow to confirm this hypothesis, it provides a useful context to explore it further.

3.3. Functional Responses Are Mediated Either by Changes in Gene Copy Number or by Expression Regulation

Given the clear patterns in the community responses to iron availability, we next wondered which molecular patterns were associated with them. We first examined the prevalence of the diatom *ISIP* genes in more detail using both metagenomics and metatranscriptomics data to detect changes in gene abundance and expression, respectively. We found that both the abundance and expression of this gene family displayed a strong negative correlation with iron (Figure 5a). Figure 5a shows a strong hyperbolic profile of *ISIP* gene abundance and mRNA levels with respect to iron concentrations (nonlinear regression fitness of 97.01 and 98.14, respectively; Table S1c). Furthermore, density clustering algorithms detected two types of responses—stations in which *ISIP* was only increased in metagenomics data (denoted group 0) and others in which both metagenomic and metatranscriptomic data showed increases in *ISIP* levels (denoted group 1; Figure 5a). The former likely correspond to locales where *ISIP* copy numbers vary in diatom genomes as a function of iron, implying that the diatoms at these stations display permanent genetic adaptations to the ambient iron concentrations, whereas the latter display transcriptional variation, indicative of more flexible short-term acclimatory rather than permanent adaptive evolutionary processes. Taxonomic analyses revealed that diatoms from the *Thalassiosira* genus were typical of group 0, whereas *Pseudo-nitzschia* was found largely in Group 1 (Figure 5b). Representatives from both these genera are well known to respond to fluctuations in iron (Cohen et al., 2017; see supporting information S1 - Claustre et al., 2008), so these different iron-response strategies may underlie why they are present in different IAAs; *Thalassiosira* is present in the Black and Turquoise IAAs whereas *Pseudo-nitzschia* is only present in the DarkRed module, where it is negatively correlated to iron (Figure 2a; Table S1d).

It is interesting to note that sampling sites can be grouped in a similar way according to either their picocyanobacterial community or diatom *ISIP* patterns in relation to iron levels (Figure 6). HLIV and HLIII codominate the *Prochlorococcus* community in group 1 stations, and these sites are also characterized by the presence of LLI, as well as the *Synechococcus* clades CRD1 and EnvB. Based on picocyanobacteria community composition, these stations tend to cluster together in a group of low-iron stations from Indian and Pacific Oceans (TARA_52, 100, 102, 110, 111, 122, 124, 125, 128, and 137). On the contrary, group 0 ISIP stations were dominated by either *Prochlorococcus* HLI or HLII and by *Synechococcus* clades II or III. Among these stations, those from the high-iron Mediterranean Sea (TARA_7, 9, 18, 23, 25, and 30) clustered together based on picocyanobacteria community composition (Figure 6).

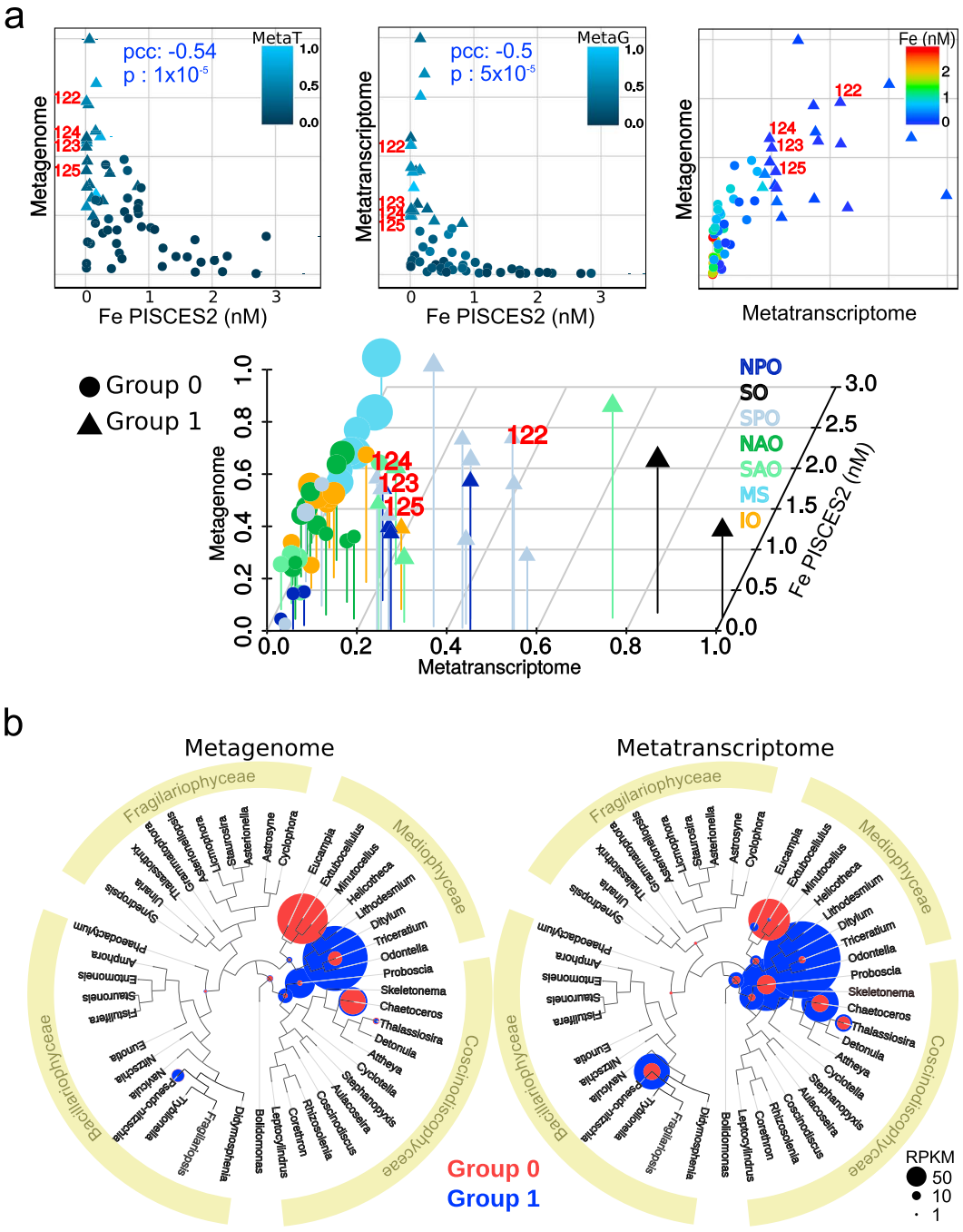


Figure 5. Abundance and expression of diatom *ISIP* genes with respect to iron concentration estimates. (a) 2-D scatter plots correspond to the correlation between gene abundance and iron (left), gene expression and iron (middle), and abundance and expression of *ISIP* genes (right). Pearson correlation coefficients (pcc) and *p* values are indicated in blue. Iron concentrations were estimated using PISCES2 model (Table S1a). In all cases, the abundance and expression of *ISIP* genes were normalized by the total diatom unigene abundance and expression, respectively, and were then scaled to the unit interval. The 3-D plot shown below is derived from the three 2-D scatter plots, with the color gradient representing the third dimension. The data were clustered using density clustering algorithms, resulting in a group of *Tara* Oceans sampling sites in which *ISIP* was only increased in metagenomics data (denoted Group 0 stations [40 stations; circles]) and others in which both metagenomic and metatranscriptomic data showed increases in *ISIP* levels (denoted group 1 stations; 21 stations; triangles). The values corresponding to *Tara* Oceans stations in the Marquesas archipelago are labeled (122–125). *Tara* Oceans sampling sites are colored according to the ocean region in the 3-D plot: NPO = North Pacific Ocean; SO = Southern Ocean; SPO = South Pacific Ocean; NAO = North Atlantic Ocean; SAO = South Atlantic Ocean; MS = Mediterranean Sea; IO = Indian Ocean. (b) Relative abundance (left) and expression (right) of *ISIP* genes assigned at different levels of resolution in a diatom phylogenetic tree. The color code corresponds to the two clusters of stations defined in panel a based on *ISIP* patterns (red for group 0 with variations only at metagenome levels; blue for group 1 with variations in both metagenome and metatranscriptome levels).

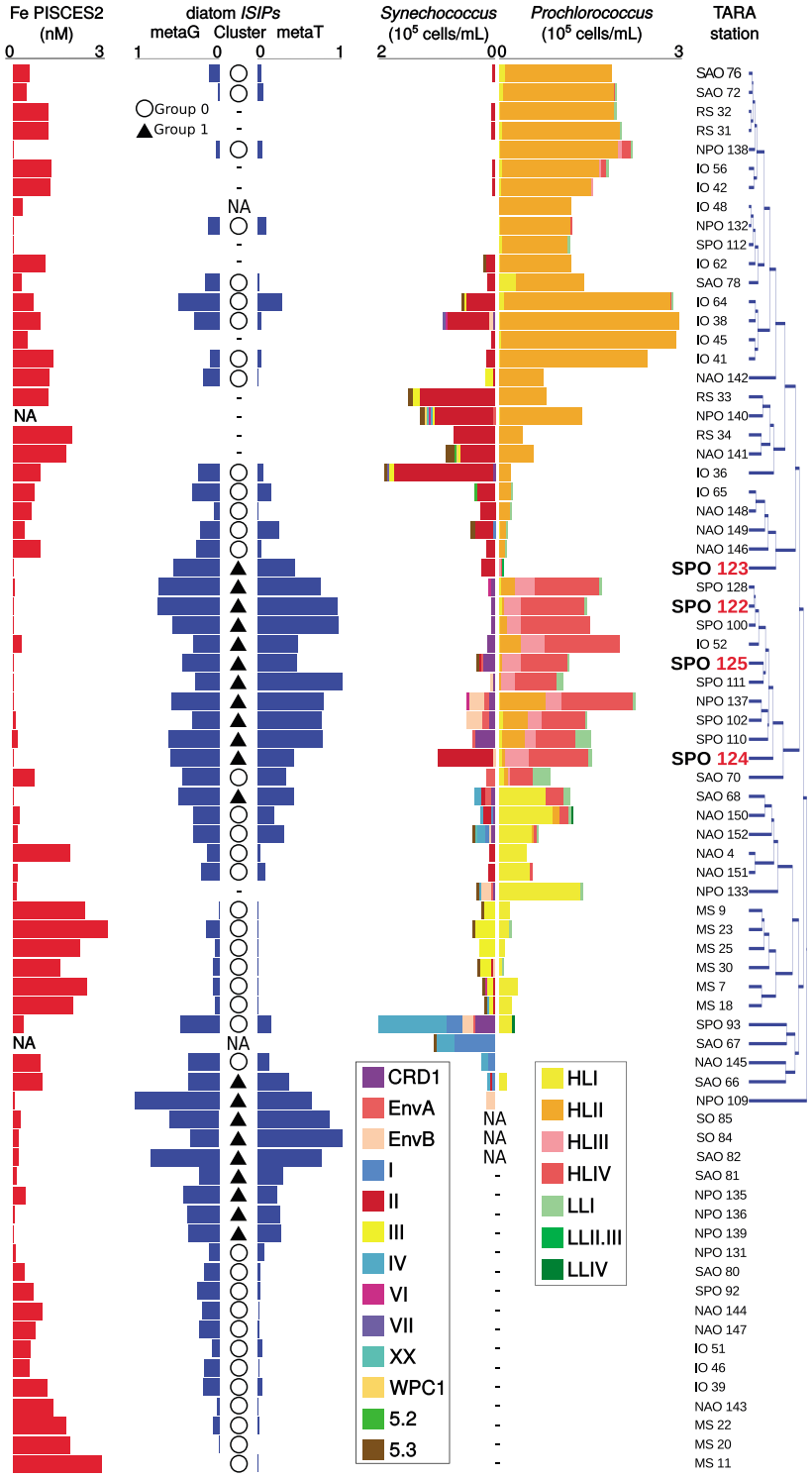


Figure 6. Comparison of iron-driven changes in diatom *ISIP* gene abundance and expression and in the picocyanobacterial community from surface waters. Histograms of cell abundance of *Synechococcus* and *Prochlororococcus* clades at each *Tara* Oceans station are displayed, with stations sorted by hierarchical clustering of a Bray-Curtis distance matrix. The left panels indicate iron concentration estimates from PISCES2 model, and metagenome and metatranscriptome levels of diatom *ISIP* genes, including the resulting cluster type (circles and triangles as described in Figure 5).

Besides diatoms, we carried out a detailed analysis of *ISIP* distributions among other phytoplankton taxa. We found that in chlorophytes Fea1-domain-encoding genes (related to *ISPI2a*; Marchetti et al., 2017) vary in copy number as a function of predicted iron levels and that *ISIP* expression also varies in haptophytes and pelagophytes (Figure S2). Dinoflagellates display the lowest correlations of *ISIP* gene abundance and expression with respect to iron. This may indicate that dinoflagellates respond differently to iron concentrations or with different genes.

A similar analysis was performed to examine the abundance and expression of type I (metal-free) and type II (containing iron or other divalent cations) FBAs (Allen et al., 2012). We found that the *FBAII* gene showed a clear up-regulation at high-iron stations in all groups, while diatoms showed a concomitant reduction in *FBAI* gene abundance and mRNA levels, pelagophytes and dinoflagellates displayed decreased gene abundance, and haptophytes displayed a response at the mRNA level (Figure S2). The chlorophytes displayed no consistent trends.

Ferritin is another important protein of iron metabolism that was relatively recently identified in diatoms (Marchetti et al., 2009) and in other phytoplankton functional groups (Botebol et al., 2015). Although it appears to be involved in long-term iron storage in *Pseudo-nitzschia* (Marchetti et al., 2009), other studies have suggested that its principal role could be in cellular iron buffering and temporal storage over shorter timescales such as during diurnal cycles (Botebol et al., 2015; Cohen et al., 2018; Pfaffen et al., 2015). Our analysis revealed no clear pattern in ferritin gene abundance or expression and estimated iron levels (Figure S2), suggesting that iron storage may not be the main function of ferritin in most eukaryotic marine phytoplankton. The exceptions are haptophytes, in which an iron-driven increase in copy number is observed (Figure S2), and the diatom genus *Pseudo-nitzschia*, in which the biogeographical patterns of ferritin gene abundance and expression suggest a positive correlation with iron (Figure S3).

We additionally examined the levels of genes encoding proteorhodopsin, a light-driven proton pump for the generation of ATP that has been proposed to supplement ATP generation from photosynthesis in iron-limiting conditions, when photosynthetic electron transport is suboptimal (Marchetti et al., 2015). According to our results, abundance of the gene is negatively correlated with iron in pelagophytes and dinoflagellates (as well as in diatoms, albeit without statistical support), and mRNA levels are negatively correlated with iron availability in pelagophytes and haptophytes (Figures S2).

We also examined the interaction between iron and other nutrients in diatoms. Particularly, we focused on silicate metabolism because iron bioavailability has been found to play a role in silicon utilization in these organisms (Durkin et al., 2012, 2016; Mock et al., 2008). The analysis of the different clades of Si transporter (SIT) multigene family support the strong interaction between iron and silicate in diatoms and suggest that the diversification of SITs has led to specialized adaptations to deal with it (see supporting information S1 and Table S1h).

Collectively, our results indicate that individual genes implicated in iron metabolism in specific organismal groups do not provide an unequivocal evaluation of iron availability in the environment and are thus of only limited use as sentinel genes of iron bioavailability. Instead, the integration of all these iron-driven patterns, spanning from genes to ecosystems, is a promising strategy for designing omics-enabled tools that can improve the representation of key nutrients in biogeochemical models. In this sense, the covariation of picocyanobacterial communities with the transcriptional regulation and altered copy numbers of diatom *ISIP* genes can potentially be exploited to predict actual iron bioavailability in the ecosystem (Figure 6). A recent report on the phytoplankton transcriptional response to upwelling (Lampe et al., 2018) highlighted that diatoms express genes involved in nitrogen assimilation, probably to overcome possible autotrophic competitors, thus suggesting that different transcriptional sets of genes may be expressed under different bloom-triggering conditions.

3.4. Plankton Respond to a Resource Burst in the Marquesas Archipelago by Reorganization of IAAs

The global analyses of IAAs and iron responsive genes in the context of the ranges of geographic iron availability provide a first-order approximation of plankton community structure organization and responses for large-scale, iron-linked biogeochemical regions. In other words, they possibly reflect the integrated, albeit diversified, response to average conditions and in a stationary or quasi-stationary phase. They further

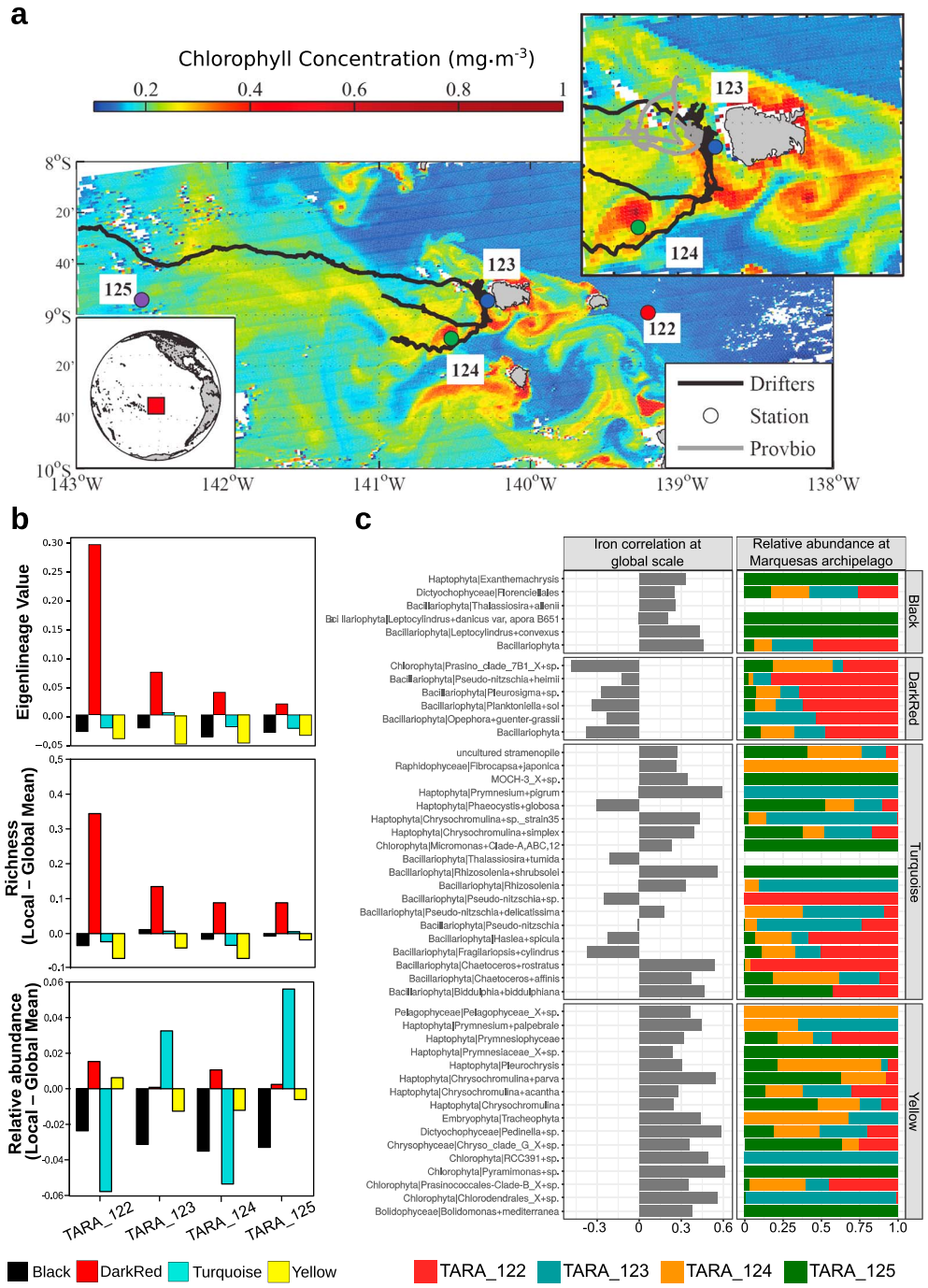


Figure 7. The Marquesas study site, showing sampling sites, surface chlorophyll concentrations, and the local dynamics of the four eukaryotic IAAs. (a) Map of surface chlorophyll in the Marquesas area. Drifter and Provbio trajectories are indicated as well as the *Tara* Oceans sampling stations, with a zoom on stations TARA_123 and TARA_124. For further details see main text and supporting information S2. (b) Analysis of the dynamics of the four IAAs at the Marquesas archipelago stations in relation to their eigenlineage values (upper), richness (middle), and relative abundance (lower). All modules show negative eigenlineage values, with the exception of the DarkRed IAA. The DarkRed module positive eigenlineage scores significantly decrease within the bloom stations. The mean IAA relative abundance calculated over the global *Tara* Oceans data set was subtracted from IAA relative abundance calculated at the Marquesas Islands. The increase in DarkRed relative abundance in station TARA_124 was due to a single Prasinophyceae OTU. The mean IAA richness calculated over the global *Tara* Oceans data set was subtracted from IAA richness calculated at the Marquesas Islands. Data indicates that the DarkRed IAA retains ~60% of its OTUs in low iron conditions, a percentage that decreases in the bloom stations. (c) Relative abundance changes at the Marquesas Islands stations for IAA photosynthetic lineages with high iron correlation. The graph shows the list of IAA autotroph lineages with the highest statistically significant correlations against PISCES2 iron estimates ($p < 0.05$) at a global scale, with the corresponding Pearson correlation coefficient, and their relative abundance at the Marquesas Islands sampling sites.

provide support for the iron products of the two biogeochemical models. We reasoned that they might also be able to indicate increases in iron in regions where biogeochemical models do not have sufficient resolution and to highlight mechanisms in action when the resource is provided in bursts that drive the community out of a previous steady state, for example, leading to blooms. One such case is the Marquesas archipelago in the subtropical Pacific Ocean, where previous studies (Martinez & Maamaatuaiahutapu, 2004) have highlighted a dynamic natural perturbation resulting in perennial plankton blooms that are visible from space. Although iron concentrations have not been measured extensively in the region, these and similar blooms (Gong et al., 2016) are triggered by different processes due to the presence of the islands (vertical mixing, horizontal stirring, local precipitation, and runoff), which are typically coupled to iron injection (Martinez & Maamaatuaiahutapu, 2004), a phenomenon that has been termed Island Mass Effect (Gove et al., 2016). We therefore focused on this region to examine the relationship between the global patterns in plankton subcommunities and iron-responsive gene abundance/expression in a more localized dynamic setting (supporting information S2–S4).

Satellite chlorophyll estimates showed that in the days preceding the visit of the *Tara* Oceans expedition to the archipelago in August 2011, the area was characterized by intense variability. Our analyses also revealed a highly turbulent environment, with mixing up to 100-m depth and strong lateral shearing downstream of the islands, which generated an area of recirculation in the wake of the main island and the formation of small eddies where the blooms were occurring (Figures 7a and S4a). Station TARA_122 sampled the HNLC prebloom waters upstream of the islands (Figure 7a). Waters of Station TARA_122 were characterized by low chlorophyll concentrations in the water column ($[\text{Chl-}a]_{\text{int:}}$ 16.6 mg/m^2) but high concentrations of nutrients ($NO_{2\_}$: 0.12 mmol/m^3, PO_4: 0.57 mmol/m^3, NO_2NO_3: 5.5 mmol/m^3, Si: 2. 2 mmol/m^3; Figure S4b and Table S2a), characteristic of an HNLC region (Quéguiner, 2013; Smetacek & Naqvi, 2008). Of note, the low concentration of silicates in this station may have acted as a limiting factor for the growth of diatoms. Station TARA_123 is coastal, 8-km downstream of Nuku Hiva island and with a seabed depth of 1,903 m and higher chlorophyll levels ($[\text{Chl-}a]_{\text{int}}$ 33.6 mg/m^2), indicative of a bloom. Nutrients were as elevated as in the prebloom HNLC area, and with a particular increase of $NO_{2\_}$ around 150-m depth (1.47 mmol/m^3). Station TARA_124 is away from the coast, 43 km from Nuku Hiva, in even deeper water (2,414-m bottom depth), and in an eddy also characterized by high chlorophyll content with respect to Station TARA_122 ($[\text{Chl-}a]_{\text{int}}$ 28.5 mg/m^2). The chlorophyll patch was possibly seeded near the islands and transported by currents far from the coast but sustained by the eddy dynamics and its interaction with underlying water. Station TARA_125 is located 300-km downstream of the islands. The chlorophyll patch was still clearly evident ($[\text{Chl-}a]_{\text{int}}$ 27.6 mg/m^2). Of note, the large $NO_{2\_}$ reservoir at the base of the mixed layer (120–180 m; Figure S4b) may indicate significant biological activity, although our data are not sufficient to discriminate, which is the relative contribution of phytoplankton, zooplankton, and bacterioplankton to establish the nitrite reservoirs.

The concentration of measured biologically relevant metals was generally reduced in Stations TARA_123 to TARA_125 with respect to HNLC station TARA_122 (Table S2b). The reduction of dissolved ions was particularly significant in the case of cobalt, nickel, copper, and cadmium, which may be considered as a potential clue for an increased uptake of biologically available trace metals in the leeward stations, although other mechanisms cannot be ruled out. Since these metals were not limiting in the HNLC conditions, it is possible that the removal of iron limitation affected the biological pathways related to metal ion uptake in general. For more information on the oceanographic context of the Marquesas Island at the time of sampling, see supporting information S2 (Blain et al., 2008; Dolan et al., 2007; Gómez et al., 2007; Guidi et al., 2008; Legeckis et al., 2004; Masquelier & Vaulot, 2007; Ras et al., 2008; Signorini et al., 1999; Stemmann et al., 2008).

At the four Marquesas sampling sites the IAAs displayed dynamic patterns (Figures 7b and 7c; Table S1f; supporting information S3). The low-iron adapted DarkRed IAA showed a progressive decrease in its prominence leeward of the islands, consistent with its negative correlations to iron at global level, while the Turquoise IAA showed increases in abundance. The Turquoise IAA is the only module containing autotrophs both positively and negatively correlated with iron, and while the latter were prominent at Station TARA_122 the former were prevalent at stations TARA_123–125 (Figure 7c). The observed changes in IAA prevalence in the Marquesas stations therefore supports a role for iron in the modulation of plankton communities in the region. Prokaryote IAAs, although not taxonomy based, are dynamically responsive at

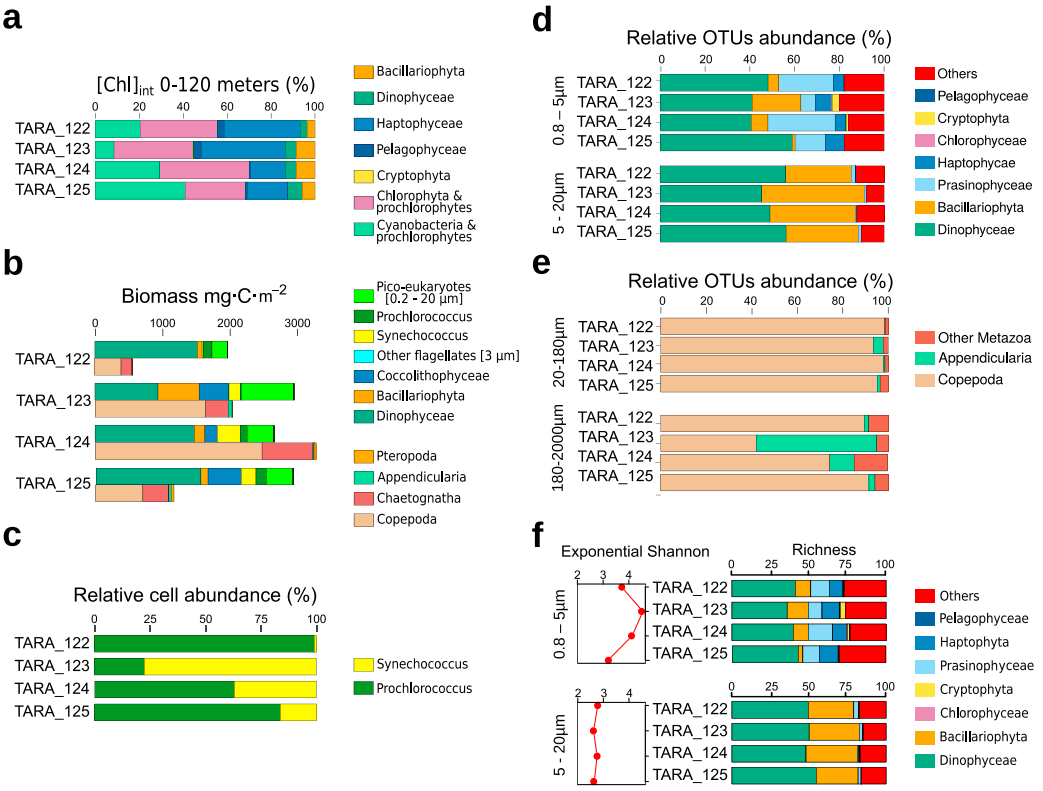


Figure 8. Variations in plankton community composition at Marquesas sampling sites. (a) Relative contribution of different autotrophic lineages to the total chlorophyll concentration in the euphotic zone (0–120 m), derived from photosynthetic pigment analysis and expressed as percent of the total measured chlorophyll. (b) Depth-integrated biomass ($mg{\cdot}C{\cdot}m^{-2}$) of autotrophs and mesozooplankton (>300 µm) in the euphotic zone (0–120 m). (c) Relative abundance of *Prochlorococcus* and *Synechococcus* picocyanobacteria expressed as percent of the total *Prochlorococcus* plus *Synechococcus* abundance estimated from flow cytometry data. Genetic markers (*petB*) showed exactly the same trends (supporting information S3). (d) Relative abundance (%) of ribotypes (18S-V9 tags) assigned to autotrophic eukaryote lineages at the surface (5-m depth). Abundances were computed for the two size fractions containing the majority of autotrophic lineages, namely, 0.8- to 5-µm and 5- to 20-µm size fractions. (e) Relative abundance (%) of ribotypes (18S-V9 tags) assigned to metazoan lineages at the surface (5-m depth). Abundances were computed for the two size fractions containing the majority of metazoans, namely, the 20- to 180-µm and 180- to 2,000-µm size fractions. (f) Richness and diversity (exponential Shannon index) of eukaryotic autotrophs in two different size fractions estimated from the metabarcode data.

the Marquesas Islands (Figure 2b, lower panel). Two types of response can be detected: (a) The prokaryote IAAs Grey60 and Plum1 show a shift from negative to positive eigenlineage scores from TARA_122 to TARA_123, and (b) the SaddleBrown, Red, and SkyBlue IAAs show eigenvalue peaks in Station TARA_123.

The dynamics of the DarkRed subnetwork at the Marquesas Islands may be used to examine our previous claim that it is the module as a whole that responds to perturbation. This is a low-iron-associated module, in which the autotrophic species are the most relevant and typically show negative correlations to iron at the global level, in contrast to most of the autotrophs in the other modules (Figure 7c). Indeed, only a few of the species associated to this subnetwork are present in high-iron conditions. In the context of the Marquesas Islands, most of the DarkRed-assigned OTUs were detected in the oligotrophic Station TARA_122 but not at Station TARA_123, where iron was not expected to be a limiting factor. Both richness and the relative abundance of the subnetwork decreased at Station TARA_123. We thus observe that an iron-responsive subnetwork changes its richness and abundance in a manner consistent with iron availability, rearranging the connectivity between its nodes. Additional details about the dynamics of IAAs in the Marquesas archipelago can be found in supporting information S3.

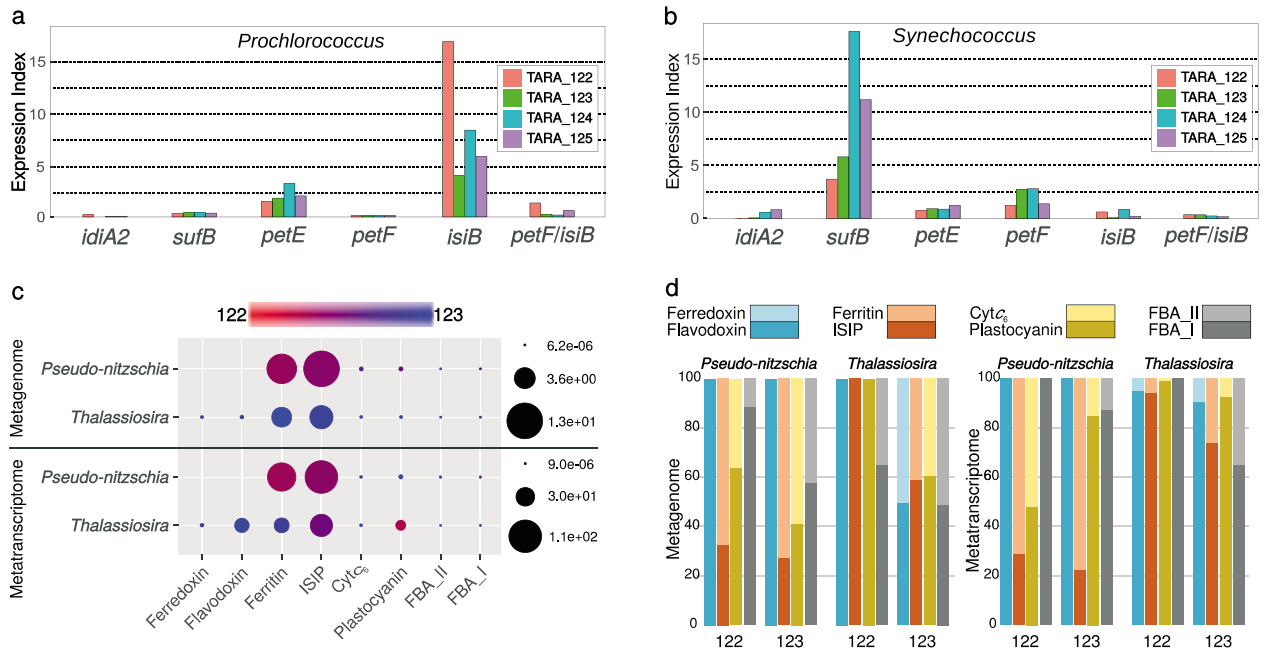


Figure 9. Variations in gene abundance and expression in cyanobacteria and diatoms at Marquesas sampling sites. (a, b) Differential expression patterns of iron-related genes from cyanobacteria *Prochlorococcus* (a) and *Synechococcus* (b) at stations TARA_122–125. Transcription values were normalized over genomic occurrence and are expressed relative to the levels observed at station TARA_122 (index 100). The flavodoxin/ferredoxin ratio is also plotted (PetF/IsiB). (c) Relative abundances and mRNA levels of diatom genes potentially responsive to iron in metagenome and metatranscriptome data sets from stations TARA_122 and 123. Values were normalized by total abundance or expression of all unigenes assigned to the corresponding taxonomic group (*Pseudo-nitzschia* and *Thalassiosira*). For clarity we focused only on changes in 5- to 20-μm size fractions. Colors indicate the contribution of each station to the total levels. (d) Relative ratios between pairs of genes whose presence in the genome or transcriptional activity has been reported previously to be potentially responsive to iron bioavailability. For clarity, *ferritin* levels have been multiplied by a factor of 10 to be comparable with *ISIP* levels, and only 5- to 20-μm size fractions from stations TARA_122 and 123 are compared.

Further analysis of the plankton communities at the Marquesas stations showed that the biomass of primary producers was around 50% higher at the leeward stations (Stations TARA_123, 124 and 125) than at HNLC Station TARA_122, with increases in diatoms, haptophytes, pelagophytes, and *Synechococcus* (Figures 8a–8d; supporting information S3, Alexander et al., 2015). The higher productivity likely fueled increases in zooplankton standing stock at these three stations, in particular copepods, chaetognaths, and appendicularians (Figures 8b and 8e).

Eukaryotic phytoplankton diversity increased at TARA_123 (Figure 8f; supporting information S3, Martin et al., 2013), likely favored by the intense physical dynamics (Barton et al., 2010; Biard et al., 2016). At these stations the increased number of diatoms was due principally to *Thalassiosira* and *Minutocellus* (supporting information S3). Increases in haptophyte and pelagophyte abundance were due to *Phaeocystis* and *Pelagomonas*, respectively. By contrast, the community at Station TARA_122 was more characteristic of an extremely oligotrophic environment, with an abundance of Rhizaria (Biard et al., 2016), *Planktoniella* diatoms (Malviya et al., 2016), *Chrysochromulina* haptophytes (Stibor & Sommer, 2003), and *Pelagococcus* pelagophytes (Guillou et al., 1999), as well as *Prochlorococcus* (Rusch et al., 2010).

Analysis of picocyanobacteria also revealed alterations consistent with increased iron bioavailability in the wake of the islands with respect to TARA_122 (Figure 6). For example, we observed an almost complete shift of *Synechococcus* community composition from clade CRD1 at TARA_122 to clade II at TARA_123 and TARA_124, while absolute abundances of *Prochlorococcus* HLIII and IV, previously shown to dominate in iron-depleted waters (Rusch et al., 2010; West et al., 2011), were significantly reduced (Figure 6; supporting information S3 and S4, Grob et al., 2007).

Using transcriptomes from MMETSP together with metatranscriptomes from *Tara* Oceans (Alberti et al., 2017; Carradec et al., 2018; Louca et al., 2016; Sunagawa et al., 2015), we could further compare the

qualitative shifts in genotypes highlighted above with changes in transcriptional outputs in cyanobacteria (Figures 9a and 9b), eukaryotic phytoplankton (Figures 9c, 9d, and S5), metazoans (Figure S6 and supporting information S4), and more specifically in diatoms (Figures 9c, 9d, and S7). Importantly, *ISIP* levels were decreased in the leeward stations (Figures 5a, 9c, and 9d), and study of gene switches proposed to be responsive to ambient iron concentrations such as ferredoxin/flavodoxin, plastocyanin/cytochrome c_6, and FBAI/FBAII (Allen et al., 2012; Mackey et al., 2015; Marchetti et al., 2012; Peers & Price, 2006; Pierella Karlusich et al., 2015; Thompson et al., 2011) revealed patterns generally consistent with increased bioavailability at Stations TARA_123–125 with respect to HNLC Station TARA_122 both in *Synechococcus* (Figures 9a and 9b) and in the major groups of eukaryotic phytoplankton (Figures 9c, 9d, S5, and S7). The abundance of proteorhodopsin and ferritin genes and mRNA in diatoms were generally also consistent with this hypothesis, with decreases in proteorhodopsin transcripts and increases in ferritin in Station TARA_123 with respect to Station TARA_122 (Figures 9c, 9d, and S7d). These patterns of known iron-responsive genes provide strong support that iron bioavailability is an important driver of the phytoplankton blooms in the Marquesas Islands (supporting information S4, Groussman et al., 2015; Kazamia, et al., 2018; Lane & Morel, 2000; McQuaid et al., 2018; Whitney et al., 2011).

Furthermore, and consistently with the global analyses, *Thalassiosira* and *Pseudo-nitzschia* appear to employ different mechanisms to respond to iron in the Marquesas stations. Specifically, small ferritin-containing *Thalassiosira* cells expressing cytochrome c_6 genes increase in abundance at Station TARA_123, replacing larger Thalassiosirales genetically adapted to low iron at Station TARA_122 by their almost exclusive expression of plastocyanin with respect to cytochrome c_6 (Figures 9c, 9d, and S7; supporting information S4). On the other hand, *Pseudo-nitzschia* cells with flavodoxin and plastocyanin genes are enriched in TARA_122 in comparison with TARA_123. For these two diatom genera, the investigation of the local response around the Marquesas Islands therefore corroborates their behavior within IAAs at the global level, and their compartmentalization into different groups based on *ISIP* gene abundance and expression (Figure 5) supports the hypothesis that they have evolved fundamentally different mechanisms to respond to iron resource availability.

The outcome of the taxon-specific responses summarized above and discussed more comprehensively in supporting information S4, (Arienzo et al., 2014; Berline et al., 2011; Gorsky et al., 1999; Probert et al., 2014; Yuasa et al., 2016) is shifts in abundance and occurrence of taxa within IAAs that change the overall structure of the food web. Our observations also reveal novel information about the genetic strategies and specialized mechanisms employed by each taxon to cope with iron availability (supporting information S4, Bundy & Kille, 2014; De Vos et al., 1992) and illustrate that these responses may ensure resilience of each IAA in a subset of conditions within a highly variable environment. Collectively, our results therefore demonstrate that the delineation of co-responsive subcommunities at global scale can provide a valuable framework for identifying key lineages whose adaptive capacities can be compared and contrasted in specific dynamic contexts. Finally, our in-depth analysis of community structure and gene expression around the Marquesas Islands illustrates how biological data can be used to inform biogeochemical models, because neither of the models used here was able to project increased iron availability in the wake of the islands. Furthermore, while the four Marquesas stations were used in the global analysis that defined the IAAs, they did not contribute to the correlation of IAAs because of the lack of resolution of the models in this area. The module responses in the Marquesas are therefore not biased but are remarkably indicative of a change in iron bioavailability in the lee of the islands.

4. Discussion

In this study we have shown how the turnover of organisms coping with ocean variability involves a combination of ontogenetic responses driven essentially by modulation of gene expression patterns, that is, acclimation, together with phylogenetic responses driven by changes in plankton community structure as well as different genotypes adapted to local conditions by altered copy numbers of iron responsive genes. Different organismal groups appear to use different strategies, meaning that they will not all respond over the same evolutionary timescales. The island mass effect in the wake of the Marquesas Islands leads to the selection of preferred genotypes at the community level and triggers acclimatory responses to fine-tune metabolic functioning via transcriptional responses. These local observations of the most affected organisms are

consistent with IAAs identified in the global ocean, suggesting that large-scale equilibria are in fact dynamic and responsive to smaller scale perturbations.

Previous studies at global scale of the effects of iron on marine plankton were focused on a specific subset of bacterial genes involved in iron metabolism using metagenomics samples from North West Atlantic, Equatorial Pacific, and Indian Oceans (Toulza et al., 2012). Our current study extends this analysis because of its broader geographical coverage and the vastly expanded sequencing data set, which has permitted us to explore both community-level and gene-level responses throughout the entire plankton community, from viruses to zooplankton. Our work thus provides an extensive global scale analysis of the different levels at which plankton biodiversity may be impacted by iron availability, although it should not be assumed that all the responses we highlighted depend solely on iron because one single resource is very unlikely to drive the physiological and structural dynamics of a community. Nonetheless, our extensive statistical analyses suggest that the responses we define do certainly involve iron bioavailability and that the responses occur at molecular, physiological, and compositional levels. Of note is the evidence of modularity in the community structure with modules of co-occurring taxa being sensitive to the resource yet displaying often contrasting strategies. This extends the results obtained by Guidi et al. (2016) who focused on a specific process, indicating that modularity is a general feature of plankton communities, which might be related to their continuous turnover. To the extent allowed by available gene catalogs and taxonomic resolution, we were able to link the subcommunity responses to the molecular toolkits of the organisms, but in many cases we emphasize that the response is not unequivocal but rather maps to a suite of strategies that had already been recorded previously in localized or laboratory experiments.

The complexity of the plankton ecosystem that emerges from the analysis of each IAA and their VIPs, whose dynamics have a certain degree of freedom with respect to the response of the others, indicates that there is some flexibility between the composition of primary producers and their consumers, even though the former are the organisms most directly impacted by nutrient availability. In particular, heterotrophic grazers appear to be central for responses to such bottom-up processes as nutrient acquisition. We interpret the VIP values versus correlation to iron and community centrality as follows: that communities are assemblages of several organisms with multiple interactions among them that cannot be reduced to just a handful of opportunistic autotrophic species able to benefit from nutrient injection and that supply organic carbon to higher trophic levels. Rather, organisms respond to resource availability according to their functional traits but also modulate interactions within their communities, thus affecting their structure. These changes will nonetheless depend on the resident community, immigration from beyond, and changes in the ambient conditions. Some organisms may thrive in different contexts and therefore not be strongly dependent on iron, but rather be good exploiters of primary production stimulated by increased nutrient bioavailability; most of the VIPs are indeed consumers. Furthermore, the relatively low subnetwork centrality of these consumers may suggest that they co-occur with only a subcomponent of the other species. Finally, the nature of the modules composed of parasitic and mixotrophic organisms further suggests that recycling of matter, for example, through remineralization, parasitism, and pathogenesis, are additional strategies within plankton communities to overcome resource limitation. Such strategies would be expected to confer further flexibility and lead to an improved capacity to respond to sporadic bursts of favorable conditions.

Taxonomy-based network analysis for the prokaryotes did not reveal significant associations with iron bioavailability, whereas their gene subnetworks did. In accordance with a recent study based largely on *Tara* Oceans data (Louca et al., 2016), this result advocates for the use of prokaryotic functional signals rather than standard taxonomic criteria to study functional responses of prokaryotes in the global ocean, at least at the level of current taxonomic resolution. In fact, picocyanobacteria displayed a remarkable strain-dependent sensitivity to iron availability. The observations further indicate the need for a better assignation of functional taxonomy, and more studies to better characterize prokaryotic genes of relevance for interpreting the mechanistic changes in prokaryotes following perturbations in iron bioavailability. Furthermore, while standard steady state analyses of ocean systems do not consider biological responses to perturbation per se, our approach of identifying steady state global IAA subnetworks and then investigating their responses to local, short-term perturbation represents a promising new approach.

Comparison of the local response to an inferred iron injection in the Marquesas archipelago with the global patterns indicates that the community response to iron availability cannot be characterized by an even

increase in biomass among existing components but involves a change in their relative weights reflecting their different adaptive solutions and the concurrent reorganization of the subcommunities. In other words, our results infer that the rate of supply of a resource is a factor that modulates the response of organisms and their communities.

Our analysis is based on iron distribution derived from two advanced biogeochemical models rather than from discrete measurements. This is because we considered them to be more representative than the instantaneous in situ measurements whose coverage is also scarce and could not be improved by our expedition, since TARA was not equipped to accurately perform iron concentration assessments. While this may be viewed as a limitation of our work, we provide evidence from independent data of the reliability of these estimates, thus providing a valuable demonstration of the utility of omics data as a tool to validate (and consequently improve) current models of earth system dynamics. The good correspondence between the molecular response and the model simulations demonstrates that metatranscriptomics is now mature enough to provide an independent, biologically based validation of ecosystem models especially when the data are scarce or hard to obtain in a reliable way. The quality and number of iron measurements are continuously improving, but metatranscriptomics may anticipate and suggest the presence of biogeochemical constraints that are still undetectable with analytical methods. In addition, it could significantly integrate the formulation of processes in current ecological models because, on the long term, it can complement the missing information about organism interactions (see above) that cannot be derived from the availability of resources (e.g., Stec et al., 2017).

In conclusion, our study reinforces the results obtained in smaller-scale studies and significantly expands the suite of indicators that can be monitored to detect responses to changes in environmental conditions, from target genes to higher levels of biological organization. Our work paves the way to a suite of possible developments in experimental design and in model formulations that prompt for the improvement of statistical tools to better characterize responses at system level. Numerical simulations of ocean processes aimed at capturing the fluxes of key elements are currently based on just a handful of plankton functional types (Le Quere et al., 2005) or functional genes (Coles et al., 2017). Our results highlight the need to incorporate the response of entire plankton assemblages to more accurately determine responses at different levels, such as gene expression, gene copy numbers, or community composition. To determine the relevance of such processes, omics should become a routine component of ocean observation, and we further demonstrate here that it can contribute to assessing the validity of ecosystem models by complementing biogeochemical measurements in the field and adding critical information about the actual bioavailability of nutrients, which is currently difficult to measure. Finally, the IAAs and other modules described herein provide a framework that is independent of taxonomic or functional groupings to tackle the complexity of natural communities, thus assisting our capacity to predict the responses and resilience of planktonic ecosystems to natural and human-induced perturbations.

Conflict of Interest

The authors declare that no competing interests exist.

Author Contribution

D. I., F. N., F. d'O., and P. T. designed the study with input from *Tara* Oceans coordinators. D. I. directed the project. F. N. directed the field work in the Marquesas archipelago. C. B. wrote the paper with substantial input from M. R. d'A., D. I., L. C., and other first authors. F. d'O., P. T., F. N., E. M., D. I., H. C., L. G., S. S., and F. K. performed oceanographic analyses; A. T. and M. J. F. provided iron concentration data from biogeochemical models; F. R. J. V., G. B., and A. T. compared iron products from different biogeochemical models; E. S., A. Z., S. M., J. V., J. L., S. C., F. V., At. T., and C. B. performed analysis of eukaryotic phytoplankton; M. G. M., J.-L. J., J.-B. R., S. G., L. C., L. S., F. L., and T. B. performed analysis of metazoans and other zooplankton; J. J. P. K., S. C. , H. D., and L. G. performed analysis of cyanobacteria; Q. C., E. P., F. R. J. V., J. J. P. K., E. V., S. G. A., A. A., S. Su., P. B., P. W., A. V., R. S., J. P., G. L.-M., and M. L. performed global omics analyses; F. R. J. V., J. J. P. K., A. K., J. P.-Y., and L. T. performed analysis of omics data from eukaryotic phytoplankton; A. K., E. P., L. C., P. S., and S. dA. performed analysis of omics data from

Acknowledgments
The *Tara* Oceans consortium acknowledges the origin of samples from Stations TARA_113-125 as French Polynesia and that they were collected under Convention number 3534 (Convention relatif à la campagne de prélévements et de mésures de Tara Oceans en Polynesie Francaise) dated 16 June 2011. We thank the commitment of the following people and sponsors who made this singular expedition possible: CNRS (in particular Groupement de Recherche GDR3280, the Mission Pour l'Interdisciplinarité – Project MEGALODOM, and the Fédération de Recherche GO-SEE FR2022), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French Government "Investissements d'Avenir" programs Oceanomics (ANR-11-BTBR-0008), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), and FRANCE GENOMIQUE (ANR-10-INBS-09), Fund for Scientific Research – Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects "PHYTBACK/ANR-2010-1709-01," POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218, SAMOSA/ANR-13-ADAP-0010, CINNAMON/ANR-17-CE02-0014-01), EU FP7 (MicroB3/No. 287589), ERC Advanced Grant Award (Diatomite: 294823), the LouisD foundation of the Institut de France, a Radcliffe Institute Fellowship from Harvard University to C. B., JSPS/MEXT KAKENHI (26430184, 16H06437, and 16KT0020), The Canon Foundation (203143100025), Gordon and Betty Moore Foundation (award #3790) and the US National Science Foundation (awards OCE#1536989 and OCE#1829831) to MBS, agnès b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the EDF Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the Fonds Français pour l'Environnement Mondial, the TARA schooner and its captain and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). This article is contribution number 85 of *Tara* Oceans. The authors have deposited the data in the following repositories: Sequencing data are archived at ENA (http://www.ebi.ac.uk/ena/) under the accession number PRJEB4352 for the metagenomics data and PRJEB6609 for the metatranscriptomics data (Carradec et al., 2018); environmental data are available at PANGAEA (https://www.pangaea.de/).

metazoans and other zooplankton; J. J. P. K., H. D., and L. G. performed gene expression analysis of cyanobacteria; J. J. P. K., J. R. B., S. R., M. B. S., and M. B. performed analysis of viruses; L. B., S. C. A.-S. B, and D. E. performed WGCNA analyses; S. R., F. N., C. D., M. P., S. K. L., S. Se., and S. P collected and managed *Tara* Oceans samples; and L. C., J. J. P. K., M. R. d'A., and C. B. assembled the manuscript. *Tara* Oceans coordinators provided a creative environment and constructive criticism throughout the study. All authors discussed the results and commented on the manuscript.

References

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., et al. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data.*, *4*. https://doi.org/10.1038/sdata.2017.93

Allen, A. E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., et al. (2008). Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(30), 10,438–10,443. https://doi.org/10.1073/pnas.0711370105

Allen, A. E., Moustafa, A., Montsant, A., Eckert, A., Kroth, P. G., & Bowler, C. (2012). Evolution and functional diversification of fructose bisphosphate aldolase genes in photosynthetic marine diatoms. *Molecular Biology and Evolution*, *29*(1), 367–379. https://doi.org/10.1093/molbev/msr223

Aumont, O., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). PISCES-v2: An ocean biogeochemical model for carbon and. *Geoscientific Model Development*, *8*(8), 2465–2513. https://doi.org/10.5194/gmd-8-2465-2015

Barton, A. D., Dutkiewicz, S., Flierl, G., Bragg, J., & Follows, M. J. (2010). Patterns of diversity in marine phytoplankton. *Science*, *327*(5972), 1509–1511. https://doi.org/10.1126/science.1184961

Bartual, S. G., Otero, J. M., Garcia-Doval, C., Llamas-Saiz, A. L., Kahn, R., Fox, G. C., & van Raaij, M. J. (2010). Structure of the bacteriophage T4 long tail fiber receptor-binding tip, *107*(47), 20,287–20,292. *Proceedings of the National Academy of Sciences* https://doi.org/10.1073/pnas.1011218107

Bernardes, J., Zaverucha, G., Vaquero, C., & Carbone, A. (2016). Improvement in protein domain identification is reached by breaking consensus, with the agreement of many profiles and domain co-occurrence. *PLoS Computational Biology*, *12*(7), e1005038–e1005039. https://doi.org/10.1371/journal.pcbi.1005038

Bernardes, J. S., Vieira, F. R. J., Zaverucha, G., & Carbone, A. (2015). A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics*, *32*(3), 345–353. https://doi.org/10.1093/bioinformatics/btv582

Biard, T., Stemmann, L., Picheral, M., Mayot, N., Vandromme, P., Hauss, H., et al. (2016). In situ imaging reveals the biomass of giant protists in the global ocean. *Nature*, *532*(7600), 504–507. https://doi.org/10.1038/nature17652

Bonnain, C., Breitbart, M., & Buck, K. N. (2016). The Ferrojan horse hypothesis: Iron-virus interactions in the ocean. *Frontiers in Marine Science*, *3*(June), 82. https://doi.org/10.3389/fmars.2016.00082

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., & Wincker, P. (2015). Tara Oceans studies plankton at planetary scale. *Science*, *348*(6237), 873. https://doi.org/10.1126/science.aac5605

Botebol, H., Lesuisse, E., Šuták, R., Six, C., Lozano, J.-C., Schatt, P., et al. (2015). Central role for ferritin in the day/night regulation of iron homeostasis in marine phytoplankton. *Proceedings of the National Academy of Sciences*, *112*(47), 14,652–14,657. https://doi.org/10.1073/pnas.1506074112

Boyd, P. W., Jickells, T., Law, C. S., Blain, S., Boyle, E. A., Buesseler, K. O., et al. (2007). Mesoscale iron enrichment experiments 1993-2005: Synthesis and future directions. *Science*, *315*(5812), 612–617. https://doi.org/10.1126/science.1131669

Browning, C., Shneider, M. M., Bowman, V. D., Schwarzer, D., & Leiman, P. G. (2012). Phage pierces the host cell membrane with the iron-loaded spike. *Structure* https://doi.org/10.1016/j.str.2011.12.009, *20*(2), 326–339.

Brum, J. R., Ignacio-espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., & Chaffron, S. (2015). Patterns and ecological drivers of ocean viral communities. *Science*, *348*(6237). https://doi.org/10.1126/science.1261498

Campbell, L., Nolla, H. A., & Vaulot, D. (1994). The importance of Prochlorococcus to community structure in the central North Pacific Ocean. *Limnology and Oceanography*, *39*(4), 954–961. https://doi.org/10.4319/lo.1994.39.4.0954

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., et al. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Chappell, P. D., Whitney, L. P., Wallace, J. R., Darer, A. I., Jean-Charles, S., & Jenkins, B. D. (2015). Genetic indicators of iron limitation in wild populations of Thalassiosira oceanica from the Northeast Pacific Ocean. *The ISME Journal*, *9*(3), 592–602. https://doi.org/10.1038/ismej.2014.171

Chen, X., Wakeham, S. G., & Fisher, N. S. (2011). Influence of iron on fatty acid and sterol composition of marine phytoplankton and copepod consumers. *Limnology and Oceanography*, *56*(2), 716–724. https://doi.org/10.4319/lo.2011.56.2.0716

Cohen, N. R., Brzezinski, M. A., Twining, K. T. B. S., Ellis, K. A., Lampe, R. H., McNair, H., et al. (2017). Diatom transcriptional and physiological responses to changes in iron bioavailability across ocean provinces. *Frontiers in Marine Science*, *4*(November), 1–20. https://doi.org/10.3389/fmars.2017.00360

Cohen, N. R., Mann, E., Stemple, B., Moreno, C. M., Rauschenberg, S., Jacquot, J. E., et al. (2018). Iron storage capacities and associated ferritin gene expression among marine diatoms. *Limnology and Oceanography*, *63*(4), 1677–1691. https://doi.org/10.1002/lno.10800

Coles, V. J., Stukel, M. R., Brooks, M. T., Burd, A., Crump, B. C., Moran, M. A., et al. (2017). Ocean biogeochemistry modeled with emergent trait-based genomics. *Science*, *358*(6367), 1149–1154. https://doi.org/10.1126/science.aan5712

de Vargas, C., Audic, S., Henry, N., Decelle, J., & Mahé, F. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(MAY), 1–12. https://doi.org/10.1007/s13398-014-0173-7.2

Dupont, C. L., Mccrow, J. P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U., Roth, R., et al. (2015). Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *ISME Journal*, *9*(5), 1076–1092. https://doi.org/10.1038/ismej.2014.198

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and amino acids* (Chap. 5). Cambridge, UK: Cambridge University Press.

Durkin, C. A., Marchetti, A., Bender, S. J., Truong, T., Morales, R., Mock, T., & Armbrust, V. E. (2012). Frustule-related gene transcription and the influence of diatom community composition on silica precipitation in an iron-limited environment. *Limnology and Oceanography*, *57*(6), 1619–1633. https://doi.org/10.4319/lo.2012.57.6.1619

Farrant, G. K., Doré, H., Cornejo-Castillo, F. M., Partensky, F., Ratin, M., Ostrowski, M., et al. (2016). Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proceedings of the National Academy of Sciences*, *113*(24), E3365–E3374. https://doi.org/10.1073/pnas.1524865113

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

Follows, M. J., Dutkiewicz, S., Grant, S., & Chisholm, S. W. (2007). Emergent biogeography of microbial. *Science*, (C), 1843–1847. https://doi.org/10.1126/science.1138544

Freeland, H. J., & Cummins, P. F. (2005). Argo: A new tool for environmental monitoring and assessment of the world's oceans, an example from the N.E. Pacific. *Progress in Oceanography*, *64*(1), 31–44. https://doi.org/10.1016/j.pocean.2004.11.002

Gong, W., Browne, J., Hall, N., Schruth, D., Paerl, H., & Marchetti, A. (2016). Molecular insights into a dinoflagellate bloom. *The ISME Journal*, *11*(2), 439–452. https://doi.org/10.1038/ismej.2016.129

Gove, J. M., McManus, M. A., Neuheimer, A. B., Polovina, J. J., Drazen, J. C., Smith, C. R., et al. (2016). Near-island biological hotspots in barren ocean basins. *Nature Communications*, *7*(1), 10,581. https://doi.org/10.1038/ncomms10581

Graff van Creveld, S., Rosenwasser, S., Levin, Y., & Vardi, A. (2016). Chronic iron limitation confers transient resistance to oxidative stress in marine diatoms. *Plant Physiology*. https://doi.org/10.1104/pp.16.00840

Greene, R. M., Geider, R. J., & Falkowksi, P. G. (1991). Effect of iron limitation on photosynthesis in a marine diatom. *Limnology and Oceanography*, *36*(8), 1772–1782. https://doi.org/10.4319/lo.1991.36.8.1772

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*(7600), 465–470. https://doi.org/10.1038/nature16942

Guillou, L., Moon-Van Der Staay, S. Y., Claustre, H., Partensky, F., & Vaulot, D. (1999). Diversity and abundance of Bolidophiceae (Heterokonta) in two oceanic regions. *Applied and Environmental Microbiology*, *65*(10), 4528–4536.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–2021. https://doi.org/10.1093/sysbio/syq010

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Lampe, R. H., Cohen, N. R., Ellis, K. A., Bruland, K. W., Maldonado, M. T., Peterson, T. D., et al. (2018). Divergent gene expression among phytoplankton taxa in response to upwelling. *Environmental Microbiology*, *20*(8), 3069–3082. https://doi.org/10.1111/1462-2920.14361

Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology*, *1*(1), 54. https://doi.org/10.1186/1752-0509-1-54

Le Quere, C., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., et al. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, *11*(11), 2016–2040. https://doi.org/10.1111/j.1365-2486.2005.01004.x

Lommer, M., Roy, A.-S., Schilhabel, M., Schreiber, S., Rosenstiel, P., & LaRoche, J. (2010). Recent transfer of an iron-regulated gene from the plastid to the nuclear genome in an oceanic diatom adapted to chronic iron limitation. *BMC Genomics*, *11*(1), 718. https://doi.org/10.1186/1471-2164-11-718

Lommer, M., Specht, M., Roy, A. S., Kraemer, L., Andreson, R., Gutowska, M. A., et al. (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biology*, *13*(7), R66. https://doi.org/10.1186/gb-2012-13-7-r66

Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science*, *353*(6305), 1272–1277. https://doi.org/10.1126/science.aaf4507

Mackey, K. R. M., Post, A. F., McIlvin, M. R., Cutter, G. A., John, S. G., & Saito, M. A. (2015). Divergent responses of Atlantic coastal and oceanic Synechococcus to iron limitation. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(32), 9944–9949. https://doi.org/10.1073/pnas.1509448112

Mahowald, N. M., Baker, A. R., Bergametti, G., Brooks, N., Duce, R. A., Jickells, T. D., et al. (2005). Atmospheric global dust cycle and iron inputs to the ocean. *Global Biogeochemical Cycles*, *19*, GB4025. https://doi.org/10.1029/2004GB002402

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., et al. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences*, *113*(11), E1516–E1525. https://doi.org/10.1073/pnas.1509523113

Marchetti, A., Catlett, D., Hopkinson, B. M., Ellis, K., & Cassar, N. (2015). Marine diatom proteorhodopsins and their potential role in coping with low iron availability. *ISME Journal*, *9*(12), 2745–2748. https://doi.org/10.1038/ismej.2015.74

Marchetti, A., Juneau, P., Whitney, F. A., Wong, C. S., & Harrison, P. J. (2006). Phytoplankton processes during a mesoscale iron enrichment in the NE subarctic Pacific: Part II-nutrient utilization. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *53*(20–22), 2114–2130. https://doi.org/10.1016/j.dsr2.2006.05.031

Marchetti, A., Moreno, C. M., Cohen, N. R., Oleinikov, I., Twining, B. S., Armbrust, E. V., & Lampe, R. H. (2017). Development of a molecular-based index for assessing iron status in bloom-forming pennate diatoms. *Journal of Phycology*, *53*(4), 820–832. https://doi.org/10.1111/jpy.12539

Marchetti, A., Parker, M. S., Moccia, L. P., Lin, E. O., Arrieta, A. L., Ribalet, F., et al. (2009). Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature*, *457*(7228), 467–470. https://doi.org/10.1038/nature07539

Marchetti, a., Schruth, D. M., Durkin, C. a., Parker, M. S., Kodner, R. B., Berthiaume, C. T., et al. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*, *109*(6), E317–E325. https://doi.org/10.1073/pnas.1118408109

Martinez, E., & Maamaatuaiahutapu, K. (2004). Island mass effect in the Marquesas Islands: Time variation. *Geophysical Research Letters*, *31*, L18307. https://doi.org/10.1029/2004GL020682

Mawji, E., Schlitzer, R., Dodas, E. M., Abadie, C., Abouchami, W., Anderson, R. F., et al. (2014). The GEOTRACES intermediate data product 2014. *Marine Chemistry*, *177*, 1–8. https://doi.org/10.1016/j.marchem.2015.04.005

Menemenlis, B. D., Campin, J., Heimbach, P., Hill, C., & Lee, T. (2008). ECCO2: High resolution global ocean and sea ice data synthesis. *Mercator Ocean Quarterly Newsletter*, *31*(October), 13–21.

Mevik, B.-H., & Wehrens, R. (2007). The pls package: Principle component and partial least squares regression in R. *Journal of Statistical Software*, *1*(1), 128–129. https://doi.org/10.1002/wics.10

Mock, T., Samanta, M. P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., et al. (2008). Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences*, *105*(5), 1579–1584. https://doi.org/10.1073/pnas.0707946105

Morrissey, J., Sutak, R., Paz-Yepes, J., Tanaka, A., Moustafa, A., Veluchamy, A., et al. (2015). A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake. *Current Biology*, *25*(3), 364–371. https://doi.org/10.1016/j.cub.2014.12.004

Peers, G., & Price, N. M. N. (2006). Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature*, *441*(7091), 341–344. https://doi.org/10.1038/nature04630

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., et al., & Tara Oceans Consortium Coordinators (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, *2*. https://doi.org/10.1038/sdata.2015.23

Pfaffen, S., Bradley, J. M., Abdulqadir, R., Firme, M. R., Moore, G. R., Brun, N. E. L., & Murphy, M. E. P. (2015). A diatom ferritin optimized for iron oxidation but not iron storage. *Journal of Biological Chemistry*, *290*(47), 28,416–28,427. https://doi.org/10.1074/jbc.M115.669713

Pierella Karlusich, J. J., Ceccoli, R. D., Graña, M., Romero, H., & Carrillo, N. (2015). Environmental selection pressures related to iron utilization are involved in the loss of the flavodoxin gene from the plant genome. *Genome Biology and Evolution*, *7*(3), 750–767. https://doi.org/10.1093/gbe/evv031

Quéguiner, B. (2013). Iron fertilization and the structure of planktonic communities in high nutrient regions of the Southern Ocean. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *90*, 43–54. https://doi.org/10.1016/j.dsr2.2012.07.024

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, *537*(7622), 689–693. https://doi.org/10.1038/nature19366

Rusch, D. B., Martiny, A. C., Dupont, C. L., Halpern, A. L., & Venter, J. C. (2010). Characterization of Prochlorococcus clades from iron-depleted oceanic regions. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16,184–16,189. https://doi.org/10.1073/pnas.1009513107

Scelfo, G. M. (1997). A practical handbook for trace metals and ancillary analyses. University of California Environmental Toxicology-Special Publication. *Physics Research Section B*, *189*, 196–201.

Smetacek, V., & Naqvi, S. W. A. (2008). The next generation of iron fertilization experiments in the Southern Ocean. *Philosophical Transactions of the Royal Society A*, *366*(1882), 3947–3967. https://doi.org/10.1098/rsta.2008.0144

Sohm, J. A., Ahlgren, N. A., Thomson, Z. J., Williams, C., Moffett, J. W., Saito, M. A., et al. (2016). Co-occurring Synechococcus ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME Journal*, *10*(2), 333–345. https://doi.org/10.1038/ismej.2015.115

Stec, K. F., Caputi, L., Buttigieg, P. L., D'Alelio, D., Ibarbalz, F. M., Sullivan, M. B., et al. (2017). Modelling plankton ecosystems in the meta-omics era. Are we ready? *Marine Genomics*, *32*, 1–17. https://doi.org/10.1016/j.margen.2017.02.006

Stibor, H., & Sommer, U. (2003). Mixotrophy of a photosynthetic flagellate viewed from an optimal foraging perspective. *Protist*, *154*(1), 91–98. https://doi.org/10.1078/143446103764928512

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237). https://doi.org/10.1126/science.1261359

Tagliabue, A., Aumont, O., DeAth, R., Dunne, J. P., Dutkiewicz, S., Galbraith, E., et al. (2016). How well do global ocean biogeochemistry models simulate dissolved iron distributions? *Global Biogeochemical Cycles*, *30*, 149–174. https://doi.org/10.1002/2015GB005289

Tagliabue, A., Bowie, A. R., Philip, W., Buck, K. N., Johnson, K. S., & Saito, M. A. (2017). The integral role of iron in ocean biogeochemistry. *Nature*, *543*(7643), 51–59. https://doi.org/10.1038/nature21058

Tagliabue, A., Mtshali, T., Aumont, O., Bowie, A. R., Klunder, M. B., Roychoudhury, A. N., & Swart, S. (2012). A global compilation of dissolved iron measurements: Focus on distributions and processes in the Southern Ocean. *Biogeosciences*, *9*(6), 2333–2349. https://doi.org/10.5194/bg-9-2333-2012

Tam, W., Pell, L. G., Bona, D., Tsai, A., Dai, X. X., Edwards, A. M., et al. (2013). Tail tip proteins related to bacteriophage λ gpL coordinate an iron-sulfur cluster. *Journal of Molecular Biology*, *425*(14), 2450–2462. https://doi.org/10.1016/j.jmb.2013.03.032

Testor, P., Meyers, G., Pattiaratchi, C., Bachmayer, R., Hayes, D., Pouliquen, S., et al. (2010). Gliders as a component of future observing systems. In J. Hall, D. E. Harrison, & D. Stammer (Eds.), *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 2), Venice, Italy, 21-25 September 2009*. ESA Publication WPP-306. https://doi.org/10.5270/OceanObs09.cwp.89

Thompson, A. W., Huang, K., Saito, M. A., & Chisholm, S. W. (2011). Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability. *The ISME Journal*, *5*(10), 1580–1594. https://doi.org/10.1038/ismej.2011.49

Toulza, E., Tagliabue, A., Blain, S., & Piganeau, G. (2012). Analysis of the global ocean sampling (GOS) project for trends in iron uptake by surface ocean microbes. *PLoS One*, *7*(2), e30931. https://doi.org/10.1371/journal.pone.0030931

Uitz, J., Claustre, H., Morel, A., & Hooker, S. B. (2006). Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research*, *111*, C08005. https://doi.org/10.1029/2005JC003207

Vandeputte, D., Kathagen, G., D'Hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature*. https://doi.org/10.1038/nature24460

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., et al. (2015). Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science*, *348*(6237). https://doi.org/10.1126/science.1261447

West, N. J., Lebaron, P., Strutton, P. G., & Suzuki, M. T. (2011). A novel clade of Prochlorococcus found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean. *The ISME Journal*, *5*(6), 933–944. https://doi.org/10.1038/ismej.2010.186

Xing, X., Morel, A., Claustre, H., D'Ortenzio, F., & Poteau, A. (2012). Combined processing and mutual interpretation of radiometry and fluorometry from autonomous profiling Bio-Argo floats: 2. Colored dissolved organic matter absorption retrieval. *Journal of Geophysical Research*, *117*, C04022. https://doi.org/10.1029/2011JC007632

Supporting Information References

Alexander, H., Rouco, M., Haley, S., Wilson, S., Karl, D., & Dyhrman, S. (2015). Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proceedings of the National Academy of Sciences*, *112*(44), E5,972–E5,979. https://doi.org/10.1073/pnas.1518165112

Arienzo, M., Toscano, F., Di Fraia, M., Caputi, L., Sordino, P., Guida, M., et al. (2014). An assessment of contamination of the Fusaro Lagoon (Campania Province, southern Italy) by trace metals. *Environmental Monitoring and Assessment*, *186*(9), 5731–5747. https://doi.org/10.1007/s10661-014-3816-4

Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F., & Covarrubias, A. A. (2008). The enigmatic LEA proteins and other hydrophilins. *Plant Physiology*, *148*(1), 6–24. https://doi.org/10.1104/pp.108.120725

Berline, L., Stemmann, L., Vichi, M., Lombard, F., & Gorsky, G. (2011). Impact of appendicularians on detritus and export fluxes: A model approach at DyFAMed site. *Journal of Plankton Research*, *33*(6), 855–872. https://doi.org/10.1093/plankt/fbq163

Blain, S., Bonnet, S., & Guieu, C. (2008). Dissolved iron distribution in the tropical and sub tropical South Eastern Pacific. *Biogeosciences*, *4*(4), 2845–2875. https://doi.org/10.5194/bgd-4-2845-2007

Bundy, J. G., & Kille, P. (2014). Metabolites and metals in metazoa—What role do phytochelatins play in animals? *Metallomics: Integrated Biometal Science*, *6*(9), 1576–1582. https://doi.org/10.1039/c4mt00078a

Claustre, H., Sciandra, A., & Vaulot, D. (2008). Introduction to the special section bio-optical and biogeochemical conditions in the South East Pacific in late 2004: The BIOSOPE program. *Biogeosciences*, *5*(3), 679–691. https://doi.org/10.5194/bg-5-679-2008

Crans, D. C., Smee, J. J., Gaidamauskas, E., & Yang, L. (2004). The chemistry and biochemistry of vanadium and the biological activities exerted by vanadium compounds. *Chemical Reviews*, *104*(2), 849–902. https://doi.org/10.1021/cr020607t

Dandonneau, Y., & Charpy, L. (1985). An empirical approach to the island mass effect in the south tropical Pacific based on sea surface chlorophyll concentrations. *Deep Sea Research Part A, Oceanographic Research Papers*, *32*(6), 707–721. https://doi.org/10.1016/0198-0149(85)90074-3

De Vos, C. H., Vonk, M. J., Vooijs, R., & Schat, H. (1992). Glutathione depletion due to copper-induced phytochelatin synthesis causes oxidative stress in Silene cucubalus. *Plant Physiology*, *98*(3), 853–858. https://doi.org/10.1104/pp.98.3.853

Dolan, J. R., Ritchie, M. E., & Ras, J. (2007). The "neutral" community structure of planktonic herbivores, tintinnid ciliates of the microzooplankton, across the SE Tropical Pacific Ocean. *Biogeosciences Discussions*, *4*(1), 561–593. https://doi.org/10.5194/bgd-4-561-2007

Durkin, C. A., Koester, J. A., Bender, S. J., Armbrust, E. V., & Kroth, P. (2016). The evolution of silicon transporters in diatoms. *Journal of Phycology*, *52*(5), 716–731. https://doi.org/10.1111/jpy.12441

Gómez, F., Claustre, H., Raimbault, P., & Souissi, S. (2007). Two high-nutrient low-chlorophyll phytoplankton assemblages: The tropical central Pacific and the offshore Perú-Chile Current. *Biogeosciences Discussions*, *4*(3), 1535–1554. https://doi.org/10.5194/bgd-4-1535-2007

Gorsky, G., Chrétiennot-Dinet, M. J., Blanchot, J., & Palazzoli, I. (1999). Picoplankton and nanoplankton aggregation by appendicularians: Fecal pellet contents of Megalocercus huxleyi in the equatorial Pacific. *Journal of Geophysical Research*, *104*, 3381–3390. https://doi.org/10.1029/98jc01850

Grob, C., Ulloa, O., Claustre, H., Huot, Y., Alarcón, G., & Marie, D. (2007). Contribution of picoplankton to the total particulate organic carbon concentration in the eastern South Pacific. *Biogeosciences*, *4*(5), 837–852. https://doi.org/10.5194/bg-4-837-2007

Groussman, R. D., Parker, M. S., & Armbrust, E. V. (2015). Diversity and evolutionary history of iron metabolism genes in diatoms. *PLoS One*, *10*(6). https://doi.org/10.1371/journal.pone.0129081

Guidi, L., Gorsky, G., Claustre, H., Miquel, J. C., Picheral, M., & Stemmann, L. (2008). Distribution and fluxes of aggregates >100 μm in the upper kilometer of the South-Eastern Pacific. *Biogeosciences*, *5*(5), 1361–1372. https://doi.org/10.5194/bg-5-1361-2008

Kazamia, E., Sutak, R., Paz-Yepes, J., Dorrell, R. G., Vieira, F. R. J., Mach, J., et al. (2018). Endocytosis-mediated siderophore uptake as a strategy for Fe acquisition in diatoms. *Science Advances*, *4*(5), eaar4536. https://doi.org/10.1126/sciadv.aar4536

Khoshnood, B., Dacklin, I., & Grabbe, C. (2016). Urm1: An essential regulator of JNK signaling and oxidative stress in Drosophila melanogaster. *Cellular and Molecular Life Sciences*, *73*(9), 1939–1954. https://doi.org/10.1007/s00018-015-2121-x

Lane, T. W., & Morel, F. M. M. (2000). A biological function for cadmium in marine diatoms. *Proceedings of the National Academy of Sciences*, *97*(9), 4627–4631. https://doi.org/10.1073/pnas.090091397

Legeckis, R., Brown, C. W., Bonjean, F., & Johnson, E. S. (2004). The influence of tropical instability waves on phytoplankton blooms in the wake of the Marquesas Islands during 1998 and on the currents observed during the drift of the Kon-Tiki in 1947. *Geophysical Research Letters*, *31*, L23307. https://doi.org/10.1029/2004GL021637

Martin, P., van der Loeff, M. R., Cassar, N., Vandromme, P., D'Ovidio, F., Stemmann, L., et al. (2013). Iron fertilization enhanced net community production but not downward particle flux during the Southern Ocean iron fertilization experiment LOHAFEX. *Global Biogeochemical Cycles*, *27*, 871–881. https://doi.org/10.1002/gbc.20077

Masquelier, S., & Vaulot, D. (2007). Distribution of micro-organisms along a transect in the South-East Pacific Ocean (BIOSOPE cruise) from epifluorescence microscopy. *Biogeosciences Discussions*, *4*(4), 2667–2697. https://doi.org/10.5194/bgd-4-2667-2007

McQuaid, J. B., Kustka, A. B., Oborník, M., Horák, A., McCrow, J. P., Karas, B. J., et al. (2018). Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. *Nature*, *555*(7697), 534–537. https://doi.org/10.1038/nature25982

Probert, I., Siano, R., Poirier, C., Decelle, J., Biard, T., Tuji, A., et al. (2014). Brandtodinium gen. nov. and B.nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *Journal of Phycology*, *50*(2), 388–399. https://doi.org/10.1111/jpy.12174

Ras, J., Claustre, H., & Uitz, J. (2008). Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: Comparison between in situ and predicted data. *Biogeosciences*, *4*(5), 3409–3451. https://doi.org/10.5194/bgd-4-3409-2007

Signorini, S. R., McClain, C. R., & Dandonneau, Y. (1999). Mixing and phytoplankton bloom in the wake of the Marquesas Islands. *Geophysical Research Letters*, *26*, 3121–3124. https://doi.org/10.1029/1999GL010470

Stemmann, L., Eloire, D., Sciandra, A., Jackson, G., Guidi, L., Picheral, M., & Gorsky, G. (2008). Volume distribution for particles between 3.5 to 2000 μm in the upper 200 m region of the South Pacific Gyre. *Biogeosciences Discussions*, *4*(5), 3377–3407. https://doi.org/10.5194/bgd-4-3377-2007

Whitney, L. A. P., Lins, J. J., Hughes, M. P., Wells, M. L., Dreux Chappell, P., & Jenkins, B. D. (2011). Characterization of putative iron responsive genes as species-specific indicators of iron stress in thalassiosiroid diatoms. *Frontiers in Microbiology*, *2*(NOV), 1–14. https://doi.org/10.3389/fmicb.2011.00234

Yuasa, T., Horiguchi, T., Mayama, S., & Takahashi, O. (2016). Gymnoxanthella radiolariae gen. et sp. nov. (Dinophyceae), a dinoflagellate symbiont from solitary polycystine radiolarians. *Journal of Phycology*, *52*(1), 89–104. https://doi.org/10.1111/jpy.12371

Appendix C

Article 6 / Co-authored manuscript 3: Chust et al. 2017

This article is a mini-review discussing future challenges for plankton diversity and macroecology research: (1) What can we learn about plankton communities from the new wealth of high-throughput "omics" data? (2) What is the link between plankton diversity and ecosystem function? (3) How can species distribution models be adapted to represent plankton biogeography? (4) How will plankton biogeography be altered due to anthropogenic climate change? and (5) Can a new unifying theory of macroecology be developed based on plankton ecology studies?

I contributed to this review as part of a workshop during my Master 2 (PlankDiv, organized by Dr Sakina Dorothée Ayata in Villefranche-sur-mer, http://plankdiv.obs-vlfr.fr/). For this co-authored manuscript, I took part in the group discussions during the workshop, and I helped to review the manuscript.

frontiers in Marine Science

MINI REVIEW
published: 10 March 2017
doi: 10.3389/fmars.2017.00068

Mare Incognitum: A Glimpse into Future Plankton Diversity and Ecology Research

Guillem Chust[1], Meike Vogt[2], Fabio Benedetti[3], Teofil Nakov[4], Sébastien Villéger[5], Anaïs Aubert[3,6], Sergio M. Vallina[7], Damiano Righetti[2], Fabrice Not[8], Tristan Biard[3,8], Lucie Bittner[9], Anne-Sophie Benoiston[9], Lionel Guidi[3], Ernesto Villarino[1], Charlie Gaborit[7], Astrid Cornils[10], Lucie Buttay[11], Jean-Olivier Irisson[3], Marlène Chiarello[5], Alessandra L. Vallim[12,13], Leocadio Blanco-Bercial[14], Laura Basconi[15] and Sakina-Dorothée Ayata[3]*

[1] Marine Research Division, AZTI, Sukarrieta, Spain, [2] Environmental Physics Group, Institute for Biogeochemistry and Pollutant Dynamics, ETH Zurich, Zurich, Switzerland, [3] Laboratoire d'Océanographie de Villefranche, Centre National de la Recherche Scientifique, Sorbonne Universités, UPMC Université Paris 06, Villefranche-sur-Mer, France, [4] Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, AR, USA, [5] Laboratoire Biodiversité Marine et ses Usages (MARBEC), UMR 9190 Centre National de la Recherche Scientifique-IRD-UM-IFREMER, Université de Montpellier, Montpellier, France, [6] Service des Stations Marines du Muséum National d'Histoire Naturelle, CRESCO, Dinard, France, [7] Institute of Marine Sciences (CSIC), Barcelona, Spain, [8] Laboratoire Adaptation et Diversité en Milieu Marin UMR7144, Station Biologique de Roscoff, Centre National de la Recherche Scientifique, Sorbonne Universités, UPMC Université Paris 06, Roscoff, France, [9] Centre National de la Recherche Scientifique, Institut de Biologie Paris-Seine, Evolution Paris Seine, Sorbonne Universités, UPMC Université Paris 06, Paris, France, [10] Polar Biological Oceanography, Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany, [11] Centro Oceanográfico de Gijón, Instituto Español de Oceanografía, Gijón, Spain, [12] Instituto de Biociências, Universidade Estadual Júlio de Mesquita Filho, São Vicente, Brazil, [13] Laboratório de Evolução e Diversidade Aquática, Universidade Estadual Júlio de Mesquita Filho, Assis, Brazil, [14] Bermuda Institute of Ocean Sciences, St. George's, Bermuda, [15] Universita' del Salento, CONISMA, Lecce, Italy*

OPEN ACCESS

Edited by:
Cosimo Solidoro, National Institute of Oceanography and Experimental Geophysics, Italy

Reviewed by:
Jan Marcin Weslawski, Institute of Oceanology (PAN), Poland
Jose M. Riascos, Universidad del Valle, Colombia
Maurizio Ribera D'Alcala', Stazione Zoologica Anton Dohrn, Italy

***Correspondence:**
Guillem Chust
gchust@azti.es

Specialty section:
This article was submitted to Marine Ecosystem Ecology, a section of the journal Frontiers in Marine Science

Received: *30 September 2016*
Accepted: *24 February 2017*
Published: *10 March 2017*

Citation:
Chust G, Vogt M, Benedetti F, Nakov T, Villéger S, Aubert A, Vallina SM, Righetti D, Not F, Biard T, Bittner L, Benoiston A-S, Guidi L, Villarino E, Gaborit C, Cornils A, Buttay L, Irisson J-O, Chiarello M, Vallim AL, Blanco-Bercial L, Basconi L and Ayata S-D (2017) Mare Incognitum: A Glimpse into Future Plankton Diversity and Ecology Research. Front. Mar. Sci. 4:68. doi: 10.3389/fmars.2017.00068

With global climate change altering marine ecosystems, research on plankton ecology is likely to navigate uncharted seas. Yet, a staggering wealth of new plankton observations, integrated with recent advances in marine ecosystem modeling, may shed light on marine ecosystem structure and functioning. A EuroMarine foresight workshop on the "*Impact of climate change on the distribution of plankton functional and phylogenetic diversity*" (PlankDiv) identified five grand challenges for future plankton diversity and macroecology research: (1) What can we learn about plankton communities from the new wealth of high-throughput "omics" data? (2) What is the link between plankton diversity and ecosystem function? (3) How can species distribution models be adapted to represent plankton biogeography? (4) How will plankton biogeography be altered due to anthropogenic climate change? and (5) Can a new unifying theory of macroecology be developed based on plankton ecology studies? In this review, we discuss potential future avenues to address these questions, and challenges that need to be tackled along the way.

Keywords: plankton, macroecology, species distribution, functional diversity, climate change, habitat modeling

INTRODUCTION

Marine ecosystems are altered by anthropogenic climate change and ocean acidification at an unprecedented rate (Waters et al., 2016). In recent years, observational studies have documented shifts in plankton biogeography and community structure in several ocean basins associated to sea warming, with changes that rank among the fastest and largest documented (Beaugrand et al., 2002; Poloczanska et al., 2013; Rivero-Calle et al., 2015). How changes in plankton distribution, phenology, and biomass may impact fisheries and other ecosystem services is poorly quantified (Cheung et al., 2013), with large uncertainties in the magnitude of potential cascading effects caused by trophic mismatch (Edwards and Richardson, 2004), trophic amplification (Chust et al., 2014a), and on global biogeochemical cycles (Doney et al., 2012). In consequence, current management policies suffer from a lack of understanding of marine systems (Borja et al., 2010), and biases arise in the perception of potential ocean calamities in the absence of robust evidence (Duarte et al., 2015).

While recent oceanographic efforts such as Tara Oceans (Pesant et al., 2015) and Malaspina (Duarte, 2015) expeditions have generated a staggering wealth of novel observational data on plankton distribution and diversity (**Figure 1**), these same data have revealed the extent of our ignorance of marine ecosystem structure and function. A large fraction of plankton diversity recorded in recent surveys cannot be assigned to known taxonomic groups (de Vargas et al., 2015), highlighting how profoundly our knowledge of the planktonic world is biased toward the taxa sampled or cultured. Not only the identity of major players, but also the drivers of community structure and interactions between organisms remain a "*mare incognitum*." In the surface ocean, plankton composed of prokaryotes (viruses, bacteria, and archaea) and eukaryotes (protists and metazoans; **Figure 1**) have been shown to form complex interaction networks driven by multiple biotic and abiotic factors (Lima-Mendez et al., 2015), and despite their key role as resource for higher trophic levels, mesopelagic plankton communities are some of the least studied on Earth (St. John et al., 2016).

Despite these gaps in our understanding, the existing data reveal the importance of community composition for marine ecosystem function. For instance, an investigation of planktonic communities at the global scale using high-throughput metagenomic sampling techniques has recently linked carbon export patterns to specific plankton interaction networks (Guidi et al., 2016), suggesting that the who's who in the plankton world is of paramount importance for the carbon cycle. Integrated with revised estimates in species abundance and biomass (Buitenhuis et al., 2013), and combined with advances in statistical (Robinson et al., 2011) and mechanistic modeling techniques (Follows et al., 2007), novel high-throughput metagenomic data may allow us to relate biogeographic patterns of plankton distribution and diversity to further ecosystem processes.

Marine plankton ecology research is thus at a crossroads: At a time where marine ecosystems reveal their nature for the first time, these transient ecosystems have already adapted to environmental changes and are continuing to do so (Waters et al., 2016), with unknown consequences for ecosystem function, and ecosystem service provision. In this context, a close collaboration between researchers belonging to various fields of plankton ecology appears timely to identify the most pressing questions, and to accelerate progress in our understanding of marine ecosystem structure and function. Recently, a EuroMarine foresight workshop on the "*Impact of climate change on the distribution of plankton functional and phylogenetic diversity*" (PlankDiv), held in March 2016 in Villefranche-sur-Mer, France, gathered experts in climate change ecology, species distribution modeling, plankton biology, as well as genomics and evolution. They identified five fundamental questions in future plankton diversity and macroecology research: (1) What can we learn about plankton communities from the new wealth of high-throughput "omics" data? (2) What is the link between plankton diversity and ecosystem function? (3) How can species distribution models be adapted to represent plankton biogeography? (4) How will plankton biogeography be altered due to anthropogenic climate change? and (5) Can a new unifying theory of macroecology be developed based on plankton ecology studies? These questions, along with their associated challenges, are the subject of this review.

THE NEW WEALTH OF PLANKTON DATA

Several recent circumpolar missions have ushered in a new era of plankton biogeography research at the planetary scale. This recent explosion of biological data is perhaps best exemplified by the output of the *Tara Oceans* expedition (Karsenti et al., 2011). While still only offering a temporal snapshot of marine communities, the 7.2 Terabites of metagenomic data gathered are a 1,000 times that generated by the previous largest marine data project, the *Sorcerer* II Global Ocean Sampling (Rusch et al., 2007). High-throughput omics data offer great potential to reveal the global structure of transient marine planktonic ecosystems, since genetic methods compare favorably to traditional observational methods such as microscopy or flow cytometry in terms of the time expenditure, expert knowledge required to identify organisms, and the cost of equipment and analysis. The growing spatial coverage of data enables researchers to estimate global-scale taxonomic diversity of unicellular eukaryotes (de Vargas et al., 2015), to identify the main environmental drivers of community structure in marine prokaryotes (Sunagawa et al., 2015), and to delve into the complexity of biotic interactions between plankton species spanning multiple domains of life, as well as their link to global biogeochemical cycling (Lima-Mendez et al., 2015; Guidi et al., 2016). Complementary to a "bulk" screening of marine biodiversity, single-cell genomics approaches allow matching of phenotype and genotype, and have been used to investigate the phylogenetic affinities of microbial dark matter (i.e., currently unculturable microbial organisms; Rinke et al., 2013; Hug et al., 2016) and to uncover niche partitioning within globally distributed lineages of marine microbes (Kashtan et al., 2014). In combination, bulk and targeted approaches could unravel the taxonomic composition of planktonic organisms, as well as aspects of their ecological function (Thrash et al., 2014; Louca

FIGURE 1 | The staggering wealth of plankton species. Diverse assemblages consist of uni- and multicellular organisms with different sizes, morphologies, feeding strategies, ecological functions, life cycle characteristics, and environmental sensitivities. Courtesy of Christian Sardet, from "Plankton—Wonders of the Drifting World" Univ Chicago Press 2015.

et al., 2016) and genome evolution to new environments (Mock et al., 2017).

Both approaches are challenged by the lack of high quality reference databases (Sunagawa et al., 2015). This highlights the need for comprehensive reference databases to guide the validation and integration of the streams of new data, and their comparison with taxonomic information (e.g., Buitenhuis et al., 2013). In addition, genomic sampling often results in temporal snapshots of one particular aspect of biodiversity [e.g., ribosomal-RNA based Operational Taxonomic Unit (OTU) richness]. Applying this approach to marine plankton communities at similarly broad geographic scales is difficult and expensive, but necessary to improve the assessments of the temporal variability of plankton diversity (Lewandowska et al., 2014). Currently, high-resolution time-series datasets are often restricted to easily-accessible, mostly coastal locations, making extrapolation to the expanses of the open ocean difficult. Therefore, the use of these data for ecological purposes may not be straightforward, especially when trying to estimate abundances of planktonic organisms from metabarcoding (e.g., Decelle et al., 2014).

While the genomic quantification of species composition has become more and more common (Bik et al., 2012; Bik, 2014), and harbors potential for marine ecosystem monitoring in times of rapid environmental and ecosystem change, the link between the identity and the functional role of species remains obscure. Genomic approaches can provide thousands of OTUs, whose metabolic state, morphology, and environmental tolerances are largely unknown. Supplementary measurements of functional traits in laboratory experiments and the quantification of spatio-temporal variability across populations is severely limited by our success in culturing the large diversity of plankton *in vitro*. Estimates that <30% of plankton are cultivable highlight the daunting task of obtaining such data across the heterogeneous plankton lineages and put alternatives, such as single-cell screens, metatranscriptomic approaches, or *in silico* method developments, to the forefront for the characterization of at least some aspects of plankton diversity.

ASSESSING FUNCTIONAL AND PHYLOGENETIC FACETS OF PLANKTON BIODIVERSITY

Traditional approaches have determined marine biodiversity using species occurrence or abundance information at the regional to global scale (e.g., Tittensor et al., 2010). However, there is a growing consensus about the need to assess other facets of biodiversity such as functional diversity, which accounts for biological traits, and phylogenetic diversity to link environmental changes, ecosystem composition and ecosystem function (Naeem et al., 2012; Mouillot et al., 2013). These two promising concepts developed for macro-organisms should be increasingly used within the marine and climate change contexts to further improve our understanding of the link between plankton

diversity, ecosystem productivity, or additional functions related to global biogeochemical cycles.

Functional diversity uses a set of complementary indices (Mouillot et al., 2013) combining measures of species abundance with selected physiological and ecological traits suggested to reflect the fitness of an organism, and which may influence ecosystem function (Violle et al., 2007). Since certain traits may occur across species pertaining to different taxa, estimates of functional diversity allow for the comparison of assemblages with little (no) taxonomic or phylogenetic overlap, but with similar responses to their environment. This metric can account for the intraspecific variability of ecological strategies (e.g., the trophic status of mixotrophic species), and it can include a diverse range of trait variables (e.g., size, feeding strategy, nutrient uptake kinetics). Although much progress has been made in understanding which characteristics of plankton determine their growth, reproduction, and survival (Litchman and Klausmeier, 2008; Litchman et al., 2013; Benedetti et al., 2016), information on traits is restricted to a few well-studied species (Barton et al., 2013). Consequently, trait choice often depends on subjective criteria such as the availability of data (Petchey and Gaston, 2006), therefore open access trait databases should be developed for marine species (Costello et al., 2015). In addition, it is challenging to measure multiple functional traits of thousands of species. Although omics data could allow identifying traits at the community level (Louca et al., 2016), more research is still needed to assign functional traits to sequences, especially for eukaryotic plankton. Despite these methodological issues, trait-based approach of marine communities opens new opportunities for a better understanding of ecosystem functioning and for the development of ecological indicators (Beauchard et al., 2017).

An alternative approach relies on the interspecific phylogenetic differences as a proxy for the overall diversity of a system, assuming that biological characteristics linked to individual fitness and ecological roles show phylogenetic conservatism, i.e., that communities consisting of species with a lower degree of relatedness differ more in their respective trait values, and are thus more diverse (Mouquet et al., 2012). Phylogenetic diversity indices (Tucker et al., 2016) measure the breadth and distribution of evolutionary history present in an assemblage (Mouquet et al., 2012; Cadotte et al., 2013), using DNA sequences to assess the phylogenetic distances between species, by aligning sequences to a reference tree, or by *de-novo* building of phylogenetic trees (Hinchliff et al., 2015).

With the advent of metagenomic data, these promising approaches need to be further explored in terms of their applicability to and relevance for the description of marine ecosystem function. However, the use of phylogenetic diversity critically depends on methodological advances: a substantial fraction of high-throughput sequences obtained by second generation sequencing for microbial communities may still lack sufficient phylogenetic information to provide a reliable phylogenetic placement. In the near future, the popularization of third generation sequencing (e.g., PacBio, Nanopore), which sequences single molecules of DNA in real time, may circumvent this problem, and will provide full opportunities to use phylogenetic diversity estimates to study present and future ecosystem function.

SPECIES DISTRIBUTION MODELING—RUNNING BEFORE WE CAN WALK?

Species Distribution Models (SDMs) are statistical tools that model a species realized niche, i.e., the environmental conditions under which the species can maintain a viable population (Hutchinson, 1957), by relating their occurrence or abundance to environmental conditions (Guisan and Zimmermann, 2000). Several key ecological attributes make planktonic species particularly well-suited for SDMs (Robinson et al., 2011): (i) their distribution reflects their environmental preferences, since plankton are short-lived organisms, with population dynamics tightly connected to climate (Sunday et al., 2012); (ii) plankton are less commercially exploited than other marine species, and thus, their spatial patterns are less biased by captures as in the case of many fish and shellfish species. These attributes make them a key group for monitoring the impacts of climate change on biodiversity and ecosystem functioning (Richardson, 2008). So far, SDMs have seldom been applied to study plankton biogeography, with only a handful of studies on phytoplankton (Irwin et al., 2012; Pinkernell and Beszteri, 2014; Brun et al., 2015; Rivero-Calle et al., 2015; Barton et al., 2016) and some more on zooplankton (e.g., Reygondeau and Beaugrand, 2011; Chust et al., 2014b; Villarino et al., 2015; Brun et al., 2016; Benedetti et al., in press). This is due not only to the limited data availability for model development, but also due to several unaddressed methodological issues.

In plankton, a major problem with SDMs is the scarcity of occurrence data, which can lead to an incomplete niche description and/or biased models. A major challenge is therefore to discern biological distribution patterns from patterns of sampling effort, especially in traditional taxonomy-based plankton data sets where reliable absences data are usually unavailable and large regions, such as the South Pacific, are chronically undersampled. Using one of the most extensive plankton data sets to date, the North Atlantic Continuous Plankton Recorder data, Brun et al. (2016) found that a suite of commonly used SDMs are unable to predict and hindcast the distribution of zooplankton and phytoplankton example-species on the decadal scale. One way to improve SDMs is either through careful methodological adjustments, such as a targeted selection of the background (Phillips et al., 2009), the reduction of environmental predictors, and model complexity (Merow et al., 2014). Another approach could be to merge existing data archives and to combine genomic data with traditional approaches in order to reduce the sampling bias. However, since SDMs apply at the species level, this will require specific identifications, either from microscopy, imaging, or sequencing, which would necessitate to keep taxonomic expertise in our laboratories and, in parallel, to develop specific tools for automatic identification.

In their basic form and most common use, classical SDMs do not account generally for three major ecological processes that may be crucial for plankton distribution: (i) the role of dispersal and its limitation, (ii) biotic interactions, and (iii) intraspecific variability, which we discuss below. The relative importance of these processes in shaping planktonic species' ranges is still being under debate (Cermeño and Falkowski, 2009; Chust et al., 2013).

Plankton dispersal is controlled by ocean currents and can impact diversity and community structure (Lévy et al., 2014). Although barriers to dispersal are fewer in the marine realm compared to the terrestrial one (Steele, 1991), coupling ocean connectivity patterns (Treml et al., 2008; Foltête et al., 2012) with niche models is likely important. Source-sink dynamics may arise frequently because of the advection of water masses (e.g., Beaugrand et al., 2007; Villar et al., 2015) that can introduce species to unsuitable regions (Pulliam, 2000), potentially biasing SDMs. Future developments for plankton could ensue from graph-based techniques (Dale and Fortin, 2010) and from SDMs coupling with simple dispersal models (Foltête et al., 2012; Zurell et al., 2016).

Furthermore, the need to account for biotic interactions when predicting species distributions has been advocated (Boulangeat et al., 2012; Wisz et al., 2013). Recently, the exploration of the plankton "*interactome*" (Lima-Mendez et al., 2015) allowed to describe how biotic interactions occur across trophic levels and relate to environmental conditions and ecosystem functioning, with many new symbiotic interactions identified (Guidi et al., 2016). When prior knowledge is too limited, food-web models could be inferred from simple size-based, or multi-traits assumptions (Albouy et al., 2014), or based on ecosystem models (e.g., Follows et al., 2007; Le Quéré et al., 2016) in combination with satellite estimates of (phyto)plankton community composition (e.g., Hirata et al., 2011).

Finally, SDMs do not consider intraspecific variability, thus assuming that genetic adaptation is negligible. However, many planktonic species exhibit local adaptation (Peijnenburg and Goetze, 2013; Sjöqvist et al., 2015) or consist of several ecotypes with different environmental preferences, and phenotypic plasticity, dispersal, and evolutionary changes could mitigate climate change impacts as they could help species to adapt to changing conditions (O'Connor et al., 2012). One possibility to account for both local adaptation and phenotypic plasticity is to include a population-dependent component in mixed effect models (e.g., Valladares et al., 2014). Furthermore, the joint use of genomic and taxonomic information may help to constrain the differences between subpopulations or ecotypes of a species, and to identify so-called cryptic species.

ADRIFT IN AN OCEAN OF CHANGE

In contrast to works on higher trophic levels (e.g., Cheung et al., 2009), the investigation of the response of plankton to future climate changes has mostly focused more on bulk variables (e.g., biomass, production), with large uncertainties associated with the simulated response of primary and secondary production (e.g., Bopp et al., 2013; Laufkötter et al., 2015). Yet, observational evidence of changes in planktonic ecosystems has been accumulating over the past decades, with ongoing efforts to attribute these changes to specific environmental drivers (e.g., Beaugrand et al., 2008; Rivero-Calle et al., 2015).

SDMs have been used to support observations of poleward plankton distribution range shifts in response to global warming in the North Atlantic (Beaugrand et al., 2002; Richardson, 2008), as well as changes in the relative abundance of certain groups (Rivero-Calle et al., 2015). However, range shifts and in particular phenological changes can vary according to region and species, leading to unexpected emergent patterns (Richardson et al., 2012; Poloczanska et al., 2013; Burrows et al., 2014; Barton et al., 2015). In fact, multiple non-exclusive and interlinked adaptation strategies at the organismal level may all operate in concert, or, alternatively, the selection of one strategy may reduce the necessity to employ another. For example, shifts in spatial distribution may preclude the necessity for phenological adjustments in a given species attempting to maintain its thermal niche. Other adaptation strategies involve species plasticity and genetic modification in order to face changing conditions (Lavergne et al., 2010; Dam, 2013), which have been documented for spatially isolated zooplankton (Peijnenburg et al., 2006; Yebra et al., 2011), but could not be confirmed for other species (Provan et al., 2009). Another alternative adaptation strategy is the change in depth-distribution, i.e., the migration to deeper waters in search for cooler temperatures carried out by fishes (Perry et al., 2005).

Given the multitude of adaptation options, future projections of ecosystem change are prone to large uncertainties. Moreover, disentangling the effects of anthropogenic climate change on plankton distribution and phenology shifts from other drivers (e.g., climate variability, population dynamics) is equally challenging (Chust et al., 2014b). In particular, the combination of controlling factors, together with systematic biases in sampling effort can lead to biases in estimated trends. The decomposition of factors using different SDMs can detect the so-called "niche tracking," which is the shift of a species distribution to follow the displacement of their habitat, e.g., poleward shifts (Monahan and Tingley, 2012; Bruge et al., 2016). At the community level, thermal biases between the average thermal affinity of assemblages and local temperature (Stuart-Smith et al., 2015) have to be considered to improve our understanding of the sensitivity of plankton reorganization with warming.

TOWARD A UNIFIED THEORY OF MACROECOLOGY

Predicting how species will respond to global environmental change requires an understanding of the processes generating their current large-scale spatio-temporal patterns of diversity and distribution, which is the essence of macroecology. One such predominant pattern on Earth is the decline in biodiversity of terrestrial and marine macroorganisms from tropical to polar areas (e.g., Tittensor et al., 2010). Hypotheses explaining this pattern often call upon evolutionary history (Mittelbach et al., 2007), diversity-area relations (Rosenzweig,

1995), temperature effects (Allen et al., 2002), or climatic stability (Fraser and Currie, 1996). Although these premises often find empirical support, their testing in the open oceans has been limited. Whereas, zooplankton likely reflect the general latitudinal trend (Beaugrand et al., 2013), bacterioplankton may form seasonal diversity peaks at high (Ladau et al., 2013) and mid (Sunagawa et al., 2015) latitudes, and for phytoplankton the validity of the global pattern itself and the processes that may explain it are still ambiguous (Rodríguez-Ramos et al., 2015; O'Brien et al., 2016). To alleviate data scarcity, which may have contributed to uncertainty, we suggest the implementation of SDMs as strategic tools to integrate novel with traditional data and to depict aspects of global diversity variation across major taxa and spatio-temporal scales.

The validity of the concept of SDM in plankton and its specific adaptation warrant further testing of the processes that determine plankton distribution, abundance, community assembly, and the maintenance of diversity at local to global scales. More than a decade after the appearance of the unified neutral theory of biodiversity (Hubbell, 2001), there is still an active debate on the relative contribution of demographic stochasticity, dispersal, and niche processes on plankton communities (Pueyo, 2006a,b; Cermeño and Falkowski, 2009; Chust et al., 2013), which promoted the revisiting of the "*Paradox of the Plankton*" (Hutchinson, 1961). Recent studies have tried to reconcile neutral and niche theories (Adler et al., 2007) and suggest that neutral combined with metabolic theory can explain macroecological patterns (Tittensor and Worm, 2016). Furthermore, neutral processes might similarly shape both population genetics and community patterns in plankton (Chust et al., 2016). The combination of data from time-series, global *in situ* observations and experiments on marine plankton provides a unique opportunity to characterize the niches of species (Brun et al., 2015) and to explore the relations between ecological niche characteristics (e.g., niche dissimilarity) and local species richness.

Thus, important open questions include: Is plankton community assembly mainly driven by niche assembly or neutral processes? Does this depend on the spatio-temporal scale of observation? Which method(s) can be used to disentangle the dominating process in community assembly and ecosystem structure? What will be the effect of the removal of geographical barriers that have long separated the Earth's biogeographical provinces on marine plankton diversity ("homogocene," Rosenzweig, 2001)? How does the evolution of microorganism dependency based on gene loss shape the structure and dynamics of communities (Mas et al., 2016)? Due to their fast duplication rates and rapid response to environmental conditions, planktonic communities assemble, dismantle, and re-assemble constantly in natural environments, thus tracking environmental disturbances. Therefore, they are optimally suited to test classical ecological theories established for terrestrial ecosystems, and to answer questions related to diversity-stability relationships, the area-diversity hypothesis, or food web interactions.

CONCLUSION

Plankton ecology research stands at a crossroads. The staggering increase in the wealth of plankton observation data coincides with a time of significant advances in marine ecosystem modeling, which allow, for the first time, the testing of important theories of macroecology in the marine realm. These achievements offer great promise to shed light on marine ecosystem functioning and ecosystem service provision within the context of global climate change. To unlock their potential, we identified a strong need for concomitant developments in the field of bioinformatics and biostatistics, ecological niche modeling, and genetic reference database assembly, thus allowing for a successful integration of these novel with traditional observations, including taxonomic expertise. Paired with the rigorous verification of new and existing macro-ecological theories in the marine realm, and the testing and application of novel biodiversity metrics that better link ecosystem composition to ecosystem function and ecosystem service provision, these theoretical and empirical advances may allow for the urgently needed quantification of potential impacts of climate change on marine ecosystems and feedbacks to higher trophic levels. Due to the complexity of the task, and the scarcity of observational evidence of these transient ecosystems, we conclude that inter-disciplinary, collaborative efforts between experts focussing on different aspects of plankton ecology will be critical in mediating this process.

AUTHOR CONTRIBUTIONS

GC, MV, FB, TN, SV, AA, SMV, DR, JI, and SA conceived and wrote the main manuscript text. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

This research was funded by the EuroMarine Network (http://www.euromarinenetwork.eu), through the organization of the PlankDiv EuroMarine Foresight workshop, held at the Observatoire Océanographique de Villefranche-sur-mer, Villefranche-sur-mer, France, in March 2016, and cofounded by the Basque Government (Department Deputy of Agriculture, Fishing and Food Policy). The PlankDiv workshop was also supported by the Laboratoire d'Océanographie de Villefranche-sur-mer (LOV, UPMC/CNRS), the PlankMed action of WP5 MERMEX/MISTRAL, and by the French national programme EC2CO-LEFE (FunOmics project). This is contribution 810 from AZTI Marine Research Division. We thank the editor and three reviewers for their insightful comments, which greatly improved the manuscript.

REFERENCES

Adler, P. B., Hillerislambers, J., and Levine, J. M. (2007). A niche for neutrality. *Ecol. Lett.* 10, 95–104. doi: 10.1111/j.1461-0248.2006.00996.x

Albouy, C., Velez, L., Coll, M., Colloca, F., Le Loc'h, F., Mouillot, D., et al. (2014). From projected species distribution to food-web structure under climate change. *Glob. Change Biol.* 20, 730–741. doi: 10.1111/gcb.12467

Allen, A. P., Brown, J. H., and Gillooly, J. F. (2002). Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. *Science* 297, 1545–1548. doi: 10.1126/science.1072380

Barton, A. D., Irwin, A. J., Finkel, Z. V., and Stock, C. A. (2016). Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proc. Natl. Acad. Sci.* 113, 2964–2969. doi: 10.1073/pnas.1519080113

Barton, A. D., Lozier, M. S., and Williams, R. G. (2015). Physical controls of variability in North Atlantic phytoplankton communities. *Limnol. Oceanogr.* 60, 181–197. doi: 10.1002/lno.10011

Barton, A. D., Pershing, A. J., Litchman, E., Record, N. R., Edwards, K. F., Finkel, Z. V., et al. (2013). The biogeography of marine plankton traits. *Ecol. Lett.* 16, 522–534. doi: 10.1111/ele.12063

Beauchard, O., Veríssimo, H., Queirós, A., and Herman, P. (2017). The use of multiple biological traits in marine community ecology and its potential in ecological indicator development. *Ecol. Indic.* 76, 81–96. doi: 10.1016/j.ecolind.2017.01.011

Beaugrand, G., Edwards, M., Brander, K., Luczak, C., and Ibanez, F. (2008). Causes and projections of abrupt climate-driven ecosystem shifts in the North Atlantic. *Ecol. Lett.* 11, 1157–1168. doi: 10.1111/j.1461-0248.2008.01218.x

Beaugrand, G., Lindley, J. A., Helaouet, P., and Bonnet, D. (2007). Macroecological study of *Centropages typius* in the North Atlantic Ocean. *Prog. Oceanogr.* 72, 259–273. doi: 10.1016/j.pocean.2007.01.002

Beaugrand, G., Mackas, D., and Goberville, E. (2013). Applying the concept of the ecological niche and a macroecological approach to understand how climate influences zooplankton: advantages, assumptions, limitations and requirements. *Prog. Oceanogr.* 111, 75–90. doi: 10.1016/j.pocean.2012.11.002

Beaugrand, G., Reid, P. C., Ibañez, F., Lindley, J. A., and Edwards, M. (2002). Reorganisation of North Atlantic marine copepod biodiversity and climate. *Science* 296, 1692–1694. doi: 10.1126/science.1071329

Benedetti, F., Gasparini, S., and Ayata, S.-D. (2016). Identifying copepod functional groups from species functional traits. *J. Plankton Res.* 38, 159–166. doi: 10.1093/plankt/fbv096

Benedetti, F., Guilhaumon, F., Adloff, F., and Ayata, S.-D. (in press). Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea. *Ecography*. doi: 10.1111/ecog.02434

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., et al. (2013). Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* 10, 6225–6245. doi: 10.5194/bg-10-6225-2013

Bik, H. M. (2014). Deciphering diversity and ecological function from marine metagenomes. *Biol. Bull.* 227, 107–116. doi: 10.1086/BBLv227n2p107

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* 27, 233–243. doi: 10.1016/j.tree.2011.11.010

Borja, A., Elliott, M., Carstensen, J., Heiskanen, A. S., and van de Bund, W. (2010). Marine management–towards an integrated implementation of the European marine strategy framework and the water framework directives. *Mar. Pollut. Bull.* 60, 2175–2186. doi: 10.1016/j.marpolbul.2010.09.026

Boulangeat, I., Gravel, D., and Thuiller, W. (2012). Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances, *Ecol. Lett.* 15, 584–593. doi: 10.1111/j.1461-0248.2012.01772.x

Bruge, A., Alvarez, P., Fontán, A., Cotano, U., and Chust, G. (2016). Thermal niche tracking and future distribution of Atlantic mackerel spawning in response to ocean warming. *Front. Mar. Sci.* 3:86. doi: 10.3389/fmars.2016.00086

Brun, P., Kiørboe, T., Licandro, P., and Payne, M. R. (2016). The predictive skill of species distribution models for plankton in a changing climate. *Global Change Biol.* 22, 3170–3181. doi: 10.1111/gcb.13274

Brun, P., Vogt, M., Payne, M. R., Gruber, N., O'Brien, C., Buitenhuis, E. T., et al. (2015). Ecological niches of open ocean phytoplankton. *Limnol. Oceanogr.* 60, 1020–1038. doi: 10.1002/lno.10074

Buitenhuis, E. T., Hashioka, T., and Quéré, C. L. (2013). Combined constraints on global ocean primary production using observations and models *Glob. Biogeochem. Cycles* 27, 847–858. doi: 10.1002/gbc.20074

Burrows, M. T., Schoeman, D. S., Richardson, A. J., Molinos, J. G., Hoffmann, A., Buckley, L. B., et al. (2014). Geographical limits to species-range shifts are suggested by climate velocity. *Nature* 507, 492–495. doi: 10.1038/nature12976

Cadotte, M., Albert, C. H., and Walker, S. C. (2013). The ecology of differences: assessing community assembly with trait and evolutionary distances. *Ecol. Lett.* 16, 1234–1244. doi: 10.1111/ele.12161

Cermeño, P., and Falkowski, P. G. (2009). Controls on diatom biogeography in the ocean. *Science* 325, 1539–1541. doi: 10.1126/science.1174159

Cheung, W. W., Lam, V. W., Sarmiento, J. L., Kearney, K., Watson, R., and Pauly, D. (2009). Projecting global marine biodiversity impacts under climate change scenarios. *Fish Fish.* 10, 235–251. doi: 10.1111/j.1467-2979.2008.00315.x

Cheung, W. W., Sarmiento, J. L., Dunne, J., Frölicher, T. L., Lam, V. W., Palomares, M. D., et al. (2013). Shrinking of fishes exacerbates impacts of global ocean changes on marine ecosystems. *Nat. Clim. Change* 3, 254–258. doi: 10.1038/nclimate1691

Chust, G., Allen, J. I., Bopp, L., Schrum, C., Holt, J., Tsiaras, K., et al. (2014a). Biomass changes and trophic amplification of plankton in a warmer ocean. *Glob. Change Biol.* 20, 2124–2139. doi: 10.1111/gcb.12562

Chust, G., Castellani, C., Licandro, P., Ibaibarriaga, L., Sagarminaga, Y., and Irigoien, X. (2014b). Are Calanus spp. shifting poleward in the North Atlantic? a habitat modelling approach. *ICES J. Mar. Sci.* 71, 241–253. doi: 10.1093/icesjms/fst147

Chust, G., Villarino, E., Chenuil, A., Irigoien, X., Bizsel, N., Bode, A., et al. (2016). Dispersal similarly shapes both population genetics and community patterns in the marine realm. *Sci. Rep.* 6:28730. doi: 10.1038/srep28730

Chust, G., Irigoien, X., Chave, J., and Harris, R. P. (2013). Latitudinal phytoplankton distribution and the neutral theory of biodiversity. *Glob. Ecol. Biogeogr.* 22, 531–543. doi: 10.1111/geb.12016

Costello, M. J., Claus, S., Dekeyzer, S., Vandepitte, L., Tuama, É. Ó., Lear, D., et al. (2015). Biological and ecological traits of marine species. *Peer J.*3:e1201. doi: 10.7717/peerj.1201

Dale, M., and Fortin, M.-J. (2010). From graphs to spatial graphs. *Annu. Rev. Ecol. Evol. Syst.* 41, 21–38. doi: 10.1146/annurev-ecolsys-102209-144718

Dam, H. G. (2013). Evolutionary adaptation of marine zooplankton to global change. *Annu. Rev. Mar. Sci.* 5, 349–370. doi: 10.1146/annurev-marine-121211-172229

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* 348:1261605. doi: 10.1126/science.1261605

Decelle, J., Romac, S., Sasaki, E., Not, F., and Mahé, F. (2014). Intracellular Diversity of the V4 and V9 Regions of the 18S rRNA in marine protists (Radiolarians) assessed by high-throughput sequencing *PLoS ONE* 9, 1–8. doi: 10.1371/journal.pone.0104297

Doney, S. C., Ruckelshaus, M., Duffy, J. E., Barry, J. P., Chan, F., English, C. A., et al. (2012). Climate change impacts on marine ecosystems. *Annu. Rev. Mar. Sci.* 4, 11–37. doi: 10.1146/annurev-marine-041911-111611

Duarte, C. M. (2015). Seafaring in the 21St Century: The malaspina 2010 circumnavigation expedition. *Limnol. Oceanogr. Bull.* 24, 11–14. doi: 10.1002/lob.10008

Duarte, C. M. R. W., Fulweiler, C. E., Lovelock, J. M., Pandolfi, P., Martinetto, M. I., Saunders, M. I., et al. (2015). Ocean calamities: delineating the boundaries between scientific evidence and belief. *Bioscience* 65, 746–747. doi: 10.1093/biosci/biv088

Edwards, M., and Richardson, A. J. (2004). Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature* 430, 881–884. doi: 10.1038/nature02808

Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W. (2007). Emergent biogeography of microbial communities in a model ocean. *Science* 315, 1843–1846. doi: 10.1126/science.1138544

Foltête, J. C., Clauzel, C., Vuidel, G., and Tournant, P. (2012). Integrating graph-based connectivity metrics into species distribution models. *Landscape Ecol.* 4, 557–569. doi: 10.1007/s10980-012-9709-4

Fraser, R. H., and Currie, D. J. (1996). The species richness-energy hypothesis in a system where historical factors are thought to prevail: coral reefs. *Am. Nat.* 148, 138–159. doi: 10.1086/285915

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465–470. doi: 10.1038/nature16942

Guisan, A., and Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecol. Model.* 135, 147–186. doi: 10.1016/S0304-3800(00)00354-9

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., et al. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl. Acad. Sci U.S.A.* 112, 12764–12769. doi: 10.1073/pnas.1423041112

Hirata, T., Hardman-Mountford, N. J., Brewin, R. J. W., Aiken, J., Barlow, R., Suzuki, K., et al. (2011). Synoptic relationships between surface Chlorophyll-*a* and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences* 8, 311–327. doi: 10.5194/bg-8-311-2011

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, NJ: Princeton University Press.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1:16048. doi: 10.1038/nmicrobiol.2016.48

Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harbor Symp. Quant. Biol.* 22, 415–427. doi: 10.1101/SQB.1957.022.01.039

Hutchinson, G. E. (1961). The paradox of the plankton. *Am. Nat.* 95, 137–146. doi: 10.1086/282171

Irwin, A. J., Nelles, A. M., and Finkel, Z. V. (2012). Phytoplankton niches estimated from field data. *Limnol. Oceanogr.* 57, 787–797. doi: 10.4319/lo.2012.57.3.0787

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., de Vargas, C., Raes, J., et al. (2011). A holistic approach to marine eco-systems biology. *PLoS Biol.* 9:e1001177. doi: 10.1371/journal.pbio.1001177

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 344, 416–420. doi: 10.1126/science.1248575

Ladau, J., Sharpton, T. J., Finucane, M. M., Jospin, G., Kembel, S. W., O'Dwyer, J., et al. (2013). Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* 7, 1669–1677. doi: 10.1038/ismej.2013.37

Laufkötter, C., Vogt, M., Gruber, N., Aita-Noguchi, M., Aumont, O., Bopp, L., et al. (2015). Drivers and uncertainties of future global marine primary production in marine ecosystem models. *Biogeosciences* 12, 6955–6984, doi: 10.5194/bg-12-6955-2015

Lavergne, S., Mouquet, N., Thuiller, W., and Ronce, O. (2010). Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annu. Rev. Ecol. Evol. Syst.* 41, 321–350. doi: 10.1146/annurev-ecolsys-102209-144628

Le Quéré, C., Andrew, R. M., Canadell, J. G., Sitch, S., Korsbakken, J. I., Peters, G. P., et al. (2016). Global carbon budget 2016. *Earth Syst. Sci. Data* 8, 605–649. doi: 10.5194/essd-8-605-2016

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

Lewandowska, A. M., Boyce, D. G., Hofmann, M., Matthiessen, B., Sommer, U., and Worm, B. (2014). Effects of sea surface warming on marine plankton. *Ecol. Lett.* 17, 614–623. doi: 10.1111/ele.12265

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015). Determinants of community structure in the global plankton interactome. *Science* 348:1262073. doi: 10.1126/science.1262073

Litchman, E., and Klausmeier, C. A. (2008). Trait-based community ecology of phytoplankton. *Annu. Rev. Ecol. Evol. Syst.* 39, 615–639. doi: 10.1146/annurev.ecolsys.39.110707.173549

Litchman, E., Ohman, M. D., and Kjorboe, T. (2013). Trait-based approaches to zooplankton communities. *J. Plankton Res.* 35, 473–484. doi: 10.1093/plankt/fbt019

Louca, S., Parfrey, L. W., and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353, 1272–1277. doi: 10.1126/science.aaf4507

Mas, A., Jamshidi, S., Lagadeuc, Y., Eveillard, D., and Vandenkoornhuyse, P., (2016). Beyond the black queen hypothesis. *ISME J.* 10, 2085–2091. doi: 10.1038/ismej.2016.22

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., McMahon, S. M., Normand, S., et al. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography* 37, 1267–1281. doi: 10.1111/ecog.00845

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., et al. (2007). Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol. Lett.* 10, 315–331. doi: 10.1111/j.1461-0248.2007.01020.x

Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., et al. (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* 541, 536–540. doi: 10.1038/nature20803

Monahan, W. B., and Tingley, M. W. (2012). Niche tracking and rapid establishment of distributional equilibrium in the house sparrow show potential responsiveness of species to climate change. *PLoS ONE* 7:e42097. doi: 10.1371/journal.pone.0042097

Mouillot, D., Graham, N. A., Villéger, S., Mason, N. W., and Bellwood, D. R. (2013). A functional approach reveals community responses to disturbances. *Trends Ecol. Evol.* 28, 167–177. doi: 10.1016/j.tree.2012.10.004

Mouquet, N., Devictor, V., Meynard, C. N., Munoz, F., Bersier, L.-F., Chave, J., et al. (2012). Ecophylogenetics: advances and perspectives. *Biol. Rev.* 87, 769–785. doi: 10.1111/j.1469-185X.2012.00224.x

Naeem, S., Duffy, J. E., and Zavaleta, E. (2012). The functions of biological diversity in an age of extinction. *Science* 336, 1401–1406. doi: 10.1126/science.1215855

O'Brien, C. J., Vogt, M., and Gruber, N. (2016). Global coccolithophore diversity: drivers and future change. *Prog. Oceanogr.* 140, 27–42. doi: 10.1016/j.pocean.2015.10.003

O'Connor, M. I., Selig, E. R., Pinsky, M. L., and Altermatt, F. (2012). Toward a conceptual synthesis for climate change responses. *Glob. Ecol. Biogeogr.* 21, 693–703. doi: 10.1111/j.1466-8238.2011.00713.x

Peijnenburg, K. T., and Goetze, E. (2013). High evolutionary potential of marine zooplankton. *Ecol. Evol.* 3, 2765–2781. doi: 10.1002/ece3.644

Peijnenburg, K. T., Fauvelot, C., Breeuwer, J. A., and Menken, S. B. (2006). Spatial and temporal genetic structure of the planktonic Sagitta setosa (Chaetognatha) in European seas as revealed by mitochondrial and nuclear DNA markers. *Mol. Ecol.* 15, 3319–3338. doi: 10.1111/j.1365-294X.2006.03002.x

Perry, A. L., Low, P. J., Ellis, J. R., and Reynolds, J. D. (2005). Climate change and distribution shifts in marine fishes. *Science* 308, 1912–1915. doi: 10.1126/science.1111322

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., et al. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* 2:150023. doi: 10.1038/sdata.2015.23

Petchey, O. L., and Gaston, K. J. (2006). Functional diversity: back to basics and looking forward. *Ecol. Lett.* 9, 741–758. doi: 10.1111/j.1461-0248.2006.00924.x

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197. doi: 10.1890/07-2153.1

Pinkernell, S., and Beszteri, B. (2014). Potential effects of climate change on the distribution range of the main silicate sinker of the Southern Ocean. *Ecol. Evol.* 4, 3147–3161. doi: 10.1002/ece3.1138

Poloczanska, E. S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P., et al. (2013). Global imprint of climate change on marine life. *Nat. Clim. Change* 3, 919–925. doi: 10.1038/nclimate1958

Provan, J., Beatty, G. E., Keating, S. L., Maggs, C. A., and Savidge, G. (2009). High dispersal potential has maintained long-term population stability in the North Atlantic copepod *Calanus finmarchicus*. *Proc. R. Soc. Lond. B Biol. Sci.* 276, 301–307. doi: 10.1098/rspb.2008.1062

Pueyo, S. (2006a). Diversity: between neutrality and structure. *Oikos* 112, 392–405. doi: 10.1111/j.0030-1299.2006.14188.x

Pueyo, S. (2006b). Self-similarity in species–area relationship and in species abundance distribution. *Oikos* 112, 156–162. doi: 10.1111/j.0030-1299.2006.14184.x

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecol. Lett.* 3, 349–361. doi: 10.1046/j.1461-0248.2000.00143.x

Reygondeau, G., and Beaugrand, G. (2011). Future climate-driven shifts in distribution of *Calanus finmarchicus*. *Glob. Change Biol.* 17, 756–766. doi: 10.1111/j.1365-2486.2010.02310.x

Richardson, A. J. (2008). In hot water: zooplankton and climate change. *ICES J. Mar. Sci.* 65, 279–295. doi: 10.1093/icesjms/fsn028

Richardson, A. J., Brown, C. J., Brander, K., Bruno, J. F., Buckley, L., Burrows, M. T., et al. (2012). Climate change and marine life. *Biol. Lett.* 8, 907–909. doi: 10.1098/rsbl.2012.0530

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Rivero-Calle, S., Gnanadesikan, A., Del Castillo, C. E., Balch, W. M., and Guikema, S. D. (2015). Multidecadal increase in North Atlantic coccolithophores and the potential role of rising CO2. *Science* 350, 1533–1537. doi: 10.1126/science.aaa8026

Robinson, L. M., Elith, J., Hobday, A. J., Pearson, R. G., Kendall, B. E., Possingham, H. P., et al. (2011). Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Glob. Ecol. Biogeogr.* 20, 789–802. doi: 10.1111/j.1466-8238.2010.00636.x

Rodríguez-Ramos, T., Maranón, E., and Cermeno, P. (2015). Marine nano- and microphytoplankton diversity: redrawing global patterns from sampling-standardized data. *Glob. Ecol. Biogeogr.* 24, 527–538. doi: 10.1111/geb.12274

Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. Cambridge: Cambridge University Press.

Rosenzweig, M. L. (2001). Four questions: what does the introduction of exotic species do to diversity? *Evol. Ecol. Res.* 3, 361–367.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., et al. (2007). The Sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS Biol.* 5:e77. doi: 10.1371/journal.pbio.0050077

Sjöqvist, C., Godhe, A., Jonsson, P. R., Sundqvist, L., and Kremp, A. (2015). Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea–Baltic Sea salinity gradient. *Mol. Ecol.* 24, 2871–2885. doi: 10.1111/mec.13208

St. John, M. A., Borja, A., Chust, G., Heath, M., Grigorov, I., Mariani, P., et al. (2016). A dark hole in our understanding of marine ecosystems and their services: perspectives from the mesopelagic community. *Front. Mar. Sci.* 3:31. doi: 10.3389/fmars.2016.00031

Steele, J. H. (1991). Can ecological theory cross the land-sea boundary? *J. Theor. Biol.* 153, 425–436. doi: 10.1016/S0022-5193(05)80579-X

Stuart-Smith, R. D., Edgar, G. J., Barrett, N. S., Kininmonth, S. J., and Bates, A. E. (2015). Thermal biases and vulnerability to warming in the world's marine fauna. *Nature* 528, 88–92. doi: 10.1038/nature16144

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348:1261359. doi: 10.1126/science.1261359

Sunday, J. M., Bates, A. E., and Dulvy, N. K. (2012). Thermal tolerance and the global redistribution of animals. *Nat. Clim. Change* 2, 686–690. doi: 10.1038/nclimate1539

Thrash, J. C., Temperton, B., Swan, B. K., Landry, Z. C., Woyke, T., DeLong, E. F., et al. (2014). Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J.* 8, 1440–1451. doi: 10.1038/ismej.2013.243

Tittensor, D. P., and Worm, B. (2016). A neutral-metabolic theory of latitudinal biodiversity. *Glob. Ecol. Biogeogr.* 25, 630–641. doi: 10.1111/geb.12451

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

Treml, E. A., Halpin, P. N., Urban, D. L., and Pratson, L. F. (2008). Modeling population connectivity by ocean currents, a graph-theoretic approach for marine conservation. *Landscape Ecol.* 23, 19–36. doi: 10.1007/s10980-007-9138-y

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., et al. (2016). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biol. Rev. Camb. Philos. Soc.* doi: 10.1111/brv.12252. [Epub ahead of print].

Valladares, F., Matesanz, S., Guilhaumon, F., Araújo, M. B., Balaguer, L., Benito-Garzón, M., et al. (2014). The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. *Ecol. Lett.* 17, 1351–1364. doi: 10.1111/ele.12348

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., et al. (2015). Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* 348:1261447. doi: 10.1126/science.1261447

Villarino, E., Chust, G., Licandro, P., Butenschön, M., Ibaibarriaga, L., Larrañaga, A., et al. (2015). Modelling the future biogeography of North Atlantic zooplankton communities in response to climate change. *Mar. Ecol. Prog. Ser.* 531, 121–142. doi: 10.3354/meps11299

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., et al. (2007). Let the concept of trait be functional! *Oikos* 116, 882–892. doi: 10.1111/j.0030-1299.2007.15559.x

Waters, C. N., Zalasiewicz, J., Summerhayes, C., Barnosky, A. D., Poirier, C., Gałuszka, A., et al. (2016). The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science* 351, 6269. doi: 10.1126/science.aad2622

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., et al. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* 88, 15–30. doi: 10.1111/j.1469-185X.2012.00235.x

Yebra, L., Bonnet, D., Harris, R. P., Lindeque, P. K., and Peijnenburg, K. T. C. A. (2011). Barriers in the pelagic: population structuring of *Calanus helgolandicus* and C. euxinus in European waters. *Mar. Ecol. Prog. Ser.* 428, 135–149. doi: 10.3354/meps09056

Zurell, D., Thuiller, W., Pagel, J., Cabral, J.S., Münkemüller, T., Gravel, D., et al. (2016). Benchmarking novel approaches for modelling species range dynamics. *Glob. Change Biol.* 22, 2651–2664. doi: 10.1111/gcb.13251

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Chust, Vogt, Benedetti, Nakov, Villéger, Aubert, Vallina, Righetti, Not, Biard, Bittner, Benoiston, Guidi, Villarino, Gaborit, Cornils, Buttay, Irisson, Chiarello, Vallim, Blanco-Bercial, Basconi and Ayata. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix D

CV - July 2019

Anne-Sophie Benoiston

Doctorante, 3ème année

UMR 7205 Institut de Systématique, Evolution, Biodiversité (ISYEB)
Muséum National d'Histoire Naturelle
+33 627398327
+33 140794807
anne-sophie.benoiston@mnhn.fr
Naissance: 5 Octobre 1989, France

Expériences de recherche (sélection)

Depuis octobre 2016 — **Thèse de recherche**, *UMR 7205 Institut de Systématique, Evolution, Biodiversité, équipe Atelier de Bioinformatique*, Muséum National d'Histoire Naturelle, Sorbonne Université, Paris, France.
Encadrants : L. Bittner (UMR 7205 Institut de Systématique, Evolution, Biodiversité) et L. Guidi (UMR 7093 Laboratoire d'Océanographie de Villefranche)
Sujet : Méta-omique et méta-données environnementales : vers une nouvelle compréhension de la pompe à carbone biologique
Description: approche basée sur la construction de réseaux de co-occurrence (bactérioplancton) pour explorer les interactions impliquées dans la pompe à carbone biologique dans l'océan oligotrophe

Mars 2016 - Août 2016 — **Stage de Master 2 (5,5 mois)**, *UMR 7138 Evolution Paris Seine, équipe Analyse des données à haut débit en génomique*, Université Pierre & Marie Curie, Paris, France.
Encadrants: L. Bittner (et S.-D. Ayata & L. Guidi, UMR 7093 Laboratoire d'océanographie de Villefranche)
Sujet: Construction de réseaux génomiques pour déchiffrer la pompe à carbone biologique dans l'océan global
Description: construction de réseaux de co-occurrence au sein des communautés planctoniques, corrélation avec les processus biogéochimiques dans l'océan

Janvier 2013 - Juin 2013 — **Stage de Master 2 (6 mois)**, *UMR 5199 de la Préhistoire à l'Actuel : Culture, Environnement et Anthropologie, équipe Anthropologie des populations passées et présentes*, Université de Bordeaux, Bordeaux, France.
Encadrantes: I. Creveoeur et P. Bayle
Sujet: Caractères externes et structure interne des dents des populations humaines mésolithique et néolithique d'El-Barga, Soudan
Description: étude bio-archéologique (comparaison morphologique et métrique) basée sur l'utilisation de techniques d'imagerie 3D (micro-tomographie)

Enseignement

Depuis octobre 2016 — **Chargée de mission d'enseignement (3 ans)** en **biostatistiques** et **bioinformatique** - 192h, *Sorbonne Université*.

Biostatistiques: enseignement en 2ème et 3ème année de Licence Sciences de la Vie, surveillance d'examens, correction de copies.
Unités d'enseignement:
- 2V314 'Mathématiques et statistiques pour la biologie', responsable: Céline Ellien.
- 3V614 'Statistiques et mathématiques pour la biologie', responsable: Dominique Lamy.

Programme : théorie des tests statistiques, échantillonnage et estimation, comparaison de distributions, indépendance de deux variables aléatoires qualitatives, ANOVA, corrélation, régression, tests non paramétriques.

Bioinformatique: enseignement en 2ème année de Licence Sciences de la Vie, évaluation d'examens oraux.
Unité d'enseignement: 2V381 'Bioinformatique: les outils indispensables', responsable: Stéphane Le Crom.
Programme : scripts bash et encadrement de projet.

Encadrement de stage

Avril-Septembre 2018 **Encadrement d'une stagiaire de Master 2 (6 mois)** en apprentissage artificiel
Marie Soret (Master Image et Son pour les Systèmes Intelligents, Sorbonne Université)
Sujet: Identification de prédicteurs impliqués dans la pompe à carbone biologique par l'utilisation d'algorithmes de machine learning
Description: implémentation et test d'algorithmes d'apprentissage artificiel pour prédire l'état de la pompe à carbone biologique à partir de données métagénomiques *(écriture d'un article en cours)*

Formation

2014-2016 **Master de Bioinformatique et Biostatistiques**, *Université de Nantes*, France.

2011-2013 **Master d'Anthropologie biologique et Préhistoire**, *Université de Bordeaux*, France.

2008-2011 **Licence de Biologie et Ecologie**, *Université de Nantes*, France.

Formation doctorale

Juin 2017 **Ecole d'été** 'Marine ecological and environmental genomics' (6 ECTS), *Station Biologique de Roscoff, Université Pierre & Marie Curie*, Roscoff, France.

Décembre 2016 **Unité d'enseignement** 'Cycle des éléments dans l'océan: forçages, perturbations, méthodes' (6 ECTS), Master 'Océanographie et environnement marin' (2ème année), *Université Pierre et Marie Curie*, Paris, France.

Août 2016 **Ecole d'été** en océanographie microbienne : 'Complexity and simplicity in microbial ecology', *Espegrend Marine Biological Station*, Bergen, Norvège.

Compétences techniques

Informatique **Langages de programmation :** R, Bash, AWK, Perl, Python, PHP, C; **Web:** HTML, CSS; **Bases de données:** MySQL, Oracle; **Systèmes d'exploitation:** Linux, OS X; **Système de composition:** LaTeX

Statistiques statistique descriptive, statistique inférentielle, analyse mutlivariée, traitement de données omiques, inférence et analyse de réseaux de co-occurrence, apprentissage artificiel.
Bonne connaissance des packages R : vegan, WGCNA, SpiecEasi, igraph.

Langues Français - langue maternelle
Anglais - courant
Espagnol - notions

Financements obtenus

Octobre 2018 **Bourse de voyage** (300 €) du *GDR Génomique Environnementale* pour assister à la conférence SFecologie2018 à Rennes, France.

Août 2018 **Bourse de voyage** (250 €) de l'*Association Francophone d'Ecologie Microbienne (AFEM)* pour assister au 17ème symposium ISME à Leipzig, Allemagne.

Août 2018 **Bourse de voyage** (300 €) de l'*International Society for Microbial Ecology (ISME)* pour assister au 17ème symposium ISME à Leipzig, Allemagne.

Juin 2017 Sélectionnée pour participer à l'**école d'été 'Marine ecological and environmental genomics'**, *Station Biologique de Roscoff, France*.

Août 2016 Sélectionnée pour participer à l'**école d'été 'Complexity and simplicity in microbial ecology'**, *Espegrend Marine Biological Station, Bergen, Norvège*.

Septembre 2016 Sélectionnée pour un poste **de chargée de mission d'enseignement** en bioinformatique et biostatistiques à Sorbonne Université.

Juillet 2016 Sélectionnée par l'école doctorale 'Complexité du vivant' pour une **bourse doctorale de 3 ans** du Ministère de l'Enseignement supérieur et de la Recherche ($\sim$ 90 k€).

Communications scientifiques

5 articles publiés (dont 2 en premier auteur et 4 évalués par des pairs), 3 communications orales (dont 2 dans une conférence internationale) et 3 posters (dont 1 dans une conférence internationale):

Articles

Publiés

1. Faure E., Not F., **Benoiston A.S.**, Labadie K., Bittner L., Ayata S.D.. Mixotrophic protists display contrasted biogeographies in the global ocean. *The ISME Journal (**IF: 9.52**)*, **jan 2019**. doi: 10.1038/s41396-018-0340-5

2. Caputi L., Carradec Q., Eveillard D., Kirilovsky A., Pelletier E., Pierella Karlusich J.J., Rocha Jimenez Vieira F., Villar E. Chaffron S., Malviya S., Scalco E., Acinas S.G., Alberti A., Aury J.M., **Benoiston A.S.** *et al.*. Community-level responses to iron availability in open ocean planktonic ecosystems. *Global Biogeochemical Cycles (**IF: 4.46**)*, **jan 2019**. doi: 10.1029/2018GB006022

3. **Benoiston A.S.**, Bayle P. et Crevecoeur I.. Biological affinity of the mesolithic and neolithic populations from El-Barga, Sudan: the dental remains. *Nubian Archaeology in the XXIst Century*, **sept 2018**. ISSN: 978-90-429-3672-0 (publication liée à mon stage de Master en anthropologie)

4. **Benoiston A.S.***, Ibarbalz F.*, Bittner L., Guidi L., Jahn O., Dutkiewicz S. et Bowler C. (*co-first authors). The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B (**IF: 5.66**)*, 372(1728):20160397, **sept 2017**. doi: 10.1098/rstb.2016.0397

5. Chust G., Vogt M., Benedetti F., Nakov T., Villéger S., Aubert A., Vallina S.M., Righetti D., Not F., Biard T., Bittner L., **Benoiston A.S.**, Guidi L., Villarino E., Gaborit C. *et al. Mare incognitum*: a glimpse into future plankton diversity and ecology research. *Frontiers in Marine Ecology*, 4, **mar 2017**. doi: 10.3389/fmars.2017.00068

En préparation

Benoiston A.S., Eveillard D., Chaffron S., Delage E., Guidi L. et Bittner L.. The microbial drivers of the biological carbon pump. Soumission prévue dans *The ISME Journal*.

Benoiston A.S., Soret M., Eveillard D., Guidi L. et Bittner L.. Random forest-based estimates of the biological pump processes from metabarcoding data. Soumission prévue dans *Biogeosciences*.

Le nom de l'orateur est souligné dans les deux sections suivantes :

Communications orales

Conférences internationales

Benoiston A.S., Eveillard D., Chaffron S., Jean G., Delage E., Ayata S.D., Bowler C., Guidi L., Bittner L. et les coordinateurs *Tara*. Ecological networks of microbial plankton influence the biological carbon pump processes in the oligotrophic ocean. *SFécologie 2018 (Conférence internationale en écologie), Rennes, France*, **22-25 octobre 2018**.

Faure E., Not, F., **Benoiston A.S.**, Labadie K., Bittner L. et Ayata S.D.. Ubiquitous yet contrasted biogeography of mixotrophic protists in the global ocean. *17th ISME (International Society for Microbial Ecology) symposium, Leipzig, Allemagne*, **12-17 août 2018**.

Conference nationale

Benoiston A.S., Bittner L., Guidi L., Chaffron S., Eveillard D., Ayata S.D., Jean G., Pelletier E., Pesant S., de Vargas C., Karsenti E., Bowler C., Gorsky G., et le consortium *Tara*. Plankton networks correlated to the biological carbon pump in the global ocean. Journée réseaux du GDR Génomique environnementale, *10èmes Journées scientifiques de l'université de Nantes, France*, **2 juin 2017**.

Posters

Conférence internationale

Benoiston A.S., Eveillard D., Chaffron S., Jean G., Delage E., Ayata S.D., Bowler C., Guidi L., Bittner L. et les coordinateurs *Tara*. Biological pump processes are driven by microbial networks in the global oligotrophic ocean. *17ème symposium ISME (International Society for Microbial Ecology), Leipzig, Allemagne*, **12-17 août 2018**.

Conferences nationales

Ayata S.D., Faure E., **Benoiston A.S.**, Da Silva O., Sonnet V., Benedetti F., Not F., Aumont O., Guidi L., Chaffron S., Eveillard D. et Bittner L.. FunOmics: Assessing functional diversity of plankton communities from high throughput -omics data. *Colloque de Bilan et de Prospective du Programme National LEFE, Clermont-Ferrand, France*, **28-30 mars 2018**.

Benoiston A.S., Bittner L., Guidi L., Chaffron S., Eveillard D., Ayata S.D., Pelletier E., Pesant S., de Vargas C., Karsenti E., Bowler C., Gorsky G., et le consortium *Tara*. Des réseaux planctoniques corrélés à l'export de carbone dans l'océan, révélés par des analyses de co-occurrence génomique. 2nd symposium *La microbiologie dans tous ses états*, *Muséum National d'Histoire Naturelle, Paris, France*, **23 mai 2017**.

Bénévolat et médiation scientifique

27 mars 2018 Sensibilisation auprès de collégiens sur le rôle du plancton et les conséquences de l'acidification des océans, *Collège Lamartine, Paris 9ème*.

Octobre 2017 Organisation et animation d'ateliers sur le plancton marin lors de la **Fête de la Science**, *Sorbonne Université*.
Mon travail consistait à contacter des partenaires pour obtenir le matériel nécessaire aux ateliers (réalité virtuelle et augmentée et microscopie), à préparer des sessions d'animation pour les élèves (école élémentaire et lycée), à concevoir des affiches et à animer les ateliers.

Octobre 2016 Animation d'ateliers sur l'évolution et la phylogénie lors de la **Fête de la Science**, *Sorbonne Université*

2011 - 2013 **Secrétaire puis présidente de l'association 'Le Chaînon Manquant'**, *Université de Bordeaux*.
Organisation d'événements culturels (conférences, excursions, ateliers), animation d'ateliers de médiation scientifique

2010 - 2018 Participation à des **fouilles paléontologiques et archéologiques** programmées (une à trois semaines par été).